agriculture

Special Issue Reprint

Digital Innovations in Agriculture

Volume II

Edited by
Gniewko Niedbała and Sebastian Kujawa

mdpi.com/journal/agriculture

MDPI

Digital Innovations in Agriculture —Volume II

Digital Innovations in Agriculture —Volume II

Editors

Gniewko Niedbała
Sebastian Kujawa

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Editors
Gniewko Niedbała
Poznań University of Life Sciences
Poznań, Poland

Sebastian Kujawa
Poznań University of Life Sciences
Poznań, Poland

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Agriculture* (ISSN 2077-0472) (available at: https://www.mdpi.com/journal/agriculture/special_issues/Digital_Innovations_Agriculture).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, A.A.; Lastname, B.B. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

Volume II
ISBN 978-3-0365-8850-6 (Hbk)
ISBN 978-3-0365-8851-3 (PDF)
doi.org/10.3390/books978-3-0365-8851-3

Set
ISBN 978-3-0365-8846-9 (Hbk)
ISBN 978-3-0365-8847-6 (PDF)

Contents

About the Editors

Gniewko Niedbała

Gniewko Niedbała is a Full Professor at the Department of Biosystems Engineering, Poznań University of Life Sciences, Poland. He defended his doctorate in 2006 at the August Cieszkowski Agricultural University of Poznań and habilitated in 2019 at the Poznań University of Life Sciences. Presently, he is working on using artificial neural networks and machine learning in many aspects of agriculture and agronomy. Between 2012 and 2016, he was a Board Member of the National Centre for Research and Development, Poland. He ranked in the World's Top 2% of most cited scientists in 2020. He was a Guest Editor for a Special Issue of Agriculture: "Neural Networks in Agriculture and Digital Innovations in Agriculture". Additionally, he is a Member of the Editorial Board of Agronomy and a Member of the Reviewer Boards of Agriculture, Water, and Land. In addition, he has authored over 150 journal and conference papers and book chapters related to artificial intelligence in agriculture.

Sebastian Kujawa

Sebastian Kujawa is employed as an Assistant Professor in the Department of Biosystems Engineering, Poznań University of Life Sciences, Poland. In 2009, he received his PhD degree from the Poznań University of Life Sciences. His scientific activity concerns the applications of computer image analysis and machine learning in developing methods for condition assessments of dynamic biosystems. Dr. Kujawa is an author of over 50 publications in scientific journals and peer-reviewed conference proceedings. He is a member of the Board of the Polish Society for Information Technology in Agriculture (POLSITA). He has served as a Guest Editor of a Special Issue in Agriculture: "Neural Networks in Agriculture, and Digital Innovations in Agriculture".

agriculture

MDPI

Article

Application of a Fractional Order Differential to the Hyperspectral Inversion of Soil Iron Oxide

Hailong Zhao [1], Shu Gan [1,2,*], Xiping Yuan [2,3], Lin Hu [1], Junjie Wang [1] and Shuai Liu [1]

[1] Faculty of Land Resources Engineering, Kunming University of Science and Technology, Kunming 650093, China
[2] Application Engineering Research Center of Spatial Information Surveying and Mapping Technology in Plateau and Mountainous Areas Set by Universities in Yunnan Province, Kunming 650093, China
[3] College of Geosciences and Engineering, West Yunnan University of Applied Sciences, Dali 671000, China
* Correspondence: gs@kust.edu.cn

Abstract: Iron oxide is the main form of iron present in soils, and its accumulation and migration activities reflect the leaching process and the degree of weathering development of the soil. Therefore, it is important to have information on the iron oxide content of soils. However, due to the overlapping characteristic spectra of iron oxide and organic matter in the visible-near infrared, appropriate spectral transformation methods are important. In this paper, we first used conventional spectral transformation (continuum removal, CR; standard normal variate, SNV; absorbance, log (1/R)), continuous wavelet transform (CWT), and fractional order differential (FOD) transform to process original spectra (OS). Secondly, competitive adaptive reweighted sampling (CARS) was used to extract characteristic wavelengths. Finally, two regression models (backpropagation neural network, BPNN; support vector regression (SVR) were used to predict the content of iron oxide. The results show that the FOD can significantly improve the correlation with iron oxide compared with the CR, SNV, log (1/R) and CWT; the baseline drift and overlapping peaks decrease with increasing the order of FOD; the CARS algorithm based on 50th averaging can select more stable characteristic wavelengths; the FOD achieves better results regardless of the modelling method, and the model based on 0.5-order differential has the best prediction performance (R^2 = 0.851, RMSE = 5.497, RPIQ = 3.686).

Keywords: soil; hyperspectral; iron oxide; spectra transform; fractional order differential

Citation: Zhao, H.; Gan, S.; Yuan, X.; Hu, L.; Wang, J.; Liu, S. Application of a Fractional Order Differential to the Hyperspectral Inversion of Soil Iron Oxide. *Agriculture* **2022**, *12*, 1163. https://doi.org/10.3390/agriculture12081163

Academic Editors: Gniewko Niedbała and Sebastian Kujawa

Received: 3 July 2022
Accepted: 1 August 2022
Published: 5 August 2022

1. Introduction

Iron oxide is the bulk of iron-bearing minerals in soils, mainly formed by the chemical weathering and the redeposition of iron-bearing silicate minerals, and is widely distributed in various types of soils around the world [1]. Due to its high activity, the morphological characteristics of iron oxide are susceptible to various environmental factors, and its aggregation and migration activities reflect the leaching process, the degree of weathering development and the zonation of the soil distribution [2]. The chemical activity of iron oxide allows it to adsorb numerous heavy metals, non-metallic ions and oxygenated anions, which greatly control the concentration, morphology and migration transformation of these elements in the soil, determining plant effectiveness, environmental toxicity, affecting crop yield and quality and human health [3].

The traditional method for the determination of iron oxide content has a high accuracy but a high determination cost and a long cycle time. Hyperspectral, with its high spectral resolution and wavelength continuity [4], is widely used for the inversion of soil physico-chemical properties [5,6]. In practice, however, many factors can affect the quality of the spectra; these include the complexity of the composition of the soil itself, the environment and the noise of the instrument itself. Therefore, suitable spectral pre-processing is an indispensable step in soil hyperspectral modelling to improve the predictive power of the

model [7]. Common spectral pre-processing methods include spectral smoothing (Savitzky–Golay filter, SG) [8], continuum removal (CR) [9], absorbance (log (1/R)), multiple scattering correction (MSC) [10], standard normal variate (SNV) [11], continuous wavelet transform (CWT) [12] and differential transformations [13]. Due to the presence of environmental and instrumental noise, spectral smoothing has become an essential step and other processing methods are based on spectral smoothing afterwards. Of these pre-treatment methods, CR, log (1/R), SNV and CWT have all been widely used. However, as soil spectra are a comprehensive reflection of soil properties, the characteristic wavelengths of iron oxide are not the same in different regions and are easily masked by organic matter. According to previous studies, the characteristic spectra of iron oxide and organic matter often overlap in the visible-near infrared band (400–1000 nm) [14]. Spectral differentials can minimize baseline drift and separate overlapping spectra. Of these, first-order and second-order differentials are two effective methods [15,16]. However, the integer order differential lacks sensitivity to the asymptotic slope or curvature of the spectra, resulting in detailed spectral information not being captured [17]. Fractional order differential (FOD) is an extension of the integer order differential, which allows us to interpolate between the original spectra (OS), the first-order differential spectra and the second-order differential spectra and even higher order differential spectra to obtain fractional order differentials. At present, FOD has been widely used in soil hyperspectral and has achieved good results. Tian et al. [18] collected soils from Xinjiang and determined the total salt content of the soils indoors. Firstly, FOD was used for five transformed spectra, and the bands whose spectra and total salt content passed the 0.01 significance test were extracted as characteristic wavelengths and finally modelled using PLSR. The prediction results showed that the best model prediction ability was obtained based on the model of 1.6-order. Hong et al. [19] collected soil samples and measured the organic matter content in the Jianghan Plain of Wuhan City, Hubei Province, while performing FOD on the original spectra at 0.25-order intervals, and the experimental results showed that the PLS-SVM model constructed based on 1.25-order had the strongest predictive power. However, so far, no studies have been carried out to estimate the iron oxide content using FOD.

Due to the high number of wavelengths in the hyperspectral, the wavelength information contained tends to be more redundant. If the full wavelength band is modeled, it not only increases the running time, but also reduces the accuracy of the model [20]. Therefore, the selection of the characteristic wavelengths before modelling is a very important step. Currently, the selection of characteristic wavelengths is mostly completed using the Pearson correlation analysis [21–23], and correlation coefficients and significance levels reflect the correlation between soil physicochemical properties and wavelength [24]. In addition, the Genetic Algorithm (GA) [25], uninformative variable elimination method (UVE) [26], successive projections algorithm (SPA) [27] and competitive adaptive reweighted sampling (CARS) algorithm [28] are the common methods used for the selection of characteristic wavelengths. The wavelengths obtained using these methods are used as the input variables to the model and the iron oxide content is used as the dependent variable for model construction. There are also many methods of model construction, such as multiple linear regression (MLR), partial least squares regression (PLSR) and principal components regression (PCR), all of which are linear regression methods and are simple to use. Of these, PLSR is the most common regression method. Xiong et al. [16] used PLSR to invert the Fe of soils and achieved a good prediction accuracy. In addition, with the popularity of machine learning, more and more methods such as random forest regression (RFR), support vector regression (SVR) and back propagation neural network (BPNN) have been applied to soil hyperspectral modelling. Qin et al. [29] used RFR to model the inversion of free iron in soil and found that the accuracy of RFR in estimating free iron in soil was significantly better than that of stepwise multiple linear regression.

However, due to the complexity of soils in different regions, there is no universal pre-processing method that is suitable for different regions. Therefore, this paper uses three types of spectral transformation methods, including conventional transform spectra

(CR, log (1/R) and SNV), CWT and FOD to transform the OS. The CARS was used to select characteristic wavelengths. Finally, the model was constructed using BPNN and SVR. Therefore, the objectives of this paper are (a) to explore the model prediction capability of the fractional order differential transformation and to compare it with the conventional spectral transform, the continuous wavelet transform; (b) to assess the capability of CARS in characteristic wavelength selection; and (c) to Compare the predictive power of BPNN and SVR models with different spectral transforms.

2. Materials and Methods

2.1. Study Area

The study area is located in Lufeng County, Chuxiong Yi Autonomous Prefecture, Yunnan Province (24°55′25″~25°22′05″ N, 102°00′00″~102°9′00″ E). The study area is about 6 km wide from east to west, 8 km long from north to south, and 6 km in diameter, covering an area of about 32 km^2, with an overall depression pit with a high elevation around and a low elevation in the middle. The area is a small Mesozoic red sedimentary basin, belonging to the Lower Ordovician Redstone Shale Formation, with a brief lithology of purplish-red and grey–green siltstone. According to two soil surveys in 1982 and 1985, there are five soil types, ten subtypes, twenty genera and forty species of brown soil, yellow–brown soil, red soil, purple soil and rice soil in Lufeng County. The purple soil accounts for 56.9% of the land area and is the most important soil type in the area, followed by red soil, which accounts for 22.8% of the land area, yellow–brown soil, which accounts for 7.8%, and the rice soil, which accounts for 6.3% [30].

2.2. Sample Collection and Data Acquisition

Soil samples were collected at the end of July 2021 from the southern rim of the Dinosaur Valley in Lufeng County, Yi Autonomous Prefecture of Chuxiong, Yunnan Province, and the sampling points were set up according to the difference in topography. Each sample was taken within 5 m × 5 m. Figure 1 shows the location of the sampling points in the study area. Within the sampling area, surface soil was collected from 0 to 20 cm according to the 5-point sampling method, and approximately 1 kg of soil was bagged and stored. The soil types were purple loam, red loam and yellow–brown loam. The collected soil was first cleaned of impurities such as weeds and stones, then naturally air dried and finally ground with an agate ball mill and sieved through 100-mesh. The aperture size of 100-mesh was 0.15 mm. Each sample was split into two, one for the determination of iron oxide content and the other for the measurement of hyperspectral data. The determination of iron oxide in soil was carried out by X-ray fluorescence spectrometry in accordance with the "Methods of Agricultural Chemical Analysis of Soil", taking into account the quality requirements of the samples and other technical specifications, as well as the limits of detection, accuracy and precision of the samples.

Soil spectroscopy was carried out in a dark room with an ASD Field Spec 3 geophysical spectrometer, using a probe with an internal halogen light source, a 21 mm inner probe diameter and a 25° front field of view and a wavelength range of 350–2500 nm. The number of wavelengths obtained by resampling the spectral interval to 1 nm was 2151. For the spectral measurements, the soil samples were placed in a 10 cm wide and 2 cm high container and scraped flat to reduce the effect of the roughness of the soil sample on the spectral measurements. The probe was held at a height of 2 cm from the soil sample and aligned vertically with the sample. Five spectral curves were measured for each sample in the same area. The actual spectral reflectance of the sample was averaged over the five spectral curves.

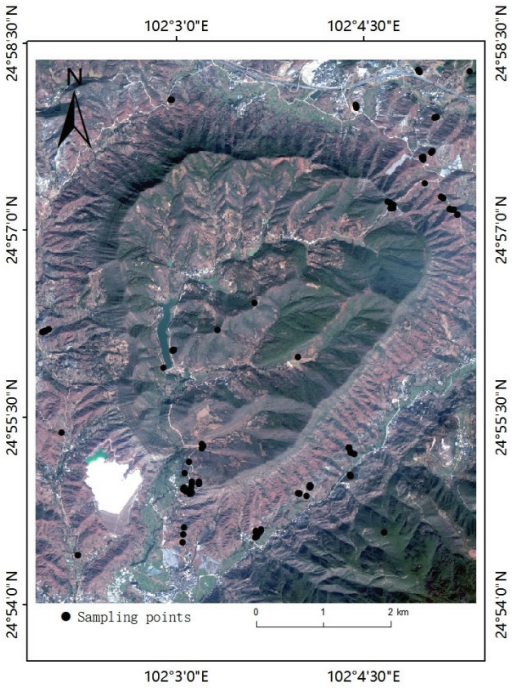


Figure 1. Distribution map of the soil sampling points.

2.3. Data Processing

The signal-to-noise ratio at 350–399 nm and 2451–2500 nm was low due to the influence of the instrument itself, so these two spectral data were removed and a total of 2051 wavelengths were obtained. To eliminate the interference of instrument noise, uneven distribution of soil particles and random factors, the Savitzky–Golay [8] smoothed curve with a window number of 9 and polynomial order of 2 was used as the OS. The CR, log (1/R), SNV and CWT were further applied to the OS. The CR can highlight the absorption and reflection features of the spectra [31]. The log (1/R) can reduce the interference of multiplicative factors caused by light transformation [32]. The CWT, on the other hand, can mine the characteristic information of the spectra at different scales [33].

Fractional order differential is an extension of integer order differential and is commonly known as Riemann–Liouville, Grünwald–Letnikov and Caputo, of which Grünwald–Letnikov is the most commonly used form.

Before giving the definition of Grünwald–Letnikov, let us observe the formula for the first order derivative:

$$\frac{d^1}{dt^1}f(t) = \lim_{h\to 0}\frac{1}{h}[f(t) - f(t-h)] \quad (1)$$

In Equation (1), the h represents the increment of the spectral variable. From the first-order differential, the second-order differential formula can be derived as follows:

$$\frac{d^2}{dt^2}f(t) = \lim_{h\to 0}\frac{1}{h^2}[f(t) - 2f(t-h) + f(t-2h)] \quad (2)$$

Looping the above method, the nth order differential of the function can be derived as follows:

$$\frac{d^n}{dt^n}f(t) = \lim_{h\to 0}\frac{1}{h^n}\sum_{j=1}^{n}(-1)^j\begin{pmatrix} n \\ j \end{pmatrix}f(t-jh) \quad (3)$$

In Equation (3), the j represents the difference between the upper and lower limits of the derivative. Using the Gamma function to replace the binomial coefficients of Equation (3), while extending the integer order to non-integer order, one can then obtain the α-order fractional order differential formula:

$$\frac{d^{\alpha}}{dt^{\alpha}}f(t) = \lim_{h\to 0}\frac{1}{h^{\alpha}}\sum_{j=0}^{[(t-t_0)/h]}\frac{(-1)^j\Gamma(\alpha+1)}{j!\Gamma(\alpha-j+1)}f(t-jh) \quad (4)$$

Since the sampling interval of the spectrum is 1, set h to 1. h represents the differential step, t represents the upper limit of the differential, t_0 represents the lower limit of the differential. The Gamma function is defined as:

$$\Gamma(z) = \int_0^{\infty} e^{-t}t^{z-1}dt = (z-1)! \quad (5)$$

Then, Equation (4) can be converted to:

$$\frac{d^{\alpha}}{dt^{\alpha}}f(t) \approx f(t) + (\alpha)f(t-1) + \frac{(-\alpha)(-\alpha+1)}{2}f(t-2) + \ldots + \frac{\Gamma(-\alpha+1)}{j!\Gamma(-\alpha+j+1)}f(t-j) \quad (6)$$

In Equation (6), α represents the order. α = 0 represents OS; α = 1 represents the first-order differential; α = 2 represents the second-order differential. The implementation of the fractional order differential in this study was implemented using the FOTF toolbox based on MATLAB 2020b [34].

The workflow for data processing is shown in Figure 2.

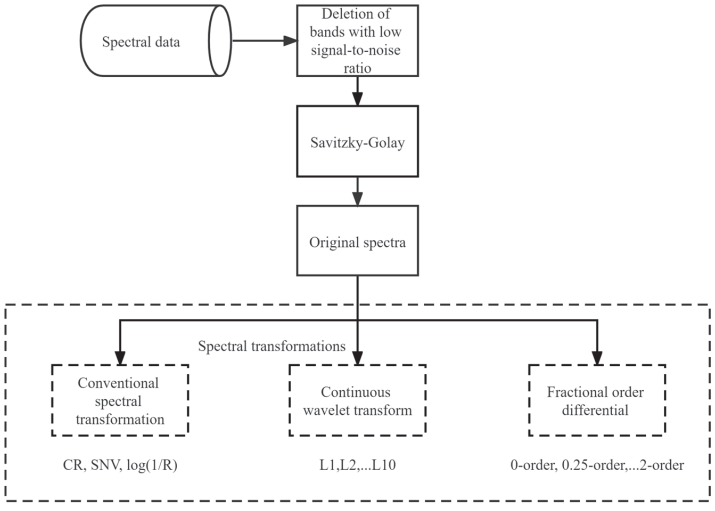


Figure 2. The workflow of data processing.

2.4. Characteristic Wavelength Selection

Due to the large redundancy of hyperspectral data, not all wavelengths are beneficial in improving the modelling accuracy when performing regression analysis. If all wavelengths are modelled and analyzed, it is not only computationally intensive, but also reduces the modelling accuracy. Therefore, characteristic wavelength selection is necessary prior to modelling.

The Competitive Adaptive Reweighted Sampling (CARS) algorithm is a characteristic wavelength selection method based on Monte Carlo sampling and PLS regression coefficients, treating each variable as an individual and selecting the one with the higher adaptive

capacity. The specific steps are: randomly select a fixed rate of samples as the calibration dataset and build a PLS model, then calculate the absolute value of the regression coefficient of the model and the weight corresponding to each wavelength, use the exponential decay function and adaptive reweighted sampling method to select the variables, while calculating the root mean square error of cross-validation, after sub-sampling, select the subset with the smallest root mean square error as the optimal subset of variables [35].

2.5. Support Vector Regression (SVR)

SVR is a non-linear modelling method based on statistical learning theory. Its basic approach is to use the support vectors in the training samples to design an optimal decision boundary to deal with linear and non-linear problems, which performs well especially when dealing with small sample data [36]. In this study, the kernel function of the support vector machine was chosen to be a Gaussian kernel function. To make the model more stable and the results more reliable, the parameters of the support vector machine: c and gamma tuning were performed during the model training process using a genetic algorithm based on differential evolution, which is more robust than the traditional genetic algorithm, has a block convergence speed and has a stronger global optimization search capability. Among the parameters of the differential evolution-based genetic algorithm, the range of values for c and gamma was set between 2^{-8} and 2^{8}, the population size was 50, the coding method was real number coding, the selection method of the basis vector was elite replication selection, the variation operator F was 0.5, the crossover operator CR = 0.5, the maximum number of evolutionary generations was 1000 and the fitness function was the average root mean square error of cross-validation. The algorithm was implemented using the Geatpy [37].

2.6. Back Propagation Neural Network (BPNN)

BPNN is a more widely used artificial neural network, with a strong non-linear processing capability [38]. The main features of BPNN are the forward transmission of input data and the backward propagation of errors. In the forward transfer process, the input data are processed progressively from the input layer through the hidden layer to the output layer. If the error in the output layer is not within the range, back propagation is performed and the weights of each layer are adjusted by gradient descent until the error is within the specified range. In this study, we used a simple three-layer network structure with an input layer, a hidden layer and an output layer, respectively. Transigmoid and purelin were chosen as the transfer functions of the hidden and output layers according to the previous study when building the BPNN prediction model. Besides, Sigmoid and trainlm were chosen as the activation function and training function, respectively [39]. Hidden layer, learning rate and maximum epoch were 8~10, 0.01 and 1000, respectively.

2.7. Model Evaluation Method

The Kennard-Stone (K-S) [40] algorithm was used to classify the calibration dataset and the validation dataset. A total of 70% of the samples were selected as the calibration dataset and the remaining 30% as the validation dataset. Since the soil samples showed non-normal distribution, the ratio of performance to interquartile spacing (RPIQ) could give a more realistic evaluation of the model [41]. The accuracy of the inverse model was therefore measured by three parameters: coefficient of determination (R^2), root mean square error (RMSE) and RPIQ.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - y_i^*)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_i^*)^2}{n}} \quad (8)$$

$$RPIQ = \frac{IQ}{RMSE} \quad (9)$$

where: y_i^* is the predicted of the ith sample; y_i is the measured value of the ith sample; $\overline{y}$ is the mean of the measured values; IQ is the difference between the third quartile and the first quartile of the sample; n is the number of samples. A larger R^2 indicates a more stable model; a smaller RMSE indicates a more accurate model; a larger RPIQ indicates a better predictive power of the model [42]. The performance of the models can be judged as follows:

RPIQ: (1) >2.5, excellent model; (2) 2.0–2.5, very good model with predictive ability; (3) 1.7–2.0, good model; (4) 1.4–1.7, fair model; and (5) <1.4, poor model [43].

3. Results

3.1. Statistical Analysis of Iron Oxide Content

The 135 samples were divided into two groups using the Kennard-Stone algorithm, with 70% being the calibration dataset (n = 95) and 30% being the validation dataset (n = 45). The obtained soil iron oxide content was counted by origin software and the relevant statistical parameters obtained are shown in Table 1. The minimum value of iron oxide in the study area was 18.293 g·kg^{-1} and the maximum value was 66.978 g·kg^{-1}, with a mean value of 41.201 g·kg^{-1} and a coefficient of variation of 28.4%, which is a medium variation. The difference between the mean and standard deviation of the calibration dataset and the validation dataset was not significant, and they can be considered as belonging to the same distribution.

Table 1. Statistical characteristics of iron oxide content.

Sample Classification	Sample Number	Max/(g·kg^{-1})	Min/(g·kg^{-1})	Mean/(g·kg^{-1})	Standard Deviation/(g·kg^{-1})	Coefficient of Variation/%
Total dataset	135	66.978	18.293	41.201	11.698	28.393
Calibration dataset	95	64.808	23.311	42.141	10.736	25.476
Validation dataset	40	66.978	18.293	38.969	13.605	34.912

3.2. Spectral Transformation Methods

3.2.1. Conventional Transform Spectra

The results of the conventional spectral transformation of the original spectra are shown in Figure 3. After the original spectra were transformed by the continuum removal, the spectral curves were normalized to a consistent spectral background and effectively highlighted the absorption features of the spectra [44], with significant iron oxide absorption features at 500 nm and 900 nm, respectively [45]. Meanwhile, near 1400 nm was the band spectra of the -OH, 1900 nm was the band of H_2O, and the absorption feature at 2200 nm was mainly due to the -OH stretching vibration and the AL-OH bending vibration [46,47].

3.2.2. Continuous Wavelet Transform

The Gaussian4 function was selected as the wavelet basis function in this study because the soil spectral curve characteristics were similar to those of the Gaussian function [48]. Based on the calibration dataset, the original spectra were first transformed into corresponding wavelet coefficients (decomposition scales were set to 2^1, 2^2, 2^3, . . . , 2^{10}), and average them. The first scale was denoted as L1, the second scale as L2, and the mth scale as Lm. The results are shown in Figure 4. It can be noticed that the absorption and reflection characteristics increased with increasing the scale in the different wavelength ranges. The wavelet coefficient curves at the L1, L2 and L3 were less distinctive and were approximately straight lines. At L4, L5, L6 and L7, distinct peaks can be observed. At L8,

L9 and L10, convex smooth curves can be observed with a smaller number of peaks. In summary, the CWT can help to highlight features of the spectra and to fully explore subtle spectral features.

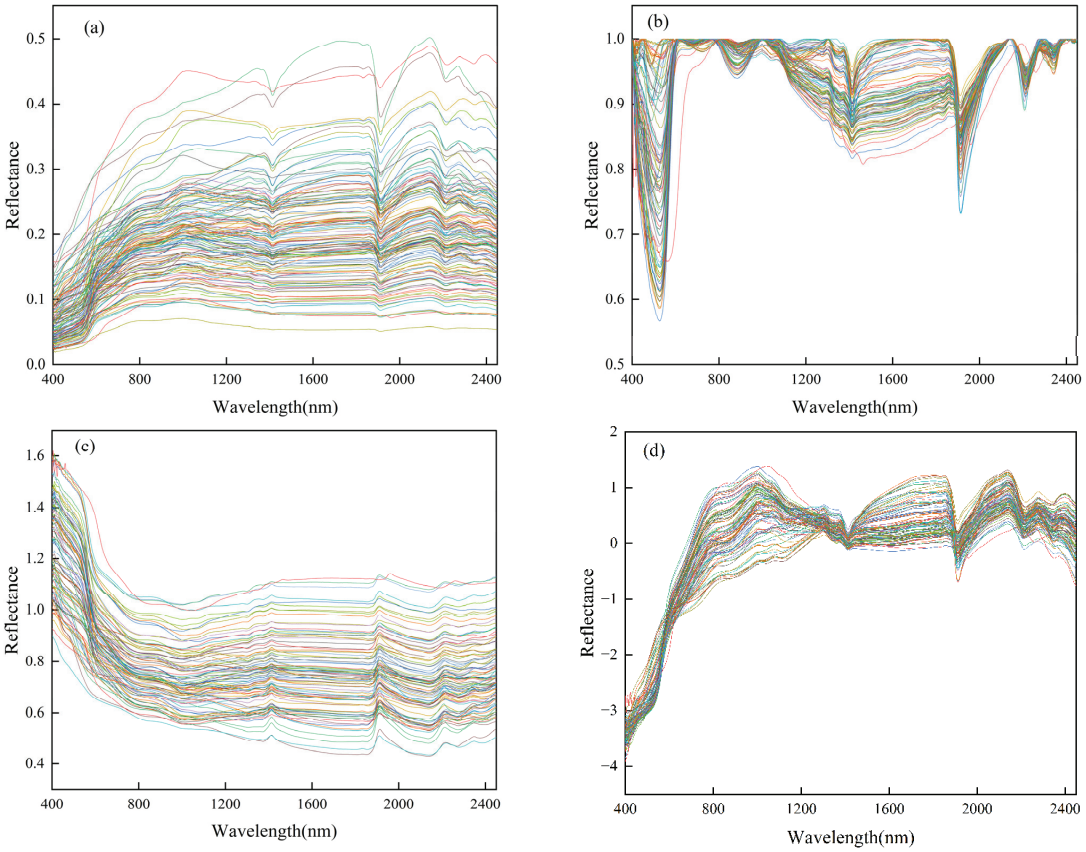


Figure 3. Transformed spectra of the soil samples: (**a**) original spectra. (**b**) continuum removal. (**c**) absorbance. (**d**) SNV transformed reflectance.

3.2.3. Fractional Order Differential

The FOD spectra of the calibration dataset are shown in Figure 5. The absorption features at 1400, 1900 and 2200 nm were more obvious, but the absorption bands were wider and overlapped. When the order was gradually increased from 0 to 1, the differential curve of each order slowly approximated the differential curve of the 1-order, and the sensitivity of the differential result to the slope of the reflectivity curve increased [49]. The three absorption features of the water molecule vibration at 1400, 1900 and 2200 nm became increasingly apparent; at the same time, there were two positive peaks at 420 and 570 nm and one negative peak at 470 nm and the absorption band at 1400 nm changed from one negative peak to one positive and one negative peak. As the order increased, the spectral reflectance values gradually approached 0, which indicates that the baseline drift and overlapping peaks were eliminated [19]. At the same time, the absorption valleys at 1900 and 2200 nm gradually changed to a positive and a negative peak, respectively. Compared to the original spectra, the FOD spectra can show changes in spectral detail and improve the resolution of the spectral curve.

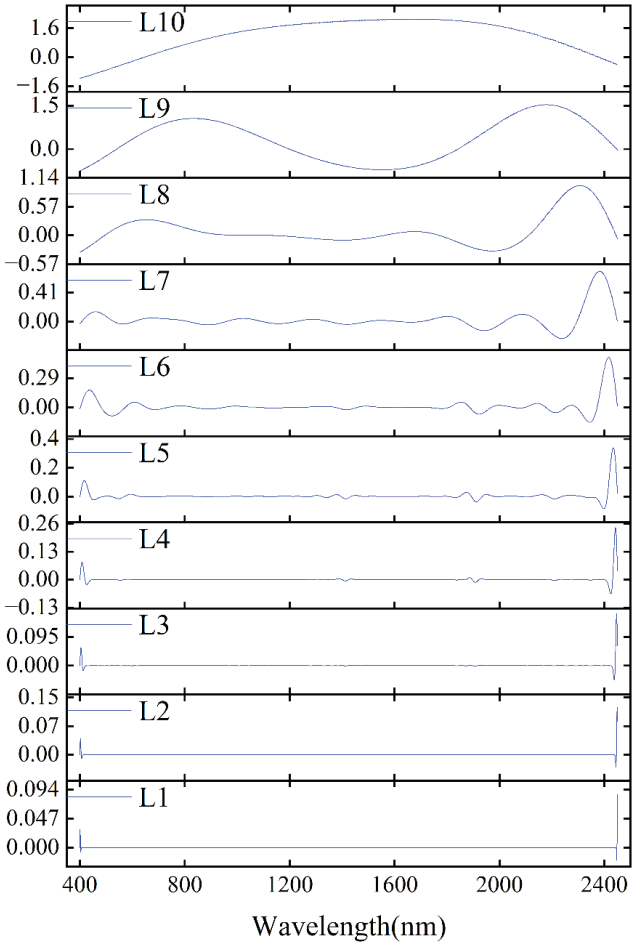


Figure 4. CWT spectra at different scales.

3.3. Correlation of Transformation Spectra with Iron Oxides

In order to observe the effect of different spectral transformations on the original spectrum, the correlation analysis of three transformed spectra with iron oxide was performed.

3.3.1. Correlation of Conventional Transformations with Iron Oxide

The conventional transformation spectra were correlated with the iron oxide content on a wavelength-by-wavelength basis. The results are shown in Figure 6. The bands that passed the 0.01 significance test for CR were mainly around 400–600, 1200–1900, 2100, 2300 and 2400 nm. The overall correlation coefficient curve was similar to that of CR, while log (1/R) passed the 0.01 significance test for all bands. Table 2 shows that log (1/R) achieved the highest correlation coefficient of 0.606 compared to SNV and CR, followed by SNV with a maximum correlation coefficient of −0.590 and 1622 wavelengths passing the 0.01 significance test. The lowest correlation coefficient was obtained for CR with a value of 0.573 and 1255 wavelengths passing the 0.01 significance test.

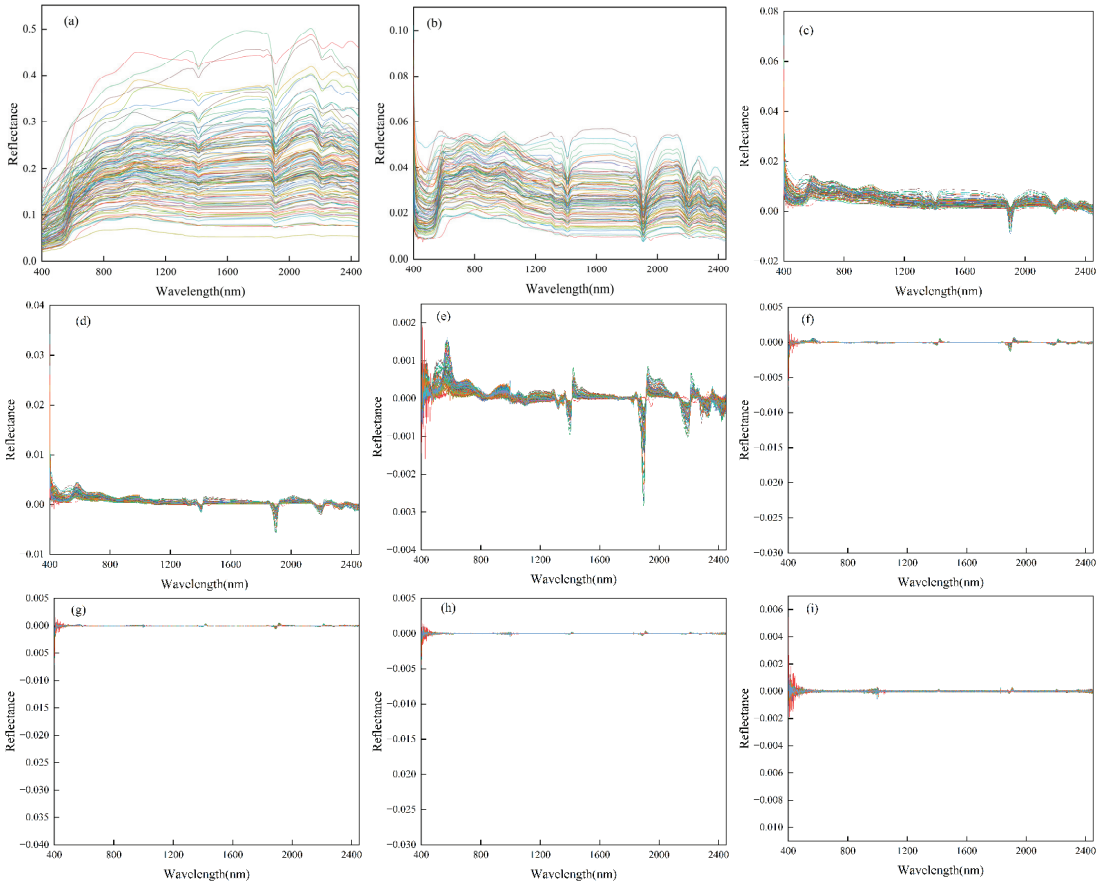


Figure 5. The FOD spectra of soil samples in the calibration dataset: (**a**) original spectra; (**b**) 0.25-order, (**c**) 0.5-order; (**d**) 0.75-order; (**e**) 1-order; (**f**) 1.25-order; (**g**) 1.5-order; (**h**) 1.75-order; (**i**) 2-order.

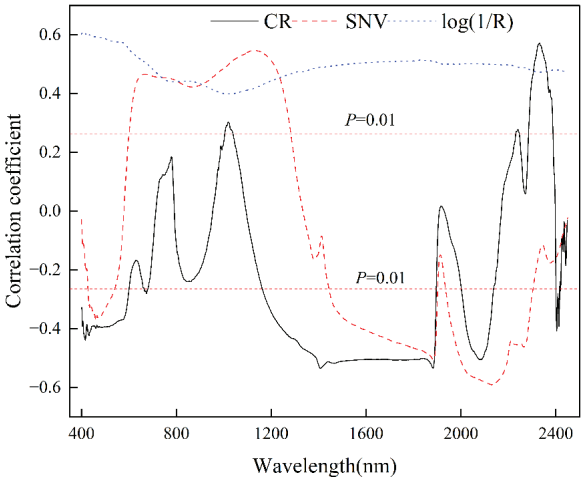


Figure 6. Correlation coefficient between spectra and iron oxide content.

Table 2. The number of wavelengths passing the significance test ($p < 0.01$) and the maximum correlation coefficient for different spectra.

Spectral Transformation Name	Number of Significant Wavelengths	Maximum Correlation Coefficient
CR	1255	0.573
Log (1/R)	2051	**0.606**
SNV	1622	−0.590

3.3.2. Correlation of Continuous Wavelet Transform with Iron Oxide

The wavelet coefficients at each scale were correlated with the iron oxide content of the soil. The heat map of the coefficient of determination of the wavelet coefficient and the iron oxide content is shown in Figure 7. The higher determination coefficients of wavelet coefficients and iron oxide content were mainly distributed in the visible band at the L3, L4, L5 and L6, with the highest determination coefficient reaching 0.367, indicating that the effective information was mainly concentrated at the L3, L4, L5 and L6. At the L1 and L2, the determination coefficients were lower, indicating that some spectral features disappeared and the effective information was less. The number of wavelengths passing the 0.01 significance test and the maximum correlation coefficient for each scale are shown in Table 3. The maximum correlation coefficient of CWT was 0.606 (L6). Meanwhile, the absolute value of the correlation coefficient between wavelet coefficient and iron oxide content of each scale showed a trend of increasing and then decreasing, and the number of its significant wavelengths showed a gradual increase.

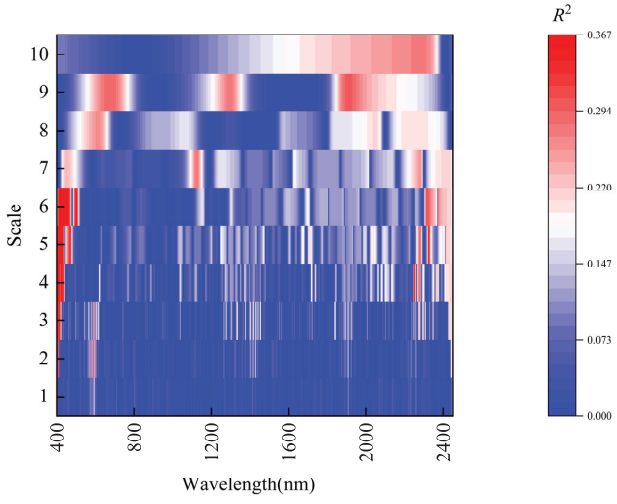


Figure 7. The heat map of the coefficient of determination of the wavelet coefficient and the iron oxide content.

3.3.3. Correlation of Fractional Order Differential with Iron Oxide

Figure 8 shows the correlation coefficients between the different fractional order spectra and the soil iron oxide content in the calibration dataset samples. All bands in the OS (Figure 8a) were negatively correlated with soil iron oxide content. The full range of bands passed the 0.01 significance test. The correlation coefficients between the original spectra and the soil iron oxide content varied smoothly with the wavelength. As the order increased, many positive and negative correlation peaks gradually appeared, and positive and negative correlations occurred for adjacent wavelengths. As can be seen in Table 4, the wavelengths that passed the 0.01 significance test gradually decreased as the order increased, with the maximum correlation reached its maximum (−0.620) at order

0.75 (Figure 8d), while the maximum absolute correlation for the original reflectance was only equal to 0.589. The 1-order and 2-order differentials showed lower correlations than the other order differentials (Figure 8e,i). The FOD provides additional detailed spectral variation information compared to the original spectra (0-order) and the commonly used integer order (1-order and 2-order).

Table 3. The number of wavelengths passing the significance test ($p < 0.01$) and the maximum correlation coefficient for different scale in CWT.

Spectral Transformation Name	Wavelet Decomposition Scale	Number of Significant Wavelengths	Maximum Correlation Coefficient
CWT	L1	92	−0.590
	L2	171	−0.593
	L3	395	**0.606**
	L4	663	−0.602
	L5	1136	0.603
	L6	1056	−0.604
	L7	1320	0.527
	L8	1357	−0.511
	L9	1273	−0.548
	L10	1447	−0.523

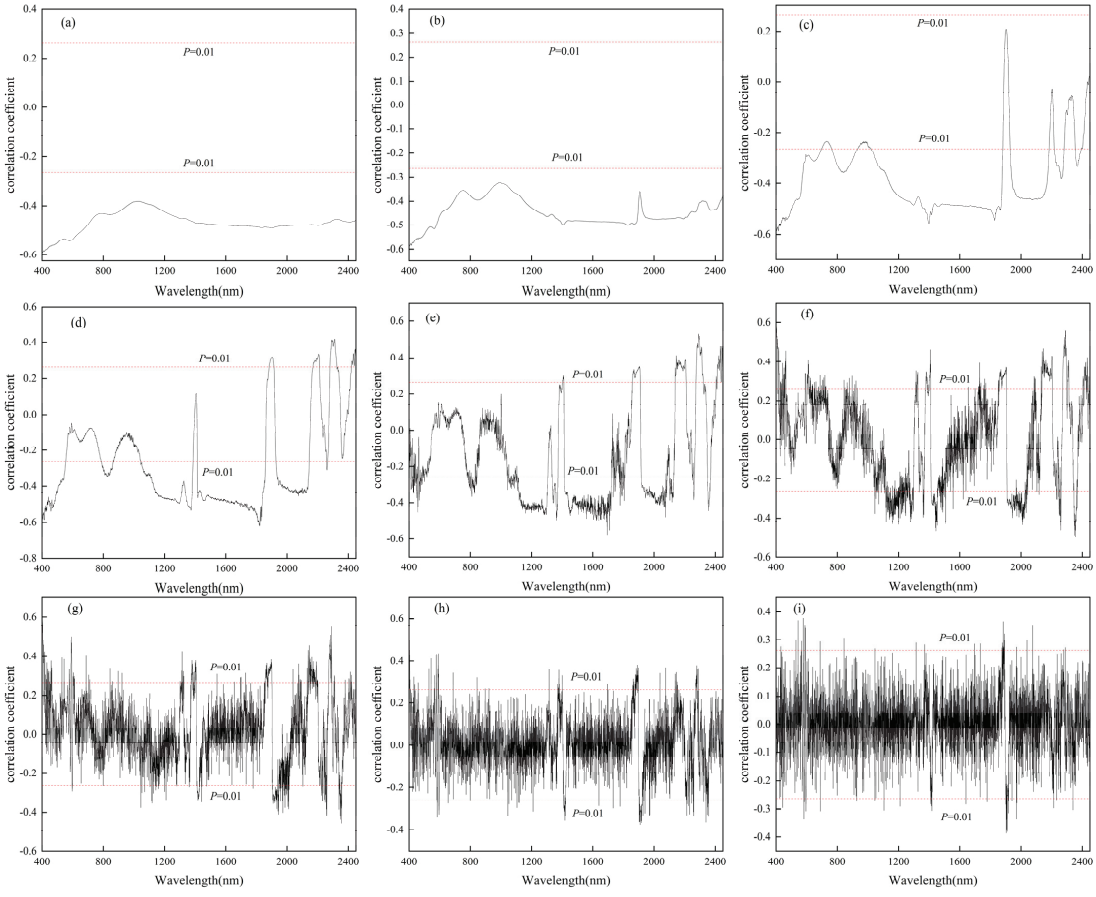


Figure 8. The correlation coefficients between the FOD spectra in the calibration dataset and the soil iron oxide content: (**a**) original spectra; (**b**) 0.25-order, (**c**) 0.5-order; (**d**) 0.75-order; (**e**) 1-order; (**f**) 1.25-order; (**g**) 1.5-order; (**h**) 1.75-order; (**i**) 2-order.

Table 4. The number of wavelengths passing the significance test ($p < 0.01$) and the maximum correlation coefficient for different order in FOD.

Spectral Transformation Name	Number of Significant Wavelengths	Maximum Correlation Coefficient
0-order	2051	−0.589
0.25-order	2051	−0.589
0.5-order	1683	−0.590
0.75-order	1394	**−0.620**
1-order	1138	−0.578
1.25-order	665	0.592
1.5-order	341	0.593
1.75-order	132	0.594
2-order	57	−0.387

3.4. Characteristic Wavelength Selection

If the full wavelength band is used directly as an input variable for modelling, not only is it too inefficient but it may also reduce the accuracy of the model. In this study, CARS was used for the selection of the characteristic wavelengths. As the Monte Carlo sampling method is unstable, the results varied over multiple runs. Therefore, in this study, CARS was cycled through 50 experiments, and the wavelengths with frequencies up to 20 or 30 times in the results obtained were used as the characteristic wavelengths, and their frequency domain thresholds were selected according to the actual situation. The results of the characteristic wavelengths selected according to the CARS algorithm are shown in Figure 9. It was found that most of the wavelengths selected using CARS under the 0.5-order differential transform were distributed around 400 nm, 440 nm and 900 nm, which is consistent with the absorption peak of iron, and the other bands were distributed at 1900 nm and 2200 nm, which was due to the influence of various functional groups. Too few wavelengths were screened at 0-order and 0.25-order, which may have led to later modelling effects being reduced. Wavelengths greater than the 1-order differential screening (1.5-order, 1.75-order and 2-order) were distributed over almost the whole waveband, especially at 600–800 nm, which is considered by previous authors to be the characteristic waveband of organic matter [50]. The L1, L2, L3, CR and log (1/R) were also distributed in the characteristic band of organic matter.

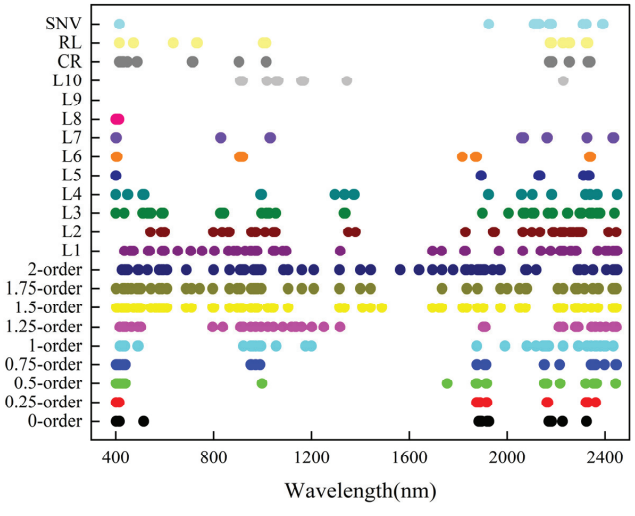


Figure 9. Diagram of the selection of characteristic wavelengths for different spectral transformations.

3.5. Model Construction and Evaluation Using the Full Spectrum

The BPNN and SVR models for estimating soil iron oxide content were constructed based on the full spectrum. As shown in Tables 5 and 6, of the conventional transformations, CR obtained the best prediction accuracy using BPNN with an RPIQ of 2.480, followed by log (1/R) and SNV with RPIQ values of 2.152 and 2.277, respectively. All three models can be considered good models. CR obtained the best prediction accuracy using SVR with an RPIQ of 2.092, followed by log (1/R) and SNV with RPIQ values of 1.813 and 1.898, respectively. In CWT, L7 obtained a good prediction accuracy using BPNN with an RPIQ value of 2.628, which can be considered as an excellent model. However, CWT used SVR to construct the model and achieved the highest accuracy at L4 with an RPIQ value of 2.440. In the FOD, a superior accuracy was obtained for 0.75-order using BPNN and SVR with RPIQ values of 3.045 and 2.529, respectively. By comparing three spectral transforms and two model construction methods, FOD and BPNN showed better performances.

Table 5. The results of the BPNN estimation of soil iron oxide content using the full spectrum.

Spectral Transformation Name	Calibration Dataset		Validation Dataset		
	R^2	RMSE/(g·kg^{-1})	R^2	RMSE/(g·kg^{-1})	RPIQ
CR	**0.795**	**5.096**	**0.671**	**8.168**	**2.480**
Log (1/R)	0.753	6.624	0.624	9.416	2.152
SNV	0.876	3.378	0.579	8.897	2.277
L1	0.621	7.367	0.161	12.399	1.634
L2	0.653	6.792	0.199	14.474	1.400
L3	0.705	6.419	0.305	15.558	1.302
L4	0.744	5.580	0.539	9.992	2.028
L5	0.819	4.602	0.463	10.172	1.992
L6	0.897	3.446	0.580	9.124	2.221
L7	**0.728**	**5.734**	**0.707**	**7.711**	**2.628**
L8	0.323	10.385	0.199	13.722	1.476
L9	0.788	4.981	0.276	12.529	1.617
L10	0.734	5.517	0.501	9.687	2.092
0-order	0.836	4.358	0.488	9.681	2.093
0.25-order	0.812	4.667	0.670	7.887	2.569
0.5-order	0.887	3.621	0.727	7.622	2.658
0.75-order	**0.876**	**3.763**	**0.782**	**6.654**	**3.045**
1-order	0.902	3.347	0.641	8.527	2.376
1.25-order	0.750	5.458	0.454	10.920	1.855
1.5-order	0.661	6.795	0.384	11.106	1.824
1.75-order	0.634	6.741	0.203	14.039	1.443
2-order	0.663	6.414	0.155	13.498	1.501

3.6. Model Construction and Evaluation Using Characteristic Wavelengths

The BPNN and SVR models for estimating soil iron oxide content were constructed based on different spectra. As shown in Tables 7 and 8, the model constructed using SNV was better than OS in BPNN, with an RPIQ of 3.151. This is related to the fact that SNV can reduce the non-specific scattering noise on the sample surface. The model performance of L4, L6 and L7 was better than that of the original spectra with RPIQ values of 3.035, 3.199 and 3.023, respectively. The performance of the L1, L2, L8 and L9 models was not as good as their RPIQ values were less than 2 and the predictive power of the models was poor. This indicates that a scale that is either too low or too high affects the accuracy of the model, with too low a scale resulting in large noise and too high a scale smoothing out some important absorption features. The prediction accuracy of the FOD spectra model varied greatly for different orders. The maximum value of the validation dataset RPIQ was 3.686 and the minimum value was only 1.729. This result shows that it is necessary to choose the right order of the FOD spectra for modelling. The model constructed using CR was better than the log (1/R), SNV and OS in the SVR with a validation set RPIQ of 2.526.

In the CWT, it was the L4 that performed better with an RPIQ of 2.748, while the FOD was better for the model constructed at the 0.75-order with an RPIQ of 2.647. Comparing all the transform spectra, it was found that the BPNN model constructed using the 0.5-order FOD spectra performed significantly better, with the highest R^2 and RPIQ and the lowest RMSE in the validation dataset. This result suggests that the FOD combined with BPNN has more potential for estimating soil iron oxide content.

Table 6. The results of the SVR estimation of soil iron oxide content using the full spectrum.

Spectral Transformation Name	Calibration Dataset		Validation Dataset		
	R^2	RMSE/(g·kg^{-1})	R^2	RMSE/(g·kg^{-1})	RPIQ
CR	**0.591**	**6.805**	**0.480**	**9.687**	**2.092**
Log (1/R)	0.579	6.932	0.308	11.174	1.813
SNV	0.522	7.386	0.368	10.677	1.898
L1	0.548	7.175	0.147	12.404	1.633
L2	0.752	5.322	0.332	10.982	1.845
L3	0.780	5.005	0.537	9.142	2.216
L4	**0.744**	**5.396**	**0.618**	**8.305**	**2.440**
L5	0.699	5.857	0.615	8.339	2.430
L6	0.651	6.309	0.591	8.594	2.358
L7	0.619	6.591	0.521	9.300	2.178
L8	0.597	6.776	0.487	9.616	2.107
L9	0.604	6.720	0.490	9.590	2.113
L10	0.403	8.247	0.264	11.522	1.758
0-order	0.594	6.805	0.346	10.860	1.866
0.25-order	0.694	5.911	0.532	9.193	2.204
0.5-order	0.808	4.681	0.621	8.274	2.449
0.75-order	**0.769**	**5.129**	**0.644**	**8.010**	**2.529**
1-order	0.733	5.519	0.572	8.791	2.304
1.25-order	0.803	4.743	0.573	8.781	2.307
1.5-order	0.814	4.602	0.348	10.845	1.868
1.75-order	0.544	7.207	0.149	12.391	1.635
2-order	0.666	6.170	0.105	12.706	1.595

Table 7. The results of the BPNN estimation of soil iron oxide content using characteristic wavelengths.

Spectral Transformation Name	Number of Characteristic Wavelengths	Calibration Dataset		Validation Dataset		
		R^2	RMSE/(g·kg^{-1})	R^2	RMSE/(g·kg^{-1})	RPIQ
CR	47	0.798	4.850	0.732	6.955	2.913
Log (1/R)	43	0.824	4.640	0.695	7.422	2.730
SNV	31	**0.795**	**4.873**	**0.772**	**6.431**	**3.151**
L1	62	0.623	7.225	0.283	12.347	1.641
L2	55	0.878	3.752	0.499	9.539	2.124
L3	68	0.913	3.247	0.607	8.903	2.276
L4	48	0.779	5.209	0.763	6.676	3.035
L5	20	0.676	6.346	0.668	8.019	2.527
L6	35	**0.788**	**5.239**	**0.787**	**6.334**	**3.199**
L7	26	0.802	4.760	0.753	6.703	3.023
L8	6	0.396	8.347	0.382	10.573	1.916
L9	6	0.568	7.103	0.448	11.618	1.744
L10	10	0.541	7.301	0.518	9.347	2.168
0-order	44	0.731	5.538	0.737	7.082	2.847
0.25-order	20	0.724	5.655	0.719	7.703	2.630
0.5-order	38	**0.903**	**3.376**	**0.851**	**5.497**	**3.686**
0.75-order	32	0.836	4.424	0.832	6.162	3.288
1-order	41	0.917	4.122	0.717	7.713	2.627
1.25-order	47	0.943	2.569	0.702	7.817	2.592
1.5-order	77	0.891	3.574	0.544	9.696	2.090
1.75-order	55	0.764	6.086	0.431	10.401	1.948
2-order	68	0.905	3.340	0.246	11.721	1.729

Table 8. The results of the SVR estimation of soil iron oxide content using characteristic wavelengths.

Spectral Transformation Name	Number of Characteristic Wavelengths	Calibration Dataset		Validation Dataset		
		R^2	RMSE/(g·kg^{-1})	R^2	RMSE/(g·kg^{-1})	RPIQ
CR	47	**0.701**	**5.838**	**0.644**	**8.022**	**2.526**
Log (1/R)	43	0.421	8.128	0.363	10.729	1.889
SNV	31	0.712	5.728	0.494	9.559	2.120
L1	62	0.628	6.511	0.203	11.993	1.689
L2	55	0.866	3.910	0.432	10.126	2.001
L3	68	0.834	4.351	0.517	9.341	2.169
L4	48	**0.716**	**4.549**	**0.652**	**7.924**	**2.748**
L5	20	0.678	6.062	0.599	8.512	2.380
L6	35	0.716	5.688	0.606	8.428	2.404
L7	26	0.695	5.897	0.602	8.471	2.392
L8	6	0.311	8.864	0.272	11.462	1.768
L9	6	0.232	9.375	0.225	11.828	1.713
L10	10	0.376	8.434	0.292	11.306	1.792
0-order	44	0.512	7.457	0.415	10.274	1.972
0.25-order	20	0.689	5.946	0.563	8.878	2.282
0.5-order	38	0.766	5.165	0.651	7.934	2.554
0.75-order	32	**0.727**	**5.576**	**0.675**	**7.655**	**2.647**
1-order	41	0.686	5.985	0.622	8.255	2.455
1.25-order	47	0.878	3.725	0.634	8.126	2.493
1.5-order	77	0.826	4.455	0.425	10.182	1.990
1.75-order	55	0.768	5.147	0.214	10.401	1.948
2-order	68	0.607	6.693	0.308	10.273	1.972

4. Discussion

Soil spectral information is a comprehensive reflection of the soil, which is mainly influenced by soil organic matter, iron oxide, soil texture and pH. The spectral features of soil organic matter and iron oxide in the visible NIR band often overlap [51], resulting in the absorption features of iron oxide in the OS being easily obscured by organic matter [14]. As a result, the inversion of iron oxide using OS may not achieve the expected accuracy. Transformation of spectra is an important tool to improve the predictive power of models [52], and different spectral transformations have different effects in enhancing correlation and highlighting spectral features. In this study, we used three types of spectral transform methods: the conventional spectral transform, CWT and FOD. Among the conventional spectral transforms, the SNV obtained a better prediction accuracy, which may be related to the fact that the SNV transform eliminates the effects of soil grain size, soil surface scattering and light range transformation on reflectance [53]. However, Tan Jie et al. [54] predicted iron oxide in mountainous red soils and found that the CR transform had the highest prediction accuracy compared to the differential transform. The CWT can perform multi-scale decomposition in the time and frequency domains [55,56], and invert the physicochemical properties of soils by finding wavelet coefficients at different scales [57]. The decomposition of the OS using the CWT reveals that the high-frequency information of the wavelet decomposition reflects the main absorption characteristics of the soil hyperspectral, and the sensitivity of the high-frequency information increases with the degree of wavelet decomposition [58]. In this study, the CWT obtained the best prediction results at the L6. However this differs from previous studies that have analyzed the copper content of chicory leaves and found that their spectra were CWT transformed to have optimum scales of L3, L4 and L5 [59]. Differential transformations can reduce the noise and enhance the spectral features of a spectrum [60]. However, traditional integer order differentials cannot capture detailed spectral information [17]. The FOD is gradually being applied to the study of soil spectra with good results [17,61]. The FOD can vary the spectral reflectance at small intervals with different degrees of curvature, thus capturing spectral features that cannot be captured by integer order differential [62]. In this study, the prediction accuracy reached its maximum at 0.5-order as the order increases. Previous studies have also found similar results when using FOD to estimate SOM and moisture content [63]. The reason may be that FOD offers a better balance between spectral resolution, spectral information and

noise than integer order spectra [64]. When the order is greater than 1, the amount of noise exceeds the amount of spectral information, which has a negative impact on the accuracy of the model [65]. Overall, all three types of transformations were effective in improving the prediction accuracy of the model, but the FOD had the best prediction accuracy, and the lower order was more advantageous than the higher order.

Due to the wavelength redundancy of hyperspectral, characteristic wavelength extraction is necessary. Different researchers have differed in their methods of characteristic wavelength selection, including the selection of characteristic wavelengths by correlation analysis using different spectral transformations with iron oxide content [22,29], or by stepwise regression and principal component analysis based on correlation analysis [66]. This study used CARS for characteristic wavelength selection. Due to the instability of CARS, the CARS algorithm was run 50 times in a loop by us to select wavelengths with a frequency of 20 or 30 times as feature wavelengths. Too many or too few characteristic wavelengths can affect the prediction accuracy of the model. From Figure 9, we can see that the number of wavelengths screened out by 0.5 order differential was 38, most of which were distributed around 400 nm, 440 nm and 900 nm, which is consistent with the absorption peak of iron, and the other wavelengths were distributed at 1900 nm and 2200 nm, which was due to the influence of various functional groups, consistent with previous studies. The 1.5-order, 1.75-order, 2-order, L1, L2 and L3 screened out too many wavelengths and the selected wavelengths were distributed in the organic matter characteristic band from 600 to 800 nm. L8, L9 and L10, on the other hand, screened out too few wavelengths and filtered out many wavelengths that were beneficial to the model, resulting in lower prediction accuracy of the model. Therefore, this method can be used as an effective wavelength screening method.

Previous studies have typically used linear models to predict the iron oxide content of soils [23,67], and this study utilized the more widely used BPNN and SVR to construct the models. Neural networks have good approximation properties and generalization capabilities, but often require a large amount of sample data to build excellent models. However, when such networks are applied to small sample data, the input to the model needs to be pre-processed to achieve good prediction accuracy. In this paper, with a small number of samples (n = 135), a series of pre-processing such as the above-mentioned data set partitioning, spectral transformation and extraction of characteristic wavelengths were performed. A BPNN with only one hidden layer was also constructed, which is a simpler network structure and belongs to a shallow neural network. This can avoid the overfitting of the model or a poor generalization performance when the sample data are small. Han lei et al. [68] used BPNN to analyze small sample data (n = 90) and found that BPNN had a higher prediction accuracy compared to PLSR, but also found that there was a corresponding increase in accuracy as the sample size increased. In this study, the results are shown in Tables 7 and 8. The BPNN achieved the best prediction results for the 0.5-order differential transformation and the use of characteristic wavelengths (R^2 = 0.851, RMSE = 5.497 and RPIQ = 3.686). This differs from the previous study [66], which concluded that the first-order differential can effectively improve the prediction accuracy of the model. The SVR achieved the best results that were obtained at 0.75-order (R^2 = 0.675, RMSE = 7.655 and RPIQ = 2.647). Comparing the two methods of constructing the model, BPNN achieved the best model prediction capability at 0.5-order.

5. Conclusions

In this paper, indoor hyperspectral data of surface soils from the southern edge of the Dinosaur Valley in Lufeng, Yunnan, were combined with laboratory data of iron oxide content to perform an inversion of iron oxide content in the region. To verify the predictive power of FOD for iron oxide, the conventional spectral transform and CWT were used for comparison. It was found that the maximum correlation of FOD was stronger than that of the conventional spectral transformation and CWT. The accuracy of the model constructed by full spectrum and characteristic wavelengths was also compared, and it was found that

the accuracy of full spectrum was lower than that using characteristic wavelengths, which indicates that it is necessary to carry out the selection of characteristic wavelengths before the model construction. The FOD achieved the best results among the different modelling methods, with the 0.5-order-BPNN having the strongest predictive power. It indicates that the FOD can obtain more detailed spectral features and effectively improve the prediction ability in soil iron oxide.

The current work was all carried out indoors, and although a high accuracy was obtained, it was limited to small scales. In the future, however, there will be a trend towards using hyperspectral satellites to explore soil spectra at large scales in estimating iron oxide content and describing its spatial distribution.

Author Contributions: Conceptualization, H.Z. and S.G.; methodology, H.Z.; software, H.Z.; validation, H.Z., L.H. and J.W.; investigation, H.Z.; resources, S.G. and X.Y.; data curation, H.Z., L.H., J.W. and S.L.; writing—original draft preparation, H.Z.; writing—review and editing, S.G.; funding acquisition, S.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 41861054.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We are very grateful to Rui Bi for his assistance in writing the thesis and to Xingping Wen for his experimental assistance.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Richter, N.; Jarmer, T.; Chabrillat, S.; Oyonarte, C.; Hostert, P.; Kaufmann, H. Free iron oxide determination in Mediterranean soils using diffuse reflectance spectroscopy. *Soil Sci. Soc. Am. J.* **2009**, *73*, 72–81. [CrossRef]
2. Fontes, M.P.; Carvalho, I.A., Jr. Color attributes and mineralogical characteristics, evaluated by radiometry, of highly weathered tropical soils. *Soil Sci. Soc. Am. J.* **2005**, *69*, 1162–1172. [CrossRef]
3. Cai, M.; Xing, C. Activation of iron oxide in soil and its environmental significance. *J. Zhejiang Normal Univ. Nat. Sci* **2004**, *27*, 279–282.
4. Lin, H.; Shu, G.; Xiping, Y.; Yan, L.; Guokun, C.; Sha, G. Spatial Differentiation Analysis of Water Quality in Dianchi Lake Based on GF-5 NDVI Characteristic Optimization. *J. Spectrosc.* **2021**, *2021*, 5542126. [CrossRef]
5. Zhao, L.; Hu, Y.-M.; Zhou, W.; Liu, Z.-H.; Pan, Y.-C.; Shi, Z.; Wang, L.; Wang, G.-X. Estimation methods for soil mercury content using hyperspectral remote sensing. *Sustainability* **2018**, *10*, 2474. [CrossRef]
6. Wei, L.; Pu, H.; Wang, Z.; Yuan, Z.; Yan, X.; Cao, L. Estimation of soil arsenic content with hyperspectral remote sensing. *Sensors* **2020**, *20*, 4056. [CrossRef]
7. Gao, Y.; Cui, L.; Lei, B.; Zhai, Y.; Shi, T.; Wang, J.; Chen, Y.; He, H.; Wu, G. Estimating soil organic carbon content with visible-near-infrared (Vis-NIR) spectroscopy. *Appl. Spectrosc.* **2014**, *68*, 712–722. [CrossRef]
8. Zhao, A.; Tang, X.; Zhang, Z.; Liu, J. Optimizing Savitzky-Golay parameters and its smoothing pretreatment for FTIR gas spectra. *Spectrosc. Spectr. Anal.* **2016**, *36*, 1340–1344.
9. Hu, L.; Gan, S.; Yuan, X.P.; Li, Y.; Lu, J.; Yang, M.L. Airborne Hyperspectral Features of Three Types of Typical Surface Vegetation in Central Yunnan. *Spectrosc. Spectr. Anal.* **2021**, *41*, 3208–3213. [CrossRef]
10. Ma, J.; Cheng, J.; Wang, J.; Pan, R.; He, F.; Yan, L.; Xiao, J. Rapid detection of total nitrogen content in soil based on hyperspectral technology. *Inf. Processing Agric.* **2021**, 2214–3173. [CrossRef]
11. Grisanti, E.; Totska, M.; Huber, S.; Krick Calderon, C.; Hohmann, M.; Lingenfelser, D.; Otto, M. Dynamic Localized SNV, Peak SNV, and Partial Peak SNV: Novel Standardization Methods for Preprocessing of Spectroscopic Data Used in Predictive Modeling. *J. Spectrosc.* **2018**, *2018*, 5037572. [CrossRef]
12. Cheng, T.; Rivard, B.; Sanchez-Azofeifa, G.A.; Feng, J.; Calvo-Polanco, M. Continuous wavelet analysis for the detection of green attack damage due to mountain pine beetle infestation. *Remote Sens. Environ.* **2010**, *114*, 899–910. [CrossRef]
13. Yang, Y.Y.; Shang, K.; Xiao, C.C.; Wang, C.K.; Tang, H.Z. Spectral Index for Mapping Topsoil Organic Matter Content Based on ZY1-02D Satellite Hyperspectral Data in Jiangsu Province, China. *Isprs Int. J. Geo-Inf.* **2022**, *11*, 111. [CrossRef]
14. Heller Pearlshtien, D.; Ben-Dor, E. Effect of organic matter content on the spectral signature of iron oxides across the VIS–NIR spectral region in artificial mixtures: An example from a red soil from Israel. *Remote Sens.* **2020**, *12*, 1960. [CrossRef]

15. Zhou, W.; Yang, H.; Xie, L.; Li, H.; Huang, L.; Zhao, Y.; Yue, T. Hyperspectral inversion of soil heavy metals in Three-River Source Region based on random forest model. *Catena* **2021**, *202*, 105222. [CrossRef]
16. Xiong, J.; Zheng, G.; Lin, C. Estimating soil iron content based on reflectance spectra. *Spectrosc. Spectr. Anal.* **2016**, *36*, 3615–3619.
17. Cui, S.; Zhou, K.; Ding, R.; Cheng, Y.; Jiang, G. Estimation of soil copper content based on fractional-order derivative spectroscopy and spectral characteristic band selection. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2022**, *275*, 121190. [CrossRef]
18. Tian, A.; Zhao, J.; Tang, B.; Zhu, D.; Fu, C.; Xiong, H. Hyperspectral Prediction of Soil Total Salt Content by Different Disturbance Degree under a Fractional-Order Differential Model with Differing Spectral Transformations. *Remote Sens.* **2021**, *13*, 4283. [CrossRef]
19. Hong, Y.; Liu, Y.; Chen, Y.; Liu, Y.; Yu, L.; Liu, Y.; Cheng, H. Application of fractional-order derivative in the quantitative estimation of soil organic matter content through visible and near-infrared spectroscopy. *Geoderma* **2019**, *337*, 758–769. [CrossRef]
20. Yu, L.; Hong, Y.; Zhou, Y.; Zhu, Q.; Xu, L.; Li, J.; Nie, Y. Wavelength variable selection methods for estimation of soil organic matter content using hyperspectral technique. *Trans. Chin. Soc. Agric. Eng.* **2016**, *32*, 95–102.
21. Vibhute, A.D.; Kale, K.V.; Gaikwad, S.V.; Dhumal, R.K. Estimation of soil nitrogen in agricultural regions by VNIR reflectance spectroscopy. *SN Appl. Sci.* **2020**, *2*, 1–8. [CrossRef]
22. Li, S.; Ma, Y.; Liu, X.; Li, C. Hyperspectral inversion of macro element content in loess based on the profile of Zaoshugou Village, Mangshan Mountain, Zhengzhou City. *Remote Sens. Land Resour.* **2021**, *33*, 121–129.
23. Ding, H.; Chen, Y.; Chen, Y. Remote Sensing Inversion Method of Soil Iron Content in the Loess Plateau. *Remote Sens. Technol. Appl.* **2019**, *34*, 275–283.
24. Shen, Q.; Xia, K.; Zhang, S.; Kong, C.; Hu, Q.; Yang, S. Hyperspectral indirect inversion of heavy-metal copper in reclaimed soil of iron ore area. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2019**, *222*, 117191. [CrossRef]
25. Zhang, H.; Luo, W.; Liu, X.; He, Y. Measurement of soil organic matter with near infrared spectroscopy combined with genetic algorithm and successive projection algorithm. *Spectrosc. Spectr. Anal.* **2017**, *37*, 584–587.
26. Yang, H.; Kuang, B.; Mouazen, A. Quantitative analysis of soil nitrogen and carbon at a farm scale using visible and near infrared spectroscopy coupled with wavelength reduction. *Eur. J. Soil Sci.* **2012**, *63*, 410–420. [CrossRef]
27. Araújo, M.C.U.; Saldanha, T.C.B.; Galvao, R.K.H.; Yoneyama, T.; Chame, H.C.; Visani, V. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemom. Intell. Lab. Syst.* **2001**, *57*, 65–73. [CrossRef]
28. Li, H.; Liang, Y.; Xu, Q.; Cao, D. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal. Chim. Acta* **2009**, *648*, 77–84. [CrossRef]
29. Qin, Q.; Qi, Y.; Wu, J.; Yang, Y.; Liu, M. Estimation of Random Forest Model of Soil Free Iron Based on Hyperspectral Data. *Chin. J. Soil Sci.* **2018**, *49*, 1286–1293.
30. Yuan, Z. Study on the Characteristics of the Geoheritages and Protection in Lufeng Dinosaur National Geopark, Yunnan. Master's Thesis, China University of Geosciences, Beijing, China, May 2015.
31. Clark, R.N.; Roush, T.L. Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. *J. Geophys. Res. Solid Earth* **1984**, *89*, 6329–6340. [CrossRef]
32. Tan, Y.; Jiang, Q.; Liu, H.; Liu, B.; Gao, X.; Zhang, B. Estimation of Organic Matter, Moisture, Total Iron and pH From Back Soil Based on Multi Scales SNV-CWT Transformation. *Spectrosc. Spectr. Anal.* **2021**, *41*, 3424–3430.
33. Yu, L.; Hong, Y.; Zhou, Y.; Zhu, Q. Inversion of soil organic matter content using hyperspectral data based on continuous wavelet transformation. *Spectrosc. Spectr. Anal.* **2016**, *36*, 1428–1433.
34. Xue, D. FOTF toolbox for fractional-order control systems. *Appl. Control* **2019**, *6*, 237–266.
35. Li, J.; Guo, Z.; Huang, W.; Zhang, B.; Zhao, C. Near-infrared spectra combining with CARS and SPA algorithms to screen the variables and samples for quantitatively determining the soluble solids content in strawberry. *Spectrosc. Spectr. Anal.* **2015**, *35*, 372–378.
36. Zhang, J.; Xi, L.; Yang, X. Construction of hyperspectral estimation model for organic matter content in sandy ginger black soil. *Trans. CSAE* **2020**, *36*, 135–141.
37. Jazzbin, E.A. Geatpy: The Genetic and Evolutionary Algorithm Toolbox with High Performance in Python. Available online: http://www.geatpy.com (accessed on 23 June 2022).
38. Zhang, B.; Guo, B.; Zou, B.; Wei, W.; Lei, Y.; Li, T. Retrieving soil heavy metals concentrations based on GaoFen-5 hyperspectral satellite image at an opencast coal mine, Inner Mongolia, China. *Environ. Pollut.* **2022**, *300*, 118981. [CrossRef]
39. Meng, X.; Bao, Y.; Liu, J.; Liu, H.; Zhang, X.; Zhang, Y.; Wang, P.; Tang, H.; Kong, F. Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *89*, 102111. [CrossRef]
40. He, Z.H.; Ma, Z.H.; Li, M.C.; Zhou, Y. Selection of a calibration sample subset by a semi-supervised method. *J. Near Infrared Spectrosc.* **2018**, *26*, 87–94. [CrossRef]
41. Bellon-Maurel, V.; Fernandez-Ahumada, E.; Palagos, B.; Roger, J.-M.; McBratney, A. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *TrAC Trends Anal. Chem.* **2010**, *29*, 1073–1081. [CrossRef]
42. Tang, H.; Meng, X.; Su, X.; Ma, T.; Liu, H.; Bao, Y.; Zhang, M.; Zhang, X.; Huo, H. Hyperspectral prediction on soil organic matter of different types using CARS algorithm. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 105–113.
43. Nawar, S.; Mouazen, A.M. Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. *Catena* **2017**, *151*, 118–129. [CrossRef]

44. Yu, L.; Hong, Y.; Geng, L.; Zhou, Y.; Zhu, Q.; Cao, J.; Nie, Y. Hyperspectral estimation of soil organic matter content based on partial least squares regression. *Trans. Chin. Soc. Agric. Eng.* **2015**, *31*, 103–109.
45. Viscarra Rossel, R.; Bui, E.; De Caritat, P.; McKenzie, N. Mapping iron oxides and the color of Australian soil using visible–near-infrared reflectance spectra. *J. Geophys. Res. Earth Surf.* **2010**, *115*, 1–13. [CrossRef]
46. Peng, J.; Zhou, Q.; Zhang, Y.; Xiang, H. Effect of soil organic matter on spectral characteristics of soil. *Acta Pedol. Sin.* **2013**, *50*, 517–524.
47. Ji, G.; Xu, B. Reflectance of soil clay minerals and its application in pedology. *Acta Pedol. Sin.* **1987**, *24*, 67–76.
48. Zhang, S.; Shen, Q.; Nie, C.; Huang, Y.; Wang, J.; Hu, Q.; Ding, X.; Zhou, Y.; Chen, Y. Hyperspectral inversion of heavy metal content in reclaimed soil from a mining wasteland based on different spectral transformation and modeling methods. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2019**, *211*, 393–400. [CrossRef]
49. Tian, A.H.; Zhao, J.S.; Tang, B.H.; Zhu, D.M.; Fu, C.B.; Xiong, H.G. Study on the Pretreatment of Soil Hyperspectral and Na+ Ion Data under Different Degrees of Human Activity Stress by Fractional-Order Derivatives. *Remote Sens.* **2021**, *13*, 3974. [CrossRef]
50. Shi, Z.; Zhou, Q.; Zhou, L.-Q. VIS NIR reflectance spectroscopy of the organic matter in several types of soils. *J. Infrared Millim. Waves* **2012**, *31*, 277–282.
51. Henderson, T.; Baumgardner, M.; Franzmeier, D.; Stott, D.; Coster, D. High dimensional reflectance analysis of soil organic matter. *Soil Sci. Soc. Am. J.* **1992**, *56*, 865–872. [CrossRef]
52. Maleki, M.; Van Holm, L.; Ramon, H.; Merckx, R.; De Baerdemaeker, J.; Mouazen, A. Phosphorus sensing for fresh soils using visible and near infrared spectroscopy. *Biosyst. Eng.* **2006**, *95*, 425–436. [CrossRef]
53. Fang, C.; Kuang, H.; Zhou, X.; Chen, X.; Wan, X.; Liu, Y. Hyperspectral Inversion Model of Soil Heavy Metals in Enshi Area, Hubei Province. *Environ. Sci. Technol.* **2021**, *44*, 154–159.
54. Tan, J.; Chen, Y.; Zhou, W.; Cui, H.; Liu, P. Inversion of Iron Oxide Contents in Forest Soils of Dawei Mountains Using Laboratory Hyperspectral Data. *Soils* **2021**, *53*, 858–864.
55. Zhang, J.; Sun, H.; Gao, D.; Qiao, L.; Liu, N.; Li, M.; Zhang, Y. Detection of canopy chlorophyll content of corn based on continuous wavelet transform analysis. *Remote Sens.* **2020**, *12*, 2741. [CrossRef]
56. Pinto, L.A.; Galvão, R.K.; Araújo, M.C.U. Influence of wavelet transform settings on NIR and MIR spectrometric analyses of diesel, gasoline, corn and wheat. *J. Braz. Chem. Soc.* **2011**, *22*, 179–186. [CrossRef]
57. Guo, J.; Zhao, X.; Guo, X.; Zhu, Q.; Luo, J.; Xu, Z.; Zhong, L.; Ye, Y. Inversion of soil properties in rare earth mining areas (southern Jiangxi, China) based on visible–near-infrared spectroscopy. *J. Soils Sediments* **2022**, 1–16. [CrossRef]
58. Gu, X.; Wang, Y.; Sun, Q.; Yang, G.; Zhang, C. Hyperspectral inversion of soil organic matter content in cultivated land based on wavelet transform. *Comput. Electron. Agric.* **2019**, *167*, 105053. [CrossRef]
59. Lin, D.; Li, G.Z.; Zhu, Y.D.; Liu, H.T.; Li, L.T.; Fahad, S.; Zhang, X.Y.; Wei, C.; Jiao, Q.J. Predicting copper content in chicory leaves using hyperspectral data with continuous wavelet transforms and partial least squares. *Comput. Electron. Agric.* **2021**, *187*. [CrossRef]
60. Schmitt, J.M. Fractional derivative analysis of diffuse reflectance spectra. *Appl. Spectrosc.* **1998**, *52*, 840–846. [CrossRef]
61. Hong, Y.; Chen, S.; Liu, Y.; Zhang, Y.; Yu, L.; Chen, Y.; Liu, Y.; Cheng, H.; Liu, Y. Combination of fractional order derivative and memory-based learning algorithm to improve the estimation accuracy of soil organic matter by visible and near-infrared spectroscopy. *Catena* **2019**, *174*, 104–116. [CrossRef]
62. Hong, Y.; Shen, R.; Cheng, H.; Chen, Y.; Zhang, Y.; Liu, Y.; Zhou, M.; Yu, L.; Liu, Y.; Liu, Y. Estimating lead and zinc concentrations in peri-urban agricultural soils through reflectance spectroscopy: Effects of fractional-order derivative and random forest. *Sci. Total Environ.* **2019**, *651*, 1969–1982. [CrossRef]
63. Ge, X.; Ding, J.; Jin, X.; Wang, J.; Chen, X.; Li, X.; Liu, J.; Xie, B. Estimating agricultural soil moisture content through UAV-based hyperspectral images in the arid region. *Remote Sens.* **2021**, *13*, 1562. [CrossRef]
64. Chen, L.; Lai, J.; Tan, K.; Wang, X.; Chen, Y.; Ding, J. Development of a soil heavy metal estimation method based on a spectral index: Combining fractional-order derivative pretreatment and the absorption mechanism. *Sci. Total Environ.* **2022**, *813*, 151882. [CrossRef] [PubMed]
65. Wang, J.; Ding, J.; Abulimiti, A.; Cai, L. Quantitative estimation of soil salinity by means of different modeling methods and visible-near infrared (VIS–NIR) spectroscopy, Ebinur Lake Wetland, Northwest China. *PeerJ* **2018**, *6*, e4703. [CrossRef] [PubMed]
66. Guo, Y.; Guo, Z.; Liu, J.; Yuan, Y.; Sun, H.; Chai, M.; Bi, R. Hyperspectral inversion of paddy soil iron oxide in typical subtropical area with Pearl River Delta, China as illustration. *J. Appl. Ecol.* **2017**, *28*, 3675–3683.
67. Yang, Y.; Huang, W.; Lu, Y.; Li, B.; Jingqiong, O.; Tang, X.; Wang, C.; Chen, Y. Spectral characteristics and quantitative retrieval of free iron content in soil. *J. South China Agric. Univ.* **2020**, *41*, 91–99.
68. Han, L.; Chen, R.; Zhu, H.; Zhao, Y.; Liu, Z.; Huo, H. Estimating soil arsenic content with visible and near-infrared hyperspectral reflectance. *Sustainability* **2020**, *12*, 1476. [CrossRef]

agriculture

MDPI

Article

Towards a Real-Time Oil Palm Fruit Maturity System Using Supervised Classifiers Based on Feature Analysis

Meftah Salem M. Alfatni [1], Siti Khairunniza-Bejo [2,3,4,*], Mohammad Hamiruce B. Marhaban [5], Osama M. Ben Saaed [1], Aouache Mustapha [6] and Abdul Rashid Mohamed Shariff [2,3,4]

[1] Libyan Authority for Scientific Research, Tripoli P.O. Box 80045, Libya
[2] Department of Biological and Agricultural Engineering, Faculty of Engineering, Universiti Putra Malaysia, Serdang 43400, Malaysia
[3] Laboratory of Plantation System Technology and Mechanization (PSTM), Institute of Plantation Studies, Universiti Putra Malaysia, Serdang 43400, Malaysia
[4] Smart Farming Technology Research Centre (SFTRC), Faculty of Engineering, Universiti Putra Malaysia, Serdang 43400, Malaysia
[5] Centre for Control System and Signal Processing, Faculty of Engineering, Universiti Putra Malaysia, Serdang 43400, Malaysia
[6] Division Télécom, Centre de Développement des Technologies Avancées, CDTA, Baba-Hassen 16303, Algeria
* Correspondence: skbejo@upm.edu.my

Abstract: Remote sensing sensors-based image processing techniques have been widely applied in non-destructive quality inspection systems of agricultural crops. Image processing and analysis were performed with computer vision and external grading systems by general and standard steps, such as image acquisition, pre-processing and segmentation, extraction and classification of image characteristics. This paper describes the design and implementation of a real-time fresh fruit bunch (FFB) maturity classification system for palm oil based on unrestricted remote sensing (CCD camera sensor) and image processing techniques using five multivariate techniques (statistics, histograms, Gabor wavelets, GLCM and BGLAM) to extract fruit image characteristics and incorporate information on palm oil species classification FFB and maturity testing. To optimize the proposed solution in terms of performance reporting and processing time, supervised classifiers, such as support vector machine (SVM), K-nearest neighbor (KNN) and artificial neural network (ANN), were performed and evaluated via ROC and AUC measurements. The experimental results showed that the FFB classification system of non-destructive palm oil maturation in real time provided a significant result. Although the SVM classifier is generally a robust classifier, ANN has better performance due to the natural noise of the data. The highest precision was obtained on the basis of the ANN and BGLAM algorithms applied to the texture of the fruit. In particular, the robust image processing algorithm based on BGLAM feature extraction technology and the ANN classifier largely provided a high AUC test accuracy of over 93% and an image-processing time of 0,44 (s) for the detection of FFB palm oil species.

Keywords: external grading system; oil palm FFB; machine learning; supervised classifiers; quality inspection; remote sensing

Citation: Alfatni, M.S.M.; Khairunniza-Bejo, S.; Marhaban, M.H.B.; Saaed, O.M.B.; Mustapha, A.; Shariff, A.R.M. Towards a Real-Time Oil Palm Fruit Maturity System Using Supervised Classifiers Based on Feature Analysis. *Agriculture* **2022**, *12*, 1461. https://doi.org/10.3390/agriculture12091461

Academic Editors: Sebastian Kujawa and Gniewko Niedbała

Received: 22 July 2022
Accepted: 2 September 2022
Published: 14 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Usually, information that can be obtained from a distance about objects or areas is a science called remote sensing [1]. Remote sensing is a procedure that can be used to measure the external physical properties of an area by receiving the energy reflected and emitted from the target area [2].

Further, an introduction and development of remote sensing was published by the authors of Refs. [3,4] using various sensors, image processing tools and techniques for remote sensing applications. In fact, the most common sensors used in remote sensing are

cameras and solid-state scanners, such as CCD (charge coupled device) images, which are available in 2D matrices for the application, and satellite image sensors [5,6].

Bakker et al. [7] indicated that a digital CCD camera is an electro-optical remote sensor made of semiconductor material, which is the most common type of detection nowadays in the range from visible to very near IR; it is used to provide area information in aerial Earth observation application for low-cost imagery. Likewise, bi-directional reflectance measurements and the method for determining the suppression of atmospheric MTF for the CCD camera on board the Huanjing 1A (HJ-1A) satellite and the digital CCD camera were designed and studied by the authors of Refs. [8,9].

In addition, a standard digital camera used to automatically monitor snow cover with high accuracy in terms of time and location as a unified index based on red, green and blue (RGB) values was developed to control errors due to lighting effects [10]. On the other hand, microcomputers have increased power as an advantage of remote sensing image processing technology. Based on their study [11], researchers used the BRIAN micro CSIRO system as an example of how to access methods for image processing, which contributed to the acceptance of remote sensing technology.

Machine learning is an evolutionary area of algorithms, hardware and storage systems working in smarter ways for several applications, such as (a) abnormal behavior proactive detection for reasonable solutions in advance; (b) creating events models based on system training in order to forecast the values of a future inquiry; (c) testing the future inquiry based on the understating of the created event model and (d) computing the individual loss reserve [12]. Thus, different researchers have used the advantage of machine learning for automated wheat diseases classification, estimation of the long-term agricultural output and prediction of soil organic carbon and available phosphorus [13–15]. Therefore, there are many benefits and advantages to using machine learning methods in computing the individual loss reserve regarding ML techniques, making such methods more feasible, with more accurate pricing, claims triage, loss prevention, a deep dive in changes in loss reserves and frequent monitoring to calculate claims reserves on individual claims data (ICR) [16–19].

In addition, Refs. [20,21] reported that machine learning techniques provide the possibility to activate and control the classification of images by remote sensing. However, multi-band deep learning, deep convolutional neural networks and modular features were implemented using limited training samples by the authors of Refs. [22–28] to classify the hyperspectral, hyperparameter, spectroradiometer and spectrometer images as those remotely detected data.

As a result, a variety of fields have been successfully populated with numerous remotely sensed images with high spectral–spatiotemporal resolution in order to identify important acoustic processes in agricultural applications [29,30]. The old method of assessing the quality of agricultural products, in general, is tedious and expensive [31]. Traditional techniques have been in use for a long time, but they are extremely tedious, expensive and out of control over time. In this context, high-tech switches are needed to use machine vision to classify the quality of agricultural food products and to assess timely and accurately [32–42].

This technique is suitable for surface or sub-surface imaging due to the incomplete penetration depth of the interrogation source. In addition, researchers from all over the world have contributed and developed an automated internal classification system as a solution for screening agricultural crops based on internal characteristics, such as sugar, moisture and acid [43–46].

Equally, the authors of Refs. [47–51] applied remote sensing and image processing technology based on texture and color measurement methods to classify images in a different application. In this paper, the remote sensing (CCD camera) and image processing (Gabor waves, GLCM and BGLAM) sensor technologies as texture property extraction techniques are based on supervised machine learning classifiers (SVMs, KNN and ANN) so as not to distract the assets and to assess the FFB quality inspection on a real-time system.

This article will review in Section 2 the fruit ripeness classification aspect, and the discussion will expand on relevant works regarding methodologies and strategies for an automated FFB grading and sorting system using different approaches, including data collection, system material, image processing, classification system and other available tools for system evaluation and assessment. Next, Section 3 will entail the results and discussion, including obtained results by modeling the different feature algorithms with classification modules. Section 4 will be the conclusion and discuss the challenges and future direction.

2. Fruit Ripeness Classification

Currently, different computer vision systems have been invented and applied to assess the quality of agricultural crops based on different color spaces through machine learning techniques. The use of such systems for a variety of fruit ripening processes based on different color spaces and classification techniques resulted in varying research accuracy, as shown in Table 1.

Table 1. The literature on accurately assessing crop quality using machine learning techniques based on a variety of color spaces. Table from Ref. [51] is cited and updated with new implementation results.

Item	Color Space	Classification Technique	Accuracy	Ref.
Oil palm	UV + RGB + NIR	KNN and SVM	93.80	[25]
Dates	JPG	CNN	99.32	[26]
Banana	RGB and GLCM	CNN and MLP	98.45	[27]
Apple	HSI	SVM	95.00	[52]
Apple	L*a*b*	MDA	100.00	[53]
Apple	RGB	SVM	96.81	[54]
Apple, pears and peaches	RGB	ANN	98.90	[55]
Papaya	LBP, HOG and GLCM	KNN, SVM and Naıve Bayes	100.00	[56]
Avocado	RGB	K-Means	82.22	[57]
Dragon Fruit	HSV + RGB	Naive Bayes	86.60	[58]
Banana	L*a*b*	LDA	98.00	[59]
Banana	RGB	ANN	96.00	[60]
Blueberry	RGB	KNN and SK-Means	85.00–98.00	[61]
Date	RGB	K-Means	99.60	[62]
Lime	RGB	ANN	100.00	[63]
Mango	RGB	SVM	96.00	[64]
Mango	L*a*b*	MDA	90.00	[65]
Mango	L*a*b*	LS-SVM	88.00	[66]
Oil palm	L*a*b*	ANN	91.67	[67]
Pepper	HSV	SVM	93.89	[68]
Paper	HIS + RGB	SIS	99.00	[69]
Persimmon	RGB + L*a*b*	QDA	90.24	[70]
Tomato	HSV	SVM	90.80	[71]
Tomato	RGB	DT	94.29	[72]
Tomato	RGB	LDA	81.00	[73]
Tomato	L*a*b*	ANN	96.00	[74]
Rice	Texture Features (Gray)	SVM	86.00	[75]
Soya	HSI	ANN	95.70	[76]
Banana	RGB	Fuzzy logic	NA	[77]
Banana	RGB + CIE L*a*b*	ANN	NA	[78]
Banana	RGB	CNN	87.00	[77]
Watermelon	YCbCr	ANN	86.51	[79]
Watermelon	VIS/NIR	ANN	80.00	[80]
Watermelon	RGB	ANN	73.33	[81]
Tomato	FTIR	SVM	99.00	[82]
Kiwi	Chemometrics MOS E-nose	PLSR, SVM and RF	99.40	[83]
Coffee	RGB + L*a*b* + Luv + YCbCr + HSV	SVM	92.00	[84]
Coffee	RGB, HIS and L*a*b*	PCA and K-Means	100.00	[85]
Cape Gooseberry	RGB + HSV + L*a*b*	ANN, DT, SVM and KNN	93.02	[86,87]

L* indicates lightness, and a* and b* are chromaticity coordinates.

As shown in Table 1, the largest generic classifier technologies are SVM, ANN, K-Means and KNN at 34%, 31%, 11% and 9%, respectively, whereas the most used color spaces in Table 1 are RGB, LAB, HSV, HIS and YCbCr with 57%, 31%, 14%, 9% and 6%,

respectively, with high output resolution. Thus, RGB color space and SVM classifier are the most popular technologies that achieved higher resolution.

To increase the production of high-quality crude palm oil, one of the challenges is to harvest the fresh fruit bunches (FFB) of oil palm at the optimal stage of ripeness. Actually, the current methods used to determine the optimal ripened stage are based on color and loose fruits observation. This traditional method relies heavily on the undiscovered technique of palm fruit size experimentation and intuition to accurately determine ripeness that cannot be easily replicated and is subject to significant human error. To address this issue and find a systematic solution to determine the oil palm fruit ripeness that is cost-efficient, fast, non-invasive, reliable and precise, researchers contributed to developing a tech-based solution using computer vision that enables auto-grading and sorting of the optimal ripened stage by integrating software (image processing, robust datasets, AI decision-making) and hardware systems (lighting system, grading and sorting system). The advancement in methods and techniques for FFB classification and grading has resulted in the development of automated computer analysis, which will aid farmers significantly in obtaining good quality in crude palm oil production, particularly in rural areas with limited access to automation facilities.

2.1. Data Collection

According to confirmation between the scientific teams of the Universiti Putra Malaysia (UPM) and Palm Oil Board of Malaysia (MPOB), knowledge and experience were shared to study the properties and future of FFB palm oil at different stages of maturity to collect valuable information. Thus, the study began with a field visit, as shown in Figure 1.

Figure 1. Excursion to MPOB Kluang: (**a**) transport, (**b**) MPOB oil palm field, (**c**) research group, (**d**) harvest method, (**e**) FFB sample and (**f**) FFB fruitlet.

The purpose of the visit was to select the study area and FFB types of oil palm according to the research needs. Accordingly, the preparation of the survey and verification of the methods and techniques for the palm oil fruit maturity grading system involved collecting 270 fruit images for each of the three types of palm oil fruit, which are (i) Nigrescens, (ii) Oleifera and (iii) Virescens, as shown in Figure 2. Each harvested fruit received a specific sheet containing its name, number, type and ripeness class. The data collection process for the oil palm system is as follows:

1. An expert in the classification of palm oil fruit maturity was appointed. The expert classified the fruits based on three grades, namely under-ripe, ripe and over-ripe;
2. A specified number of fruits per day were collected. The collection ranged from 15 to 20 fruits based on the ability of the lab capacity and the quantity available in the field;
3. Give the physical image of each fruit the name and number of the organization using the computer or during laboratory analysis;
4. Third item.

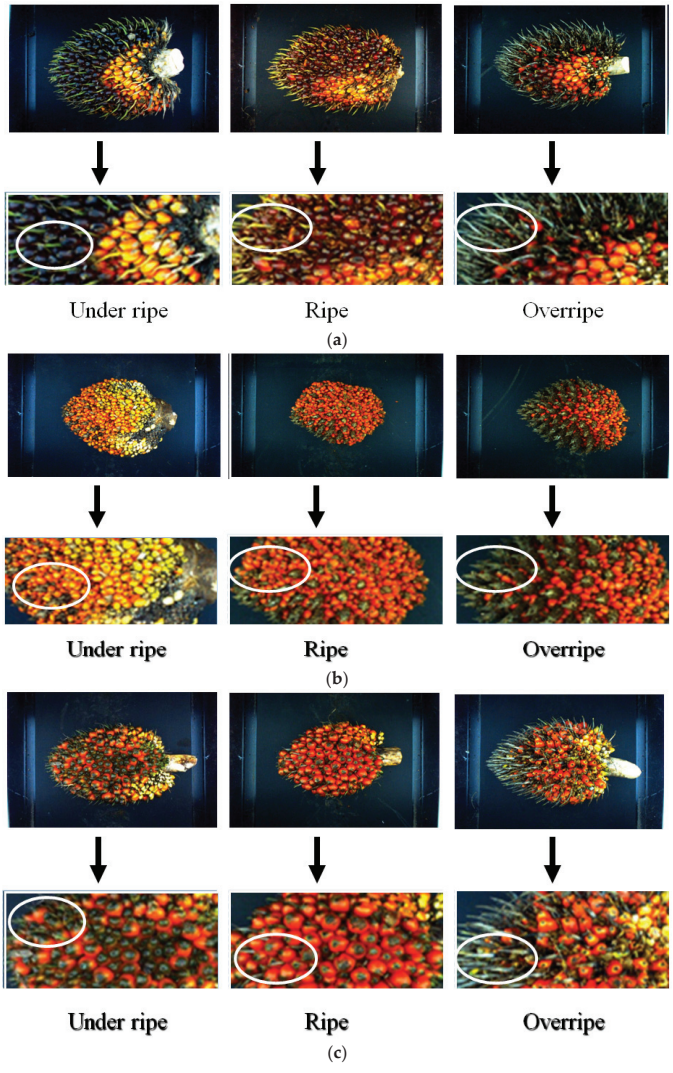


Figure 2. FFB ripening classes of oil palm: (**a**) Nigarsens, (**b**) Olifera, (**c**) Varseness.

2.2. System Material

In general, the material and process of the FFB palm oil maturity classification system is shown in Figure 3. Accordingly, the fruit ripeness grading system used computer vision application in agricultural quality inspection to ensure ripeness category of fruit. The system includes: (a) a housing having an enclosure for scanning process; (b) defused tubes of LED illumination means with optical lens illumination filter provided at the enclosure of the housing; (c) preferably, a suitable charge coupled device (CCD) digital camera DFK 41BF02.H FireWire CCD color camera is used to capture fruit sample's image, provided at top portion of the enclosure of the housing; (d) a feeding device for conveying fruit samples to the housing; (e) a processing unit to process and analyze the fruit sample image; (f) a data acquisition interface provided in between the camera and the processing unit and wherein the processing unit further provided with a disk top computational unit serves to transfer data to a computer. In fact, the fruit was obtained in real time with a controlled indoor lighting system.

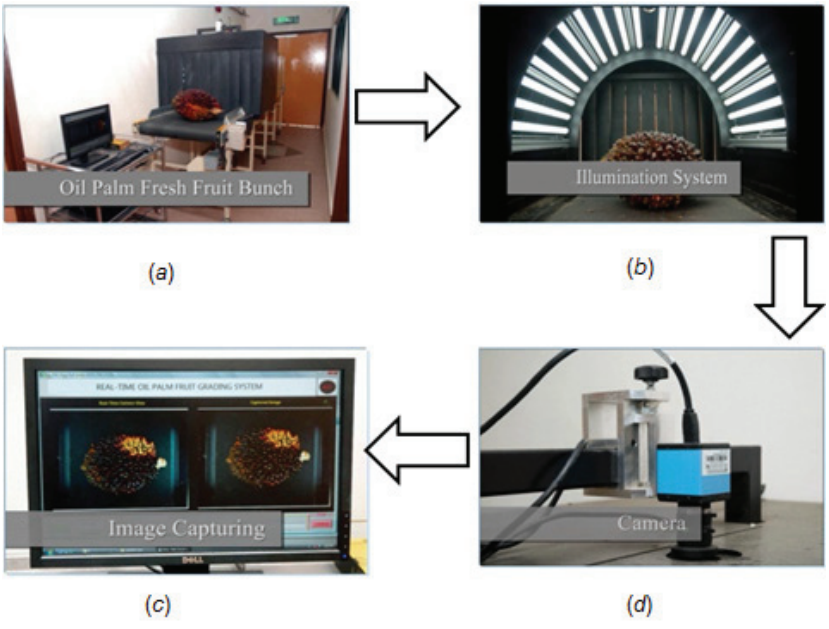


Figure 3. Process of acquiring materials and images: (**a**) oil palm grading system, (**b**) lighting system, (**c**) captured image, (**d**) camera and RGB cable.

2.3. Image Processing Approach

In general, image processing and analysis using computer vision and external file systems were performed with general and standard steps, as shown in Figure 4 [31,88,89]. Image acquisition and pre-processing include low-level processing, segmentation, representation and description as mid-level operations, while higher-level operations include object recognition and image classification.

As a result, the group of oil palm fruits went through fruit image processing stages based on various steps, as shown in Figure 5. The steps included fruit image acquisition, pre-processing and processing, treatment, segmentation and extraction of features as well as applying the retrieval methods and techniques as a decision-making system based on the similarity calculation as proposed in the future work. All images were related to the training model and a fresh fruit bunch was evaluated. The decision-making process was based on the training model.

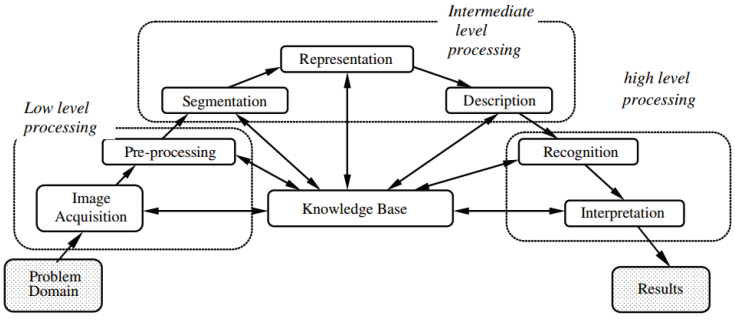


Figure 4. The three levels of image processing algorithm for external grading system process.

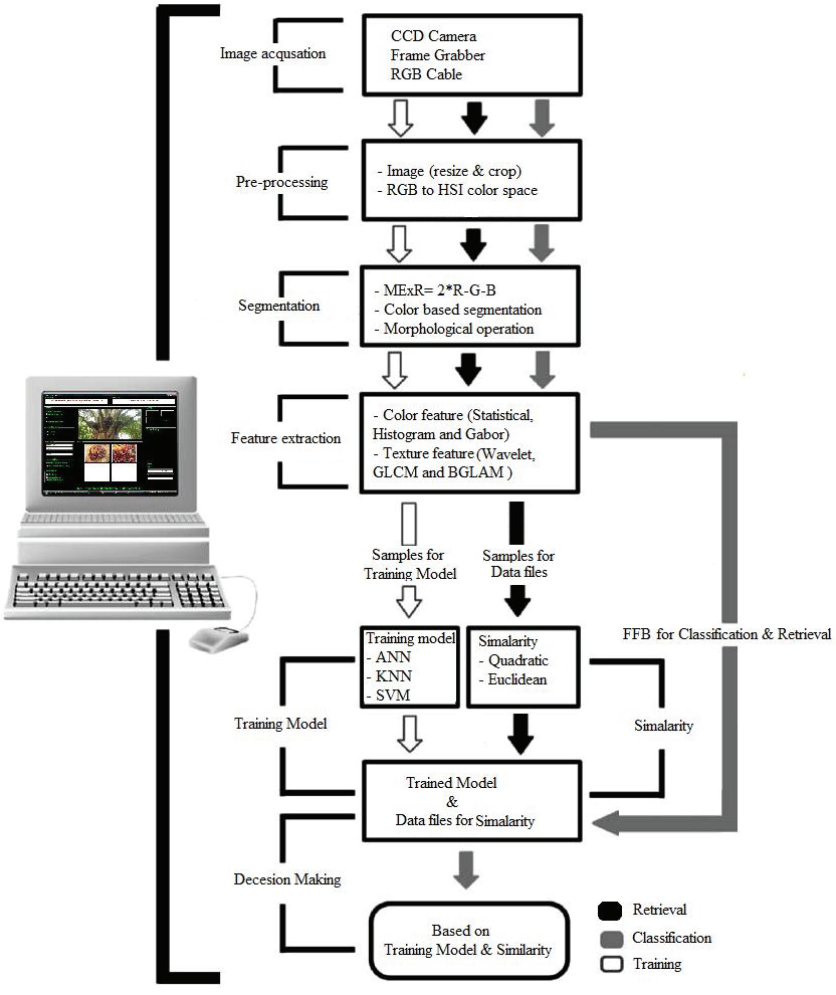


Figure 5. Steps and image processing techniques for the FFB real-time oil palm ripeness classification system.

Several experiments were performed with different models (color, texture and thorns) of the FFB palm oil classification system. The three different regions of interest (ROI1, ROI2, ROI3) were verified for the FFB maturity of the oil palm, as shown in Figure 6, using various feature extraction techniques (color feature extraction, such as mean, standard deviation and color histogram techniques) as well as texture extraction techniques (Gabor wavelet (GW), gray level co-occurrence matrix (GLCM) and basic gray level halo matrix (BGLAM)).

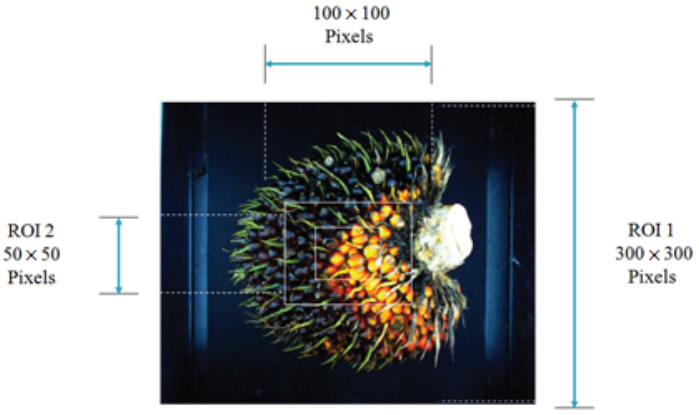


Figure 6. Three different regions of interest (ROI1, ROI2, ROI3) for the FFB ripeness of oil palm.

2.4. Classification System

Decision-making based on image classification through supervised machine learning classifiers is the last step in the process, which is a method of learning a set of rules from cases called a training set to create a classifier that can be used to create a great presentation using new cases for tests [90–92]. The classification system defines objects by classifying them in a limited set of categories [93–95]. As noted at the beginning of this article in Table 1, the most popular supervised classifiers in fruit categorization are SVM, ANN and KNN. These classifiers were used in this article for the experimental parameters.

2.4.1. Artificial Neural Network (ANN)

An artificial neural network (ANN) provides an efficient alternative for mapping complex nonlinear relationships between input and output datasets without the need for detailed knowledge of the underlying physical relationships [96]. ANNs contain connected nerve cells that mimic the work of the brain. ANN differs significantly from algorithm software due to its ability to disseminate knowledge about new data unearthed. Expert systems must collect real knowledge about the specific area. Multi-layered direct feedback neural networks are grouped into input, output and hidden layers and are used with the FFB oil palm classification system.

Each layer comprises several neurons, which are known as processing elements (PE), as illustrated in Figure 7 [67,97–99]. No pre-defined rules were needed to be set for ANN because it is able to learn and generalize from "experience" or a set of presented examples, which is called a training set. The number of optimum hidden neurons was determined experimentally from the training processes of the MLP classifiers. An in-depth description of the MLP concept was addressed by the authors of Ref. [100].

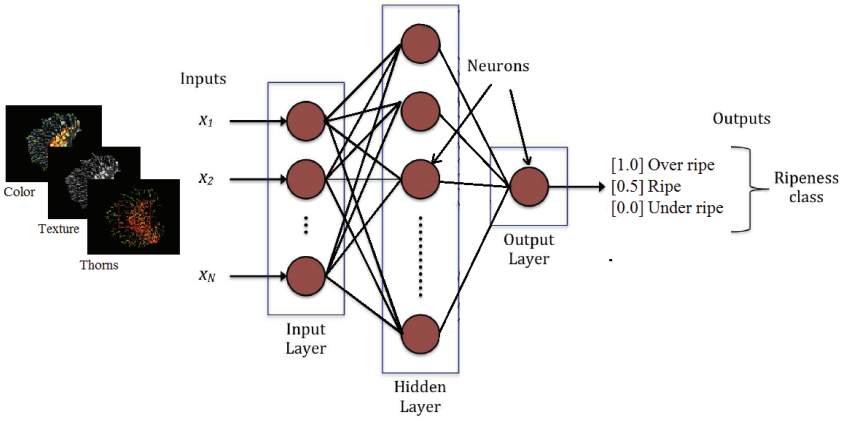


Figure 7. ANN multi-layer structure directly operating with a single port of the FFB oil palm grading system.

Figure 6 illustrates the construction of a three-layer MLP building. The general task of the PEs in the input layer of an MLP is to buffer the input signals to the PEs in the hidden layer. This step collects the products of input signals with their weighted connections by each PE.

Artificial Neural Network (ANN)

Varying the weights given to neural connections is a process of training a neural network to achieve a satisfactory result. The supervised learning procedure for multi-layered front-end power systems provides a recipe for changing the weight of elements in adjacent layers. This algorithm reduces the sum of squares errors, which have been identified as least squares.

The mean square errors (MSE) and the efficiency (EFF) of the training and testing for each classifier are calculated.

During the training phase, data were used to fit the system using the ANN model. Each category in the dataset was presented as an input sample for ANN–MLP for training assignments. In order to reduce the mean square error (MSE) between goals and outputs, a trial and error trial [100] was performed. Under-ripe, ripe and over-ripe were determined using the desired outputs as 0.5, 0 and 1, respectively, while the input characteristics were normalized within the range [0, 1]. Training effectiveness was used as an important indicator of the accuracy of rating evaluation. However, each method used different ANN constructs to result in inefficiency. The commonly used backpropagation networks were selected for the FFB classification system for oil palm trees due to their success with a variety of image processing applications in agriculture [101–104].

2.4.2. K-Nearest Neighbor (KNN)

KNN is another supervised classifier used in this work based on the concept that observations in a dataset are, in general, close to other observations with similar properties. Additionally, the metric distance and k-value play a major role in the KNN classification algorithm [105], although Ref. [106] notes that the KNN classifier is not a pre-classifier; KNN determines their location. kNN is used to query the new training space model based on the appropriate similarity distance scale.

KNN Performance

KNN regulation is one of the largest algorithms for classifying attractive patterns. In this work, different k-values and distance measurement methods were adapted to balance

the trade-off of the FFB maturity classification by excluding values and methods having low confidence accuracy levels, as shown in Figure 8.

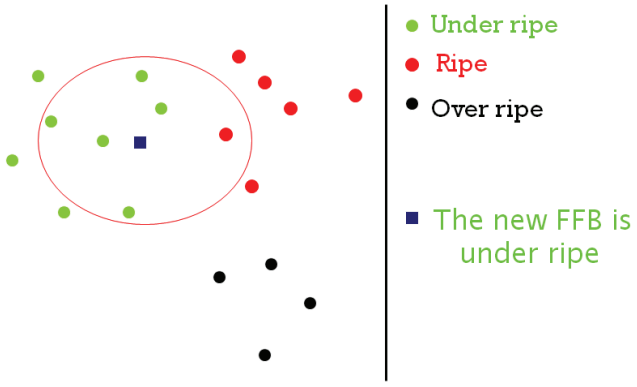


Figure 8. K-nearest neighbor (KNN) example with k = 6.

Moreover, an experimental investigation was carried out based on the values of K, which are 1, 3, 5, 7 and 9, as well as methods for measuring metric distance, namely: Euclidean, City, Cosine and Correlation, as in "Equations" by Refs. [105,107]. The research aims to determine the appropriate KNN classification coefficient for the high-precision FFB palm oil maturity classification system. The study showed that the appropriate distance measure that reduced the distance between two similar classified examples is the city-block metric. The value of k = 1 affects the performance of the KNN procedure. The results of the evaluation can be obtained next regarding applications in agriculture [101–104].

2.4.3. Support Vector Machine (SVM)

KNN SVM is a supervised machine learning classifier developed by the authors of Ref. [108] based on constructing hyper-plane as a decision line separating Class 1 from Class 2, as shown in Figure 9 [109]. A special characteristic of SVMs is that they simultaneously reduce experimental classification error and maximize geometric boundary by optimizing the superlative level of linear separation and converting the nonlinear data model into a linearly separable format in a feature space with high-dimensional [110].

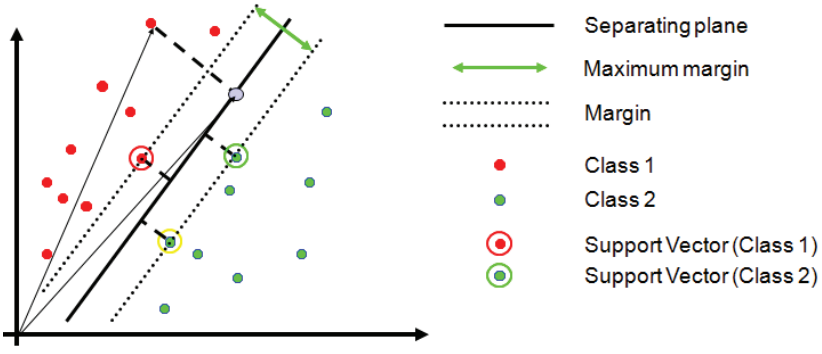


Figure 9. Support vector machine (SVM) classification system.

In the FFB maturity classification, there are three different target classes (under-ripe, ripe and over-ripe) and one against all approach (OAA), which subdivides each class and merges all the others [111]. Due to the performance efficiency and less processing time

than the multi-class SVM classifier, the OAA method was used to perform the FFB ripeness classification of oil palm.

SVM Performance

To improve the classification result for specific models, special classes of FFB palm oil had to be learned according to linear, non-linear and four-step basis. First, the input data comprise two sets of vectors in an n-dimensional space. SVM will build a separate hyperplane in that space that increases the "margin" between the two datasets. Second, when calculating the margin, we construct two parallel planes parallel, one on each side of the separator planes, that are "pushed up" for the two datasets. Third, instinctively, a fine separation is reached by means of the hyper-plane that has the largest distance to the data points adjacent to both classes. Finally, the classifier's best generalization error will depend on the largest margin or distance between these parallel hyperplanes.

The parameter tuning is the most important factor in the SVM model-building process. In SVM, tests were accomplished with different kernel types, such as linear, polynomial and radial basis function kernels, to achieve the classification task. Furthermore, to control the trade-off between maximizing the margin and minimizing the training error, the sigma of RBF was tuned from 1 to 100 and the polynomial distance was also tuned from 1 to 4. The regularization parameter magnitude C was tuned from 1 to 1000 for both polynomial and RBF kernels.

2.5. Training and Testing

As with Kotsiantis, three techniques are used to estimate the accuracy of the classifier [105]. First is the cross-validation technique, by dividing the training set into subgroups of equal selection and size. For each subgroup, the classifier is trained on one of all other subgroups. The second is the leave-one-out validation. The third is the most common one, which is used in this work with the FFB palm oil grading system. Two-thirds of the data are for training and the remaining is for performance appraisal.

Numerous statistical measurements of efficiency and mean square error (MSE) were applied as indexes to validate the performance of the classifier. In particular, an automatic parameter tuning procedure as in Ref. [112] is implemented for the system to dynamic adaptive thresholding algorithm for the oil palm FFB ripeness grading. The objective of supervised learning is to create a concise model of the distribution of class labels in terms of predictor features.

Training and Testing Stage

The training stage includes data collection, data analysis and a training model analyzing 270 fruit samples of three different ripeness categories for the three different oil palm FFB types that were collected, analyzed and then a training model for fruit image type and ripeness classification was created. Meanwhile, the testing stage included testing the grading system initially in the lab. Testing the grading system in the field ensured that the system provided a high percentage of internal validity for findings obtained using the system design. Furthermore, 90 samples for each class were used to test the oil palm FFB ripeness grading system. Figure 10 illustrates the main approaches considered in the classification module for oil palm FFB types and ripeness.

In general, the classification of FFB type and ripeness of oil palm was successfully carried out based on the performance of three levels of image processing and subsequent analysis, as shown in Figure 11.

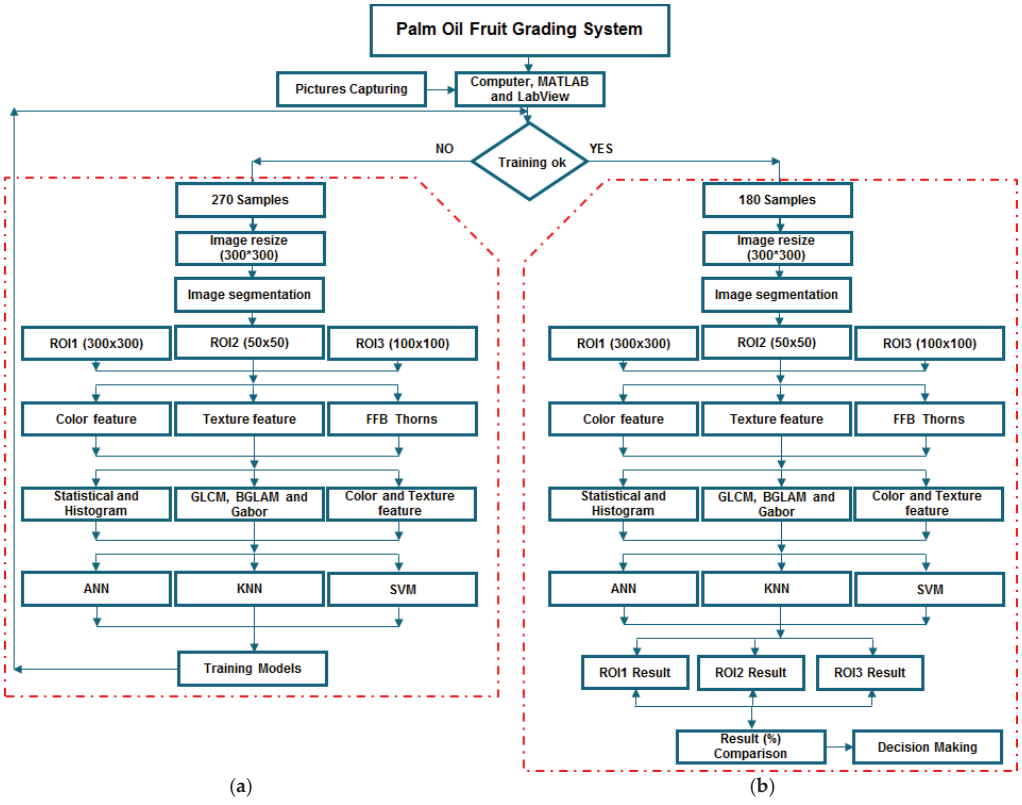


Figure 10. Training and testing stages of the oil palm ripeness grading system: (**a**) training stage and (**b**) testing stage.

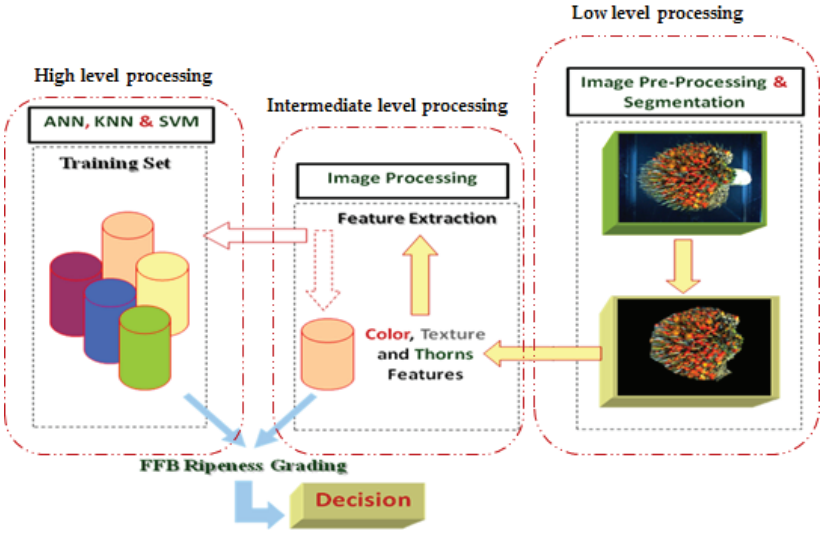


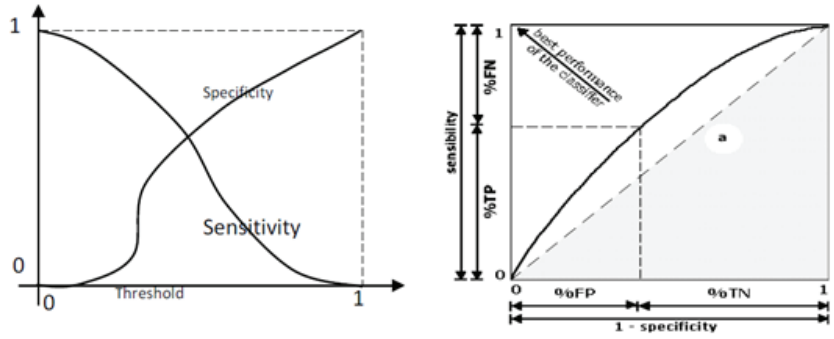
Figure 11. Image processing algorithms levels for FFB ripeness classification.

2.6. Classifier Performance Evaluation

The performance measurement of a classifier independently is conducted according to its sensitivity and specificity. The analysis of the ROC of a classifier is a solution to limit the empirical precision of binary classification. Results significantly greater than 50% could be due to a biased classifier tested on an unbalanced dataset, and overall precision does not differentiate between forms of error [113]. The experiments aimed to infer the crucial architecture with the selected color, texture and spine models using the ROC as a statistical measurement analysis. This analysis provides a quantitative assessment using AUC.

Receiver Operating Characteristic Curve

Figure 12 shows the ROC curve, which has become the standard tool for evaluating predictive accuracy to evaluate and compare models and prediction algorithms. ROC analysis offers a methodical analysis of the sensitivity and specificity of judgment [114,115].



(**a**) (**b**)

Figure 12. (**a**) Graphs of sensibility versus threshold, specificity versus threshold and (**b**) ROC curve with a hypothetical example [114,115].

Sensitivity is the capability of the classifier to recognize the positive pattern amongst the truly positive patterns. Specificity is the ability of the classifier to recognize the negative patterns amongst the truly negative patterns. Figure 12 shows that point (0,1) is the ideal classifier, which categorizes all the positive and negative cases appropriately. In this instance, the false positive rate is none (0), and the true positive rate is all (1). In addition, point (0, 0) indicates that the classifier predicted all the cases to be negative, while point (1, 1) matches a classifier with all the cases that are positive. Point (1,0) means the classifier fails to implement the correct classification for all the cases, as shown in Figure 9. The given n test samples are constructed according to the confusion matrix as illustrated in Table 2 that resulted from classification [113,115–117]. The calculation of accuracy, sensitivity or true positive rate (TPR) and 1- specificity or false positive rate (FPR) are given by 'Equations (1)–(3)', respectively.

$$Accuarcy = \frac{TP + TN}{n} \tag{1}$$

$$TPR = Sensitivity = \frac{\sum TP}{\sum TP + \sum FN} \tag{2}$$

$$FPR = Specificity = \frac{\sum TN}{\sum TN + \sum FP} \tag{3}$$

where the true TP positives are the number of correctly classified maturities; true negatives TN is the number of incorrectly classified maturities; false positives FP is the number of maturities classified as non-maturities and false negative FN is the number of non-maturities classified as maturities. Finally, the performance evaluation of the oil palm FFB

maturity classification system classifier typically includes the measurement of sensitivity and specificity as performance results based on the ROC curve and measurement of the area under the ROC curve (AUC).

Table 2. Confusion matrix.

Test	Actual Positive	Actual Negative
Predicted Positive	*TP*	*FP*
Predicted Negative	*FN*	*TN*
Total	*P*	*N*

3. Results and Discussion

The FFB characteristics of the oil palm (color, texture and thorns) were extracted using the algorithms of the color model, the texture model and the thorn model. Three different supervised machine learning techniques, ANN, KNN and SVM, were incorporated into the extracted features based on the three different models to make decisions regarding FFB type and maturity. Experiments were carried out on the classifiers to select the appropriate model for the FFB oil palm grading system and to ensure high-quality grading results. The best possible classification accuracy can be achieved by selecting the highest AUC measured from the ROC curve.

3.1. Classification Based on ANN–MLP

This section discusses MLP models as classifying FFB maturity of oil palm based on statistical color function, color histogram, Gabor wavelets, GLCM and BGLAM functionality. The different ANN models selected on the basis of the experimental results were performed with different feature extraction techniques implemented in the oil palm grading system, as shown in Figure 13. A comparison between the MSE and the effectiveness of the training results and test steps was performed to validate the parameters of the ANN supervised learning classifier, as shown in Table 3.

Table 3. MSE and efficiency result comparison of the training and testing stages based on the FFB feature techniques.

FET	Models	Training Stage		Testing Stage	
		MSE	Eff	MSE	Eff
Statistical	[40 × 10 × 1]	0.0080	0.9523	0.0190	0.8865
	[40 × 20 × 1]	0.0072	0.9568	0.0197	0.8820
	[40 × 30 × 1]	0.0038 *	0.9775 *	0.0182 *	0.8914 *
Histogram	[25 × 10 × 1]	9.9402×10^{-5} *	0.9994 *	0.0136 *	0.9189 *
	[25 × 15 × 1]	9.8412×10^{-5}	0.9994	0.0163	0.9024
	[25 × 20 × 1]	9.5647×10^{-5}	0.9994	0.0144	0.9140
GLCM	[40 × 10 × 1]	9.9991×10^{-5}	0.9994	0.0291	0.8263
	[40 × 20 × 1]	1.2126×10^{-4}	0.9993	0.0330	0.8030
	[40 × 30 × 1]	9.9943×10^{-5} *	0.9994 *	0.0278 *	0.8338 *
BGLAM	[45 × 11 × 1]	8.9770×10^{-5}	0.9995	0.0242	0.8556
	[45 × 22 × 1]	9.6629×10^{-5} *	0.9994 *	0.0177 *	0.8942 *
	[45 × 33 × 1]	9.5860×10^{-5}	0.9994	0.0215	0.8712
Gabor	[40 × 10 × 1]	9.9573×10^{-5}	0.9994	0.0922 *	0.4489 *
	[40 × 20 × 1]	9.9895×10^{-5} *	0.9994 *	0.0966	0.4228
	[40 × 30 × 1]	9.9873×10^{-5}	0.9994	0.1330	0.2048

Notes: FET = Feature extraction techniques, MSE = Mean square error, Eff = Efficiency, * = The best result.

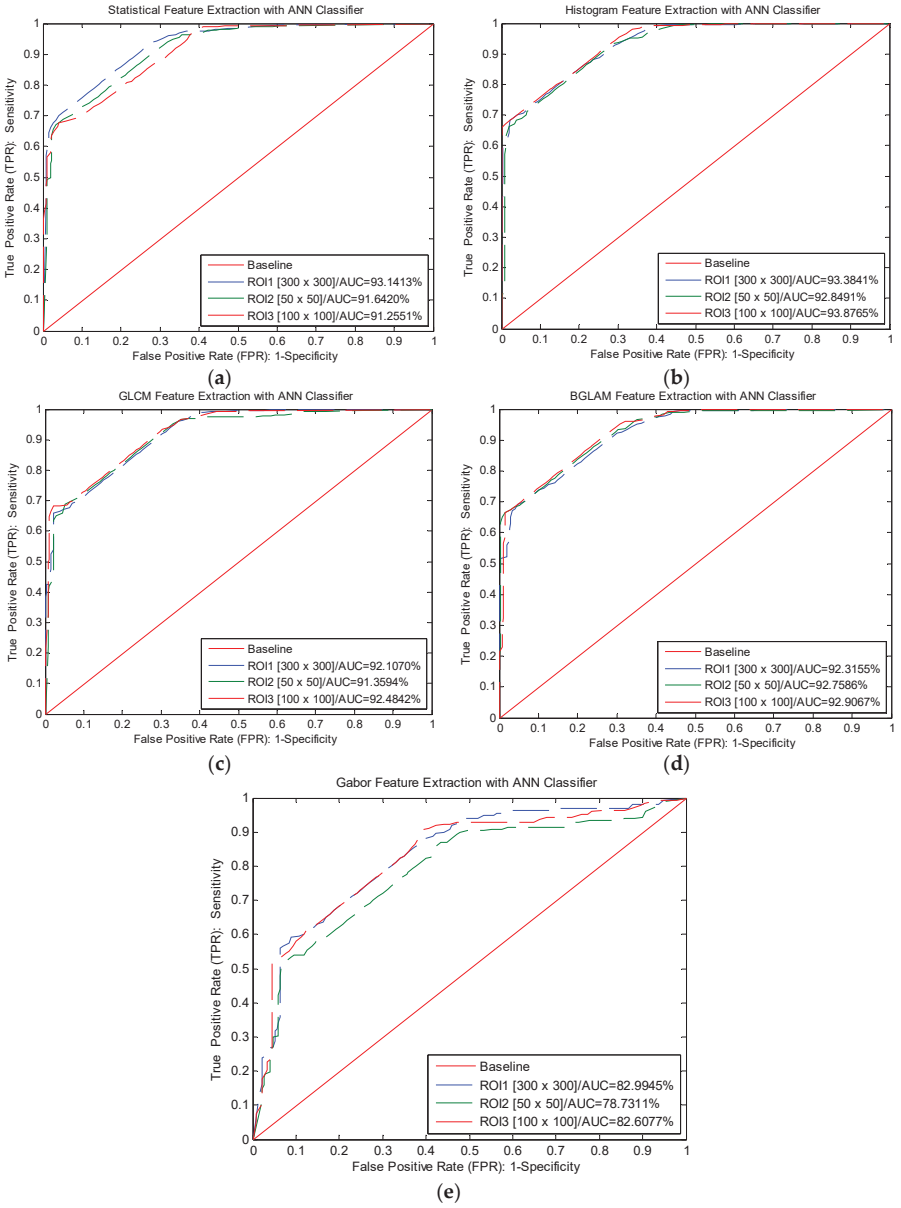


Figure 13. AUC score of FFB maturity based on feature extraction and ANN: (**a**) statistical, (**b**) histogram, (**c**) GLCM, (**d**) BGLAM and (**e**) Gabor.

Table 3 indicates that the MLP and MSE learning stage of the learning procedures did not exceed 0.003. The higher proficiency score observed revealed a selection scale of the architectural MLP model [40 × 30 × 1], [25 × 10 × 1], [40 × 30 × 1], [45 × 22 × 1] and [40 × 20 × 1] over all the tracks with statistical color function, color histogram, GLCM, BGLAM and GW, respectively, for the FFB recording system. After several training sessions, the MLP model was able to learn and perfectly match the target in the training phase with extreme efficiency and with complete FFB palm oil training datasets. During the test phase,

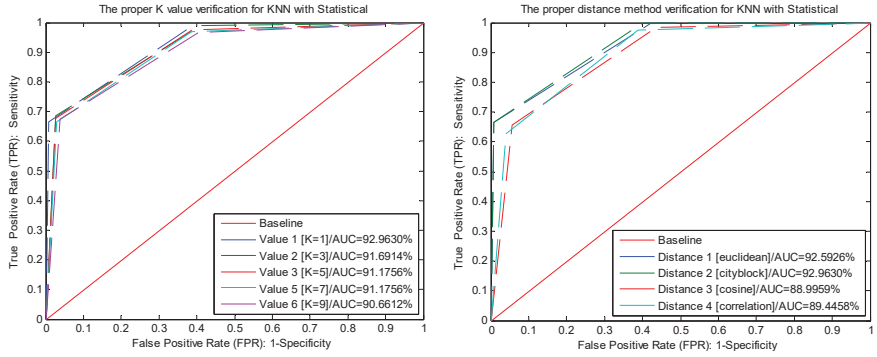
Figure 13a–e shows the classification of ROC plots performed by the FFB oil palm maturity classification system, with a higher AUC score observed in the MLP models.

3.2. Classification Based on KNN

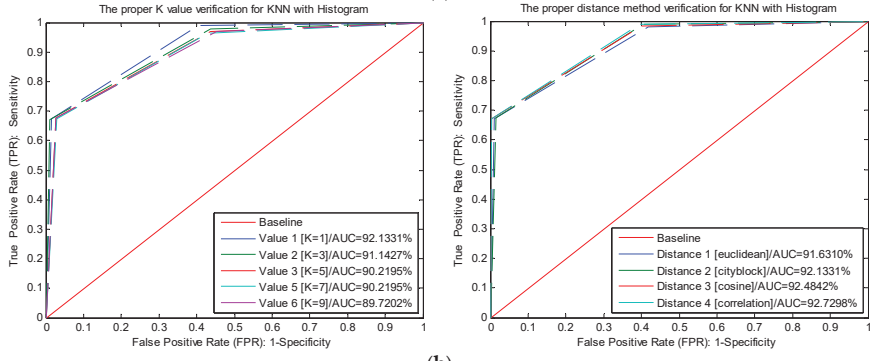
The basic principle of the oil palm grading system based on nearest neighbor (NN) approximation is that two FFB images with similar color, texture and thorn features should reveal similar classes and grades. Thus, using the FFB images of similar ripeness is sensible when identifying the new FFB image. All images in the database can be grouped based on their ripeness features. The nearest neighbor technique is defined as dividing a sample set into categories, with each category holding similar samples that share the same features. The testing sample is determined by the known classifications of the training samples.

Based on the samples' characteristics, five main steps were described to classify FFB images of oil palm into their categories (under-ripe, ripe and over-ripe). Indeed, choosing the best k-values and appropriate distance measurements ensures the accuracy of the results of the KNN classifier, which were usually chosen experimentally by static validation with a set of k-values and distance measurements. Thus, the best k-value that can be used with feature extraction techniques (statistical color feature, color histogram, Gabor wavelet, GLCM and BGLAM) was verified.

Figure 14 shows the ROC area for the best results performed by KNN with different values of k = 1, 3, 5, 7 and 9 and with different distance metrics, Euclidean, city-block, cosine and correlation, for the FFB oil palm maturity grading system with feature extraction techniques. Therefore, the experimental results show that k-value = 1 with the city-block distance technique provides the greatest AUC scores equal to 93.00%, 92.00%, 91.00%, 92.00% and 80% using feature extraction techniques, including statistical, color histogram, GLCM, BGLAM and Gabor wavelet, respectively, based on the KNN algorithm, as shown in Figure 14.

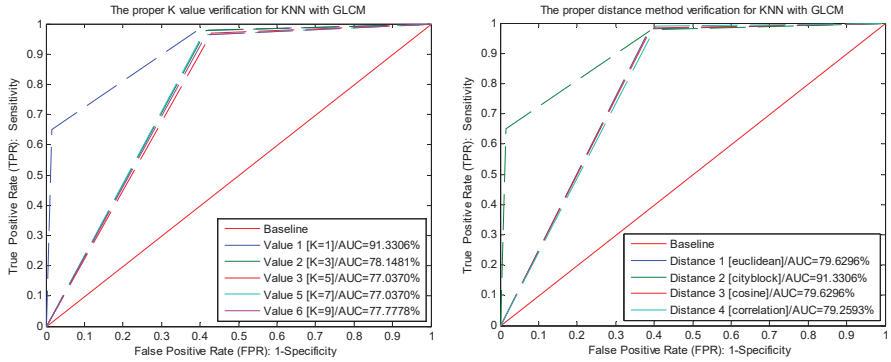


(**a**)

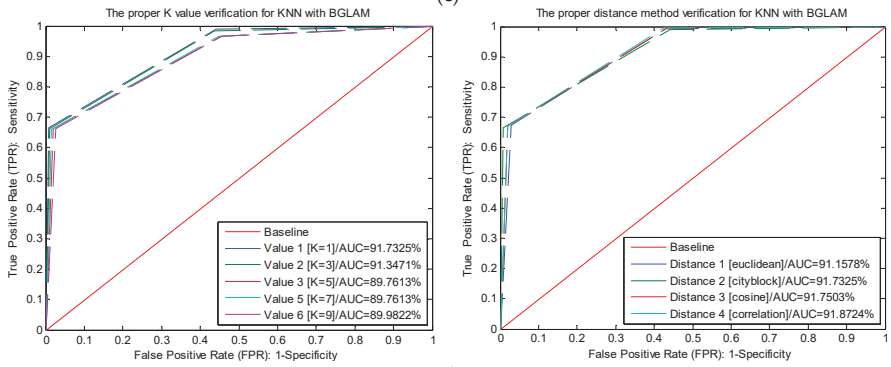


(**b**)

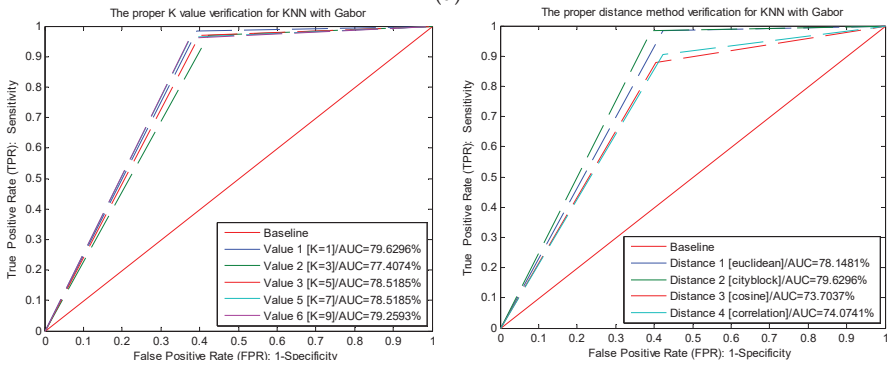
Figure 14. *Cont.*



(**c**)



(**d**)



(**e**)

Figure 14. K-value and distance metric for KNN and feature extraction techniques: (**a**) statistical, (**b**) histogram, (**c**) GLCM, (**d**) BGLAM and (**e**) Gabor.

3.3. Classification Based on SVM

The SVM algorithm is implemented in the FFB maturity classification of oil palm, and the input data include three sets of vectors in the n-dimensional SVM space. These data create a discrete hyper-plane in this space, which increases the "margin" between the three datasets and reduces the expected generalization error. In the case of oil palm FFB ripeness grading, three target categories exist, namely under-ripe, ripe and over-ripe. In this case, OAA is used, in which each class is split out and all the other classes are merged in the oil palm FFB grading system to solve multiclass issues with less computation time. An important aspect of the SVM model-building process is parameter tuning.

Three different types of kernel functions, linear, polynomial and radial (RBF), were used to perform the classification task. To control the trade-off between maximizing the headroom and minimizing the training error, the sigma of RBF was set from 1 to 100, while the polynomial distance was also set from 1 to 4. The magnitude of the regularization parameter C was set from 1 to 1000 for polynomial kernels and RBF, as explained in Table 4.

Table 4. Best results of RBF kernel function based on sigma values and c with FFB ripeness grading.

FET	RBF-Sigma	C	Accuracy %		
			ROI1	ROI2	ROI3
Statistical	1	1000	90	90	82
Histogram	50	100	89	90	91
GLCM	1	500	75	78	79
BGLAM	10	500	92	90	92
Gabor	10	500	76	87	89

Note: FET = Feature extraction techniques.

As shown in Figure 15, the kernel function provided a significantly higher accuracy rate for the FFB maturity classification of oil palm. The results are based on different values of sigma and c, as examined by other research [111], and a comparison of linear and nonlinear polynomial kernel functions. Therefore, as demonstrated in Table 4, the experimental results show that RBF-sigma = 10 with C = 500 provides the greatest results of 92% using BGLAM with ROI3 based on the SVM algorithm, as shown in Figure 15.

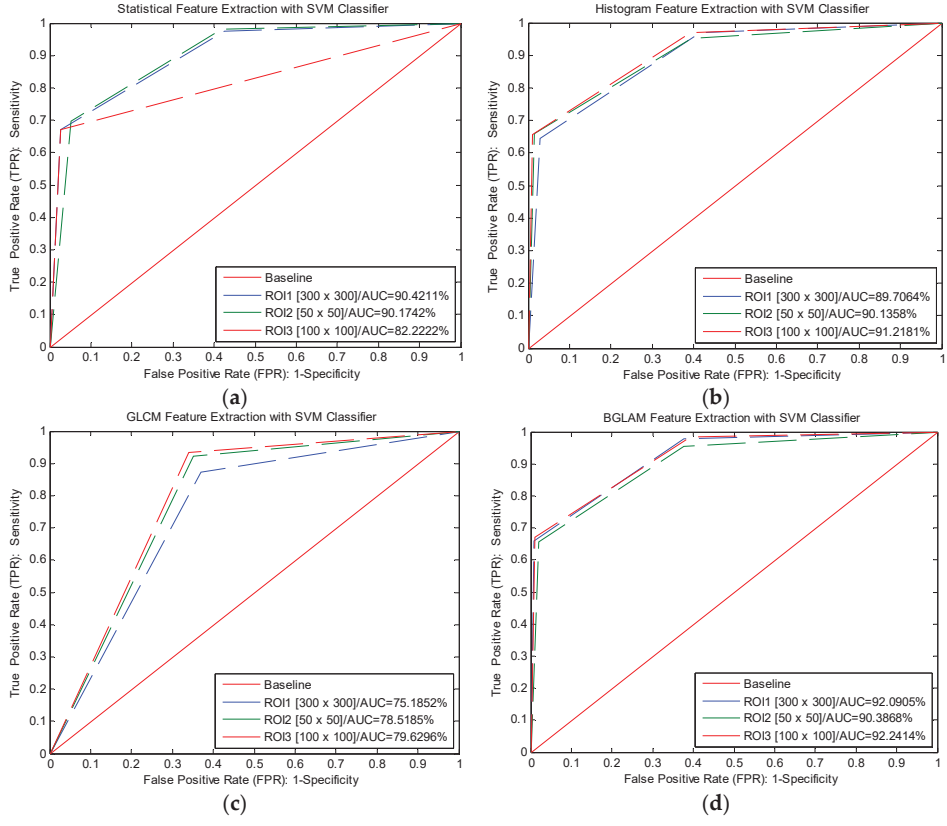
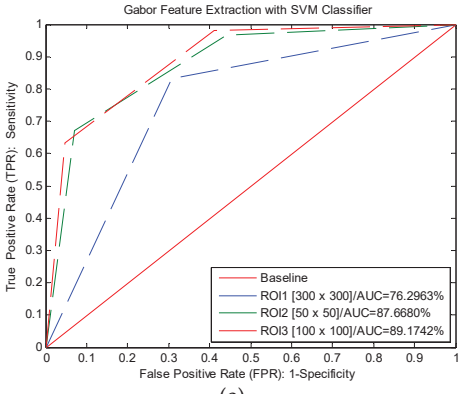


Figure 15. *Cont.*



(**e**)

Figure 15. RBF parameters for SVM with feature extraction techniques: (**a**) statistical feature extraction, (**b**) histogram feature extraction, (**c**) GLCM feature extraction, (**d**) BGLAM feature extraction and (**e**) Gabor feature extraction.

3.4. Experimental Results

Four experiments were carried out. In experiment 1, the texture characteristics of the oil palm were extracted and the classification was performed for the FFB type classification. In experiments 2, 3 and 4, oil palm color, texture and thorn features were extracted. The classification was then conducted for the Nigrescens, Oleifera and Virescens FFB ripeness grading.

The complete picture of the threshold between the sensitivity and 1- specificity is displayed by plotting the ROC curve across a series of threshold points. The AUC is considered to be an effective measurement of the inherent validity of a grading system test. This curve is suitable for (a) assessing the discriminatory ability of a test to pick correctly the under-ripe, ripe and over-ripe classes; (b) finding the optimal threshold point to minimize class misclassification and (c) comparing the efficacy of ROI1, ROI2 and ROI3 for assessing the same sample or class, as illustrated in Figure 16.

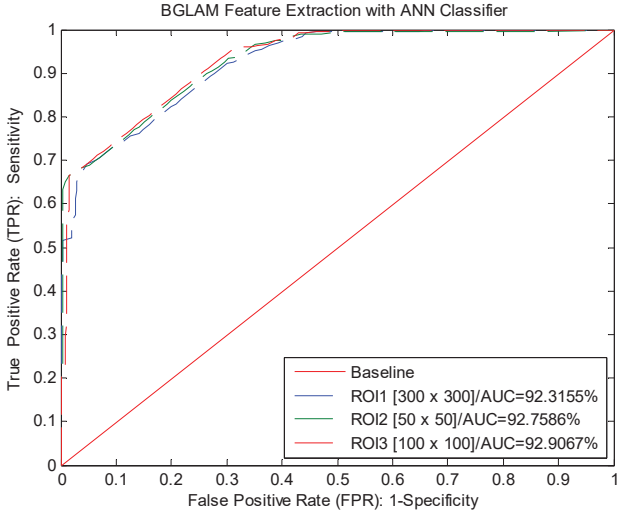


Figure 16. Oil palm FFB type classification based on BGLAM and ANN classifier.

3.4.1. FFB Type Grading System Results

The oil palm grading system was able to accurately classify the three different oil palm FFB types based on the external texture features and properties by using feature extraction techniques GLCM and BGLAM and supervised machine learning classifiers ANN, KNN and SVM, as critically explained in Table 5.

Table 5. Results of test computing FFB type classification based on GLCM and BGLAM using ANN, KNN and SVM.

Image Size	Classifiers	Texture Feature Extraction Techniques			
		GLCM		BGLAM	
		Testing Accuracy (%)	Time (s)	Testing Accuracy (%)	Time (s)
ROI1	ANN	89.00	4.08	91.00	1.02
	KNN	87.00	4.06	75.00	0.98
	SVM	79.00	4.7	90.00	1.74
ROI2	ANN	86.00	0.553	89.00	0.43
	KNN	84.00	0.50	76.00	0.38
	SVM	67.00	1.13	77.00	1.14
ROI3	ANN	86.00	0.56	93.00 **	0.44 **
	KNN	77.00	0.51	89.00	0.40
	SVM	69.00	1.22	82.00	1.32

Note: ** = The best result.

Table 5 indicates that the fastest and most accurate method and technique for the oil palm type grading system is the BGLAM feature extraction technique combined with the ANN supervised machine learning technique applied on pruning a 100 × 100-pixel FFB image with the ROI3. This finding achieved an optimal accuracy of 93.00% and an image processing speed of 0.44 s in the test performance.

3.4.2. FFB Ripeness Grading System Results

The maturity classification task was trained and tested for the three closest classes, over-ripe, ripe and under-ripe, based on the three FFB maturity models of oil palms: color, texture and thorns.

Color Model

The ripeness grading system testing performance based on the color model for different FFB image ROIs was evaluated. The results are clearly illustrated in Table 6.

Table 6 indicated the optimal methods and techniques that are the fastest and most accurate for the ripeness grading system. The data are based on the color histogram feature extracted combined with the ANN technique applied to the 100 × 100-pixel FFB image size with ROI3. The results achieved 93.00% accuracy and 1.6 s image processing speed in terms of testing performance for Nigrescens and Oleifera and 100%, 93% testing performance accuracy and 1.4 s image processing speed based on the ANN technique applied with ROI2. For Virescens, the statistical color feature accurately obtained 93% testing performance based on ANN for the different oil palm types. However, the results were limited by the slow processing time compared with the color histogram performance and the oil palm system objectives.

Table 6. Results of test computing FFB ripeness classification based on statistical and histogram using ANN, KNN and SVM.

T	C	Image Size	Color Feature Extraction Techniques			
			Statistical Color Feature		Color Histogram	
			Testing Accuracy (%)	Time (s)	Testing Accuracy (%)	Time (s)
T1	ANN	ROI1	93.00	3	92.00	2.7
		ROI2	92.00	2.5	92.00	1.4
		ROI3	92.00	2.65	94.00 **	1.6 **
	KNN	ROI1	81.00	2.3	81.00	2.5
		ROI2	82.00	1.2	82.00	1.2
		ROI3	82.00	1.02	82.00	1.3
	SVM	ROI1	81.00	5.3	81.00	7
		ROI2	81.00	4.5	82.00	6
		ROI3	80.00	5	80.00	7
T2	ANN	ROI1	93.00	3	93.00	2.7
		ROI2	92.00	2.5	93.00	1.4
		ROI3	91.00	2.65	94.00 **	1.6 **
	KNN	ROI1	92.00	2.3	93.00	2.5
		ROI2	91.00	1.2	90.00	1.2
		ROI3	92.00	1.02	92.00	1.3
	SVM	ROI1	90.00	5.3	90.00	7
		ROI2	90.00	4.5	90.00	6
		ROI3	82.00	5	91.00	7
T3	ANN	ROI1	91.00	3	90.00	2.7
		ROI2	88.00	2.5	93.00 **	1.4 **
		ROI3	90.00	2.65	92.00	1.6
	KNN	ROI1	74.00	2.3	78.00	2.5
		ROI2	74.00	1.2	79.00	1.2
		ROI3	87.00	1.02	79.00	1.3
	SVM	ROI1	69.00	5.3	78.00	7
		ROI2	73.00	4.5	78.00	6
		ROI3	72.00	5	79.00	7

Notes: T = Types, T1 = Nigrescens, T2 = Oleifera, T3 = Virescens, C = Classifier, ** = The best result.

Texture Model

The ripeness grading system testing performance based on the texture model for different FFB image ROIs was evaluated. The results are clearly illustrated in Table 7.

As indicated in Table 7, the fast and accurate method and techniques used for the oil palm FFB ripeness grading system based on the texture model were primarily the BGLAM combined with the ANN technique. This technique was applied to the ROI3 with 92.00% testing performance accuracy with a 0.43 s image processing speed for Nigrescens. Moreover, the BGLAM combined with the ANN technique applied to the ROI2 achieved 93.00% testing performance accuracy with a significant image processing speed of 0.40 s for Oleifera and Virescens. Due to the sensitivity of SVM to noise and the weakness of the Gabor wavelet and GLCM techniques with texture features and processing time, the limitations of these methods and techniques are clearly stated in the testing result tables.

Thorn Model

The ripeness grading system performance of the oil palm FFB types for testing based on the thorn model for the different ROIs was evaluated. The results are clearly illustrated in Table 8.

Table 7. Results of test computing FFB ripeness classification based on GLCM, GLAM and Gabor by using ANN, KNN and SVM.

T	C	Image Size	Texture Feature Extraction Techniques					
			GLCM		BGLAM		Gabor	
			Testing Accuracy (%)	Time (s)	Testing Accuracy (%)	Time (s)	Testing Accuracy (%)	Time (s)
T1	ANN	ROI1	91.00	3.6	89.00	1	86.00	1.03
		ROI2	89.00	1.7	90.00	0.40	81.00	0.43
		ROI3	91.00	2.2	92.00 **	0.43 **	80.00	0.44
	KNN	ROI1	79.00	2.8	80.00	1	74.00	1.04
		ROI2	77.00	1.5	82.00	0.39	77.00	0.39
		ROI3	78.00	1.6	81.00	0.40	77.00	0.41
	SVM	ROI1	80.00	7.7	80.00	1.9	76.00	1.97
		ROI2	76.00	5.5	81.00	1	82.00	1.79
		ROI3	79.00	6	81.00	0.85	74.00	2.00
T2	ANN	ROI1	92.00	3.6	92.00	1	83.00	1.03
		ROI2	91.00	1.7	93.00 **	0.40 **	79.00	0.43
		ROI3	92.00	2.2	93.00	0.43	83.00	0.44
	KNN	ROI1	79.00	2.8	90.00	1	78.00	1.04
		ROI2	80.00	1.5	91.00	0.39	90.00	0.39
		ROI3	91.00	1.6	92.00	0.40	79.00	0.41
	SVM	ROI1	75.00	7.7	92.00	1.9	76.00	1.97
		ROI2	79.00	5.5	90.00	1	88.00	1.79
		ROI3	80.00	6	92.00	0.85	89.00	2.00
T3	ANN	ROI1	87.00	3.6	88.00	1	82.00	1.03
		ROI2	89.00	1.7	93.00 **	0.40 **	81.00	0.43
		ROI3	86.00	2.2	91.00	0.43	77.00	0.44
	KNN	ROI1	76.00	2.8	77.00	1	74.00	1.04
		ROI2	72.00	1.5	79.00	0.39	73.00	0.39
		ROI3	72.00	1.6	90.00	0.40	78.00	0.41
	SVM	ROI1	65.00	7.7	76.00	1.9	72.00	1.97
		ROI2	64.00	5.5	79.00	1	67.00	1.79
		ROI3	67.00	6	80.00	0.85	68.00	2.00

Notes: T = Types, T1 = Nigrescens, T2 = Oleifera, T3 = Virescens, C = Classifier, ** = The best result.

Table 8. Results of test computing FFB ripeness classification with statistical, histogram, GLCM, BGLAM and Gabor using ANN, KNN and SVM.

T	Technique	Image Size					
		ROI1		ROI2		ROI3	
		Testing Accuracy (%)	Time (s)	Testing Accuracy (%)	Time (s)	Testing Accuracy (%)	Time (s)
T1	Statistical and KNN	79.00	4.06	78.00	2.3	79.00	2.5
	Histogram and SVM	80.00	9	81.00	8	81.00	8.5
	GLCM and ANN	87.00 **	3.7 **	84.00	2.6	85.00	2.9
	BGLAM and ANN	76.00	1.6	79.00	0.8	78.00	0.83
	Gabor and KNN	78.00	1.6	74.00	0.78	74.00	0.84
T2	Statistical and KNN	87.00	4.06	90.00	2.3	79.00	2.5
	Histogram and SVM	88.00	9	80.00	8	79.00	8.5
	GLCM and ANN	91.00	3.7	89.00	2.6	88.00	2.9
	BGLAM and SVM	91.00	2.5	89.00	1.36	91.00 **	1.20 **
	Gabor and KNN	89.00	1.6	82.00	0.78	86.00	0.84
T3	Statistical and ANN	84.00	4.06	86.00	2.3	87.00	2.5
	Histogram and SVM	73.00	9	77.00	8	79.00	8.5
	GLCM and ANN	88.00	3.7	84.00	2.6	87.00	8.5
	BGLAM and ANN	84.00	1.6	82.00	0.8	87.00 **	0.83 **
	Gabor and KNN	74.00	1.6	73.00	0.78	78.00	0.84

Notes: T = Types, T1 = Nigrescens, T2 = Oleifera, T3 = Virescens, ** = The best result.

Due to data noise, the thorn model shows poor results based on performance and processing time, while BGLAM combined with ANN technology with SVM technology applied to ROI3 achieved a test performance of 91.00% and an Oleifera image processing speed of 1.20 s.

4. Conclusions

An FFB fruit palm oil ripeness classification system was designed based on remote sensing sensors (CCD camera) and image processing technologies as computer vision applications for inspection of agricultural crop quality.

The system aims to ensure the maturity class of different types of FFB palm oil based on external characteristics, such as color, texture and thorns. Image processing methods and techniques, including the acquisition and segmentation of images in ROI1, ROI2 and ROI3 and the extraction of image properties as a function of the statistical function of the color, histogram color, GLCM, BGLAM and Gabor wavelet, were implemented.

In addition, decision-making for image classification through training and testing of the system based on the different algorithms, SVM, KNN and ANN, was implemented on a maturity classification system. The training and testing of oil palm FFB species (Nigrescens, Oleifera and Virescens) and maturation (under-ripe, ripe and over-ripe) depending on the color, texture and pattern of the thorns were extracted.

AUC and ROC were used to accurately estimate and evaluate the performance of different classifiers based on system performance, processing time and system cost. The results showed that the texture models were improved with ANN classifiers as the best result of the algorithm classifier, ANN-based BGLAM with ROI3, provides 93.00% accuracy with a shorter image processing time of 0.44 (s) for FFB type recognition. Meanwhile, the BGLAM algorithm that relies on ANN and ROI3 obtained 92.00% accuracy and a short processing time of 0.43 (s) for Nigrescens, plus the algorithm BGLAM based on ANN and ROI2 obtained 93.00% accuracy and a short processing time of 0.40 (s) for Oleifera and Virescens for maturity classification.

In the final analysis, different predictions were used. Maximum accuracy was obtained using an ANN classifier with the highest prediction accuracy observed compared to all the other classifiers. The following more accurate prediction is indicated by the different classifiers: KNN and SVM, respectively. The scope of the existing work is limited to investigation of the relationship between oil palm fruit ripeness level and image processing approach and AI.

As mentioned above, the authors have implemented several experiments based on different methods and techniques for automation of a real-time oil palm FFB ripeness grading system that carried out satisfactory results, but, in the future, the existing work can be extended to include some recommended practical actions and scientific studies of the system's hardware and software for developing the current system and improving the results.

In terms of hardware development, proper hardware design and development make it easier for the programmer to set his algorithm for a high-accuracy performance result. (1) Since the illumination system is one of the most important hardware parts in the oil palm grading system, in order to control the lighting beam incident on FFB to be reflected to the camera, a linear polarizer (LP) filter should be fixed at the camera and each light source, and (2) it is important to use other types of sensors, such as a thermal camera, to collect valuable information about the oil palm FFB ripeness and build grading system models based on the obtained information.

Regarding software development, the real-time oil palm FFB ripeness grading system was implemented as a solution for effective oil palm FFB ripeness grading. However, in order to improve the oil palm system functionality and performance, different methods and techniques should be proposed based on the system software, such as (1) using the oil palm FFB internal feature lab analysis information, such as oil content and free fatty acid, to correlate with external features of FFB, such as color and texture features, to validate and support the oil palm FFB ripeness result; (2) applying the retrieval methods and techniques as a decision-making system based on the similarity calculation as proposed and illustrated in Figure 5; (3) further research is needed to generalize the system for other agriculture applications by considering the size, weight and shape of FFB during the system design. Hence, that assembles the system to be a multipurpose application system, which can be

used in similar applications for different agricultural crops. Although the existing study utilizes image processing, similar results are expected to be obtained using a portable device. The proposed method has the potential to be a rapid on-site assessment tool for ripeness classification in the oil palm industry.

Author Contributions: Conceptualization, M.S.M.A., A.R.M.S., S.K.-B., M.H.B.M., O.M.B.S. and A.M.; methodology, M.S.M.A., A.R.M.S., S.K.-B., M.H.B.M., O.M.B.S. and A.M.; software, M.S.M.A. and A.R.M.S.; validation, M.S.M.A. and A.R.M.S.; formal analysis, M.S.M.A., A.R.M.S., M.H.B.M. and O.M.B.S.; investigation, M.S.M.A., A.R.M.S., S.K.-B., M.H.B.M., O.M.B.S. and A.M.; data curation, M.S.M.A., A.R.M.S. and O.M.B.S.; writing—original draft preparation, M.S.M.A. and A.R.M.S.; writing—review and editing, M.S.M.A., A.R.M.S., S.K.-B., M.H.B.M., O.M.B.S. and A.M.; supervision, A.R.M.S. and M.H.B.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Science, Technology and Innovation Malaysia (MOSTI); grant titled "Development of an Oil Palm Fresh Fruit Bunches (FFB) Image Analyser" (Grant Number 5450426) is hereby acknowledged in supporting this research. Publication of this paper was supported by: Universiti Putra Malaysia Journal Publication Fund (JPF) administered by the Research Planning & Knowledge Management Division, Research Management Centre (RMC), Universiti Putra Malaysia.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: All data will be made available on request to the corresponding author's email with appropriate justification.

Acknowledgments: The authors would like to thank the Libyan Embassy in Malaysia, University Putra Malaysia (UPM), Faculty of Engineering, Geospatial Information Science Research Centre (GISRC) and Department of Biological and Agricultural Engineering for providing support, infrastructure and laboratory facilities. The authors thank Research Station—Kluang, Malaysian Palm Oil Board (MPOB), Sime Darby Plantation Sdn. Bhd, Agriculture Park UPM (Taman Pertanian Universiti) and Spatial Research Group, UPM, for assistance.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. NOAA. What Is Remote Sensing? National Ocean Service Website. 25 June 2018. Available online: https://oceanservice.noaa.gov/facts/remotesensing.html (accessed on 26 February 2021).
2. USGS. What Is Remote Sensing and What Is It Used for? Mapping, Remote Sensing, and Geospatial Data. 18 August 2016. Available online: https://www.usgs.gov/faqs/what-remote-sensing-and-what-it-used?qt-_news_science_products=7&qt-news_science_products=7#qt-news_science_products (accessed on 30 September 2019).
3. Cracknell, A.; Hayes, L. Introduction to remote sensing. In *Geocarto International. 40*; Taylor & Francis: London, UK, 2008; Volume 7.
4. Cracknell, A.P. The development of remote sensing in the last 40 years. *Int. J. Remote Sens.* **2018**, *39*, 8387–8427. [CrossRef]
5. Murai, S. *Remote Sensing Notes*; Sensing, J.A.o.R., Ed.; National Space Development Agency of Japan (NASDA): Tokyo, Japan; Remote Sensing Technology Center of Japan (RESTEC): Tokyo, Japan, 1999.
6. Richards, J.A. *Remote Sensing Digital Image Analysis—An Introduction*, 5th ed.; Springer: Berlin/Heidelberg, Germany, 2013.
7. Tempfli, K.; Huurneman, G.C.; Bakker, W.H.; Janssen, L.L.F.; Feringa, W.F.; Gieske, A.S.M.; Grabmaier, K.A.; Hecker, C.A.; Horn, J.A.; Kerle, N.; et al. *Principles of Remote Sensing-An Introductory Textbook*, 4th ed.; Janssen, L.L.F., Huurneman, G.C., Eds.; The International Institute for Aerospace Survey and Earth Sciences (ITC): Enschede, The Netherlands, 1999; Volume 2, p. 591.
8. Li, X.; Gu, X.; Yu, T.; Cheng, T.; Li, J.; Gao, H.; Wang, Z. Atmospheric scattering and turbulence modulation transfer function for CCD cameras on CBERS-02b and HJ-1A/1B. *Int. J. Remote Sens.* **2012**, *33*, 1413–1427. [CrossRef]
9. Demircan, A.; Geiger, B.; Radke, M.; Von Schönermark, M. Bi-directional reflectance measurements with the CCD line camera WAAC. *Remote Sens. Rev.* **2009**, *19*, 95–110. [CrossRef]
10. Hinkler, J.; Pedersen, S.B.; Rasch, M.; Hansen, B.U. Automatic snow cover monitoring at high temporal and spatial resolution, using images taken by a standard digital camera. *Int. J. Remote Sens.* **2010**, *23*, 4669–4682. [CrossRef]
11. Harrison, B.A.; Jupp, D.L.B.; Hutton, P.G.; Mayo, K.K. Accessing remote sensing technology The microBRIAN example. *Int. J. Remote Sens.* **2007**, *10*, 301–309. [CrossRef]
12. Kherwa, P.; Ahmed, S.; Berry, P.; Khurana, S.; Singh, S.; Sen, J.; Mehtab, S.; Cadotte, D.W.W.; Anderson, D.W.; Ost, K.J.; et al. Machine Learning Algorithms, Models and Applications. In *Artificial Intelligence*; Sen, J., Ed.; IntechOpen: London, UK, 2022; p. 155.
13. Khan, H.; Haq, I.U.; Munsif, M.; Khan, S.U.; Lee, M.Y. Automated Wheat Diseases Classification Framework Using Advanced Machine Learning Technique. *Agriculture* **2022**, *12*, 1226. [CrossRef]

14. Kuan, C.-H.; Leu, Y.; Lin, W.-S.; Lee, C.-P. The Estimation of the Long-Term Agricultural Output with a Robust Machine Learning Prediction Model. *Agriculture* **2022**, *12*, 1075. [CrossRef]
15. Kaya, F.; Keshavarzi, A.; Francaviglia, R.; Kaplan, G.; Başayiğit, L.; Dedeoğlu, M. Assessing Machine Learning-Based Prediction under Different Agricultural Practices for Digital Mapping of Soil Organic Carbon and Available Phosphorus. *Agriculture* **2022**, *12*, 1062. [CrossRef]
16. Wüthrich, M.V. Machine learning in individual claims reserving. *Scand. Actuar. J.* **2018**, *2*, 465–480. [CrossRef]
17. Qiu, D. Individual Claims Reserving: Using Machine Learning Methods. In *Mathematics and Statistics*; Concordia University: Montreal, QC, Canada, 2019; p. 90.
18. Härkönen, V. On Claims Reserving with Machine Learning Techniques. In *Mathematical Statistics*; Stockholms Universitet: Stockholm, Sweden, 2021; p. 69.
19. Liu, X.; He, L.; He, Z.; Wei, Y. Estimation of Broadleaf Tree Canopy Height of Wolong Nature Reserve Based on InSAR and Machine Learning Methods. *Forests* **2022**, *13*, 1282. [CrossRef]
20. Maxwell, A.E.; Warner, T.A.; Fang, F. Implementation of machine-learning classification in remote sensing: An applied review. *Int. J. Remote Sens.* **2018**, *39*, 2784–2817. [CrossRef]
21. Dawid, L.; Tomza, M.; Dawid, A. Estimation of Usable Area of Flat-Roof Residential Buildings Using Topographic Data with Machine Learning Methods. *Remote Sens.* **2019**, *11*, 2382. [CrossRef]
22. Zhao, W.; Guo, Z.; Yue, J.; Zhang, X.; Luo, L. On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery. *Int. J. Remote Sens.* **2015**, *36*, 3368–3379. [CrossRef]
23. Yue, J.; Zhao, W.; Mao, S.; Liu, H. Spectral–spatial classification of hyperspectral images using deep convolutional neural networks. *Remote Sens. Lett.* **2015**, *6*, 468–477. [CrossRef]
24. Zhao, W.; Li, S.; Li, A.; Zhang, B.; Li, Y. Hyperspectral images classification with convolutional neural network and textural feature using limited training samples. *Remote Sens. Lett.* **2019**, *10*, 449–458. [CrossRef]
25. Goh, J.Q.; Shariff, A.R.M.; Nawi, N.M. Application of Optical Spectrometer to Determine Maturity Level of Oil Palm Fresh Fruit Bunches Based on Analysis of the Front Equatorial, Front Basil, Back Equatorial, Back Basil and Apical Parts of the Oil Palm Bunches. *Agriculture* **2021**, *11*, 1179. [CrossRef]
26. Pérez-Pérez, B.D.; Vázquez, J.P.G.; Salomón-Torres, R. Evaluation of Convolutional Neural Networks' Hyperparameters with Transfer Learning to Determine Sorting of Ripe Medjool Dates. *Agriculture* **2021**, *11*, 115. [CrossRef]
27. Mesa, A.R.; Chiang, J.Y. Multi-Input Deep Learning Model with RGB and Hyperspectral Imaging for Banana Grading. *Agriculture* **2021**, *11*, 687. [CrossRef]
28. Zhao, F.; Yang, G.; Yang, H.; Long, H.; Xu, W.; Zhu, Y.; Meng, Y.; Han, S.; Liu, M. A Method for Prediction of Winter Wheat Maturity Date Based on MODIS Time Series and Accumulated Temperature. *Agriculture* **2022**, *12*, 945. [CrossRef]
29. Hufkens, K.; Melaas, E.K.; Mann, M.L.; Foster, T.; Ceballos, F.; Robles, M.; Kramer, B. Monitoring crop phenology using a smartphone based near-surface remote sensing approach. *Agric. For. Meteorol.* **2019**, *265*, 327–337. [CrossRef]
30. Zhong, Y.; Ma, A.; Ong, Y.S.; Zhu, Z.; Zhang, L. Computational intelligence in optical remote sensing image processing. *Appl. Soft Comput.* **2018**, *64*, 75–93. [CrossRef]
31. Alfatni, M.S.; Shariff, A.R.M.; Abdullah, M.Z.; Saeed, O.; Ceesay, O.M. Recent Methods and Techniques of External Grading Systems for Agricultural Crops Quality Inspection—Review. *Int. J. Food Eng.* **2011**, *7*, 1–40. [CrossRef]
32. Malamasa, E.N.; Petrakis, E.G.M.; Zervakis, M.; Petit, L.; Legat, J.D. A survey on industrial vision systems, applications and tools. *Image Vis. Comput.* **2003**, *21*, 171–188. [CrossRef]
33. Pamornnak, B.; Limsiroratana, S.; Khaorapapong, T.; Chongcheawchamnan, M.; Ruckelshausen, A. An automatic and rapid system for grading palm bunch using a Kinect camera. *Comput. Electron. Agric.* **2017**, *143*, 227–237. [CrossRef]
34. Prakasa, E.; Rosiyadi, D.; Ni'mah, D.F.I. Automatic Region-of-Interest Selection for Corn Seed Grading. In Proceedings of the International Conference on Computer, Control, Informatics and its Applications (IC3INA), Jakarta, Indonesia, 23–26 October 2017; pp. 23–28.
35. López, Y.Y.; Martínez-García, A.; Gómez, S.J. Apple quality study using fringe projection and colorimetry techniques. *Opt.—Int. J. Light Electron Opt.* **2017**, *147*, 401–413. [CrossRef]
36. Tretola, M.; Ottoboni, M.; Di Rosa, A.R.; Giromini, C.; Fusi, E.; Rebucci, R.; Leone, F.; Dell'Orto, V.; Chiofalo, V.; Pinotti, L. Former Food Products Safety Evaluation: Computer Vision as an Innovative Approach for the Packaging Remnants Detection. *J. Food Qual.* **2017**, *2017*, 1–6. [CrossRef]
37. Sabri, N.; Ibrahim, Z.; Syahlan, S.; Jamil, N.; Mangshor, N.N.A. Palm Oil Fresh Fruit Bunch Ripeness Grading Identification Using Color Features. *J. Fundam. Appl. Sci.* **2017**, *9*, 563–579. [CrossRef]
38. Khoje, S.A.; Bodhe, S.K. A Comprehensive Survey of Fruit Grading Systems for Tropical Fruits of Maharashtra. *J. Crit. Rev. Food Sci. Nutr.* **2015**, *55*, 1658–1671. [CrossRef] [PubMed]
39. Beek, J.V.; Tits, L.; Somers, B.; Deckers, T.; Verjans, W.; Bylemans, D.; Janssens, P.; Coppin, P. Temporal Dependency of Yield and Quality Estimation through Spectral Vegetation Indices in Pear Orchards. *Remote Sens.* **2015**, *7*, 9886–9903. [CrossRef]
40. Wang, P.; Niu, T.; He, D. Tomato Young Fruits Detection Method under Near Color Background Based on Improved Faster R-CNN with Attention Mechanism. *Agriculture* **2021**, *11*, 1059. [CrossRef]
41. Plasquy, E.; Garcia, J.M.; Florido, M.C.; Sola-Guirado, R.R. Estimation of the Cooling Rate of Six Olive Cultivars Using Thermal Imaging. *Agriculture* **2021**, *11*, 164. [CrossRef]

42. J Bird, J.; Barnes, C.M.; Manso, L.J.; Ekárt, A.; Faria, D.R. Fruit quality and defect image classification with conditional GAN data augmentation. *Sci. Hortic.* **2022**, *293*, 110684. [CrossRef]
43. Leemans, V.; Destain, M.-F. A real-time grading method of apples based on features extracted from defects. *J. Food Eng.* **2004**, *61*, 83–89. [CrossRef]
44. Njoroge, J.B.; Ninomiya, K.; Kondo, N.; Toita, H. Automated Fruit Grading System using Image Processing. In Proceedings of the 41st SICE Annual Conference. SICE 2002, Osaks, Japan, 5–7 August 2002; pp. 1346–1351.
45. Thang, Y.M.; A Ariffin, A.; Appleton, D.R.; Asis, A.J.; Mokhtar, M.N.; Yunus, R. Determination of sugars composition in abscission zone of oil palm fruit. *Ser. Mater. Sci. Eng.* **2017**, *206*, 12034. [CrossRef]
46. Xuan, G.; Gao, C.; Shao, Y. Spectral and image analysis of hyperspectral data for internal and external quality assessment of peach fruit. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2022**, *272*, 121016. [CrossRef] [PubMed]
47. Chuah, H.T.; Kam, S.W.; Chye, Y.H. Microwave dielectric properties of rubber and oil palm leaf samples: Measurement and modelling, International Journal of Remote Sensing. *Int. J. Remote Sens.* **1997**, *18*, 2623–2639. [CrossRef]
48. Tan, K.P.; Kanniah, K.D.; Cracknell, A.P. On the upstream inputs into the MODIS primary productivity products using biometric data from oil palm plantations. *Int. J. Remote Sens.* **2014**, *35*, 2215–2246. [CrossRef]
49. Hamsa, C.S.; Kanniah, K.D.; Muharam, F.M.; Idris, N.H.; Abdullah, Z.; Mohamed, L. Textural measures for estimating oil palm age. *Int. J. Remote Sens.* **2019**, *40*, 7516–7537. [CrossRef]
50. Yu, H.; Yang, W.; Xia, G.-S.; Liu, G. A Color-Texture-Structure Descriptor for High-Resolution Satellite Image Classification. *Remote Sens.* **2016**, *8*, 259.
51. De-la-Torre, M.; Zatarain, O.; Avila-George, H.; Muñoz, M.; Oblitas, J.; Lozada, R.; Mejía, J.; Castro, W. Multivariate Analysis and Machine Learning for Ripeness Classification of Cape Gooseberry Fruits. *Processes* **2019**, *7*, 928.
52. Xiaobo, Z.; Jiewen, Z.; Yanxiao, L. Apple color grading based on organization feature parameters. *Pattern Recognit. Lett.* **2007**, *28*, 2046–2053. [CrossRef]
53. Cárdenas-Pérez, S.; Chanona-Pérez, J.; Méndez-Méndez, J.V.; Calderón-Domínguez, G.; López-Santiago, R.; Perea-Flores, M.J.; Arzate-Vázquez, I. Evaluation of the ripening stages of apple (Golden Delicious) by means of computer vision system. *Biosyst. Eng.* **2017**, *159*, 46–58. [CrossRef]
54. Bhargava, A.; Bansal, A. Machine learning based quality evaluation of mono-colored apples. *Multimed. Tools Appl.* **2020**, *79*, 22989–23006. [CrossRef]
55. Wu, L.; Zhang, H.; Chen, R.; Yi, J. Fruit Classification using Convolutional Neural Network via Adjust Parameter and Data Enhancement. In Proceedings of the 12th International Conference on Advanced Computational Intelligence (ICACI), Dali, China, 14–16 March 2020.
56. Behera, S.K.; Rath, A.K.; Sethy, P.K. Maturity status classification of papaya fruits based on machine learning and transfer learning approach. *Inf. Processing Agric.* **2020**, *8*, 244–250. [CrossRef]
57. Guerrero, E.R.; Benavides, G.M. Automated system for classifying Hass avocados based on image processing techniques. In Proceedings of the 2014 IEEE Colombian Conference on Communications and Computing (COLCOM), Bogota, Colombia, 4–6 June 2014.
58. Khisanudin, I.S. Dragon Fruit Maturity Detection Based-HSV Space Color Using Naive Bayes Classifier Method. *IOP Conf. Series Mater. Sci. Eng.* **2020**, *771*, 1–6. [CrossRef]
59. Mendoza, F.; Aguilera, J.M. Application of Image Analysis for Classification of Ripening Bananas. *J. Food Sci.* **2004**, *69*, 471–477. [CrossRef]
60. Paulraj, M.; Hema, C.R.; Sofiah, S.; Radzi, M. Color Recognition Algorithm using a Neural Network Model in Determining the Ripeness of a Banana. In Proceedings of the International Conference on Man-Machine Systems (ICoMMS), Batu Ferringhi, Penang, Malaysia, 11–13 October 2009; pp. 2B71–2B74.
61. Li, H.; Lee, W.S.; Wang, K. Identifying blueberry fruit of different growth stages using natural outdoor color images. *Comput. Electron. Agric.* **2014**, *106*, 91–101. [CrossRef]
62. Pourdarbani, R.; Ghassemzadeh, H.R.; Seyedarabi, H.; Nahand, F.Z.; Vahed, M.M. Study on an automatic sorting system for Date fruits. *J. Saudi Soc. Agric. Sci.* **2015**, *14*, 83–90. [CrossRef]
63. Damiri, D.J.; Slamet, C. Application of Image Processing and Artificial Neural Networks to Identify Ripeness and Maturity of the Lime (citrus medica). *Int. J. Basic Appl. Sci.* **2012**, *1*, 171–179. [CrossRef]
64. Nandi, C.S.; Tudu, B.; Koley, C. A Machine Vision-Based Maturity Prediction System for Sorting of Harvested Mangoes. *IEEE Trans. Instrum. Meas.* **2014**, *63*, 1722–1730. [CrossRef]
65. Vélez-Rivera, N.; Blasco, J.; Chanona-Pérez, J.; Calderón-Domínguez, G.; Perea, M.D.J.; Arzate-Vázquez, I.; Cubero, S.; Farrera-Rebollo, R. Computer Vision System Applied to Classification of "Manila" Mangoes During Ripening Process. *Food Bioprocess Technol.* **2014**, *7*, 1183–1194. [CrossRef]
66. Zheng, H.; Lu, H. A least-squares support vector machine (LS-SVM) based on fractal analysis and CIELab parameters for the detection of browning degree on mango (*Mangifera indica* L.). *Comput. Electron. Agric.* **2012**, *83*, 47–51. [CrossRef]
67. Fadilah, N.; Mohamad-Saleh, J.; Halim, Z.A.; Ibrahim, H.; Ali, S.S.S. Intelligent Color Vision System for Ripeness Classification of Oil Palm Fresh Fruit Bunch. *Sensors* **2012**, *12*, 14179–14195. [CrossRef] [PubMed]
68. Elhariri, E.; El-Bendary, N.; Hussein, A.M.; Hassanien, A.E.; Badr, A. Bell pepper ripeness classification based on support vector machine. In Proceedings of the in 2nd International Conference on Engineering and Technology, Cairo, Egypt, 19–21 August 2014.

69. Rahman, M.O.; Hussain, A.; Basri, H. Automated sorting of recycled paper using smart image processing. *At-Automatisierungstechnik* **2020**, *68*, 277–293. [CrossRef]
70. Mohammadi, V.; Kheiralipour, K.; Ghasemi-Varnamkhasti, M. Detecting maturity of persimmon fruit based on image processing technique. *Sci. Hortic.* **2015**, *184*, 123–128. [CrossRef]
71. El-Bendary, N.; El Hariri, E.; Hassanien, A.E.; Badr, A. Using machine learning techniques for evaluating tomato ripeness. *Expert Syst. Appl.* **2015**, *42*, 1892–1905. [CrossRef]
72. Goel, N.; Sehgal, P. Fuzzy classification of pre-harvest tomatoes for ripeness estimation–An approach based on automatic rule learning using decision tree. *Appl. Soft Comput.* **2015**, *36*, 45–56. [CrossRef]
73. Polder, G.; der Heijden, G.v. Measuring ripening of tomatoes using imaging spectrometry. In *Hyperspectral Imaging for Food Quality Analysis and Control*; Elsevier: Amsterdam, The Netherlands, 2010; pp. 369–402.
74. Rafiq, A.; Makroo, H.A.; Hazarika, M.K. Neural Network-Based Image Analysis for Evaluation of Quality Attributes of Agricultural Produce. *Food Processing Preserv.* **2016**, *40*, 1010–1019. [CrossRef]
75. Ashraf, T.; NiazKhan, Y. Weed density classification in rice crop using computer vision. *Comput. Electron. Agric.* **2020**, *175*, 105590. [CrossRef]
76. Abdulhamid, U.F.; Aminu, M.A.; Daniel, S. Detection of Soya Beans Ripeness Using Image Processing Techniques and Artificial Neural Network. *Asian J. Phys. Chem. Sci.* **2018**, *5*, 1–9. [CrossRef]
77. Hadfi, I.H.; Yusoh, Z.I.M. Banana ripeness detection and servings recommendation system using artificial intelligence techniques. *J. Telecommun. Electron. Comput. Eng.* **2018**, *10*, 83–87.
78. Nagvanshi, S.; Goswami, T.K. Development of a system to measure color in fresh and microwave dried banana slices. *J. Food Sci. Technol.* **2020**, *41*, 1673–1681.
79. Rizam, S.; Yasmin, F.; Ihsan, A.; Shazana, K. Non-destructive Watermelon Ripeness Determination Using Image Processing and Artificial Neural Network (ANN). *Int. J. Comput. Inf. Eng.* **2009**, *3*, 332–336.
80. Abdullah, N.E.; Madzhi, N.K.; Yahya, A.M.A.A.; Rahim, A.A.A.; Rosli, A.D. Diagnostic System for Various Grades of Yellow Flesh Watermelon based on the Visible light and NIR properties. In Proceedings of the 4th International Conference on Electrical, Electronics and System Engineering (ICEESE), Kuala Lumpur, Malaysia, 8–9 November 2018.
81. Syazwan, N.A.; Rizam, M.S.B.S.; Nooritawati, M.T. Categorization of watermelon maturity level based on rind features. *Procedia Eng.* **2012**, *41*, 1398–1404. [CrossRef]
82. Skolik, P.; Morais, C.L.M.; Martin, F.L.; McAinsh, M.R. Determination of developmental and ripening stages of whole tomato fruit using portable infrared spectroscopy and Chemometrics. *BMC Plant Biol.* **2019**, *19*, 236. [CrossRef] [PubMed]
83. Du, D.; Wang, J.; Wang, B.; Zhu, L.; Hong, X. Ripeness Prediction of Postharvest Kiwifruit Using a MOS E-Nose Combined with Chemometrics. *Sensors* **2019**, *19*, 419. [CrossRef] [PubMed]
84. Ramos, P.J.; Avendaño, J.; Prieto, F.A. Measurement of the ripening rate on coffee branches by using 3d images in outdoor environments. *Comput. Ind.* **2018**, *99*, 83–95. [CrossRef]
85. Costa, A.G.; De Sousa, D.A.G.; Paes, J.L.; Cunha, J.P.B.; De Oliveira, M.V.M. Classification of Robusta Coffee Fruits at Different Maturation Stages Using Colorimetric Characteristics. *Eng. Agrícola Jaboticabal* **2020**, *40*, 518–525. [CrossRef]
86. Castro, W.; Oblitas, J.; De-La-Torre, M.; Cotrina, C.; Bazan, K.; Avila-George, H. Classification of Cape Gooseberry Fruit According to its Level of Ripeness Using Machine Learning Techniques and Different Color Spaces. *IEEE Access* **2019**, *7*, 27389–27400. [CrossRef]
87. De-la-Torre, M.; Avila-George, H.; Oblitas, J.; Castro, W. Selection and Fusion of Color Channels for Ripeness Classification of Cape Gooseberry Fruits. In *Trends and Applications in Software Engineering*; Part of the Advances in Intelligent Systems and Computing Book Series; Mejia, J., Muñoz, M., Rocha, Á., Calvo-Manzano, A.J., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; Volume 1071, pp. 219–233.
88. Brosnan, T.; Sun, D.-W. Improving quality inspection of food products by computer vision—A review. *J. Food Eng.* **2004**, *61*, 3–16. [CrossRef]
89. Chithra, P.L.; Henila, M. Defect Identification in the Fruit Apple Using K-Means Color Image Segmentation Algorithm. *Int. J. Adv. Res. Comput. Sci.* **2017**, *8*, 381–388. [CrossRef]
90. Riese, F.M.; Keller, S.; Hinz, S. Supervised and Semi-Supervised Self-Organizing Maps for Regression and Classification Focusing on Hyperspectral Data. *Remote Sens.* **2020**, *12*, 7. [CrossRef]
91. Alfatni, M.S.M.; Shariff, A.R.M.; Abdullah, M.Z.; Marhaban, M.H.; Shafie, S.B.; Bamiruddin, M.D.; Saaed, O.M.B. Oil palm fresh fruit bunch ripeness classification based on rule-based expert system of ROI image processing technique results. *IOP Conf. Ser. Earth Environ. Sci.* **2014**, *20*, 12018. [CrossRef]
92. Guo, R.; Liu, J.; Li, N.; Liu, S.; Chen, F.; Cheng, B.; Duan, J.; Li, X.; Ma, C. Pixel-Wise Classification Method for High Resolution Remote Sensing Imagery Using Deep Neural Networks. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 110. [CrossRef]
93. Dimililer, K.; Bush, I.J. Automated Classification of Fruits: Pawpaw Fruit as a Case Study. In *Advances in Intelligent Systems and Computing*; Springer: Cham, Switzerland, 2017; pp. 365–374.
94. Mekhalfa, F.; Nacereddine, N. Gentle Adaboost algorithm for weld defect classification. In *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*; IEEE: Poznan, Poland, 2017.
95. Iqbal, S.M.; Gopal, A.; Sankaranarayanan, P.; Nair, A.B. Classification of Selected Citrus Fruits Based on Color Using Machine Vision System. *Int. J. Food Prop.* **2016**, *19*, 272–288. [CrossRef]

96. Sudheer, K.P.; Gowda, P.; Chaubey, I.; Howell, T. Artificial Neural Network Approach for Mapping Contrasting Tillage Practices. *Remote Sens.* **2010**, *2*, 579–590. [CrossRef]
97. Yang, C.-C.; Prasher, S.O.; Landry, J.-A.; Ramaswamy, H.S.; Ditommaso, A. Application of artificial neural networks in image recognition and classification of crop and weeds. *Can. Agric. Eng.* **2000**, *42*, 147–152.
98. Unay, D. Multispectral Image Processing and Pattern Recognition Techniques for Quality Inspection of Apple Fruits. In *Facult'e Polytechnique de Mons in Applied Sciences*; Facult'e Polytechnique de Mons: Mons, Belgium, 2006; p. 159.
99. Ranjbarardestani, M. Determining the ripeness of fruit juices based on image processing technology and neural network classification. *Eur. Online J. Nat. Soc. Sci.* **2016**, *5*, 846–850.
100. Rafiq, M.Y.; Bugmann, G.; Easterbrook, D.J. Neural Network Design for Enginnering Applications. *Comput. Struct.* **2001**, *79*, 1541–1552. [CrossRef]
101. Deck, S.H.; Morrow, C.T.; Heinemann, P.H.; Iii, H.J.S. Comparison of a neural network and traditional classifier for machine vision inspection of potatoes. *Appl. Eng. Agric.* **1995**, *11*, 319–326. [CrossRef]
102. Schmoldt, D.L.; Li, P.; Abbott, A.L. Machine vision using artificial neural networks with local 3D neighbourhoods. *Comput. Electron. Agric.* **1997**, *16*, 225–271. [CrossRef]
103. Timmermans, A.J.M.; Hulzebosch, A.A. Computer vision system for on-line sorting of pot plants using an artificial neural network classifier. *Comput. Electron. Agric.* **1996**, *15*, 41–55. [CrossRef]
104. Wang, D.; Dowell, F.E.; Lacey, R.E. Single wheat kernel color classification using neural networks. *Trans. ASAE* **1999**, *42*, 233–240. [CrossRef]
105. Kotsiantis, S.B. Supervised Machine Learning: A Review of Classification Techniques. *Informatica* **2007**, *31*, 249–268.
106. Khan, M.A.M. Fast Distance Metric Based Data Mining Techniques Using P-trees: K-Nearest-Neighbor Classification and k-Clustering. In *Computer Science*; North Dakota State University of Agriculture and Applied Science: Fargo, ND, USA, 2001; p. 67.
107. Sudha, L.R.; Bhavani, R. Gait based Gender Identification using Statistical Pattern Classifiers. *Int. J. Comput. Appl.* **2012**, *40*, 30–35. [CrossRef]
108. Vapnik, V.N. (Ed.) *Statistical Learning Theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control*; Wiley: New York, NY, USA, 1998; p. 768.
109. Nayak, J.; Naik, B.; Behera, H.S. A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications & Challenges. *Int. J. Database Theory Appl.* **2015**, *8*, 169–186.
110. Arabameri, A.; Roy, J.; Saha, S.; Blaschke, T.; Ghorbanzadeh, O.; Bui, D.T. Application of Probabilistic and Machine Learning Models for Groundwater Potentiality Mapping in Damghan Sedimentary Plain, Iran. *Remote Sens.* **2019**, *11*, 3035. [CrossRef]
111. Nashat, S.; Abdullah, M.Z. Multi-class colour inspection of baked foods featuring support vector machine and Wilk's k analysis. *J. Food Eng.* **2010**, *101*, 370–380. [CrossRef]
112. Zemmour, E.; Kurtser, P.; Edan, Y. Automatic Parameter Tuning for Adaptive Thresholding in Fruit Detection. *Sensors* **2019**, *19*, 2130. [CrossRef] [PubMed]
113. Brodersen, K.H.; Ong, C.S.; Stephan, K.E.; Buhmann, J.M. The binormal assumption on precision-recall curves. In Proceedings of the International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 4263–4266.
114. Gönen, M. Receiver Operating Characteristic (ROC) Curves. In Proceedings of the SAS Users Group International 31(SUGI 31), San Francisco, CA, USA, 26–29 March 2006.
115. Mustapha, A.; Hussain, A.; Samad, S.A. A new approach for noise reduction in spine radiograph images using a non-linear contrast adjustment scheme based adaptive factor. *Sci. Res. Essays* **2011**, *6*, 4246–4258.
116. Brown, C.D.; Davis, H.T. Receiver operating characteristics curves and related decision measures: A tutorial. *Chemom. Intell. Lab. Syst.* **2006**, *80*, 24–38. [CrossRef]
117. Dimopoulos, T.; Bakas, N. Sensitivity Analysis of Machine Learning Models for the Mass Appraisal of Real Estate. Case Study of Residential Units in Nicosia, Cyprus. *Remote Sens.* **2019**, *11*, 3047. [CrossRef]

agriculture

MDPI

Article

An Improved Multi-Objective Optimization Decision Method Using NSGA-III for a Bivariate Precision Fertilizer Applicator

Yugong Dang [1], Hongen Ma [1], Jun Wang [2,*], Zhigang Zhou [1] and Zhidong Xu [3]

[1] School of Vehicle and Traffic Engineering, Henan University of Science and Technology, Luoyang 471000, China
[2] School of Electrical Engineering, Henan University of Science and Technology, Luoyang 471000, China
[3] China Petroleum First Construction Corporation Ltd., Luoyang 471023, China
* Correspondence: wj@haust.edu.cn

Abstract: In order to boost the performance of a bivariable granular fertilizer applicator and simplify the control methodology of fertilization rate regulation, this paper proposed a fertilization decision method to obtain the optimal combination of rotational speed and opening length by selecting the accuracy, uniformity, adjustment time, and breakage rate as the optimization objectives. We processed the outlier data collected using the indoor bench test, segmented the data with the fertilization growth rate as the index, and proved the rationality of the data segmentation by an independent sample *t*-test. SVM, BPNN, ELM, and RVM were used to train the two data sections to create the fertilization rate prediction model, and the models with the highest accuracy in the two data sections were selected for the assembly of the final prediction model used to describe the fertilization process of the bivariate fertilizer applicator. Moreover, the fertilization performance problem model was established with the objectives of accuracy, uniformity, adjustment time, and breakage rate and was solved using the NSGA-III algorithm to gain an optimal fertilization decision. Compared with GA and MOEA-D-DE methods, the results show that, using the new method, the average relative error declines from 8.64% and 6.05% to 3.09%, and the average coefficient of variation reduces from 6.67% and 6.81% to 6.41%, respectively. In addition, the adjustment time lowers from 2.01 s and 1.33 s to 0.78 s, and the average breakage rate drops from 1.084% and 0.845% to 0.803%, respectively. It is indicated that the presented method offers the most notable improvements in accuracy and adjustment time, while the advancements in regard to uniformity and breakage rate is slight, but both are within a reasonable range.

Keywords: granular fertilizer applicator; fertilization decision; multi-objective optimization; NSGA-III; breakage rate

Citation: Dang, Y.; Ma, H.; Wang, J.; Zhou, Z.; Xu, Z. An Improved Multi-Objective Optimization Decision Method Using NSGA-III for a Bivariate Precision Fertilizer Applicator. *Agriculture* **2022**, *12*, 1492. https://doi.org/10.3390/agriculture12091492

Academic Editors: Gniewko Niedbała and Sebastian Kujawa

Received: 8 August 2022
Accepted: 14 September 2022
Published: 17 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

From the perspective of environmental protection, precision agriculture (PA) can manage crop production inputs in an environmentally friendly way by determining the fertilizer, seed, and pesticide utilization rates through site-specific knowledge [1]. As an essential element of PA, variable rate fertilization can supply the most suitable fertilizer input for the fertilizer application areas in the field, resulting in an increased crop yield, reduced environmental pollution, decreased agricultural costs, and boosted economic income [2,3].

The variable granular fertilizer applicators can be divided into single variable fertilizer applicators and bivariate fertilizer applicators according to the number of control variables. The single variable fertilizer applicator primarily adjusts the degree of fertilizer application by regulating the rotational speed of the fertilizer discharging shaft or the opening length of the feeding inlet [4]. For instance, a subsoiling variable rate fertilization machine for conservation tillage developed by Qi et al. [5] can alter the rotational speed of the fertilizer

discharging shaft through a direct current (DC) motor to achieve the variation in the fertilization rate. Moreover, Su et al. [6] designed a variable fertilization system using a servo-driven actuator to modify the opening length of the feed inlet and obtained an average fertilization uniformity of 8.4%, and the time needed to switch the fertilization rate from 204 kg ha^{-1} to 314 kg ha^{-1} was identified as about 4.2 s. Since the single-variable fertilizer applicator requires the adaption of only one parameter, the control mode is definitely more straightforward than that of the bivariate fertilizer applicator, but the adjustment range of the fertilizer application rate is also limited. In addition, in the case of low fertilization requirements, the pulse phenomenon of fertilizer discharge unavoidably occurs due to the low speed of the fertilizer discharging shaft, thus affecting the uniformity of the fertilization [7,8]. In contrast, the bivariate fertilizer applicator modifies the fertilization rate by changing these two control parameters simultaneously. For example, a control system of a bivariable fertilizer applicator introduced by Liu et al. [9] can manipulate the change in the fertilization rate by accommodating the rotational speed of the fertilizer discharging shaft and the opening length of the feeding inlet, leading to the increased scope of the fertilization rate. However, the bivariate fertilizer applicator not only enhances the fertilization performance, but also transforms the fertilization rate regulation into a nonlinear and strongly coupled process, causing the control of the fertilizer applicator to become more complicated [10]. Meanwhile, an individual fertilization rate may correspond to multiple combinations of rotational speeds and opening lengths. Once the fertilization operation is carried out with a non-optimal strategy, the accuracy and uniformity of the fertilization will be inevitably affected [11]. Therefore, it is necessary to adopt the optimal fertilization decision to perform efficiently under the best-coordinated parameters [12].

To date, a significant number of studies have systematically investigated the control optimization techniques of bivariate fertilizer applicators. For example, Yuan et al. [13] obtained the optimal control parameters by the iterative multi-objective optimization method based on the GA (genetic algorithm), which combined fertilizing accuracy with energy-saving and fertilizing consistency objectives. Zhang et al. [10] established a three-objective problem model with accuracy, uniformity, and adjustment time as the objectives and solved the problem of the optimal fertilization control decision through the MOEA-D (multi-objective evolutionary algorithm based on decomposition) based on DE (differential evolution) algorithm. Nevertheless, data distortion may be introduced by manual errors or data entry errors during the bench test of a fertilizer applicator and will directly impact the quality of the machine learning technology and predicted results [14,15]. Therefore, it is critical to properly preprocess the original experimental data before using machine learning models. Although the control methods in the above studies have enhanced the fertilization performance compared with the traditional variable rate fertilization, they all neglect the importance of the pretreatment of the original data. Moreover, the accuracy of the fertilization rate prediction model obtained by the experimental data essentially determines the final fertilization performance. The changing trend of the data caused by the diverse range of rotational speeds and opening lengths will affect the forecast accuracy, and the adequate segmentation of data can be helpful for advancing the generalization ability of a model by improving the accuracy [16].

Furthermore, the high rotational speed of the fertilizer discharging shaft inevitably reduces the precision of the fertilization and increases the breakage of fertilizer particles [17,18]. Accordingly, the breakage rate should also be used as one of the evaluation parameters of the fertilization performance. However, these studies on the decision-making methods did not include the breakage rate in the optimization objectives. With the expansion of the objective dimension of the optimization problem, the proportion of non-dominant solutions in the population grows exponentially, slowing down the calculation process and making it challenging to distinguish the good and bad individuals through the Pareto dominance relationship [19]. The non-dominated sorting genetic algorithm III (NSGA-III), recommended by Jain et al. [20], has an excellent performance in unravelling optimization problems with

more than three objectives and has been widely applied to practical problems, providing a potential solution to this problem [21].

In view of the above-mentioned problems, a multi-objective optimization decision method based on NSGA-III for a bivariate granular fertilizer applicator is proposed in this study. Firstly, after the outliers are excluded from the data acquired in the bench experiment using Grubbs' criterion, the residual data are divided into two sections according to the variation trend of the fertilization growth rate and the differences between the two sections are verified by the independent sample *t*-test. Secondly, four machine learning algorithms, including a support vector machine (SVM) [22], back propagation neural network (BPNN) [23], extreme learning machine (ELM) [24], and relevance vector machine (RVM) [25], are used to train the prediction model of the fertilization rate by adopting the segmented data, respectively, and the multi-objective problem model is constructed with the accuracy, uniformity, adjustment time, and breakage rate as the optimization objectives. Finally, the optimal combination of the rotational speed of the fertilizer discharging shaft and the opening length of the feeding inlet is achieved by resolving the problem model through NSGA-III, and the diversity and convergence of the solution are verified using the hypervolume indicator. Moreover, the practicality and feasibility of the proposed method are proved by bench experiments.

The main contributions and innovations of this study are as follows:

1. The data utilized to build the prediction model of the fertilization rate is more consistent with the actual operation, eliminating outliers and using data segment modeling to effectively enhance the prediction accuracy of the model.
2. The breakage rate is suggested as one of the optimization objectives to reduce the impairment level in the case of high fertilizer demands.
3. NSGA-III is used to calculate the multi-objective optimization problem in this paper, so that the feasible solution has a more remarkable diversity and convergence.

2. Materials and Methods

2.1. Design of the Variable Granular Fertilizer Applicator

The designed fertilizer distributor is shown in Figure 1, which is mainly composed of a feeding box, a fertilizer particle baffle, a baffle connector, a screw conveyor, a fertilizer discharging shaft, a rolling bearing, and a bearing cap. The overall structure of the developed fertilization platform is illustrated in Figure 2. There, the baffle connector is linked with the guide screw, and the guide screw is connected to the opening-length-regulating motor (Motor 1) through a coupler. The fertilizer discharging shaft is fastened to the rotational-speed-regulating motor (Motor 2) by another coupler. In the fertilization phase, the PLC (programmable logic controller) controls the motor operation to adjust the rotational speed of the fertilizer discharging shaft N and the opening length of the fertilizer particle baffle L. Once the fertilizer particle baffle is opened, the fertilizer particles drop into the fertilizer distributor from the fertilizer hopper, and Motor 2 forces the screw conveyor to rotate spirally so that the fertilizer particles are transported to the fertilizer outlet. Finally, the fertilizer particles fall into the weighing container through the fertilizer feeding pipe to complete the fertilization process.

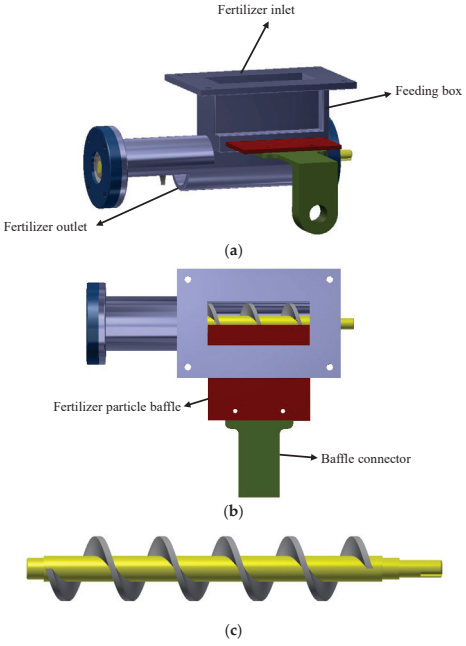


Figure 1. Structure of the granular fertilizer distributor: (**a**) front view of the fertilizer distributor; (**b**) top view of the fertilizer distributor; (**c**) screw conveyor.

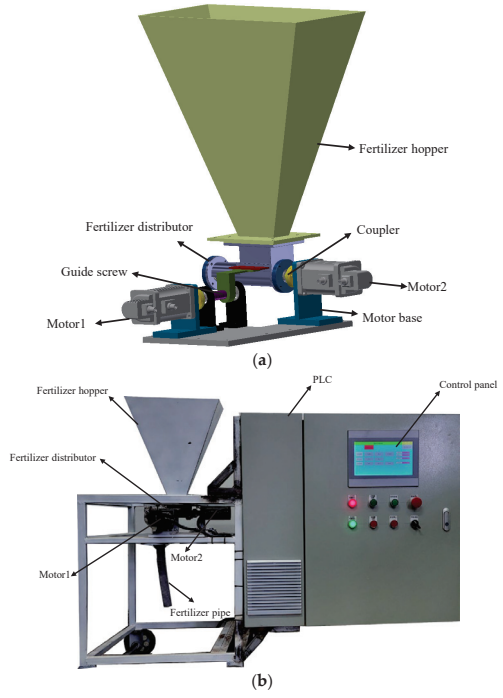


Figure 2. Overall structure of the fertilization platform: (**a**) structure of the test platform; (**b**) real image of the test platform.

2.2. Acquisition of the Fertilization Rate Data

The physical characteristics of a granular fertilizer play an influential impact on the fertilization performance [13], and the fertilizer used in the data acquisition stage should match the fertilizer used in the actual fertilization process in order to prevent the obtained fertilization control decision from being unsuitable for the fertilization of various crops. In this study, a compound fertilizer for wheat, corn, soybean, fruit trees, and other crops was used to collect experimental data. The total nutrient mass fraction of the fertilizer used is ≥45% (nitrogen-phosphorus-potassium: 15-15-15), and the particle density is 1716.3 kg m^{-3}. According to our method, an indoor bench test is conducted to obtain fertilization data under different rotational speeds and opening lengths. The ranges of the rotational speed and opening length are listed in Table 1. The change range in the rotational speed is 10–150 r min^{-1}, and the step size is set to 10 r min^{-1}. In addition, due to the considerable growth variation in the fertilization rate at a small opening length, to accurately acquire the characteristics of the fertilization data, the size step of the opening length is set to 0.5 mm, 1 mm, and 2 mm for the ranges of 6–12 mm, 12–16 mm, and 16–28 mm, respectively. Each combination of the rotational speed and opening length is sampled 20 times, lasting for 30 s each time. The discharged fertilizer for each test is weighed, and the corresponding fertilization rate is recorded. Finally, 345 groups of fertilization rate data with 20 samples in each group are obtained, of which 20 samples are the fertilization rate data measured under a fixed rotation speed and opening length.

Table 1. Parameters of the bench experiment.

Parameter	Group	Range	Step Size
Opening length	23	6–12 mm	0.5
		12–16 mm	1
		16–28 mm	2
Rotational speed	15	10–150 r min^{-1}	10

2.3. Data Preprocessing

In the method used in our study, the prediction model of the fertilization rate is established using the data from the bench test. Since the model is directly applied to the subsequent objective-solving process, the data reliability significantly affects the final fertilization performance. Therefore, to elevate the prediction accuracy, a series of processing steps, such as an outlier test, data segmentation, and normalization, are carried out using the original data.

2.3.1. Outlier Test

In the method used in this paper, Grubbs' criterion is used to test the outliers in order to prevent the distortion of the fertilization rate data caused by manual measurement errors or data entry errors during the data acquisition and decrease the influence of outliers on the overall data. Since Grubbs' criterion requires the data to be inspected in order to obey the normal distribution, the SW (Shapiro–Wilks) test method is used to prove the normality of the data at the significance level of 0.05 before the outlier test. If the result is $p > 0.05$, the data set is assumed to follow the normal distribution. Conversely, for data subject to the normal distribution, if a particular value x_p in a set of data satisfies the condition of Equation (1), then x_p is considered as an anomalous data item and should be excluded from the set of data, and the removed data will be replaced by the mean of the set of data:

$$\left| x_p - \overline{x} \right| > G_{(\alpha,n)S} \tag{1}$$

where $\overline{x}$ expresses the mean of the data set that x_p belongs to, $G_{(\alpha,n)S}$ indicates the critical value of Grubbs' test, α refers to the significance level, n signifies the number of samples in the set of data, and S is the standard deviation of this data set. Particularly, in the case

of $\alpha = 0.05$ and $n = 20$, the corresponding $G_{(\alpha,n)S}$ is equal to 2.557. Furthermore, after the outlier test, 345 samples of the fertilization rate with the averages of 20 measured data under every combination of the rotation speed and opening length are obtained.

2.3.2. Data Segmentation

It is observed that the fertilization rate is more sensitive to the small opening length operation, and the variation in the fertilization rate gradually becomes stable with the increase in the opening length. Thus, to relieve the impact of the growth rate difference on the accuracy of the prediction model, the average growth rate of the fertilization rate corresponding to each opening length and its previous opening length is calculated, and the fertilization rate data are divided into two segments, labelled as A and B, according to the changing trend of the average growth rate. The average growth rate of the fertilization rate can be obtained by:

$$R_{meani} = \frac{\sum_{j=1}^{n} \frac{r_j - r_{pj}}{r_{pj}}}{n} \quad (2)$$

where R_{meani} is the average fertilization growth rate of the ith opening length compared with the i-1th opening length, r_j is the fertilization rate corresponding to the ith opening length and the jth rotational speed, r_{pj} is the fertilization rate corresponding to the i-1th opening length and the jth rotational speed, and n is the total number of rotational speeds.

The average growth rate R_{mean} is grouped according to the set range of the opening length. The number of samples in each group is increased by one compared with the previous group, and the first group has at least two samples. During the calculation of the variance in the samples in each group, we chose the median value of the maximum variance and minimum variance as the segmented reference and selected the data group with the minimum difference between the variance and reference value as segment A and the remaining data as segment B.

After the data segmentation, the independent sample t-test is performed on the segmented data to determine whether there is a significant difference at the significance level of 0.05. If the p-value achieved from the test result is less than 0.05, the data are supposed to be split reasonably.

2.4. Establishment of the Fertilization Rate Prediction Model

The machine learning (ML) technique addresses the question of how we can build computers that self-improve automatically through experience [26]. It is one of today's most rapidly growing technical fields, lying at the intersection of computer science and statistics and at the core of artificial intelligence and data science. Currently, the mainstream ML algorithms are primarily used to complete the classification and regression of data and to implement predictions for data outside the sample set through model training. In the method used in this paper, the data obtained from the bench experiment of the fertilizer applicator are used as the sample set, the rotational speed and opening length are regarded as inputs, and the actual fertilization rate is adopted as the output. Then, the model trained by the ML algorithms is used to forecast the unknown fertilization rate. The four algorithms, including SVM, BPNN, ELM, and RVM, are selected to build the prediction model of the fertilization rate, and the model with the most suitable generalization ability for each data segment is accepted as the final model.

2.4.1. Dataset Partition and Performance Evaluation

The data sets in the two data segments (A, B) are separated before the model training (Figure 3), and each data segment is divided into 10 groups, one of which is used as a test set to assess the generalization faculty of the established model. Among them, there are 54 samples in the training set, 14 samples in the verification set, and 7 samples in the test set of the data of section A. There are 205 samples in the training set, 52 samples in the validation set, and 28 samples in the test set of the data of segment B. Since the samples, at the point of segmentation, are the commonly shared by the two sections of data, the

total amount of the two sections of data reaches 360 samples. For the remaining 9 sets, 20% of the data are utilized as the validation set in the training process, and 80% of the data are applied for the model training. Moreover, the mean absolute percentage error (MAPE) and the coefficient of determination R^2 are employed as the evaluation indicators for the prediction effect. MAPE is defined as follows:

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{3}$$

R^2 can be expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\overline{y} - y_i)^2} \tag{4}$$

where n is the total number of samples, y_i is the original value, $\hat{y}_i$ is the predicted value, and $\overline{y}$ is the mean of the original samples. According to the above dataset division, these four algorithms are trained and tested using the five combinations of the training and validation sets of each data segment, respectively, and the built model of each algorithm with the most minor errors for the validation set is chosen to participate in the subsequent error comparison using the test set.

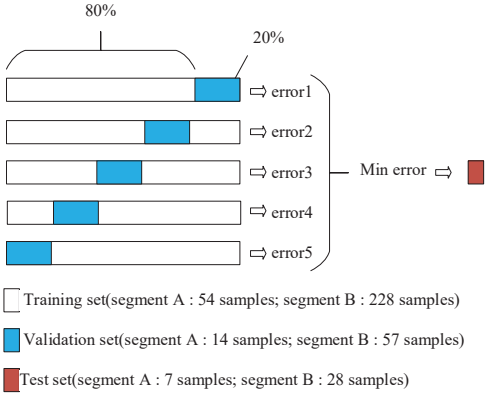


Figure 3. Dataset partitioning.

2.4.2. Comparison of the Fertilization Rate Prediction Models

By utilizing the identical test sets to compare the prediction errors of the four selected models for a data segment, we can choose the model with the minimum error as the final model of an individual data segment. In the model training phase, the samples corresponding to the segment point of the opening length are the co-shared samples of the two data segments, and the models constructed by various segments are obviously different. Therefore, the segment point is a repeat point for the fertilization rate prediction model, resulting in the two sets of different forecasted fertilization rates at this point. To solve this problem, we determine each data segment's opening length variation range by comparing the average relative errors between the predicted value and the actual value, corresponding to the prediction models of the two data segments at the specific point.

In addition, SVM, BPNN, ELM, and RVM are used to train the unsegmented data to evaluate whether the accuracy of the fertilization rate prediction model has been effectively improved before and after the data segmentation. The test sets of the models established by two algorithms are consistent with those used for the performance assessment of the fertilization rate prediction model, and the average prediction error of the test sets is used as the assessment index.

2.5. Modeling and Optimization of the Fertilization Decision

For a preset fertilization rate, there may be multiple combinations of rotational speeds and opening lengths used to satisfy the requirement of the fertilizer application rate [13]. However, the bivariate granular fertilizer applicator is affected by discharge pulsation at a low rotational speed of the fertilizer discharging shaft, resulting in an uneven fertilization. Meanwhile, at a high rotational speed of the fertilizer discharging shaft, the fertilization accuracy will decline, and the breakage rate of the fertilizer particles will conversely increase [11]. Furthermore, in the case where the target fertilization rate varies dramatically, if merely one variable is adjusted, the transition time of the fertilization rate will be unavoidably prolonged, thereby lowering the fertilization accuracy during the regulation process. Through these analyses, we aimed to use NSGA-III to obtain the optimal fertilization decision by solving the combinatorial optimization problem of the accuracy, uniformity, adjustment time, and breakage rate.

2.5.1. Objective Model

In the method used in this study, the target fertilization rate is derived from the fertilization prescription map, and the fertilization accuracy is determined by the absolute value of the difference between the target value and the predicted value. Therefore, the objective function of the fertilization accuracy is defined as follows:

$$\text{Min } f_1(L, N) = ||q - \hat{q}| - \varepsilon| \tag{5}$$

where q is the target fertilization rate, $\hat{q}$ is the predicted fertilization rate, and ε is the maximum permissible error of the fertilizer applicator.

As it is affected by the structure of the fertilizer distributor, the fertilization operation is not a continuous process. For the fertilizer distributor developed in this study, once the blade of the fertilizer discharge shaft is situated at the fertilizer outlet, the bivariate fertilizer applicator has a non-discharge status. Only when the gap between the two blades is aligned with the fertilizer outlet will the fertilizer particles be discharged, resulting in the pulse phenomenon in the fertilization process. Specifically, the lower the rotational speed of the fertilizer discharging shaft is, the more observable the phenomenon will be. Therefore, a greater rotational speed should be selected for a target fertilization rate to promote the uniformity of the fertilization. In this paper, the uniformity is depicted by the magnitude of the polar angle θ corresponding to the rotational speed and opening length.

For instance, both point a and point b correspond to the fertilization rate of 500 g 30 s^{-1}, as illustrated in Figure 4. We assume that the horizontal axis is the polar axis; thus, the included angles formed by the connection lines between the two points and the origin of the coordinates and the horizontal axis are defined as the polar angles. Apparently, under a fixed fertilization rate, the polar angle will increase with the growing rotational speed, while the polar angle will drop with the rise in the opening length. Hence, increasing the polar angle is helpful for obtaining a more suitable uniformity. Similarly, subsequent objective models of the adjustment time and breakage rate are built on the basis of this approach. The polar angle θ is expressed as follows:

$$\theta = \arctan\frac{N}{L} \tag{6}$$

and the objective model of uniformity can be given by:

$$\text{Min } f_2(L, N) = \left|\frac{1}{\theta}\right| \tag{7}$$

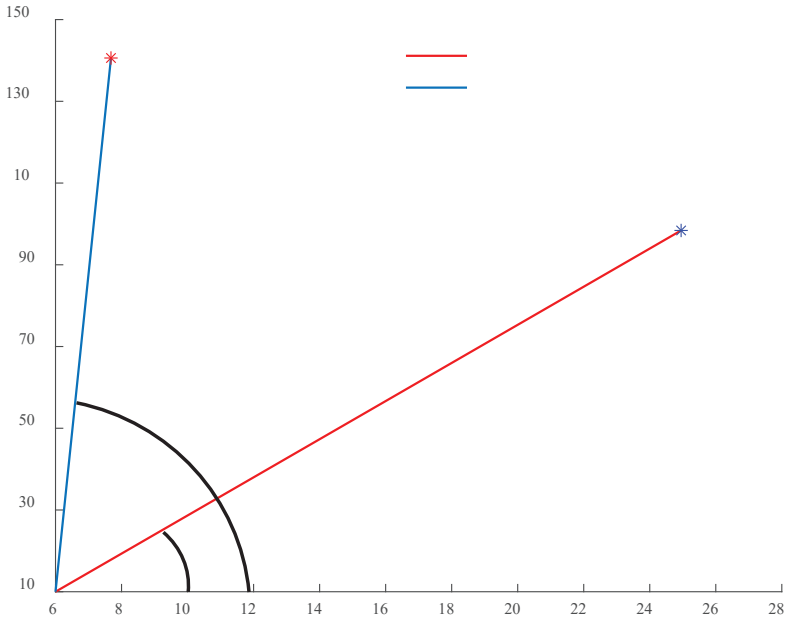


Figure 4. Schematic diagram of uniformity.

The adjustment time refers to the duration of time required for the fertilizer applicator to transition from one target fertilization rate to another [10]. The length of the adjustment time is directly related to the efficiency, timeliness, and precision of the variable fertilization. Accordingly, it is essential to minimize the variation range of the rotational speed and opening length corresponding to adjacent fertilization rates. The objective model of the adjustment time is described as follows:

$$\text{Min } f_3(L, N) = |\theta_1 - \theta_0| \tag{8}$$

where θ_1 and θ_0 are the polar angles corresponding to the present and previous fertilization rates, respectively.

To determine the object model of the breakage rate, the breakage rate experiments at the speeds of 150 r min^{-1}, 125 r min^{-1}, and 100 r min^{-1} are conducted to determine the trend of the breakage rate with the rotational speed in order to select a recommended speed that can satisfy both the breakage rate and uniformity. Moreover, the opening length has an insignificant effect on the breakage rate. To gain a shorter adjustment time, the opening length can be quadratically constrained simultaneously with the breakage rate optimization by minimizing the difference in the opening length between adjacent fertilization rates. The objective model of the breakage rate can be depicted by:

$$\text{Min } f_4(L, N) = \left|\theta_1 - \theta_{Obj}\right| \tag{9}$$

where θ_1 is the polar angle of the current fertilization rate and θ_{Obj} is the polar angle of the opening length corresponding to the previous fertilization rate at a recommended target rotational speed.

2.5.2. Multi-Objective Optimization Solution Based on NSGA-III

NSGA-III was proposed by Deb et al. [19] in 2014 on the basis of NSGA-II. The essential architecture of this algorithm is identical to NSGA-II, and the primary difference between NSGA-II and NSGA-III relates to the individual selection mechanisms. In particular, the

individual selection method of NSGA-II is based on the crowding distance, while that of NSGA-III is founded on reference points. This paper uses the method proposed by Das and Dennis [27] to generate reference points. Assuming that the hyperplane boundary corresponding to each objective is evenly divided into p parts, then the total number of reference points H corresponding to M objectives is given by:

$$H = \binom{M+p-1}{p} \tag{10}$$

For the t-th generation population, the archive set S_t is established after the completion of the reference point setting. Furthermore, the genetic algorithm is used to generate offspring population Q_t from parent population P_t, and population R_t is formed by merging P_t and Q_t. Through non-dominated sorting, the population R_t is split into different levels of non-dominated layers, and individuals with higher non-dominated levels are stored in S_t. If $|S_t| > K$, it is necessary to carry out an adaptive normalization, reference point correlation, and individual retention operations for the individuals of the last non-dominated layer in s until $|S_t| = K$, where K means the population size. Among these factors, adaptive normalization constitutes the ideal point $\overline{z} = \left(z_1^{min}, z_2^{min}, \cdots, z_M^{min}\right)$ of the population S_t by selecting the minimum value of each objective and translates the population S_t through Equation (11), causing the ideal point become a zero vector:

$$f_i'(x) = f_i(x) - z_i^{min} \tag{11}$$

where $f_i(x)$ is the i-th objective value of the population, z_i^{min} is the ideal value of the i-th objective, and $f_i'(x)$ is the i-th objective value after translation.

The extremum point corresponding to each objective can be obtained by Equation (12), which is expressed as:

$$\mathrm{AFS}(x, w) = max_{i=1}^{M} \frac{f_i'(x)}{w_i},\ x \in S_t. \tag{12}$$

where w is the normalized direction vector of the coordinate axes. For $w_i = 0$, we replace it with a smaller number 10^{-6}.

An M dimensional linear hyperplane is formed by connecting the extreme points corresponding to the number of fertilizer performance objectives M. Then, the intercept a_i of the i-th objective axis and the linear hyperplane can be computed, and the objective functions can be normalized as follows:

$$f_i^n(x) = \frac{f_i'(x)}{a_i - z_i^{min}} = \frac{f_i(x) - z_i^{min}}{a_i - z_i^{min}},\ i = 1, 2, \cdots, M. \tag{13}$$

where $\sum_{i=1}^{M} f_i^n(x) = 1$.

The detailed implementation process is shown in Algorithm 1. In this paper, the maximum number of iterations T of NSGA-III is set to 100, the objective number M of the problem to be solved is 4, and p is assigned as 6. Then, according to Equation (12), the total number of reference points is 84. Meanwhile, the population size K is designated as 84.

Algorithm 1 Optimization of the fertilization decision based on NSGA-III

Input: The variation range of the rotational speed and opening length, population size K, maximum iterations T, the number of optimization objectives M, and the objective models.
Output: The approximate Pareto set of rotational speeds and opening lengths: $\{v_1, v_2, \ldots, v_K\}$. The approximate Pareto front of the optimization objective: $\{F(v_1), F(v_2), \ldots, F(v_K)\}$

1: H-structured reference points Z
2: **for** $i = 1 \rightarrow K$ **do**
3: Initialize population to acquire P_1
4: **end for**
5: **for** $t = 1 \rightarrow T$ **do**
6: Set archive set $S_t = \emptyset$
7: Genetic operation to generate new population Q_t
8: Merge to generate new population $R_t = P_t \cup Q_t$
9: Non-dominated sorting of population R_t to obtain non-dominated layer F_1, F_2, ...
10: **repeat**
11: $S_t = S_t \cup F_t$, $t = t + 1$
12: **until** $|S_t| \geq K$
13: Last front to be included: $F_l = F_t$
14: **if** $|S_t| = K$ **then**
15: $P_{t+1} = S_t$
16: **or else**
17: $P_{t+1} = \cup_{j=1}^{l-1} F_j$
18: The number of individuals selected from F_l : $G = K - |P_{t+1}|$
19: **for** $j = 1 \rightarrow M$ **do**
20: Compute ideal point: $z_j^{min} = min_{s \in s_t} f_j(s)$
21: Translate objective points $f_j'(s) \Leftarrow f_j(s)$ using Equation (11), where $\forall s \in S_t$
22: Compute extreme points $z^{j,max}$ using Equation (12)
23: **end for**
24: Compute intercepts a_j of the coordinate axis corresponding to each objective
25: Normalize objectives using Equation (13)
26: **for** each reference point $z \in Z$ **do**
27: Compute reference line $w = z$
28: **end for**
29: **for** each $s \in S_t$ **do**
30: **for** each $w \in Z$ **do**
31: Compute $d_\perp(s, w) = s - w^T / w$
32: **end for**
33: $\pi(s) = w : argmin_{w \in Z} d_\perp(s, w)$
34: $d(s) = d_\perp(s, \pi(s))$
35: **end for**
36: Compute $\rho_{\bar{j}}$ (the number of individuals associated with the reference point $j \in Z$)
37: **for** $g = 1 \rightarrow G$ **do**
38: $J_{min} = \{j : argmin_{j \in Z} \rho_{\bar{j}}\}$
39: **if** number$(J_{min} > 1)$ **then**
40: $\bar{j} = random(J_{min})$
41: **end if**
42: $I_{\bar{j}} = \{s : \pi(s) = \bar{j}, \ s \in F_l\}$
43: **if** $I_{\bar{j}} \neq \emptyset$ **then**
44: **if** $\rho_{\bar{j}} = 0$ **then**
45: $P_{t+1} = P_{t+1} \cup (s : argmin_{s \in I_{\bar{j}}} d(s))$
46: **else**
47: $P_{t+1} = P_{t+1} \cup random\left(I_{\bar{j}}\right)$
48: **end if**
49: $\rho_{\bar{j}} = \rho_{\bar{j}} + 1$, $F_l = F_l \backslash s$
50: **else**
51: $Z = Z_{\bar{j}}$
52: **end if**
53: **end for**
54: **end if**
55: **end for**
56: **return** $\{v_1, v_2, \ldots, v_K\}, \{F(v_1), F(v_2), \ldots, F(v_K)\}$

2.5.3. Performance Comparison of the Multi-Objective Optimization Results

The hypervolume (HV) indicator is a metric approach commonly used to compare the results of an evolutionary multi-objective optimization algorithm (EMOA) [28]. This method was first proposed by Zitzler et al. [29] and is employed to calculate the volume of the space surrounded by the Pareto solution set and reference points [30]. The greater the value of the HV indicator is, the more satisfactory the convergence and diversity of the solution set will be [31]. In the method used in this paper, the NSGA-III and MOEA-D algorithms are used to calculate the solution sets of different target fertilization rates under the same conditions, and the mean values of the HV indicators of these two algorithms are computed. Thereafter, the convergence and diversity of the solution set acquired by NSGA-III are estimated by comparing the differences between the mean values.

2.6. Evaluation Criteria of the Fertilization Performance

The proposed method is compared with the GA, suggested by Yuan et al. [13], and MOEA-D-DE, recommended by Zhang et al. [10], to prove its feasibility and practicality. Using the accuracy, uniformity, adjustment time, and breakage rate as evaluation criteria, eight groups of fertilization decisions acquired by the three methods corresponding to the target fertilization rates are verified and compared using the fertilization platform under identical testing situations. The specific descriptions of the assessment criteria are as follows.

2.6.1. Accuracy

Using the fertilization platform to verify the optimal rotational speed and opening length solved by these algorithms, the accuracy is defined as the relative error between the measured value and the target value of the fertilization rate and is expressed explicitly as:

$$RE = \frac{|y - y^*|}{y^*} \times 100\% \tag{14}$$

where y is the fertilization rate obtained from the actual experiment and y^* is the target fertilization rate.

2.6.2. Uniformity

The uniformity test is conducted 15 times under the optimal rotational speed and opening length determined by these three algorithms, and the duration of the fertilization each time is 3 s. The coefficient of variation (CV) is used to measure the uniformity of the fertilization as follows:

$$CV = \frac{\sigma}{\overline{y^t}} = \frac{\sqrt{\frac{\sum_{i=1}^{15}\left(y_i^t - \overline{y^t}\right)^2}{15-1}}}{\overline{y^t}} \tag{15}$$

where y_i^t is the fertilization rate acquired by the i-th test and $\overline{y^t}$ is the average fertilization rate.

2.6.3. Adjustment Time

In the method used in this paper, the maximum time required to transition from the current fertilization rate to the following fertilization rate, so as to alter the rotational speed and opening length, is used to express the adjustment time. Obviously, the shorter the adjustment time is, the faster the response of the fertilizer applicator will be.

2.6.4. Breakage Rate

The breakage rate test is carried out at the optimal rotational speed corresponding to the target fertilization rate. During the experiment, a sieve with holes of 1.5 mm is used to ensure that the measured fertilizer is not damaged. After documenting the weight, the

treated fertilizer is poured into the fertilizer applicator, and the discharged fertilizer is sifted and weighed again. The breakage rate is calculated as follows:

$$B = \frac{b_1 - b_2}{b_1} \times 100\% \quad (16)$$

where B is the breakage rate, b_1 is the weight of the fertilizer before the fertilizer discharge, and b_2 is the weight of the fertilizer after the fertilizer application.

3. Results and Discussion

3.1. Data Pretreatment Effect

A total of 345 groups of data must be tested for outliers, and each group comprises 20 samples. A normal test is performed for each group of data. For the data subject to a normal distribution, Grubbs' criterion is adopted to screen the data measured during the indoor bench test under the significance level of 0.05, as shown in Table 2. Thirteen abnormal data are filtered out, and every excluded outlier datum is replaced by the corresponding intra-group mean.

Table 2. Elimination and replacement of outliers.

No.	Rotational Speed (r min^{-1})	Opening Length (mm)	Abnormal Data (g 30 s^{-1})	Replace Data (g 30 s^{-1})
1	30	6.5	44	19
2	40	6	33	11
3	40	7	121	39
4	50	7	82	129
5	70	15	325	336
6	90	9	394	401
7	120	6	7	21
8	120	8	494	511
9	130	9	523	534
10	130	11.5	569	558
11	140	8.5	549	567
12	150	8.5	566	582
13	150	16	645	655

The data are sectioned according to the variation trend of the fertilization growth rate. The maximum and minimum variances for all the possible groups of the average fertilization growth rate are 0.6965 and 0.2452, respectively, and the reference value is 0.4708. The variance is 0.5427 in the case where the opening range varies from 6 to 8 mm, which is the smallest compared with the reference value. Thus, the corresponding fertilization rate data are classified as segment A, and the remaining data are designated as segment B. The data in segment A ($L \leq 8$ mm) include 75 samples, and the data in segment B ($L \geq 8$ mm) consist of 285 samples in total. The rationality of the data segmentation is verified by the independent sample *t*-test, and the SW test is first used to test the normality of the two segments of fertilization growth rate data, as shown in Table 3. At the significance level of 0.05, the significance of the data of segment A and segment B is 0.873 and 0.325, respectively, and the significance of the data of both segments is greater than 0.05. Thus, the data of these two segments follow the normal distribution.

Table 3. Normality test results of the two data segments by the SW approach.

Groups	Statistics	df	Sig.
A	0.975	4	0.873
B	0.943	18	0.325

The results of the independent sample *t*-test are shown in Table 4. First, Levin's homogeneity test of variance is applied to inspect the homogeneity of the variances in the two data segments. The outcome ($p = 0$) indicates that the variances exhibit heterogeneity. In

that case, the *p*-value is 0.049, less than the given significance level ($p = 0.05$). Consequently, we can infer that there is a significant difference between the two data sections, and it is reasonable to segment them in terms of this range.

Table 4. The *t*-test results of the two data segments.

Levene's Test for Equality of Variances			*t*-Test for Equality of Means					95% Confidence Interval	
	F	Sig.	*t*	df	Sig. (2-Tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Equal variances assumed	95.199	0	7.49	20	0	1.19174	0.15911	0.85985	1.52364
Equal variances not assumed			3.207	3	0.049	1.19174	0.37156	0.00930	2.37419

In this study, the main purpose of the data preprocessing is to improve the degree of approximation between the data and the actual fertilization process and reduce the impact of the data variation on the accuracy of the fertilization rate prediction model. However, the two methods that we compared do not include the relevant technique. In addition, for the MOEA/D-DE method, only three repeated tests are conducted for each group of data during the data acquisition, and the average is taken as the true fertilization rate of a certain combination of the opening length and rotation speed. It should be noted that, once there is a serious deviation between the measured data and the actual fertilization rate caused by operational errors or other reasons, a subsequent erroneous result will inevitably be generated. Therefore, a feasible method is to enhance the accuracy of the data by increasing the number of tests and detect the outliers after the measurements.

3.2. Performance of the Fertilization Rate Prediction Model

The performance of the fertilization rate prediction model based on SVM, BPNN, ELM, and RVM is evaluated using the two test sets of segments A and B, respectively. The numbers of samples in the test sets of segment A and segment B are 7 and 28, respectively, as shown in Tables 5 and 6.

Table 5. Test samples of segment A.

No.	Rotational Speed (r min^{-1})	Opening Length (mm)	Fertilization Rate Q (g $30\ s^{-1}$)
1	10	6	11.65
2	40	7	42.00
3	50	6.5	21.95
4	60	7	66.85
5	100	8	436.20
6	110	6	18.15
7	140	8	550.30

Table 6. Test samples of segment B.

No.	Rotational Speed (r min^{-1})	Opening Length (mm)	Fertilization Rate (g $30\ s^{-1}$)	No.	Rotational Speed (r min^{-1})	Opening Length (mm)	Fertilization Rate (g $30\ s^{-1}$)
1	10	10	47.30	15	100	22	501.85
2	10	12	47.10	16	110	10	482.05
3	20	10.5	100.45	17	110	12	491.75
4	20	11.5	100.65	18	110	18	507.35
5	20	26	104.25	19	110	20	510.70
6	40	8.5	194.45	20	110	28	515.90
7	50	9.5	245.00	21	120	14	556.40
8	60	8.5	274.95	22	120	28	588.25
9	60	9	280.25	23	140	8	550.30
10	60	15	294.35	24	140	9	575.00
11	70	8	315.65	25	140	10.5	600.05
12	80	10	367.55	26	140	11	605.15
13	80	16	384.10	27	140	14	618.50
14	80	20	392.90	28	150	26	685.85

In addition, the prediction results of segment A are shown in Figure 5. The effect of SVM is the most acceptable among these four algorithms, corresponding to an R^2 of 0.9995 and MAPE of 5.7209%. The data forecast results of segment B are illustrated in Figure 6, and BPNN demonstrates a remarkable performance, with an R^2 of 0.9997 and MAPE of 0.8658%.

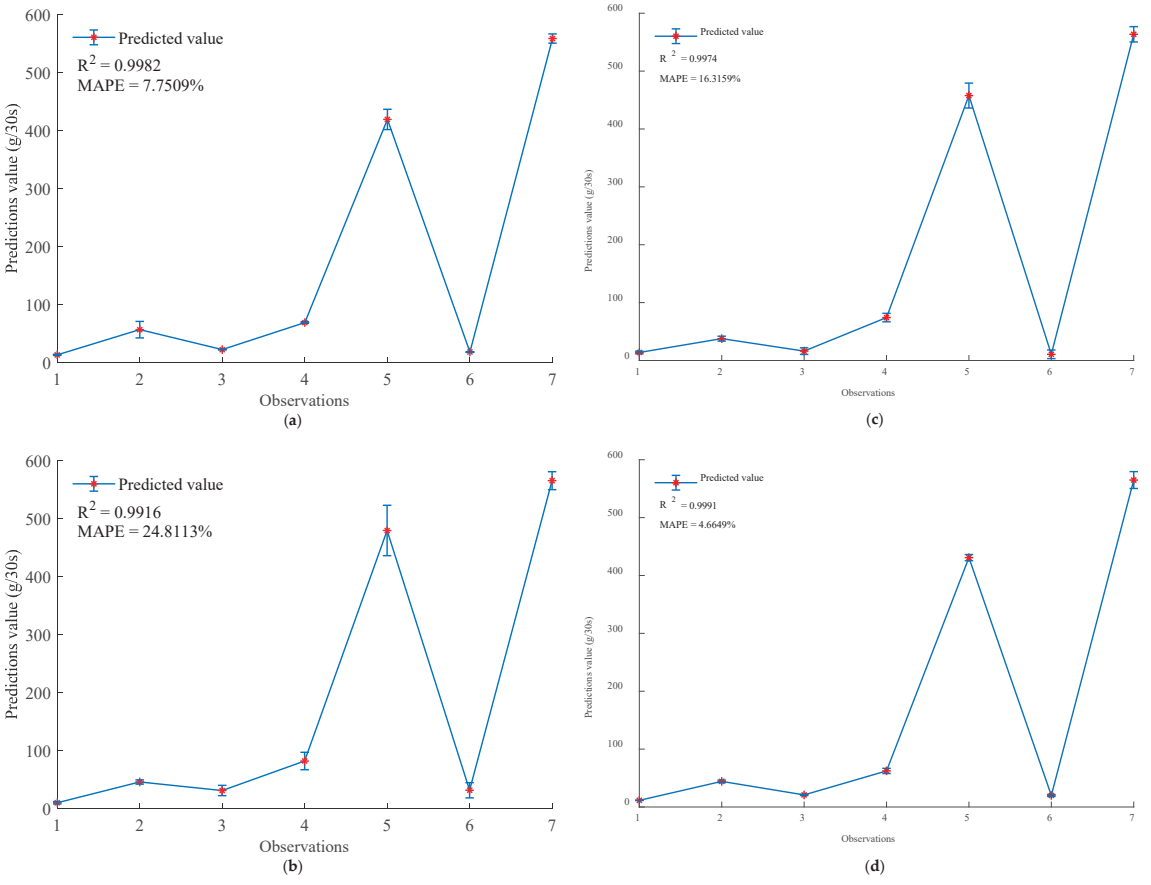


Figure 5. Differences between the experimental data and predicted data for four machine learning algorithms with an opening length range of 6–8 mm: (**a**) BPNN; (**b**) ELM; (**c**) RVM; (**d**) SVM.

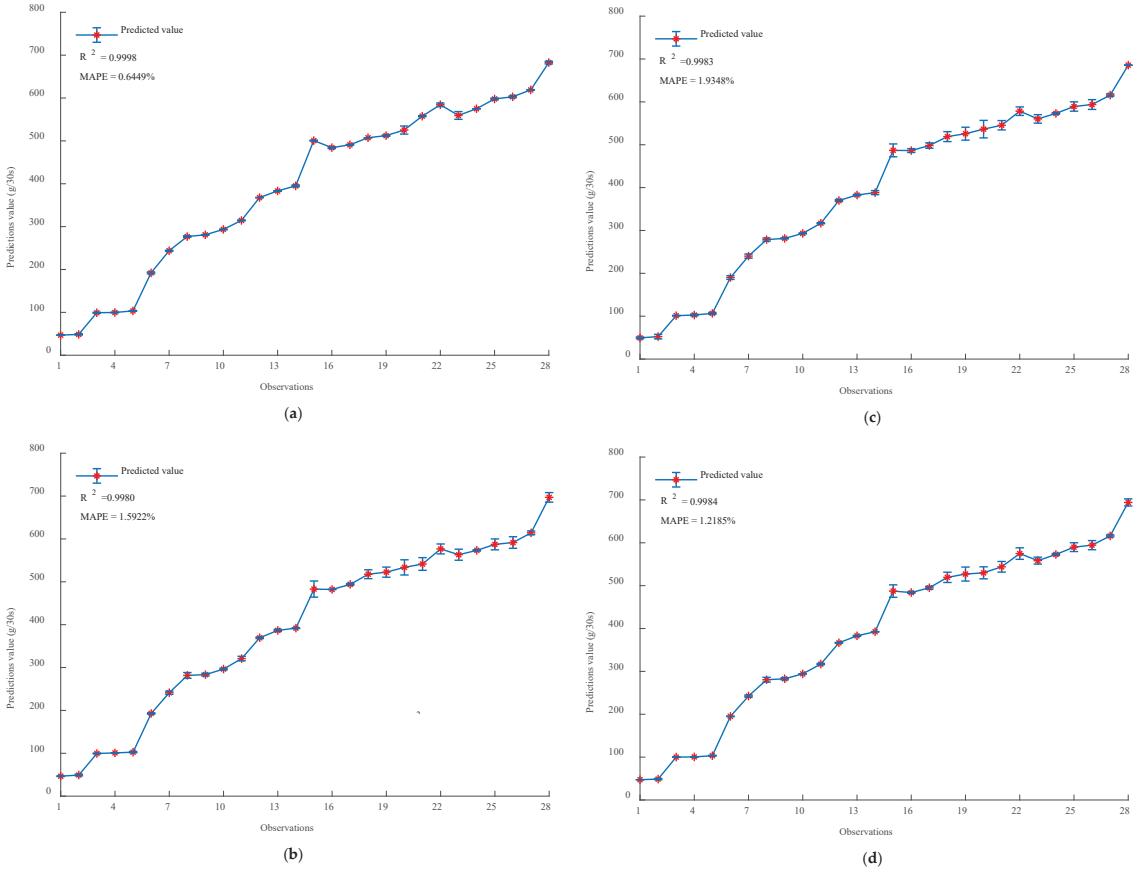


Figure 6. Deviations between the experimental data and predicted data for four machine learning algorithms with an opening length range of 8–28 mm: (**a**) BPNN; (**b**) ELM; (**c**) RVM; (**d**) SVM.

The optimal models (SVM and BPNN) for the above two data segments are selected to predict the fertilization rate at the opening length of 8 mm, respectively, and the relative errors between the predicted value and the actual fertilization rate of the two models are compared, as shown in Table 7. The average relative errors of the SVM-based model for segment A and the BPNN-based model for segment B are 0.019824% and 0.007116%, respectively. It can be seen that the deviation between the predicted value and the actual value of the data in segment B is less than that of segment A. Therefore, the final fertilization rate prediction model is established as follows:

$$\mathrm{Q}(L,N) = \begin{cases} SVM - based\ Model & 6\mathrm{mm} \le L < 8\mathrm{mm} \\ BPNN - based\ Model & 8\mathrm{mm} \le L \le 28\mathrm{mm} \end{cases} \quad (17)$$

where Q signifies the predicted fertilization rate.

Table 7. Comparison of the prediction accuracies at the opening length of 8 mm.

Rotational Speed (r min^{-1})	Experimental Value (g 30 s^{-1})	Predict Value (g 30 s^{-1})		Relative Error (%)	
		Segment A	Segment B	Segment A	Segment B
10	45.15	46.79	44.65	0.036323	0.011074
20	95.45	95.47	96.16	0.000210	0.007438
30	146.10	142.75	144.71	0.022930	0.009514
40	193.80	187.92	188.31	0.030341	0.028328
50	230.55	230.85	232.56	0.001301	0.008718
60	271.80	271.91	271.18	0.000405	0.002281
70	315.65	311.76	314.46	0.012324	0.00377
80	350.75	351.17	351.14	0.001197	0.001112
90	390.75	390.77	390.69	0.000051	0.000154
100	436.20	430.82	441.70	0.012334	0.012609
110	468.60	471.07	467.81	0.005271	0.001686
120	509.80	510.65	509.80	0.001667	0
130	520.40	548.19	519.84	0.053401	0.001076
140	550.30	581.87	559.38	0.057369	0.016500
150	573.95	609.67	575.37	0.062235	0.002474
Average relative error				0.019824	0.007116

SVM, BPNN, ELM, and RVM are used to train the unsegmented data in order to set up the fertilization rate prediction models, and the performance of the trained models is evaluated using the test sets used for the two data segments, as shown in Table 8. For the prediction model training using the unsegmented and segmented data, the averages of R^2 and MAPE correspond to 6–8 mm opening length increases from 0.9734 and 24.4065% to 0.9966 and 13.3858%, respectively, and the averages of R^2 and MAPE correspond to 8–28 mm opening length increases from 0.9824 and 17.5236% to 0.9986 and 1.3476%, respectively. It can clearly be concluded that the forecast accuracy can be effectively improved by adequately processing the data before training the fertilization rate prediction model.

Table 8. Prediction performance before and after the data segmentation.

Algorithm	Before Segmentation				After Segmentation			
	6–8 mm		8–28 mm		6–8 mm		8–28 mm	
	R^2	MAPE	R^2	MAPE	R^2	MAPE	R^2	MAPE
SVM	0.9789	23.4581%	0.9833	16.7325%	0.9991	4.6649%	0.9984	1.2185%
BPNN	0.9721	24.1568%	0.9892	15.4239%	0.9982	7.7509%	0.9998	0.6449%
ELM	0.9735	24.3684%	0.9749	19.3415%	0.9916	24.8113%	0.9980	1.5922%
RVM	0.9689	25.6428%	0.9821	18.5963%	0.9974	16.3159%	0.9983	1.9348%
Mean value	0.9734	24.4065%	0.9824	17.5236%	0.9966	13.3858%	0.9986	1.3476%

By using different machine learning algorithms to train the fertilizer rate prediction models corresponding to the two data segments and combining the models of the two data segments, the prediction accuracy is obviously improved. Moreover, BPNN and SVM, which were adopted in this study, can be further improved by hyper-parameter optimization.

3.3. Performance of the Multi-Objective Optimization Based on NSGA-III

The fertilizer breakage rate is measured at the speeds of 150 r min^{-1}, 125 r min^{-1}, and 100 r min^{-1}. The breakage rate of each rotational speed is measured using three different amounts of fertilizer, and the results of the breakage rate for the three rotating speeds are shown in Table 9. The results indicated that the maximum average breakage rate is 3.5% at 150 r min^{-1}, 1.87% at 125 r min^{-1}, and 0.69% at 100 r min^{-1}. Hence, the rotational speed

should be maintained at around 100 r min^{-1} to lower the breakage rate and ensure the uniformity of the fertilization.

Table 9. Results of the breakage rate at various rotational speeds.

Rotational Speed (r min^{-1})	Before Measurement (g)	After Measurement (g)	Error Amount (g)	Breakage Rate (%)	Average Breakage Rate (%)
100	626	621	5	0.80	0.69
	362	360	2	0.55	
	410	407	3	0.73	
125	660	648	12	1.82	1.87
	469	460	9	1.92	
	638	626	12	1.88	
150	529	510	19	3.59	3.5
	510	489	21	4.12	
	679	660	19	2.80	

In the method used in this study, NSGA-III is used to solve the multi-objective optimization problem to obtain the optimized combination of the rotational speed and opening length of the bivariate fertilizer applicator. Furthermore, the optimization results are compared with those of the algorithms proposed by Yuan et al. and Zhang et al. The optimal combination of the three methods corresponding to the target fertilization rate is shown in Table 10.

Table 10. Fertilization decisions acquired by the three methods for different target fertilization rates.

No.	Target Fertilizer Rate (kg ha^{-1})	Improved Method		GA		MOEA-D-DE	
		Rotational Speed (r min^{-1})	Opening Length (mm)	Rotational Speed (r min^{-1})	Opening Length (mm)	Rotational Speed (r min^{-1})	Opening Length (mm)
1	138.25	33.9	7.65	23.07	9.33	25	8.79
2	239.78	40.92	9.22	40.64	13.56	45	15.09
3	342.92	59.75	10.33	59.57	12.9	63.73	10.25
4	437.95	78.82	9.91	92.57	17.32	79.18	9.75
5	522.77	92.61	14.38	108.29	18.59	95.01	9.21
6	600.24	104.54	17.79	114.53	9.47	106.36	15.5
7	700.75	123.18	18.78	133.34	11.79	125.01	16.96
8	757.91	138.5	17.55	148.42	18.29	135.01	16.90

As demonstrated in Table 11, the HV indicators of NSGA-III and MOEA/D under the four objectives (accuracy, uniformity, adjustment time, and breakage rate) are estimated. The average of the various HV indicators of NSGA-III is greater than that of MOEA-D, implying that the application of NSGA-III shows better convergence and diversity so as to determine the fertilization decision model. In addition, it is worth noting that both algorithms show a significant downward trend in regard to HV5. This is mainly because, in the case of a greater target fertilization rate, it is necessary to increase the variation range of the rotational speed and opening length to ensure the accuracy of the fertilization. However, once the rotational speed exceeds 100 r min^{-1}, the breakage rate will unavoidably increase. The rotational speed needs to be decreased to lower the breakage rate, leading to a decline in accuracy and uniformity. Therefore, the solution convergence used for fertilization decisions at a high target fertilization rate is inferior to that at a low target fertilization rate.

Table 11. Comparison of the HV indicators of the evolutionary multi-objective optimization solution set.

Algorithms	HV1	HV2	HV3	HV4	HV5	Mean HV
NSGA-III	0.2024	0.2381	0.1905	0.1986	0.1667	0.1993
MOEA-D	0.1824	0.2156	0.1852	0.1914	0.1629	0.1937

The difference in accuracy between the three methods is illustrated in Figure 7. The average relative error of the fertilization rate obtained by the method proposed in this paper is 3.09%, 8.64% for GA and 6.05% for MOEA-D-DE. It can be noticed that the method proposed in this paper clearly and dramatically improved the fertilization accuracy. Because these three methods were carried out under the same testing conditions, except for the rotational speed and opening length, the difference in accuracy is mainly generated by the diverse forecast results of the fertilization rate prediction models. Therefore, it can be concluded that the method proposed in this paper can effectively enhance the predictive accuracy by improving the generalization ability of the prediction model. In addition, segmented training is valuable for increasing the model's accuracy.

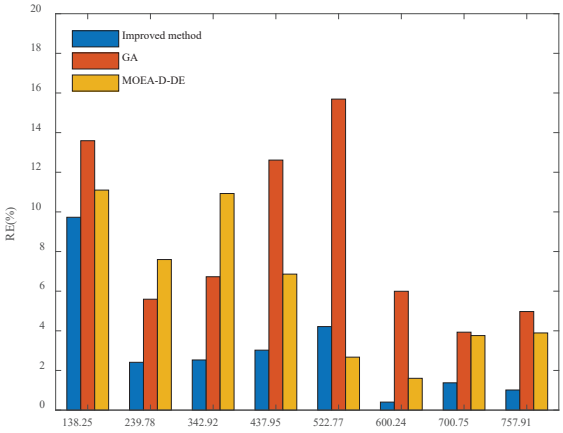


Figure 7. Comparison of the fertilization accuracies.

As shown in Figure 8, the average CV of the fertilization measured by the proposed method in the bench test is 6.41%, slightly higher than that of GA (6.67%) and MOEA-D-DE (6.81%). It is worth noting that the CV of the proposed method is relatively low for lower fertilization rate demands, while the uniformity of the proposed method is insufficient compared with that of GA for higher fertilization rate demands. The possible reason for this is that this study considered that the increase in the rotational speed would unavoidably lead to an upsurge in the breakage rate. Hence, to decrease the breakage rate, the high rotational speed is restricted, with the aim of optimizing the fertilization decision while maintaining a more appropriate uniformity of the fertilization.

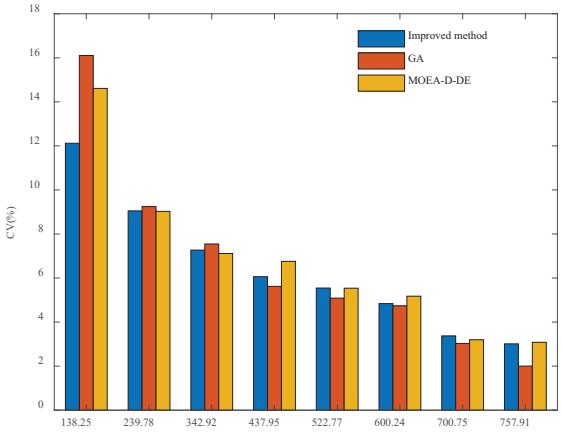


Figure 8. Comparison of the fertilization uniformities.

As displayed in Figure 9, the average adjustment time measured by the proposed method is 0.78 s, while that of GA is 2.01 s and that of MOEA-D-DE is 1.33 s. Since the experiment used for the adjustment time in this study is a static test conducted through an indoor bench test, the influencing factor of the adjusting time is primarily mechanical lag, which is affected by the regulation of the rotational speed and opening length during the change in the fertilization rate. Since the method proposed in this paper, combined with MOEA-D-DE, optimizes the adjustment range of the rotational speed and opening length, it can effectively speed up the adjustment process and lower the mechanical lag compared with GA. Moreover, in optimizing the breakage rate, the proposed method enforced a quadratic constraint on the variation range of the opening length to further shorten the adjustment time of the equipment.

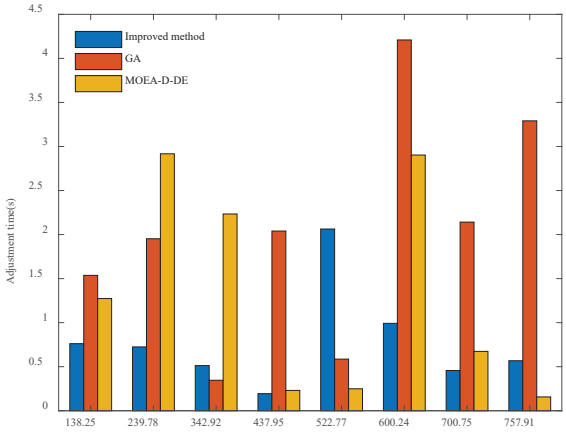


Figure 9. Comparison of the adjustment times.

As exhibited in Figure 10, in the case of lower fertilization rate demands, there is little difference in the breakage rate between the three methods due to the low speed of the fertilizer discharging shaft. However, with the increase in the target fertilization rate, the corresponding speed and breakage rate also increase markedly. The maximum breakage rates of the method proposed, GA, and MOEA-D-DE are 2.51%, 3.45%, and 2.7%, respectively. Furthermore, the averages of the presented method, GA, and MOEA-D-DE are 0.803%, 1.084%, and 0.845%, respectively. From the perspective of the average breakage rate,

there is a slight dissimilarity between the three methods, and the difference is more apparent at a higher fertilization rate. GA has the highest breakage rate among the three methods because it does not consider the optimization of the breakage rate. Contrarily, MOEA-D-DE defines the objective function of uniformity as the minimum Euclidean distance between the rotational speed and opening length to be optimized and the corresponding rotational speed and opening length of the center of the adjusted region, which limits the rotational speed to a certain extent and causes a decline in the breakage rate.

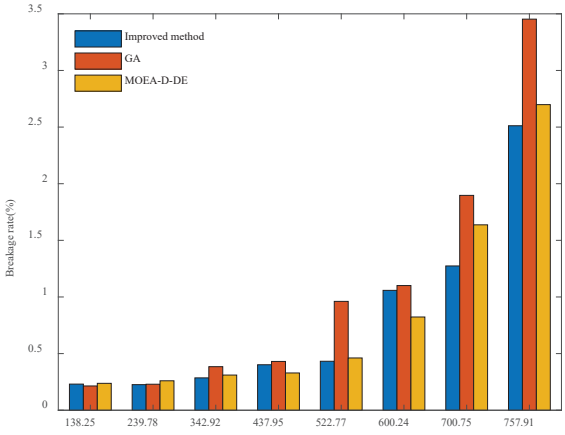


Figure 10. Comparison of the breakage rates.

Through the analysis of the above results, it can be concluded that the accuracy of the fertilization mainly depends on the accuracy of the prediction model, and the uniformity of the fertilization can be enhanced by increasing the rotational speed. Meanwhile, the adjustment time is indirectly affected by the limitation of the breakage rate. In addition, the selection of the recommended target speed in the objective model of the breakage rate is not confined to 100 r min^{-1} and can be readjusted according to the actual situation. Once the breakage rate is relatively lower at a high rotation speed due to the physical characteristics of the fertilizer particles, the recommended target rotational speed can be appropriately increased to improve the uniformity of the fertilization.

4. Conclusions

This paper presents a novel decision method designed for a developed bivariate granular fertilizer applicator based on a screw conveyor. Among the four machine learning algorithms for segment B, the prediction accuracy of BPNN is comparatively greater. For segment A, the prediction accuracy of SVM is relatively higher, benefiting from the advantage of SVM in the context of a small sample size. By combining the best models for the two data segments, the accuracy of the prediction model is significantly improved, indicating that the segmentation method proposed in this paper is valuable for the subsequent fertilizer decision optimization. Moreover, the high-precision prediction model of the fertilization rate can reduce the influence of prediction errors on the fertilization performance and accurately reproduce the actual fertilizer discharge process as a substitute for the online measurement of the fertilization rate. Through the optimization of the fertilization performance by NSGA-III, the optimal solution set matching the target fertilization rate is obtained, and the HV indicators are used as the evaluation indicators of the solution set. The results of the HV indicators show that the solution set generated by NSGA-III has better convergence and diversity. Compared with the experimental results of GA and MOEA-D-DE, the method proposed in this paper is most effective in improving the accuracy of the fertilization due to the outstanding performance of the prediction model. Meanwhile, this outcome further verifies the importance of data pretreatment. In this

study, the breakage rate was used as one of the optimization objectives to restrict the high rotational speed of the fertilizer discharging shaft. In the face of a higher target fertilizer rate, the target fertilizer rate is more effectively achieved by boosting the opening length to decrease the breakage rate. With the transformation of the target fertilization rate, the decreased change range of the opening length accelerates the adjustment process and is helpful for reducing the influence of mechanical lag on the accuracy of the fertilization.

In future studies, the index table of the optimal rotational speed and opening length corresponding to the required fertilization rate in each area of the prescription map can be further established in order to guide the operation of the fertilizer applicator for the purpose of greatly reducing the control complexity of the process of fertilization.

Author Contributions: Conceptualization, Y.D.; methodology, H.M. and J.W.; software, Z.Z.; validation, Z.Z. and Z.X.; formal analysis, H.M.; investigation, Y.D. and J.W.; resources, J.W.; data curation, Z.X. and H.M.; writing—original draft preparation, Y.D., H.M., and J.W.; writing—review and editing, H.M. and J.W.; visualization, H.M.; supervision, J.W.; project administration, J.W.; funding acquisition, Y.D. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Program for Science & Technology Innovation Talents of the Universities of Henan Province (grant no. 20HASTIT029), Key Scientific Research Projects of the Universities of Henan Province (grant no. 19A460021), and the Key Science and Technology Project of Henan Province (grant no. 2221022102164, 212102210352).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bongiovanni, R.; Lowenberg-DeBoer, J. Precision agriculture and sustainability. *Precis. Agric.* **2004**, *5*, 359–387. [CrossRef]
2. Tola, E.; Kataoka, T.; Burce, M.; Okamoto, H.; Hata, S. Granular fertiliser application rate control system with integrated output volume measurement. *Biosyst. Eng.* **2008**, *101*, 411–416. [CrossRef]
3. Forouzanmehr, E.; Loghavi, M. Design, development and field evaluation of a map-based variable rate granular fertilizer application control system. *Agric. Eng. Int. CIGR J.* **2012**, *14*, 255–261. [CrossRef]
4. Jafari, M.; Hemmat, A.; Sadeghi, M. Development and performance assessment of a DC electric variable-rate controller for use on grain drills. *Comput. Electron. Agric.* **2010**, *73*, 56–65. [CrossRef]
5. Qi, J.; Tian, X.; Li, Y.; Fan, X.; Yuan, H.; Zhao, J.; Jia, H. Design and experiment of a subsoiling variable rate fertilization machine. *Int. J. Agric. Biol. Eng.* **2020**, *13*, 118–124. [CrossRef]
6. Su, N.; Xu, T.S.; Song, L.T.; Wang, R.; Wei, Y.Y. Variable rate fertilization system with adjustable active feed-roll length. *Int. J. Agric. Biol. Eng.* **2015**, *8*, 19–26. [CrossRef]
7. Shi, Y.Y.; Hu, Z.C.; Wang, X.C.; Odhiambo, M.O.; Sun, G.X. Fertilization strategy and application model using a centrifugal variable-rate fertilizer spreader. *Int. J. Agric. Biol. Eng.* **2018**, *11*, 41–48. [CrossRef]
8. Sugirbay, A.M.; Zhao, J.; Nukeshev, S.O.; Chen, J. Determination of pin-roller parameters and evaluation of the uniformity of granular fertilizer application metering devices in precision farming. *Comput. Electron. Agric.* **2020**, *179*, 105835. [CrossRef]
9. Liu, C.; Yuan, J.; Liu, J.; Li, C.; Zhou, Z.; Gu, Y. ARM and DSP-based bivariable fertilizing control system design and implementation. *Trans. CSAM* **2010**, *41*, 233–238.
10. Zhang, J.Q.; Liu, G.; Luo, C.M.; Hu, H.; Huang, J.Y. MOEA/D-DE based bivariate control sequence optimization of a variable-rate fertilizer applicator. *Comput. Electron. Agric.* **2019**, *167*, 105063. [CrossRef]
11. Yuan, J.; Liu, C.; Gu, Y.; Miao, Z. Bivariate fertilization control sequence optimization based on relevance vector machine. *Trans. Chin. Soc. Agric. Mach.* **2011**, *42*, 184–189.
12. Zhang, J.Q.; Liu, G. Effects of control sequence optimisation on the performance of bivariate fertiliser applicator. *Comput. Electron. Agric.* **2022**, *192*, 106594. [CrossRef]
13. Yuan, J.; Liu, C.L.; Li, Y.M.; Zeng, Q.B.; Zha, X.F. Gaussian processes based bivariate control parameters optimization of variable-rate granular fertilizer applicator. *Comput. Electron. Agric.* **2010**, *70*, 33–41. [CrossRef]
14. Kartashov, O.O.; Chernov, A.V.; Polyanichenko, D.S.; Butakova, M.A. XAS Data Preprocessing of Nanocatalysts for Machine Learning Applications. *Materials* **2021**, *14*, 7884. [CrossRef]

15. Khalil, A.; Salahuddin; Mashwani, W.K.; Shafiq, M.; Hassan, S.; Kumam, W. New advanced outliers detection tests. *Commun. Stat.-Theory Methods* **2021**, *50*, 1640–1655. [CrossRef]
16. Guilizzoni, M.; Eizaguirre, P.M. Trend Lines and Japanese Candlesticks Applied to the Forecasting of Wind Speed Data Series. *Forecasting* **2022**, *4*, 165–181. [CrossRef]
17. Gu, Y.; Yuan, J.; Liu, C. FIS-based method to generate bivariate control parameters regulation sequence for fertilization. *Trans. Chin. Soc. Agric. Eng.* **2011**, *27*, 134–139. [CrossRef]
18. Wang, L.; Liao, Q.; Liao, Y.; Gao, L.; Xiao, W.; Chen, H. Effects of distributor types on fertilizing performance in an air-assisted applicator. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 24–34. [CrossRef]
19. Deb, K.; Jain, H. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. *IEEE Trans. Evol. Comput.* **2014**, *18*, 577–601. [CrossRef]
20. Jain, H.; Deb, K. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point Based Nondominated Sorting Approach, Part II: Handling Constraints and Extending to an Adaptive Approach. *IEEE Trans. Evol. Comput.* **2014**, *18*, 602–622. [CrossRef]
21. Wang, C.; Ji, Z.C.; Wang, Y. Many-objective flexible job shop scheduling using NSGA-III combined with multi-attribute decision making. *Mod. Phys. Lett. B* **2018**, *32*, 1840110. [CrossRef]
22. Yeganefar, A.; Niknam, S.A.; Asadi, R. The use of support vector machine, neural network, and regression analysis to predict and optimize surface roughness and cutting forces in milling. *Int. J. Adv. Manuf. Technol.* **2019**, *105*, 951–965. [CrossRef]
23. Liang, W.; Wang, G.W.; Ning, X.J.; Zhang, J.L.; Li, Y.J.; Jiang, C.H.; Zhang, N. Application of BP neural network to the prediction of coal ash melting characteristic temperature. *Fuel* **2020**, *260*, 116324. [CrossRef]
24. Ding, S.F.; Zhao, H.; Zhang, Y.N.; Xu, X.Z.; Nie, R. Extreme learning machine: Algorithm, theory and applications. *Artif. Intell. Rev.* **2015**, *44*, 103–115. [CrossRef]
25. Tipping, M.E. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **2001**, *1*, 211–244. [CrossRef]
26. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef]
27. Das, I.; Dennis, J.E. Normal-boundary intersection: A new method for generating the Pareto surface in nonlinear multicriteria optimization problems. *SIAM J. Optim.* **1998**, *8*, 631–657. [CrossRef]
28. Beume, N.; Naujoks, B.; Emmerich, M. SMS-EMOA: Multiobjective selection based on dominated hypervolume. *Eur. J. Oper. Res.* **2007**, *181*, 1653–1669. [CrossRef]
29. Zizler, E. Multiobjective optimization using evolutionary algorithms: A comparative study. *Parallel Probl. Solving Nat.* **1998**, *1498*, 292–301. [CrossRef]
30. Zitzler, E.; Thiele, L. Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto approach. *IEEE Trans. Evol. Comput.* **1999**, *3*, 257–271. [CrossRef]
31. Tiwari, A.; Sharma, K.; Trivedi, M.K. NSGA-III-Based Time–Cost–Environmental Impact Trade-Off Optimization Model for Construction Projects. In *Artificial Intelligence and Sustainable Computing*; Springer: Singapore, 2022; pp. 11–25.

agriculture

MDPI

Article

Hyperspectral Imaging-Based Multiple Predicting Models for Functional Component Contents in *Brassica juncea*

Jae-Hyeong Choi [1,2], Soo Hyun Park [1], Dae-Hyun Jung [1,3], Yun Ji Park [1], Jung-Seok Yang [1], Jai-Eok Park [1], Hyein Lee [1] and Sang Min Kim [1,2,*]

[1] Smart Farm Research Center, KIST Gengneung Institute of Natural Products, Gangneung 25451, Korea
[2] Department of Bio-Medical Science & Technology, KIST School, University of Science and Technology, Seoul 02792, Korea
[3] Department of Smart Farm Science, Kyung Hee University, Yongin 17104, Korea
* Correspondence: kimsm@kist.re.kr; Tel.: +82-33-650-3640; Fax: +82-33-650-3679

Abstract: Partial least squares regression (PLSR) prediction models were developed using hyperspectral imaging for noninvasive detection of the five most representative functional components in *Brassica juncea* leaves: chlorophyll, carotenoid, phenolic, glucosinolate, and anthocyanin contents. The region of interest for functional component analysis was chosen by polygon selection and the extracted average spectra were used for model development. For pre-processing, 10 combinations of Savitzky–Golay filter (S. G. filter), standard normal variate (SNV), multiplicative scatter correction (MSC), 1st-order derivative (1st-Der), 2nd-order derivative (2nd-Der), and normalization were applied. Root mean square errors of calibration (RMSEP) was used to assess the performance accuracy of the constructed prediction models. The prediction model for total anthocyanins exhibited the highest prediction level (R_V^2 = 0.8273; RMSEP = 2.4277). Pre-processing combination of SNV and 1st-Der with spectral data resulted in high-performance prediction models for total chlorophyll, carotenoid, and glucosinolate contents. Pre-processing combination of S. G. filter and SNV gave the highest prediction rate for total phenolics. SNV inclusion in the pre-processing conditions was essential for developing high-performance accurate prediction models for functional components. By enabling visualization of the distribution of functional components on the hyperspectral images, PLSR prediction models will prove valuable in determining the harvest time.

Keywords: hyperspectral image; partial least squares regression; prediction models; root mean square error of prediction; standard normal variate; total anthycyanins; total carotenoids; total chlorophylls; total glucosinolates; total phenolics

Citation: Choi, J.-H.; Park, S.H.; Jung, D.-H.; Park, Y.J.; Yang, J.-S.; Park, J.-E.; Lee, H.; Kim, S.M. Hyperspectral Imaging-Based Multiple Predicting Models for Functional Component Contents in *Brassica juncea*. *Agriculture* **2022**, *12*, 1515. https://doi.org/10.3390/agriculture12101515

Academic Editors: Gniewko Niedbała and Sebastian Kujawa

Received: 17 August 2022
Accepted: 19 September 2022
Published: 21 September 2022

1. Introduction

Brassica juncea is a member of the family Brassicaceae whose leaves contain a variety of functional components including chlorophyll, carotenoid, flavonoid, and phenolic components, as well as glucosinolate and anthocyanin components [1–5]. Functional component content in crop plants can vary extensively within a wide range, depending on the cultivation conditions. The bioactivities of such functional components and their applications as pharmacological agents, health functional foods [6], and cosmetic materials [7] have been extensively studied. For example, chlorophylls and anthocyanins are components that confer the characteristic green and red colors to plant organs, respectively, and they may also indicate crop management or growth conditions in wilted or diseased plant organs [8,9]. In addition, carotenoids have been investigated with respect to their biological effects in anti-obesity activity [10], and flavonoids, phenolics, and anthocyanins have been investigated for anti-obesity [11], antioxidant, and antimicrobial activities [12,13]. The glucosinolate derivative, isothiocyanate, has been investigated for anticancer [14,15] and antimicrobial activities [16].

Recently, there has been growing interest in detection and quantification of specific bioactive substances for targeted applications. Concomitantly, studies involving the development of faster and efficient detection methods have attracted attention. The most widely used detection method involves the collection of plant samples coupled with invasive extraction using organic solvents, followed by analysis of the extracts. Such detection methods involving organic solvents and chemical reagents are time consuming, less efficient, and adversely affect the environment. To overcome these disadvantages of chemical methods, many scientists have conducted chemometrics studies based on spectroscopic knowledge to analyze the functional components in plants, such as High-Performance Liquid Chromatography (HPLC) and Near-Infrared (NIR) [17], or Fourier Transform-NIR (FT-NIR) spectroscopy [18,19]. However, these methods are invasive and inefficient, and data processing is time consuming. Alternatively, recent studies have focused on non-invasive methods. For example, a hyperspectral imaging system was used to explore various parameters including plant-derived functional components, such as total chlorophylls and total carotenoids [20], total capsaicinoids [17], total glucosinolates [21], total flavonoids [22–24], and total polysaccharides [24], as well as microbial contamination of fish flesh [25], fruit moisture content [26], and sulfite dioxide residues on fruit surfaces [27]. Among the studies aimed at developing prediction models for functional components, prediction of total anthocyanins [28] at $R_C^2 = 0.883$ and $R_V^2 = 0.830$ and total polyphenols [29] at $R_C^2 = 0.820$ and $R_V^2 = 0.551$ have been reported. However, in these studies, only one or two functional components were targeted for prediction model construction by using hyperspectral images.

In this study, it was hypothesized that multiple components could be predicted non-invasively and spontaneously from hyperspectral images from the leaves of *B. juncea*. In order to develop a prediction model for multiple components, a partial least squares regression (PLSR) method was used, with 10 pre-processing combinations. Compared to other regression models, the PLSR method enables the development of a predictive model with high accuracy using only a relatively small amount of data. For this reason, it is a representative method widely used to make industrialized products because it reduces the data acquisition time, and the model application is simple. Total chlorophylls, total carotenoids, total phenolics, total glucosinolates, and total anthocyanins were selected as target functional components for prediction, because these components are the most general functional components in plants which have a beneficial effect for human health and can be estimated by spectrophotometric method after simple solvent extraction of plant samples. *B. brassica* was cultivated under various conditions to obtain functional components with a wide range of concentrations. The harvested leaves were used for hyperspectral image and functional component quantification. In addition, visualization software was developed to detect real-time distribution of target components. The results in this study suggest the possibility of noninvasive multiple component prediction from one hyperspectral image. In addition, models and visualization software were applied to agricultural systems (such as growth chamber, indoor farm, and greenhouse) to monitor real-time functional components.

2. Materials and Methods

2.1. Plant Growth Conditions

To obtain *Brassica juncea* (L.) Czern. leaf samples with varying concentrations of functional components, three cultivation environments were implemented: an indoor farm, a greenhouse, and an open field (Figure 1). *B. juncea* was cultivated in an indoor farm in a hydroponic system under red, blue, and green LED light combination at 18–23 °C; Hoagland nutrient solution with an electrical conductivity (EC) value of 1.5 dS/m was used as a growth medium. Greenhouse and open field cultivation were carried out on the soil with fertilizer components at 20–28 °C for greenhouse and 15–20 °C under the sunlight for open field, respectively. The leaves of *B. juncea* were collected after 6 weeks of cultivation in each environment. Considering the growth phase and varied distributions of

leaf colors, 15–20 full-grown leaves from each cultivation environment were harvested. A total of 55 leaves were sampled and stored at −20 °C until further analysis. Spectral data were obtained for all the samples by using hyperspectral imaging. After four days of freeze-drying and subsequent grinding, each sample's powder was divided into 20 mg portions in triplicate for analysis of the five functional components. Content values with a large deviation for each functional component were excluded, and the mean of the remaining values was used as the component content value.

Figure 1. Cultivation features of *B. juncea* in indoor farm (**A**), greenhouse (**B**), and outdoor field (**C**). The plants were harvested at the sixth week of cultivation. The growth temperature for crop cultivation was maintained at 18–23 °C in indoor farm (**A**), 20–28 °C in greenhouse (**B**), and 15–20 °C in outdoor field.

2.2. Total Chlorophyll and Carotenoid Contents

Measurements were made using extracts prepared in triplicate by adding 2 mL of 90% MeOH containing 10% water (*v*/*v*) to 20 mg of each sample, followed by sonication for 1 h at 40 °C. The resulting crude extract was centrifuged at 4000 rpm and 4 °C for 20 min to separate plant debris and the supernatant. The supernatant was filtered using a 0.45 µm membrane filter; 1.5 mL of this filtrate was collected to use in quantification of functional components. For analysis, 150 µL from the 1.5 mL of the filtered supernatant was mixed with 90% MeOH containing 10% water to prepare 1.5 mL of a 10× diluted solution. Following the method described previously [30], the absorbance of the diluted sample solution was measured at 665.2, 652.4, and 470.0 nm wavelengths using a spectrophotometer (Cary 60 UV-Vis, Agilent Technologies, Santa Clara, CA, USA). The specification of spectrophotometer includes Xenon Flash Lamp (80 Hz) as a light source, measuring wavelengths from 190 nm to 1100 nm with a resolution of 1.5 nm. The scanning speed of this equipment is 24,000 nm per min. Absorbance (A) at each wavelength was used in Equations (1)–(3) to calculate total chlorophyll a, total chlorophyll b, and total carotenoids respectively.

$$\text{Chl}a\ (\mu\text{g mL}^{-1}) = 16.82\ A_{665.2} - 9.28\ A_{652.4} \quad (1)$$

$$\text{Chl}b\ (\mu\text{g mL}^{-1}) = 36.92\ A_{652.4} - 16.54\ A_{665.2} \quad (2)$$

$$\text{Car}\ (\mu\text{g mL}^{-1}) = \frac{(1000\ A_{470.0} - 1.91\ \text{Chl}a - 95.15\ \text{Chl}b)}{225} \quad (3)$$

where $A_{665.2}$ is absorbance at 665.2 nm; $A_{652.4}$ is absorbance at 652.4 nm; $A_{470.0}$ is absorbance at 470 nm; Chl*a* stands for total chlorophyll a; Chl*b* stands for total chlorophyll b; Chl*a* + Chl*b* stands for total chlorophyll content; and Car stands for total carotenoid content. The data are expressed as mean ± standard deviation mg g^{-1} dry weight (DW) from biological triplicates.

2.3. Total Phenolic Contents

A previously described method [31] was used after modification for the detection of total phenolic content. For analysis, 100 µL from the 1.5 mL of supernatant (Section 2.2) was mixed with 100 µL of Folin-Ciocalteu reagent and 1.5 mL of distilled water; the mixture was incubated for 5 min at room temperature; then, 300 µL of 7.5% Na_2CO_3 solution was added and the mixture was allowed to react for 1 h at room temperature. Absorbance was

measured at 765.0 nm using a spectrophotometer. A standard curve prepared using 25, 50, 100, and 250 ppm gallic acid standard solutions was used to calculate total phenolic content in the samples from their measured absorbance. The data are expressed as mean ± standard deviation mg g^{-1} DW from biological triplicates.

2.4. Total Glucosinolate Contents

The extract prepared by the method described in 2.2 was used to obtain 1.5 mL of supernatant of the crude extract. For analysis, 50 µL from the 1.5 mL of supernatant was mixed with 150 µL of distilled water and 1.5 mL of 2 mM sodium tetrachloropalladate (II); the mixture was left to react for 1 h at room temperature. The absorbance was measured at 425.0 nm using a spectrophotometer. The absorbance value was then used to calculate total glucosinolate content according to Equation (4) [32]:

$$y\ (\mu\text{mol g}^{-1}) = 1.40 + 118.86\ A_{425.0} \quad (4)$$

where $A_{425.0}$ is the absorbance at 425.0 nm. The data are expressed as mean ± standard deviation µmol g^{-1} DW from biological triplicates.

2.5. Total Anthocyanin Contents

The method described in a previous study [33] was modified and used for the estimation of the total anthocyanins. Sample extracts were prepared in triplicate. Briefly, 2 mL of acidic MeOH containing 1% HCl (*v*/*v*) was added to 20 mg of powder sample (Section 2.1) and sonicated for 1 h at 60 °C. The resulting crude extract was centrifuged and filtered in the same way as in the pre-processing of samples for total chlorophylls (Section 2.2). Then, 300 µL of the 1.5 mL of filtered supernatant was diluted by adding MeOH containing 1% HCl, and a 5× diluted solution of 1.5 mL volume was prepared. Absorbance of this solution was measured at 530.0 and 600.0 nm wavelengths, and the values obtained were used to calculate total anthocyanin contents according to Equation (5), as reported previously [33].

$$y\left(\text{mg g}^{-1}\right) = (A_{530.0} - A_{600.0})\ \frac{V \times n \times Mw}{\varepsilon \times m} \quad (5)$$

where $A_{530.0}$ is the absorbance at 530.0 nm; $A_{600.0}$ is the absorbance at 600.0 nm; V is the total volume of the extracted solution; n is the dilution ratio; Mw is the molecular weight of cyanidin-3-glucoside (i.e., 449.4); ε is the molar extinction coefficient of anthocyanin (29,600 M^{-1} cm^{-1}); and m is the mass of the sample. The data are expressed as mean ± standard deviation mg g^{-1} DW from biological triplicates.

2.6. Hyperspectral Imaging

A hyperspectral imaging camera (MicroHSI 410 SHARK, Corning Inc., Corning, NY, USA) was used. The detailed specifications of the hyperspectral camera used in this study are summarized in Table 1. As shown in Figure 2, the hyperspectral imaging system was equipped with eight halogen lamps (15 W × 8) as light sources; for data scanning, the hyperspectral camera was moved by the conveyor at the top of a dark chamber that blocked all external light. Considering the time and speed for a single scan, two or three samples were measured per one hyperspectral image at a moving speed of 100 mm s^{-1}.

Table 1. MicroHSI™ 410-SHARK specifications.

Item.	Specification
No. of spatial pixels	1408 spatial pixels
Focal Length, f-number	16 mm, f/1.4 standard
Spectral Range	400–1000 nm
Full FOV	28.6 degrees (500 mrad) standard

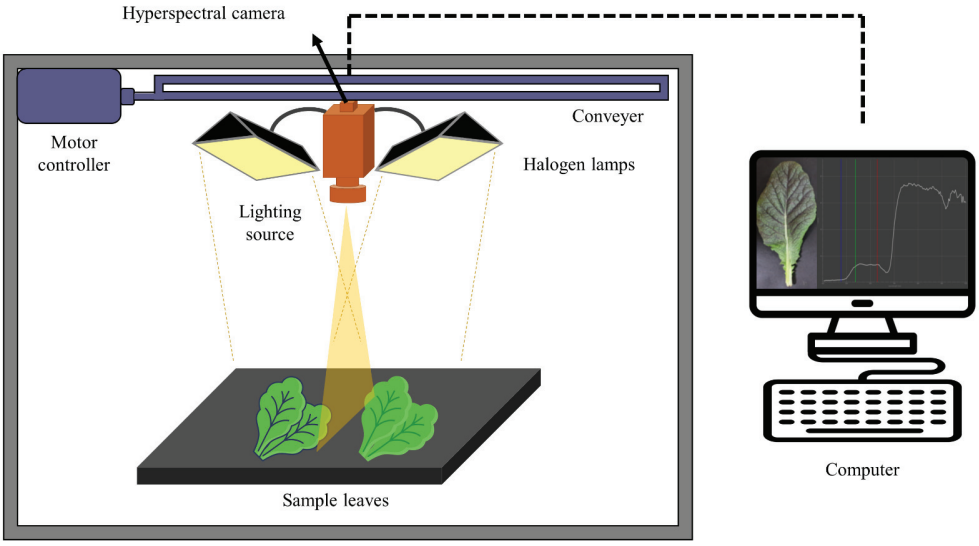


Figure 2. Hyperspectral imaging system (HSI) and extraction of spectral data from computer.

2.7. Data Processing and Prediction Models

The spectral data required for model development were extracted from the hyperspectral image data of *B. juncea* leaves, for which the spectral library in the Python 3.8 environment was used. After obtaining the RGB images of three bands (Blue: 456.55 nm, Red: 544.61 nm, and Green: 660.67 nm), the area of the sample for functional component analysis was chosen by polygon selection and set as the region of interest (ROI) (Figure 3). An area of 1 cm width on the leaf edge was not included in the ROI or used in the analysis, as it was presumed to contain much background noise. Further, the area of the leaf vein was also excluded in the experiments due to the relatively low content of functional components. The average spectra of the ROI across all samples was used for PLSR model development after checking the correlation with each functional component based on absorbance wavelengths.

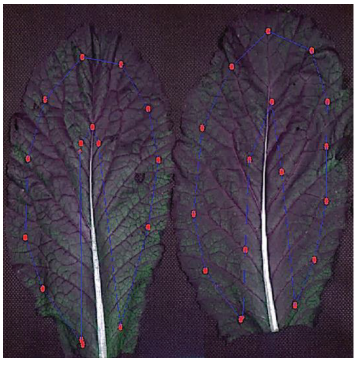


Figure 3. Extraction of the region of interest (ROI) by polygon selection.

To predict each functional component, the extracted average spectra were applied to the Unscrambler X software v.11 (CAMO, Oslo, Norway) for model development through PLSR with full cross validation. The maximum number of PLS components used was limited up to 20. Various data pre-processing protocols were combined with PLSR and tested to ensure outstanding prediction performance of the developed model with reduced

noise signal in the spectral data, including the Savitzky–Golay filter (S. G. filter) with 3 or 7 smoothing points, standard normal variate (SNV), multiplicative scatter correction (MSC), mean normalization, 1st-order derivative (1st-Der), and 2nd-order derivative (2nd-Der). The applied pre-processing combinations are listed in Table 2.

Table 2. Pre-processing combinations tested in this study.

Methods	Pre-Processing Conditions
1	Raw data
2	Raw data, S. G. filter (interval = 3)
3	Raw data, S. G. filter (interval = 7)
4	Raw data, S. G. filter (interval = 3), SNV
5	Raw data, S. G. filter (interval = 3), MSC
6	Raw data, 1st-Der
7	Raw data, 2nd-Der
8	Raw data, SNV, 1st-Der
9	Raw data, SNV, 2nd-Der
10	Raw data, Normalization

The performance of the PLSR-based prediction models developed in this study was evaluated based on the R^2 and root mean square errors (*RMSE*) of calibration and validation, respectively, using Equations (6) and (7). From the prediction models developed with various pre-processing combinations, the model exhibiting comparatively high $R_V{}^2$ and low RMSE for validation was selected.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \tag{7}$$

where y_i is the measured value of component obtained by analysis; $\hat{y}_i$ is the value predicted by the model; $\overline{y}$ is the mean value of component from the analysis; and n is the number of samples, which was a total of 55 numbers for this experiment.

2.8. Development of Visualization Software for Applying Predictive Models

Visualization software was produced and utilized so that the developed model could be recognized intuitively. PyQT5, OpenCV, Pillow, Spectral, Matplotlib, Numpy, Loguru, and Pandas libraries were used in a Python 3.8 environment. The GUI was constructed by using PyQT5. The predicted values of the components were calculated by multiplying the weights at each wavelength on the spectrum obtained at each pixel from hyperspectral image data. The predicted values for each pixel were visualized with a jet colormap. Color bar means the predicted value of functional component by PLSR models.

3. Results and Discussion

3.1. Analysis of Plant Pigments and Metabolites

Table 3 summarizes the results of the quantification of five main functional components found in the leaves of *B. juncea* (i.e., total chlorophyll, total carotenoid, total phenolic, total glucosinolate, and total anthocyanin content). With respect to total chlorophyll and total anthocyanin content, which are responsible for leaf color, the former was relatively low when the anthocyanin content was high. Total chlorophyll content ranged from 2.13 to 11.70 mg g^{-1} DW, with a mean value of 6.33 $\pm$ 2.21 mg g^{-1} DW. The amount of total carotenoids in *B. juncea* is generally low; it had the lowest mean value (0.91 mg g^{-1} DW) among the five functional components under study herein. The total phenolic contents ranged from 2.11 to 9.56 mg g^{-1} DW, with a mean of 4.85 $\pm$ 1.94 mg g^{-1} DW. A common feature of the family Brassicaceae is the abundance of glucosinolate components; therefore,

B. juncea leaves exhibited the highest minimum, maximum, and mean values for total glucosinolate content among all five functional components analyzed. Anthocyanins are responsible for the red and blue color of *B. juncea* leaves, and the color intensity varies depending on the cultivation conditions. A high level of anthocyanins reportedly requires blue light, cool climate, and a large daily temperature range [34–36]. Consistently, in the field experiment, *B. juncea* cultivated in the cool outdoor field environment exhibited a relatively higher level of anthocyanin production, whereas *B. juncea* cultivated in the glass greenhouse under higher temperatures resulted in intense light-green-to-green leaves according to hyperspectral imaging. Thus, total anthocyanin content varied greatly from 0 to 33.80 mg g^{-1} DW, with a mean of 5.41 ± 6.75 mg g^{-1} DW.

Table 3. Minimum, maximum, and mean content for the five functional components in *B. juncea*.

Parameter	Min *	Max *	Mean *	Standard Deviation *
total chlorophylls [a]	2.13 *	11.70 *	6.33 *	2.21 *
total carotenoids [a]	0.21	1.59	0.91	0.29
total phenolics [a]	2.11	9.56	4.85	1.94
total glucosinolates [b]	8.62	52.89	24.49	9.41
total anthocyanins [a]	0	33.80	5.41	6.75

* Units of measurement, [a]: mg g^{-1} DW; [b]: µmol g^{-1} DW.

3.2. Average Spectra and Correlation Analysis

The average spectra are shown in Figure 4A. The graphs showing the correlation of the contents of the five functional components at each wavelength of spectral data are shown in Figure 4B–F. Except for total phenolic contents, the other four functional components studied herein exhibited the highest negative correlation coefficient in the range of 400–600 nm of the visible light region of the spectrum. Total phenolic content showed a positive correlation in the wavelength bands corresponding to blue and green regions. This accounted for the high total phenolic content in *B. juncea* leaves with a relative increase in green areas, and an increase in total chlorophylls, total carotenoids, total glucosinolates, and total anthocyanins with a relative reduction in green areas. The potential use of prediction models for functional components quantitation via PLSR analysis was verified using correlation coefficient values of approximately 0.5 for all components, evenly distributed across the wavelength bands.

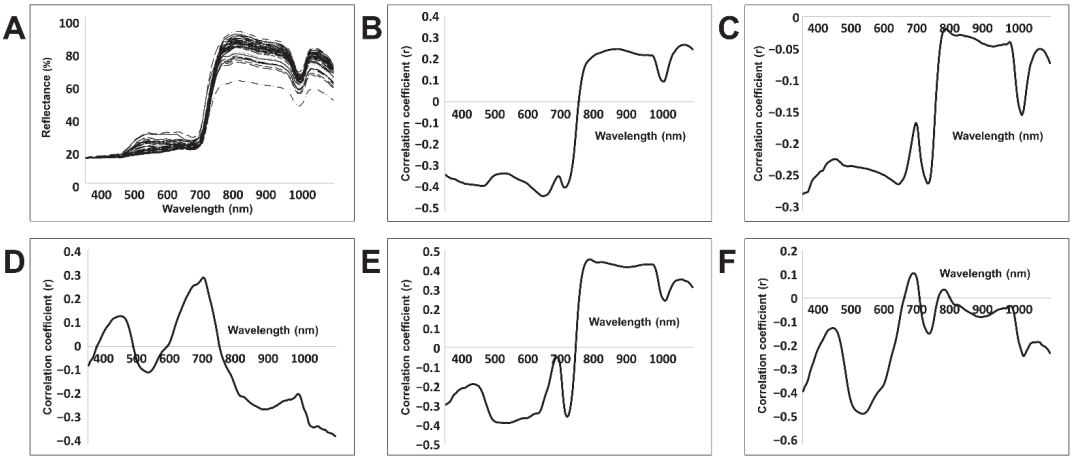


Figure 4. Average spectra (**A**) and correlation coefficient graphs for the contents of the main functional components present in leaves of *B. juncea* (total chlorophylls, (**B**); total carotenoids (**C**); total phenolics (**D**); total glucosinolates (**E**); total anthocyanins (**F**)).

3.3. Development of PLSR Models Using Spectral Data Extracted from HSI

The PLSR prediction models for the prediction of five important functional components in plant leaf tissues were developed by applying various pre-processing combinations. The performance of each prediction model according to the pre-processing combination for each of the five functional components is shown in Table 4. Of all the five functional components studied, the highest prediction accuracy was recorded for total anthocyanins in almost all pre-processing methods. Notably, among the ten pre-processing combinations tested, the 6th combination registered R_V^2 and RMSEP values of 0.8273 and 2.4277, respectively, indicating the highest level of prediction performance. This may be because the distribution of anthocyanins, referred to as 'purple magic' by Kim et al. [37], accounted for the largest purple areas over the plant body. When dark-red areas on *B. juncea* leaves were visibly greater, the anthocyanin content was higher, which may have influenced the spectrum extracted from hyperspectral images. In turn, the highest prediction performance for total phenolic contents was registered for the 4th pre-processing combination, i.e., S. G. filter and SNV. Apart from these two components, the highest prediction performance for total chlorophyll, total carotenoid, and total glucosinolate content was observed for the 8th pre-processing conditions with SNV and 1st-Der. Pre-processing condition SNV is shared by the 4th and the 8th combinations, which is a well-known pre-processing method of normalization based on the standard deviation of the overall spectrum to eliminate the influence of light scattering. Presumably, the variability of the spectrum—which may arise from the vibration of the migration module while producing the hyperspectral image data—had been calibrated. In addition, pre-processing by 1st-Der, which is shared by the 6th and the 8th combinations, is effective in calibrating the baseline variations originating from the difference in relative intensity of the light sources. This is because the method differentiates the spectrum to place an emphasis on the changes within the absorption bands to amplify the spectral variation, while only the variation is shown. This is presumed to account for the better performance of the prediction model with the 8th pre-processing combination for total chlorophylls, total carotenoids, and total glucosinolates.

Table 4. PLSR prediction model outcomes on functional components for each pre-processing combination. The prediction model with the best performance among each pre-processing method was determined to have the lowest RMESP values (marked in bold).

Pre-Processing Method	Total Chlorophylls				Total Carotenoids			
	Calibration		Validation		Calibration		Validation	
	R_C^2 *	RMSEC *	R_V^2 *	RMSEP *	R_C^2	RMSEC	R_V^2	RMSEP
1	0.6475	1.3538	0.4066	1.6701	0.6419	0.1267	0.4485	0.1525
2	0.6747	1.2651	0.4145	1.6100	0.6403	0.1270	0.4570	0.1514
3	0.6732	1.2437	0.3893	1.6406	0.6455	0.1272	0.4073	0.1574
4	0.7001	1.2236	0.4818	1.5154	0.6569	0.1056	0.4331	0.1343
5	0.6795	1.2340	0.4589	1.5475	0.6346	0.1077	0.4165	0.1363
6	0.6546	1.3105	0.4278	1.6300	0.5979	0.1086	0.3172	0.1431
7	0.2853	1.7787	0.2121	1.8574	0.2853	1.7787	0.2121	1.8574
8	**0.6842**	**1.1832**	**0.5350**	**1.4045**	**0.6775**	**0.1040**	**0.4683**	**0.1315**
9	0.2755	1.7886	0.2160	1.8528	0.1807	0.1500	0.1080	0.1555
10	0.6822	1.2316	0.4384	1.5721	0.6512	0.1087	0.3983	0.1398

Pre-Processing Method	Total Phenolics				Total Glucosinolates				Total Anthocyanins			
	Calibration		Validation		Calibration		Validation		Calibration		Validation	
	R_C^2	RMSEC	R_V^2	RMSEP	R_C^2	RMSEC	R_V^2	RMSEP	R_C^2	RMSEC	R_V^2	RMSEP
1	0.7657	0.9461	0.6625	1.1195	0.8587	3.4785	0.7429	4.5713	0.9296	1.5415	0.8194	2.5925
2	0.8322	0.7837	0.6955	1.0571	0.8525	3.5799	0.7365	4.6769	0.7903	2.6361	0.6861	3.2571
3	0.7620	0.9911	0.6258	1.2247	0.8352	3.8748	0.7226	4.8702	0.6910	3.3728	0.6255	3.7414
4	**0.8204**	**0.8151**	**0.6909**	**1.0474**	0.8465	3.6502	0.7635	4.4275	0.9312	1.5469	0.8378	2.6025
5	0.7563	0.9475	0.6213	1.1641	0.8481	3.6354	0.7350	4.6763	0.8835	1.9652	0.7871	2.7335
6	0.7869	0.8908	0.6398	1.1517	0.8237	3.8265	0.7119	4.8708	**0.9144**	**1.6814**	**0.8273**	**2.4277**
7	0.6466	1.0884	0.4591	1.3577	0.9591	1.8771	0.6729	5.2963	0.9273	1.8615	0.7718	3.3002
8	0.8157	0.8652	0.6983	1.0794	**0.8508**	**3.4289**	**0.7827**	**4.0647**	0.9072	2.1886	0.7938	3.1842
9	0.7981	0.8477	0.4419	1.4260	0.7414	4.7484	0.6175	5.7491	0.8060	2.6448	0.7150	3.2259
10	0.7659	0.9329	0.6647	1.0989	0.8552	3.5580	0.7419	4.9688	0.9102	1.7609	0.8034	2.7358

* Rc^2—R^2 value in calibration; RMSEC—RMSE value in calibration; Rv^2—R^2 value in validation; RMSEP—RMSE value in validation. Bold means the values showing the best performance.

Notably, the prediction models for total chlorophyll and total carotenoid content led to lower performance in prediction performance, with $R^2 < 0.3$ for the 7th and 9th pre-processing combinations sharing pre-processing by 2nd-Der, which shows the characteristics of the changes in spectral slope to effectively calibrate the baseline and remove any micro-noise that may appear in the system, along with pre-processing by 1st-Der. Nevertheless, it is possible to interpret that the removal of noise resulted in the removal of the effects of absorption bands created by micro-components, thereby lowering prediction performance. In previous studies, total phenolics had been predicted at $R_C^2 = 0.820$ and $R_V^2 = 0.551$ by Caporaso et al. [29], and total anthocyanins had been predicted at $R_C^2 = 0.883$ and $R_V^2 = 0.830$ by Liu et al. [28]. However, the results reported herein were higher, at $R_C^2 = 0.8204$ and $R_V^2 = 0.6909$ for total phenolics, and at $R_C^2 = 0.9144$ and $R_V^2 = 0.8273$ for total anthocyanins. Furthermore, the 4th and 8th pre-processing combinations in the PLSR models for the five functional components resulted in the most outstanding prediction performance, implying that the use of pre-processing with SNV in the prediction spectrum for the five components is essential for the development of high-prediction performance models. Figure 5 depicts the results of the prediction model with the highest prediction performance among all models developed. The selected models did not have a very high prediction rate for functional components; nevertheless, the results verify their potential for predicting the concentrations of multiple functional components in actual cultivation conditions of *B. juncea*. In addition, the development of other regression models such as machine learning, mixed model, or principal components analysis have the potential to lead better predictive performance. It is judged that the combination of pre-processing methods used in this study is sufficiently worthy of reference for application to other regression models.

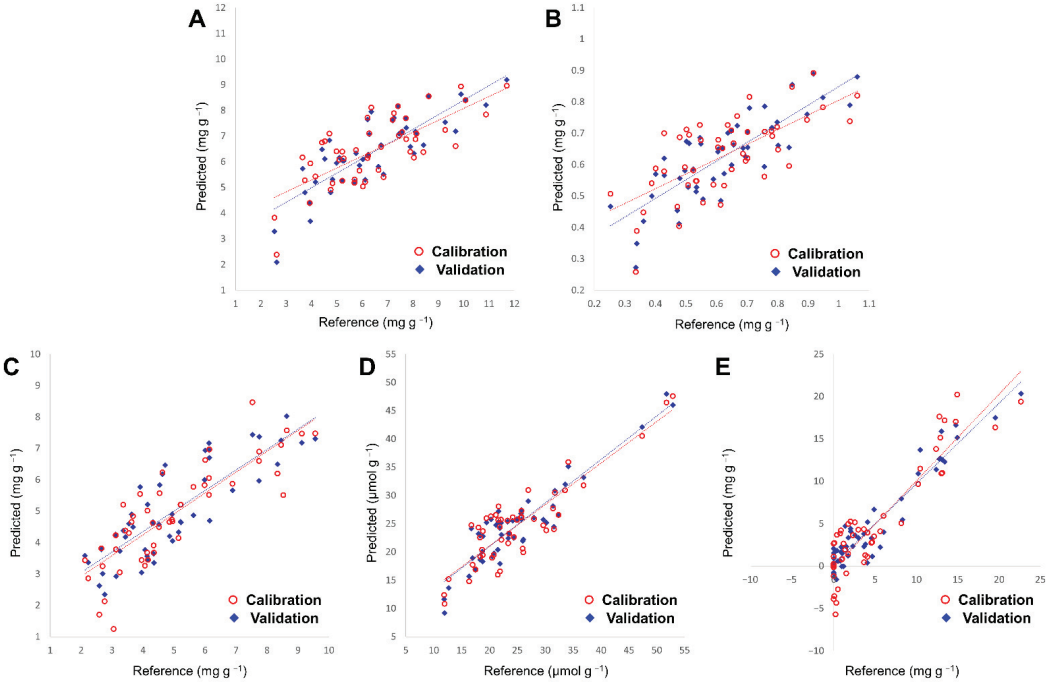


Figure 5. Prediction models for each functional component: (**A**) total chlorophylls, (**B**) total carotenoids, (**C**) total phenolics, and (**D**) total glucosinolates with the 8th pre-processing combination; (**E**) total anthocyanins with the 4th pre-processing combination. Fifth-five samples were used to establish each prediction model.

Hyperspectral imagery spectroscopy is a technology that uses spectroscopic techniques and imaging which can offer the sample's optical information. Therefore, even if the amounts of components are the same, the spectral properties will not be the same if the distribution of these components in the plant are different. Thus, direct application of models in this study to other plants is difficult. New leaning data from other plants are required to make prediction model of new plants. However, it is noteworthy that the kind of pre-processing method and the methodology presented in this paper can be applied to new plants to develop prediction models for multiple components quickly. These trials can expand the scientific meaning and industrial uses related to engineering based on a hyperspectral imaging system.

3.4. Application of the Functional Component Prediction Model for Visualization

For the prediction models developed for the five plant functional components under study, visualization could probably be achieved in the form of a distribution map based on the prediction of component values by the prediction model from the spectrum values in the unit of pixels obtained from the hyperspectral image. Figure 6 shows such visualization based on PLSR prediction models for each pixel representation of the concentrations of functional components according to the colors on the color map. Hence, the variation of different colors from high to low concentrations of functional components in the leaf can be detected intuitively with respect to the distribution using the color map. Similarly, cultivation in the open air under actual sunlight conditions is likely to allow enhanced accuracy of the prediction model with training data, which in turn is likely to enable monitoring of the functional components during cultivation. This should enable farmers to determine the best time for harvest based on the prediction of key functional components.

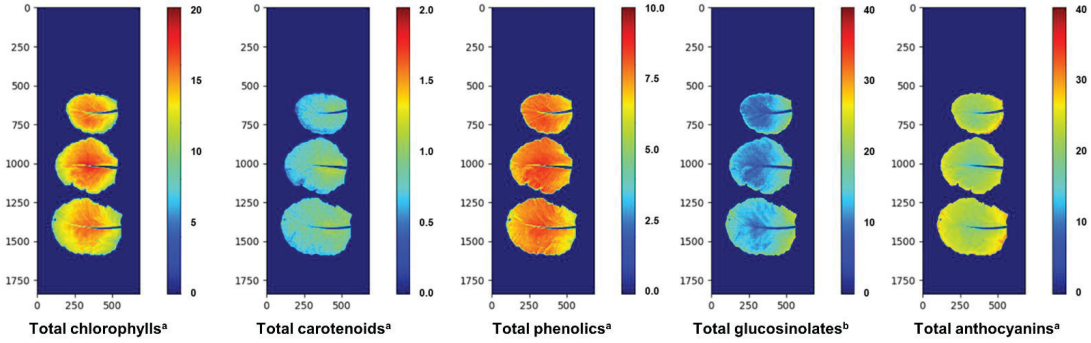


Figure 6. Distribution maps of plant functional components predicted by the developed PLSR models. *X*- and *Y*-axes indicate the size of images in pixels. Color bar means the predicted value of functional component by the developed PLSR model (units of measurement, [a]: mg g^{-1} DW; [b]: μmol g^{-1} DW).

4. Conclusions

Using hyperspectral imaging, PLSR models were developed for the prediction of the multiple contents of five key functional components in the leaves of *B. juncea*: chlorophylls, carotenoids, phenolics, glucosinolates, and anthocyanins. To develop the models, the region of interest for analysis of the functional components was uniformly chosen by polygon selection from the hyperspectral image data, and the average spectra were extracted. Various pre-processing combinations were applied to the spectral data to develop a model which showed the most outstanding prediction performance. The resulting PLSR prediction models had $R^2 \geq 0.8$ for total phenolic, total glucosinolate, and total anthocyanin contents, and valid models were thus obtained for each functional component. In addition, the models for total chlorophylls and total carotenoids had $R_C{}^2$ of 0.6842 and 0.6775, respectively, which implied the potential for development of more efficient models via further data processing in the future. Overall, among the ten pre-processing condi-

tions tested here, the 8th pre-processing condition combining SNV and 1st-Der resulted in the highest performance of prediction for functional components. Additionally, the 4th combination containing SNV exhibited outstanding prediction performance for total phenolic content. Hence, the most efficient pre-processing condition for the development of a prediction model exhibiting a high level of performance was SNV, shared by the 4th and 8th pre-processing combinations. In addition, the 'multiple-chemical' analysis from one spectral-image data could be effectively performed. Considering the application of the models developed here, it is possible to draw a distribution map for functional components by applying the spectral data in pixels from the hyperspectral images to the prediction models. The variation of different colors from high to low concentrations of functional components in the leaf can thus be detected intuitively with respect to their distribution, as the data for the content value of each functional component was associated with the color map. Based on our findings, the application of the model to measure hyperspectral images under actual cultivation in natural sunlight conditions is likely to allow real-time monitoring of the components of interest. Furthermore, the prediction for the functional components will contribute to accurate determination of the best time for harvest.

Author Contributions: Conceptualization, S.H.P. and S.M.K.; funding acquisition, S.M.K.; investigation, S.H.P. and S.M.K.; methodology, J.-H.C., D.-H.J., J.-S.Y., J.-E.P. and Y.J.P.; project administration, S.H.P.; software, H.L. and J.-H.C.; validation, J.-H.C. and S.H.P.; writing—original draft, J.-H.C. and S.H.P.; writing—review and editing, S.M.K. and S.H.P.; supervision, S.M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Korean Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET) and by the Korean Smart Farm R&D Foundation (KosFarm) through the Smart Farm Innovation Technology Development Program, funded by the Ministry of Agriculture, Food and Rural Affairs (MAFRA); the Ministry of Science and ICT (MSIT); and the Rural Development Administration (RDA) (421034-04).

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Kumar, V.; Thakur, A.K.; Barothia, N.D.; Chatterjee, S.S. Therapeutic potentials of *Brassica juncea*: An overview. *Cell Med.* **2011**, *1*, 2.1–2.16. [CrossRef]
2. Tian, Y.; Deng, F. Phytochemistry and biological activity of mustard (*Brassica juncea*): A review. *CyTA-J. Food* **2020**, *18*, 704–718. [CrossRef]
3. Malabed, R.S.; Noel, M.G.; Aton, B.C., III; Toribio, E.A.F. Characterization of the glucosinolates and isothiocyanates in mustard (*Brassica juncea* L.) extracts and determination of its myrosinase activity and antioxidant capacity. In Proceedings of the De La Salle University Research Congress 2014, Manila, Philippines, 6–8 March 2014; pp. 1–7.
4. Oulad El Majdoub, Y.; Alibrando, F.; Cacciola, F.; Arena, K.; Pagnotta, E.; Matteo, R.; Micalizzi, G.; Dugo, L.; Dugo, P.; Mondello, L. Chemical characterization of three accessions of *Brassica juncea* L. extracts from different plant tissues. *Molecules* **2020**, *25*, 5421. [CrossRef]
5. Cartea, M.E.; Francisco, M.; Soengas, P.; Velasco, P. Phenolic compounds in *Brassica* vegetables. *Molecules* **2010**, *16*, 251–280. [CrossRef] [PubMed]
6. Manohar, P.R.; Pushpan, R.; Rohini, S. Mustard and its uses in Ayurveda. *Indian J. Tradit. Knowl.* **2009**, *8*, 400–404.
7. Lee, J.E.; Kim, A.J. Antioxidant activity, whitening and anti-wrinkle effects of leaf and seed extracts of *Brassica juncea* L. Czern. *Asian J. Beauty Cosmetol.* **2020**, *18*, 283–295. [CrossRef]
8. Méthy, M.; Olioso, A.; Trabaud, L. Chlorophyll fluorescence as a tool for management of plant resources. *Remote Sens. Environ.* **1994**, *47*, 2–9. [CrossRef]
9. Jezek, M.; Zörb, C.; Merkt, N.; Geilfus, C.M. Anthocyanin management in fruits by fertilization. *J. Agric. Food Chem.* **2018**, *66*, 753–764. [CrossRef]
10. Mounien, L.; Tourniaire, F.; Landrier, J.F. Anti-obesity effect of carotenoids: Direct impact on adipose tissue and adipose tissue-driven indirect effects. *Nutrients* **2019**, *11*, 1562. [CrossRef]
11. Kawser Hossain, M.; Abdal Dayem, A.; Han, J.; Yin, Y.; Kim, K.; Kumar Saha, S.; Yang, G.M.; Choi, H.Y.; Cho, S.G. Molecular mechanisms of the anti-obesity and anti-diabetic properties of flavonoids. *Inter. J. Mol. Sci.* **2016**, *17*, 569. [CrossRef] [PubMed]
12. Kim, M.H. Antioxidant and antibacterial activity of extracts from *Brassica juncea* Czerniak et coss., *Celosia cristata* L., and *Beta vulgaris* L. *J. Korean Soc. Food Cult.* **2012**, *27*, 719–729. [CrossRef]

13. Nawaz, H.; Shad, M.A.; Muzaffar, S. Phytochemical composition and antioxidant potential of Brassica. In *Brassica Germplasm: Characterization, Breeding and Utilization*; IntechOpen: London, UK, 2018; Volume 1, pp. 7–26. [CrossRef]
14. Okulicz, M. Multidirectional time-dependent effect of sinigrin and allyl isothiocyanate on metabolic parameters in rats. *Plant Foods Hum. Nutr.* **2010**, *65*, 217–224. [CrossRef]
15. Zhang, Y. Allyl isothiocyanate as a cancer chemopreventive phytochemical. *Mol. Nutr. Food Res.* **2010**, *54*, 127–135. [CrossRef] [PubMed]
16. Luciano, F.B.; Holley, R.A. Enzymatic inhibition by allyl isothiocyanate and factors affecting its antimicrobial action against *Escherichia coli*. *Inter. J. Food Microbiol.* **2009**, *131*, 240–245. [CrossRef]
17. Mo, C.; Hasegawa, M.; Lee, K.; Lim, J.G.; Kim, M.S.; Kang, S.; Lee, H.D.; Bae, H.; Kim, D.Y.; Cho, B.K. Development of a non–destructive on–line pungency measurement system for red–pepper powder. *J. Fac. Agric. Kyushu Univ.* **2013**, *58*, 137–144. [CrossRef]
18. Teye, E.; Huang, X.; Sam-Amoah, L.K.; Takrama, J.; Boison, D.; Botchway, F.; Kumi, F. Estimating cocoa bean parameters by FT-NIRS and chemometrics analysis. *Food Chem.* **2015**, *176*, 403–410. [CrossRef]
19. Sunoj, S.; Igathinathane, C.; Visvanathan, R. Nondestructive determination of cocoa bean quality using FT-NIR spectroscopy. *Comput. Electron. Agric.* **2016**, *124*, 234–242. [CrossRef]
20. Zhao, Y.R.; Li, X.; Yu, K.Q.; Cheng, F.; He, Y. Hyperspectral imaging for determining pigment contents in cucumber leaves in response to angular leaf spot disease. *Sci. Rep.* **2016**, *6*, 27790. [CrossRef]
21. Hernández-Hierro, J.M.; Esquerre, C.; Valverde, J.; Villacreces, S.; Reilly, K.; Gaffney, M.; González-Miret, M.L.; Heredia, F.J.; O'Donnell, C.P.; Downey, G. Preliminary study on the use of near infrared hyperspectral imaging for quantitation and localisation of total glucosinolates in freeze-dried broccoli. *J. Food Eng.* **2014**, *126*, 107–112. [CrossRef]
22. Onivogui, G.; Zhang, H.; Mlyuka, E.; Diaby, M.; Song, Y. Chemical composition, nutritional properties and antioxidant activity of monkey apple (*Anisophyllea laurina* R. Br. ex Sabine). *J. Food Nutr. Res.* **2014**, *2*, 281–287. [CrossRef]
23. Ihsan, M.; Saputro, A.H.; Handayani, W. Flavonoid distribution mapping system of velvet apple leaf based on hyperspectral imaging. In Proceedings of the 2019 International Conference on Advanced Computer Science and information Systems (ICACSIS), Bali, Indonesia, 12–13 October 2019; pp. 157–162.
24. He, J.; Chen, L.; Chu, B.; Zhang, C. Determination of total polysaccharides and total flavonoids in *Chrysanthemum morifolium* using near-infrared hyperspectral imaging and multivariate analysis. *Molecules* **2018**, *23*, 2395. [CrossRef]
25. Cheng, J.H.; Sun, D.W. Rapid quantification analysis and visualization of *Escherichia coli* loads in grass carp fish flesh by hyperspectral imaging method. *Food Bioproc. Technol.* **2015**, *8*, 951–959. [CrossRef]
26. Choi, J.Y.; Kim, J.; Kim, J.; Jeong, S.; Kim, M.; Park, S.; Moon, K.D. Hyperspectral imaging technique for monitoring moisture content of blueberry during the drying process. *Korean J. Food Preserv.* **2021**, *28*, 445–455. [CrossRef]
27. Bai, X.; Xiao, Q.; Zhou, L.; Tang, Y.; He, Y. Detection of sulfite dioxide residue on the surface of fresh-cut potato slices using near-infrared hyperspectral imaging system and portable near-infrared spectrometer. *Molecules* **2020**, *25*, 1651. [CrossRef] [PubMed]
28. Liu, Y.; Sun, Y.; Xie, A.; Yu, H.; Yin, Y.; Li, X.; Duan, X. Potential of hyperspectral imaging for rapid prediction of anthocyanin content of purple-fleshed sweet potato slices during drying process. *Food Anal. Methods* **2017**, *10*, 3836–3846. [CrossRef]
29. Caporaso, N.; Whitworth, M.B.; Fowler, M.S.; Fisk, I.D. Hyperspectral imaging for non-destructive prediction of fermentation index, polyphenol content and antioxidant activity in single cocoa beans. *Food Chem.* **2018**, *258*, 343–351. [CrossRef]
30. Lichtenthaler, H.K.; Buschmann, C. Chlorophylls and carotenoids: Measurement and characterization by UV-VIS spectroscopy. *Curr. Protoc. Food Anal. Chem.* **2001**, *1*, F4.3.1–F4.3.8. [CrossRef]
31. Thomas, M.; Badr, A.; Desjardins, Y.; Gosselin, A.; Angers, P. Characterization of industrial broccoli discards (*Brassica oleracea* var. italica) for their glucosinolate, polyphenol and flavonoid contents using UPLC MS/MS and spectrophotometric methods. *Food Chem.* **2018**, *245*, 1204–1211. [CrossRef]
32. Mawlong, I.; Sujith Kumar, M.; Gurung, B.; Singh, K.; Singh, D. A simple spectrophotometric method for estimating total glucosinolates in mustard de-oiled cake. *Inter. J. Food Prop.* **2017**, *20*, 3274–3281. [CrossRef]
33. Yang, Y.C.; Sun, D.W.; Pu, H.; Wang, N.N.; Zhu, Z. Rapid detection of anthocyanin content in lychee pericarp during storage using hyperspectral imaging coupled with model fusion. *Postharvest Biol. Technol.* **2015**, *103*, 55–65. [CrossRef]
34. Jeong, J.C.; Kim, S.J.; Hong, S.Y.; Nam, J.H.; Sohn, H.B.; Kim, Y.H.; Mekapogu, M. Growing environment influence the anthocyanin content in purple-and red-fleshed potatoes during tuber development. *Korean J. Crop. Sci.* **2015**, *60*, 231–238. [CrossRef]
35. Ninu, L.; Ahmad, M.; Miarelli, C.; Cashmore, A.R.; Giuliano, G. Cryptochrome 1 controls tomato development in response to blue light. *Plant J.* **1999**, *18*, 551–556. [CrossRef]
36. Giliberto, L.; Perrotta, G.; Pallara, P.; Weller, J.L.; Fraser, P.D.; Bramley, P.M.; Fiore, A.; Tavazza, M.; Giuliano, G. Manipulation of the blue light photoreceptor cryptochrome 2 in tomato affects vegetative development, flowering time, and fruit antioxidant content. *Plant Physiol.* **2005**, *137*, 199–208. [CrossRef] [PubMed]
37. Kim, H.S.; Yoo, J.H.; Park, S.H.; Kim, J.S.; Chung, Y.; Kim, J.H.; Kim, H.S. Measurement of environmentally influenced variations in anthocyanin accumulations in *Brassica rapa* subsp. Chinensis (Bok Choy) using hyperspectral imaging. *Front. Plant Sci.* **2021**, *12*, 693854. [CrossRef]

agriculture

MDPI

Article

Compact Convolutional Transformer (CCT)-Based Approach for Whitefly Attack Detection in Cotton Crops

Aqeel Iftikhar Jajja [1], Assad Abbas [1], Hasan Ali Khattak [2,*], Gniewko Niedbała [3,*], Abbas Khalid [4], Hafiz Tayyab Rauf [5] and Sebastian Kujawa [3]

[1] Department of Computer Science, COMSATS University Islamabad, Islamabad 45500, Pakistan
[2] School of Electrical Engineering & Computer Science (SEECS), National University of Sciences & Technology (NUST), H12, Islamabad 44000, Pakistan
[3] Department of Biosystems Engineering, Faculty of Environmental and Mechanical Engineering, Poznań University of Life Sciences, Wojska Polskiego 50, 60-627 Poznań, Poland
[4] Department of Computer Science and IT, The University of Lahore, Lahore 54590, Pakistan
[5] Centre for Smart Systems, AI and Cybersecurity, Staffordshire University, Stoke-on-Trent ST18 0YB, UK
* Correspondence: hasan.alikhattak@seecs.edu.pk (H.A.K.); gniewko.niedbala@up.poznan.pl (G.N.)

Abstract: Cotton is one of the world's most economically significant agricultural products; however, it is susceptible to numerous pest and virus attacks during the growing season. Pests (whitefly) can significantly affect a cotton crop, but timely disease detection can help pest control. Deep learning models are best suited for plant disease classification. However, data scarcity remains a critical bottleneck for rapidly growing computer vision applications. Several deep learning models have demonstrated remarkable results in disease classification. However, these models have been trained on small datasets that are not reliable due to model generalization issues. In this study, we first developed a dataset on whitefly attacked leaves containing 5135 images that are divided into two main classes, namely, (i) healthy and (ii) unhealthy. Subsequently, we proposed a Compact Convolutional Transformer (CCT)-based approach to classify the image dataset. Experimental results demonstrate the proposed CCT-based approach's effectiveness compared to the state-of-the-art approaches. Our proposed model achieved an accuracy of 97.2%, whereas Mobile Net, ResNet152v2, and VGG-16 achieved accuracies of 95%, 92%, and 90%, respectively.

Keywords: computer vision; CCT; cotton pest attack; whitefly attack; deep learning; precision agriculture

Citation: Jajja, A.I.; Abbas, A.; Khattak, H.A.; Niedbała, G.; Khalid, A.; Rauf, H.T.; Kujawa, S. Compact Convolutional Transformer (CCT)-Based Approach for Whitefly Attack Detection in Cotton Crops. *Agriculture* **2022**, *12*, 1529. https://doi.org/10.3390/agriculture12101529

Academic Editor: Wei Ji

Received: 11 August 2022
Accepted: 20 September 2022
Published: 23 September 2022

1. Introduction

Agriculture yield in recent years has declined dramatically, posing a threat to global food security. With the world's population predicted to approach 9.7 billion by 2050, there is a strong exigency to boost productivity through use of new technology. Pakistan is ranked sixth in the world in terms of cotton production [1]. Cotton accounts for around 0.6% of Pakistan's GDP. Its output has steadily fallen in recent years, falling by up to 22% [2], and a further fall at this rate will eventually have a negative impact on productivity. Weather, humidity, insect attacks, numerous viruses, and poor pesticides are all elements which impair cotton production.

The whitefly (Bemisia tabaci) is considered as one of the conventional pests that attack the plant and is a carrier of various viruses, for example, cotton leaf curl disease (CLCUD). The whitefly is one of the world's 100 worst invasive alien species [3]. Whiteflies can infect a cotton field and adjacent crops, restricting plant development by up to 50% [4]. As a consequence, this minuscule insect causes harm to both global food and domestic cash crops.

With technological progress, field inspection is uniformly upgrading to more automated sources, such as state-of-the-art automated disease detection algorithms and spatial

drones for detection and prediction. Since beforehand there were no suitable mechanisms to identify plant diseases, one had to manually inspect each plant and prescribe appropriate treatment, which was exceedingly laborious, time-consuming, and required more excellent professional knowledge and skills.

Due to the seriousness and implications of whiteflies for plants, more effective, efficient, and sustainable strategies are needed. Researchers have used machine learning, image processing, and computer vision techniques to develop several solutions to these problems [5,6]. The applicability of image processing to precisely identifying and categorizing disease has arisen in numerous domains.

Recently, several works have employed images of multi-class disease-based private datasets from online dataset repositories. Image acquisition is a critical challenge, as it is hard to acquire photographs from a real-world context. However, Legaspi et al. [7] proposed a framework for detecting and classifying whiteflies using the YOLO-V3 pre-trained model. Initially, 400 images were gathered manually from the fields. Tulshan et al. [8] suggested a multi-class disease classification method based on k-nearest neighbors (KNN) and compared its performance with the support vector machine (SVM). The dataset used for experiments was small, exhibiting various diseases, including mosaic virus, leaf miner, whitefly, mildew, and early blight. Using the SVM methodology, Rajan et al. [9] suggested a strategy for early pest identification. The SVM was trained and evaluated using a sample dataset of 100 images, using slack variable and the threshold value.

The scarcity of adequate cotton datasets is a hindrance to the aggrandizing of deep learning approaches in pest and disease detection. To eradicate this problem, we collected images from a real environment. A total of 5135 image samples were utilized for deep learning model training. The AgriPK dataset (AgriPK Dataset https://doi.org/10.34740/KAGGLE/DSV/2927481) is the largest among the available datasets. Moreover, it is the first dataset for whitefly pest-affected plants to best of our knowledge.

Deep learning models are regarded as pivotal models for classification and detection. Attention-based feature extraction layers applied to models enable them to concentrate on the region of interest (ROI) optimally. We used the latest version of the visual transformer, the Compact Convolutional Transformer (CCT), for the medium-sized dataset. Convolutional layers are utilized in the input layer of the proposed model to build feature maps. The experimental results demonstrate the effectiveness of the developed dataset and the proposed CCT-based approach.

The main contributions of the paper are presented below:

1. A dataset for whitefly attacks in cotton crops comprising 5135 images was developed and published to help future researchers.
2. A multi-class dataset with ground truth annotation was prepared for our model.
3. A Compact Convolutional Transformer (CCT)-based approach is proposed for the classification.
4. The performance of the proposed CCT-based approach is compared with those of various state-of-the-art models, such as MobileNet, ResNet152v2, VGG-16, and SVM. Experimental results demonstrate that the CCT approach outperformed the compared approaches.

This paper is organized as follows: Section 2 presents the related work, whereas Section 3 describes the dataset used in this work. The proposed CCT-based approach is presented in Section 4, whereas experimental results are discussed in Section 5. Section 6 concludes the paper and highlights the directions for future work.

2. Related Work

Precision agriculture highly depends on image datasets when dealing with computer vision applications. The modern applications work effectively for classification and detection, and the state-of-art models have already demonstrated remarkable accuracy with the benchmark datasets. Manual inspection is not only time-consuming and laborious,

but also, late inspection can lead to yield losses. Several deep learning-based techniques have been proposed for leaf disease identification in recent years.

Sujatha et al. [10] compared various machine learning algorithms (random forest (RF), support vector machine (SVM), stochastic gradient descent (SGD)) with deep learning algorithms (VGG-19, VGG-16, Inception-v3) in cotton disease classification. Results showed that VGG-16 outdistanced all other models with an accuracy of 89.5%. However, the number of input samples was very low. Azath et al. [11] proposed a mechanism to detect cotton disease and pest control. The authors used a deep learning technique, predominantly a CNN, for segmentation and classification. The model used a dataset of 2400 images divided into four classes of leaf, namely, minor, spider mite, healthy leaf, and bacterial blight. The model achieved an accuracy of 98% on an apportioned imagery dataset.

Caldeira et al. [12] stated that cotton leaf lesion is a common disease that can affect plant growth. The authors gathered a dataset of 60,000 images mainly based on leaf lesions, and the dataset was further divided into two classes, such as healthy and lesioned leaves. This paper mostly focused on comparing deep learning models with conventional models to validate and substantiate their performances. The depicted results showed that all models achieved accuracy over 70%, and SVM's was above 80%. Using a radial basis function neural network (RBFN) algorithm, Saleem et al. [13] suggested an IoT-based smart system for whitefly detection. The proposed mechanism was tested in the field using IoT sensors. The devices collected 12,896 images, which were connected to web servers.

Pechuho et al. [14] used the YOLO-V3 pre-trained deep learning model. The author used a multi-class dataset of cotton disease from ImageNet's open repository. The model achieved an accuracy of 90% following multiple experiments.

Rothe et al. [15] used a back propagation model to classify multi-class cotton leaves disease. The imagery dataset was collected from different sources physically. The acquired dataset was implemented on Gaussian filters after removing noise using preprocessing techniques.

The above-mentioned studies provide a succinct overview of cotton plant diseases. However, there are no detailed solutions for infectious plant diseases. The datasets employed for the studies are either private or have small sample sizes. Furthermore, because the images are taken by farmers, the image sample data are susceptible to background noise. The deep learning models implemented in the studies performed well on the small datasets. However, they fared badly on large datasets.

The application of state-of-art machine learning algorithms has drawn the attention of researchers. Mojjada et al. [16] worked on multi-class classification using five different types of wheat diseases from the plant village dataset. The authors employed K-means for dataset segmentation using region-based and threshold value techniques.

Neural networks are the baseline architectures for deep learning models. The CNN is a significant classification model with a higher learning rate and low parameters. It learns the spatial features of input images. Furthermore, several studies utilized self-supervised algorithms such as autoencoders [17] to compress data into low dimensions. The convolutional autoencoder employs N layers to learn those low-dimensional data using convolutional layers. Therefore, Bedi [18] proposed a hybrid model of CAE and CNN for plant disease detection. They used a dataset consisting of 4415 images: 3342 images were used for training and 1115 for testing. The computer-aided engineering (CAE) algorithm was employed to reduce the dimensionality of the images, and CNN was used for image classification. The training and testing accuracy were claimed to be 99.35% and 98.38%, respectively.

Chowdhury et al. [19] implemented different segmentation models for the tomato leaf. The plant village database was used for image acquisition. The unbalanced dataset consisted of 18,161 images divided into binary classes, namely, healthy and unhealthy, which were further divided into ten categories. Of the ten classes, one class was healthy; all other classes were labeled as diseased. Efficient Net-B7 outperformed all other models, gaining accuracy of over 99% in binary classification, and Efficient Net-B4 demonstrated accuracy of 99.89%.

Singh [20] collected images of diseased sunflower leaves manually by using digital cameras and other image capturing tools. Particle swarm optimization (PSO) was used for detection and classification. The model provided accuracy of 98%. Bernardes et al. [21] suggested that pathogens are disruptive in plantations and can affect crops' substantially.

The authors developed a framework based on two distinct datasets that were integrated to form a single dataset. The image samples were transformed into HSV and grayscale. SVM was utilized for RGB image classification. The model achieved an accuracy of 96%.

The indiscernible region of interest (RoI) may affect the performances of deep learning models. Even with excellent accuracy, dealing with noisy data may result in the underfitting and overfitting of the model. Plant diseases indicate distinct patterns and dots in the early stages, which require a more robust model with an attention mechanism to detect the disease. However, in the related studies, no attention mechanism was used for detection. We applied an attention-based mechanism for the classification of whitefly attack classification.

Naeem et al. [22] used multiple classification models on a plant-leaf dataset. The labeled classes were then passed to five different models to dissect the accuracy. Multi-layer perceptron (MLP) gained an accuracy of 95% by utilizing 1200 samples.

Zhang et al. [23] implemented a SIFT algorithm to detect corn ear tests and sequence them according to their appearance. The model achieved a maximum accuracy of 97%. Islam et al. [24] implemented different deep learning models for multi-class papaya-leaf classification. The CNN outperformed all other state-of-the-art algorithms, achieving an accuracy of 98.04%. The validation and training loss was 0.79%, which is very low because any CNN requires a large dataset to train.

Arsenovic et al. [25] articulated that the lack of appropriate datasets is a major obstacle in implementing deep learning and computer vision models in the agriculture field. The author proposed a dataset of fourteen different diseases containing over 70,000 images to overcome this issue. Moreover, a hybrid model called Plant Disease Net (PDN) was proposed for classification, which achieved an accuracy of 93%.

Ngugi et al. [26] developed an application for automatic disease detection on the limited number of datasets gathered manually. The MobileNet-V2 model was implemented along with kijani Net, which achieved an accuracy of 90%. Mobilenet-v2 is a pre-trained model used to address classification and detection tasks. Mobilenet-v2 is widely used in agriculture disease detection, and it has shown exquisite results.

3. Materials

Due to the unavailability of a dataset of reasonable size, in this research, we developed a dataset called the AgriPK dataset containing images of the cotton crops. The details of the dataset are presented in the sub-section below.

3.1. AgriPK Dataset

The sample images of the leaves utilized for the proposed research were collected from Bahawalpur, the Southern city of the Punjab province in Pakistan, between August and October 2021.

The South Punjab region is deemed to be the best cotton-growing region. As a result, the whitefly infestation is intense and prominent. We collected 5135 images from Bahawalpur farms for this study. Our dataset is categorized into five classes, as shown in Table 1. The gleaned dataset is unbiased, not very noisy, and unambiguous. On-field images are vulnerable to natural phenomena such as multiple leaves appearing simultaneously, air resistance, and bright sunlight. We established a controlled environment to mitigate environmental factors to eliminate the challenges above. The different classes are shown in Figure 1.

Table 1. Dataset used for this study.

Our AgriPK Dataset				
No. of Classes	**Categories**	**No. of Images**	**Test/Train**	**Total No. of Images**
1	Healthy	2213	1600/713	5137
2	Unhealthy	2852	2110/741	
3	Mild	210	152/58	
4	Nutrition Deficiency	235	160/75	
5	Severe	2407	1801/675	
Cotton Diseased Dataset				
1	Diseased Cotton Leaves	288	235/53	1951
2	Diseased Cotton Plant	815	602/203	
3	Fresh Cotton Leaves	427	324/104	
4	Fresh Cotton Plant	421	321/101	

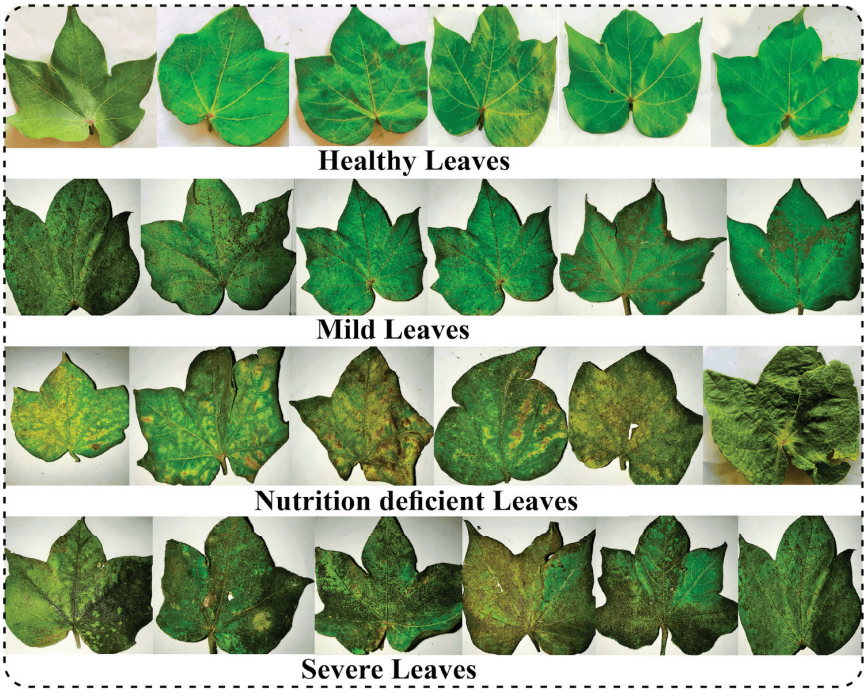


Figure 1. Our proposed AgriPK dataset.

3.1.1. Image Collection

Image capture and labeling is a time-consuming, laborious, and expensive procedure, as it requires significant resources. Multiple devices, including the mobile phones and DSLR cameras, were utilized to acquire the sample images, which ranged in resolution from 5 to 12 megapixels. To ensure the effectiveness of the process, agriculture professionals were involved, who supervised the image gathering and classification processes. After examining numerous cotton fields, around 8000 images were initially captured. To reduce noise, all images were captured in a special environment with artificial light, manually adjusting angles and capturing them at the same distance.

3.1.2. Professional Annotation

After collecting the image data, we annotated them so that we could effectively perform the task of classification. As professional knowledge is required for manual data annotation, the samples were labeled by agriculture professionals from the Islamia University of Bahawalpur, Pakistan. The dataset given to the annotator was first classified into five categories: (a) healthy, (b) unhealthy, (c) nutritional deficit, (d) mild, and (e) severe, as illustrated in Figure 1. Moreover, after experts verification, we placed each image in the corresponding folders. To train the dataset, we distributed the samples into five classes and annotated each sample with a class label. Each image data sample is represented with a label vector of five values.

3.2. Existing Cotton Disease Dataset

There are various public datasets of cotton diseases available. We used the Cotton Disease Dataset [27] to compare our study to the existing dataset. The Cotton Disease Dataset Figure 2 is a free resource found in the Kaggle repository. The four classes in the dataset are: (i) diseased plants, (ii) diseased leaves, (iii) healthy plants, and (iv) healthy leaves. The dataset includes 1918 samples. Real-time cotton photos were captured in the field during the daytime. The Cotton Disease Dataset is one of the few cotton datasets that are currently available online. We contrasted our dataset with a dataset of infected cotton.

Figure 2. Cotton Disease Dataset [27].

4. Methods

This study presents a classification approach for whitefly attacks on cotton plants. We introduced the AgriPK dataset and implemented a Compact Convolutional Transformer (CCT) for classification. The workflow diagram of the presented research is shown in Figure 3.

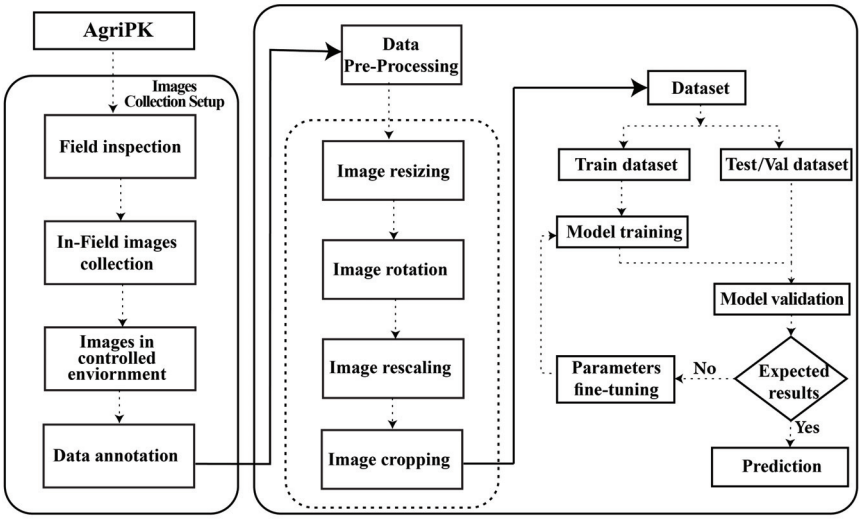


Figure 3. Work flow diagram of proposed scheme.

4.1. Data Pre-Processing Module

Data pre-processing is the preliminary procedure for improving and adjusting input image quality, size, and flipping. For data cleaning, many pre-processing techniques, such as rotation, scaling, cropping, and resizing, were used which were imported from the Python framework. Many re-scaling techniques are mostly used, but data normalization, also known as Min-Max scaling, is a preferred technique. Noise removal is an important aspect of deep learning models where model accuracy is highly dependable on the image background. While dealing with more acute and sensitive diseases and finding the region of interest (RoI), background noise can affect model performance.

Image rotation is required to fix the input images to the same length and angle to enable the model to smoothly classify the required output. We used data augmentation techniques, such as image translation, scaling, and image rotation, to balance the dataset during the experiment. Image scaling was used to resize an image, whereas image rotation was used to rotate an image to a specific degree from its central axis. Several other techniques, such as image cropping, horizontal rotation, adjusting angles, removing noise, and changing the image dimension and input size, were also applied as shown in Figure 4.

4.2. Classification Model

Compact Convolutional Transformer

In this study, a Compact Convolutional Transformer (CCT) was used for whitefly classification. The CCT [28] is the most recent version of the Vision Transformer (Vit), which is a compact model for image processing problems. The conventional transformer is considered data hungry for image processing tasks. Many authors proposed different techniques to address this issue, such as DeiT [29], ConViT [30], CvT [31], and T2T-ViT [32]. However, the CCT model outperformed all existing state-of-art techniques. The model was trained and evaluated on three types of datasets: small-scale and low-resolution images (FashionMNIST, MNIST, and CIFAR-10/100), medium-sized (Image Net), and small-scale high-resolution images (Flowers-102). It utilized few parameters and thus minimized the time complexity of the model. The model architecture as shown in Figure 4 includes various novel frameworks that distinguish it from prior approaches. The model architecture is described in the following paragraphs.

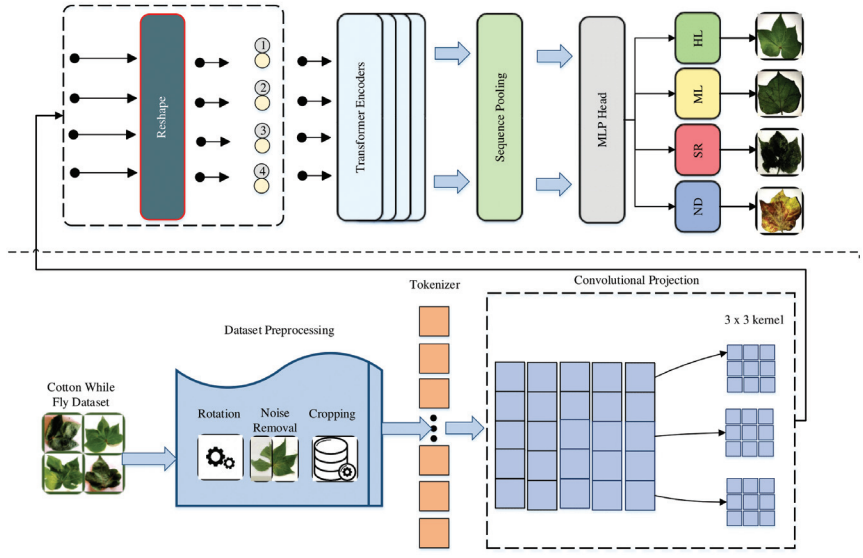


Figure 4. CCT system model proposed in this work.

First, the augmented RGB images are fed to an input layer with dimensions of 150 × 150 × 3. A dual tokenizer with a kernel size of 3 × 3 convolutional layers and a stride size of 1 is applied to preserve the boundary-level information. The resultant feature map attained from convolutional layer helps to alleviate self-attention computation complexity. The rectified linear activation (Relu) is used as an activation function in conjunction with the "he normal" kernel-initializer with padding size 1. The mathematical representation of convolutional layers employed in the CCT tokenizer is given as follows:

$$Xo = MaxPool(ReLU(conv2d(x)) \tag{1}$$

where MaxPool in Equation (5) [28] represents the convolutional layer. The feature map

$$x \epsilon R^{H \times W \times C}, \tag{2}$$

as shown in Equation (2), represents the extraction of local features. C is the number of channels, and H represents the resolution of images. Due to convolutional blocks, the model is not dependent on image resolution, as it creates a feature map and preserves locally partial information. Then, an attention-based encoder–decoder mechanism is applied on a high-dimensional feature map area to extract the spatial and intensity-based pertinent features in the multi-class cotton dataset. All the features are flattened into a 1D array and fed to the sequence embedding layer. It locates information within the sequence regarding the relative locations of image patches. After passing through the self-attention module, this class embedding predicts each class of an input image.

$$MultiHead(Q, K, V) = [Head_1 ..., Head_h]W_o \tag{3}$$

where queries, keys, and values are represented by Q, K, and V, respectively, as shown in Equation (6) [29]. After positional embedding, we applied multi-layer perceptron (MLP) stacking on cotton image patches with their ground truths for whitefly attack classification. The deep learning models correspond to various hyperparameter tuning techniques during pre-processing and model training. We kept an input size of 224, as shown in Table 2; the input size directly affects the model performance. The input image is the sample image on

which the model will predict. Therefore, the algorithm required a fixed size to learn the temporal and spatial features. Similarly, the batch size is also a hyperparameter where a total number of sample images is processed prior to model updates. The batch size can be increased or decreased depending on the dataset. For our experiments, we used a batch size of 64 with 100 epochs. Usually, training a neural network would require a series of epochs where training data will pass through every single epoch.

Table 2. Parameters used during training.

Parameter	Value
Learning rate	0.06
Batch size	64
Input size	224
No. of epochs	100
Weight decay	0.006

The accuracy, precision, recall, and F1-score were among the evaluation measures used to assess the suggested method's performance. Precision is the ratio of correct positive results to all positive instances predicted by the classifier. The ratio of true-positive predictions over the sum of true-positive and false-negative predictions is used to calculate recall. The following are the mathematical representations of these measures.

$$Accuracy = \frac{No.\,of\ correct\ instances}{total\ no.\,of\ input\ samples} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1\text{-}score = \frac{2(PrecisionRecall)}{Precision + Recall} \tag{7}$$

Positive and negative class samples are represented by P and N, respectively. The rightly identified positive class is referred to as true positive (TP). Similarly, true negative (TN) depicts the identification of an abnormal class. FN and FP, on the other hand, stand for the misclassified normal and abnormal classes.

5. Results and Discussion

To demonstrate the effectiveness of CCT-based approach on the AgriPK dataset, extensive experimentation was performed. During the experiments, the images were collected in the natural environment and labeled under the supervision of experts. Google Colab and Kaggle were used to train the deep learning and CCT models using Python 3.0. All the deep learning models were implemented using Keras, TensorFlow, and scikit learn libraries, which are open-source libraries. We compared the proposed CCT-based approach with several state-of-the-art deep learning models, namely, MobileNet, VGG16, and Resnet 152 V2. The brief overview of the compared models is given below.

5.1. MOBILE NET

MobileNet [33] is reliable and efficient when applied to real-world applications. It comes up with many frameworks. The standard convolutional layer is replaced by depth-wise separable convolution to make it a lighter model. MobileNet-v2 is based on 53 hidden layers with a ReLU activation function. These are already pre-trained on millions of images from various imagery repositories.

5.2. VGG-16

VGG-16 [34] is a vision model, which is being utilized in multiple disease classification and detection scenarios. The model architecture is based on a complex 16-layer structure. The activation layer is present after the block of the input convolutional layer. The model randomly updates weights after each training layer to minimize error.

5.3. RESNET 152-V2

Resnet152-V2 [35] is the state-of-the-art model used for image classification. It consists of multiple hidden layers. The input data pass through the reshaped layer and then to the flattened layer. Moreover, the dense layer in the model consists of 128 neurons. To avoid model overfitting, a drop-out layer is added. Finally, a SoftMax function for image classification is used for prediction.

5.4. Yolo V5

Yolo V5 [36] is based on the conventional architecture of the Yolo series, which was put forward by ultralytics. It is mainly based on three steps known as the model backbone, neck, and head for one-step object detection. To mitigate the time consumption problems and to attribute the duplicate features, a CSP Net (cross-stage partial networks) is employed. The input image fed to the model is processed by the model backbone, where the important feature is extracted using CSP Nets. PANet (path aggregation network) [37] is utilized for image scaling and to formulate feature pyramids in the model neck module. However, the model head of Yolo v5 is similar to those of Yolo v3 and Yolo v4. The model head primarily focuses on anchoring boxes and predicting final outputs with bounding boxes. The model follows regression approaches for detection inducing fewer parameters.

5.5. Performance on AgriPK

To demonstrate the efficacy of our proposed AgriPK dataset we experimented several state-of-the-art deep learning models and the latest version of the Vision Transformer, namely, the CCT. We conducted several experiments on the small and medium-sized datasets to evaluate their performances. The results show the effectiveness of proposed model on small and medium-sized datasets as shown in Table 3. A total of 5135 images were used for training and testing, as shown in Figure 5. We used parameter tuning mechanisms to test the models' performances, as shown in Figure 6. The efficacy of our proposed dataset was evaluated based on the true positive and false negative rates as shown in Figure 7.

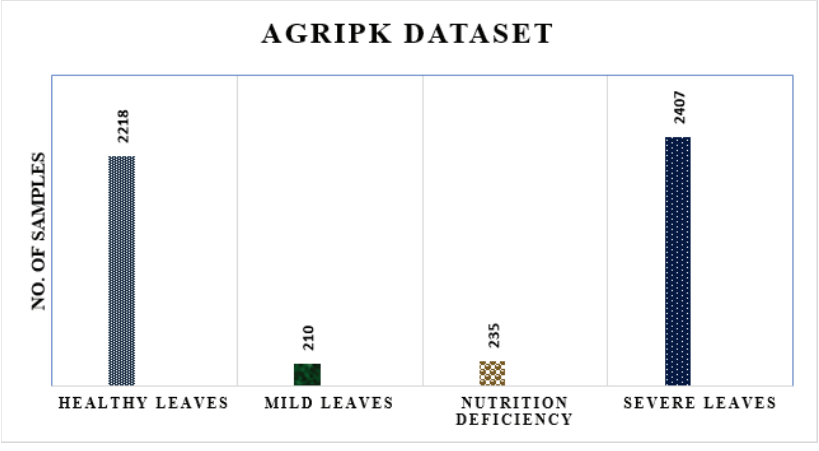
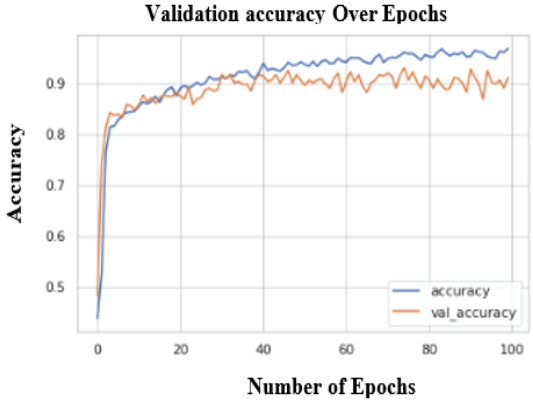
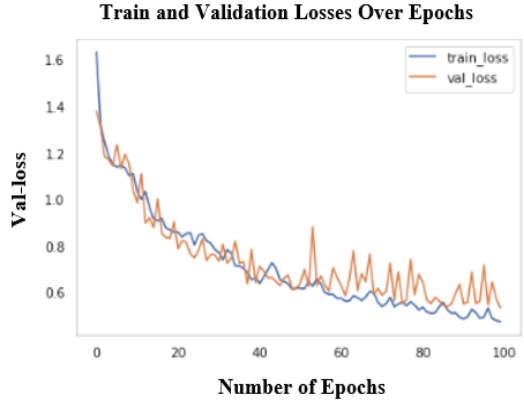


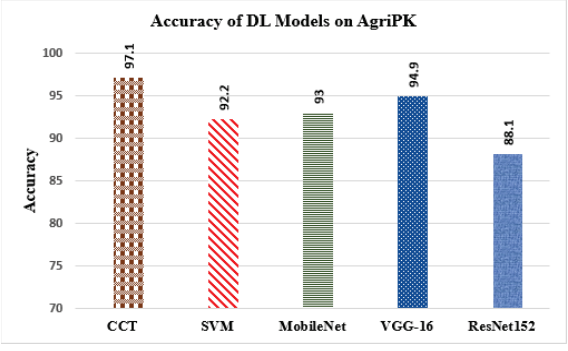
Figure 5. AgriPK sample distribution.



(**A**)



(**B**)



(**C**)

Figure 6. (**A**) Validation accuracy over epochs. (**B**) Training and validation loss epochs. (**C**) Accuracy of all DL models.

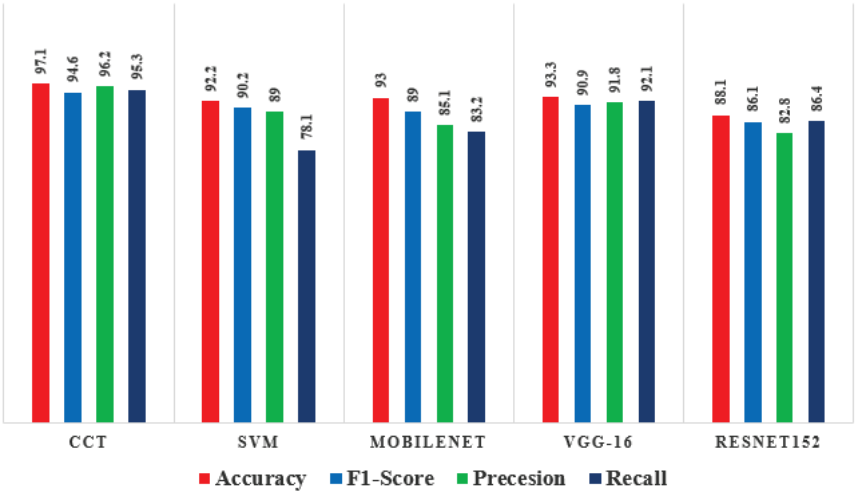


Figure 7. Benchmark models' performances on AgriPK as shown in Table 3.

Table 3. Benchmark models' performances on AgriPK.

Parameters	Mobile Net	VGG-16	ResNet-152	YoloV5	SVM	CCT
Accuracy (%)	93	93.3	88.1	95.1	92.2	97.1
F1-Score	89	90.9	86.1	93.2	90.2	94.6
Precision	85.1	91.8	82.8	91.1	89	96.2
Recall	83.2	92.1	86.4	85.7	78.1	95.3

We utilized five classes in the first experiments, namely, (i) healthy, (ii) unhealthy, (iii) severe, (iv) mild, and (v) nutrition deficit. The final experiment used a 120 × 120 × 3 input image size for 100 epochs. A learning rate of 0.001 and a batch size of 64 were utilized for the experiment. To improve the model's accuracy, the image size was lowered to 120, leading to an accuracy of 97%, as shown in Figure 7. We initiated the trials with an image size of 550 × 550 × 3. In the first stage, the model was explicit 88 percent of the time. Moreover, the batch size was lowered to 32 to comprehend the varied CCT parameters better as given in Figure 8.

The CCT demonstrated the highest accuracy over 100 epochs, with the least validation and training losses. When compared to MobileNet and VGG-16, the CCT took more time. During our trails, we noticed that CCT's time consumption was slightly higher than those of Mobilenet and VGG-16. This is because CCT is based on complex transformer architectures, as shown in Table 4.

Table 4. Comparison of computational time and number of parameters.

Model	No. of Params.	Training Time
CCT	897,413	7200 ms
VGG-16	750,567	6600 ms
MobileNet	699,156	6300 ms
Resnet 152	950,567	8900 ms

The Resnet model achieved the lowest accuracy of 88% among all the models. The SVM model achieved 92% accuracy, as shown in Table 5. Resnet has a complex structures; therefore, it required more training time as compared to other models. On the other hand, VGG-16 achieved around 95% accuracy. However, the validation loss as compared to the CCT model was greater than 0.9% for VGG-16. We learned that varying the training epochs

also effects the performance of classification modelas shown in Figure 8. Mobilenet and SVM showed accuracies of 93% and 92.2%, respectively.A real-time detection model, Yolov5, was employed on the proposed AgriPK dataset to evaluate its efficacy. The model showed an accuracy of 95.1%. It can be noted that the CCT resulted in balanced recall and precision, whereas the Yolo v5 lagged when applied to similar domains. Reasons for the tangible decreases in recall and precision were the indistinguishable circles and constant colors in the input images. Although all state-of-art model depicted good accuracy, for whitefly diseased leaf classification, CCT showed the best classification accuracy.

Table 5. Performance evaluation of AgriPK and Cotton Disease Dataset.

Model	Cotton Disease Dataset	AgriPK Dataset	F1-Score Increment
CCT	91.8	94.6	2.8%
SVM	80.2	90.2	9.6%
Mobile Net	80.4	89	8.6%
VGG-16	89.8	90.6	0.8%
ResNet152-v2	80.7	86.1	5.7%
Yolo V5	92.6	93.2	0.5%

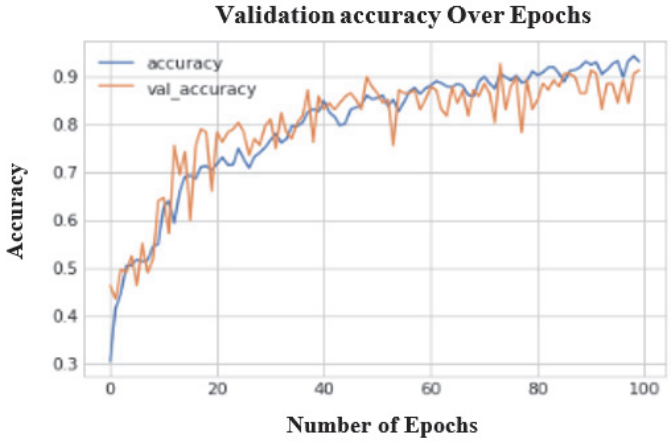


Figure 8. Training loss.

5.6. Performance on Cotton Disease Dataset

We also evaluated the accuracy of our proposed CCT model on the existing Cotton Disease Dataset. The CCT model showed an accuracy of 95% on the Cotton Disease Dataset. As shown in Figure 8, the simulations were conducted over 100 epochs to test the model. However, we kept the corresponding parameters consistent during the comparison session. Initially, we conducted experiments over 20 epochs for CCT, and the accuracy was 85%. Subsequently, after adjusting the epochs to 50, we attained the best results.

The highest accuracies attained by using the CCT were 97.1% and 95.4% on the AgriPK dataset and Cotton Disease Dataset, respectively, as demonstrated in Figure 9. The reason CCT performed well on both datasets is that it is highly adaptive and has an effective learning rate, ascribed to the presence of MLP layers and convolutional blocks in the model architecture. We evaluated the performances of CCT and deep learning models using various evaluation metrics. We noticed that sufficient trainable data are required for complex models for classification and detection.

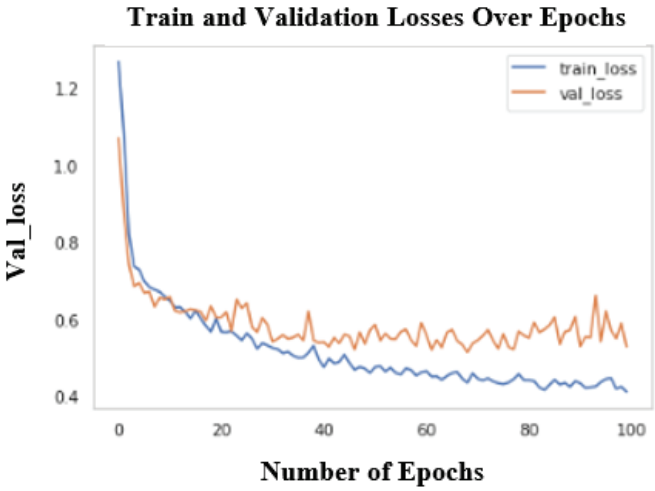


Figure 9. Validation accuracy.

We utilized F1-score, precision, recall, and accuracy as key metrics for model evaluation, as stated earlier in Table 3. To train and test the models, the data were split into 25% and 75%, respectively (X and Y). Throughout the simulations, we kept the training and testing settings consistent. The SVM and Mobile Net performed poorly on both parts of the Cotton Disease Dataset, whereas the F1-score on our AgriPK dataset was substantially higher as shown in Figure 10. Yolov5 showed a slight increase in the F1-score in both datasets. The model utilizes a real-time detection mechanism. However, the baseline datasets are prone to having similar features; therefore, model performance deteriorated when new data were given to the model. The top F1-score of 94.6% of CCT on both datasets demonstrated its effectiveness. We employed transfer learning and pre-trained layers, with customized layers replacing the last 15% of training data as shown in Figure 11.

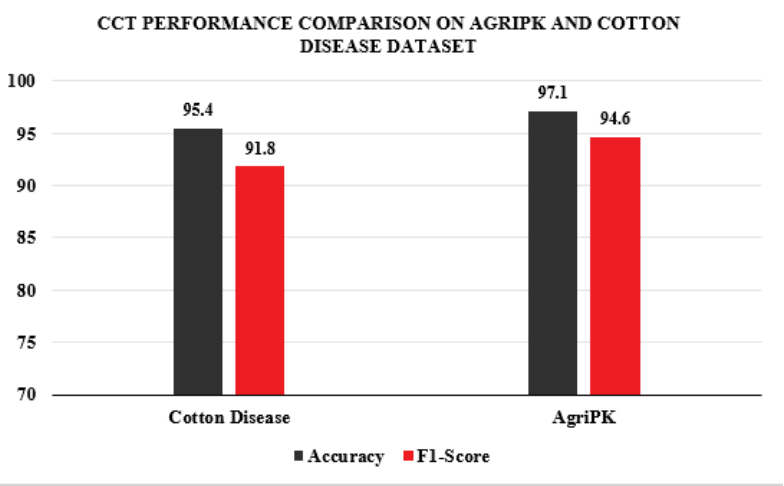


Figure 10. Accuracy and F1-score.

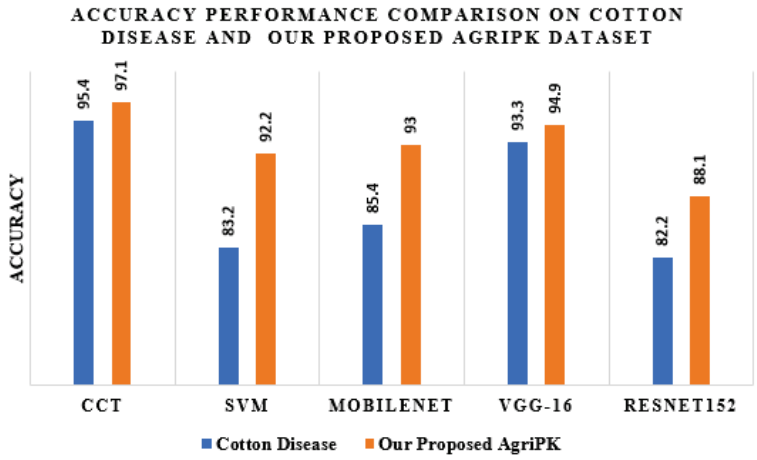


Figure 11. Performance comparison.

6. Conclusions and Future Work

The agriculture domain lacks enough datasets, and there are not enough imagery samples present online to help future research. As a result, there is a need to generate new datasets for the future of AI in agriculture. To that end, we presented AgriPK, a dataset based on cotton leaves damaged by whiteflies. The dataset was created in supervised controlled conditions to eliminate noise and irrelevant components from the image. The dataset contains 5137 images and is publicly available. Furthermore, we used a Compact Convolutional Transformer on the developed dataset to ascertain its generalizability. Despite the model's intricacy, it showed strong performance when compared to other deep learning models. However, other state-of-the-art models also demonstrated substantially better accuracy.

In future research, we will focus on enhancing our ApriPK dataset's samples. Furthermore, the proposed model can be utilized on other large public datasets for performance evaluation. In the future, we will work on different cotton diseases and will work to develop a hybrid model for effective pest classification and detection.

Author Contributions: Conceptualization, A.I.J., A.A., H.A.K. and A.K.; methodology, A.I.J., A.A., H.A.K., G.N., A.K., H.T.R. and S.K.; software, A.I.J., A.A. and H.A.K.; visualization, A.I.J., A.A. and H.A.K.; writing—original draft, A.I.J., A.A., H.A.K., G.N., A.K., H.T.R. and S.K.; data curation, A.I.J., A.A., H.A.K., G.N., A.K., H.T.R. and S.K.; supervision, A.A., H.A.K. and A.K.; writing—review and editing, A.I.J., A.A., H.A.K., G.N., A.K., H.T.R. and S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The dataset is available at Kaggle https://doi.org/10.34740/KAGGLE/DSV/2927481 (accessed on 10 August 2022), and the code is available upon request.

Acknowledgments: The AgriPK dataset was collected and labeled with the help of Naveed Iftikhar and M. Irfan Akram , Entomology Department Islamia University Bahawalpur, Pakistan.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shuli, F.; Jarwar, A.H.; Wang, X.; Wang, L.; Ma, Q. Overview of the cotton in Pakistan and its future prospects. *Pak. J. Agric. Res.* **2018**, *31*, 291–418. [CrossRef]
2. Ali, A.; Ahmed, Z. Revival of Cotton Pest Management Strategies in Pakistan. *Outlooks Pest Manag.* **2021**, *32*, 144–148. [CrossRef]

3. Poorter, M.d.; Browne, M. The Global Invasive Species Database (GISD) and international information exchange: Using global expertise to help in the fight against invasive alien species. In *Plant Protection and Plant Health in Europe: Introduction and Spread of Invasive Species, Held at Humboldt University, Berlin, Germany, 9–11 June 2005*; British Crop Protection Council: Alton, UK, 2005; pp. 49–54.
4. Zia, K.; Hafeez, F.; Bashir, M.H.; Khan, B.S.; Khan, R.R.; Khan, H.A.A. Severity of cotton whitefly (*Bemisia tabaci* Genn.) population with special reference to abiotic factors. *Pak. J. Agric. Sci.* **2013**, *50*, 217–222.
5. Hara, P.; Piekutowska, M.; Niedbała, G. Selection of independent variables for crop yield prediction using artificial neural network models with remote sensing data. *Land* **2021**, *10*, 609. [CrossRef]
6. Sedri, M.H.; Niedbała, G.; Roohi, E.; Niazian, M.; Szulc, P.; Rahmani, H.A.; Feiziasl, V. Comparative Analysis of Plant Growth-Promoting Rhizobacteria (PGPR) and Chemical Fertilizers on Quantitative and Qualitative Characteristics of Rainfed Wheat. *Agronomy* **2022**, *12*, 1524. [CrossRef]
7. Legaspi, K.R.B.; Sison, N.W.S.; Villaverde, J.F. Detection and Classification of Whiteflies and Fruit Flies Using YOLO. In Proceedings of the 2021 13th International Conference on Computer and Automation Engineering (ICCAE), Melbourne, Australia, 20–22 March 2021; pp. 1–4.
8. Tulshan, A.S.; Raul, N. Plant leaf disease detection using machine learning. In Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 6–8 July 2019; pp. 1–6.
9. Nesarajan, D.; Kunalan, L.; Logeswaran, M.; Kasthuriarachchi, S.; Lungalage, D. Coconut disease prediction system using image processing and deep learning techniques. In Proceedings of the 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS), Genova, Italy, 9–11 December 2020; pp. 212–217.
10. Sujatha, R.; Chatterjee, J.M.; Jhanjhi, N.; Brohi, S.N. Performance of deep learning vs machine learning in plant leaf disease detection. *Microprocess. Microsyst.* **2021**, *80*, 103615. [CrossRef]
11. Azath, M.; Zekiwos, M.; Bruck, A. Deep learning-based image processing for cotton leaf disease and pest diagnosis. *J. Electr. Comput. Eng.* **2021**, *2021*.
12. Caldeira, R.F.; Santiago, W.E.; Teruel, B. Identification of cotton leaf lesions using deep learning techniques. *Sensors* **2021**, *21*, 3169. [CrossRef] [PubMed]
13. Saleem, R.M.; Kazmi, R.; Bajwa, I.S.; Ashraf, A.; Ramzan, S.; Anwar, W. IOT-Based Cotton Whitefly Prediction Using Deep Learning. *Sci. Program.* **2021**, *2021*, 8824601. [CrossRef]
14. Pechuho, N.; Khan, Q.; Kalwar, S. Cotton Crop Disease Detection using Machine Learning via Tensorflow. *Pak. J. Eng. Technol.* **2020**, *3*, 126–130.
15. Rothe, P.; Kshirsagar, R. Cotton leaf disease identification using pattern recognition techniques. In Proceedings of the 2015 International Conference on Pervasive Computing (ICPC), Pune, India, 8–10 January 2015; pp. 1–6.
16. Mojjada, R.K.; Kumar, K.K.; Yadav, A.; Prasad, B.S.V. Detection of plant leaf disease using digital image processing. *Mater. Today Proc.* **2020**. [CrossRef]
17. Bisong, E. Autoencoders. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*; Springer: Cham, Switzerland, 2019; pp. 475–482.
18. Bedi, P.; Gole, P. Plant disease detection using hybrid model based on convolutional autoencoder and convolutional neural network. *Artif. Intell. Agric.* **2021**, *5*, 90–101. [CrossRef]
19. Chowdhury, M.E.; Rahman, T.; Khandakar, A.; Ayari, M.A.; Khan, A.U.; Khan, M.S.; Al-Emadi, N.; Reaz, M.B.I.; Islam, M.T.; Ali, S.H.M. Automatic and reliable leaf disease detection using deep learning techniques. *AgriEngineering* **2021**, *3*, 294–312. [CrossRef]
20. Singh, V. Sunflower leaf diseases detection using image segmentation based on particle swarm optimization. *Artif. Intell. Agric.* **2019**, *3*, 62–68. [CrossRef]
21. Bernardes, A.A.; Rogeri, J.G.; Oliveira, R.B.; Marranghello, N.; Pereira, A.S.; Araujo, A.F.; Tavares, J.M.R. Identification of foliar diseases in cotton crop. In *Topics in Medical Image Processing and Computational Vision*; Springer: Cham, Switzerland, 2013; pp. 67–85.
22. Naeem, S.; Ali, A.; Chesneau, C.; Tahir, M.H.; Jamal, F.; Sherwani, R.A.K.; Ul Hassan, M. The classification of medicinal plant leaves based on multispectral and texture feature using machine learning approach. *Agronomy* **2021**, *11*, 263. [CrossRef]
23. Zhang, X.; Liu, J.; Song, H. Corn ear test using SIFT-based panoramic photography and machine vision technology. *Artif. Intell. Agric.* **2020**, *4*, 162–171. [CrossRef]
24. Islam, M.A.; Islam, M.S.; Hossen, M.S.; Emon, M.U.; Keya, M.S.; Habib, A. Machine learning based image classification of papaya disease recognition. In Proceedings of the 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 5–7 November 2020; pp. 1353–1360.
25. Arsenovic, M.; Karanovic, M.; Sladojevic, S.; Anderla, A.; Stefanovic, D. Solving current limitations of deep learning based approaches for plant disease detection. *Symmetry* **2019**, *11*, 939. [CrossRef]
26. Ngugi, L.C.; Abelwahab, M.; Abo-Zahhad, M. Recent advances in image processing techniques for automated leaf pest and disease recognition—A review. *Inf. Process. Agric.* **2021**, *8*, 27–51. [CrossRef]
27. D3v. Cotton Disease Dataset, Version 1. 2020. Available online: https://www.kaggle.com/datasets/janmejaybhoi/cotton-disease-dataset (accessed on 6 January 2022).

28. Hassani, A.; Walton, S.; Shah, N.; Abuduweili, A.; Li, J.; Shi, H. Escaping the big data paradigm with compact transformers. *arXiv* **2021**, arXiv:2104.05704.
29. d'Ascoli, S.; Touvron, H.; Leavitt, M.L.; Morcos, A.S.; Biroli, G.; Sagun, L. Convit: Improving vision transformers with soft convolutional inductive biases. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 2286–2296.
30. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 10347–10357.
31. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 22–31.
32. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 558–567.
33. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
36. ultralytics/yolov5: v6.2. YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai Integrations. Available online: https://github.com/ultralytics/yolov5/releases (accessed on 10 August 2022).
37. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.

agriculture

MDPI

Communication

Fuzzy Quality Certification of Wheat

Cristian Silviu Simionescu [1], Ciprian Petrisor Plenovici [1], Constanta Laura Augustin [1,*], Maria Magdalena Turek Rahoveanu [1], Adrian Turek Rahoveanu [2] and Gheorghe Adrian Zugravu [1]

[1] Faculty of Engineering and Agronomy in Braila, "Dunarea de Jos" University of Galati, 800008 Galati, Romania
[2] Faculty of Management and Rural Development, University of Agronomic Sciences and Veterinary Medicine Bucharest, 011464 Bucharest, Romania
* Correspondence: laura.zugravu@ugal.ro; Tel.: +40-722-676-202

Abstract: This paper presents a fuzzy quality certification of wheat. This analysis is based on the fuzzy analysis model of wheat. We developed a Matlab application with the help of which we modeled the perceptions in relation to the main quality physical and chemical characteristics of wheat obtaining a quality index of wheat lots. The algorithm presented in this article allows for obtaining and using the global quality index, generating applicability not only to the commercial sphere as a quality reference and price setting, but also a measure of appreciation of processing opportunities. Indices of fuzzy quality associated with wheat lots using a fuzzy model offer the opportunity to develop local markets through quality certification.

Keywords: wheat quality; fuzzy quality certification model

Citation: Simionescu, C.S.; Plenovici, C.P.; Augustin, C.L.; Rahoveanu, M.M.T.; Rahoveanu, A.T.; Zugravu, G.A. Fuzzy Quality Certification of Wheat. *Agriculture* **2022**, *12*, 1640. https://doi.org/10.3390/agriculture12101640

Academic Editors: Gniewko Niedbała and Sebastian Kujawa

Received: 14 September 2022
Accepted: 4 October 2022
Published: 8 October 2022

1. Introduction

The main objective of this paper is to develop an informational model for the standardization of wheat quality in international contracts and obtain a global quality index of the lots analyzed. Decisions about the quality of wheat lots involve imprecise and vague information [1,2]. The development of the information model is based on the assessment by experts of the results obtained from the analysis of 25 batches of wheat, through the reference methods, and the transposition of these results into a mathematical modeling system, called a fuzzy system. By using the fuzzy system, the quality of the analyzed lots will be translated into a global quality index related to each lot. Fuzzy systems are oriented towards managing uncertain or imprecise information [3,4]. The fuzzy system is used in fields where the input variables do not have fixed values or their value and importance may vary [5,6]. The fuzzy concept was first introduced by Zadeh in 1965, to alleviate uncertainties and fuzziness problems. The technique relies on human subjectivity in decision-making due to linguistic variables that allow precise modeling of imprecise entrances [7,8] and has been applied to many engineering problems [9,10]. This concept offers a number of advantages for users, such as reducing the costs of production, transport, storage, and recovery as well as improving the costs of the companies using it [11,12]. Fuzzy is used to represent process uncertainty and simulation of final product quality determination [13–15]. Fuzzy logic is a broad field of study and various tools have been developed in recent years. Food quality is a fuzzy category that could be evaluated using fuzzy logic [16–18]. Its implementation in food quality control for the food industry has been highlighted by several authors who have focused on different applications designed specifically for this field [19,20]. This is especially true when considering the reasoning process, expressed in linguistic terms, of operators and experts [21,22]. However, applications are still limited and few reviews are available on the subject [23–25]. By using fuzzy logic, obtaining the global wheat quality index can be determined by going through the following stages in the development of the informational model:

- Formation of the knowledge base;
- Fuzzy inference;
- Fuzzification;
- Defuzzification.

The adopted approach regarding the evaluation of wheat quality is based on the expertise of 20 specialists in the field, with competences in the determination and evaluation of wheat quality as well as in the field of grain trading. The allocation of qualifications by experts regarding the quality of the lots analyzed, as well as the establishment of the weight of the importance of the analyzed parameters, was carried out based on international and European standards, but also on the quality specifications used in commercial contracts, together with the professional experience of each specialist. Mapping quality attributes into a fuzzy domain as multidimensional fuzzy sets results in a quality index associated with the entire lot [26–28]. In quality control, specialists may face uncertain and unclear concepts. By using the fuzzy concept and developing the information model, the quality of the wheat is qualified according to the strictness of the decision factors and the values of the obtained quality parameters. Natural activities and human thought form the basis for fuzzy logic that presents itself based on various application perspectives [29,30]. At present, food safety incidents have occurred frequently in China and customer trust has declined rapidly;therefore, food quality and safety issues have attracted more and more social attention [31,32]. Considering the concern about ensuring food quality and improving consumer confidence, many companies have developed a traceability system based on the fuzzy concept to visualize the supply chain and avoid food safety incidents [33–35]. Fuzzification processes are implemented in the fields of health care, education, career selection, real estate, and financial markets [36–38].The parameters used for systems analysis are the input factors, the type of membership function used for fuzzification, defuzzification of the generated fuzzy sets [39–41]. Based on the results generated by this system, information is generated that derives recommendations for the selection and optimization of processes. Agriculture is considered as a system that provides products of value and indeterminate returns. It is essential to choose an appropriate technique to maximize yield and minimize losses in the trade chain.

2. Methodology

In this paper, we proposed a model, which can not only perform a quality assessment at all control points, but also assess the quality of wheat that is the subject of an international contract. The quantities that are the subject of commercial transactions usually come from several farmers and are stored before delivery in several storage areas (platforms, warehouses, and silo cells). The use of lot mapping based on the resulting quality index is an advantage for traders and storekeepers, generating a clear and objective overview.

Wheat quality is a complex and widely used term to describe the ability and general potential of wheat to be used in a wide variety of finished products by milling and obtaining quality flours for the production of bread and bakery productsand pastry, semolina, as well as the use in various processes in the extractive, fermentation industries, or in the animal husbandry industry.

The main determinants of wheat quality are endosperm texture (grain hardness), protein content and gluten concentration.Endosperm texture in wheat is the single most important and defining quality characteristic, as it facilitates wheat classification and influences milling, baking, and end-use quality [42].

For millers, wheat quality is considered to be the ability of a wheat variety to produce high quantities and qualities of flour or semolina during the extraction process.In this process, the level of contamination of flour or semolina with bran fractions is also important and is related in most cases to undesirable characteristics for the end-use quality of the productas well as grain hardness [43].

Millers prefer large, uniform, whole, unpolished, and full grains.These physical characteristics, along with chemical and rheological properties, are objectives for wheat growers to increase yield quality and production [44].

Sampling represents the operation that consists of taking and constituting a sample in order to determine the quality by analyzing the monitored parameters.The sample must be as representative as possible for the sampled lot.Sampling is carried out both for grain in motion and for batches of stationary grain or in packaged units (bags).

In the case of bulk, stationary grains, elementary samples are taken with the manual cylindrical probe from different points. From the point of view of the regulations regarding the approach to the sampling process, ISO launched the specific standard for sampling, a document in circulation under the name ISO 24333:2010—Cereals and cereal products. Sampling has beentaken up by most National Standardization Bodies. Depending on the size of the lot, the standard provides the mass of the elementary sample, the minimum number of elementary samples, the minimum mass of the laboratory sample, and limits the maximum size of the lot to 1500 tons [45].

Depending on the analyzes requested, the mass of the laboratory sample may be higher, taking into account that for the determination of aflatoxin and ochratoxin a quantity of 10 kg is needed. The determination of other contaminants such as heavy metals, pesticides, DON (deoxynivalenol), or dioxins can also be identified in samples of at least one kilogram, and for fumonisins and zearalenone a quantity of 3 kg is required. The mass of the laboratory sample is determined according to the required determinations of contaminants to which is added the minimum mass provided by the standard.

The evaluation of the organoleptic and sanitary characteristics is established already in the pre-harvest phase, to avoid possible contamination of the installations or storage spaces. Organoleptic characteristics consist of appearance, color, smell, and taste.Determining the organoleptic characteristics of wheat grains are the examination through the sensory organs of qualified personnel.

International standard ISO 7971—Common wheat specifies and establishes the organoleptic conditions of the wheat grains and stipulates that the grain mass subject to evaluation must be free-flowing, without foreign smell and taste that would indicate a change in the product mass (moldiness or burning), with a normal appearance and a characteristic color [46].

The appearance is determined by visual analysis of the laboratory sample spread in a uniform layer on a white tray to allow the observation of deviations from the specific appearance. The evaluation is based on the shape of the seeds, if the grains are well developed, mature, and healthy or if they are shriveled, burnt, sprouted, altered, attacked by insects or diseases, etc. The uniformity of the grains and the appearance of the skin are monitored.

Maturity means reaching the complete and stable physiological stage. The normal appearance of the bean is considered when the covering of the bean has not undergone changes due to adverse weather conditions, improper storage conditions or attack by insects or other pests. Determining the color consists of assessing it in natural light, observing any changes compared to the characteristic color of the product. The change in the color of the berries can be influenced by excessive humidity, heat, spoilage, mold, drying or improper storage, and contact with chemical substances. Odor determination can be performedboth for whole grains and ground seeds. Determining the smell of whole grains is performedby heating and rubbing in the palms of about 100 g of seeds and inhaling immediately. Another method consists of putting 10–20g of wheat in a glass of warm water with a temperature of 60 °C, which is covered and left to rest for 2–3 min, after which the resulting vapors are inhaled, and then the water is removed from the glass and examine the smell of the remaining grains. Following the procedure used, it is assessed if the smell of the sample is characteristic, in accordance with the specification in the product standard or if it presents certain changes that may come from an inadequate storage without ventilation, from a heating of the product mass or from mold colonies. It is also possible to identify

the smell of putrefaction, of decay, of rancidity, of fermentation, honey (in case of mite infestation), musty, of foreign substances such as phosphine, fuels or sulfur, the smell of strongly aromatic plants if they are present seeds of these plants in the mass of the product or other foreign smells. Appreciating the taste is performed by chewing a few grains of wheat, preferably ground, after removing impurities and spoiled grains. The analysis aims at the specificity of the taste and if it corresponds to the specifications of the quality standards or, on the contrary, is it bitter, sour, hot, or rancid. This determination is not performed on altered, moldy grains, on those that show traces of entomological attack, or on those treated for the purpose of seeding or to combat pests. It is important to know the origin of the batches in order to avoid ingesting chemicals from fertilizers or other agrochemical products and also to identify weed seeds that may contain toxic alkaloids (e.g., ricin) [47].

Experts' Assessment of the Quality of Wheat Lots

Using fuzzy logic to describe abstract concepts and design decision-making systems much closer to the way a human does is an interesting and useful area to explore. To effectively implement these types of systems, expert knowledge of the domain in which the application is being used is required [48–50]. For each linguistic term that a linguistic variable implies, a fuzzy set described by a relevance function will be created. The semantic properties of the (linguistic) concept are described by the outline of the respective fuzzy set [51,52]. Therefore, the closer the behavior of the phenomenon under study is to the curve of the relevance function, the more accurate or performing the fuzzy model is in representing the real world [53–55]. In the database formation stage, the physic-chemical quality indicators obtained from laboratory determinations by analyzing the samples of the 25 batches of wheat were based on the evaluations of 20 experts using a scale with 5 linguistic terms, associated with the qualifiers:

- N = unsatisfactory;
- S = satisfactory;
- M = medium;
- B = good;
- FB = very good.

The terms N, S, M, B, and FB represent the linguistic variables to which the values of the analyzed parameters are associated. Each expert evaluated the 25 batches of wheat and assigned a value of 1 on the N-S-M-B-FB qualification scale to each analyzed parameter once, the assessment being expressed in accordance with the international standards and specifications used in international trade, as well as with their own expertise. Based on relevance, the system sets each value in the fuzzy set to a value between 0 and 1, a measure that represents the degree of relevance of the fuzzy set element.

The second stage in the development of the informational model consisted in defining a scale of three qualifications, respectively little important (PI), important (I), and very important (FI), which was made available to the experts in order to assign a qualification to each parameter that was determined in order to evaluate the quality of the wheat. Based on the standards, the international specifications used in the international wheat trade, as well as their own expertise from professional activity, the experts assigned a qualifier to the quality indicators.

By using the Matlab R2020 program and the Fuzzy Logic Designer function, the association of triplets was achieved following the evaluation of the quality of wheat lots by experts, as well as for the weights established regarding the importance of quality parameters in the evaluation of a wheat lot. Fuzzy triplets were associated with the linguistic terms used to assess the quality of the wheat batches with the help of left-right triangular membership functions, as follows:

- Unsatisfactory, (N) = [0 0 25];
- Satisfactory, (S) = [25 25 25];
- Medium, (M) = [50 25 25];

- Good, (B) = [75 25 25];
- Very good, (FB) = [100 25 0].

Linguistic terms used to determine the weights of the analysis indicators in the value of the global quality index were associated with fuzzy triplets with the help of left-right triangular membership functions as follows:

- Slightly important, (PI) = [0 0 50];
- Important, (I) = [50 50 50];
- Very important, (FI) = [100 50 0];

3. Results

The relative weight of the physical–chemical analyses is logged in calculation of the global wheat quality index. Additionally, it was determined using the function of the application based on the following relationship [56]:

$$\mathrm{Ft} = \mathrm{F} \times [\mathrm{PI};\ \mathrm{I};\ \mathrm{FI}]/20; \tag{1}$$

The fuzzy triplets associated based on the relative weight of the physic-chemical analyzes were transposed into the matrix calculation function of the fuzzy application and subjected to modeling as follows [56,57]:

$$\mathrm{Qt} = \mathrm{sum}(\mathrm{Ft}(:,1)); \tag{2}$$

$$\mathrm{Ftrel} = \mathrm{Ft}/\mathrm{Qt}; \tag{3}$$

Ftrel—represents the weight matrix of each quality indicator in the calculation of the global quality index.

The calculation of the global quality index of the analyzed wheat lots includes the fuzzification and defuzzification phase. With the help of the relative weights in the form of fuzzy triplets, the global quality index was calculated for each wheat lot, using the extended pe.mat product. The mathematical model used is [58]:

$$\mathrm{Clti} = \Sigma\, \mathrm{lti} \otimes \mathrm{Ftreli},\ \text{where i} = 1:30. \tag{4}$$

The mathematical model uses the extended product that was introduced in the form of the pe.mat function [58]:

$$\begin{gathered} \%\text{the extended product} \\ \text{function C} = \mathrm{pe}(\mathrm{A},\ \mathrm{B}) \\ \mathrm{C}(1) = \mathrm{A}(1) \times \mathrm{B}(1); \\ \mathrm{C}(2) = \mathrm{A}(1) \times \mathrm{B}(2) + \mathrm{B}(1) \times \mathrm{A}(2); \\ \mathrm{C}(3) = \mathrm{A}(1) \times \mathrm{B}(3) + \mathrm{B}(1) \times \mathrm{A}(3); \end{gathered} \tag{5}$$

To calculate the quality of the wheat batch, we used the cg.mat function, which transposes the above mathematical model into the Matlab language [59]:

$$\begin{gathered} \%\text{quality of the wheat batch wheat} \\ \text{function C} = \mathrm{cg}(\mathrm{A},\ \mathrm{B}) \\ \mathrm{C} = [0\ 0\ 0]; \\ \text{for i} = 1:30, \mathrm{C} = \mathrm{C} + \mathrm{pe}(\mathrm{A}(\mathrm{i},:), \mathrm{B}(\mathrm{i},:)); \\ \text{end;} \end{gathered} \tag{6}$$

The complexity of the algorithm is given by the matrix analysis of the 30 physic-chemical indicators of the wheat batches, indicators that are translated into fuzzy triplets. Thus, the ICG vector is obtained, with quality indices for each batch.

Correlation of the global quality index by reporting to the Grading Plan for common wheat in Romania was achieved by ordering the values of the ICG index in descending

order and the association with the assigned grade. In the grading operation, the fraction of grains attacked by black point was eliminated from the value of total impurities, their identification and highlighting had the role of making a complex assessment taking into account all the fractions provided by the standard.

After analyzing the two methods of assessing the quality of some wheat batches, it can be concluded that the minimum values of ICG obtained in the case of the 25 analyzed batches were mainly attributed to the infested batches and whose determining parameters in the grading recorded values below the limits imposed on the RO 1 degree (Figure 1). The criteria that are the basis of the grading operation are the sanitary characteristics, the content of total impurities, the hectoliter mass, and the protein content relative to the dry substance.

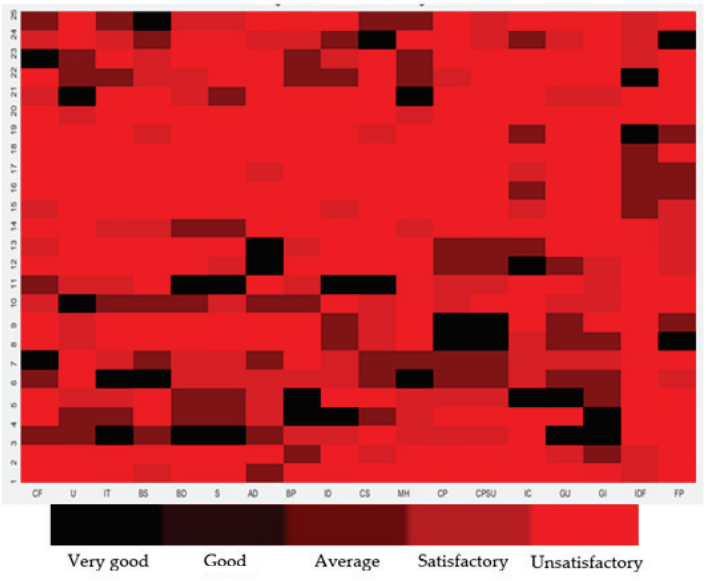


Figure 1. Map of the global quality indices in the analyzed commodity.

The global quality index can also be implemented in credit contracts where wheat stocks are brought as a guarantee, credit institutions having an image not only of the quantity. Researchers can use the global quality index in the study of culture technologies or breeding processes in the commercial chain; the global quality index can highlight the critical points of some batches and facilitate differentiated storage in warehouses, barges, ship holds, silos, wagons or other means of transport. From the point of view of continuing the research and development of the global quality index, the proposed informational model can be extended to other types of agricultural seeds and can be associated with an application that generates an informational model for setting the price. In the study of wheat quality, the global quality index can also be studied by resizing the number of parameters or limiting it to certain indicators of interest.

The weights of the analysis indicators in the value of the global quality index form a neutrosophic statistic, which is a generalization of the classical statistic.What distinguishes neutrosophics from other fields can be indeterminacy, neutrality, even game, unknown, contradiction, ignorance, imprecision, etc. The neutrosophic approach as a generalization of fuzzy logic is a generalization of classical probability and imprecise probability [60–62].

The substantiation of the importance of wheat from an economic and cultural point of view, as well as the quality conditions stipulated in the specific contracts used in the international wheat trade, highlighted the different perception on the evaluation of wheat quality. The concept of wheat quality in large wheat-producing countries is differently

perceived and adapted to the socio-cultural life of the analyzed regions. The appreciation and importance of quality parameters emphasizes the geographical character, the natural and economic resources available to each state, these aspects differentiate and limit the quality of certain wheat lots through national norms related to the international standards and regulations in force. In the stage proceeding the generation of the information model, applied research was carried out regarding the determination of the main quality indicators in order to identify the critical points and the aspects related to an objective, correct and representative evaluation of the analyzed wheat lots.

Methods used, as well as the results obtained, were described in relation to European and international product specifications. Throughout the work, the conditions and limitations that can classify wheat batches as unfit for human consumption, as well as aspects related to food safety through potential contamination with toxins or impurities difficult to eliminate in the conditioning, handling processes, were taken into account or transportation. All the lots analyzed can be used in baking, the resulting flour being suitable for making bread, some lots requiring mixing with quantities of a higher quality or adding additives to capitalize the flour in pastry products or other bakery specialties. The qualitative analysis of the lots reveals the need for quick intervention on the lots where live infestation was identified, by taking the necessary measures, namely their gassing or fumigation. Improving the quality of the analyzed batches can be achieved through additional conditioning, thus reducing the fractions of the category of total impurities and improving the hectoliter mass parameter. Changing the hectoliter mass involves increasing the content of extracted flour and can positively influence the parameters aimed at the rheological behavior of the dough, but also the content and quality of gluten.

In recent decades, rapid wheat quality determination methods and devices have become widely adopted for the determination of moisture content, hectoliter mass and protein in general, but also for continuous monitoring of stock quality. The analyzers are based on near-infrared (NIR) transmission technology, which can be used for the simultaneous and precise determination of several parameters, such as moisture, protein content, gluten content, fat content, but also hectoliter weight. From the halogen lamp housing on the back of the instrument, light is guided through an optical fiber into the monochromator inside the instruments. The monochromator provides monochromatic light in the spectrum from 850 nm to 1050 nm, and by means of an optical fiber, the light is guided to the collimator lens system, which is placed above the sample cup in the sample cup chamber. After the light is transmitted through the sample, the unabsorbed light reaches the detector. The detector measures the amount of light and sends the result to the digital signal processor which communicates with the computer, calculating the result. Rotating the sample cup between sample scans (called sub-samples) allows more parts of the sample to be analyzed. Sub-samples are chosen from one or two concentric circles in the sample cup, providing a more representative result from an inhomogeneous sample. Devices can be calibrated by using certified reference materials. Hyperspectral imaging (HSI) combines near-infrared (NIR) spectroscopy and digital imaging to provide information on the chemical properties of wheat grains. In order to establish and identify the wheat grains whose germination has started, studies were carried out using hyperspectral near infrared (NIR) imaging (HSI) for their detection. Experiments were conducted to determine which spectral bands have the best potential to discriminate between sound and sprouted grains. Two wavelengths were selected and combined into an index that was used to indicate the presence or absence of germination. Experiments have shown that the proposed method is effective in identifying the grains for which the germination process has started, achieving 100% accuracy for the samples used in this study. An imperfect correlation with the fall index was also observed, making it difficult to accurately determine the degree of germination, especially if the sprouts are not yet visible. These results confirm the utility of the near-infrared spectral range for detecting chemical alterations in wheat grains [63]. Protein content is one of the most important quality factors in wheat and can be determined using this technique. To solve the recognition and classification problems associated with

impurities in wheat, the researchers developed a recognition method that uses a convolution neural network. The development of this network consisted in the construction of a data set of wheat without impurities and of five impurities, with which the filtering algorithm and the enhancement algorithm were used for image pre-processing. Based on research, the testing accuracy was between 98.59% and 99.98%, respectively. Consequently, the developed network, named WheNet can be a useful tool in the recognition of impurities in wheat. In addition, this method can be used to detect impurities in other domains [64]. The determination of the total nitrogen content by combustion according to the Dumas method and the calculation of the crude protein content is based on the quantitative digestion by burning the sample at about 900 °C in excess of oxygen [65]. The sample is burned, and the organic elements are oxidized. Combustion gases (O_2, CO_2, H_2O, and N_2) and nitrogen oxides (NOx) are removed, except nitrogen and nitrogen oxides. Carbon dioxide and water are removed by passing the gases through special columns. The nitrogen content is determined by gas chromatography, and the crude protein content is calculated by multiplying the amount of nitrogen measured by the appropriate factor and expressed as a percentage.

A comparative study on the accuracy of protein determination methods, namely the Kjeldhal method, the Dumas method and the NIR technique revealed the precision error rate below 2% for the Kjeldahl method, while the precision error rate for the Dumas method varied in a range of 2–4%. The NIR method proved to be the fastest in determining protein content; however, the error rate varied between 3% and 6%. The Kjeldahl method, due to its high precision and very small ranges of variation, has made it the major method for estimating protein in food. The Dumas method for the quantitative determination of organic nitrogen was at least as accurate as the Kjeldahl method, but considerably faster. The NIR method has a relatively large standard deviation and is particularly useful for rapid analysis of protein content [64]. Both the Kjeldahl and Dumas methods for protein determination in foods are currently used, but the empirical nitrogen factors used to convert determined nitrogen content to protein content are based only on the Kjeldahl method [65].

The main objective of the research described in this article, represents an innovative method of approaching the quality of wheat and can be a landmark in the calculation of penalties and bonuses within international commercial contracts. Going through the stages described in this research has generated a global wheat quality index that can be extended for use in several commercial, governmental, or scientific segments. The research transposed in this article combined the international standards that regulate the reference methods for wheat quality determinations frequently used in international contracts, as well as national and contractual specifications in terms of quality determination. In addition to the specialized literature, which mostly includes studies on changes and behavior of wheat in different phases of culture, storage or processing, personal experience in the field of quality and the appreciation of experts on the weight of the important quality parameters and the evaluation of quality based on the results obtained, have led to the configuration of the proposed informational model. Obtaining and using the global quality index generates applicability not only to the commercial sphere as a quality reference and price setting, but also a measure of appreciation of processing opportunities. The global quality index can also be implemented in credit contracts where wheat stocks are brought as a guarantee, credit institutions having an image not only of the quantity. Researchers can use the global quality index in the study of culture technologies or breeding processes in the commercial chain, the global quality index can highlight the critical points of some batches and facilitate differentiated storage in warehouses, barges, ship holds, silos, wagons, or other means of transport. From the point of view of continuing the research and development of the global quality index, the proposed informational model can be extended to other types of agricultural seeds and can be associated with an application that generates an informational model for setting the price. In the study of wheat quality, the global quality index can also be studied by resizing the number of parameters or limiting it to certain indicators of interest.

In all lots, a higher share of seeds belonging to other plants is noted, but none of the subject lots exceed the limits of this category in terms of this sub-parameter. The values of the content of foreign bodies (chaff, dust) are reduced, which indicates a good conditioning before storage or a correct adjustment of the harvester. No toxic seeds, rye horn, or grains attacked by common wheat blight or Fusarium spp. were identified. A remarkable aspect is the lack of burnt-hot berries, which leads to the hypothesis of a moderate dryness after reception.

Sampling was carried out in accordance with the specific standard and during sampling it was possible to assess the mass of the product, not identifying agglomerations in layers or in the extracted elementary samples, the extracted wheat is free flowing. From the point of view of the organoleptic characteristics, all the lots analyzed fell within the specificity of the healthy product in terms of smell, appearance, and taste, and no alterations were identified in the mass of the product. Live infestation was observed in seven batches, the identified species being Sitophilus zeamais, Rhyzoperta domnicasi, and Cryptolestes ferrugineus. The moisture content values are between 11.27% and 13.42%; no batches with moisture above the maximum allowed limit of 14% being identified. The hectoliter mass recorded values below 77 kg/hl. Five lots (16, 17, 18, 19, and 20) have values above 77 kg/hl, reaching a maximum of 80.5 kg/hl. From the point of view of total impurity content, five samples exceed the 6% limit related to grade RO1 wheat, thus ranking the lots in terms of total impurity content in grade RO2. In the sum of total impurities, the presence of broken grains and defective grains can be noted in all analyzed batches. Sprouted grains were identified in 9 of the 25 lots analyzed, and the maximum percentage resulting from the analysis is 0.10%, so that none of the lots presents a risk of damage or a risk of failure for this reason. In all lots, a higher share of seeds belonging to other plants is noted, but none of the subject lots exceed the limits of this category in terms of this sub-parameter. The values of the content of foreign bodies (chaff or dust) are reduced in all analyzed samples, which indicate a good conditioning before storage or a correct adjustment of the harvester. No toxic seeds, rye horn, or grains attacked by common wheat blight or Fusarium spp. were identified. A remarkable aspect is the lack of burnt-hot berries, which leads to the hypothesis of a moderate dryness after reception. The use of the reference method to determine the protein content led to obtaining reliable results regarding the value of this parameter. The determination of the crude protein and then reporting to the percentage of moisture revealed variable percentages, the minimum values being 11.69 in the case of batch 19, respectively, and 11.98% in the case of batch 17. All other batches recorded values above 12%; the maximum value was13.75%. In the optimal range of 22–25% wet gluten content there are only three lots, thus characterizing the related quantities with an average wet gluten content, and in the range of 25–31% there are 22 lots of wheat with a high gluten content, thus placing all lots in higher quality classes in terms of this parameter. The values obtained after determining the gluten index parameter fall into the category of normal gluten with values between 30–80%. In the 25 lots analyzed, the drop index is above the minimum limit of 220 s. The values obtained after establishing the deformation index of this determination place 23 batches in the optimal range between 5 and 13 mm; batch 19 and 23 hadvalues below 5 mm. From the point of view of the rheological properties and the parameters determined to generate an image of these properties, only batches 4, 21, and 22 reach the optimum values, respectively, for W, G, and can be considered batches with excellent baking properties. The determination of the DON content revealed the presence of the mycotoxin in 12 analyzed lots, but the values obtained do not endanger public health, being at most 1/3 of the maximum allowed limit. In the case of 13 batches, the values obtained were below the detection limit of the device. All the results of the performed tests fell within the repeatability limit provided in the method standards used.

4. Conclusions

The yield and efficiency of the wheat crop is quantified both by the quantity harvested and by the quality obtained. The quality of the wheat is determined from the pre-harvest

phase by determining the moisture content to identify maturity and the optimal harvesting period, as well as by identifying possible microbial contamination or other aspects related to the physical structure of the grain. Post-harvest, quality determination is carried out in several sequences prior to processing. The primary evaluation is carried out immediately after harvesting and it is important to know the physico-chemical aspects of the grains to be able to intervene and subject the quantities received or stored to immediate drying and conditioning operations. In establishing the quality of some wheat lots, an important role is played by sampling, which must generate a sample representative of the whole lot or informative, depending on the parameters being pursued. Sampling rules are established by international, national, or trade association standards, all with the major objective of obtaining a representative sample to provide a clear overview of quality. Quality assessment is initially carried out by evaluating the organoleptic and sanitary characteristics that must correspond to the healthy product, present a free-flowing appearance of the seed mass, without agglomerations or modified color, smell, and specific taste. Humidity plays an important role, being the first determination that is made after the organoleptic analysis and determination of infestation, its value being taken as a reference in the calculation of subsequent determinations. From the point of view of physical characteristics, the hectoliter weight is a useful indicator in the milling and baking industry, and at the same time a low value of this parameter can classify the wheat in lower quality classes or can be considered fodder. Studies have shown that from a quantity of wheat with a high hectoliter mass, the amount of flour extracted is greater compared to the amount of flour obtained from wheat with a low hectoliter mass. The chemical analyzes consist of determining the protein content, the quantity and quality of gluten, but also the behavior of the flour during kneading by determining the falling index and alveographic properties. The frequent use of irrigation water from various sources and plant protection products in an uncontrolled and sometimes irrational way, the determination of the content of heavy metals and pesticides has become in recent years an important criterion for determining the quality of wheat. Non-compliance with culture technologies and climatic conditions favors the development of some species of fungi such as Fusarium spp. which, due to their toxic nature, produce secondary metabolites generically known as mycotoxins. Vomitoxin or deoxynivalenol is considered the wheat-associated mycotoxin, with studies showing little presence of aflatoxins, ochratoxin or zearalenone in wheat. The methods for determining quality indicators are varied and are regulated by international standards and regulations, these having a mandatory character in the case of heavy metals, mycotoxins and pesticides. By contributing the experts in evaluating the quality of the wheat lots based on the results of analyzes made available and by ranking the importance of parameters within a global appreciation, the database required for the mathematical modeling system was constituted.

The use of fuzzy logic in the configuration of the informational model was achieved by using the Matlab 2020 program and the fuzzy logic designer function. The Matlab program with fuzzy functions is frequently used and there are numerous scientific articles that are based on statistical data processed by this method. The specialized literature emphasizes the potential of fuzzy applications used to render the uncertainty of the process and simulate determining the quality of a final product. The principle of fuzzy logic is based on the transposition of clear results into an unclear fuzzy system, by associating triplets with value from 0 to 1 in the fuzzy stage and subsequently subjected to defuzzification, resulting in an associated value in the form of an index. The map of quality attributes by fuzzy technique as multidimensional fuzzy sets and later defuzzification resulted in obtaining a global quality index associated with the whole lot (IGC).

The limits of these parameters with an impact on food safety are established both for unprocessed wheat and for products obtained from it and intended for human and animal consumption.

The evolution of technology has allowed the development of methods for rapid determination of wheat quality in order to automate and obtain quick results, but in case of litigation the reference methods provide the most accurate results for settlement and

arbitration. The expansion in time of quality determinations, the interest of standardization organizations regarding the methods and limits of wheat quality parameters, as well as innovation, position wheat in an area of major interest, generating competitiveness in the sphere of quality and its determination as a future concern.

Author Contributions: Conceptualization, C.S.S., C.P.P. and M.M.T.R. investigation, C.P.P. and A.T.R.; methodology, G.A.Z. and C.L.A.; software, G.A.Z. and A.T.R.; validation, C.P.P. and M.M.T.R.; writing—original draft, C.L.A. and C.P.P.; writing—review and editing, C.S.S. and G.A.Z.; supervision, G.A.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ali, G.; Alolaiyan, H.; Pamučar, D.; Asif, M.; Lateef, N. A novel madm framework under q-rung orthopair fuzzy bipolar soft sets. *Mathematics* **2021**, *9*, 2163. [CrossRef]
2. Tchier, F.; Ali, G.; Gulzar, M.; Pamučar, D.; Ghorai, G. A new group decision-making technique under picture fuzzy soft expert information. *Entropy* **2021**, *23*, 1176. [CrossRef]
3. Hussain, H.I.; Slusarczyk, B.; Kamarudin, F.; Thaker, H.M.T.; Szczepańska-Woszczyna, K. An investigation of an adaptive neuro-fuzzy inference system to predict the relationship among energy intensity, globalization, and financial development in major ASEAN economies. *Energies* **2020**, *13*, 850. [CrossRef]
4. Thaker, S.; Nagori, V. Analysis of Fuzzification Process in Fuzzy Expert System. *Procedia Comput. Sci.* **2018**, *132*, 1308–1316. [CrossRef]
5. Chistol, L.T.; Bandini, L.G.; Must, A.; Phillips, S.; Cermak, S.A.; Curtin, C. Sensory Sensitivity and Food Selectivity in Children with Autism Spectrum Disorder. *J. Autism Dev. Disord.* **2018**, *48*, 583–591. [CrossRef]
6. Grimen, H.L.; Diseth, Å. Sensory Processing Sensitivity. *Compr. Psychol.* **2016**. [CrossRef]
7. Sharif, M.K.; Butt, M.S.; Sharif, H.R.; Nasir, M. Sensory Evaluation and Consumer Acceptability. In *Handbook of Food Science and Technology*; CRC Press: Boca Raton, FL, USA, 2017; pp. 361–386.
8. Qiong, O.U. A Brief Introduction to Perception. *Stud. Lit. Lang.* **2017**, *15*, 18–28.
9. Rosyidi, C.N.; Murtisari, R.; Jauhari, W. A concurrent optimization model for suppliers selection, tolerance and component allocation with fuzzy quality loss. *Cogent Eng.* **2016**, *3*, 1222043. [CrossRef]
10. Rosyidi, C.N.; Murtisari, R.; Jauhari, W.A. A concurrent optimization model for supplier selection with fuzzy quality loss. *J. Ind. Eng. Manag.* **2017**, *10*, 98–110. [CrossRef]
11. Zhang, S.; Chen, M.; Zhang, W.; Zhuang, X. Fuzzy optimization model for electric vehicle routing problem with time windows and recharging stations. *Expert Syst. Appl.* **2020**, *145*, 113123. [CrossRef]
12. Wang, X.; Yang, Z. Application of fuzzy optimization model based on entropy weight method in atmospheric quality evaluation: A case study of Zhejiang province, China. *Sustainability* **2019**, *11*, 2143. [CrossRef]
13. Garitta, L.; Langohr, K.; Gómez, G.; Hough, G.; Beeren, C. Sensory cut-off point obtained from survival analysis statistics. *Food Qual. Prefer.* **2015**, *43*, 135–140. [CrossRef]
14. Ciappini, M.; Gatti, M.; Cabreriso, M.; Chaín, P. Modificaciones fisicoquímicas y sensoriales producidas durante las frituras domésticas sobre aceite de girasol refinado y aceite de oliva virgen extra. *Inven. Rev. Investig. Acad.* **2016**, *37*, 153–161.
15. Ciappini, M.C. Polyhenolic profile of floral honeys in correlation with their pollen spectrum. *J. Apic. Res.* **2019**, *58*, 772–779. [CrossRef]
16. Garitta, L.; Langohr, K.; Elizagoyen, E.; Gugole Ottaviano, F.; Gómez, G.; Hough, G. Survival analysis model to estimate sensory shelf life with temperature and illumination as accelerating factors. *Food Qual. Prefer.* **2018**, *68*, 371–376. [CrossRef]
17. Esmerino, E.A.; Paixão, J.A.; Cruz, A.G.; Garitta, L.; Hough, G.; Bolini, H.M.A. Survival analysis: A consumer-friendly method to estimate the optimum sucrose level in probiotic petit suisse. *J. Dairy Sci.* **2015**, *98*, 7544–7551. [CrossRef]
18. Elizagoyen, E.S.; Hough, G.; Garitta, L.; Fiszman, S.; Bravo Vasquez, J.E. Consumer's expectation of changes in fruit based on their sensory properties at purchase. The case of banana (Musa Cavendish) appearance evaluated on two occasions: Purchase and home consumption. *J. Sens. Stud.* **2017**, *32*, e12278. [CrossRef]
19. Patrignani, M.; Ciappini, M.C.; Tananaki, C.; Fagúndez, G.A.; Thrasyvoulou, A.; Lupano, C.E. Correlations of sensory parameters with physicochemical characteristics of Argentinean honeys by multivariate statistical techniques. *Int. J. Food Sci. Technol.* **2018**, *53*, 1176–1184. [CrossRef]

20. Ciappini, M.C.; Calviño, A. A Holistic View to Develop Descriptive Sheets for Argentinean Clover and Eucalyptus Unifloral Honey. *Curr. Nutr. Food Sci.* **2020**, *16*, 919–927. [CrossRef]
21. Crabtree, R.H. NHC ligands versus cyclopentadienyls and phosphines as spectator ligands in organometallic catalysis. *J. Organomet. Chem.* **2005**, *690*, 5451–5457. [CrossRef]
22. Lee, S.Y.; Allgeyer, L.; Neely, E.; Kreger, J. Sensory analysis of fruit and fermented fruit product flavors. In *Handbook of Plant-Based Fermented Food and Beverage Technology*, 2nd ed.; Routledge: Abingdon-on-Thames, UK, 2012.
23. Zeng, S.; Shoaib, M.; Ali, S.; Smarandache, F.; Rashmanlou, H.; Mofidnakhaei, F. Certain properties of single-valued neutrosophic graph with application in food and agriculture organization. *Int. J. Comput. Intell. Syst.* **2021**, *14*, 1516–1540. [CrossRef]
24. Ziv, G.; Watson, E.; Young, D.; Howard, D.C.; Larcom, S.T.; Tanentzap, A.J. The potential impact of Brexit on the energy, water and food nexus in the UK: A fuzzy cognitive mapping approach. *Appl. Energy* **2018**, *210*, 487–498. [CrossRef]
25. Perrot, N.; Ioannou, I.; Allais, I.; Curt, C.; Hossenlopp, J.; Trystram, G. Fuzzy concepts applied to food product quality control: A review. *Fuzzy Sets Syst.* **2006**, *157*, 1145–1154. [CrossRef]
26. Heymann, H. A personal history of sensory science. *Food Cult. Soc.* **2019**, *22*, 203–223. [CrossRef]
27. Stone, H.; Sidel, J. Sensory Science and Consumer Behavior. In *Global Issues in Food Science and Technology*; Elsevier: Amsterdam, The Netherlands, 2009.
28. Civille, G.V.; Carr, B.T. *Sensory Evaluation Techniques*, 5th ed.; CRC Press: Boca Raton, FL, USA, 2016.
29. David, F.I.; Akpado, K.A.; Alumona, T.L. Design and Implementation of a Fuzzy Logic Controller for Power Plant Temperature Monitoring and Control using Fuzzylite. *J. Eng. Res. Rep.* **2021**, *20*, 33–51. [CrossRef]
30. Urbieta Parrazales, R.; Zagaceta Álvarez, M.T.; Aguilar Cruz, K.A.; Palma Orozco, R.; Fernández Muñoz, J.L. Implementation of a fuzzy logic controller for the irrigation of rose cultivation in mexico. *Agriculture* **2021**, *11*, 576. [CrossRef]
31. Kacprzyk, J. Fuzzy dynamic programming: Interpolative reasoning for an efficient derivation of optimal control policies. *Control Cybern.* **2013**, *42*, 63–84.
32. Kacprzyk, J.; Zadeh, L. Foreword. In Proceedings of the 6th IEEE International Conference Intelligent Systems, Sofia, Bulgaria, 6–8 September 2012. [CrossRef]
33. Xia, X.; Qiu, Y.; Hu, L.; Zhou, G. Application of information technology on traceability system for agro-food quality and safety. In *IFIP Advances in Information and Communication Technology*; Springer: New York, NY, USA, 2015; Volume 452. [CrossRef]
34. Zhao, J.; Li, A.; Jin, X.; Pan, L. Technologies in individual animal identification and meat products traceability. *Biotechnol. Biotechnol. Equip.* **2020**, *34*, 48–57. [CrossRef]
35. Wang, J.; Yue, H.; Zhou, Z. An improved traceability system for food quality assurance and evaluation based on fuzzy classification and neural network. *Food Control* **2017**, *79*, 363–370. [CrossRef]
36. Keviczky, L.; Bars, R.; Hetthéssy, J.; Bányász, C. Introduction to MATLAB. In *Advanced Textbooks in Control and Signal Processing*; Springer: Singapore, 2019. [CrossRef]
37. Dorfman, K.D.; Daoutidis, P. MATLAB 'Tutorial'. In *Numerical Methods with Chemical Engineering Applications*; Cambridge University Press: Cambridge, UK, 2018.
38. Kim, P. *MATLAB Deep Learning*; APress: New York, NY, USA, 2017.
39. Bryniarska, A. Mathematical Models of Diagnostic Information Granules Generated by Scaling Intuitionistic Fuzzy Sets. *Appl. Sci.* **2022**, *12*, 2597. [CrossRef]
40. Davvaz, B.; Mukhlash, I.; Soleha, S. Himpunan Fuzzy dan Rough Sets. *Limits J. Math. Appl.* **2021**, *18*, 79–94. [CrossRef]
41. Pérez-Fernández, R.; Alonso, P.; Bustince, H.; Díaz, I.; Jurio, A.; Montes, S. Ordering finitely generated sets and finite interval-valued hesitant fuzzy sets. *Inf. Sci.* **2015**, *325*, 375–392. [CrossRef]
42. Pasha, I.; Anjum, F.M.; Morris, C.F. Grain Hardness: A Major Determinant of Wheat Quality. *Food Sci. Technol. Int.* **2010**, *16*, 511–522. [CrossRef]
43. Edwards, M.A.; Osborne, B.G.; Henry, R.J. Puroindoline genotype, starch granule size distribution and milling quality of wheat. *J. Cereal Sci.* **2010**, *52*, 314–320. [CrossRef]
44. Matsuo, R.R.; Dexter, J.E. Relationship between some durum wheat physical characteristics and semolina milling properties. *Can. J. Plant Sci.* **1980**, *60*, 49–53. [CrossRef]
45. *SR EN ISO 24333:2010*; Cereale si Produse din Cereale. Esantionare. Organismul National de Standardizare: Bucharest, Romania, 2010.
46. *ISOInternational Standard 7970:2021*; Wheat (*Triticum aestivum* L.)—Specification. ISO International Standard: Geneva, Switzerland, 2021.
47. *STAS 6253-80*; Seminţe pentru Consum. Determinarea Caracteristicilor Organoleptice. Organismul National de Standardizare: Bucharest, Romania, 1980.
48. Nawaz, U.; Ali, A.; Raza, U.A.; Shehzadi, K. A Survey: Sentimental Analysis on Product Reviews Using (MLT) Machine Learning Techniques and Approaches. *Int. J. Adv. Trends Comput. Sci. Eng.* **2021**, *10*, 1253–1263. [CrossRef]
49. Venkateswara Rao, K.; Srilatha, D.; Mary Gladence, L. Disease prediction and diagnosis implementing fuzzy neural classifier based on IoT and cloud. *Int. J. Adv. Sci. Technol.* **2019**, *29*, 737–745.
50. Svensson, S.Å. Implementing a Fuzzy Classifier and Improving its Accuracy using Genetic Algorithms. *53rd Annu. Conf. Stat. Comput. Sci. Oper. Res.* **2020**, *3*.

51. Ramos-Calderer, S.; Bellini, E.; Latorre, J.I.; Manzano, M.; Mateu, V. Quantum search for scaled hash function preimages. *Quantum Inf. Process.* **2021**, *20*, 180. [CrossRef]
52. Saez, Y.; Estebanez, C.; Quintana, D.; Isasi, P. Evolutionary hash functions for specific domains. *Appl. Soft Comput. J.* **2019**, *78*, 58–69. [CrossRef]
53. França, L.V.; Bressane, A.; Silva, F.N.; Peche Filho, A.; Medeiros, G.A.; Ribeiro, A.I.; Roveda, J.A.; Roveda, S.R. Modelagem Fuzzy Aplicada à Análise da Paisagem: Uma proposta para o diagnóstico ambiental participativo. *Front. J. Soc. Technol. Environ. Sci.* **2014**, *3*, 124–141. [CrossRef]
54. Pessoa, M.A.R.; de Souza, F.J.; Domingos, P.; de Azevedo, J.P.S. Índice fuzzy de qualidade de água para ambiente lótico—IQAFAL TT—IQA FAL—Fuzzy water quality index for lotic environments. *Eng. Sanit. Ambient* **2020**, *25*, 21–30. [CrossRef]
55. Torfi, F.; Farahani, R.Z.; Rezapour, S. Fuzzy AHP to determine the relative weights of evaluation criteria and Fuzzy TOPSIS to rank the alternatives. *Appl. Soft Comput. J.* **2010**, *10*, 520–528. [CrossRef]
56. Yang, C.C. An evaluation of the FRWMA chart for dependent interval-valued data. *Cluster Comput.* **2019**, *22*, 10325–10332. [CrossRef]
57. Kalaiarasi, K.; Sabina Begum, M.; Sumathi, M. Optimization of unconstrained multi-item (EPQ) model using fuzzy geometric programming with varying fuzzification and defuzzification methods by applying python. *Mater. Today Proc.* **2021**. [CrossRef]
58. Rondeau, L.; Ruelas, R.; Levrat, L.; Lamotte, M. A defuzzification method respecting the fuzzification. *Fuzzy Sets Syst.* **1997**, *86*, 311–320. [CrossRef]
59. Salama, A.A.; Hanafy, I.M.; Elghawalby, H.; Dabash, M.S. Neutrosophic Sets and Systems. *Neutrosophic Sets Syst.* **2016**, *12*.
60. Giridhar, K.S. Evaluation of Supplier in Lean Manufacturing Environment using Neutrosophic Sets and Systems. *Int. J. Res. Appl. Sci. Eng. Technol.* **2018**, *6*. [CrossRef]
61. Barbedo, J.G.A.; Guarienti, E.M.; Tibola, C.S. Detection of sprout damage in wheat kernels using NIR hyperspectral imaging. *Biosyst. Eng.* **2018**, *175*, 124–132. [CrossRef]
62. Shen, Y.; Yin, Y.; Zhao, C.; Li, B.; Wang, J.; Li, G.; Zhang, Z. Image Recognition Method Based on an Improved Convolutional Neural Network to Detect Impurities in Wheat. *IEEE Access* **2019**, *7*, 162206–162218. [CrossRef]
63. Nielsen, S.S. Protein Nitrogen Determination. In *Food Science Text Series*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 131–135.
64. Mihaljev, Ž.A.; Jakšić, S.M.; Prica, N.B.; Ćupić, Ž.N.; Živkov-Baloš, M.M. Comparison of the Kjeldahl method, Dumas method and NIR method for total nitrogen determination in meat and meat products. *J. Agroaliment. Process. Technol.* **2015**, *21*, 365–370.
65. Thompson, M.; Owen, L.; Wilkinson, K.; Wood, R.; Damant, A. A comparison of the Kjeldahl and Dumas methods for the determination of protein in foods, using data from a proficiency testing scheme. *Analyst* **2002**, *127*, 1666–1668. [CrossRef]

agriculture

MDPI

Article

Research on the Optimization of Fresh Agricultural Products Trade Distribution Path Based on Genetic Algorithm

Jun Sun [1], Tianhang Jiang [1], Yufei Song [1,*], Hao Guo [1] and Yushi Zhang [2,*]

[1] School of Management, Dalian Polytechnic University, Dalian 116034, China
[2] School of Marxism, Liaoning Normal University, Dalian 116029, China
* Correspondence: songyufeifei@outlook.com (Y.S.); zhangyushi07@outlook.com (Y.Z.)

Abstract: This study, taking the R fresh agricultural products distribution center (R-FAPDC) as an example, constructs a multi-objective optimization model of a logistics distribution path with time window constraints, and uses a genetic algorithm to optimize the optimal trade distribution path of fresh agricultural products. By combining the genetic algorithm with the actual case to explore, this study aims to solve enterprises' narrow distribution paths and promote the model's application in similar enterprises with similar characteristics. The results reveal that: (1) The trade distribution path scheme optimized by the genetic algorithm can reduce the distribution cost of distribution centers and improve customer satisfaction. (2) The genetic algorithm can bring economic benefits and reduce transportation losses in trade for trade distribution centers with the same spatial and quality characteristics as R fresh agricultural products distribution centers. According to our study, fresh agricultural products distribution enterprises should emphasize the use of genetic algorithms in planning distribution paths, develop a highly adaptable planning system of trade distribution routes, strengthen organizational and operational management, and establish a standard system for high-quality logistics services to improve distribution efficiency and customer satisfaction.

Keywords: fresh agricultural products; time window; path optimization; genetic algorithm

Citation: Sun, J.; Jiang, T.; Song, Y.; Guo, H.; Zhang, Y. Research on the Optimization of Fresh Agricultural Products Trade Distribution Path Based on Genetic Algorithm. *Agriculture* **2022**, *12*, 1669. https://doi.org/10.3390/agriculture12101669

Academic Editors: Gniewko Niedbała and Sebastian Kujawa

Received: 2 September 2022
Accepted: 5 October 2022
Published: 11 October 2022

1. Introduction

With the rapid development of modern society and the economy, people's living and consumption levels are increasing, and the consumption concept and consumption ideals are also gradually being upgraded. In addition, the demand level for personal dimension is rising. In terms of diet, people pursue a diversity of food choices and high-quality food. Therefore, health food, green food, and other food ration products catering to consumers came into being (as shown in Figures 1 and 2) (Source from "Food consumption trends in leading world markets". www.statista.com, accessed on 1 October 2022). In reality, consumers' requirements for fresh produce are gradually changing. The market has shown a trend toward personalized and diversified development. However, fresh agricultural products' corrosive, regional, and seasonal characteristics have put forward very demanding requirements for logistics and distribution. These characteristics restrict the choice of distribution routes and the transportation time of vehicles. The distribution center is an essential link to modern logistics trade activities. It can improve logistics and distribution efficiency by optimizing logistics and trade distribution paths to reduce logistics and distribution costs. According to incomplete statistics, more than 500 million tons of fresh agricultural products need to be distributed in China, and this value is gradually increasing. As the business scale of various fresh agricultural products logistics and distribution companies continues to expand, consumers' requirements for logistics service quality are growing higher and higher. At the same time, while emphasizing the freshness and quality of agricultural products, companies require further reduction of logistics and distribution costs, thus improving distribution efficiency. Therefore, studying

fresh produce distribution path optimization with a time constraint is conducive to reducing the operation cost of fresh produce distribution companies and improving logistics and distribution efficiency. Moreover, this study has important theoretical significance and application value for logistics and distribution companies for fresh agricultural products.

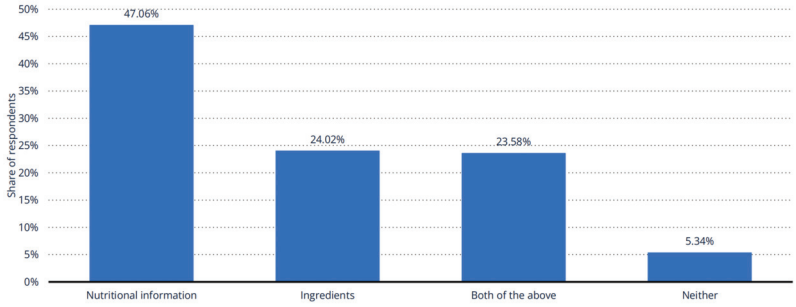


Figure 1. Do you pay attention to any of the following while grocery shopping? (Relevant product information in China 2019).

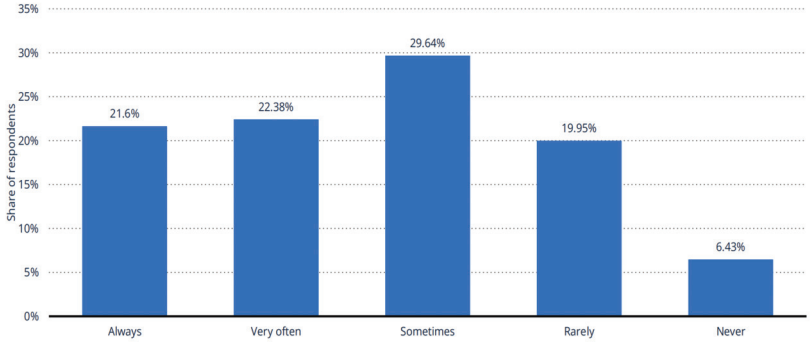


Figure 2. Would you say that you eat a healthy and balanced diet? (Individual following a healthy and balanced diet in China 2019, by frequency).

The planting pattern has recently changed from traditional family-based garden-style vegetable cultivation to large-scale facility farming. The agricultural logistics market is also changing and developing with it. After more than a decade of development, larger wholesale markets for fresh agricultural products have been established. Distribution centers specializing in fresh agricultural products' processing, storage, and logistics have also been established. According to this survey, most distribution centers for fresh agricultural products have severe problems with unreasonable distribution route design. The distribution routes of each distribution center are often chosen by the distribution drivers themselves under the regulation of the restricted time, and the maximum time limit. This approach lacks proper planning and scientific arrangements. This has led to rising distribution costs, wasted resources, and decreased customer satisfaction.

Therefore, optimizing the logistics distribution route is of great significance to improving the overall efficiency of the distribution of fresh produce distribution centers. A scientific and reasonable distribution route optimization measure can effectively reduce the distribution costs of the distribution center and reasonably arrange the distribution vehicles. Moreover, this measure can also effectively supervise distribution personnel and improve customer service. Therefore, exploring distribution path optimization of fresh agricultural product distribution centers can build a good brand image of fresh agricultural products. Furthermore, while satisfying the local people's requirements for the safety and freshness

of agricultural products, it also promotes the healthy development of fresh agricultural product distribution enterprises.

Based on the research results at the forefront of academia, this paper innovates on two aspects of research content and research objects, respectively, to improve distribution centers' distribution efficiency and plan scientific distribution paths.

1. Regarding the research content, mathematical models have solved previous optimization problems of distribution paths and are rarely linked with the operation of distribution centers. In this paper, we select real-life examples to investigate the problems of distribution paths and construct optimization models of distribution paths. The case study method can combine the genetic algorithm with the actual situation more effectively, thus making up for the disconnect between theory and practice in the previous study.
2. Regarding the research object, this paper aims to address the distribution path optimization problem of fresh agricultural product distribution centers. This research will help promote the application of this model in similar fresh agricultural product distribution enterprises with similar characteristics to solve the problem of narrow distribution paths and high distribution costs.

2. Literature Review

In recent years, more and more scholars have focused on researching fresh produce logistics trade distribution and path optimization problems, which are of great significance in terms of the economical use of agricultural trade resources. From the existing literature, the existing studies are mainly distinguished from two aspects: the study of the influence factors of fresh agricultural product distribution and the diversity of research methods used.

The existing literature is mainly focused on fresh agricultural product logistics distribution. The relevant studies mainly focus on three aspects.

The first aspect is the research on the supply chain. Some scholars focus their research perspective on the impact of the supply chain of fresh agricultural products on distribution efficiency. Dabbene's research group took optimization of the fresh produce supply chain as a starting point. From a supply chain optimization perspective, the team focused on the impact of variable factors such as crop growth maturity, storage temperature, and humidity during transportation on fresh produce logistics and distribution efficiency [1]. Ruhe Xie performed a SWOT analysis on the wholesale market model, agricultural leading enterprise mode, third-party logistics model, E-commerce model, supermarket docking model, and farmer direct sales model. On this basis, the study constructed a system dynamics model of regional circulation of fresh agricultural products to enhance the stability of urban fresh agricultural products supply and improve circulation efficiency [2]. Ying Ji researched the decision optimization problem of cold chain logistics from the perspective of cost-benefit to address the problem of severe loss of fresh agricultural products in circulation and solved it using an analytical model. The study proves that using IoT technology in the logistics practice of fresh agricultural products can strengthen the construction of cold chain logistics and improve logistics efficiency [3].

Another part of the scholars focuses on the design of the supply chain [4]. Akbari Kasgari designed a copper network and backed up suppliers as a resilience strategy. This study proposed models without backup and with backup for multiple objectives and compares them with each other. The result shows that using the backup model can increase supply chain responsiveness and improve the economic and social performance of the supply chain [5]. M. Kaviyani-Charati developed two Bi-Level Stackelberg Models (BLSMs) under non- and agile conditions in the presence of strategic customers. The team considered retailers and manufacturers competing with each other in a sequential game to determine the optimal production and order quantities and prices with and without agile capabilities [6].

The second aspect is the study of the impact of delivery vehicles and delivery time on fresh produce logistics distribution. Some scholars believe that delivery vehicles and delivery time should be considered [7]. Jiang et al. proposed a mixed integer nonlinear programming model (MINLP) to optimize the harvest time and vehicle-to-consumer routes, significantly reducing the time from harvest to distribution of agricultural products [8]. Song et al. examined the selection of logistics models and their impact on fresh produce logistics and distribution. This team confirmed the effectiveness of standard refrigerated vehicle models in fresh produce logistics and distribution [9].

The third aspect is the study of fresh agricultural product logistics and distribution from the perspective of freshness. Banerjee et al. studied the decrease in the freshness of fresh produce in the retail sector with increasing production dates. The scholars studied the inventory model of perishable items from the initial selling price to the future preservation conditions. Based on this, they constructed a daily demand function and distribution model for perishable goods to reduce sales problems [10]. Yan Bo and others construct a dual-channel fresh agricultural product (FAP) supply chain consisting of retailers and suppliers. This study thoroughly considered the impact of freshness level on the freshness of perishable products and constructed a time-varying demand function based on freshness to achieve the optimal pricing strategy and profitability of the supply chain members under dual-channel [11].

Most of the research methods used in the existing literature focus on distribution path optimization models and solution algorithms.

The first aspect is the research on vehicle capacity and time window constraints. The research team of Hop Van Nguyen considered two objectives, delivery lead time and total transportation cost, under the constraints of vehicle capacity and time window. The team proposed an adaptive inertia weighted particle swarm optimization (AIWPSO) algorithm to handle the case of a large number of delivery customers [12]. Bauernhansl et al. studied the vehicle routing problem with constraints such as time windows and vehicle capacity in depth. They designed a vehicle route preference model to suit the needs of each customer service outlet for simultaneous distribution and pick-up. They used distribution models scientifically and rationally. The model minimizes the cost of logistics and distribution operations and enhances customer satisfaction [13].

The second aspect is the study of logistics route optimization calculation. Exact algorithms have a long tradition in distribution route optimization [14]. Archetti et al. use the branch pricing shear approach to perform an integrated service simultaneously, providing multiple logistics points and different logistics products. This approach subdivides the vehicle routing problem into a vehicle routing optimization model with partitioning to demonstrate the model's effectiveness [15]. Annelieke C. Baller and others used the branch-and-price-and-cut solution method to solve the vehicle routing problem with outsourcing. The method is used to achieve the minimization of fixed, variable, and outsourced costs of the outsourced fleet [16].

At the same time, new models and research methods have been proposed [17–19]. Ellabib and other developers have pioneered the idea of parallel data processing and improved the search characteristics of the algorithm in the use of ant colony computing. The computational search performance was also improved using parallel processing concepts for the ACO algorithm [20]. The team also proposes a new method for solving the fully intuitive fuzzy multi-objective fractional transportation problem. The method transforms the problem into a linear problem through some transformations. The linearized model is then reduced to a concise multi-objective transportation problem using the accuracy function of each objective [21]. M.A. Elsisy and others have proposed a new algorithm for generating the Pareto frontier for bi-level multi-objective rough nonlinear programming problems [22]. E.M. et al. systematically investigated the effect of different road selection strategies (including interchange strategies commonly used in practice) on response times and proposed a novel road selection strategy. The model examines not only the transport

of cars individually but also the transport transit process between different cars at selected customer locations, known as intermediate transport consolidation. [23].

In conclusion, most scholars' main research directions are the multi-objective distribution path optimization models with time window constraints and other hybrid algorithms such as modern heuristics. However, due to the unpredictable nature of the actual distribution environment, theoretical models and algorithms are confronted with new requirements during their practical implementation. In contrast, intelligent hybrid algorithms often have to traverse the entire search space, which can very easily lead to combinatorial destruction of the search, making the problem impossible to realize in a polynomial time frame.

Based on this, this paper attempts to use genetic algorithms to solve the problem of fresh agricultural product distribution path optimization. From the literature mentioned above, cutting-edge scholars have researched respectively fresh agricultural product distribution problems and specific algorithms for path optimization. However, little content has been explored for combining genetic algorithms with practical cases. We hope our study will help fill the gap in this area.

3. Theoretical Mechanism

The distribution path generally refers to the distribution path between the logistics distribution center and different distribution locations [24]. The distribution path problem is also known as the vehicle route problem (VRP), which mainly includes distribution centers, customer points, and distribution networks [25]. Distribution path optimization refers to optimizing the route taken by distribution vehicles [26]. The distribution center will face multiple scattered distribution points in performing distribution tasks and choosing several distribution paths. The trade distribution path optimization problem is characterized by complex conditions, time-varying tasks, and non-linearity between variables, which require specific analysis [27]. The distribution path optimization problems are carefully classified according to different classification features, and the various types are shown in Table 1 [28].

Table 1. Classification of Distribution Route Optimization Problems.

Classification Standards	Type	Definition
Type of vehicle	Single-type distribution	Delivery by one vehicle type
	Multi-type distribution	Delivery by multiple vehicle types
Number of distribution centers	Vehicle routing with a single depot	One distribution center
	Vehicle routing with multiple depots	Multiple distribution centers
Number of optimization targets	Single target	One optimization target
	Multiple targets	Multiple optimization targets
Time windows	Hard time windows	Delivery within the time window
	Soft time windows	Deliveries can be made outside the time window, but with penalties.
	No time windows	No limitation on delivery time
	Fuzzy time windows	Setting a time window that can influence customer satisfaction
Return Restrictions	Open vehicle routing	Vehicles do not need to return to the distribution center.
	Enclosed vehicle routing	Vehicles need to be returned to the distribution center.
Whether the decision information is certain	Static	Information on distribution environment and demand point conditions is certain.
	Dynamic	Information on distribution environment and demand point conditions is uncertain.
Vehicle loading	Fully loaded	The load of the vehicle is less than the single demand point requirement.
	Non-full load	The load of the vehicle is greater than the single demand point needs.

There are also many ways to solve the optimization problem of distribution routes in logistics. A classification of the main exact and heuristic algorithms is shown in Figure 3 [29].

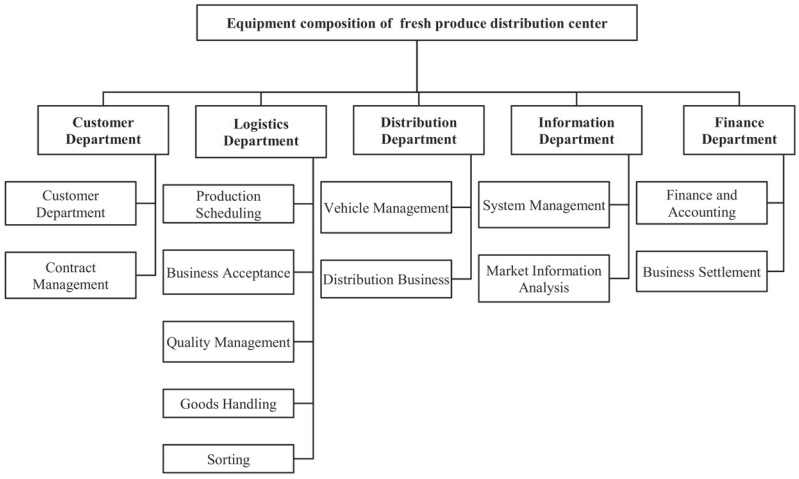


Figure 3. Vehicle routing problem-solving algorithm.

The solution of the forbidden search algorithm is very dependent on the initial solution of the model and is not suitable for solving large-scale problems [30,31]. The simulated annealing algorithm has the advantage of short time consumption and eases handling the combined path solution. However, its ability to calculate the optimal solution is poor, and the parameter settings influence its solution results [32]. The optimal path operation of the ant colony algorithm uses a positive and negative mechanism, which is time-consuming. The algorithm is not conducive to the processing and computation of large-scale data and is prone to stagnation while the algorithm is being computed. The accuracy of its solutions depends on operational experience [33]. In this paper, the genetic algorithm was chosen because it better fits the realistic scenario of this paper. However, the genetic algorithm does not quickly obtain a local best path in solving the best vehicle path due to the slow speed of the optimization operation process.

Nevertheless, the overall best result is better obtained using the genetic algorithm. Due to the slow speed of the optimization operation process, the genetic algorithm is more difficult to obtain the local best path in solving the best path. However, it is easy to arrive at the overall best result. From a holistic perspective, the genetic algorithm has the highest correctness. Combined with the specific situation of this paper's research, this paper uses genetic algorithms in heuristic algorithms as the research method.

4. Analysis of the Current Situation

4.1. Basic Situation

Jinpu New District of Dalian City has unique geographical advantages. Located in the geographical center of Northeast Asia, the district has traded with more than 300 ports in more than 160 countries and regions, with a total area of 2299 square kilometers and a population of 1.61 million. Dalian Jinpu New Area Fresh Agricultural Product Distribution Center (DJ-FAPDC) is located in the Dalian Jinfadi Comprehensive Wholesale Market (Due to the space limitation, the Dalian Jinpu New Area fresh agricultural product distribution center will be named "DJ-FAPDC" in the following contents of this article for the convenience of reading and explanation. This is to clarify.). Dalian Jinfadi Comprehensive Wholesale Market is the largest comprehensive wholesale market in Dalian. It is located at the intersection of Huaihe West Road and Wutun Airport in Jinzhou District, Dalian,

covering an area of more than 400,000 square meters. The wholesale market is developed in two phases, and the agricultural and sideline products trading market is the first phase of construction. It covers an area of 180,000 square meters, with a construction area of 60,000 square meters and 120,000 square meters of roads, parking, and ground hardening. Among them are ten trading halls with an area of 35,520 square meters and six buildings with warehouse gatehouses of 9000 square meters. Other than this, there are four comprehensive supporting buildings for refrigeration and quick freezing, with an area of 14,600 square meters.

In the wholesale market, dozens of small agricultural and sideline products companies and five large fresh agricultural products distribution companies. Among them, R Fresh Product Distribution Center (R-FAPDC), as one of DJ-FAPDC, is the earliest, largest, and most representative distribution center that was established.

To obtain direct data, we conducted a site survey and learned that DJ-FAPDC has a similar composition structure. The organizational management system and equipment composition are shown in Figures 4 and 5.

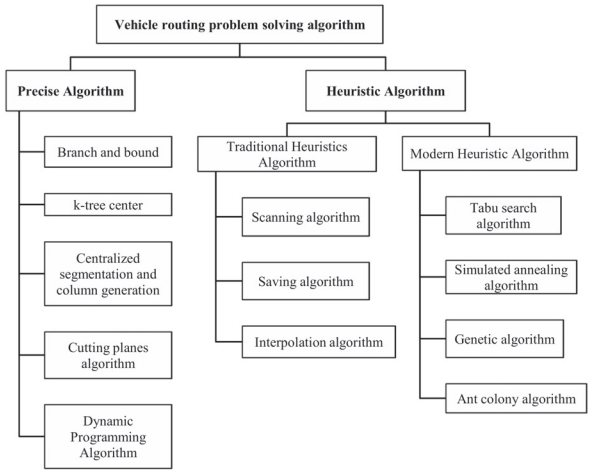


Figure 4. Organization management system.

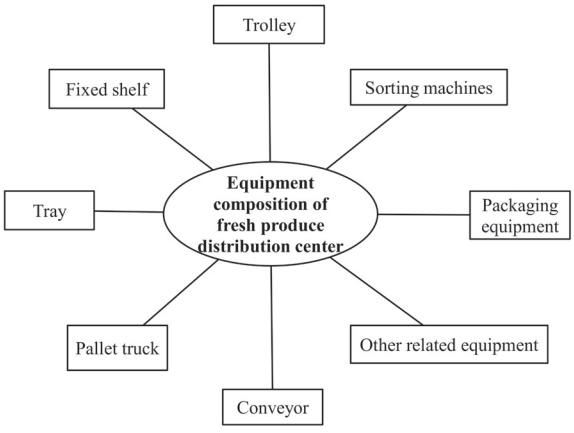


Figure 5. Equipment composition diagram.

4.2. Current Situation and Problems

Through an extensive survey, it was learned that DJ-FAPDC lacks a modern intelligent distribution route planning system. The trade distribution path selection of DJ-FAPDC has the characteristics of randomness and subjectivity, mainly relying on the empirical selection of distribution personnel. This leads to the following problems in the distribution of DJ-FAPDC.

4.2.1. Lack of Standardization in Trade Distribution Route Planning Leads to Higher Distribution Costs

Random logistics and distribution route planning can significantly increase the distribution costs of logistics and distribution centers. A large number of distribution points in each logistics and distribution center results in an excessive number of logistics and distribution routes being chosen. When selecting routes manually will inevitably lead to uncertainty in vehicle mileage and higher distribution costs for logistics and distribution centers.

4.2.2. Lack of Regulation in Distribution Route Planning Leads to Lower Customer Satisfaction

DJ-FAPDC's customer demands are increasingly characterized by multiple batches, small lots, and urgency. Irrational trade distribution route planning can lead to difficulty handling emergencies and heavy load on distribution vehicles during the distribution process. At the same time, the problem of poor customer service is increasingly exposed, leading to lower customer satisfaction.

4.2.3. Lack of Standardization in Distribution Route Planning Makes it Difficult to Guarantee the Distribution Time

DJ-FAPDC has not planned and designed the distribution routes of distribution vehicles. Therefore, strict time limits cannot be imposed on the distribution time of vehicles. This has led to difficulties for DJ-FAPDC to meet the requirements of vehicle recycling. The difficulty of its management is increased. No vehicle may be available for dispatch when there is a temporary urgent delivery task. In addition, it is also difficult to improve customer satisfaction in such situations.

This shows that DJ-FAPDC pays insufficient attention to distribution route planning and lacks a comprehensive logistics distribution and distribution route planning system. The logistics and distribution center was unable to monitor the status of the implementation of the distribution plan, nor was it able to effectively control the distribution vehicles and staff. DJ-FAPDC's logistics concept is backward. The specific performance includes the following three aspects. (1) The distribution center did not invest enough in logistics activities, focusing only on producing and selling fresh agricultural products. (2) The distribution center lacks good logistics and trade distribution planning. The distribution companies did not adopt a modern logistics management system. (3) The logistics and distribution department staff lacked knowledge in the logistics field. In their daily work, they can only rely on their work to accumulate experience, which makes them lack scientific theoretical support and unable to make scientific management decisions.

5. Construction of Distribution Path Optimization Model

This paper will start from the perspective of trade distribution path optimization of fresh agricultural products distribution enterprises, taking R-FAPDC as an example. After considering the timeliness of fresh agricultural products, a distribution path optimization model with time window constraints is constructed and solved in this study. The ultimate objective of this study is to help the fresh produce distribution enterprises in the Jinpu New Area of Dalian find the optimal distribution route with the lowest cost, highest efficiency, and excellent customer satisfaction. At the same time, this paper reduces human decision-making errors by employing scientific planning of distribution routes, realizing real-time control of personnel and vehicle dynamics by enterprises, and closing the loopholes of in-transit supervision.

The elements of the simulation scenario in this paper are: to complete the sorting, loading, and unloading of goods at distribution points within the distribution center, plan good routes for distribution personnel, and complete home delivery services. Among them are known: the distribution network, including the geographic location of each customer node and the distance of each section of the route; cargo information, including attributes such as cargo specifications and weight, as well as the delivery time requested by the customer and the acceptable time window; and dispatch information, including the number of vehicles and the maximum load capacity.

5.1. Model Assumptions

In this paper, the model is constructed based on the research of cutting-edge scholars. The goal is to achieve a high degree of integration between genetic algorithms and actual cases, thus improving the model's practical utility. Reasonable assumptions can simplify the model construction process, reduce the consideration of secondary factors, and reduce the difficulty of model construction. Therefore, to control some unexplained variables and directly reflect the fundamental problems of the model, this study simplifies the actual situation. This paper proposes setting conditions for the appropriate environment required for model construction.

1. Distribution point: each distribution point demand is satisfied, and only one distribution transport vehicle is allowed to serve it and be visited only once.
2. Distribution vehicles: The distribution vehicles are of uniform type and have the same load and speed.
3. Distribution routes: The distribution center has sufficient vehicles to serve each route in time, according to the demand.
4. Distribution center: The distribution center is unique and is the starting, and final destination of each distribution vehicle, and no more midway assignments are made.
5. Load limit: The total amount of all customer demands on the same path shall not exceed the vehicle's rated load.
6. Time window constraint: The arrival time of delivery vehicles and loading operation shall be satisfied within the acceptable time window of the delivery point. Otherwise, it is invalid.
7. Distribution process without consideration of road congestion, bad weather conditions, or vehicle breakdown.

5.2. Symbol Description

The symbols appearing in this paper and their practical meanings are shown in Table 2.

Table 2. Model Symbol.

Symbols	Practical Meanings	Symbols	Practical Meanings
n	The number of vehicles put into the distribution chain.	$[e_i,l_i]$	Customer i specified deliverable time period.
k	The kth vehicle.k$\in${1,2,$\cdots$,n}	E_i	The earliest delivery time tolerated by customer i.
i, j, p	Distribution center or customer location.	L_i	The latest delivery time that can be tolerated by customer i.
R	R = {i, j, p | i, j, p = 1,$\cdots$,m}	$[E_i,L_i]$	The delivery time period that can be tolerated by customer i.
f_1	Fixed costs incurred per vehicle in the distribution chain.	T_i	Customer i defined time point for satisfactory service.
D_{ij}	Miles traveled between customer i and j.	A_{ik}	Actual time of arrival of the kth vehicle at customer i's location

Table 2. *Cont.*

Symbols	Practical Meanings	Symbols	Practical Meanings
f_2	Transportation cost per unit distance	f_3	Waiting costs incurred per unit of time in case of early arrival
X_{ijk}	0,1 variables, indicating that the kth vehicle completes the delivery operation for customer *i,j*.	f_4	Delayed costs incurred per unit of time in case of delayed arrival
E_{ik}	Earliest time for the kth vehicle to arrive at customer *i*'s location	X_{ik}	0,1 variables, indicating that the delivery operation of customer *i* is completed by the kth vehicle
L_{ik}	Latest time for the kth vehicle to arrive at customer i's location	Q_k	Actual load of the kth vehicle
$[E_{ik},L_{ik}]$	Time period for the arrival of the kth vehicle at customer i's location	q_i	Actual demand of customer i
e_i	The earliest time that can be delivered as specified by customer i	Q	Maximum load of distribution vehicles
l_i	The latest deliverable time specified by customer i		

5.3. Distribution Cost Analysis

The distribution cost of the distribution center includes several aspects: transportation, sorting, assembly, and distribution processing. Based on the actual situation of fresh produce distribution in this paper, the sorting, assembly, and distribution processing costs are not the main costs in this paper.

Therefore, they can be disregarded. This paper will consider distribution costs from the following three aspects [34,35].

5.3.1. Fixed Costs

Numbered lists can be added as follows: It refers to costs unrelated to the distance traveled by the vehicle and the number of goods transported within the distribution chain but will undoubtedly be incurred. It includes management costs, repair and maintenance costs, road maintenance fees, bridge fees, and other miscellaneous costs. This part of the cost is only related to the number of vehicles. Thus, the calculation formula is shown below.

$$C_1 = n \times f_1 \tag{1}$$

where *e n* denotes the number of vehicles invested in the distribution segment; f_1 denotes the fixed cost incurred by each vehicle in the distribution segment.

5.3.2. Transportation Costs

It refers to the part of the cost that increases or decreases with the mileage and vehicle load in the distribution process. It is also known as a vehicle-kilometer variable cost, including fuel consumption and depreciation costs. This cost can be obtained by multiplying the mileage of the distribution vehicle and the transportation cost per unit distance. The calculation formula is shown below.

$$C_2 = \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{n} D_{ij} \times f_2 \times X_{ijk} \tag{2}$$

where D_{ij} denotes the mileage traveled between customer *i*, *j*; f_2 denotes the transportation cost per unit distance; X_{ijk} is a 0,1 variable, indicating that the kth vehicle completes the delivery operation of customer *i*, *j*.

5.3.3. Penalty Costs

Customers often have personalized requirements for delivery time and want to complete the delivery within the specified time window. However, in the actual delivery

operation, the delivery operation cannot be completed within the specified time due to objective reasons such as bad weather, traffic jams, or subjective factors such as mistakes in route selection by delivery personnel. Early or late delivery can affect customer satisfaction and bring penalty costs to the delivery company. In this paper, penalty cost is defined as any additional cost to delivery companies due to the impact of delivery time [36]. According to the times, delivery vehicles arrive at customer locations. This paper will analyze penalty costs in three cases: early arrival, on-time arrival, and delayed arrival.

1. Early arrival.

Usually, the customer specifies the delivery time, which means that the delivery can only be signed for within a specific period. Therefore, when the vehicle arrives at the customer's location in advance, it needs to wait until the set time for door-to-door delivery service. In the waiting process, vehicles and human resources will be wasted, equivalent to delivery companies paying extra waiting costs.

2. On-time arrival.

The door-to-door delivery service can be carried out on time when the vehicle arrives at the customer's location according to the agreed time. After the customer signs for the delivery, the distribution staff will go to the next destination. There will be no additional cost to the delivery company in this ideal situation.

3. Delayed arrival.

Due to traffic congestion and human error, the delivery company must pay additional vehicle in-transit costs when the vehicle arrives later than the specified delivery time. Furthermore, the company also needs to bear the cost of possible secondary delivery.

Suppose the time for the kth vehicle to arrive at customer i's location is $[E_{ik}, L_{ik}]$. The deliverable time specified by customer i is $[e_i, l_i]$ and is a subset of $[E_{ik}, L_{ik}]$. The actual vehicle arrival time $A_{ik} \in [E_{ik}, L_{ik}]$. Based on the above three cases for penalty cost generation analysis, a line graph of penalty cost with distribution time is derived, as shown in Figure 6.

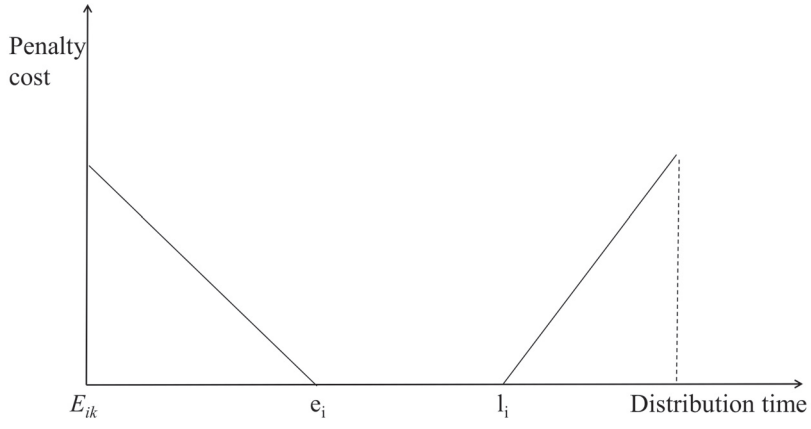


Figure 6. Penalty cost line chart.

In summary, the expression for calculating the penalty cost of the distribution chain is:

$$C_3 = f_3 \sum_{i=1}^{m} \sum_{k=1}^{n} X_{ik} \times \max(0, e_i - A_{ik}) + f_4 \sum_{i=1}^{m} \sum_{k=1}^{n} X_{ik} \times \max(0, A_{ik} - l_i) \quad (3)$$

f_3 indicates the waiting cost incurred per unit of time in case of early arrival.

f_4 denotes the delayed cost incurred per unit time when arriving late.
X_{ik} is a 0, 1 variable indicating that the kth vehicle performs the delivery operation for customer i.

5.4. Analysis of Customer Satisfaction

This paper's distribution path optimization model is a multi-objective function model to achieve the minimum distribution cost and the highest customer satisfaction simultaneously. Therefore, this paper needs to analyze customer satisfaction briefly.

Customer satisfaction refers to the subjective feelings of customers about the products or services they receive. When portraying satisfaction, it is generally the ratio between the expected value and the final achieved value, which is between [0,1]. In the case of delivery services, satisfaction is mainly influenced by whether the delivery company can meet the customer's requirements for delivery time [37].

5.4.1. Fuzzy Appointment Time

The concept of fuzzy appointment time has emerged in academia because of the uncertainty in real life regarding the setting of delivery time windows by customers. Fuzzy appointment time is the concept of time interval, which contains the customer's desired service point or time and the range of service time that the customer can tolerate to overrun or delay [38].

When the adequate service time of customer i is a specific time point T_i, the linear image representation of customer satisfaction is shown in Figure 7.

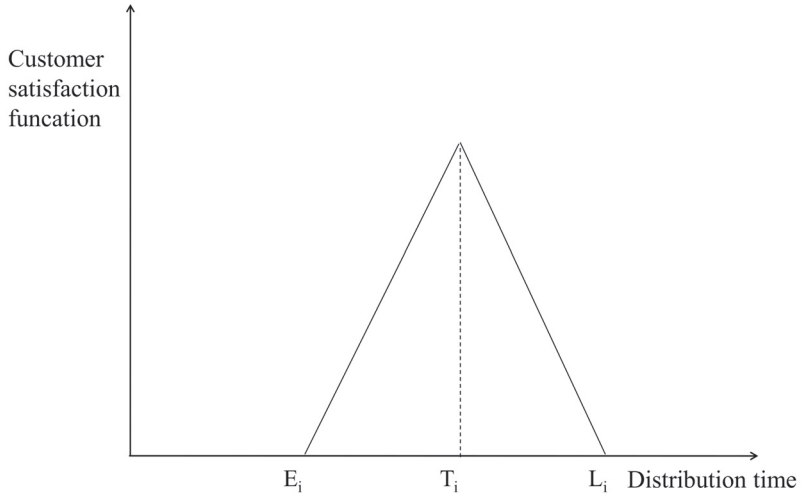


Figure 7. Customer satisfaction function graph (1).

The linear image of customer satisfaction when the satisfied service time of customer i is a specific time $[e_i, l_i]$ is shown in Figure 8.

Customer satisfaction at different time points or periods is shown as follows.

1. When the delivery service occurs at the reservation time point T_i or within the time $[e_i, l_i]$, the customer expectation is met, and the satisfaction level reaches the highest.
2. When the delivery service does not occur at the reservation time point T_i or within the time $[e_i, l_i]$ but is within the tolerable delivery time window $[E_i, L_i]$, the customer's expectations are not achieved. However, the customer still chooses to accept the service. At this time, customer satisfaction is average.

3. When the delivery service occurs outside the tolerable delivery window $[E_i, L_i]$, it is far from meeting the customer's expectations, and customer satisfaction is the lowest.

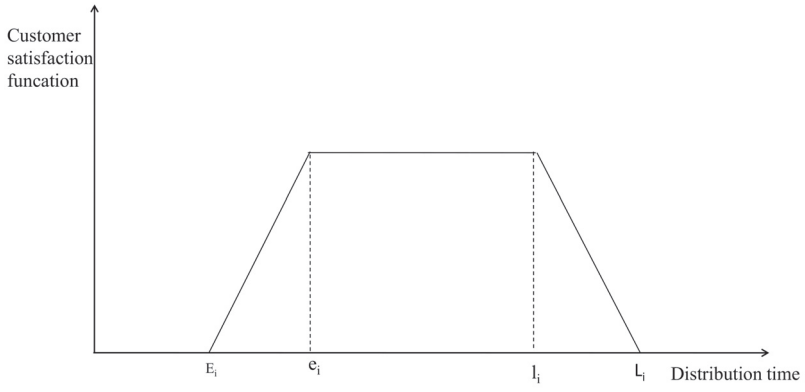


Figure 8. Customer satisfaction function graph (2).

5.4.2. Time Window-Based Customer Satisfaction Measurement

In the actual delivery process, in most cases, customers will pre-set the delivery time window instead of the scheduled time point. Therefore, this paper's time window is fuzzy reservation time processing. Concurrently, this paper uses the fuzzy post-affiliation function shown in the following equation to represent customer satisfaction [39–41].

$$F(A_{ik}) = \begin{cases} 0 & A_{ik} \le E_i \\ \frac{A_{ik}-E_i}{e_i-E_i} E_i \le A_{ik} \le e_i \\ 1 & e_i \le A_{ik} \le l_i \\ \frac{L_i-A_{ik}}{L_i-l_i} l_i \le A_{ik} \le L_i \\ 0 & A_{ik} \ge L_i \end{cases} \quad (4)$$

The expression for the average customer satisfaction is:

$$F(A) = \frac{\sum_{i=1}^{m} F(A_{ik})}{m} \quad (5)$$

5.5. Model Building

5.5.1. Objective Function

1. Treatment of multi-objective function

This paper studies how to achieve the optimal balance between the two objectives of minimizing distribution cost and maximizing customer satisfaction based on satisfying distribution timeliness. This problem belongs to the multi-objective function optimization problem [42]. In order to make the solution more convenient, this study converts the two objective function optimization problems into achieving a single objective function optimization problem. The special treatment is shown below.

$$F'(A) = 1 - F(A) \quad (6)$$

In the following, to be consistent with solving the distribution cost minimization objective, this paper converts achieving the maximum average customer satisfaction into solving the minimum average customer dissatisfaction.

2. Final objective function

$$\begin{aligned} MinZ = C1 + C2 \quad & +C3 + F'(A) \\ & = n \times f_1 + \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{n} D_{ij} \times f_2 \times X_{ijk} \\ & + f_3 \sum_{i=1}^{m} \sum_{k=1}^{n} X_{ik} \times \max(0, e_i - A_{ik}) \\ & + f_4 \sum_{i=1}^{m} \sum_{k=1}^{n} X_{ik} \times \max(0, A_{ik} - l_i) + 1 - \frac{\sum_{i=1}^{m} F(A_{ik})}{m} \end{aligned} \tag{7}$$

5.5.2. Constraints

All models have corresponding conditions of applicability. In order to constrain the generation of its wide range of results and facilitate analytical processing, this study constructs the following constraints for the above models.

1. 0-1 variables

$$X_{ik} = \begin{cases} 1, \text{ the delivery operation of customer} \\ \text{i is carried out by the kth vehicle} \\ 0, \text{ others} \end{cases} \tag{8}$$

$$X_{ik} = \begin{cases} 1, \text{ the delivery operation of customer i} \\ \text{and j are completed by the kth vehicle} \\ 0, \text{ others} \end{cases} \tag{9}$$

2. The actual weight of the vehicle shall not exceed the maximum weight of the vehicle.

$$Q_k = \sum_{i=1}^{m} X_{ik} \times Q_i \tag{10}$$

$$Q_k \leq Q \tag{11}$$

3. Each customer has only one vehicle to deliver for him/her.

$$\sum_{k=1}^{n} X_{ik} = 1 \tag{12}$$

4. After a delivery task is completed, the delivery vehicle must go to the next delivery point.

$$\sum_{i=1}^{m} X_{ipk} = \sum_{j=1}^{m} X_{pjk} \tag{13}$$

5. The vehicle returns to the distribution center after completing all tasks.

$$\sum_{i=1}^{m} \sum_{k=1}^{n} X_{i1k} = \sum_{j=1}^{m} \sum_{k=1}^{n} X_{1jk} \tag{14}$$

6. The vehicle capacity fulfills the distribution needs of all customers.

$$\sum_{i=1}^{m} \sum_{j=1}^{m} X_{ij} = m \tag{15}$$

7. The number of vehicles involved in distribution does not exceed the number of all distribution vehicles.

$$\sum_{k=1}^{n} X_{1i} \leq n \tag{16}$$

8. Meeting the time window constraints of all customers.

$$e_i \leq A_{ik} \leq l_i \tag{17}$$

6. Case Study of Fresh Produce Distribution Center

6.1. Background Introduction

R-FAPDC is one of the earliest fresh produce distribution enterprises established in the Jinpu New Area. Furthermore, it is also one of the critical leading companies of agricultural industrialization in the Dalian Jinpu New Area. The distribution center was built in 2005. It has a freezing processing warehouse of 1000 m^2, a distribution center of 2000 m^2, a vegetable planting base of 333,335 m^2, and 86 employees. In addition, the distribution center has 13 vehicles in charge of distribution, and the vehicle types are all light vans. Among them, the total vehicle mass is 4500 kg, and the load capacity is 1500 kg.

6.2. Data Collection and Processing

6.2.1. Constraints

The main distribution area of R-FAPDC is 13 distribution areas in Jinpu New District. The locations of distribution points are shown in Figure 9.

Figure 9. TheR Distribution center customer points.

6.2.2. Addresses and Codes of Distribution Points

The addresses and codes of R-FAPDC and 13 distribution points are shown in Table 3.

Table 3. Distribution point address and code.

Name	Address	Code
R Fresh produce distribution center	No. 1, Yingjun Road, Jinzhou District, Dalian	0—distribution center
Distribution point 1	No. 10 Xuefu Street, Jinzhou District, Dalian	1—P1
Distribution point 2	Huaihe Middle Road, Jinzhou District, Dalian	2—P2
Distribution point 3	East Liahe Road and Shuang D1 Street, Jinzhou District, Dalian	3—P3
Distribution point 4	Intersection of Dong'an Road and Donghui Street, Jinzhou District, Dalian	4—P4
Distribution point 5	No. 50–66, Pengyun Home, Jinzhou District, Dalian	5—P5
Distribution point 6	Near Kushan Middle Road, Jinzhou District, Dalian	6—P6
Distribution point 7	Near Maqiaozi Street, Jinzhou District, Dalian	7—P7
Distribution point 8	No. 19, Tonghui Road, Jinzhou District, Dalian	8—P8
Distribution point 9	No. 18, West Liahe Road, Jinzhou District, Dalian	9—P9
Distribution point 10	No. 31 Tieshan West Road, Jinzhou District, Dalian	10—P10
Distribution point 11	No. 14 Tieshan West Road, Jinzhou District, Dalian	11—P11
Distribution point 12	No. 288 Jingang Road, Jinzhou District, Dalian	12—P12
Distribution point 13	Intersection of Jingang Road and Longwan Road, Jinzhou District, Dalian	13—P13

6.2.3. Distribution Point Data Information

This paper retrieved relevant information through the distribution center sales system during the research period. This paper can obtain the distance and demand information of 13 distribution outlets in R-FAPDC. The summary is shown in Table 4.

Table 4. Send point data information.

Node Coordinates	Horizontal Coordinate (Latitude)	Vertical Coordinate (Longitude)	Distance (km)	Demand (t)
1	39.106662	121.826772	10.3	0.43
2	39.077017	121.856698	7.8	0.55
3	39.070833	121.871083	11.5	0.26
4	39.046255	121.818441	7.0	0.5
5	39.031466	121.813384	7.9	0.47
6	38.969756	121.852591	17.5	0.70
7	39.054987	121.808604	5.8	0.485
8	39.104574	121.840646	10.9	0.37
9	39.051454	121.781714	3.7	0.44
10	39.075549	121.775135	2.5	0.5
11	39.071659	121.765949	3.0	0.55
12	39.083375	121.747865	4.2	0.45
13	39.082096	121.730709	4.0	0.44

The distance information between distribution outlets is shown in Table 5.

Table 5. Distance table between distribution points.

Node Label (km)	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0												
2	5.1	0											
3	8.6	4.8	0										
4	7.8	5.4	8.2	0									
5	9.7	7.3	10.1	3.2	0								
6	17.8	15.4	15.7	11.4	10.1	0							
7	7.8	5.4	8.2	2.6	5.1	13.2	0						

Table 5. *Cont.*

Node Label (km)	1	2	3	4	5	6	7	8	9	10	11	12	13
8	7.0	4.6	5.5	9.0	11.1	19.3	8.4	0					
9	10.5	7.7	10.4	3.8	5.1	14.4	8.7	10.7	0				
10	7.1	7.2	9.9	5.4	7.8	16	3.4	8.2	4.5	0			
11	8.3	7.7	10.5	5.9	8.4	16.5	4.4	9.4	5.0	1.0	0		
12	10.5	10.0	12.7	8.2	10.6	18.8	6.6	16.1	7.3	3.2	2.2	0	
13	12.1	11.6	14.3	9.8	11.0	20.4	8.2	13.2	7.5	4.8	3.8	1.7	0

6.2.4. Demand Time Window

It refers to the best service time acceptable to the customer and the tolerable delivery time. This method requires converting the time format of the time window to a decimal format that is convenient for algorithmic operations. Example: "06:00" is converted to "06.00". The results of the conversion are shown in Table 6.

Table 6. Customer best service time window and tolerable time window statistics table.

Node Coordinates	Optimal Service Time Window $[e_i,l_i]$		Tolerable Time Window $[E_i,L_i]$	
	Earliest Time e_i	Latest Time l_i	Earliest Time E_i	Latest Time L_i
1	06.00	09.00	05.00	10.00
2	06.00	09.00	05.00	10.00
3	05.30	08.30	04.30	09.30
4	06.30	09.30	05.30	10.30
5	05.30	08.30	04.30	09.30
6	06.30	09.30	05.30	10.30
7	06.00	09.00	05.00	10.00
8	06.30	09.30	05.30	10.30
9	07.00	10.00	06.00	11.00
10	05.30	08.30	04.30	09.30
11	06.30	09.30	05.30	10.30
12	06.00	09.00	05.00	10.00
13	05.30	08.30	04.30	09.30

6.2.5. Other Parameters

To simplify the operation, this study takes the median value of each parameter interval to fix the parameters, as shown in Table 7.

Table 7. Other values in the model.

Parameter	Takes Values
Unit vehicle fixed cost (CNY/vehicle)	150.0
Unit distance transportation cost (CNY/km)	2.0
Early arrival waiting fee (CNY/h)	10.0
Delayed arrival delay fee (CNY/h)	20.0
Vehicle maximum load capacity (t)	1.5
Average vehicle speed (km/h)	30.0
Average customer service time (h)	0.5

6.3. Algorithm Parameter Setting

According to the actual situation of the calculation case, in this paper, if we want to simplify the operation based on the validity of the solution results, we need to set the algorithm parameters. The specific parameter settings are shown in Table 8.

Table 8. Genetic algorithm parameter setting.

Parameter	Initial Population Size	Crossover Probability	Mutation Probability	Maximum Number of Iterations
Takes values	200	0.9	0.1	200

6.4. Model Solving

We substitute the collated data and the set parameters into the model constructed in the previous chapter and use MATLAB computer software to program the solution. Through simulation experiments on the distribution path of R-FAPDC, the final cost iteration diagram and path diagram after path optimization are derived, as shown in Figures 10 and 11.

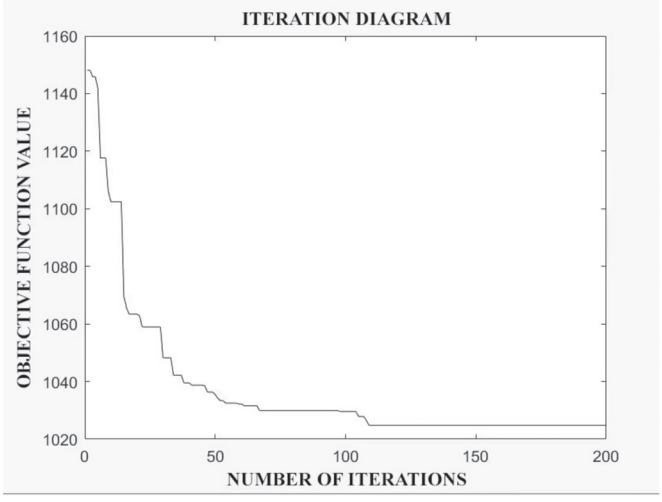


Figure 10. Cost search iteration chart.

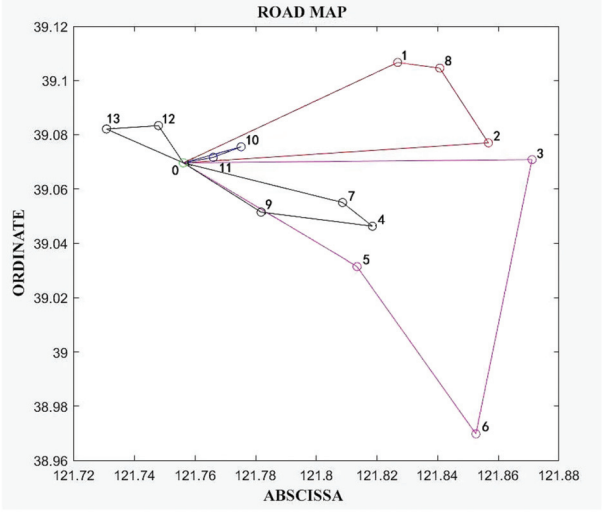


Figure 11. Optimized path planning diagram.

The specific route selection after optimization is shown in Table 9.

Table 9. Optimized distribution path table.

Route Number	Driving Path
1	0-2-8-1-0
2	0-11-10-0
3	0-3-6-5-0
4	0-7-4-9-0
5	0-13-12-0

6.5. Comparative Analysis before and after Optimization

The initial distribution routes before optimization are shown in Table 10.

Table 10. Distribution path table before optimization.

Route Number	Driving Path
1	0-10-1-0
2	0-3-2-0
3	0-7-8-0
4	0-4-9-0
5	0-5-6-0
6	0-11-12-13-0

Before optimization, the distribution vehicle dispatching details are shown in Tables 11–13.

Table 11. List of distribution vehicles before optimization.

Vehicle Number	Order of Visit	Load Capacity (t)	Route (km)	Time (h)	Full Load Rate (%)
1	0-10-1-0	0.93	19.9	1.66	62
2	0-3-2-0	0.81	24.1	1.80	54
3	0-7-8-0	0.85	25.1	1.84	57
4	0-4-9-0	0.94	8.2	1.27	63
5	0-5-6-0	1.17	35.5	2.18	78
6	0-11-12-13-0	1.44	10.9	1.86	96

Table 12. Optimization of vehicle travel time nodes before.

Route Number	Time Nodes
1	0(05:00)-10(05:05–05:35)-1(05:50–06:20)-0(06:40)
2	0(05:00)-3(05:23–05:53)-2(06:03–06:33)-0(06:49)
3	0(05:00)-7(05:12–05:42)-8(05:59–06:29)-0(06:51)
4	0(05:00)-4(05:14–05:44)-9(05:52–06:22)-0(07:00)
5	0(05:00)-5(05:17–05:47)-6(06:07–06:37)-0(07:12)
6	0(05:00)-11(05:06–05:36)-12(05:41–06:11)-13(06:15–06:45)-0(06:53)

The optimized path diagram clearly shows that the distribution center needs 5 vehicles to complete the distribution tasks in the distribution chain. The scheduling and usage details of the corresponding distribution vehicles are shown in Tables 14–16.

Table 13. Optimize the pre-delivery vehicle cost table.

Vehicle Number	Fixed Cost (CNY)	Transportation Cost (CNY)	Penalty Cost (CNY)	Total
1	150	39.8	20	209.8
2	150	48.2	10	208.2
3	150	50.2	20	220.2
4	150	16.4	40	206.4
5	150	71.0	20	241.0
6	150	21.8	40	211.8

Table 14. List of optimized delivery vehicles.

Vehicle Number	Order of Visit	Load Capacity (t)	Route (km)	Time (h)	Full Load Rate (%)
1	0-2-8-1-0	1.35	29.7	2.49	90
2	0-11-10-0	1.05	6.5	1.22	70
3	0-3-6-5-0	1.43	45.2	3.00	95
4	0-7-4-9-0	1.42	15.9	2.03	95
5	0-13-12-0	0.89	9.9	1.33	60

Table 15. The optimized vehicle travel time node.

Route Number	Time Nodes
1	0(05:00)-2(05:16–05:46)-8(05:56–06:26)-1(06:40–07:10)-0(07:30)
2	0(05:00)-11(05:06–05:36)-10(05:38–06:08)-0(06:13)
3	0(05:00)-3(05:23–05:53)-6(06:25–06:55)-5(06:15–06:45)-0(07:00)
4	0(05:00)-7(05:20–05:50)-4(05:56–06:26)-9(06:34–07:04)-0(07:12)
5	0(05:00)-13(05:08–05:38)-12(05:42–06:12)-0(06:20)

Table 16. Optimized delivery vehicle cost table.

Vehicle Number	Fixed Cost (CNY)	Transportation Cost (CNY)	Penalty Cost (CNY)	Total
1	150	59.4	20	229.4
2	150	13.0	20	183.0
3	150	90.4	20	260.4
4	150	31.8	30	211.8
5	150	19.8	20	189.8

The detailed comparison results before and after distribution path optimization are shown in Table 17.

Table 17. Comparison table of results before and after optimization.

	Before Optimization	After Optimization	Comparison before and after Optimization
Number of vehicles (vehicles)	6.0	5.0	−1.0
Fixed cost (CNY)	900.0	750.0	−150.0
Transportation cost (CNY)	247.4	214.4	−33.0
Penalty cost (CNY)	150.0	110.0	−40.0
Distribution cost (CNY)	1294.4	1074.4	−220.0
Distribution mileage (km)	123.7	107.2	−16.5
Customer satisfaction (%)	87.9	88.3	+0.4

After various cost calculations, the results show that the distribution center must dispatch five transport vehicles in the distribution chain. The total cost of distribution at this stage is 1074.4 RMB, compared with the total cost corresponding to the distribution path before optimization, which is 1294.4 RMB, a reduction of 220 RMB. Among them, the fixed cost savings accounted for 68% of the total cost savings; the transportation cost savings accounted for 14% of the total cost savings; the penalty cost savings accounted for 18% of the total cost savings. The improved delivery path can save 33 min compared to the pre-improved delivery path. The average loading rate of the six transport vehicles before optimization was 68%, and the average loading rate of the five transport vehicles after optimization was 82%. The vacancy rate of distribution vehicles is low.

The comparison results show that the model constructed in this paper is reasonable and practical. Optimized trade distribution routes can help distribution companies reduce costs, including transportation costs and penalty costs. Concurrently, customer satisfaction can be improved.

7. Conclusions and Prospects

7.1. Conclusions

With the growing demand for fresh agricultural products trade, the research about fresh agricultural products distribution and the business of fresh agricultural products logistics distribution has been developing rapidly. What cannot be ignored is that trade distribution has gradually shifted from a single route to an interactive network. This has also led the fresh produce distribution centers to enter a critical period of rapid development in multiple ways. They have more paths and ways of choice possibility. Through the case analysis and comparison results derived from the above model, this study draws the main conclusions based on the theory related to distribution path optimization. The results of the study include that:

1. Due to the expansion of the distribution network of distribution centers, the previous method of judging distribution paths based on manual experience can no longer adapt to the growing fresh agricultural product distribution centers.
2. R-FAPDC has its value. It has the typical characteristics of fresh trade distribution in coastal areas. Suppose the distribution problem research of such object considers the particular timeliness of fresh agricultural products distribution, constructs a multi-objective distribution path optimization model with time window constraints, and uses the genetic algorithm to derive a reasonable distribution route that meets the independent characteristics. In that case, it will be of great significance to improve the efficiency of trade distribution.

The results of this paper reveal that:

1. The trade distribution path scheme optimized by the genetic algorithm can reduce the distribution cost of fresh agricultural products distribution centers and improve customer satisfaction.
2. The genetic algorithm can bring economic benefits and reduce transportation losses in trade for the trade distribution centers with the same spatial characteristics and quality characteristics as R-FAPDC.

Therefore, the above conclusion shows that the genetic algorithm is effective for the optimization scheme of trade distribution path for fresh agricultural products distribution centers and has reference significance for other similar problems.

7.2. Data Collection and Processing

7.2.1. Emphasis on the Use of Genetic Algorithms in Planning Trade Distribution Path

The genetic algorithm has excellent advantages. We have analyzed land transportation traffic in a vital development zone in the Northeast Asia International Shipping Center compared with other studies, considering various complex possibilities. Moreover, in this study, the genetic algorithm ensures the diversity of the distribution path population and

the overall population quality and suppresses the algorithm's premature search. The use of this algorithm is more conducive to obtaining the optimal global solution and optimal distribution paths. The comparison with the original internal data of the company confirms the idea of "effectiveness." Therefore, we suggest paying attention to the use of genetic algorithms in planning trade distribution paths and applying them to a broader range of problems. This is a trend to be considered.

7.2.2. Development of a Planning System of Trade Distribution Routes with High Adaptability

The distribution route planning system is crucial for a fresh produce distribution company. At a time when the modern trade process cycle is growing shorter and shorter, each trade link's efficiency is becoming more and more demanding. Companies largely ignore distribution problems, and their random and unexpected nature makes it even less easy to control costs. Therefore, a stable planning system of trade distribution route is related to the cost saving and profit growth or even the quality of the whole trade link completion. The function of the distribution route planning system is not only route planning but also covers all the links in the trade distribution supply chain. Distribution route planning is only one of the functions, but it is also essential. When used well by enterprises, it can significantly improve fresh food companies' delivery speed and efficiency [43]. Compared to manual experience, scientific system planning is more reasonable. This paper provides a model framework and demonstrates its initial effectiveness in improving trade distribution efficiency and reducing distribution costs. It also provides a reference for developing respective planning systems of trade distribution routes.

7.2.3. Strengthening the Rational Setting of Worker Performance in the Management of Organizational Operations

People are the root of all productive forces. This means that people can determine productivity and production relations. That is not to mention the power of a single worker in handling the resources of the company to which he belongs without constraints [44]. If workers are motivated through a performance setting, it will surely increase resource-saving efficiency. Therefore, fresh produce distribution centers should form a reasonable organizational structure. Based on an effective organizational structure, the distribution company should use a performance management model throughout each link, from production to sales. For example, distribution centers ought to implement operational handling performance management in the sorting segment, performance management of trade distribution efficiency in the distribution segment, and sales performance management in the sales segment [45]. In this way, it helps the staff be willing to work and think about company profitability and saving. In addition, the distribution center should regularly train the staff with professional knowledge and skills to help them combine their professional knowledge and work experience effectively and thus improve their distribution efficiency.

7.2.4. Establishing a High-Quality Logistics Service Standard System to Realize the Benign Development of Enterprise Distribution Work in the Process of Winning Customers

A high-quality logistics trade service standard system can not only improve the service level of distribution centers and optimize the business process of the distribution system but also significantly improve the efficiency of logistics and distribution to promote the smooth and sound development of fresh agricultural products distribution centers. It is the next development trend in logistics and distribution of distribution centers. Through standardized means, enterprises standardize service concepts, integrate service functions, and promote service equipment, to finally realize the standardized management of fresh agricultural products distribution centers.

8. Deficiencies and Prospects

8.1. Deficiencies

In this paper, a representative R-FAPDC in Jinpu New District, Dalian, is selected as an arithmetic example, and its distribution path is optimized using a genetic algorithm. Satisfactory results are finally obtained. However, this is an experimental result obtained by simulation under certain assumptions and constraints. Furthermore, the experiment eventually has to return to reality. In real life, it is often disturbed by many external factors. The subsequent research needs to consider various external influences more comprehensively.

1. Some specific conditional restrictions are proposed in the model assumptions section. For example, all delivery vehicles are the same model, and the customer demand at the delivery point is always the same.
2. In the data collection part, the customer demand information is mainly derived from the system records of R-FAPDC. This is a real-time data record. The distribution of data information may change at any time, which in turn affects the optimization effect. In this paper, the supply information of distribution nodes is known in advance. However, in reality, the supply quantities of distribution points are uncertain. This may lead to unforeseen new problems when implementing specific solutions.
3. The optimized distribution path scheme is not implemented in practice in this paper due to the time constraints of the study and personal and professional levels. Therefore, the specific implementation effects in the distribution environment have not been studied in depth.

8.2. Prospects

This paper constructs a multi-objective distribution path optimization model with time windows based on the research of frontier scholars. Preliminary research results have been achieved using a genetic algorithm optimization solution. However, with the continuous development of society, it is necessary to continuously enrich, improve, and develop the theory to carry out subsequent research work.

1. Subsequent research should continuously optimize the algorithm. This paper only uses the genetic algorithm for path optimization solutions. There are many path optimization algorithms. Each algorithm has its advantages and disadvantages. Solving such problems with a single algorithm will inevitably have drawbacks and defects. The next step can be to improve the genetic algorithm or combine the genetic algorithm with other heuristic algorithms.
2. The distribution center should use modern intelligent technology to monitor the distribution process. In the future implementation of the distribution plan, consider in-depth research in supervising the distribution process and grasping the optimization effect. In particular, with the wide application of IoT ("Internet of Things") technology, fresh agricultural product distribution centers should use modern intelligent methods to monitor the distribution process effectively.
3. Researchers should consider the freshness of agricultural products in the optimization model constraints. The confirmation and measurement of the freshness of the agricultural products are not considered in the distribution process, and no detailed constraints are established. In the following research, it is necessary to further quantify and reflect the freshness of fresh agricultural products in the model to improve the optimization effect comprehensively.
4. Researchers should continue to study the main reasons affecting the distribution efficiency in the distribution process, such as congestion or the distribution efficiency affected by the receiving and inspection link. If the delivery time is delayed by the receiving and inspection process, blind matching can be considered. Compare the cost of blind distribution with the previous distribution cost to choose a more reasonable distribution method.

Author Contributions: Writing—original draft preparation, J.S.; writing—review and editing, T.J., Y.S., H.G. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Due to the confidentiality agreement requirements for the information sources used in this study, only the results are published, and the data sources for comparison are not explained at this time.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dabbene, F.; Gay, P.; Sacco, N. Optimisation of fresh-food supply chains in uncertain environments, Part I: Background and methodology. *Biosyst. Eng.* **2008**, *99*, 348–359. [CrossRef]
2. Xie, R.; Han, S.; Jiang, Y.; Peng, Z. Comparison and Optimization of Circulation Modes of Fresh Agricultural Products Based on System Dynamics—The Case of China. *J. Serv. Sci. Manag.* **2018**, *11*, 297–322. [CrossRef]
3. Ying, J. Decision Optimization for Cold Chain Logistics of Fresh Agricultural Products under the Perspective of Cost-Benefit. *Libr. J.* **2019**, *6*, 1–17.
4. Taghipour, A.; Khazaei, M.; Azar, A.; Ghatari, A.R.; Hajiaghaei-Keshteli, M.; Ramezani, M. Creating Shared Value and Strategic Corporate Social Responsibility through Outsourcing within Supply Chain Management. *Sustainability* **2022**, *14*, 1940. [CrossRef]
5. Akbari-Kasgari, M.; Khademi-Zare, H.; Fakhrzad, M.B.; Hajiaghaei-Keshteli, M.; Honarvar, M. De-signing a resilient and sustainable closed-loop supply chain network in copper industry. *Clean Technol. Environ. Policy* **2022**, *24*, 1553–1580. [CrossRef]
6. Kaviyani, C.M.; Ghodsypour, S.H.; Hajiaghaei, K.M. Impact of adopting quick response and agility on supply chain compe-tition with strategic customer behavior. *Sci. Iran.* **2022**, *29*, 387–411.
7. Li, M.; He, L.; Yang, G.; Lian, Z. Profit-Sharing Contracts for Fresh Agricultural Products Supply Chain Considering Spatio-Temporal Costs. *Sustainability* **2022**, *14*, 2315. [CrossRef]
8. Jiang, Y.; Chen, L.; Fang, Y. Integrated Harvest and Distribution Scheduling with Time Windows of Perishable Agri-Products in One-Belt and One-Road Context. *Sustainability* **2018**, *10*, 1570. [CrossRef]
9. Song, B.D.; Ko, Y.D. A vehicle routing problem of both refrigerated- and general-type vehicles for perishable food products delivery. *J. Food Eng.* **2016**, *169*, 61–71. [CrossRef]
10. Banerjee, S.; Agrawal, S. Inventory model for deteriorating items with freshness and price dependent demand: Optimal discounting and ordering policies. *Appl. Math. Model.* **2017**, *52*, 53–64. [CrossRef]
11. Yan, B.; Fan, J.; Wu, J.-W. Channel choice and coordination of fresh agricultural product supply chain. *RAIRO—Oper. Res.* **2021**, *55*, 679–699. [CrossRef]
12. Hop, V.N.; Phan, P.P. Adaptive inertia weight particle swarm optimisation for a multi-objective capacitated vehicle routing problem with time window in air freight forwarding. *Int. J. Syst. Manag.* **2021**, *40*, 423–442.
13. Oesterle, J.; Bauernhansl, T. Exact Method for the Vehicle Routing Problem with Mixed Linehaul and Backhaul Customers, Heterogeneous Fleet, time Window and Manufacturing Capacity. *Procedia CIRP* **2016**, *41*, 573–578. [CrossRef]
14. Archetti, C.; Bianchessi, N.; Grazia, S.M. A branch-price-and-cut algorithm for the commodity constrained split delivery ve-hicle routing problem. *Comput. Oper. Res.* **2015**, *64*, 1–10. [CrossRef]
15. Dimitrakos, T.; Kyriakidis, E. A single vehicle routing problem with pickups and deliveries, continuous random demands and predefined customer order. *Eur. J. Oper. Res.* **2015**, *244*, 990–993. [CrossRef]
16. Annelieke, C.B.; Said, D.B.; Wout, E.H.; Dullaert, D.V. The Vehicle Routing Problem with Partial Outsourcing. *Transport. Sci.* **2020**, *54*, 855–1152.
17. El Sayed, M.; Farahat, F.; Elsisy, M. A novel interactive approach for solving uncertain bi-level multi-objective supply chain model. *Comput. Ind. Eng.* **2022**, *169*, 108225. [CrossRef]
18. Elsisy, M.A.; Elsaadany, A.S.; El Sayed, M.A. Using Interval Operations in the Hungarian Method to Solve the Fuzzy Assignment Problem and Its Application in the Rehabilitation Problem of Valuable Buildings in Egypt. *Complexity* **2020**, *2020*, 1–11. [CrossRef]
19. Sadri, E.; Harsej, F.; Hajiaghaei-Keshteli, M.; Siyahbalaii, J. Evaluation of the components of intelligence and greenness in Iranian ports based on network data envelopment analysis (DEA) approach. *J. Model. Manag.* **2022**, *17*, 1008–1027. [CrossRef]
20. El Sayed, M.A.; Ibrahim, A.B.; Pitam, S. A modified TOPSIS approach for solving stochastic fuzzy multi-level multi-objective fractional decision-making problem. *Opsearch* **2020**, *57*, 1374–1403. [CrossRef]
21. El Sayed, M.A.; Abo, S.M.A. A novel approach for fully intuitionistic fuzzy multi-objective fractional transportation problem. *Alex. Eng. J.* **2021**, *60*, 1447–1463. [CrossRef]
22. Elsisy, M.A.; El Sayed, M.A.; Abo, E.Y. A novel algorithm for generating Pareto frontier of bi-level multi-objective rough non-linear programming problem. *Ain Shams Eng. J.* **2021**, *12*, 2125–2133. [CrossRef]
23. Juan, D.C.; Yoshinori, S. Vehicle Routing with Shipment Consolidation. *Int. J. Prod. Econ.* **2020**, *227*, 167–181.

24. Ning, T.; An, L.; Duan, X. Optimization of cold chain distribution path of fresh agricultural products under carbon tax mechanism: A case study in China. *J. Intell. Fuzzy Syst.* **2021**, *40*, 10549–10558. [CrossRef]
25. Huang, R.; Ning, J.; Mei, Z.; Fang, X.; Yi, X.; Gao, Y.; Hui, G. Study of delivery path optimization solution based on improved ant colony model. *Multimed. Tools Appl.* **2021**, *80*, 28975–28987. [CrossRef]
26. Jia, X. Research on the Optimization of Cold Chain Logistics Distribution Path of Agricultural Products E-Commerce in Urban Ecosystem from the Perspective of Carbon Neutrality. *Front. Ecol. Evol.* **2022**, *10*, 966111. [CrossRef]
27. Liu, X.; Peng, X.; Gu, G. Logistics Distribution Route Optimization Based on Genetic Algorithm. *Comput. Intell. Neurosci.* **2022**, *2022*, 8468438. [CrossRef]
28. Zhang, W.Q.; Li, H.R.; Yang, W.D.; Zhang, G.H.; Gen, M. Hybrid multiobjective evolutionary algorithm considering combi-nation timing for multi-type vehicle routing problem with time windows. *Comput. Ind. Eng.* **2022**, *171*, 108435. [CrossRef]
29. He, D. Intelligent Selection Algorithm of Optimal Logistics Distribution Path Based on Supply Chain Technology. *Comput. Intell. Neurosci.* **2022**, *2022*, 9955726. [CrossRef]
30. Daneshdoost, F.; Hajiaghaei, K.M.; Sahin, R.; Niroomand, S. Tabu Search Based Hybrid Meta-Heuristic Approaches for Schedule-Based Production Cost Minimization Problem for the Case of Cable Manufacturing Systems. *Informatica* **2022**, *33*, 499–522. [CrossRef]
31. Zheng, C.; Sun, K.; Gu, Y.; Shen, J.; Du, M. Multimodal Transport Path Selection of Cold Chain Logistics Based on Improved Particle Swarm Optimization Algorithm. *J. Adv. Transp.* **2022**, *2022*, 5458760. [CrossRef]
32. Fan, Q.; Nie, X.X.; Yu, K.; Zuo, X.L. Optimization of Logistics Distribution Route Based on the Save Mileage Method and the Ant Colony Algorithm. *Appl. Mech. Mater.* **2013**, *448–453*, 3683–3687. [CrossRef]
33. Sun, R.; Liu, M.; Zhao, L. Research on logistics distribution path optimization based on PSO and IoT. *Int. J. Wavelets Multiresolution Inf. Process.* **2019**, *17*, 1950051. [CrossRef]
34. Zhai, R. Solving the Optimization of Physical Distribution Routing Problem with Hybrid Genetic Algorithm. *J. Phys. Conf. Ser.* **2020**, *1550*, 022001. [CrossRef]
35. Wu, D.Q.; Wu, C.X. TDGVRPSTW of Fresh Agricultural Products Distribution: Considering Both Economic Cost and Environmental Cost. *Appl. Sci.* **2021**, *11*, 10579. [CrossRef]
36. Zhang, B. The Optimization of Distribution Path of Fresh Cold Chain Logistics Based on Genetic Algorithm. *Comput. Intell. Neurosci.* **2022**, *2022*, 4667010. [CrossRef]
37. Wang, X.; Cao, W. Research on optimization of distribution route for cold chain logistics cooperative distribution of fresh e-commerce based on price discount. *J. Phys. Conf. Ser.* **2021**, *1732*, 012041. [CrossRef]
38. Zhao, Z.; Li, X.; Zhou, X. Distribution Route Optimization for Electric Vehicles in Urban Cold Chain Logistics for Fresh Products under Time-Varying Traffic Conditions. *Math. Probl. Eng.* **2020**, *2020*, 9864935. [CrossRef]
39. Wu, F. Contactless Distribution Path Optimization Based on Improved Ant Colony Algorithm. *Math. Probl. Eng.* **2021**, *2021*, 5517778. [CrossRef]
40. Wang, Z.; Wen, P. Optimization of a Low-Carbon Two-Echelon Heterogeneous-Fleet Vehicle Routing for Cold Chain Logistics under Mixed Time Window. *Sustainability* **2020**, *12*, 1967. [CrossRef]
41. Xia, Y.K.; Fu, Z. Improved tabu search algorithm for the open vehicle routing problem with soft time windows and satis-faction rate. *Clust. Comput.* **2019**, *22*, 8725–8733. [CrossRef]
42. Liu, H. Optimization of Dairy Distribution Path Based on Genetic Algorithm. *J. Phys. Conf. Ser.* **2019**, *1345*, 042054. [CrossRef]
43. Olaniyi, O.S.; James, A.K. On the Application of a Modified Genetic Algorithm for Solving Vehicle Routing Problems with Time Windows and Split Delivery. *IAENG Int. J. Appl. Math.* **2022**, *52*, 1–9.
44. Zhang, Y.; Jiang, T.; Sun, J.; Fu, Z.; Yu, Y. Sustainable Development of Urbanization: From the Perspective of Social Security and Social Attitude for Migration. *Sustainability* **2022**, *14*, 10777. [CrossRef]
45. Dou, S.; Liu, G.; Yang, Y. A New Hybrid Algorithm for Cold Chain Logistics Distribution Center Location Problem. *IEEE Access* **2020**, *8*, 88769–88776. [CrossRef]

agriculture

MDPI

Review

Smart Farming: Internet of Things (IoT)-Based Sustainable Agriculture

Muthumanickam Dhanaraju [1,*], Poongodi Chenniappan [2], Kumaraperumal Ramalingam [1], Sellaperumal Pazhanivelan [3] and Ragunath Kaliaperumal [3]

[1] Department of Remote Sensing and GIS, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamilnadu, India
[2] Department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Sathyamangalam 638 401, Tamilnadu, India
[3] Water Technology Centre, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamilnadu, India
* Correspondence: muthumanickam.d@tnau.ac.in

Abstract: Smart farming is a development that has emphasized information and communication technology used in machinery, equipment, and sensors in network-based hi-tech farm supervision cycles. Innovative technologies, the Internet of Things (IoT), and cloud computing are anticipated to inspire growth and initiate the use of robots and artificial intelligence in farming. Such groundbreaking deviations are unsettling current agriculture approaches, while also presenting a range of challenges. This paper investigates the tools and equipment used in applications of wireless sensors in IoT agriculture, and the anticipated challenges faced when merging technology with conventional farming activities. Furthermore, this technical knowledge is helpful to growers during crop periods from sowing to harvest; and applications in both packing and transport are also investigated.

Keywords: crop management; sustainable agriculture; smart farming; internet-of-things (IoT); advanced agriculture practices; issues and problems

Citation: Dhanaraju, M.; Chenniappan, P.; Ramalingam, K.; Pazhanivelan, S.; Kaliaperumal, R. Smart Farming: Internet of Things (IoT)-Based Sustainable Agriculture. *Agriculture* **2022**, *12*, 1745. https://doi.org/10.3390/agriculture12101745

Academic Editors: Gniewko Niedbała and Sebastian Kujawa

Received: 7 September 2022
Accepted: 12 October 2022
Published: 21 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Sustainable agriculture is a measure of the endurance and sustenance of food grains produced in an eco-friendly manner [1]. Sustainable agriculture helps in the encouragement of farming practices and approaches to help sustain farmers and resources. It is economically feasible and maintains soil quality, reduces soil degradation, saves water resources, improves land biodiversity, and ensures a natural and healthy environment [2]. Sustainable agriculture plays a significant role in preserving natural resources, halting biodiversity loss, and reducing greenhouse gas emissions [3].

Sustainable agriculture farming is a method of preserving nature without compromising the future generation's basic needs, whilst also improving the effectiveness of farming. The basic accomplishments of smart farming in terms of sustainable agriculture are crop rotation, the control of nutrient deficiency in crops, the control of pests and diseases, recycling, and water harvesting, leading to an overall safer environment. Living organisms depend on the nature of biodiversity, and are contaminated by waste emissions, the use of fertilizers and pesticides, degraded dead plants, etc. The emission of greenhouse gases affects plants, animals, humans, and the environment; hence, it necessitates a better environment for living things [4] (Figure 1).

Agriculture is the largest contributor in India, with an 18% gross domestic product involving approximately 57% of people in rural areas. Over the years, although India's total agronomic output has increased, the number of growers has fallen from 71.9% in 1951 to 45.1% in 2011 [5]. The Economic Survey 2018 revealed that the number of agricultural workers in the total workforce will drop to 25.7% in 2050. In rural areas, farming families gradually lose the next generation of farmers, overwhelmed by higher costs of cultivation,

low per capita productivity, inadequate soil maintenance, and migrations to a non-farming or higher remunerative occupation. Presently, the world is on the verge of a digital revolution, and so it is the appropriate time to connect the agricultural landform with wireless technology to introduce and accommodate digital connectivity with farmers.

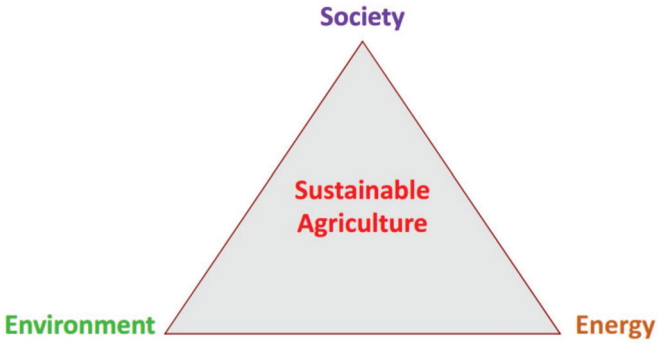


Figure 1. Factors of sustainable agriculture.

Regrettably, not all parts of the Earth's surface are suitable for agriculture due to various restrictions, such as: soil quality, topography, temperature, climate, and most relevant cultivable areas are also not homogenous [6]. Further, existing farming land is fragmented by political and fiscal features, and rapid urbanization, which consistently increases pressure on arable land availability (Figure 2). Recently, total agricultural land used for food production has declined [7]. Furthermore, every crop field has different critical characteristics, such as soil type, flow of irrigation, presence of nutrients, and pest resistance, which are all measured separately both in quality and quantity regarding a specific crop. Both spatial and temporal differences are necessary for optimizing crop production in the same field by crop rotation and an annual crop growth development cycle [8].

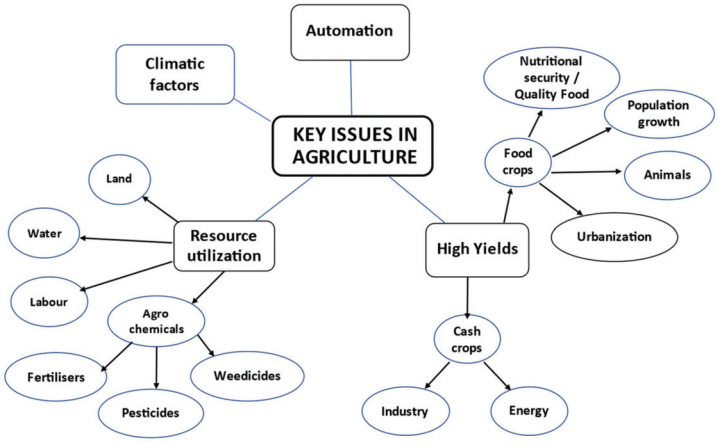


Figure 2. Key issues of technology in the agriculture industry.

In most cases, variations in characteristics occur within a single crop, or the same crop is grown on the whole farm and requires site-specific analyses for optimum yield production. New technology-based approaches are needed to produce more from less land, and to address these various issues. In traditional farming practices, farmers frequently

visit their fields throughout the crop's life in routine farming activities to better understand the crop conditions [9]. The current sensor and communication technologies offer an precise view of the field, from which farmers can detect ongoing field activities without being in the field in person. Wireless sensors monitor the crops with higher accuracy and detect issues at early stages, often facilitating the use of smart tools from initial sowing to the harvest of crops [10].

The timely use of sensors has made the entire farming operation smart and cost-effective, due to precise monitoring. The various autonomous harvesters, robotic weeders, and drones have sensors attached to collect data over short intervals. However, the vastness of agriculture puts extreme demands on technological solutions for sustainability with minimum ecological impact. Sensor technology through wireless communication helps farmers to know the various needs and requirements of crops without being in the fields, and they are then able to take remote action [11].

2. Smart Farming

Historically, ancient agriculture practices were related to the production of food in cultivated lands for the survival of humans and the breeding of animals [12], and was called the traditional agricultural era 1.0. This mainly resorted to using manpower and animals. Simple tools were used for farming activities, such as sickles and shovels. Work was mainly conducted through manual labor, and subsequently, productivity continued at a low level (Figure 3).

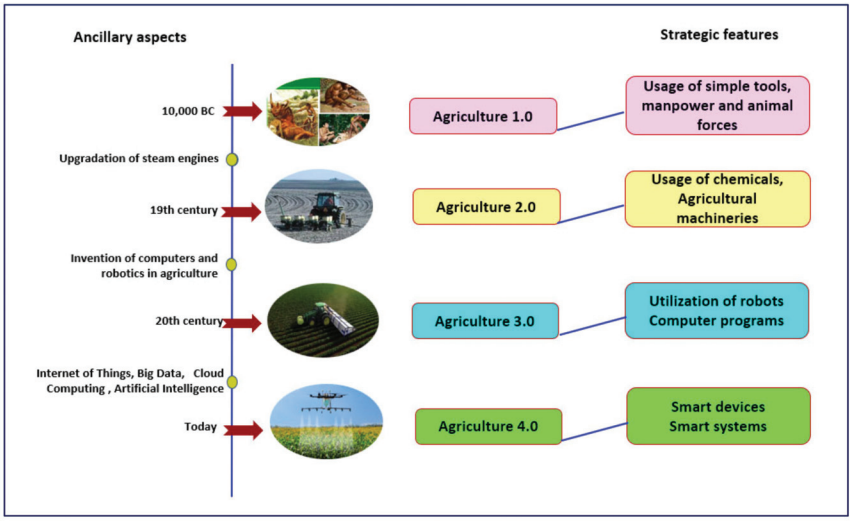


Figure 3. Agricultural decision support system framework.

During the 19th Century, new types of machinery appeared in the agricultural industries, in the form of steam engines. The wide use of agricultural machinery and abundant chemicals by farmers signaled the start of the agricultural era 2.0, and outwardly improved effectiveness and productivity of farmers and farms. However, considerably harmful implications, such as chemical pollution, environmental devastation, waste of natural resources, and excess utilization of energy, simultaneously developed.

The agricultural era 3.0 emerged during the 20th Century, due to the rapid growth of computation and electronics. Robotic techniques, programed agricultural machinery, and other technologies enhanced the agricultural processes efficiently. The issues that had arisen during agricultural era 2.0 were solved, and policies were readapted to the agricultural era

3.0 through work distribution, precise irrigation, the reduced use of chemicals, site-specific nutrient application, and efficient pest control technologies, etc.

The next agricultural era is also the current iteration of agriculture, the agricultural era 4.0, involving the engagement of recent technologies, such as the Internet of Things, big data analysis, artificial intelligence, cloud computing and remote sensing, etc. The adoption of new technologies has significantly improved agricultural activities by developing low-cost sensor and network platforms, aimed towards the optimization of production efficiency, along with reductions in the usage of water resources and energy with minimum environmental effects [13]. Big data in smart farming provides extrapolative overviews of real-time agricultural situations, allowing farmers to make effective decisions [14]. Real-time programming is developed with artificial intelligence concepts and embedded in IoT devices, helping farmers make the most suitable decisions [15].

Smart farming promotes precision agriculture with modern, sophisticated technology and enables farmers to remotely monitor the plants. Smart farming helps agricultural processes, such as harvesting and crop yields, as the automation of sensors and machinery has made the farming workforce more efficient [16]. The technologies convert traditional farming methods to automatic devices, causing a technological revolution in agriculture. Today, the technology in agriculture has altered the way farming is conducted, and conventional techniques have been transformed by the Internet of Things [17].

In terms of optimizing farm labor requirements and increasing the quantity and quality of products, smart farming is an emerging modern technique implemented with information and communication technologies (ICT) [16]. Modern ICT technologies, such as the Internet of Things, GPS (Global Positioning Systems), sensors, robotics, drones, precision equipment, actuators, and data analytics, are used to identify the farmers' needs and select suitable solutions to their problems. These innovations increase the accuracy and timeliness of decisions taken, and improve crop productivity. Several multilateral organizations and developing countries around the world have proposed smart farming technologies to increase agricultural output [18].

Sensors are constantly monitoring crops with greater accuracy, detecting any undesirable conditions during the early stages of the crop's lifecycle. Current farming incorporates smart tools from crop sowing to harvest, storing, and conveyance. The appropriate use of a wide variety of sensors has made the entire operation both more efficient and profitable, due to its accurate monitoring competencies. In addition, sensors that collect data quickly are directly available online for further evaluation, and they provide crop and site-specific agriculture for every site.

The many issues related to crop production are addressed by smart agriculture and monitoring, particularly regarding changes in soil characteristics, climate factors, soil moisture, etc., to improve the spatial management practices that increase crop production and avoid the excess use of fertilizers and pesticides [19]. The ANN models in smart irrigation water management (SIWM) regulate irrigation scheduling support systems (DSS) and offer data on irrigation efficiency, water productivity index, and irrigation water demand and supply on a real-time basis. Climate-smart agriculture (CSA) is an upcoming technology, especially in developing countries, due to its potential to improve food security, farm system resilience, and lower greenhouse gas emissions [20]. Smart agriculture technology based on IoT technologies has many advantages in all agricultural processes and practices in real-time, including irrigation, plant protection, improving product quality, fertilization, disease prediction, etc. [21]. The benefit of smart agriculture lies in its collection of real-time data on crops, the precise assessment of soil and crops, remote monitoring by farmers, supervising water and other natural resources, and improving livestock and agricultural production. Therefore, smart agriculture is considered to be the progression of precision agriculture through modernization and smart methods to attain various information of farm activities that are then remotely managed, and reinforced by suitable alternative real-time farm maintenance solutions.

3. Internet of Things

The Internet of Things (IoT) is a new technology that allows devices to connect remotely to achieve smart farming [22]. The IoT has begun to influence a vast range of industries, from health, trade, communications, energy and agriculture, to enhance efficiency and performance across all markets [23–25].

Current applications provide information on the IoT's effects, and its practices that are yet to be observed. However, by considering the advancement of technologies, one can envisage the IoT technologies perform a crucial role in numerous activities of farming, such as the utilization of communication infrastructure, data acquisition, smart objects, sensors, mobile devices, cloud-based intelligent information, decision-making, and the automation of agricultural operations (Figure 4).

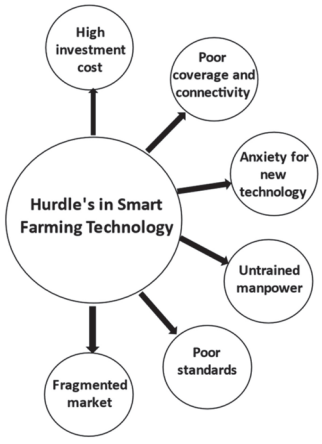


Figure 4. Barriers in the implementation of smart agriculture technology.

The IoT technology monitors plants and animals and remotely retrieves information from mobile phones and devices. Sensors and instruments empower farmers to assess the weather and to anticipate production levels. The IoT plays a role in water harvesting, monitoring and controlling the flow amount, assessing crops' water requirements, time of supply, and the saving of water, more than ever before [26]. Sensors and cloud connectivity through the gateway can remotely monitor the status and water supply based on soil and plant needs [27]. To correct nutrient deficiencies, pests, and diseases, farmers cannot monitor and observe every plant manually, but IoT technology is still beneficial and has led farmers to a new milestone in modern agriculture [28].

Recently, the development of IoT technologies has played a major role throughout the farming sector, particularly through its communication infrastructure. This has included connecting smart objects, remote data acquisition, using vehicles and sensors through mobile devices and the internet, cloud-based intelligent analysis, interfacing, decision formation, and the automation of agricultural operations. These proficiencies have revolutionized the agriculture industry in terms of resource optimization, controlling climate effects, and improving crop yields.

Researchers have proposed different methods, architectures, and various equipment to monitor and convey crop information at different growth stages, based on several crop and field types. Many manufacturers provide communication devices, multiple sensors, robots, heavy machinery, and drones to collect and then distribute data. Food and agriculture organizations, along with other government organizations, develop guidelines and policies for regulating the use of technologies to preserve food and environment safety [29,30].

Fundamentals of IoT Applications in Agriculture

The accessible, inexpensive and interactive tracking platform provides consolidated information on traditional agricultural methods, techniques, implements, crop pests and diseases, etc., collected from various sources for sustainable agriculture. Interactive agriculture allows easy access to the data by users through multiple devices, such as computers and mobile phones [31].

1. Robust Models: The distinctive features of the agriculture sector are diversity, complexity, spatio-temporal variability, and uncertainties of the right types of harvests and facilities.
2. Scalability: The variation in farm size from smaller to larger; hence, the results should be scalable. The placement and testing planning should be progressively scaled up with fewer expenses.
3. Affordability: Affordability is vital to farming achievement, and therefore price should be suitable with significant assistance. Standardized platforms, products, tools, and facilities could obtain a satisfactory price.
4. Sustainability: The problem of sustainability is a vital issue due to strong economic pressure and intense competition worldwide.

4. Technologies Used in Smart Farming

4.1. Global Positioning System (GPS)

GPS accurately records latitude, longitude, and elevation information [32]. Global Positioning System satellites transmit signals and permit GPS receivers to compute their location in real-time, and provide continuous positions while moving. The exact location information offers farmers the opportunity to discover the precise position of field data, such as pest occurrence, type of soil, weeds, and other barriers. The system facilitates the recognition of various field locations in order to then apply the necessary inputs (seed, fertilizer, herbicide, pesticide, and water) to a particular field [33].

4.2. Sensor Technologies

Technics, such as photo electricity, electromagnetics, conductivity, and ultrasound, are used to estimate soil texture and structure, nutrient level, vegetation, humidity, vapor, air, temperature, etc. Remote sensing data can differentiate between crop types, categorize pests and weeds, locate stress in soil and plant conditions, and monitor drought [34].

Plant health is affected by many factors, such as soil moisture, nutrient availability, exposure to light, humidity, the amount of rainfall, the color of leaves, etc. The plants are monitored by maintaining the optimum temperature and light intensity, and conserving water and energy through micro-irrigation. Different sensors are used to detect many parameters. If they cross a threshold, the sensor senses the changes and transmits them to the microcontroller to perform the required actions until the parameter returns to its optimum level [1].

The temperature, humidity, soil pattern monitoring, airflow sensor, location, CO_2, pressure, light, and moisture sensors are generally used in sensing technologies. Prominent sensor characteristics include reliability, memory, portability, durability, coverage, and computational efficiency, and make them suitable for agriculture [35]. Currently available wireless sensors play a vital role in collecting data on crop conditions and providing other information. These sensors are standalone types and can be integrated with advanced agricultural tools and heavy machinery, based on application necessities.

The major sensor types with their corresponding working procedure and purpose are represented in Table 1.

Table 1. Sensor types and their applications.

Sensors	Applications	Working Procedure
Acoustic sensors	Pest monitoring and detection classifying seed varieties, fruit harvesting [36].	Measuring the variations in noise level when intermingling with other materials, i.e., soil particles [37].
Airflow sensors	Measuring soil air permeability, moisture, and structure in a static position or mobile mode [38].	Based on various soil properties, unique identifying signatures [38].
Eddy covariance-based sensors	Quantifying exchanges of CO_2, water vapor, methane, or other gases. Measuring surface atmosphere and trace gas fluxes in various agricultural ecosystems [39].	Measuring continuous flux over large areas [40].
Electrochemical sensors	To analyze soil nutrient levels and pH [41].	Nutrients in soil, salinity, and pH are measured using sensors [42]
Electromagnetic sensors	Recording electrical conductivity, electromagnetic responses, residual nitrates, and organic matter in soil [43].	Electrical circuits measure the capability of soil particles to conduct or accumulate electrical charge [44].
Field programmable gate array (FAAA) based sensors	Measuring real-time plant transpiration, irrigation, and humidity [45].	Programmable silicon chips and logic blocks are surrounded together by programmable interconnected resources of the digital circuit [46].
Light detection and ranging (LIDAR)	Land mapping, soil type determination, farm 3D modelling, erosion monitoring and soil loss, and yield forecasting [47].	Sensors emit pulsed light waves and bounce off when colliding with objects and are returned to the sensor. The time taken for each pulse to return is used for assessment [47].
Mass flow sensors	Yield monitoring based on the amount of grain flow through a combine harvester [48].	Sensing the mass flow of grain with modules, e.g., grain moisture sensor, data storage device, and an internal software [48]
Mechanical sensors	Soil compaction or mechanical resistance	Sensors record the force assessed by strain gauges or load cells [48].
Optical sensors	Soil organic substances, soil moisture, color, minerals, composition, clay content, etc. Fluorescence-based optical sensors are used to supervise fruit maturation [49]. Integrating optical sensors with microwave scattering to characterize orchard canopies [50]	Sensors use light reflectance phenomena to measure changes in wave reflections [44].
Optoelectronic sensors	Differentiate plant types to detect weeds in wide-row crops [51].	Sensors differentiate based on reflection spectra [51].
Soft water level-based (SWLB) sensors	Used in catchments to characterize hydrological behaviors (water level and flow, time-step acquisitions) [52]	Measuring rainfall, stream flow, and other water presence options [52].
Telematics sensors	Assessing location, travel routes, and machine and farm operation activities [53].	Telecommunication between places (especially inaccessible points) [53].
Ultrasonic ranging sensors	Tank monitoring, spray distance measurement, uniform spray coverage, object detection, monitoring crop canopy [54], and weed detection [55].	An ultrasonic sensor uses a transducer to send and receive ultrasonic pulses that relay information about an object's proximity [56].
Remote sensing	Crop assessment, yield modeling, forecasting yield date, land cover and degradation mapping, forecasting, the identification of plants and pests, etc. [57].	Satellite-based sensor systems collect, process, and disseminate environmental data from fixed and mobile platforms [57].

4.3. Variable-Rate of Technology (VRT) and Grid Soil Sampling

Variable-rate technologies (VRT) are used in farming to predict the delivery rate of inputs based on a predetermined map extrapolated from GIS for the placement of inputs at variable amounts in the right place and at the right time [16,33]. Grid soil sampling is soil collection from a systematic grid to establish a map for every parameter. These maps are the basis for VRT and are loaded into a variable-rate applicator. The computer and GPS receiver direct and control the changes in the delivery amount or fertilizer product, based on map features [58,59].

New technologies, such as variable rate technology and associated practices (grid soil sampling), potentially improve soil fertility management and assess the spatial distribution of nutrients and yields [60]. In grid sampling, samples are collected from sampled sections based on the subdivision of a field into small areas, or cells, by superimposing the grid lines onto the field. Composite samples represent an entire area of each much smaller area (grid-point sampling) at the intersections of grid lines. Soil-test values from grid sampling are mapped by interpolating methods from non-measured locations between sampled points. The variability of phosphorus and potassium is field-specific, and each field should be fertilized differently to improve nutrient management practices by uniform applications of fertilizers and manure for better precision agriculture [61].

4.4. Geographic Information System (GIS)

The GIS comprises hardware and software designed to provide compilation, storage, retrieval, attributes analysis, and location data to generate maps and analyze characters and geography for statistics and spatial methods [62]. The GIS database provides information on field soil types, nutrient status, topography, irrigation, surface and subsurface drainage, quantity of chemical applications, and crop production, and also establishes the relationship between elements that affect a crop on a particular farming field [63]. Apart from data storage and display, the GIS is used to assess present and alternative management by compounding and altering data layers for decision-making.

4.5. Crop Management

Satellite images provide information on variations in soil conditions, as well as crop performances affected by topography within the field. Therefore, farmers can exactly monitor production factors, such as seeds, fertilizers, and pesticides, that are responsible for yield increase and efficiency.

The spatial coverage and temporal revisit frequency of satellite images provide the information in near real-time at a regional scale. The relationship between the spectral properties of crops and their biomass/yield experiments [64] is predicted by spectral reflectance properties of vegetation, especially in red and near-infrared combinations (vegetation indices) to monitor green foliage. Among the different indices, the normalized difference vegetation index (NDVI) is the most popular indicator to assess vegetation health and crop production, due to the closely related leaf area index (LAI) and photosynthetic activity of green vegetation [25]. Crop monitoring methods are based on the interpretation of remote-sensing-derived indicators by comparing actual crop status to previous or normal seasons [65]. The relationship between vegetation indices and biomass permits early crop yield estimation in certain periods before harvest [66]. The automated data acquisition, processing, monitoring, decision-making, and management of farm operations [67], including the basic functions of crop production (yields), profits and losses, farm weather prediction, field mapping, soil nutrients tracking, are the more complicated functionalities available through automated field management.

4.6. Soil and Plant Sensors

Sensor technology, a significant constituent of precision agriculture, provides soil properties information, fertility, and water status. Hence, new sensors have been developed based on desirable features and established apart from currently available sensors [68].

Soil sensors and plant wearables monitor real-time physical and chemical signals in soil, such as moisture, pH, temperature, and pollutants, and provide information to optimize crop growth conditions, fight against biotic and abiotic stresses, and increase crop yields. Soil organic matters (SOMs), nitrogen (N), phosphorus (P), and potassium (K) are the most important nutrients for crop production. The NIR reflectance-based sensors measure the spatial variation of surface and subsurface soil nitrogen [69]. SOM is predicted based on optimal wavelengths by assessing soil spectral reflectance in IR and visible wavelength regions [70]. The soil nitrogen and phosphorus are predicted using NIR spectrophotometry technology [71–73]. The soil apparent electrical conductivity (ECa) sensors collect information continuously on the field surface, since ECa is sensitive to changes in soil texture and salinity. Soil insects/pests are detected using optoelectronic, acoustic, impedance sensors, and nanostructured biosensors [74].

4.7. Rate Controllers

Rate controllers are designed to control the delivery rate of inputs by monitoring the speed of vehicles across the field, and altering the flow rate of material on a real-time basis at the target rate. Rate controllers are commonly used as stand-alone systems [75].

4.8. Precision Irrigation in Pressurized Systems

Recent developments in irrigation systems have introduced irrigation machines, devoted to motion control, GPS-based controllers, sensor technologies, and wireless communication to monitor soil and climatic conditions together with an assessment of irrigation parameters, i.e., flow and pressure, to attain greater water utilization efficiency by crop. These technologies show significant potential; however, further progress is required before they can become commercially available [76].

4.9. Yield Monitor

Yield monitors are the combination of sensors and components, including a data storage device, a computer, and user interface, that control integration and interaction components. The sensor measures yield continuously by evaluating the force of mass or volume of grain flow. The mass flow sensor was based on the principle of transmitting microwave energy beams and measuring the energy that bounces back after hitting. In yield monitors, GPS receivers create yield maps based on the location yield data [77].

The yield monitor is mounted on a harvester and connected with the mobile app for displaying live harvest data, and automatically uploads to the web-based platform. The app can generate and share high-quality yield maps with an agronomist, and farmers can export other farm management data for analysis. In horticultural crops, to precisely determine the yield quantity and quality of produce, fruit growth is considered one of the most relevant parameters in the crop progressing period [78]. Color images are used to track fruit conditions for estimating fruit maturation, making decisions for harvesting, and targeting the right market [79]. Satellite images are one of the options for real-time monitoring of the yield of crops over vast areas; for example, Sentinel-1A images are used to map the rice yield and crop intensity in Myanmar [80].

The crop yield estimation system was designed using both software and hardware components. Based on a Bluetooth terminal android application and yield estimator software program, crop yield is estimated using a mathematical calculation through a mobile application [81]. Satellite-based crop yield predictions based on spectral signatures reveal the estimated yields are as reliable as actual yields. The maize yield predictions were successfully carried out under varying environments using machine learning and satellite-derived data assimilation in crop models [82].

4.10. Software

The software has multiple tasks, such as mapping, display controller interfacing, data processing, analysis, and interpretation, etc. Most commonly, software is used to generate

the maps for soil properties and nutrient status, yield maps, variable rate applications maps for inputs, and overlaying different kinds of maps with advanced geostatistical features [83].

5. Applications in Agriculture

By adopting the current sensor and IoT technologies in agriculture, each characteristic of conventional farming practices is rehabilitated. The incorporation of wireless sensors and IoT in smart farming answers many of the issues facing conventional agriculture; for example, land suitability, drought monitoring, irrigation, pest control, and yield maximization. Figure 5 demonstrates the order of main applications, facilities, and devices for smart agriculture applications. Using advanced technologies at various stages in the following few applications enhances efficiency and revolutionizes agriculture.

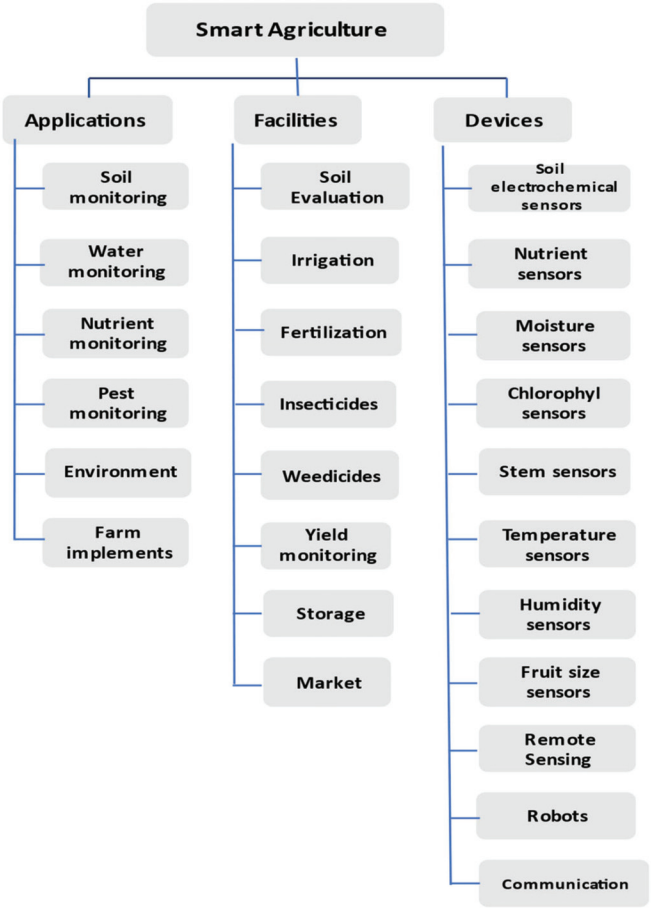


Figure 5. Hierarchy of probable applications, facilities and devices for smart agriculture.

5.1. Soil Mapping and Plant Monitoring

Soil analysis estimates the nutrient status of the field based on GPS position and field-specific information, and critical decisions are then taken according to the nutrient deficiencies at different stages of the crop. The factors controlling soil fertility status are topography, type and texture, cropping pattern, application of fertilizer, irrigation, etc. [84]. Soil mapping is useful for assessing crop suitability or varieties in a specific field, as well as planting depth, the physical, chemical and biological properties of soil, in order to best utilize resources. Presently, a wide range of sensors and tools are used to monitor soil properties, such as water-holding capacity, texture, and absorption rate, which assists farmers in tracking the soil quality and adopting suitable remedies to avoid soil degradation such as erosion, alkalization, acidification, salinization, and pollution. Drought is another concern that affects plant productivity and crop yield. Remote sensing techniques that can obtain soil moisture data frequently assist in analyzing agricultural drought in remote regions. Soil moisture maps generated from satellite data are used to estimate the soil water deficit index (SWDI), which enables the development of prediction models based on soil physical properties [85,86].

Various factors, such as soil type, soil nutrients, irrigation, and pests, affect rice yield and quality. The IoT-based mobile application aids crop management and provides real-time information on soil nutrition and characteristics. The system consists of electrical conductivity (EC), temperature sensors with a T-Beam microcontroller, and IoT connectivity, and the estimated EC value near the calibration solution is 12.88 mS/cm, and 150 mS/cm is less than 2% of the calibration solution's value. The measured EC values are linearly proportional to temperature and depth, and values of 1.04 and 3.86 mS/cm were noticed with and without fertilizer at 5 cm depth, while it was 0.656 and 420 mS/cm at 10 cm depth, respectively [87].

Plant monitoring conducted through the IoT ADCON-based station, with sensors and mobile devices (smartphones and tablets), farmers are able to collect data on soil and ambient parameters, such as leaf wetness, air and soil temperature, soil and air humidity to improve the grape productivity, and crop quality from seeding to harvest. Further, the data transmission system highlights the soil-plant-atmosphere interactions needed to optimize agricultural production [88]. By analyzing the data from soil moisture, carbon dioxide, light, and temperature sensors in bell peppers grown in a greenhouse were compared with day and night CO_2, rolling the doors and windows of the greenhouse open and closed, based on soil moisture [89].

5.2. Irrigation

According to the UN Convention to Combat Desertification (UNCCD), 168 countries will be inundated with desertification by 2030, and nearly 50% of the world population lives in high water shortage areas [90]. Considering the water crises and increasing demand for farming and other activities, it must be provided to regions with water quantities. Water resources are conserved by adopting more controlled and efficient irrigation systems; for example, drip and sprinkler irrigations. Water demand estimation for crops is controlled by soil type, precipitation, irrigation method, crop type, and requirement, as well as soil moisture retention. Using air and soil moisture control systems with wireless sensors optimizes water resources and improves crop health. In the current scenario, a substantial increase in crop productivity is anticipated using IoT techniques, namely CWSI (crop water stress index)-based water management [91], calculated from the crop canopy at varying crop growth stages and air temperatures. The information from climate data, sensors, and satellite imaging are related to the CWSI model for water requirement calculation, and predictions using the irrigation index values can be used for every field based on slope or soil variability to improve water usage efficiency.

5.3. Site-Specific Nutrient Management

Fertilizer is a either natural or synthetic chemical substance that provides nutrients for plant growth and soil fertility. Both nutrient deficiency and excessive fertilizer use harms soil, plant health, and the environment [92]. The site-specific soil nutrient fertilization under smart agriculture estimates the required quantity of nutrients precisely, and minimizes their negative effects through excessive use on soil and in the environment. The site-specific soil, nutrient measurements are influenced by soil types, crop type, yield targets, exchange capacity, use efficiency, the type of fertilizer, weather conditions, etc. The IoT-based fertilizing technique estimates the nutrient's spatial patterns of distribution [93,94]. The normalized difference vegetation index (NDVI) was obtained from satellite images to observe crop nutrient status [95,96], crop health, vegetation vigor, and plant density, as well as soil nutrient level. Recent technologies, like GPS [97], geo mapping [98], variable rate technology (VRT) [99,100], and autonomous vehicles [101] strongly contribute to IoT-based smart fertilization. Apart from these, fertigation [102] and chemigation [103,104], i.e., the use of water-soluble fertilizers in soil amendments and pesticides, are considered effective management practices to improve fertilization efficiency.

5.4. Crop Pest and Disease Management

The Food and Agriculture Organization (FAO) concluded that an annual global crop yield loss of 20- 40% was only due to pests and diseases [105], and these losses are controlled by the use of pesticides and other agrochemicals [106]. Most of them are harmful to human and animal health, and ultimately cause contamination of environmental systems [107,108]. The IoT-based devices, such as robots, wireless sensors, and drones, precisely spot and control the crop opponents by real-time monitoring, modelling, and disease forecasting, increasing overall effectiveness [109,110] more than traditional pest control procedures. The IoT-based disease and pest management process depends on detection and image processing. The remote sensing imagery and field sensors are used to collect data, such as plant health and pest incidence, in every field for the entire crop period. IoT-based automated traps [111,112] capture, count, and describe insect types, and further upload data to the Cloud for complete analysis. Due to advancements in robotic technology, an agricultural robot with multispectral image sensing devices and precision spraying nozzles is utilized to detect and control pest problems more accurately under the IoT management system.

5.5. Yield Monitoring and Forecasting

The yield monitoring mechanism conforms to yield, moisture content, and quality of produce. The quality depends on pollination with good pollen, especially under changing environmental circumstances [113–115]. Crop forecasting predicts the yield before the crop harvest, and assists the farmer in future planning, decision-making, and further analysis of the yield quality. Maturity determines the right harvesting time by monitoring the crop at different development stages, including factors such as fruit color, size, etc. Predictions of the correct harvesting time aids in maximizing crop quality and production, and regulates market management strategies. Therefore, farmers should know the exact harvest time of crops to obtain profit. Figure 6 outlines the idea of a farm area network, representing the whole farm in real-time conditions.

The development and installation of a yield monitor [116] on a harvester, connected with a mobile app, shows real-time crop harvest, and automatically transmits data to the manufacturer's web-based platform. To estimate crop production and monitoring, satellite images are exploited to cover vast areas [80]. For fruit crops, multicolor (RGB) satellite images [79] are utilized to track the diverse fruit conditions, especially fruit size and color, and plays a major role in estimating its maturation, making decisions on harvest, and market opportunities. Similarly, multiple optical sensors are used [117] to monitor shrinking fruits during drying conditions.

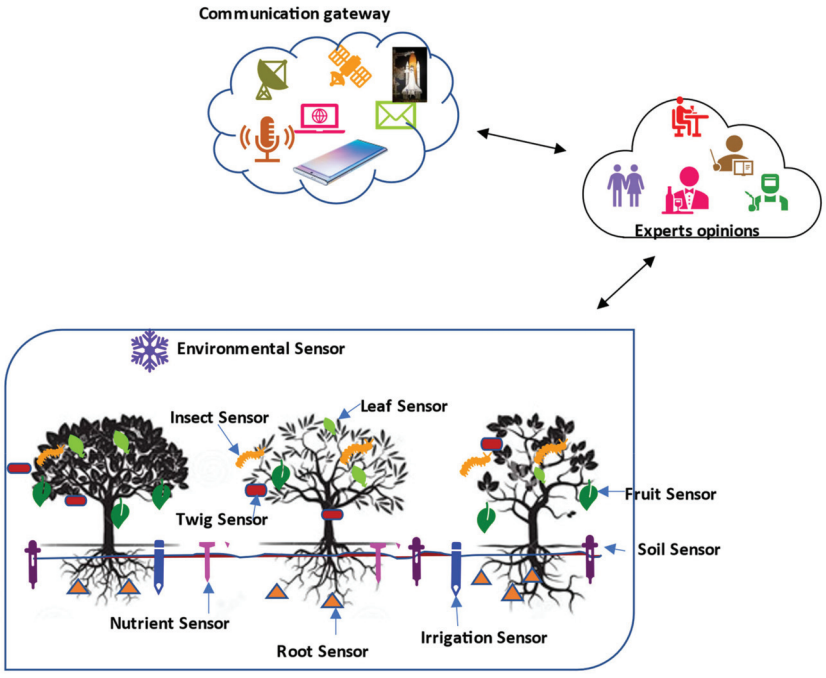


Figure 6. An Internet-of-Things-based network for smart farming.

6. Role of IoT in Advanced Farming Practices

Adopting the new methods based on sensor and IoT-based technologies improved the yield of crops more than conventional agriculture processes. The involvement of new sophisticated sensor-based technologies in controlled environments plays an important role in enhancing the quality and quantity of produce.

6.1. Greenhouse Farming and Protected Cultivation

Growing plants in a controlled environment gained popularity in the 19th Century, and is considered one of the oldest methods of smart farming. These practices further accelerated during the 20th Century in countries facing severe weather conditions [118]. Crops grown in indoor conditions are less affected by the environment. As a result, crops grown traditionally under suitable conditions are today being raised at anytime and anywhere by the use of sensors and communication devices. The success of crop production under a controlled environment depends on various factors, such as shed structures and material for controlling wind effects, aeration systems, accuracy of monitoring parameters, decision support system, etc. [119]. One of the greatest challenges in greenhouses is the precise monitoring of environmental parameters; hence, it requires several measurement points to predict the various parameters for controlling and ensuring the local climate. In an IoT-based greenhouse, sensors are used to measure and monitor the internal parameters, such as humidity, temperature, light, and pressure [120].

The smart greenhouse has helped farmers automatically conduct farm work, without manual inspection, and protects the plants from hailstorms, winds, ultraviolet radiation, and insect and pest attacks. Hibiscus plants are grown with the required wavelength during the night using lights, temperature, and air humidity sensors. A study revealed a reduction in 70–80% water requirement, and the IoT enables direct contact between the farmer and consumer to make farming as efficient and profitable as possible [121]. The IoT-enabled automated system increased the productivity of rose plants grown in a greenhouse by

monitoring and controlling various parameters, such as humidity, mist, CO_2 level, UV light intensity, pH and EC value, water nutrients solution level, temperature, and amount of pesticides, through sensors for further efficient detection and diagnosis [122].

6.2. Hydroponics

Hydroponics, a subdivision of hydroculture, is growing plants without soil to improve greenhouse farming benefits. Hydroponic-based irrigation systems enable a balanced rate of application of dissolved nutrients in the water to crop roots as a solution. Presently, the available systems and sensors [123] detect a wide range of parameters and perform data analysis at predetermined intervals. Precise measurement and monitoring of nutrient content in solution is crucial for plant growth and considers its demands. On a real-time basis, the wireless-sensor-based prototype [124] has delivered a solution for soilless cultivation, and measures the concentration of numerous nutrients and water levels [125].

An automated smart hydroponics system integrated with IoT consists of three major components: input data, cloud server and output data. These monitor lettuce cultivation from anywhere through the internet by analyzing parameters, such as pH level, water, nutrient-rich water-based solution, room temperature, and humidity, on a real-time basis [126]. The hydroponic system of the deep flow technique is a method for cultivating plants by placing roots in deep water layers, and ensuring the continuous circulation of plant nutrient solution. The plant growth elements data, such as pH, temperature, humidity, and water level in the hydroponic reservoir, are acquired by sensors integrated into Raspberry Pi, and data are processed and monitored automatically on a real-time basis to ensure proper water circulation [127].

6.3. Vertical Farming

The industrial-based agricultural farming practices damage soil quality at a faster rate than nature can reconstruct. The alarming erosion rate and use of fresh water for agriculture has led to the reduction of arable land, and increased the overburden on present water reservoirs [128]. Vertical farming (VF) offers an opportunity to keep the plants in a precisely controlled environment, significantly reducing resource consumption and, at the same time, increasing production at varied times; and only a portion of the ground surface is needed depending on the number of stacks. VF is also extremely effective in higher yields and reducing water consumption compared to traditional farming [129]. The carbon dioxide measurement is the most critical parameter; hence, nondispersive infrared (NDIR) CO_2 sensors play a vital part in tracking and controlling the conditions in vertical farms.

6.4. Phenotyping

Phenotyping is an emerging crop engineering technique, relating plant genomics with ecophysiology and agronomy. The advancement of genetic and molecular tools is significant for crop breeding; however, quantitative analysis of crop behaviors, such as pathogen resistance, grain weight, etc., is inadequate due to the absence of effective technologies and efficient techniques. In this condition, [130] reported that plant phenotyping is highly useful in investigating the quantitative characteristics responsible for growth, resistance to various stresses, yield quality, and quantity. The sensing technologies and image-based phenotyping describe screening of biostimulants and an understanding of their mode of action [131]. IoT-based phenotyping is intended to observe the crop and related trait measurements, and offer facilities for the breeding of crops and digital agriculture [132]. The trait analysis algorithms and modelling support determine the relationships among genotypes, phenotypes, and their growing condition.

7. The Role of the Engineer in Smart Farming

Farmers face many issues when they adopt IoT-based agriculture. Therefore, engineers must develop solutions for specific problems related to smart farming techniques. An engineering role concerns the application and use of innovative technologies and methods

for precision agricultural machinery, and smart farming is a creative way to mechanize agricultural engineering through means different from conventional mechanization [133]. The concepts and synergy-based information are obtained from different technology areas, such as agricultural mechanization, mechatronics, instrumentation, control systems, and knowledge in artificial and computational intelligence [134]. Big data, satellite, and aerial images have revolutionized precision agriculture, and these new technologies increase production efficiency by creating a balance between productivity and environmental protection. As a system integrator, engineering combines technical experience and strong business skills in both the public and private sectors [135].

At the same time, engineering exploits the rewards of digital transformation in the entire agri-food chain, from day-to-day farming activities to supporting sales operations, logistics, and the maintenance of farm assets. For example, knowledge of the IoT, AI, mobile, precision farming technologies, remote sensing, advanced analytics, the Cloud, RPA, and blockchain technologies are necessary [136]. The data collected from the various types of machinery through sensors and other devices generates responses concerning cereals, viticulture, fruit, and vegetables, as well as soil and monitoring [137].

The use of digital technologies and control systems to automate production processes also reduces manual human intervention. The production process, from field to final product, is carried out by planning, organizing, and analyzing data received from machines. The data acquired are stored in historical archives and correlated with each other to retrieve useful information for products through traceable systems working based on radio-frequency signals [138].

7.1. Purpose

Purpose is based on the user's final requirement, and influences the monitoring of crops during the growth period. Sensors provide the IoT solutions to their problems. For example, the end-user is a corn farmer, faced by problems mainly concerning water usage and ensuring that a crop gets adequate water; therefore, water level and moisture monitoring sensors are accommodated to prevent water wastage.

7.2. Technology

Distance plays an important role in technology selection because the sensors collect data and send to the server; hence, similar technology cannot be used for varying distances. For example, radio frequency identification (RFID) or near field communication (NFC) and low power, wide area network (LPWAN) technologies could send data over a distance of hundreds or even thousands of meters.

7.3. Power Requirements

Most IoT solutions are spread across a large farm, so it is better to develop low-power applications. On the other hand, more data transmission requires huge data costs and power consumption; hence, designers need to consider developing cost-effective IoT solutions for farmers. Usually, engineers save costs with customized IoT-based farming solutions, and develop apps for sending the data less frequently.

7.4. Data Frequency

The end user's necessities are critical in deciding the number of sensors and data packets. Sometimes, a farmer does not require information frequently, but developer design an IoT application to function on a continual and real-time basis, with very high data frequency.

7.5. Placement of Sensors

Sensors are placed in such a way that they provide optimal performance, even if the farm has all the essential sensors with proper placement.

8. Barriers to Implementing Smart Farming Technologies

Technology adoption is a method with a certain level of heterogeneity factors that are affective [139]. Technology implemented in farming systems has provided accuracy, efficiency, and eased time pressures. Although smart farming increases the productivity of crops, there are still problems in adopting these technologies

8.1. Cost of Technology

Existing technologies minimize the workforce and perform extremely fast with high accuracy. Therefore, it is anticipated that machines would probably replace a human workforce in the near future. However, it is impossible, since many countries have experienced poverty wherever the workforce was the main source for the agriculture sector. The implementation of devices and technologies requires a huge amount of money; therefore, farmers face difficulties in terms of affordability when they look beyond conventional tools.

8.2. Lack of Financial Resources

Financial supporters could provide adequate loans to farmers if farmers did not get the anticipated yield, perhaps because unexpected calamities like drought, flood, pests, and diseases impacted the crops.

8.3. Literacy Status of Farmers

The education level among the farmers is one of the greater challenges in implementing technologies in developing countries. The knowledge needed encompasses educational and technical abilities to manage the tools. The level of education increases a farmers' aptitude to process information, and thus make decisions using smart farming technologies [140], facilitating farmers' use of computers [141]. Farmers in developing nations are mostly uneducated and unskilled because of a lack of desire to gain knowledge, or any new technology awareness [142]. Hence, it is a reason for farmers in choosing traditional farming over smart farming [143]. Farmers have considered that usage is too complex, sometimes incapable of recognizing the icons used in a mobile application as the farmers use general icons based on traditional understanding. Farmers need to be digitally literate to reinforce the advantages of smart farming technologies and, simultaneously, agri-tech companies should ensure farmers easily understand the limitations of the technology.

8.4. Lack of Integration between the Systems

Integration across systems is one of the areas where smart farming technologies needs to be advanced further by incorporating production, property management, and decision-making tools. The communication between academics and interdisciplinary groups must overcome the gap between agricultural and information science. More emphasis has been given to increasing user effectiveness during the development of an information system [144]. The basis for improved decision-making is based on the timely obtainability of superior quality data; hence, data must be integrated to generate information and knowledge.

8.5. Telecommunications Infrastructure

Farming activities mostly occur in rural areas more effectively in arable land than contaminated land. However, poor telecommunication infrastructure makes data transmission unreliable, especially through mobile phones and tablets. Smart farming necessitates a real-time connection with the internet to enable the use of information. In addition, various operation control systems, such as fertilizers, pesticides, and seed volume, requires high-quality internet connection to produce outcomes. Recently, with the expansion of mobile phones, rural producers have gained to access mobile internet; however, signal quality and input speed are limited.

8.6. Data Management

Farmers are facing problems in organizing and manipulating data obtained by the sensors. The weather stations are generating data; however, farmers do not recognize how to use the information and how to change the data into a more available form. Its complex systems, alongside issues of acceptability and usability, lead to incorrect calculations. Farmers, consultants, and others involved in the production process must provide greater accessibility to data and information in productive systems.

9. Current Challenges and Future Expectations

In the 2030 Agenda for Sustainable Development, the United Nations and international community established a goal to end hunger by 2030. Currently, the World Health Organization reports that more than 800 million people are facing food shortages worldwise [145]. In addition, the increasing global population is increasing the demand for quality food; therefore, food and cash crops could improve overall crop production.

Figure 7 represents the future challenges agriculture is anticipated to face in 2050. This illustration offers three major problems: (1) feeding 10 billion people, (2) limitations in the expansion of land, and (3) the reduction of greenhouse gases emissions. These challenges lead to new thinking about water scarcity, shrinking arable land, rural labor, climate conditions, and much more. The diminishment in rural populations due to urbanization is not only shrinking communities, but is also leading to ageing populations; therefore, younger growers must step forward to take responsibility. The generation shift and population imbalance create further implications for the workforce and production.

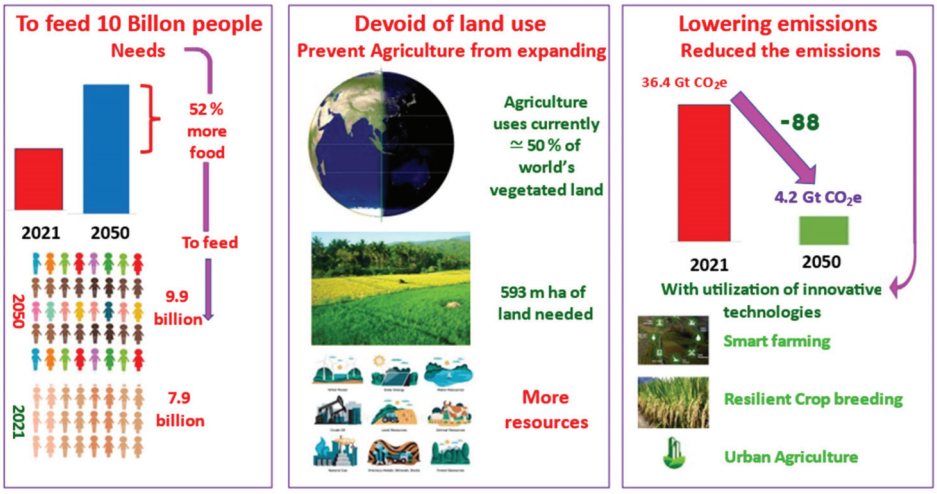


Figure 7. Challenges in sustainable future agriculture.

The further shrinking of arable land and the suitability of particular crops in specific regions are due to geographic and ecological conditions. Abrupt weather changes enhance the intensity of environmental issues, such as drought, groundwater depletion, and soil degradation, affecting crop production. Moreover, traditional agricultural methods have historically met food demands by employing fertilizers and pesticides; however, it increases food production only to a certain level and negligent use of chemical deteriorate the environment. In developing countries, various problems facing the agricultural sector include no suitable crop selection, soil testing, efficient irrigation systems, weather forecasting, animal husbandry, etc. Technological advancements have proved beneficial in developed countries, both quantitively and qualitatively, but, in developing countries, 50% of the population is already engaged in agriculture.

The future of agriculture is expected to interconnect with artificial intelligence and big data services. As a result, the systems will converge into a single unit, where farm machinery and management start from seeding to production forecasting. Few of the key technologies and methods are focused on achieving sustainable future agriculture.

9.1. Communication

The achievement of the IoT in agriculture mostly depends on connectivity between devices [146]. Most telecom operators provide connectivity services, but represent a small percentage of smart farming as a whole. Cellular operators offer new services to target growers and enhance market facilities, especially in rural areas. The success of cellular technology is feasible when service providers guarantee its benefits, such as flexibility, portability, and extravagance, of both-way communication at low cost. In developing countries, mobile services and smartphone technology offer a hopeful future for farmers to enhance their crop yields. The low power wide area technology (LPWA) is anticipated to play a major role in smart farming agriculture, due to its improved facilities, efficient coverage, low power consumption, and cost economics. The cellular operators with robust IoT create significant returns by offering smart agriculture facilities in collaboration with LPWA technology.

9.2. Wireless Sensors and IoT

Placing wireless sensors around the field provides timely information on a real-time basis to farmers in order to make decisions and act in order to obtain higher crop yields. Wireless sensor networks (WSNs) with GPS technology update all information on crop growth and terrain features. Recently, digital images and signal processing offer additional competencies to WSN, and precisely ascertain crop quality and health. The IoT technology can streamline tasks in a predictable manner by diagnosing crop requirements at each stage to maximize their effectiveness. In the future, IoT will be upgraded to the fifth-generation (5G) cellular mobile communication technologies, to provide real-time data to farmers at any time and everywhere. Based on this achievement, around 29 billion IoT-based components are expected to operate in the agriculture sector by the end of 2022. Furthermore, it is expected to create 4.1 million data points daily from farms by 2050 [147].

9.3. Drones and Unarmed Vehicles

Farmers widely use drones for crop growth monitoring, spraying nutrient solutions and water, and pesticides in tough terrains and for different crop heights. Drones have proven their value for spraying speed, area coverage, and precision, compared to traditional machinery. Due to advancements in technology, drones are equipped with various sensors, and 3D cameras provide comprehensive capabilities in land management by farmers. With the addition of UAVs in agriculture, many challenges, particularly the incorporation of technologies and use in inclement weather conditions, are addressed by farmers. Other than drones, robotics in agriculture has also enhanced productivity due to higher yields achieved by spraying and weeding without human intervention. The seeding, transplanting, and fruit harvest/picking robots have recently added a new efficiency level to traditional methods.

The UAV technology in smart agriculture provides information on fertilization, irrigation, use of pesticides, plant growth monitoring, weed management, crop disease management, and field-level phenotyping to enhance cultivation practices. A new method of 3D modeling has been used to monitor crop growth parameters to determine the height of maize and sorghum plants under field conditions using UAV, and the average root mean square error (RMSE) of sorghum height with hand sampling field data was 0.33 m [148]. The UAV and 3D models were also restored to extract leaf area index (LAI) in soybean plants, the measured LAI predicted accuracy corresponding to the handheld device (R^2 = 0.92) was correlated with destructive LAI measurements (R^2 = 0.89) [149].

Weed detection and management were assessed by integrating low-resolution multi-spectral high-resolution RGB images [150] using the Random Forest (RF) technique in field-grown rice and sugar beet crops [151]. Multi-spectral digital images obtained by UAVs are used for evaluating vegetation indices (VIs) and multi-temporal VIs to predict grain yield in wheat [152]. The indices, including the normalized difference vegetation index (NDVI), spectral vegetation index (SVI), and green area index (GAI), are evaluated in wheat crops to predict grain yield [153], monitor breeding [154], detect plant stress caused by yellow rust disease [155], and quantify plant density [156]. The usage of pesticides in agriculture is crucial for crop yields and the environment, and efforts have been made to develop and evaluate an algorithm to self-adjust UAV routes during chemical spraying in a crop field to reduce the waste of pesticides and fertilizers [157].

9.4. Vertical Farming and Hydroponics

The shrinking of arable land and rapid urbanization results in greater pressure on the present resources [158], which causes hardships for food production with current agriculture practices. Vertical farming (VF) navigates land and water shortage challenges, and is highly suitable for adoption in nearby cities. Hydroponics plays a key role in lowering water requirements. Hydroponics, along with VF, increases available arable land without distressing forests and other natural habitats. The presence of advanced technologies, especially the IoT, makes the agriculture industry highly remunerative with a reduction in labor requirements and other resources, in addition to minimizing environmental impact.

9.5. Performance Analysis Using Machine Learning

Data analytics and machine learning concepts are applied to analyze the real-time data. In crop production, identifying the best genes is an important process that can be conducted using machine learning techniques. In agriculture, machine learning is used to envisage the best genes suited for crop production, especially for selecting seed varieties that are highly suitable to specific climate conditions and locations. Machine learning algorithms identify high demand products and currently unavailable products. Recent developments in machine learning and analytics allow farmers to correctly categorize their harvests before it is processes and delivered to customers.

Machine learning (ML) in big data systems solves the issues related to farmers' decision-making, crops, animal research, land, food availability and security, weather and climate change, and weeds [159]. ML-based applications accommodate a large number of agricultural activities, such as yield prediction based on a deep memory model for maize [160], binary classification model with logistic regression technique to assess rainfall intensity [161], and a short-term memory model to predict soil water content with data parameters of rainfall, temperature, water diversion, evaporation, and time for the next 1, 2, and 7 days with greater R^2 compared to artificial neural networks [162]. As a result, the agricultural sector is increasing farmers' incomes, and so communities are further integrated into the agricultural value chain to reduce poverty and provide access to health care, education, and nutritious food for their families [163].

9.6. Renewable Energy, Microgrids and Smart Grids

Smart farming requires extensive energy due to power consumption by long-standing sensor placement, use of GPS, and data transmission. Traditionally, using renewable energy sources in remote areas solves long-term power issues. Smart grids and microgrids are integrated into distributed energy sources (DERs). Recent advances in storage devices combine electricity and heat systems to stock energy and use the heat produced.

Globally, smart grid technology enables a smooth transition from traditional to smart energy systems, ensuring energy security. In developing countries, power-strengthening systems integrated with renewable sources have enhanced the transport sector, and increased bioenergy use in the power sector through profuse renewable energy sources identified using smart technologies, such as, energy storage devices, smart appliances,

computational intelligence, and the IoT. For example, the smart grid provides a broad range of opportunities for power sector reform in Nepal, alleviating the rural electricity problem by implementing smart microgrids, and subsequently, connecting to the national grid [164]. The Dayalbagh renewable energy smart microgrid in India is a small-scale electricity system comprising distributed loads and renewable energy resources, acting as a single controllable entity in the grid. The smart microgrids are integrated into renewable resources and form building blocks of smart grids, especially for the dairy plant to produce various dairy products [165].

The mixed integer linear programme (MILP) systematically and efficiently managed energy consumption and subsequently lowered the cost, especially in residential areas, by scheduling the use of smart appliances and charging/discharging electric vehicles (EVs). The model generates its own energy from a microgrid containing solar panels and wind turbines, and forecasts wind speed and solar radiation for effective energy management. MILP-based energy planning sustains the effectiveness and productiveness of energy-efficient techniques [166].

10. Conclusions

Smarter and more efficient crop production methodologies are needed to address the issues of shrinking arable land and the food demands of an increasing world population. There is a necessity for everyone to be aware of food security in terms of sustainable agriculture. The growth of new technologies for increasing crop yield and encouraging the adoption of farming by innovative young people as a legitimate profession. This paper emphasized the role of many technologies used for farming, particularly the IoT, in making agriculture smarter and more effective in meeting future requirements. The current challenges faced by the industry and future prospects are noted to guide scholars and engineers. Hence, every piece of farmland is important to enhance crop production by dealing with every inch of land using sustainable IoT-based sensors and communication technologies.

Author Contributions: Conceptualization, M.D.; investigation, P.C., K.R.; methodology, M.D., S.P.; resources, K.R., R.K.; supervision, S.P.; visualization, R.K.; writing—original draft, M.D.; writing—review & editing, P.C.; funding acquisition, S.P., validation, R.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research and APC were funded by GIZ, Germany by Deutsche Gesellschaft für Internationale Zusammenarbeit (Grant number 81278637).

Conflicts of Interest: The authors declare that no competing financial interests or personal relationships could have appeared to influence the work reported in this paper.

References

1. Srisruthi, S.; Swarna, N.; Ros, G.M.S.; Elizabeth, E. Sustainable agriculture using eco-friendly and energy efficient sensor technology. In Proceedings of the 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 20–21 May 2016; IEEE: Bangalore, India, 2016; pp. 1442–1446. [CrossRef]
2. Brodt, S.; Six, J.; Feenstra, G.; Ingels, C.; Campbell, D. Sustainable Agriculture. *Nat. Educ. Knowl.* **2011**, *3*, 1.
3. Obaisi, A.I.; Adegbeye, M.J.; Elghandour, M.M.M.Y.; Barbabosa-Pliego, A.; Salem, A.Z.M. Natural Resource Management and Sustainable Agriculture. In *Handbook of Climate Change Mitigation and Adaptation*; Lackner, M., Sajjadi, B., Chen, W.Y., Eds.; Springer: Cham, Switzerland, 2022. [CrossRef]
4. Latake, P.T.; Pawar, P.; Ranveer, A.C. The Greenhouse Effect and Its Impacts on Environment. *Int. J. Innov. Res. Creat. Technol.* **2015**, *1*, 333–337.
5. Reddy, T.; Dutta, M. Impact of Agricultural Inputs on Agricultural GDP in Indian Economy. *Theor. Econ. Lett.* **2018**, *8*, 1840–1853. [CrossRef]
6. *World Agriculture: Towards 2015/2030: An FAO Perspective and Summary Report*; FAO: Rome, Italy, 2002; Available online: www.fao.org/3/a-y4252e.pdf (accessed on 1 August 2022).
7. Roser, M.; Ritchie, H.; Ortiz-Ospina, E. World Population Growth. 2013. Available online: https://ourworldindata.org/world-population-growth (accessed on 1 August 2022).

8. Hernández-Ochoa, I.M.; Gaiser, T.; Kersebaum, K.C.; Webber, H.; Seidel, S.J.; Grahmann, K.; Ewert, F. Model-based design of crop diversification through new field arrangements in spatially heterogeneous landscapes. A review. *Agron. Sustain. Dev.* **2022**, *42*, 74. [CrossRef]
9. Navulur, S.; Sastry, A.S.C.S.; Giri Prasad, M.N. Agricultural Management through Wireless Sensors and Internet of Things. *Int. J. Electr. Comput. Eng.* **2017**, *7*, 3492–3499. [CrossRef]
10. Ayaz, M.; Ammad-uddin, M.; Baig, I.; Aggoune, E.M. Wireless Sensor's Civil Applications, Prototypes, and Future Integration Possibilities: A Review. *IEEE Sens. J.* **2018**, *18*, 4–30. [CrossRef]
11. Lin, J.; Yu, W.; Zhang, N.; Yang, X.; Zhang, H.; Zhao, W. A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications. *IEEE Internet Things J.* **2017**, *4*, 1125–1142. [CrossRef]
12. Tekinerdogan, B. *Strategies for Technological Innovation in Agriculture 4.0. Reports*; Wageningen University: Wageningen, The Netherlands, 2018.
13. Ferrandez-Pastor, F.J.; Garcia-Chamizo, J.M.; Nieto-Hidalgo, M.; Mora-Pascual, J.; MoraMartinez, J. Developing ubiquitous sensor network platform using Internet of Things: Application in precision agriculture. *Sensors* **2016**, *16*, 1141. [CrossRef]
14. Wolfert, S.; Ge, L.; Verdouw, C.; Bogaardt, M.J. Big data in smart farming—A review. *Agric. Syst.* **2017**, *153*, 69–80. [CrossRef]
15. Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine learning in agriculture: A review. *Sensors* **2018**, *18*, 2674. [CrossRef]
16. O'Grady, M.J.; O'Hare, G.M.P. Modelling the smart farm. *Inf. Process. Agric.* **2017**, *4*, 179–187. [CrossRef]
17. Quy, V.K.; Hau, N.V.; Anh, D.V.; Quy, N.M.; Ban, N.T.; Lanza, S.; Randazzo, G.; Muzirafuti, A. IoT-Enabled Smart Agriculture: Architecture, Applications, and Challenges. *Appl. Sci.* **2022**, *12*, 3396. [CrossRef]
18. Raj Kumar, G.; Chandra Shekhar, Y.; Shweta, V.; Ritesh, R. Smart agriculture—Urgent need of the day in developing countries. *Sustain. Comput. Inform. Syst.* **2021**, *30*, 100512.
19. El Nahry, A.H.; Mohamed, E.S. Potentiality of land and water resources in African Sahara: A case study of south Egypt. *Environ. Earth Sci.* **2011**, *63*, 1263–1275. [CrossRef]
20. Palombi, L.; Sessa, R. *Climate-Smart Agriculture: Source Book*; Food and Agriculture Organization: Rome, Italy, 2013.
21. Adamides, G.; Kalatzis, N.; Stylianou, A.; Marianos, N.; Chatzipapadopoulos, F.; Giannakopoulou, M.; Papadavid, G.; Vassiliou, V.; Neocleous, D. Smart Farming Techniques for Climate Change Adaptation in Cyprus. *Atmosphere* **2020**, *11*, 557. [CrossRef]
22. Patil, K.A.; Kale, N.R. A model for smart agriculture using IoT. In Proceedings of the 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, Jalgaon, India, 22–24 December 2016; IEEE: Jalgaon, India, 2016; pp. 543–545. [CrossRef]
23. Sisinni, E.; Saifullah, A.; Han, S.; Jennehag, U.; Gidlund, M. Industrial Internet of Things: Challenges, Opportunities, and Directions. *IEEE Trans. Ind. Inform.* **2018**, *14*, 4724–4734. [CrossRef]
24. Shi, X.; An, X.; Zhao, Q.; Liu, H.; Xia, L.; Sun, X.; Guo, Y. State- of- the- Art Internet of Things in Protected Agriculture. *Sensors* **2019**, *19*, 1833. [CrossRef]
25. Elijah, O.; Rahman, T.A.; Orikumhi, I.; Leow, C.Y.; Hindia, M.N. An Overview of Internet of Things (IoT) and Data Analytics in Agriculture: Benefits and Challenges. *IEEE Internet Things J.* **2018**, *5*, 3758–3773. [CrossRef]
26. Yong, W.; Shuaishuai, L.; Li, L.; Minzan, L.; Ming, L.; Arvanitis, K.G.; Grorgieva, C.; Sigrimis, N. Smart Sensors from Ground to Cloud and Web Intelligence. *IFAC Pap. OnLine* **2018**, *51*, 31–38. [CrossRef]
27. Mekala, M.S.; Viswanathan, P. A Survey: Smart agriculture IoT with cloud computing. In Proceedings of the 2017 International Conference on Microelectronic Devices, Circuits and Systems (ICMDCS), Vellore, India, 10–12 August 2017; IEEE: Vellore, India, 2017; pp. 1–7. [CrossRef]
28. Mittal, A.; Singh, A. Microcontroller based pest management system. In Proceedings of the Second International Conference on Systems (ICONS'07), Martinique, France, 22–28 April 2007; IEEE: Martinique, France, 2007; p. 43. [CrossRef]
29. Bonneau, V.; Copigneaux, B. Industry 4.0 in Agriculture: Focus on IoT Aspects, European Commission, Digital Transformation Monitor. 2017. Available online: https://ec.europa.eu/growth/tools-databases/dem/monitor/content/industry-40-agriculture-focus-iot-aspects (accessed on 30 December 2020).
30. King, T.; Cole, M.; Farber, J.M.; Eisenbrand, G.; Zabaras, D.; Fox, E.M.; Hill, J.P. Food safety for food security: Relationship between global megatrends and developments in food safety. *Trends Food Sci. Technol.* **2017**, *68*, 160–175. [CrossRef]
31. Chandhini, K. A Literature Study on Agricultural Production System Using IoT as Inclusive Technology. *Int. J. Innov. Technol. Res.* **2016**, *4*, 2727–2731.
32. Lang, L. GPS + GIS + remote sensing: An overview. *Earth Obs. Mag.* **1992**, *1*, 23–26.
33. Batte, M.T.; VanBuren, F.N. Precision farming—Factor influencing productivity. In Proceedings of the Northern Ohio Crops Day Meeting, Wood County, OH, USA, 21 January 1999.
34. Chen, F.; Kissel, D.E.; West, L.T.; Adkin, W.; Clark, R.; Rickman, D.; Luvall, J.C. Field Scale Mapping of Surface Soil Clay Concentration. *Precis. Agric.* **2004**, *5*, 7–26. [CrossRef]
35. Muhammad, S.F.; Shamyla, R.; Adnan, A.; Tariq, U.; Yousaf, B.Z. Role of IoT Technology in Agriculture: A Systematic Literature Review. *Electronics* **2020**, *9*, 319. [CrossRef]
36. Srivastava, N.; Chopra, G.; Jain, P.; Khatter, B. Pest Monitor and Control System Using Wireless Sensor Network (With Special Reference to Acoustic Device Wireless Sensor). In Proceedings of the International Conference on Electrical and Electronics Engineering, Khartoum, Sudan Goa, 26–28 August 2013. ISBN: 978-93-82208-58-7.

37. Kong, Q.; Chen, H.; Mo, Y.L.; Song, G. Real-time monitoring of water content in sandy soil using shear mode piezoceramic transducers and active sensing-A feasibility study. *Sensors* **2017**, *17*, 2395. [CrossRef]
38. García-Ramos, F.J.; Vidal, M.; Boné, A.; Malón, H.; Aguirre, J. Analysis of the Air Flow Generated by an Air-Assisted Sprayer Equipped with Two Axial Fans Using a 3D Sonic Anemometer. *Sensors* **2012**, *12*, 7598–7613. [CrossRef]
39. Moureaux, C.; Ceschia, E.; Arriga, N.; Béziat, P.; Eugster, W.; Kutsch, W.L.; Pattey, E. Eddy covariance measurements over crops. In *Eddy Covariance: A Practical Guide to Measurement and Data Analysis*; Aubinet, M., Vesala, T., Papale, D., Eds.; Springer: Dordrecht, The Netherlands, 2012.
40. Kumar, A.; Bhatia, A.; Fagodiya, R.K. Eddy covariance flux tower: A promising technique for greenhouse gases measurement. *Adv. Plants Agric. Res.* **2017**, *7*, 337–340.
41. Yew, T.K.; Yusoff, Y.; Sieng, L.K.; Lah, H.C.; Majid, H.; Shelida, N. An electrochemical sensor ASIC for agriculture applications. In Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 26–30 May 2014; pp. 85–90.
42. Cocovi-Solberg, D.J.; Rosende, M.; Miro, M. Automatic kinetic bioaccessibility assay of lead in soil environments using flow-through micro dialysis as a front end to electrothermal atomic absorption spectrometry. *Environ. Sci. Technol.* **2014**, *48*, 6282–6290. [CrossRef]
43. Yunus, M.A.M.; Mukhopadhyay, S.C. Novel Planar Electromagnetic Sensors for Detection of Nitrates and Contamination in Natural Water Sources. *IEEE Sens. J.* **2011**, *11*, 1440–1447. [CrossRef]
44. Millan-Almaraz, J.R.; Romero-Troncoso, R.J.; Guevara-Gonzalez, R.G.; Contreras-Medina, L.M.; Carrillo-Serrano, R.V.; Osornio Rios, R.A.; Duarte-Galvan, C.; Rios-Alcaraz, M.A.; Torres Pacheco, I. FPGA-based fused smart sensor for real-time plant transpiration dynamic estimation. *Sensors* **2010**, *10*, 8316–8331. [CrossRef]
45. Weiss, U.; Biber, P. Plant detection and mapping for agricultural robots using a 3D-LIDAR sensor. *Robot. Auton. Syst.* **2011**, *59*, 265–273. [CrossRef]
46. Montagnoli, A.; Fusco, S.; Terzaghi, M.; Kirschbaum, A.; Pflugmacher, D.; Cohen, W.B.; Scippa, G.S.; Chiatante, D. Estimating forest aboveground biomass by low-density LiDAR data in mixed broad-leaved forests in the Italian Pre-Alps. *For. Ecosyst.* **2015**, *2*, 10. [CrossRef]
47. Schuster, J.N.; Darr, M.J.; McNaull, R.P. Performance benchmark of yield monitors for mechanical and environmental influences. In *Agricultural and Biosystems Engineering Conference Proceedings and Presentations*; IOWA State University: Ames, IA, USA, 2017.
48. Hemmat, A.; Binandeh, A.R.; Ghaisari, J.; Khorsandi, A. Development and field testing of an integrated sensor for on-the-go measurement of soil mechanical resistance. *Sens. Actuators A Phys.* **2013**, *198*, 61–68. [CrossRef]
49. Murray, S.C. Optical Sensors Advancing Precision In Agricultural Production. *Photonics Spectra* **2018**, *51*, 48.
50. Molina, I.; Morillo, C.; García-Meléndez, E.; Guadalupe, R.; Roman, M.I. Characterizing olive grove canopies by means of ground-based hemispherical photography and spaceborne RADAR data. *Sensors* **2011**, *11*, 7476–7501. [CrossRef]
51. Andújar, D.; Ribeiro, Á.; Fernández-Quintanilla, C.; Dorado, J. Accuracy and feasibility of optoelectronic sensors for weed mapping in wide row crops. *Sensors* **2011**, *11*, 2304–2318. [CrossRef]
52. Crabit, A.; Colin, F.; Bailly, J.S.; Ayroles, H.; Garnier, F. Soft water level sensors for characterizing the hydrological behaviour of agricultural catchments. *Sensors* **2011**, *11*, 4656–4673. [CrossRef]
53. Mark, T.; Griffin, T. Defining the Barriers to Telematics for Precision Agriculture: Connectivity Supply and Demand. In Proceedings of the SAEA Annual Meeting, San Antonio, TX, USA, 6–9 February 2016.
54. Dvorak, J.S.; Stone, M.L.; Self, K.P. Objct Detection for Agricultural and Construction Environments Using an Ultrasonic Sensor. *J. Agric. Saf. Health* **2016**, *22*, 107–119.
55. Pajares, G.; Peruzzi, A.; Gonzalez-de-Santos, P. Sensors in agriculture and forestry. *Sensors* **2013**, *13*, 12132–12139. [CrossRef]
56. Zhmud, V.A.; Kondratiev, N.O.; Kuznetsov, K.A.; Trubin, V.G.; Dimitrov, L.V. Application of ultrasonic sensor for measuring distances in robotics. *J. Phys. Conf. Ser.* **2018**, *1015*, 032189. [CrossRef]
57. Yalew, S.G.; van Griensven, A.; Mul, M.L.; van der Zaag, P. Land suitability analysis for agriculture in the Abbay basin using remote sensing, GIS and AHP techniques. *Model Earth Syst. Environ.* **2016**, *2*, 101. [CrossRef]
58. Berntsen, J.; Thomsen, A.; Schelde, K.; Hansen, O.M.; Knudsen, L.; Broge, N.; Hougaard, H.; Horfarter, R. Algorithms for sensor-based redistribution of nitrogen fertilizer in winter wheat. *Precis. Agric.* **2006**, *7*, 65–83. [CrossRef]
59. Ferguson, R.B.; Hergert, G.W.; Schepers, J.S.; Gotway, C.A.; Cahoon, J.E.; Peterson, T.A. Site-specific nitrogen management of irrigated maize; Yield and soil residual nitrate effects. *Soil Sci. Soc. Am. J.* **2002**, *66*, 544–553.
60. Fleming, K.L.; Westfall, D.G.; Bausch, W.C. Evaluating management zone technology and grid soil sampling for variable rate nitrogen application. In Proceedings of the 5th International Conference on Precision Agriculture, Bloomington, MN, USA, 16–19 July 2000; pp. 1–13.
61. Mallarino, A.P.; Wittry, D.J. Use of DGPS, yield monitors, soil testing and variable rate technology to improve phosphorus and potassium management. In Proceedings of the Integrated Crop Management Conference; Iowa State University Extension and Outreach: Ames, IA, USA, 1997; pp. 267–275.
62. Ehlers, M. Geoinformatics and digital earth initiatives: A German perspective. *Int. J. Digit. Earth* **2008**, *1*, 17–30. [CrossRef]
63. Ojo, O.I.; Ilunga, M.F. Geospatial Analysis for Irrigated Land Assessment Modeling and Mapping. In *Multi-Purposeful Application of Geospatial Data*; Rustamov, R.B., Ed.; IntechOpen: London, UK, 2018; pp. 65–84. [CrossRef]

64. Tucker, C.J.; Holben, B.N.; Elgin, J.H., Jr.; McMurtrey, J.E., III. Relationship of spectral data to grain yield variation. *Photogramm. Eng. Remote Sens.* **1980**, *46*, 657–666.
65. Muthumanickam, D.; Kannan, P.; Kumaraperumal, R.; Natarajan, S.; Sivasamy, R.; Poongodi, C. Drought assessment and monitoring through remote sensing and GIS in western tracts of Tamil Nadu, India. *Int. J. Remote Sens.* **2011**, *32*, 5157–5176. [CrossRef]
66. Felix, R.; Clement, A.; Igor, S.; Oscar, R. Using Low Resolution Satellite Imagery for Yield Prediction and Yield Anomaly Detection. *Remote Sens.* **2013**, *5*, 1704–1733.
67. Chowdhury, M.E.H.; Khandakar, A.; Ahmed, S.; Al-Khuzaei, F.; Hamdalla, J.; Haque, F.; Reaz, M.B.I.; Shafei, A.A.; Emadi, N.A. Design, Construction and Testing of IoT Based Automated Indoor Vertical Hydroponics Farming Test-Bed in Qatar. *Sensors* **2020**, *20*, 5637. [CrossRef]
68. Adamchuk, V.I.; Hummel, J.W.; Morgan, M.T.; Upadhyaya, S.K. On-the-go soil sensors for precision agriculture. *Comput. Electron. Agric.* **2004**, *44*, 71–91. [CrossRef]
69. Sudduth, K.A.; Hummel, J.W. Soil Organic Matter, CEC, and Moisture Sensing with a Portable NIR Spectrophotometer. *Trans. ASAE* **1993**, *36*, 1571–1582. [CrossRef]
70. Daniel, K.; Tripathi, N.K.; Honda, K.; Apisit, E. Analysis of spectral reflectance and absorption patterns of soil organic matter. In Proceedings of the 22nd Asian Conference on Remote Sensing, Singapore, 5–9 November 2011.
71. Kuang, B.; Mouazen, A.M. Non-biased prediction of soil organic carbon and total nitrogen with vis-NIR spectroscopy, as affected by soil moisture content and texture. *Biosyst. Eng.* **2013**, *114*, 249–258. [CrossRef]
72. Maleki, M.R.; Van Holm, L.; Ramon, H.; Merckx, R.; De Baerdemaeker, J.; Mouazen, A.M. Phosphorus Sensing for Fresh Soils using Visible and Near Infrared Spectroscopy. *Biosyst. Eng.* **2006**, *95*, 425–436. [CrossRef]
73. Lvova, L.; Nadporozhskaya, M. Chemical sensors for soil analysis: Principles and applications. In *Series Nanotechnology in the Agri-Food Industry*; New Pesticides and Soil Sensors; Grumezescu, A.M., Ed.; Elsevier: Amsterdam, The Netherlands, 2017; Volume 10, pp. 637–678. [CrossRef]
74. Potamitis, I.; Rigakis, I.; Tatlas, N.A.; Potirakis, S. In-Vivo Vibroacoustic Surveillance of Trees in the Context of the IoT. *Sensors* **2019**, *19*, 1366. [CrossRef]
75. Sushil, S.; Radha Mohan, S.; Manhas, S.S.; Shiv Kumar, L. Potential of Variable Rate Application Technology in India. *AMA Agric. Mech. Asia Afr. Lat. Am.* **2014**, *45*, 74–89.
76. Hassan, A.; Aitazaz, A.F.; Farhat, A.; Bishnu, A.; Travis, E. Precision Irrigation Strategies for Sustainable Water Budgeting of Potato Crop in Prince Edward Island. *Sustainability* **2020**, *12*, 2419. [CrossRef]
77. Naorem, A.; Rani, A.; Roy, D.; Kundu, S.; Rao, N.S.; Sreekanth, P.D.; Kumar, A.; Manjaiah, A.M.; Rao, C.S. Frontier Soil Technologies for Sustainable Development Goals (SDGs) in India. In *Challenges and Emerging Opportunities in Indian Agriculture*; Rao, C.S., Senthil, V., Meena, P.C., Eds.; National Academy of Agricultural Research Management: Hyderabad, India, 2019; pp. 113–152.
78. Luigi, L.M.; Emanuele, E.P.; Zibordi, M.; Morandi, B.; Muzzi, E.; Losciale, P.; Corelli, L.; Grappadelli, L.C. Monitoring Strategies for Precise Production of high quality Fruit and Yield in Apple in Emilia Romagna. *Chem. Eng. Trans.* **2015**, *44*, 301–306.
79. Wang, Z.; Walsh, K.B.; Verma, B. On-tree mango fruit size estimation using RGB-D images. *Sensors* **2017**, *17*, 2738. [CrossRef]
80. Torbick, N.; Chowdhury, D.; Salas, W.; Qi, J. Monitoring Rice Agriculture across Myanmar Using Time Series Sentinel-1 Assisted by Landsat-8 and PALSAR-2. *Remote Sens.* **2017**, *9*, 119. [CrossRef]
81. Mishachandar, B.; Vairamuthu, S. Crop Yield Estimation Using the Internet of Things. *J. Inf. Knowl. Manag.* **2021**, *20*, 2140006. [CrossRef]
82. Olipa, N.L.; Lydia, M.C.; Chabala1, S.; Chizumba, S. Satellite-Based Crop Monitoring and Yield Estimation—A Review. *J. Agric. Sci.* **2021**, *13*, 180–194.
83. Ferrández-Pastor, F.J.; García-Chamizo, J.M.; Nieto-Hidalgo, M.; Mora-Martínez, J. Precision Agriculture Design Method Using a Distributed Computing Architecture on Internet of Things Context. *Sensors* **2018**, *18*, 1731. [CrossRef]
84. Dinkins, C.P.; Jones, C. *Interpretation of Soil Test Results for Agriculture*; MontGuide. Publication no. MT200702AG; Montana State University Extension: Bozeman, MT, USA, 2013.
85. Martínez-Fernández, J.; González-Zamora, A.; Sánchez, N.; Gumuzzio, A.; Herrero-Jiménez, C.M. Satellite soil moisture for agricultural drought monitoring: Assessment of the SMOS derived Soil Water Deficit Index. *Remote Sens.* **2016**, *177*, 277–286. [CrossRef]
86. Vågen, T.G.; Winowiecki, L.A.; Tondoh, J.E.; Desta, L.T.; Gumbricht, T. Mapping of soil properties and land degradation risk in Africa using MODIS reflectance. *Geoderma* **2016**, *263*, 216–225. [CrossRef]
87. Othaman, N.N.C.; Md Isa, M.N.; Hussin, R.; Zakaria, S.M.M.S.; Isa, M.M. IoT Based Soil Nutrient Sensing System for Agriculture Application. *Int. J. Nanoelectron. Mater.* **2021**, *14*, 279–288.
88. Ioana, M.; George, S.; Cristina, M.B.; Ana-Maria, D.; Marius, A.D. IoT Solution for Plant Monitoring in Smart Agriculture. In Proceedings of the IEEE 25th International Symposium for Design and Technology in Electronic Packaging, Cluj-Napoca, Romania, 23–26 October 2019; pp. 194–197.
89. Pallavi, S.; Mallapur, J.D.; Bendigeri, K.Y. Remote sensing and controlling of greenhouse agriculture parameters based on IoT. In Proceedings of the International Conference on Big Data, IoT and Data Science (BID), Pune, India, 20–22 December 2017; pp. 44–48.

90. Rubio, V.S.; Ma, F.R. From Smart Farming towards Agriculture 5.0: A Review on Crop Data Management. *Agronomy* **2020**, *10*, 207. [CrossRef]
91. Yuan, G.; Luo, Y.; Sun, X.; Tang, D. Evaluation of a crop water stress index fordetecting water stress in winter wheat in the North China Plain. *Agric. Water Manag.* **2004**, *64*, 29–40. [CrossRef]
92. Köksal, Ö.; Tekinerdogan, B. Architecture design approach for IoT-based farm management information systems. *Precis. Agric.* **2019**, *20*, 926–958. [CrossRef]
93. Xue, J.; Su, B. Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications. *J. Sens.* **2017**, *2017*, 1353691. [CrossRef]
94. Lavanya, G.; Rani, C.; Ganeshkumar, P. An automated low cost IoT based Fertilizer Intimation System for smart agriculture. *Sustain. Comput. Inform. Syst.* **2020**, *28*, 100300. [CrossRef]
95. Benincasa, P.; Antognelli, S.; Brunetti, L.; Fabbri, C.; Natale, A.; Sartoretti, V.; Vizzari, M. Reliability of NDVI Derived by High Resolution Satellite and UAV Compared to In-Field Methods for the Evaluation of Early Crop N Status and Grain Yield in Wheat. *Exp. Agric.* **2018**, *54*, 604–622. [CrossRef]
96. Pinheiro Lisboa, I.; Melo Damian, J.; Roberto Cherubin, M.; Silva Barros, P.P.; Ricardo Fiorio, P.; Cerri, C.C.; Eduardo Pellegrino Cerri, C. Prediction of Sugarcane Yield Based on NDVI and Concentration of Leaf Tissue Nutrients in Fields Managed with Straw Removal. *Agronomy* **2018**, *8*, 196. [CrossRef]
97. Sishodia, R.P.; Ray, R.L.; Singh, S.K. Applications of Remote Sensing in Precision Agriculture: A Review. *Remote Sens.* **2020**, *12*, 3136. [CrossRef]
98. Suradhaniwar, S.; Kar, S.; Nandan, R.; Raj, R.; Jagarlapudi, A. Geo-ICDTs: Principles and Applications in Agriculture. In *Geospatial Technologies in Land Resources Mapping, Monitoring and Management*; Obi Reddy, G.P., Singh, S.K., Eds.; Geotechnologies and the Environment; Springer: Cham, Switzerland, 2018; Volume 21, pp. 75–99. [CrossRef]
99. Colaço, A.F.; Molin, J.P. Variable rate fertilization in citrus: A long term study. *Precis. Agric.* **2017**, *18*, 169–191. [CrossRef]
100. Bruno, B.; Benjamin, D.; Davide, C.; Andrea, P.; Francesco, M.; Luigi, S. Environmental and Economic benefits of variable rate nitrogen fertilization in a nitrate vulnerable zone. *Sci. Total Environ.* **2016**, *545–546*, 227–235.
101. Khan, N.; Medlock, G.; Graves, S.; Anwar, S. *GPS Guided Autonomous Navigation of a Small Agricultural Robot with Automated Fertilizing System*; SAE Technical Paper 2018-01-0031; SAE International: Warrendale, PA, USA, 2018. [CrossRef]
102. Raut, R.; Varma, H.; Mulla, C.; Pawar, V.R. Soil Monitoring, Fertigation, and Irrigation System Using IoT for Agricultural Application. In *Intelligent Communication and Computational Technologies*; Springer: Singapore, 2017; pp. 67–73.
103. Briones, A.G.; Castellanos-Garzón, J.A.; Martín, Y.M.; Prieto, J.; Corchado, J.M. A Framework for Knowledge Discovery from Wireless Sensor Networks in Rural Environments: A Crop Irrigation Systems Case Study. *Wirel. Commun. Mob. Comput.* **2018**, *2018*, 6089280. [CrossRef]
104. Villarrubia, G.; De Paz, J.F.; De La Iglesia, D.H.; Bajo, J. Combining Multi-Agent Systems and Wireless Sensor Networks for Monitoring Crop Irrigation. *Sensors* **2017**, *17*, 1775. [CrossRef]
105. Newlands, N.K. Model-Based Forecasting of Agricultural Crop Disease Risk at the Regional Scale, Integrating Airborne Inoculum, Environmental, and Satellite-Based Monitoring Data. *Front. Environ. Sci.* **2018**, *6*, 63. [CrossRef]
106. Khattaba, A.S.; Habiba, E.D.; Ismail, H.; Zayanc, S.; Fahmya, Y.; Khairya, M.M. An IoT-based cognitive monitoring system for early plant disease forecast. *Comput. Electron. Agric.* **2019**, *166*, 105028. [CrossRef]
107. Carvalho, F.P. Pesticides, environment, and food safety. *Food Energy Secur.* **2017**, *6*, 48–60. [CrossRef]
108. Ramalingam, B.; Mohan, R.E.; Pookkuttath, S.; Gómez, B.F.; Sairam Borusu, C.S.C.; Wee Teng, T.W.; Tamilselvam, Y.K. Remote Insects Trap Monitoring System Using Deep Learning Framework and IoT. *Sensors* **2020**, *20*, 5280. [CrossRef]
109. Kim, S.; Lee, M.; Shin, C. IoT-Based Strawberry Disease Prediction System for Smart Farming. *Sensors* **2018**, *18*, 4051. [CrossRef] [PubMed]
110. Venkatesan, R.; Kathrine, G.; Jaspher, W.; Ramalakshmi, K. Internet of Things Based Pest Management Using Natural Pesticides for Small Scale Organic Gardens. *J. Comput. Theor. Nanosci.* **2018**, *15*, 2742–2747. [CrossRef]
111. Ennouri, K.; Kallel, A. Remote Sensing: An Advanced Technique for Crop Condition Assessment. *Math. Probl. Eng.* **2019**, *2019*, 9404565. [CrossRef]
112. Marinelli, M.C.; Scavuzzo, C.M.; Giobellina, B.L.; Scavuzzo, C.M. Geoscience and Remote Sensing on Horticulture as Support for Management and Planning. *J. Agron. Res.* **2019**, *2*, 43–54. [CrossRef]
113. Wietzke, A.; Westphal, C.; Gras, P.; Kraft, M.; Pfohl, K.; Karlovsky, P.; Pawelzik, E.; Tscharntke, T.; Smit, I. Insect pollination as a key factor for strawberry physiology and marketable fruit quality. *Agric. Ecosyst. Environ.* **2018**, *258*, 197–204. [CrossRef]
114. Chung, S.O.; Choi, M.C.; Lee, K.H.; Kim, Y.J.; Hong, S.J.; Li, M. Sensing Technologies for Grain Crop Yield Monitoring Systems: A Review. *J. Biosyst. Eng.* **2016**, *41*, 408–417. [CrossRef]
115. Talaei, G.H.T.H.; Gholami, S.; Pishva, Z.K.; Dehaghi, M.A. Effects of Biological and Chemical Fertilizers Nitrogen on Yield Quality and Quantity in Cumin (*Cuminum cyminum* L.). *J. Chem. Health Risks* **2014**, *4*, 55–64.
116. Singh, R.; Singh, G.S. Traditional agriculture: A climate-smart approach for sustainable food production. *Energy Ecol. Environ.* **2017**, *2*, 296–316. [CrossRef]
117. Udomkun, P.; Nagle, M.; Argyropoulos, D.; Mahayothee, B.; Müller, J. Multi-sensor approach to improve optical monitoring of papaya shrinkage during drying. *J. Food Eng.* **2016**, *189*, 82–89. [CrossRef]

118. Theopoulos, A.; Boursianis, A.; Koukounaras, A.; Samaras, T. Prototype wireless sensor network for real-time measurements in hydroponics cultivation. In Proceedings of the 7th International Conference on Modern Circuits and Systems Technologies (MOCAST), Thessaloniki, Greece, 7–9 May 2018. [CrossRef]
119. Shamshiri, R.R.; Kalantari, F.; Ting, K.C.; Thorp, K.R.; Hameed, I.A.; Weltzien, C.; Ahmad, D.; Shad, Z. Advances in greenhouse automation and controlled environment agriculture: A transition to plant factories and urban agriculture. *Int. J. Agric. Biol. Eng.* **2018**, *11*, 1–22. [CrossRef]
120. Akkaş, M.A.; Sokullu, R. An IoT-based greenhouse monitoring system with Micaz motes. *Procedia Comput. Sci.* **2017**, *113*, 603–608. [CrossRef]
121. Kodali, R.K.; Jain, V.; Karagwal, S. IoT based smart greenhouse. In Proceedings of the 2016 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Agra, India, 21–23 December 2016. [CrossRef]
122. Tripathy, P.K.; Tripathy, A.K.; Agarwal, A.; Mohanty, S.P. MyGreen: An IoT-Enabled Smart Greenhouse for Sustainable Agriculture. *IEEE Consum. Electron. Mag.* **2021**, *10*, 57–62. [CrossRef]
123. Sambo, P.; Nicoletto, C.; Giro, A.; Pii, Y.; Valentinuzzi, F.; Mimmo, T.; Lugli, P.; Orzes, G.; Mazzetto, F.; Astolfi, S.; et al. Hydroponic Solutions for Soilless Production Systems: Issues and Opportunities in a Smart Agriculture Perspective. *Front. Plant Sci.* **2019**, *10*, 923. [CrossRef] [PubMed]
124. Yang, W.; Feng, H.; Zhang, X.; Zhang, J.; Doonan, J.H.; Batchelor, W.D.; Xiong, L.; Yan, J. Crop Phenomics and High-Throughput Phenotyping: Past Decades, Current Challenges, and Future Perspectives. *Mol. Plant* **2020**, *13*, 187–214. [CrossRef]
125. Rouphael, Y.; Spíchal, L.; Panzarová, K.; Casa, R.; Colla, G. High-Throughput Plant Phenotyping for Developing Novel Biostimulants: From Lab to Field or From Field to Lab? *Front. Plant Sci.* **2018**, *9*, 1197. [CrossRef]
126. Lakshmanan, R.; Djama, M.; Selvaperumal, S.; Abdulla, R. Automated smart hydroponics system using internet of things. *Int. J. Electr. Comput. Eng.* **2020**, *10*, 6389–6398. [CrossRef]
127. Usman, N.; Arief, P.; Gilang, L.; Erfan, R.; Hendra, P. Implementation IoT in System Monitoring Hydroponic Plant Water Circulation and Control. *Int. J. Eng. Technol.* **2018**, *7*, 122–126.
128. Pimentel, D.; Burgess, M. Soil erosion threatens food production. *Agriculture* **2013**, *3*, 443–463. [CrossRef]
129. Benke, K.; Tomkins, B. Future food-production systems: Vertical farming and controlled-environment agriculture. *Sustain. Sci. Pract. Policy* **2017**, *13*, 13–26. [CrossRef]
130. Tripodi, P.; Massa, D.; Venezia, A.; Cardi, T. Sensing Technologies for Precision Phenotyping in Vegetable Crops: Current Status and Future Challenges. *Agronomy* **2018**, *8*, 57. [CrossRef]
131. Paul, K.; Sorrentino, M.; Lucini, L.; Rouphael, Y.; Cardarelli, M.; Bonini, P.; Reynaud, H.; Canaguier, R.; Trtílek, M.; Panzarová, K.; et al. Understanding the Biostimulant Action of Vegetal-Derived Protein Hydrolysates by High-Throughput Plant Phenotyping and Metabolomics: A Case Study on Tomato. *Front. Plant Sci.* **2019**, *10*, 47. [CrossRef]
132. Zhou, J.; Reynolds, D.; Websdale, D.; Le Cornu, T.; Gonzalez Navarro, O.; Lister, C.; Orford, S.; Laycock, S.; Finlayson, G.; Stitt, T.; et al. Cropquant: An automated and scalable field phenotyping platform for crop monitoring and trait measurements to facilitate breeding and digital agriculture. *bioRxiv* **2017**. [CrossRef]
133. Bochtis, D.; Sørensen, C.A.G.; Kateris, D. *Operations Management in Agriculture*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 1–18. [CrossRef]
134. Terence, S.; Purushothaman, G. Systematic review of Internet of Things in smart farming. *Trans. Emerg. Telecommun. Technol.* **2020**, *31*, e3958. [CrossRef]
135. James, A.; Saji, A.; Nair, A.; Joseph, D. CropSense–A Smart Agricultural System using IoT. *J. Electron. Des. Eng.* **2019**, *5*, 1–7.
136. Bacco, M.; Barsocchi, P.; Ferro, E.; Gotta, A.; Ruggeri, M. The digitization of agriculture: A survey of research activities on smart farming. *Array* **2019**, *3–4*, 100009. [CrossRef]
137. Ahmed, A.T.; El Gohary, F.; Tzanakakis, V.A.; Angelakis, A.N. Egyptian and Greek Water Cultures and Hydro-Technologies in Ancient Times. *Sustainability* **2020**, *12*, 9760. [CrossRef]
138. Adebayo, S.; Ogunti, E.O.; Akingbade, F.K.; Oladimeji, O. A review of decision support system using mobile applications in the provision of day-to-day information about farm status for improved crop yield. *Period. Eng. Nat. Sci.* **2018**, *6*, 89–99. [CrossRef]
139. Foster, A.D.; Mark, R.R. Microeconomics of technology adoption. *Annu. Rev. Econ.* **2010**, *2*, 395–424. [CrossRef]
140. Feder, G.; Just, R.E.; Zilberman, D. Adoption of agricultural innovations in developing countries: A survey. *Econ. Dev. Cult. Chang.* **1985**, *33*, 255–298. [CrossRef]
141. Alvarez, J.; Peter, N. Adoption of computer based information systems: The case of dairy farmers in Canterbury, NZ, and Florida, Uruguay. *Comput. Electron. Agric.* **2006**, *50*, 48. [CrossRef]
142. Kimiti, J.M.; Odee, D.W.; Vanlauwe, B. *Area under Grain Legumes Cultivation and Problems Faced by Smallholder Farmers in Legume Production in the Semi-Arid Eastern Kenya*; Academic Press: Cambridge, MA, USA, 2009.
143. Khan, A.R.; Dubey, M.K.; Bisen, P.K.; Saxena, K.K. Constraints faced by farmers of Narsing Kheda village of Sihore district. *Young* **2007**, *8*, 16.
144. Abdul Hakkim, V.M.; Abhilash Joseph, E.; Ajay Gokul, A.J.; Mufeedha, K. Precision Farming: The Future of Indian Agriculture. *J. Appl. Biol. Biotechnol.* **2016**, *4*, 068–072. [CrossRef]
145. Fróna, D.; Szenderák, J.; Rákos, M.H. The Challenge of Feeding the World. *Sustainability* **2019**, *11*, 5816. [CrossRef]
146. Tzounisa, A.; Katsoulasa, N.; Bartzanasb, T.; Kittas, C. Internet of Things in agriculture, recent advances and future challenges. *Biosyst. Eng.* **2017**, *164*, 31–48. [CrossRef]

147. Henriksen, A.V.; Edwards, T.C.G.; Pesonen, L.A.; Green, O.; Sørensen, C.A.G. Internet of Things in arable farming: Implementation, applications, challenges and potential. *Biosyst. Eng.* **2019**, *191*, 60–84. [CrossRef]
148. Roth, L.; Aasen, H.; Walter, A.; Liebisch, F. Extracting leaf area index using viewing geometry effects new perspective on high-resolution unmanned aerial system photography. *ISPRS J. Photogramm. Remote Sens.* **2018**, *141*, 161–175. [CrossRef]
149. Chang, A.; Jung, J.; Maeda, M.; Landivar, J. Crop height monitoring with digital imagery from unmanned aerial system (UAV). *Comput. Electron. Agric.* **2017**, *141*, 232–237. [CrossRef]
150. Barrero, O.; Perdomo, S.A. RGB and multispectral UAV image fusion for Gramineae weed detection in rice fields. *Precis. Agric.* **2018**, *19*, 809–822. [CrossRef]
151. Lottes, P.; Khanna, R.; Pfeifer, J.; Siegwart, R.; Stachniss, C. UAV-based crop and weed classification for smart farming. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 3024–3031.
152. Stroppiana, D.; Migliazzi, M.; Chiarabini, V.; Crema, A.; Musanti, M.; Franchino, C.; Villa, P. Rice yield estimation using multispectral data from UAV: A preliminary experiment in northern Italy. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 4664–4667.
153. Hassan, M.A.; Yang, M.; Rasheed, A.; Yang, G.; Reynolds, M.; Xia, X.; Xiao, Y.; He, Z. A rapid monitoring of NDVI across the wheat growth cycle for grain yield prediction using a multi-spectral UAV platform. *Plant Sci.* **2019**, *282*, 95–103. [CrossRef]
154. Duan, T.; Chapman, S.; Guo, Y.; Zheng, B. Dynamic monitoring of NDVI in wheat agronomy and breeding trials using an unmanned aerial vehicle. *Field Crops Res.* **2017**, *210*, 71–80. [CrossRef]
155. Su, J.; Liu, C.; Coombes, M.; Hu, X.; Wang, C.; Xu, Z.; Li, Q.; Guo, L.; Chen, W.H. Wheat yellow rust monitoring by learning from multispectral UAV aerial imagery. *Comput. Electron. Agric.* **2018**, *155*, 157–166. [CrossRef]
156. Jin, X.; Liu, S.; Baret, F.; Hemerl, M.; Comar, A. Estimates of plant density of wheat crops at emergence from very low altitude UAV imagery. *Remote Sens. Environ.* **2017**, *198*, 105–114. [CrossRef]
157. Faial, B.S.; Costa, F.G.; Pessin, G.; Ueyama, J.; Freitas, H.; Colombo, A.; Fini, P.H.; Villas, L.; Osrio, F.S.; Vargas, P.A.; et al. The use of unmanned aerial vehicles and wireless sensor networks for spraying pesticides. *J. Syst. Archit.* **2014**, *60*, 393–404.
158. Al-Kodmany, K. The Vertical Farm: A Review of Developments and Implications for the Vertical City. *Buildings* **2018**, *8*, 24. [CrossRef]
159. Cravero, A.; Pardo, S.; Sepúlveda, S.; Muñoz, L. Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review. *Agronomy* **2022**, *12*, 748. [CrossRef]
160. Kaneko, A.; Kennedy, T.W.; Mei, L.; Sintek, C.; Burke, M.; Ermon, S.; Lobell, D.B. Deep Learning for Crop Yield Prediction in Africa. In Proceedings of the International Conference on Machine Learning AI for Social Good Workshop, LongBeach, CA, USA, 10–15 June 2019.
161. Oswal, N. Predicting rainfall using Machine Learning Techniques. *arXiv* **2019**, arXiv:1910.13827.
162. Zhang, J.; Zhu, Y.; Zhang, X.; Ye, M.; Yang, J. Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. *J. Hydrol.* **2018**, *561*, 918–929. [CrossRef]
163. Ouyang, H.; Wei, X.; Wu, Q. Agricultural commodity futures prices prediction via long- and short-term time series network. *J. Appl. Econ.* **2019**, *22*, 468–483. [CrossRef]
164. Bhattarai, T.N.; Ghimire, S.; Mainali, B.; Gorjian, S.; Treiche, H.; Paudel, S.R. Applications of smart grid technology in Nepal: Status, challenges, and opportunities. *Environ. Sci. Pollut. Res.* **2022**, 1–25. [CrossRef] [PubMed]
165. Kedri, J. Simulation and Validation of SPV Micro Grid Comprising 518.2 KWp Distributed Solar Power Plants at Dayalbagh Educational Institute. Ph.D. Thesis, Dayalbagh Educational Institute, Dayalbagh, Agra, India, 2014.
166. Aslam, S.; Khalid, A.; Javaid, N. Towards Efficient Energy Management in Smart Grids Considering Microgrids with Day-ahead Energy Forecasting. *Electr. Power Syst. Res.* **2020**, *182*, 106232. [CrossRef]

agriculture

MDPI

Article

Modeling the Agricultural Soil Landscape of Germany—A Data Science Approach Involving Spatially Allocated Functional Soil Process Units

Mareike Ließ

Department of Soil System Science, Helmholtz Centre for Environmental Research—UFZ, D-06120 Halle (Saale), Germany; mareike.liess@ufz.de

Abstract: The national-scale evaluation and modeling of the impact of agricultural management and climate change on soils, crop growth, and the environment require soil information at a spatial resolution addressing individual agricultural fields. This manuscript presents a data science approach that agglomerates the soil parameter space into a limited number of functional soil process units (SPUs) that may be used to run agricultural process models. In fact, two unsupervised classification methods were developed to generate a multivariate 3D data product consisting of SPUs, each being defined by a multivariate parameter distribution along the depth profile from 0 to 100 cm. The two methods account for differences in variable types and distributions and involve genetic algorithm optimization to identify those SPUs with the lowest internal variability and maximum inter-unit difference with regards to both their soil characteristics and landscape setting. The high potential of the methods was demonstrated by applying them to the agricultural German soil landscape. The resulting data product consists of 20 SPUs. It has a 100 m raster resolution in the 2D mapping space, and its resolution along the depth profile is 1 cm. It includes the soil properties texture, stone content, bulk density, hydromorphic properties, total organic carbon content, and pH.

Keywords: digital soil mapping; soil process units; soil parameter space; machine learning; unsupervised classification

Citation: Ließ, M. Modeling the Agricultural Soil Landscape of Germany—A Data Science Approach Involving Spatially Allocated Functional Soil Process Units. *Agriculture* **2022**, *12*, 1784. https://doi.org/10.3390/agriculture12111784

Academic Editors: Gniewko Niedbała and Sebastian Kujawa

Received: 20 September 2022
Accepted: 24 October 2022
Published: 27 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Global food security, the protection of our groundwater resources, and our efforts to combat climate change largely depend on the sustainable use of soils. This concerns the strategic planning of an adequate crop rotation, the careful use of fertilizers, and the restricted use of pesticides. To maintain the soils' high productivity, we need to provide crops with sufficient and easily accessible nutrients. However, the soils' storage potential is limited. Surplus fertilizer contaminates valuable water resources when it percolates to the groundwater. It enhances global warming while released as greenhouse gases into the atmosphere. Furthermore, crops also require sufficient plant-available soil water resources in their respective development stages. Irrigation needs to be crop- and soil-specific but may not be the best solution as it restricts water for other uses. In consequence, it requires thoughtful planning of an adapted crop cycle involving drought-tolerant cultivars [1] and respective soil water management by alternative means [2,3].

All decisions and their consequences with regards to soil productivity and environmental impact ultimately depend on the soil characteristics on site. Accordingly, the national-scale evaluation and modeling of the impact of agricultural management and climate change on agricultural soils, yields, and the environment require information on the multivariate 3D soil parameter space at a spatial resolution addressing individual agricultural fields [4,5]. This concerns the assessment of the soils' agricultural productivity [6] and the restrictions and required adaptations due to prolonged drought periods. Crop phenology models [7] and the evaluation and modeling of soil-related drought [8–10] and

corresponding irrigation requirements [11] could be improved to a large extent by adequate soil information at a high spatial resolution. The same applies to the evaluation of the soils' storage potential for soil organic carbon [12,13], the modeling of the complex processes causing the release of greenhouse gases to combat climate change [14], and the modeling of mitigation options to reduce nitrate pollution [15,16].

Running agricultural process models at national scale requires information about the multivariate 3D soil parameter space at a spatial resolution targeting individual agricultural fields. With a spatial resolution of 100 m, this already amounts to about 20 million raster cells for the agricultural soils of Germany. Process models require high computing capacities to run repeated simulations considering agricultural management and climate scenarios on this number of raster cells. Unfortunately, this also goes along with an unnecessarily high amount of energy consumption, counteracting our efforts to combat climate change. Hence, a creative data science approach is required to agglomerate the information contained in the raster cells to a limited number of spatially allocated functional soil process units (SPUs). This enables us to reduce the required resources without having to accept a lower spatial resolution.

One might argue why not rather use the spatial map units (SMUs) contained in conventional soil maps as SPUs? For Germany, there are mainly three reasons why the contained soil information is inappropriate: (1) The best conventional soil map available at national scale for Germany is the BÜK at a map scale of 1:250.000 [17]. Its SMUs each define a paragenesis of soil systematic units (SUs) with highly differing characteristics. The spatial allocation of these SUs within the SMUs is unknown. Hence, the contained information is not site-specific when it comes to addressing individual agricultural fields. (2) Important soil properties guiding soil functionality are only distinguished at a low hierarchical level of the German soil classification system KA [18]. Rather similar soils concerning their properties and functionality are assigned to different upper-level SUs. This particularly applies to the particle size distribution, which is one of the most important properties guiding soil functionality. (3) Last but not least, the BÜK is uncertain. All soil maps are. However, on the one hand, the BÜK's uncertainty is unknown. On the other hand, its uncertainty likely differs between the federal states as the map was developed by slightly differing approaches at the regional soil survey institutions and then later joined and harmonized concerning inconsistencies at the regional boundaries.

The development of creative data science approaches to provide spatially continuous soil information relates to the research field pedometrics. Pedometrics is an interdisciplinary science that integrates soil science with geoinformatics and data science. Pedometric modeling approaches are used to investigate the spatial-temporal variation of the soil landscape and derive spatially continuous soil information from soil profile data. They rely on the conceptual model of pedogenesis, with soils and their vertical profile differentiation and characteristics being the product of the site-specific interaction of the soil-forming factors through long periods of time [19]. The conceptual approach was extended by McBratney et al. [20] to include geographic location and proxies for soil itself. The resulting SCORPAN factors include proxies to soil (S), climate (C), organisms including land use, agricultural management, etc. (O), relief (R), parent material (P), age (A), and geographic location (N). They are each approximated by spatially continuous gridded data proxies from either remote sensing, by conducting a digital terrain analysis, and/or by including expert knowledge. Padarian et al., Arrouays et al., and Chen et al. [21–23] provide recent reviews. Many studies refer to pedometric modeling for landscape-scale predictions by the terms 'digital soil mapping' or 'predictive soil mapping'. I prefer the term pedometric modeling since digital soil maps are also created by other approaches, and any map is two-dimensional and, therefore, does not necessarily include 3D data products.

Current approaches in pedometric modeling to generate nationwide soil information predominantly address the prediction of individual soil properties. Žížala et al. and Gebauer et al. [24,25] provide recent 2D applications, Malone and Searle and Reddy et al. [26,27] 2.5D applications, and Padarian et al. and Ma et al. [28,29] 3D applica-

tions. However, the separate modeling of individual soil properties and their respective joint consideration as input to agricultural process models may result in constructed soil profile information that does not occur in reality and may be unrealistic according to the underlying pedogenetic processes and dependencies between the properties. Ließ et al. [4] provide a promising alternative for the joint modeling of multiple soil properties in 3D. The resulting data product represents the multivariate 3D soil parameter space of the nationwide agricultural landscape of Germany in terms of spatially allocated SPUs, each being described by a multivariate parameter distribution along the depth profile from 0 to 100 cm. It includes depth- and property-wise uncertainty estimates.

Here, a data science approach shall be developed that serves to generate such multivariate 3D data products consisting of spatially allocated functional SPUs. In contrast to Ließ et al. [4], it involves the development of unsupervised classification methods that account for differences in variable types and distributions and involve optimization to identify those SPUs with the lowest internal variability and maximum inter-unit difference with regards to both their soil characteristics and landscape setting. The approach shall be evaluated by applying it to the German agricultural soil landscape to improve the previously mentioned data product.

2. Materials and Methods

2.1. Data

2.1.1. Soil Profile Data—Consistency Check and Gap Filling

The soil profile data from the agricultural soil inventory of Germany [30] were used for this study. The data were collected by systematic sampling along an 8 km × 8 km grid at 3104 sites. Each soil profile has an identifier and geographic coordinates. The data comprise field data (dataF) in terms of a soil profile description according to the German soil survey system KA5 [18], and laboratory data (dataL). From dataF, the horizon-wise texture class, stone content, and the horizon symbol of all profiles were considered. From dataL, the particle-size distribution (3 particle-size separates), the bulk density, stone content, total organic carbon content (TOC), and the pH value of all profiles were considered. In the following, I describe the consistency check, subsequent data modification, and gap-filling procedure that were applied prior to any further analysis.

The sampling protocol for the Agricultural Soil Inventory states that samples for subsequent laboratory analysis ought to be taken for the depth increments 0–10, 10–30, 30–50, 50–70, and 70–100 cm while taking into account horizon boundaries, i.e., including multiple samples per depth increment for each corresponding soil horizon present with five or more centimeters [31]. However, as could be expected for such a large soil survey campaign involving multiple teams, the dataset contains some inconsistencies. To combine dataF and dataL, the two datasets were checked for mismatches in absolute profile depth and horizon sequence notation (term used for dataF and dataL), as well as non-compliant data entries, duplicates or gaps in the horizon sequence notation. After correcting non-compliant data entries, the next correction step concerned the mismatches in profile depth and horizon sequence notation. For their correction, I tested whether mismatches concerning depth and horizon sequence notation corresponded to additional layers (or horizons) and whether the difference was minor, up to 5 cm, i.e., mismatches in line with the sampling protocol. After adjusting the layer boundaries accordingly, all other mismatches were corrected stepwise by favoring the profile depth of dataF over dataL in case the difference was not caused by additional layers (or horizons), and by splitting layers of dataL if they included one or more horizon boundaries that differed from the upper or lower layer boundary by five or more centimeters. This procedure resulted in matching horizon sequence notation and profile depth between dataF and dataL, and the two datasets were combined using the profile identifier. From now on, these joint depth divisions will be referred to as horizons.

The modifications in the horizon sequence notation in dataL and dataF resulted in data gaps concerning all laboratory or field data of a certain depth interval. Further, using interpolation methods to fill these gaps may not be the best option due to the

geological stratification, i.e., discontinuities in the soil profiles. In addition, the data gaps relating to the uppermost and last soil horizon cannot be filled in this way. Therefore, the following procedures were applied: The resulting texture data gaps in dataL were filled by additionally considering texture data from dataF. The mean value of the sand, silt and clay content (dataL) from other horizons with matching texture classes (dataF) was used. This happened stepwise. If the prerequisites were met, only data from the same profile were used. Otherwise the complete dataset's respective class-wise mean values were assigned. Finally, the remaining texture classes were filled by the KA5 texture class's mean sand, silt, and clay content. The latter corresponds to layers with uncommon soil texture classes and hence too few data entries (less than five). For data gaps in the TOC of organic soil horizons, a similar approach was followed considering horizon symbols and organic texture classes. For TOC in the mineral soil horizons, as well as the pH, bulk density and stone content of all horizons, random forest (RF) models were trained. Model training, tuning, and evaluation were conducted with nested stratified cross-validation (CV), as explained in Section 2.3.2. As predictors, the same property's values from upper and lower horizons as well as related soil properties of over- and underlying horizons were used. Related properties of the same horizon could not be used unless for those where dataF was used to fill gaps in dataL.

After gap filling, some additional variables were created. For the stone content, the data from dataF and dataL were combined by assigning the maximum of the two values. This was done since on the one hand, dataL underestimates the stone content with regards to large rock fragments beyond the size of the steel cores used for sampling. On the other hand, the visual method applied to estimate the stone content in dataF may neglect smaller rock fragments. Concerning hydromorphic features, one variable was created for each, the presence (value = 1) or absence (value = 0) of stagnic and gleyig properties, and named symbol_S and symbol_G. The information was derived from the horizon symbology of dataF. An additional variable 'mob' was included in the dataset assigning each horizon to either 'mineral', 'organic', or 'bedrock' by considering the TOC, horizon symbology, and the availability of texture data. Each profile was then subdivided into 1 cm slices up to a depth of 100 cm.

2.1.2. Data Cube of Covariates

The covariates included to train and apply the machine learning models for nationwide spatial prediction were grouped according to the SCORPAN factor they represent. Table 1 gives an overview. Ließ et al. [4] provide a description of the German landscape setting.

Concerning SCORPAN C, seasonal averages of air temperature and drought and the sum of precipitation of the winter (Dec., Jan., and Feb.) and the summer (Jun., Jul., and Aug.) months were derived from the German Weather Service (DWD). The seasonal averages of the drought index were calculated from DWD temperature in degrees centigrade (T) and precipitation in millimeters (P) grids as $P/(T + 10)$.

Table 1. Covariates.

Soil Forming Factor	Abbreviation	Description	Data Source
Climate	PRESU PREWI	Average seasonal precipitation (summer) [raster, 1000 m] Average seasonal precipitation (winter) [raster, 1000 m]	[32]
	TEMSU TEMWI	Average seasonal temperature (summer) [raster, 1000 m] Average seasonal temperature (winter) [raster, 1000 m]	[33]
	DINSU DINWI	Average seasonal drought index (summer) [raster, 1000 m] Average seasonal drought index (winter) [raster, 1000 m]	[34]
Organisms/Soil	B0118, 0218, … B0818, B8A18, B1118, B1218	Sentinel-2 spectral bands B1, B2, … B8, B8A, B11, and B12 composites of the 2nd yearly quartile of the year 2018	
	B0121, 0221, … B0821, B8A21, B1121, B1221	Sentinel-2 spectral bands B1, B2, … B8, B8A, B11, and B12 composites of the 2nd yearly quartile of the year 2021	
	EVI18, EVI21	Enhanced vegetation index, calculated from Sentinel 2 band composites of 2nd quartile 2018 & 2021 (S2-Q2-18/21), $EVI = G * (B8A - B04)/(B8A + C1 * B04 - C2 * B02 + L)$, with G = 2.5, C1 = 6, C2 = 7.5 and L = 1	
	MSI18, MSI21	Moisture index: S2-Q2-18/21, MSI = B11/B08	
	NDM18, NDM21	Normalized difference moisture index: S2-Q2-18/21, NDMI = (B08 − B11)/(B08 + B11)	
	NDV18, NDV21	Normalized difference vegetation index: S2-Q2-18/21, NDVI = (B08 − B04)/(B08 + B04)	
	NDW18, NDW21	Normalized difference water index: S2-Q2-18/21, NDWI = (B03 − B08)/(B03 + B08)	
	PSR18, PSR21	Plant senescence reflectance index: S2-Q2-18/21, PSRI = (B04 − B02)/B06	
	DMP16 DMP18	Dry matter productivity, June 2016 [raster, 300 m] Dry matter productivity, June 2018 [raster, 300]	[35]
	VPI16 VPI18	Vegetation Productivity Index, June 2016 [raster, 300 m] Vegetation Productivity Index, June 2018 [raster, 300 m]	[36]

Table 1. *Cont.*

Soil Forming Factor	Abbreviation	Description	Data Source
Topography	GMK00	Geomorphographic map of Germany [raster, 250 m resolution, map scale 1:1,000,000]	[37]
	DEM00	Digital elevation model [raster, 25 m resolution]	[38]
	SLO01, SLO05, SLO10	Slope: calculated from DEM (cfD) with a search radius of 1, 5, 10 cells, using SAGA module Morphometric features	
	NOR01, NOR05, NOR10	Northness: derived from aspect cfD with a search radius of 1, 5, 10 cells, using SAGA module Morphometric features	
	EAS01, EAS05, EAS10	Eastness: derived from aspect cfD with a search radius of 1, 5, 10 cells, using SAGA module Morphometric features	
	TST01, TST05, TST10	Terrain surface texture: cfD with a search radius of 1, 5, 10 cells, using SAGA module Terrain Surface Texture	
	TSR01, TSR05, TSR10	Terrain surface ruggedness: cfD with a search radius of 1, 5, 10 cells, using SAGA module Terrain Ruggedness Index	
	CON01, CON05, CON10	Convergence index: cfD with a search radius of 1, 5, 10 cells, using SAGA module Convergence Index (Search Radius)	
	SLH00	Slope height: cfD using SAGA module Relative Heights and Slope Positions	
	VAD00	Valley depth: cfD using SAGA module Relative Heights and Slope Positions	
	NOH00	Normalized height: cfD using SAGA module Relative Heights and Slope Positions	
	WIN00	Wind exposure: cfD using SAGA module Wind Effect	
	NOP00	Negative openness: cfD using SAGA module Topographic Openness	
	POP00	Positive openness: cfD using SAGA module Topographic Openness	
	VOF0S	Vertical overland flow distance to all river segments: cfD using SAGA module Terrain analysis/Channels	
	VOF0M	Vertical overland flow distance to major rivers: cfD using SAGA module Terrain analysis/Channels	
	HOF0S	Horizontal overland flow distance to all river segments: cfD using SAGA module Terrain analysis/Channels	
	HOFOM	Horizontal overland flow distance to major rivers: cfD using SAGA module Terrain analysis/Channels	
	SWI00	SAGA wetness index: cfD using SAGA module SAGA Wetness Index	

Table 1. *Cont.*

Soil Forming Factor	Abbreviation	Description	Data Source
Parent material	LIT00	Lithology, Hydrogeological map of Germany, HÜK [polygon shapefile, map scale 1:250,000]	[39]
	STR00	Stratigraphy, Hydrogeological map of Germany, HÜK [polygon shapefile, map scale 1:250,000]	
	BAG00	Groups of soil parent material in Germany [polygon shapefile, map scale 1:5,000,000]	[40]
Soil	BGL00	Soil scapes in Germany [map scale 1:5,000,000]	[41]
	DMP86	Dry matter productivity, DMP18–DMP16 [raster, 300 m]	
	VPI86	Vegetation Productivity Index, VPI18–VPI16 [raster, 300 m]	
Geographic location	LAT00	INSPIRE Latitude	[42]
	LON00	INSPIRE Longitude	

To approximate SCORPAN O, the following covariates were included: Sentinel-2 data composites of the second yearly quartile of 2018 and 2021 of the bands B01, B02, B03, B04, B05, B06, B07, B08, B8a, B11, and B12, as well as the vegetation indices EVI, MSI, NDMI, NDVI, NDWI, and PSRI (please see Table 1 for the details). The composites were compiled using the Sentinel-Hub on behalf of the surface reflectance values, from the Level 2A product. The composites were downloaded as multiple tiles in 20 m spatial resolution, then mosaicked and resampled to the 100 m Infrastructure for Spatial Information in Europe (INSPIRE) grid topology [42] before calculating the vegetation indices. Additionally, remote sensing products on dry matter productivity (DMP) and the Vegetation Productivity Index (VPI) of the time slot June 11th–20th of the years 2016 and 2018 were derived from the Copernicus Global Land Service. All SCORPAN O covariates seek to capture the main annual phase of agricultural productivity.

SCORPAN R was represented by the geomorphographic map of Germany and terrain parameters derived by digital terrain analysis with the System for Automated Geoscientific Analyses (SAGA) [43] from the EU–DEM digital elevation model.

The map of the "Groups of soil parent material" was included to approximate SCORPAN P. Lithology and stratigraphy according to the hydrogeological map of Germany were additionally incorporated.

Proxies to soil itself (SCORPAN S) can generally be included in the form of conventional soil polygon maps, and remote sensing products relating to soil properties. Regarding the former, the map of the German soil scapes was included. Concerning the latter, differences in DMP and VPI between the dry year 2018 and the rather wet year 2016 were included. They relate to crop phenology affected by drought and, therefore, to the root zone plant-available soil water capacity.

All covariates were resampled to the INSPIRE grid topology at 100 m resolution [42]. This resolution was chosen as a compromise between the ambition to provide soil information for individual agricultural fields and a restrictive use of computing capacities. The nearest-neighbor method was used for categorical predictors, and B-spline interpolation was applied for numeric predictors. INSPIRE latitude and longitude were additionally included to represent the geographic location (SCORPAN N), and particularly to represent spatial patterns not captured by the other data proxies. The national border and coastline of Germany were derived from the digital land model at map scale of 1:250,000 (version 2.0) provided by the Federal Agency for Cartography and Geodesy (©GeoBasis-DE/BKG, 2020).

2.2. Differentiation of Functional SPUs

The nationwide data product is composed of a limited number of spatially allocated functional SPUs, each being defined by a multivariate parameter distribution along the depth profile. Each SPU's internal variability is described by a probability density distribution of all considered soil properties in all 1 cm depth slices. Two data science approaches were developed to derive SPUs with the lowest possible internal variability and maximum inter-unit difference with regards to both their soil characteristics and landscape setting. They are unsupervised classification methods, rely on the partitioning-around-medoids (PAM) algorithm [44] and involve optimization. Furthermore, they address the major concern that the joint consideration of mixed variable types (categorical and numerical) and variables of different distribution and scale have on the clustering result. Ahmad and Khan [45] and Van Mechelen et al. [46] provide an overview. In this particular case, there are variables with 1–0 coding for presence–absence type variables (symbol_S, symbol_G), variables with many zero values (stone content), variables with a threefold distribution (texture represented by sand, silt, and clay content), and variables with a bimodal distribution (TOC, bulk density) due to the inclusion of profiles that are all-mineral and profiles composed of mineral and organic horizons. PAM clustering after a mere data transformation did not yield satisfying results.

The two approaches will be described in the following sections. However, two aspects concern the methodology of both approaches:

1. The gap-filled, sliced (1 cm slices) profile data were used to calculate individual property distance matrices. First, the data were normalized to a range between 0 and 1, considering all slices in all profiles except for texture. For texture, the composites' relation of sand, silt, and clay content were kept summing up to 1. Then, the mean of the slice-wise Euclidian profile distance was calculated for each variable and stored in separate distance matrices. Non-defined distances in case of differences in soil material causing missing data, e.g., missing texture data for organic horizons or slices assigned to bedrock, were assigned the maximum distance occurring between any two profile slices for the respective soil property. These property-wise distance matrices were then again normalized, resulting in a minimum distance of 0 and a maximum distance of 1. Hereafter, they will be referred to as normalized single-property distance matrices (*nSPdist*).
2. The respective input parameter vectors of the involved optimization process to extract the SPUs are evaluated on behalf of a complex objective function. It seeks to identify those SPUs with the lowest possible internal variability and maximum inter-unit difference with regards to both their soil characteristics and landscape setting. The former is evaluated by using the Silhouette Index [47]. The latter requires the training of machine learning models to capture the soil-landscape relation and evaluate their predictive performance. A simple and fast learner is required to reduce the required computation time. The random forest (RF) algorithm [48] was chosen to suit this purpose. It is described in Section 2.3.1.

2.2.1. Approach 1 (PAMp)—SPU Extraction by P Weights Optimization

Approach 1 seeks to obtain the optimal SPUs in terms of the lowest property-wise predictive RMSE from pedometric model training by simultaneously optimizing the number of clusters *nclus* and the weights Pw_1, Pw_2, Pw_3, . . . , Pw_p (p = number of soil properties) applied to the *nSPdist*. The weights give the inter-profile distances with regards to certain soil properties higher or lower importance compared to others. To avoid confusion, the weights will hereafter be termed P weights (property weights). Approach 1 will, therefore, be named PAMp. The objective function evaluated for each of the number of n parameter vectors of $z = 8$ components (seven P weights and *nclus*) evaluated in each iteration step of the optimization is shown in Figure 1. It consists of the following parts:

1. Pw_1, Pw_2, Pw_3, . . . , Pw_p in the range [0.1, 1] are assigned to each *nSPdist*, which are then combined by calculating the weighted average (*dist*). The values of the resulting distance matrix are normalized to the range [0, 1].
2. PAM clustering is conducted on the normalized distance matrix (*ndist*) with *nclus* = 8, 9, . . . , 100. The *nclus* minimum value was selected according to Ließ et al. [4]. For each input parameter vector including Pw_1, Pw_2, Pw_3, . . . , Pw_p and $nclus$, the best cluster solution is selected on behalf of the Silhouette Index.

3.1. The resulting clustering solution $Rdata_{in}$, which assigns each soil profile to one cluster, is then combined with the respective l covariates' values x_1, x_2, . . . , x_l of each profile (*Pdata*) to compile the predictor-response dataset (*PRdata*). The data were subdivided into 5 folds for a stratified 5-fold CV (Section 2.3.2). Categorical covariate values with zero data instances in any of the folds were removed.

3.2. Each profile's property-wise mean along the depth profile, $Rdata_{in}$ $[y_1, y_2, \ldots, y_p]$, was used to compute property-wise means per cluster.

4. An RF model was trained by 5-fold stratified CV using the *PRdata* [3.1] as input. The function 'rfsrc' of R package 'randomForestSRC' [49] was used with 1000 trees, a node size of five, and the default setting for the *mtry* parameter, while imputing no data values.

5.1. The previously computed property-wise cluster means [3.2] were assigned to each profile on behalf of the test set RF predictions ($Rdata_{pred}$) generating $Rdata_{pred}$ $[y_1, y_2, \ldots, y_p]$.

5.2. The property-wise RMSE was calculated using $Rdata_{in}$ $[y_1, y_2, \ldots, y_p]$ and $Rdata_{pred}$ $[y_1, y_2, \ldots, y_p]$. The objective function value corresponds to the negative mean of the property-wise RMSE values. It is maximized in the optimization process.

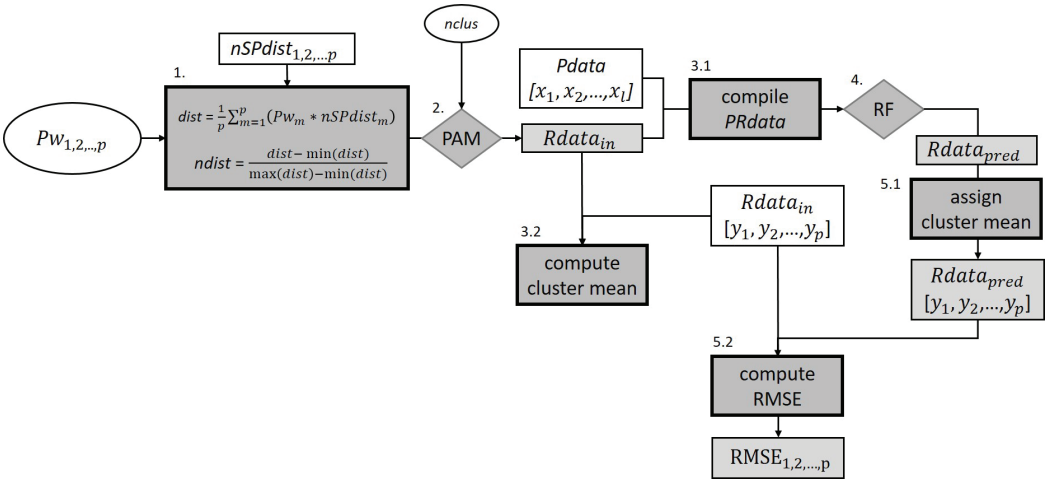


Figure 1. Objective function of the optimization process for SPU differentiation with PAMp. All white boxes are required input data. White ovals reflect parameters that are optimized. *Pw* = vector of P weights, *dist* = distance matrix, *ndist* = normalized distance matrix, *nSPdist* = normalized single-property distance matrices, *nclus* = number of clusters, PAM = partitioning around medoids clustering, *Pdata* = predictor data, *Rdata* = response data, *PRdata* = predictor-response data, RF = random forest.

2.2.2. Approach 2 (PAMm)—SPU Extraction by Optimized Multistep Clustering

Approach 2 seeks to obtain the optimal SPUs in terms of the lowest property-wise predictive RMSE from pedometric model training by applying a multistep clustering with PAM. It will, therefore, be termed PAMm. In this approach, Part 1 and Part 2 of the objective function of PAMp are replaced by the multistep approach (Figure 2). The other subsequent parts remain the same.

The properties considered at each step need to be selected in advance. Optimizing their selection would have increased the complexity of the optimization task and hence required more iterations before convergence. Multistep clustering was conducted in the following way: Step 1 (texture), Step 2 (symbol_S, symbol_G), Step 3 (stone content, bulk density), and Step 4 (TOC, pH). The normalized distance matrices $ndist_1$, $ndist_2$, $ndist_3$, and $ndist_4$ for each step were prepared in advance and then provided as input to the objective function. Each *ndist* was calculated as the normalized average of the *nSPdist* of the soil properties considered in the respective step.

In Step 1, PAM is applied to $ndist_1$ testing a number of 2 to ncl_u clusters. The cluster solution with the best Silhouette Index value is chosen unless there are cluster solutions with a sufficiently good Silhouette Index value equal to or above the threshold sil_1. In that case, the cluster solution with the maximum number of clusters from all cluster solutions with a Silhouette Index value greater than or equal to sil_1 is chosen. In Step 2, PAM is conducted for each cluster resulting from Step 1. This requires subsetting $ndist_2$ according to the profile IDs that were assigned to the respective higher-level Step 1 clusters cl_1, cl_2, ... and normalizing the distance matrix subsets, which were then named nd_{cl1}, nd_{cl2}, etc. The clusters resulting from Step 2 receive a 2nd cluster identifier, e. g., $cl_{1|1}$, $cl_{1|2}$, $cL_{2|1}$, $cl_{2|2}$ indicate that the two clusters from Step 1 were each subdivided into two clusters in Step 2. This procedure is repeated likewise for Step 3 and Step 4.

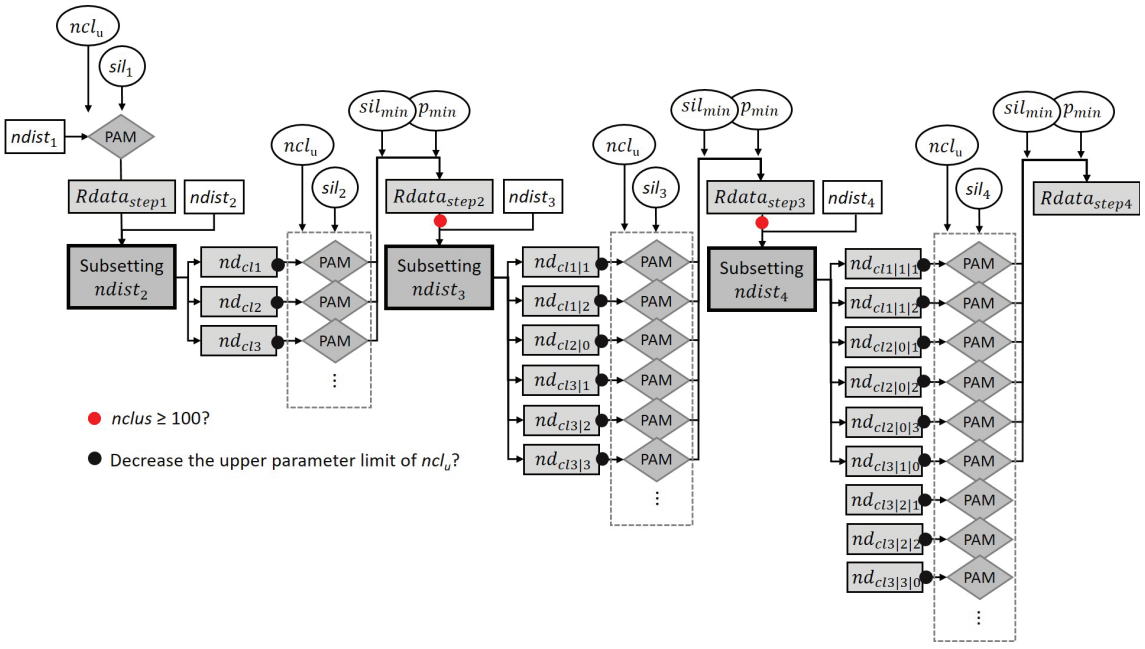


Figure 2. Multistep clustering part of the objective function of the optimization process for SPU differentiation with PAMm. All white boxes are required input data. White ovals reflect parameters that are optimized. *ndist* = normalized distance matrix, ncl_u = maximum number of clusters to test in each step, *sil* = threshold of the Silhouette Index, sil_{min} minimum Silhouette Index value, p_{min} minimum number of profiles in each cluster, *nd* = normalized distance matrix subset, PAM = partitioning around medoids clustering, *Rdata* = response data, *nclus* number of clusters.

In order not to force unreasonable splitting into a high number of clusters supported by only a low number of profiles, two criteria are tested after each step: (1) The Silhouette Index value of the nd_z cluster solution needs to be greater than or equal to the threshold value sil_{min}, and (2) the number of profiles in each resulting cluster from nd_z needs to have a minimum number of profiles p_{min}. If any of the criteria are not fulfilled, then no subdivision is conducted for the respective higher-level cluster in this step, and all profiles receive the identifier 0. p_{min} is also considered to check whether the upper parameter limit of ncl_u needs to be reduced before running PAM on nd_z. PAM is run in parallel for the respective profile subsets starting from Step 2. A stopping criterion is included to stop in case Step 2 or Step 3 leads to an overall number of clusters *nclus* of 100 or more. Seven parameters were optimized in PAMm:

- ncl_u: The maximum number of clusters considered in each step.
- $sil_{1,2,3,4}$: One Silhouette Index threshold value per step.
- sil_{min}: The minimum Silhouette Index value tolerated to accept a lower-level clustering solution.
- p_{min}: The minimum number of profiles per cluster.

Table 2 displays the respective parameter ranges. The ranges were chosen according to some test runs.

Table 2. Parameter ranges for SPU differentiation by Approach 2.

Parameter	Lower Limit	Upper Limit
ncl_u	3	10
$sil_{1,}$	0.3	0.4
$sil_{2,3,4}$	0.4	0.8
sil_{min}	0.25	0.4
p_{min}	5	10

2.3. Modeling

The multivariate parameter distributions of the SPUs obtained by PAMp and PAMm are defined by the respective groups of assigned soil profiles. To regionalize the SPUs to the continuous space and to further enhance the extraction of the already-considered soil-landscape relation, two machine learning models were trained for each of the PAMp and PAMm results using the RF algorithm and the support vector machine (SVM) algorithm. Thus, model training by machine learning was applied for three scopes:

1. for gap filling,
2. for SPU differentiation, and
3. to train the pedometric model fathoming the soil–landscape relation to obtain nation-wide and spatially continuous predictions (regionalization task).

2.3.1. Machine Learning Algorithms

The RF algorithm [48] was applied for all three scopes. It is a recursive partitioning method. Depending on the supervised learning task at hand, it grows either multiple regression or classification trees. The results of all trees are averaged. In each tree, the data are subsequently partitioned by the predictor variables into preferably homogeneous subsets regarding the response variable. The mean of each data subset (regression task) or the dominating class (classification task) is then used as the predicted response value. A partition gateway is defined by the predictor and the threshold value in its range, which achieves the most homogeneous partition into two subsets (tree branches). Overall, the stability of the tree ensemble is obtained by training each tree model with a data subset and by using a subset of all predictors. RF is known to achieve reasonable results without tuning, an important characteristic to make it the perfect choice to act as the simple and fast learner for the objective function of the optimization task for Scope (2).

The function 'cforest' of R package 'party', an RF implementation employing conditional inference trees as base learners [50], was used to train the models for gap filling. Model training involved 500 trees (training 1000 trees did not improve model performance in this particular case). The size of the predictor subset (*mtry*) was tuned via a one-dimensional grid search including one to all predictors. The function 'rfsrc' of R package 'randomForestSRC' [49] was used for the tasks of Scopes (2) and (3). It provides a fast parallel computing implementation of RF. In both cases, 1000 trees were trained. However, while for Scopes (1) and (3) the *mtry* parameter was tuned, for Scope (2), the *mtry* parameter was set to the default to speed up computation time, i.e., use RF as a fast and simple learner.

The SVM algorithm [51] was applied for the regionalization task (Scope (3)) and compared to the RF models. While RF was applied to pay tribute to the fact that the optimization might have favored an SPU differentiation whose soil–landscape relation is well captured with RF (learner in the objective function), the SVM algorithm was chosen as a powerful algorithm, which led to promising results when capturing the soil–landscape relation to generate the data product of Ließ et al. [4].

SVMs were developed by Cortes and Vapnik [51]. In binary classification tasks, they search for the hyperplane that maximizes the margin between the two classes' closest points. The properties of this decision surface ensure the SVM's high generalization ability. Points along the boundary are called support vectors. The data are projected to the higher

dimensional space via kernel techniques to allow for separation in case of nonlinearity. The radial basis function kernel was applied for this purpose. It helps to build complex decision boundaries and includes two parameters: C and γ, which need to be tuned. The γ parameter can be interpreted as the inverse of the radius of influence of the support vectors. C is the cost or penalty parameter. With a small C, the penalty for misclassified points is low; high values increase the risk of overfitting. Finally, it balances the misclassification of training samples against the simplicity of the hyperplane. R package "e1071" provides the R interface to the LIBSVM library for SVM [52,53]. To allow for multi-class classification, it uses the one-against-one technique by fitting all binary classifiers and finding the correct class by a voting mechanism. The two-dimensional parameter space to search for the optimal parameter combination expands in the following ranges: C [0.01, 100], γ [0.01, 10].

2.3.2. Model Training, Tuning, and Evaluation

For the gap-filling task, the predictor-response dataset consists of horizon-wise data (horizon sequence notation after combining dataL and dataF). For the SPU differentiation and regionalization tasks, it consists of profile-wise data. All numerical predictors were scaled to the range 0, 1 to avoid misbalance. Categorical data were kept for RF and recoded into dummy variables for SVM. To generate the predictor-response dataset for Scopes (2) and (3), the predictor values were extracted at the soil profile sites, and each soil profile was assigned to an SPU. Concerning SPU differentiation, the latter was performed in each iteration step of the objective function, as explained in Figure 1. Concerning the regionalization task, the final SPUs obtained respectively by PAMp and PAMm were used.

Model training and evaluation were conducted by a 5-times repeated 5-fold stratified CV [54] to obtain robust models. For the machine learning applications (Scope (1) and Scope (3)) involving model tuning via grid search (RF) or optimization (SVM), the CV was nested. The predictor-response dataset was subdivided into five folds of equal size using the response variable for stratification. Of these five folds, then always one fold was kept out as a test set while the other four were combined to form the model training set, leading to five separate test set evaluations (one per data instance). Each of the outer CVs' training sets was again subdivided to provide the datasets for parameter tuning in the inner CV cycle. Concerning the categorical predictors, categories not present in all data subsets were removed before model training, tuning, and evaluation. To evaluate model performance, the test set predictions were compared to the measured data to calculate the slice-wise RMSE for each of the considered soil properties. The interquartile ranges of the SPUs' multivariate distributions were used for this purpose, i.e., for each considered soil property and depth slice, it was tested whether the test set profile measurements fall within the interquartile range of the slice- and property-wise density distributions of the predicted SPU (residual of zero), whether they are smaller than the 25% quantile and how much (positive residual), or whether they are larger than the 75% quantile (negative residual). The five repetitions of the 5-fold CV resulted in 25 models and five RMSE values.

For the RF models to conduct gap-filling, a repeated 5-fold stratified group CV was applied, i.e., all horizons of a profile were assigned to the same fold to avoid overoptimistic test set estimates due to spatial autocorrelation. Concerning the regionalization task with SVM, the parameter tuning involving optimization was in a first step only conducted on behalf of one out of the 25 training sets of the outer CV cycle to check whether this provided satisfying results, while the obtained tuning parameter values were applied to all other training sets. Altogether, for the regionalization with RF and SVM of the SPUs obtained by PAMp and PAMm, four pedometric models were trained. They will be referred to as RF–PAMp, RF–PAMm, SVM–PAMp, and SVM–PAMm.

2.3.3. Variable Importance

Concerning gap filling (Scope (1)) with cforest, the package's internal variable importance (VI) measures were used. Concerning the regionalization task (Scope (3)), a different procedure was followed to allow for the comparison between SVM and RF. For model

interpretation, each predictor's importance was obtained by permuting the predictor in the test set before model application. In this way, any predictor-response relationship with regards to that predictor was eliminated. The resulting relative decrease in model performance was then attributed as vVI to the respective predictor. Values of five permutations were averaged. The VI values for the dummy variables created from each of the categorical predictors (SVM) were summed. Due to the five times repeated 5-fold CV approach (outer CV cycle), the VI plots display boxplots of 25 VI values for each predictor.

2.4. Genetic Algorithm Optimization

Genetic algorithm (GA) optimization was applied to differentiate the SPUs (Scope (2)) and to conduct parameter tuning in machine learning (Scope (3)). The GAs' operational structure is inspired by the general principles of biological evolution involving mutation, crossover, selection, and elitism [55]. The objective function for Scope (2) was described in Section 2.2 (Figures 1 and 2). The objective function for SVM parameter tuning (Scope (3)) was defined as indicated by Figure 1, Parts 3–5 while replacing Part 4 with SVM. It corresponds to the inner CV cycle (Section 2.3.2). RF (Scope (1) and Scope (3)) does not require optimization for parameter tuning [4].

The parameter space to be searched for the optimal combination of parameter values had to be predefined by providing a minimum and maximum value for each parameter. Then, a random number of *n* parameter vectors, the parent population, was evaluated by a problem-specific objective function. Weights were assigned to each parameter vector according to its objective function value before starting to modify them by conducting 'selection', 'mutation' and 'crossover' to form a new population of parameter vectors, which was again evaluated. This process was iterated until either (1) an initially defined objective function value was achieved by any of the vectors, (2) a maximum number of iterations was reached, or (3) the overall best objective function value did not improve for a certain number of consecutive iterations. GA optimization was run in parallel, subdividing the parent population of size 500 into subpopulations and allowing for limited exchange of population individuals (parameter vectors) between the so-defined islands. Twenty-five islands (20 parameter vectors per island) were used for the differentiation of the SPUs with PAMp (Scope (2)) and the tuning of the SVM models (Scope (3)). For the differentiation of the SPUs with PAMm (Scope (2)), the number of islands was reduced to 5, resulting in a subpopulation size of 100 per island. The search on the islands was not run in parallel but sequentially due to conflicts that were otherwise caused by the parallelization of the objective function.

3. Results and Discussion

3.1. Gap Filling

Gap-filling of the soil profile data was needed to calculate the slice-wise distance matrices and run PAMp and PAMm. The gaps originated from the correction for horizon sequence notation mismatches between dataL and dataF. With an average R^2 between 0.86 and 0.95, all gap-filling models displayed very good predictive performance (Figure 3B). The RMSE amounted to a mean value of 0.12 g cm^{-3} for bulk density, 5.9 Vol-% for stonesF, 3.7 Vol-% for stonesL, 5.9 g kg^{-1} for TOC, and 0.22 for pH (Figure 3A1–A4). The respective gap-filling of dataL with dataF for the particle size distribution and TOC of organic horizons remains unevaluated. It consists of the consideration of field estimates for those depth increments where laboratory data is missing, a common practice in soil science. The data are of course less precise since the KA5 soil survey instructions identify property classes instead of precise values. Errors in the class assignment were corrected by the approach presented here.

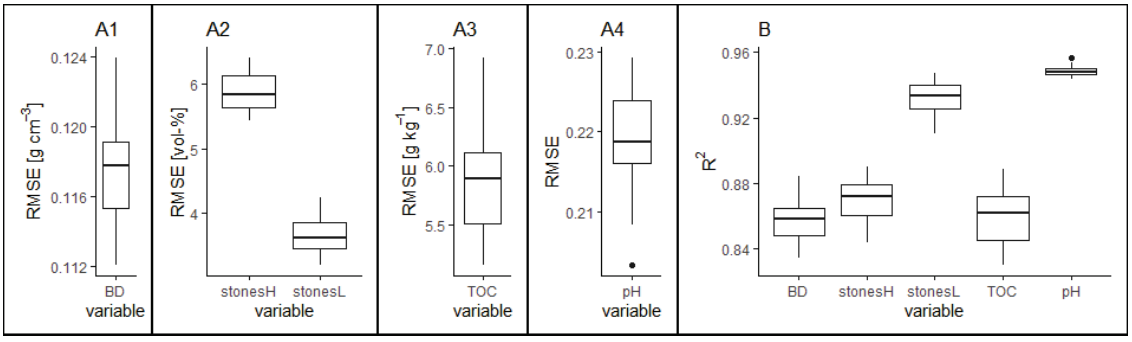


Figure 3. Predictive model performance of the RF models for gap filling. (**A**) RMSE boxplots of 25 models, (**B**) R^2 boxplots of 25 models. BD = bulk density, stonesF = stone content from dataF, stonesL = stone content from dataL, and TOC = total organic carbon content.

Data gaps in soil profile data are a common feature. Multiple approaches have been applied, including extrapolation to estimate soil properties in deeper soil horizons, gap-filling to provide estimates on behalf of expert knowledge, or assigning values from associated databases [56,57]. I am unaware, though, of any other publication documenting the use of multiple soil properties from over- and underlying horizons to train machine learning models to conduct gap filling. However, machine learning algorithms are readily applied to fill spatial data gaps in remote sensing data [58,59] and temporal gaps in time series data [60,61]. Another related field is the development of pedotransfer functions to estimate missing data of soil properties that are laborious to determine from other, readily available properties using machine learning. Ghanbarian and Pachepsky [62] provide a review.

Figure 4 displays the relative VI values for the respective gap-filling models for bulk density (Figure 4a), stonesL (Figure 4b), stonesF (Figure 4c), TOC (Figure 4d), and pH (Figure 4e). The minimum and maximum profile values, the horizon's material, the horizon's sand and silt content, as well as the underlying horizon's TOC value, were the most important predictors for gap-filling bulk density data. The stonesF data were gap-filled detecting the horizon's dataL stone content, the horizon's material and the stonesF values of the over- and underlying horizons as main predictors. For stonesL, the most important predictors were the horizon material and the horizon's dataF stone content. Gap-filling the TOC data indicated the first horizon's TOC value, the underlying horizon's TOC value (below gap), the profile's minimum TOC value, and the horizon's sand content as the most important predictors, followed by the horizon's symbol annotation as A-horizon or H-horizon. Although the gap filling was applied for mineral horizons only, there were still horizons assigned as organic (symbol_H), indicating some questionable assignments during soil profile description in the field. Gap-filling pH data indicated the horizon's sand content, the underlying horizon's TOC value, and the profile's maximum total inorganic carbon content as the most important predictors. Overall, several soil properties related to the target property were detected as important predictors in all cases. Still, for each of the target properties, there were some non-important predictors or predictors with very low VI values. Ultimately, all information which could be of any help for filling gaps with regards to the respective property were included to make sure the result with the lowest predictive uncertainty was obtained.

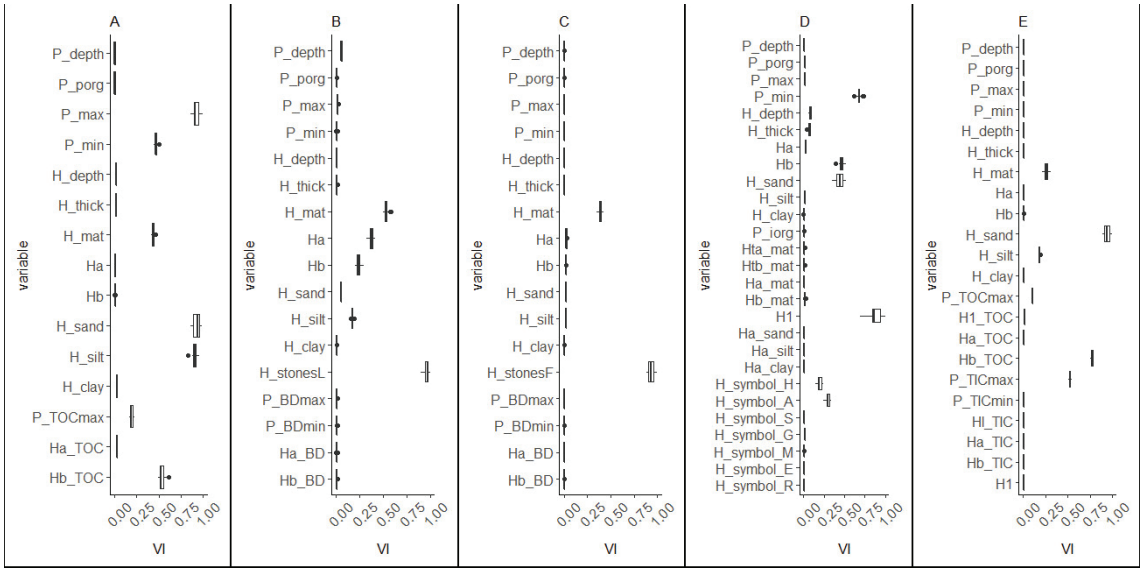


Figure 4. Relative variable importance values (VI) of the RF models for gap filling. (**A**) Bulk density, (**B**) stonesF (stone content from dataF), (**C**) stonesL (stone content from dataL), (**D**) TOC, and (**E**) pH. P_ * = value corresponding to the whole profile (* stands for any following variable indicator). H_ * = property values of the horizon to be gap-filled. Ha_ * property values of the overlying horizon, Hb_ * property values of the underlying horizon, H1 = value of the uppermost horizon, Hl = value of the last horizon, porg = percentage of organic horizons, thick = thickness, mat = horizon material (mineral, organic, bedrock).

3.2. Differentiation of Functional SPUs

The optimization to differentiate the SPUs resulted in 20 SPUs for PAMp and 47 SPUs for PAMm. Table 3 displays the resulting parameter values for PAMp, and Table 4 reports the values for PAMm. None of the parameter values was close to the upper or lower boundary of the respective parameter range, indicating that they were chosen well.

Table 3. PAMp parameters resulting from optimization to differentiate SPUs.

Parameter	P weights							*nclus*
	Texture	Stone Content	Bulk Density	Symbol_S	Symbol_G	TOC	pH	
value	0.24	0.56	0.70	0.51	0.44	0.64	0.86	20

Table 4. PAMm parameters resulting from optimization to differentiate SPUs.

Parameter	ncl_u	sil_1	sil_2	sil_3	sil_4	sil_{min}	p_{min}
value	6	0.31	0.74	0.62	0.53	0.34	13

The different P weights indicate that the profile distances with regards to the respective soil properties were assigned differing importance by PAMp. The profile distance with regards to texture was given the overall lowest importance, the distance with regards to TOC, bulk density and pH the highest, and the importance of the distance with regards to stone content, symbol_S, and symbol_G ranged somewhere in between. The P weights as such were a result of three aspects: (1) the variable types and multivariate distribution in the available soil profile data and considered soil properties, (2) the importance of the

profile distances concerning the respective properties for differentiating the clusters, and (3) how well the clusters separate in space on behalf of the available data proxies of the soil-forming factors. Aspect (1) was the reason to develop PAMp, Aspect (2) was due to the fact that for each PAMp input parameter vector, the best PAM clustering solution was chosen according to the Silhouette Index, and Aspect (3) concerned the evaluation of the respective cluster solution by the RF predictive performance. As a consequence, the P weights cannot be interpreted as a mere soil property importance for clustering.

The optimized parameter values in the second approach, PAMm, did not allow for such a direct interpretation, either. The corresponding parameters sil_1, sil_2, sil_3, and sil_4 merely provided the chance to increase the number of clusters in the respective clustering step of the multistep clustering procedure. Instead of choosing the best cluster solution in each step according to the Silhouette Index, solutions with a sufficiently good Silhouette Index value were accepted. This then, of course, also had an impact on the clustering in all subsequent steps. Figure 5 displays the subdivision tree of the step-wise procedure. Step 1 subdivided the profile data into six clusters. The best Silhouette value for this step would have led to a cluster solution with two clusters only. Hence, the sil_1 threshold of 0.31 led to this higher number of clusters obtained on behalf of the profiles' texture data. In Step 2, the subdivision with regards to symbol_S and symbol_G resulted in six clusters for Cluster 1, four for Cluster 3, three for Cluster 4, and six for Cluster 5, while there was no subdivision for Clusters 2 and 6. Six of the overall 21 clusters present after Step 2 were not further subdivided in the subsequent steps. Then, after Step 3, the dataset was already that much subdivided that further subdivision resulted in a maximum of two clusters for each of the Step 3 clusters in Step 4. During the optimization process, very different tree structures were tested, leading to this overall result. The variables in each step were selected according to their estimated importance for soil functionality. Furthermore, only variables of similar variable type and distribution were considered in each step. Applying the four steps in a different sequence would certainly have resulted in a different solution. However, previous test runs had shown this sequence to be the most promising.

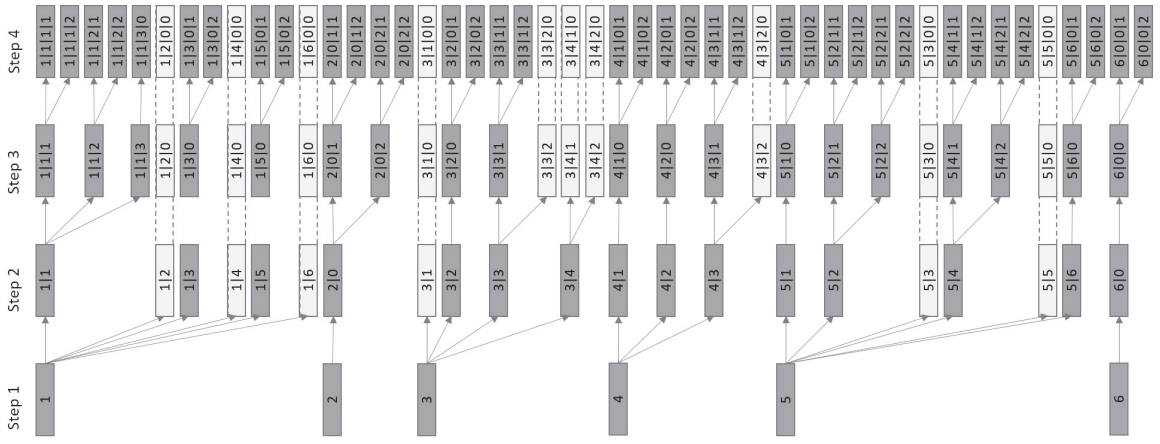


Figure 5. PAMm subdivision tree of the step-wise procedure. The light grey color indicates that the cluster obtained by the respective step is already a final cluster, which will not be further subdivided in the subsequent clustering steps.

Figure 6 shows the multivariate parameter distributions along the depth profile for the 20 SPUs resulting from PAMp. The SPUs were sorted to facilitate their description: one SPU including organic horizons (SPU 1), three leptic–skeletic SPUs (SPU 2–SPU 4) having a high stone content and depth limitation in the top 100 cm, three skeletic SPUs (SPU 5–SPU 7), four SPUs differentiated on behalf of their texture and other soil properties

(SPU 8–SPU 11), four stagnic SPUs (SPU 12–SPU 15), and five gleyic SPUs (SPU 16–SPU 20). Figure 6A1–A20 display the percentage of soil profiles composed of organic, mineral or bedrock material in the respective depth slice of the SPUs. The corresponding perc_o, perc_m, and perc_b values of the data product published alongside this manuscript replace the symbol_H, symbol_C, and symbol_mC variables of the multivariate distributions of the data product from Ließ et al. [4] in an elegant way.

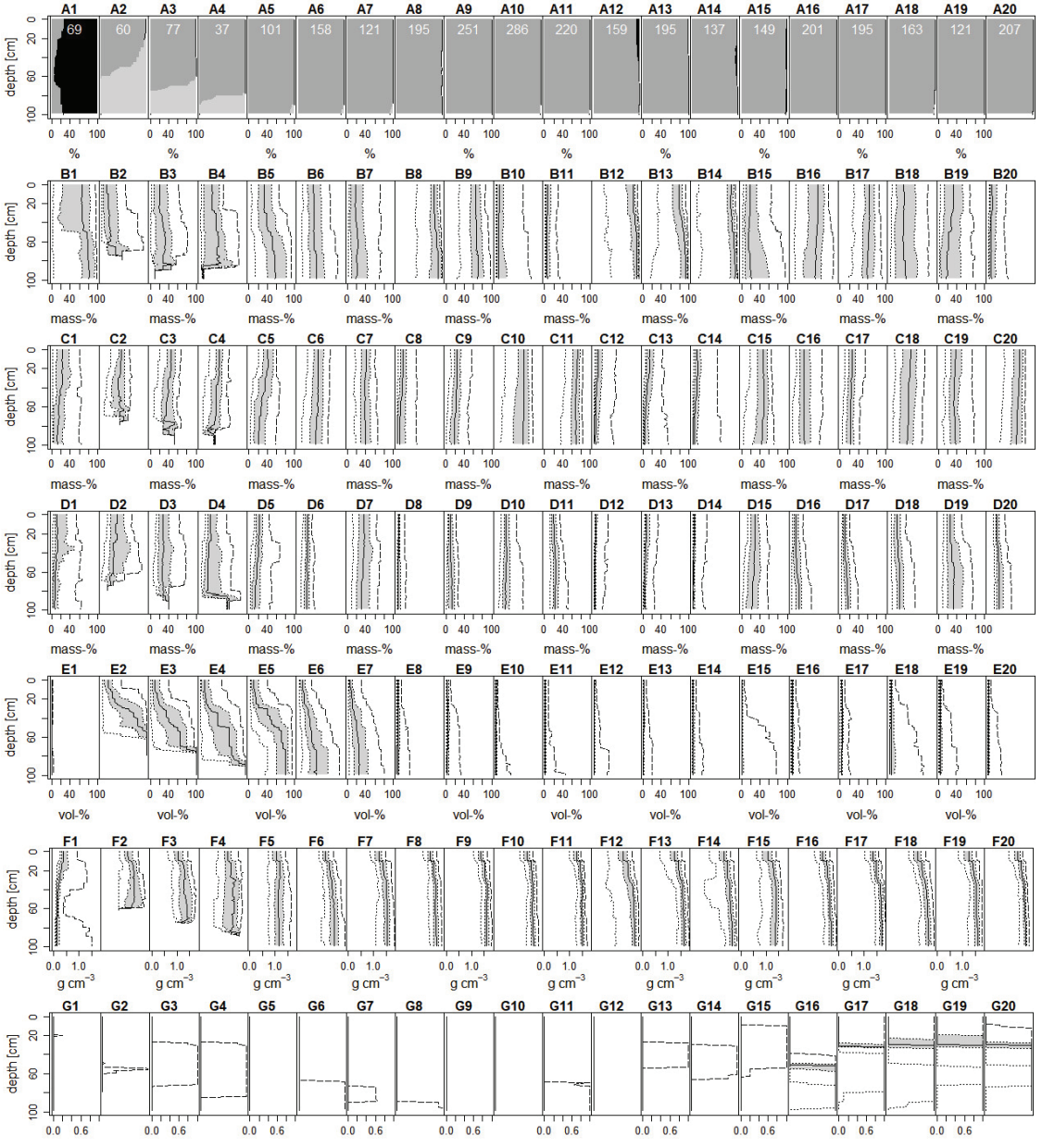


Figure 6. *Cont.*

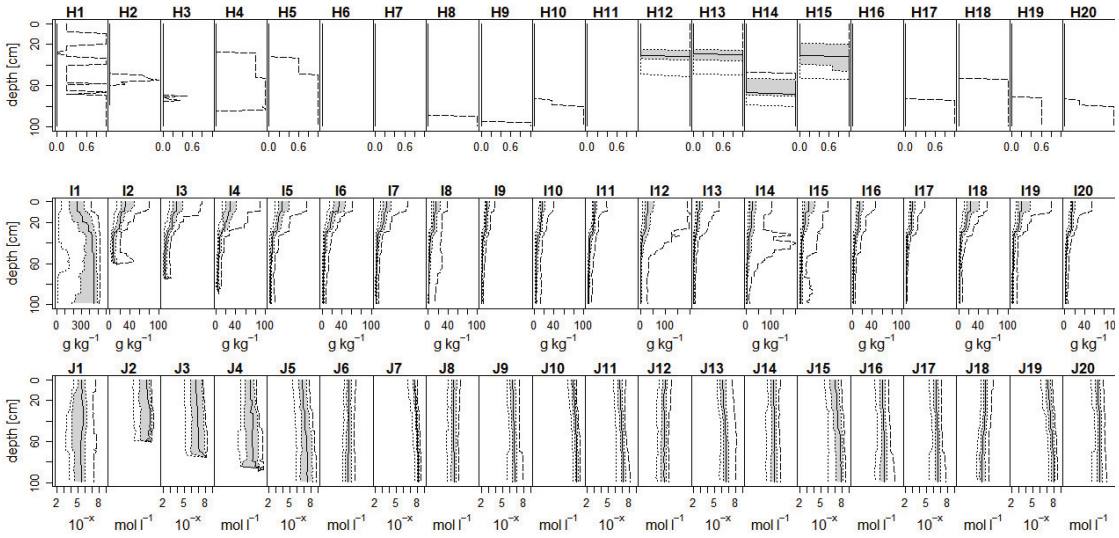


Figure 6. Multivariate soil parameter distributions along the depth profile of the SPUs obtained with PAMp. The figure columns reflect the respective SPUs 1 to 20, figure lines refer to the various soil properties. (**A**) The slice-wise contribution of profiles with mineral properties (dark grey), organic properties (black) or bedrock (light-grey). The white numbers indicate the number of profiles supporting the respective SPU. (**B**) Sand content, (**C**) silt content, (**D**) clay content, (**E**) stone content, (**F**) bulk density, (**G**) symbol_S, (**H**) symbol_G, (**I**) TOC, and (**J**) pH. In figure (**B**) to (**J**), the solid line indicates the median of the distribution, the shaded area between dotted lines reflects the interquartile range, the other dotted line reflects the 5% quantile, and the dashed line reflects the 95% quantile. Please be aware that figures (**I**) have different X-axis ranges, namely (**I**1) = 0–600, (**I**12) to (**I**15) = 0–200, and all others = 0–100.

SPU 1 corresponds to agricultural soils that are made up of organic material in one or more horizons along their profile (Figure 6A1). The particle size distribution in its mineral horizons shows the maximum variation among all SPUs in terms of sand content (Figure 6B1). It lies between 0–5% and 96–98%, taking into account the slice-wise 5 and 95% quantiles of the distribution along the depth profile. Looking at the interquartile range, the variation in sand content in the top 49 cm still ranges between 12–25% and 82–84%. The overall median TOC and also the variation in TOC are the highest among all SPUs. Considering the interquartile range, the TOC ranges between 27–358 and 331–492 g kg^{-1} throughout the profile. Regarding the low number of profiles with organic horizons contained in the dataset, this high variation in TOC and soil texture is not surprising. The high variability of soils in SPU 1 cannot be further subdivided by PAMp, allowing for a maximum of 100 clusters. Additionally, some of the profiles including organic horizons are still included in the other SPUs (compare, e.g., Figure 6A12,A14). The same was also reported by Ließ et al. [4]. Likewise, a perfect separation into all-mineral and partly mineral soils in the first step of PAMm was also not successful, while the mere assignment to organic or non-organic of the respective slice was considered, or additional soil properties such as TOC (previous test runs) were included. However, a further subdivision of this SPU could likely be achieved by increasing the dataset of these partly mineral soils. Meanwhile, an alternative could be to conduct a previous subdivision into all-mineral and partly mineral soils, and then apply PAMp and PAMm to each of the two groups separately.

The SPUs 2–7 have a rather high stone content increasing with depth (Figure 6E2–E7). Of these six SPUs, SPUs 2–4 have a depth limitation within the top 100 cm (Figure 6A2–A4). They differ in the strength of this depth limitation, though. SPU 5 displays the same strong increase in stone content with depth comparable to the SPUs 2–4, whereas SPU 6 and

SPU 7 have a smaller increase. Furthermore, the SPUs 5–7 also differ in their particle size distribution: their sand content is decreasing from SPU 5 to SPU 7 (Figure 6B5–B7).

The SPUs 8–11 also have a decreasing sand content (Figure 6B8–B11). I will refer to SPU 8 and SPU 9 as sandy and to SPU 10 and SPU 11 as silty SPUs. Three of these SPUs (SPU 9, SPU 10, and SPU 11) are also the SPUs with the overall highest number of profiles (Figure 6A9–A11). Apart from their texture, these four SPUs differ in their pH (Figure 6J8–J11), with SPU 8 having the lowest and SPU 10 the highest pH value. For SPU 8, this corresponds to a pH between 5.2–5.6 and 5.9–6.0 throughout the profile; for SPU 10, it corresponds to a pH between 7.3–7.9 and 7.8–8.3 (interquartile range). A similarly high pH value is attributed to SPU 7, SPU 15, and SPU 19, indicating that there is one such SPU in each group: the skeletic SPUs, the texture SPUs, the stagnic SPUs, and the gleyic SPUs.

The SPUs 12–20 have hydromorphic properties in some part of their profile. Of these, the SPUs 12–15 have a horizon with stagnic properties (Figure 6H12–H15), and the SPUs 16–20 indicate ground water influence (Figure 6G16–G20). Still, the presence of the 95% quantile in most of the other SPUs indicates that a few soil profiles with hydromorphic properties have also been assigned to these SPUs. PAM clustering to separate soils with and without stagnic properties and soils with and without gleyic properties merely on the *nSPdist* of symbol_S or symbol_G, respectively, also did not succeed in providing a perfect separation (test runs). Ließ et al. [4] did not achieve this, either. However, it has to be noted that the two SPUs with gleyic and two SPUs with stagnic properties of the data product by Ließ et al. [4] were now extended to five and four SPUs, respectively. The SPUs 12–15 indicate a high TOC consistent with hydromorphic conditions that reduce organic matter decomposition (Figure 6I12–I15). The median TOC in the top 20 cm ranges between 16 and 38 g kg^{-1} for these SPUs, while it lies between 10 and 16 g kg^{-1} for SPUs 8–11. SPUs 2–7 and 18–19 have a comparatively higher variation in the TOC in their top 10 cm, indicating that they include grassland soils. This is reasonable given that SPUs 2–7 have high stone contents and are likely to occur in inclined areas, and SPUs 18–19 have groundwater influence at shallow depth. Furthermore, due to their comparatively lower topsoil TOC values, it is likely that most of the soil profiles assigned to SPU 16, SPU 17 and SPU 20 were drained to be used for crop cultivation or were cultivated with crops that do not mind waterlogging at a low rooting depth. While SPUs 12–14 have a rather high median sand content and differ due to the depth of their stagnic horizon and their pH value (Figure 6J12–J14), SPU 15 has a low median sand content and correspondingly higher silt and clay contents (Figure 6B15,C15,D15).

Compared to the data product from Ließ et al. [4], the ranges between the 5 and 95% quantiles and the interquartile ranges of the SPUs' multivariate parameter distributions regarding the particle size distribution, bulk density and stone content were reduced. With regard to the stagnic and gleyic properties, Ließ et al. [4] included prediction probabilities instead of quantiles. These were low in the upper part of the profile, then increased with depth in a transition zone of 30 cm and were high in the lower part of the profile. Considering the interquartile ranges of the multivariate distributions related to symbol_S and symbol_G, these transition zones were smaller for all gleyic SPUs and the stagnic SPUs 12–14 but similar for the stagnic SPU 15.

3.3. Pedometric Modeling to Capture the Soil–Landscape Relation

3.3.1. Model Performance

Figure 7 displays the property-wise predictive model performance for the four models RF–PAMp, RF–PAMm, SVM–PAMp, and SVM–PAMm. The performance measure of the approach always depends on two aspects: (1) the statistical dispersion of the multivariate parameter distributions of the SPUs resulting from PAMp or PAMm and (2) the performance of the machine learning algorithm to extract the soil–landscape relation. Consequently, the evaluation of the data product was best achieved in a sense of the predictive RMSE of the individual soil properties.

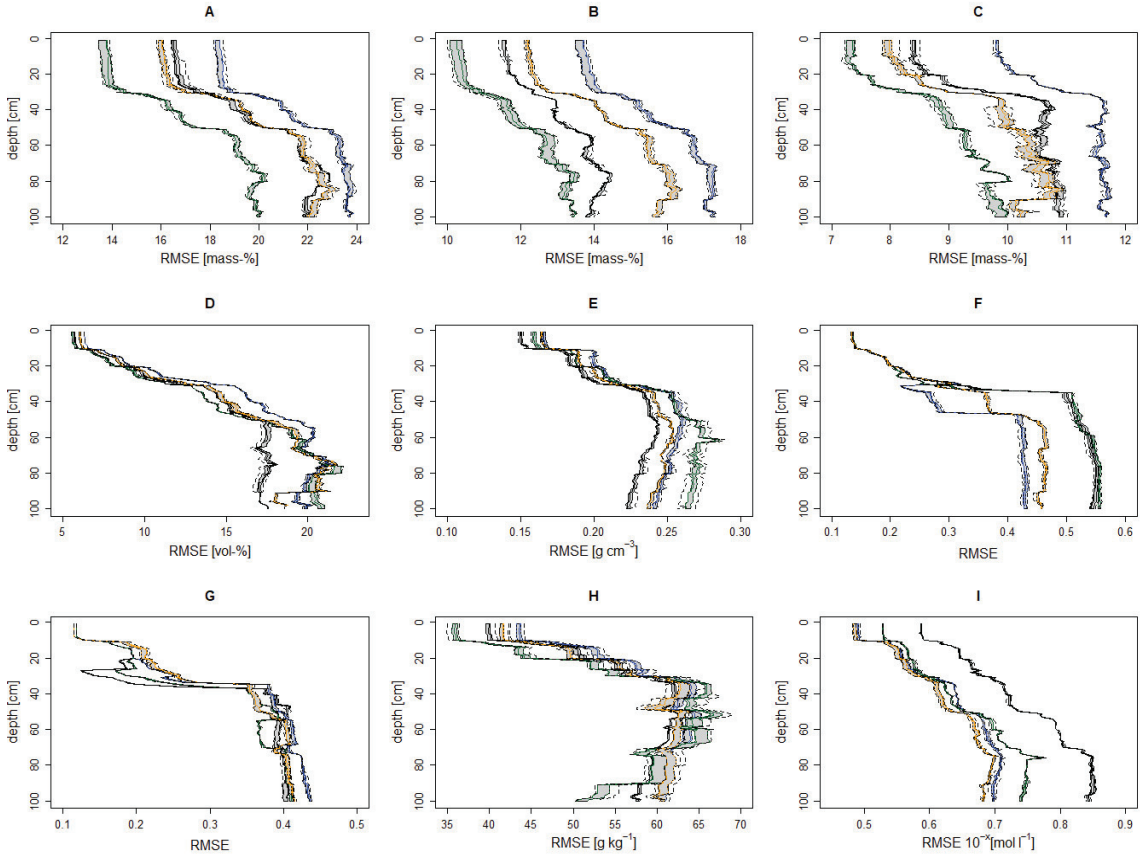


Figure 7. Predictive model performance considering the interquartile range of the SPUs' multivariate distribution along the depth profile. (**A**) sand content, (**B**) silt content, (**C**) clay content, (**D**) stone content, (**E**) bulk density, (**F**) symbol_S, (**G**) symbol_G, (**H**) TOC, and (**I**) pH. The colors reflect the different models: black = RF−PAMp, blue = RF−PAMm, green = SVM−PAMp, yellow = SVM−PAMm. The lines along the shaded area correspond to the lower and upper hinges of the five predicted values (repeated CV), the solid line to the median, and the dotted lines to the upper and lower whiskers.

Regarding soil texture, predictive model performance always detects SVM–PAMp as the best model and RF–PAMm as the least promising, whereas the priority between SVM–PAMm and RF–PAMp favors SVM–PAMm for sand and clay content and RF–PAMp for silt (Figure 7A–C). SVM–PAMp is also the most promising among the four models concerning its predictive performance in terms of the stone content up to a depth of 60 cm (Figure 7D), the prediction of gleyic properties, and the TOC (Figure 7H). Below 60 cm, RF–PAMp shows the best performance for the stone content (Figure 7D). Additionally, this model has the best performance concerning bulk density (Figure 7E). Predicting pH, SVM–PAMm shows the best performance. However, the RMSE of RF–PAMm and SVM–PAMp are only slightly higher. Model performance in reference to stagnic properties is hardly distinguishable between the four models until a depth of 30 cm. This similarity continues for SVM–PAMp and RF–PAMp in the subsoil, while the RMSE of RF–PAMm and SVM–PAMm does not increase as much, resulting in RF–PAMm being the overall best for this property. Altogether, this makes SVM–PAMp the best model for three out of seven soil properties and minor differences for a fourth property.

This indicates the high power of the SVM algorithm when combined with GA optimization for parameter tuning. In contrast, it was expected that RF might result in the overall better algorithm due to its usage in the objective function for SPU optimization. However, the results were ambivalent. RF resulted in the better algorithm for two properties, and SVM for four properties. Overall, this enhances the critical discussion on the common perception that RF is often stated to have the best predictive performance when comparing multiple machine learning algorithms in pedometric modeling applications [63]. The comparison is usually not conducted appropriately since RF does not require much tuning and its most important parameters are natural numbers and, therefore, the common grid-search approach is sufficient. In contrast, the training of SVMs requires thorough tuning of real-valued parameters [4,25]. A fair comparison of the two algorithms is, therefore, only possible if optimization is applied for tuning SVM models.

The overall model performance was decreasing with depth concerning all soil properties, as is commonly perceived in pedometric modeling (e.g.). Figure 7 shows that this decrease was non-linear. For the topsoil, it usually had very good performance, which then rapidly decreased at a certain soil depth. The threshold value differed between the soil properties, though. For the particle size distribution (Figure 7A–C) it ranged around 25 cm, and for the other soil properties, around 10 cm depth (Figure 7D–I). Some of the latter had two steps in the performance decrease, one at 10 cm and another at 25 or 30 cm (bulk density, Figure 7E), 30 or 50 cm (symbol_S, Figure 7F), at 40 cm (symbol_G, Figure 7G), or 25 cm (TOC, Figure 7H) depth. The good topsoil performance with regards to the hydromorphic features was probably due to their onset at a certain soil depth. The other step was likely caused by grassland soils not being separated from cropland soils in the SPU differentiation. This could mean that the difference between grassland soils and cropland soils was minor either with regards to the vertical soil profile differentiation and characteristics or regarding the soil–landscape relation. Concerning the latter, the high number of SCORPAN O predictors from remote sensing data provides a good representation of the land cover and would, therefore, easily allow for this separation between the grassland and cropland soils. With the former, it must be taken into account that the difference between the two only affected a limited number of the considered properties, and then only the respective topsoil. However, this aspect could only be addressed while the calculation of the property-wise profile difference assigned a higher weight to the topsoil differences for these soil properties. The decision on assigning different weights along the depth profile was not trivial, though. A few test runs were conducted with an exponential weight decay function and a step-wise approach. Additionally, optimizing the weights along the depth profile in addition to the already-implemented optimization tasks in PAMp and PAMm would add to the complexity of the objective function and prolong the optimization process to differentiate the SPUs. I would further like to note that the comparison of the RMSE values along the soil profile for certain soil properties can be misleading, as the respective value ranges differed between the various soil depths. This was clearly visible for TOC (Figure 7H), where the predictive model performance seemed to improve at a certain soil depth. However, the lower RMSE values were likely caused by the lower TOC range at this higher soil depth.

In the following, the multivariate 3D data product will be compared to other readily available data products. This is achieved by referring to the predictive median RMSE with regards to the interquartile range of the multivariate parameter distributions along the depth profile. On the one hand, the property- and depth-wise uncertainty will be compared to its first version from Ließ et al. [4]. On the other hand, the national performance estimates (considering agricultural soils) for other spatially continuous data products covering the entirety of Germany were calculated. Table 5 provides an overview. They were evaluated by extracting the predicted property values at the soil survey sites of the test set profile data, which had been used to evaluate the data product developed here. The weighted mean was calculated for the respective depth layer before calculating the RMSE. The compared

data products had the following spatial raster resolutions: national scale—100 m [25,64], European scale—500 m [65] and 1000 m [66], and global scale—250 m [67].

Table 5. National-scale evaluation (RMSE) of existing national, European and global-scale data products (considering agricultural soils). The predictive uncertainty was evaluated on behalf of the test set profile data. The values of the raster data products were extracted at the profile sites. A weighted average was calculated for the respective depth interval of the measured data.

Scale of the Data Product	Depth Interval [cm]	Sand Content [Mass-%]	Silt Content [Mass-%]	Clay Content [Mass-%]	Stone Content [Vol-%]	Bulk Density [g cm^{-3}]	TOC [g kg^{-1}]	pH 10^{-x} mol L^{-1}
National	0–30	15.0 [25]	11.8 [25]	8.2 [25]	-	-	22 [64]	-
European	0–20	17.6 [65]	13.8 [65]	9.8 [65]	9 [65]	0.26 [65]	48.3 [66,68]	-
Global [67]	0–5	19.3	16.5	11.4	7.1	0.30	43.6	1.2
	5–15	19.4	16.4	11.0	7.8	0.30	46.2	1.2
	15–30	19.9	17.6	11.7	10.5	0.31	57.6	1.2
	30–60	22.9	18.7	13.8	17.5	0.35	62.4	1.3
	60–100	25.9	19.6	14.3	21.2	0.36	60.7	1.4

With regards to the particle size distribution, the predictive performance improved compared to Ließ et al. [4]. For the sand content, it improved from 14.8 to 13.8 mass-% at 20 cm depth, from 17.5 to 16.3 mass-% at 40 cm depth, and from 20.2 to 19 mass-% at 60 cm depth. Respectively, it improved from 10.7 to 10.4, from 12.2 to 11.6, and from 14.3 to 12.7 mass-% for the silt content, and from 8.2 to 7.5, from 10.1 to 8.9, and from 10.1 to 9.3 mass-% for the clay content. Figure 7A–C show the continuous performance estimates. Concerning the topsoil, the national scale 0–30 cm [25], the European scale 0–20 cm [65], and the global scale 15–30 cm predictions [67] had a higher uncertainty with an RMSE of 15.0, 17.6, and 19.9 mass-% for sand, 11.8, 13.8, and 17.6 mass-% for silt, and 8.2, 9.8, and 11.7 mass-% for clay (Table 5), respectively. For the subsoil, the global scale 30–60 cm predictions [67] also had a higher RMSE. They amounted to 22.9 mass-% for sand, 18.7 mass-% for silt, and 13.8 mass-% for clay (Table 5).

Compared to Ließ et al. [4], the predictive performance concerning the stone content remained more or less the same in the 20 cm depth with 8.1 versus 8.0 vol-%, improved for 40 cm depth from 14.8 to 13.9 vol-%, but was impaired in the 60 cm depth from 16.9 to 19.1 vol-%. For the topsoil, the European (0–20 cm) and global scale (15–30 cm) predictions had a slightly higher uncertainty, with an RMSE of 9 and 10.5 vol-% (Table 5), respectively. Considering the same depth intervals, the RMSE of the data product created here corresponded to an average RMSE of 6.5 vol-% for the 0–20 depth interval and 9.1 vol-% for the 15–30 cm depth interval. This even higher difference in reference to the European data product is due to the overall decrease in uncertainty with lower soil depth (Figure 7D).

For bulk density, the predictive performance was impaired at 20 and 60 cm depths from 0.15 to 0.19, and 0.25 to 0.27 g cm^{-3}, but remained the same at the 40 cm depth compared to Ließ et al. [4]. The predictive topsoil uncertainty was still higher for the European and global data products with an RMSE of 0.26 and 0.31 g cm^{-3}, respectively. The same applied to the subsoil with an RMSE of 0.35 g cm^{-3} (global predictions 30–60 cm, Table 5).

The predictive model performance along the depth profile with regards to the TOC is displayed in Figure 7H. TOC was not part of the data product generated by Ließ et al. [4]. The averaged RMSE for the respective depth interval was 39.3 compared to 48.3 g kg^{-1} for Aksoy et al. [66] in the 0–20 cm interval, 43.8 compared to 21 g kg^{-1} for Sakhaee et al. [64] in the 0–30 cm interval, 38.8 compared to 46.2 g kg^{-1} in the 5–15 cm, and 49.8 compared to 57. 6 g kg^{-1} in the 15–30 cm interval for Poggio et al. [67]. This means the data product developed here had a lower predictive topsoil uncertainty compared to the global and European data products, but a higher uncertainty compared to the national data product. One of the reasons for the latter is the high diversity in the soils containing an organic horizon in some part of their profile. The low number of soil profiles representing these soils in

the dataset of the agricultural soil inventory had also caused trouble for Sakhaee et al. [64]. They addressed this aspect by training separate models for organic and mineral topsoil, which resulted in an RMSE decrease from 31.6 to 21.0 g kg^{-1}. The complexity increases, though, while multiple properties are jointly considered in 3D. The optimization to differentiate the SPUs merged all these soils into a single SPU (SPU1, Figure 6A1). Compared to the high difference in TOC content between these soils and the all-mineral soils, the TOC differences among the all-mineral soils were minor. Conducting the cluster analysis while applying data transformation to this and other soil properties before calculating the distance matrices did not solve the issue, either. A solution might be to subdivide the data into all-mineral and partly mineral soils and then conduct two separate optimization processes to differentiate the SPUs in each subgroup, as suggested earlier.

The predictive model performance along the depth profile with regards to the pH is displayed in Figure 7I. The pH was not part of the data product generated by Ließ et al. [4]. The averaged RMSE for the respective depth interval was $10^{-0.55}$ compared to $10^{-1.2}$ mol l^{-1} in the 5–15 cm, and $10^{-0.58}$ compared to $10^{-1.2}$ mol l^{-1} in the 15–30 cm interval for Poggio et al. [67].

Overall, the models presented here deal with high complexity: They address the multivariate soil variability in 3D compared to the models trained to obtain the univariate 2D data products. It is impressive that a lower predictive uncertainty was still achieved. The lower uncertainty compared to the European and global data products is likely because at national scale for Germany, there are many more data proxies available to approximate the soil-forming factors, namely the expert information contained in the national map products providing information on the soil distribution [41] and parent material [39,40]. This helps in capturing the soil–landscape relation by machine learning. In reference to the national scale data products, a higher performance was achieved for texture, but a lower performance for TOC due to the previously mentioned reasons. Finally, it has to be emphasized that the data product presented here differs from the others. The univariate predictions (single soil property) considered in the comparison provide single-cell predictions for a certain depth interval. In contrast, the data product developed here provides 3D soil information in terms of the multivariate distributions. Its spatial resolution in the 2D mapping space is 100 m, and the resolution along the depth profile is 1 cm. Accordingly, for each raster cell, it provides the slice-wise multivariate distribution of the respective soil properties. It would be inappropriate to consider the median of these distributions for each raster cell. The benefit lies in considering these distributions, which are the consequence of condensing the information contained in the raster cells to a limited number of functional SPUs.

3.3.2. Variable Importance

The VI values (Figure 8) indicate that all predictors were important to a certain extent for all four models. The values are relative, and not comparable between the models.

However, what separates the SVM models (Figure 8C,D) from the RF models (Figure 8A,B) is the high importance they assign to the categorical predictors in comparison to the other predictors. These categorical predictors reflect the inclusion of expert knowledge with regards to parent material and soils included in conventional map products (BAG00, LIT00, STR00, and BGL00) as well as the classified topography (GMK00). Categorical predictors had also proved highly important for the models of the first implementation to represent the agricultural soil landscape of Germany by SPUs [4]. It is unfortunate in this regard that further categorical SCORPAN S and SCORPAN P predictors available at a larger map scale could not be included (e.g. [17,69]). The soil profile database of the agricultural soil inventory does not include sufficient data entries to represent the high number of SMUs included in these maps. The RF models do not prioritize the categorical information, though. This is surprising, as they are known to generally favor categorical predictors [70,71]. In contrast to the latter, they assign comparatively higher importance to the DEM.

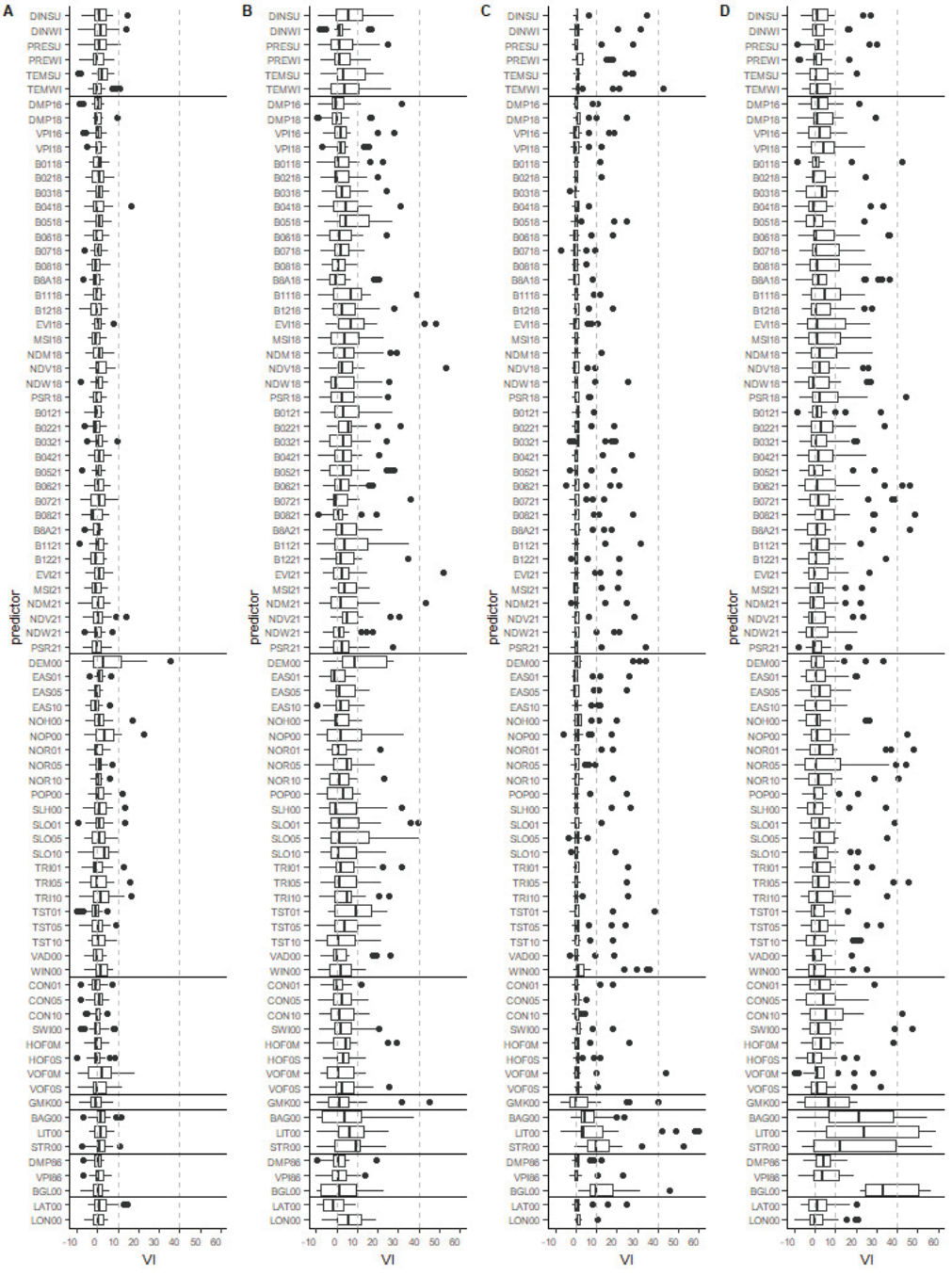


Figure 8. Variable importance (VI) boxplots of the models for SPU regionalization. (**A**) RF–PAMp, (**B**) RF–PAMm, (**C**) SVM–PAMp, and (**D**) SVM–PAMm. The horizontal lines separate the respective predictor groups corresponding to the SCORPAN factors: climate, organisms, relief (topography), relief (hydrology), relief (categorical), parent material, soil, and latitude and longitude. Please refer to Table 1 for the predictor abbreviations.

3.3.3. Nationwide Prediction

Figure 9 displays the map of the nationwide prediction of the SPUs with model SVM–PAMp. In the following, it will be described from north to south according to the four morphologic regions of Germany: the North German Lowland, the Central Germany Uplands, the Alpine Foreland, and the Alps.

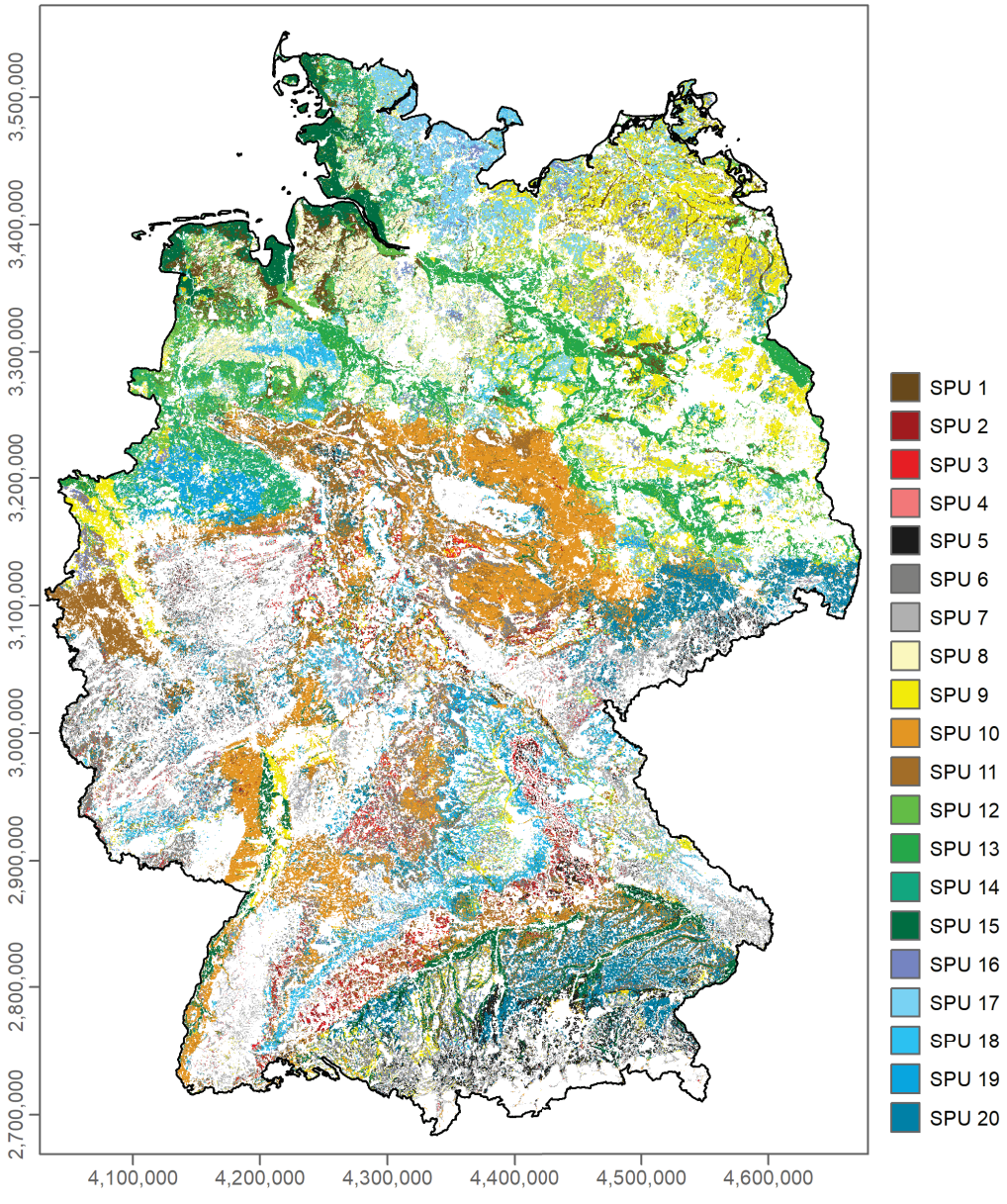


Figure 9. Map of Germany displaying the distribution of the SPUs corresponding to model SVM–PAMp. Colors were selected to emphasize the groups: SPU 1 organic, SPU 2—SPU 4 leptic–skeletic, SPU 5—SPU 7 skeletic, SPU 8—SPU 9 sandy, SPU 10—SPU 11 silty, SPU 12—SPU 15 stagnic, and SPU 16—SPU 20 gleyic SPUs. Non-agricultural areas are masked. Coordinate reference system EPSG 3035.

The North German Lowland presents a mixture of sandy soils (SPU8, SPU 9), stagnic soils (SPU 12–SPU 15) gleyic soils (SPU 16–SPU 19) and patches of the organic SPU 1. Of the sandy soils, SPU 8 dominates in the west, and SPU 9 in the east. SPU 8 has higher sand and correspondingly lower pH values (Figure 6B8,B9,J8,J9). The higher topsoil TOC values of SPU 8 likely originate from the land-use history in this region. Nutrient-poor, sandy topsoil was often improved by mixing it with grass or heather plagues [4]. Stagnic SPU 15 is found along the North Sea coast in the marshland under tidal influence. It is this stagnic SPU that differs from the other stagnic SPUs due to its much lower sand and correspondingly higher silt and clay contents. SPU 14 dominates in the northernmost part, right between the North and Baltic Seas. It is the stagnic SPU whose stagnic properties start at a higher soil depth compared to the others. SPU 12 and SPU 13 are found in the floodplains and lower terraces of the rivers Weser, Elbe, and Oder. The gleyic soils in the north are dominated by SPU 17 along the east coast (Baltic Sea), with patches of this SPU as well as SPU 16, SPU 18, and SPU 19 further inland. SPU 14 and SPU 19 also dominate the area in the southwestern-most part of the North German lowland corresponding to the lowlands of the glacial valleys of the old moraine area [41].

In the Central German Uplands, the Loess plains are represented by SPU 10 and SPU 11. Considering their multivariate distributions, they are mainly differentiated by their pH, with SPU 10 having the higher pH values (Figure 6J10,J11). The gleyic SPU 20 dominates the loess plains in Saxony. It has high silt contents similar to those of SPU 10 and SPU 11. However, large parts of the Central German Uplands are covered by the leptic–skeletic SPUs 2–4 and skeletic SPUs 5–7, which are distinguished by their high stone contents. Of these, SPU 7, with much lower sand contents and correspondingly higher pH values (compared to SPU 5 and SPU 6), dominates. Still, large parts along the Swabian Alp, the Franconian Alp, Spessart, and Franconian Switzerland display high coverage by leptic–skeletic SPU 2, the SPU with the highest depth limitation. The gleyic SPU 18 covers large parts along these mountain ranges. The lower Rhine valley stands out by the domination of sandy SPU 9. Between the cities Karlsruhe and Mainz, SPU 9 is then accompanied by the stagnic SPU 15 with its much lower sand contents. SPU 15 also dominates along the floodplains of the Danube and tributary rivers, which separate the Central German Uplands from the Alpine Foreland. Regarding the considered soil properties, these soils are similar to those along the North Sea coast. To distinguish them from one another, additional soil properties would have to be included. The soils might differ in their electrical conductivity due to the tidal influence along the North Sea coast.

Large parts of the northeast of the Alpine Foreland are covered by the siltic SPU 10 as well as the gleyic SPU 20, having a similar texture. This indicates the similarity of these soils to the Loess plains. Additionally, they co-occur with gleyic SPU 16, which has higher sand contents. Large parts of the remaining region are dominated by the leptic–skeletic SPU 2 and SPU 4, while patches of the sandy SPU 9 and organic SPU 1 are also clearly distinguishable. Large parts of the Alps are not under agricultural use. Those that are often contain high stone contents (SPU 3, SPU 5, and SPU 6) and are partly limited in depth (SPU 3). Still, the sandy SPU 9, silty SPU 10, stagnic SPU 15, and gleyic SPU 16 also occur.

Overall the number of SPUs increased from 8 to 20 in comparison to Ließ et al. [4], providing a more detailed spatial differentiation. The previous single SPU with a high stone content and a depth limitation in the top 100 cm is now augmented to six SPUs with a high stone content, of which three additionally have a depth limitation in their top 100 cm. The two SPUs with stagnic and two with gleyic properties were augmented to four and five, respectively. The SPUs with a predominantly sandy or silty texture were augmented from one SPU to two SPUs in both cases. Simply, the SPU including soils with organic horizons remained only one, another hint to consider the separate differentiation into SPUs for the all-mineral and partly mineral soils.

The pattern of the spatial allocation of the SPUs shows some similarity with regards to the national-scale soil map products BÜK200 and BÜK1000 [17,72]. This was expected considering the high importance of the SCORPAN P, SCORPAN R, and SCORPAN S

predictors. Ultimately, the national soil maps also heavily rely on topography and parent material. As mentioned previously, the information contained in the spatial units differs. Complex SMUs composed of multiple co-occurring soils differing largely in their profile characteristics are by no means comparable to spatially allocated SPUs, each being described by a multivariate parameter distribution along the depth profile. It is interesting to note, though, that the data product provided here is a national-scale representation with much fewer SPUs than the SMUs in these soil maps.

4. Conclusions

The national-scale evaluation and modeling of the impact of agricultural management and climate change on soils, crop growth, and the environment require soil information at a spatial resolution addressing individual agricultural fields. The agglomeration of the soil parameter space into a limited number of functional SPUs allows for reducing the required resources to run agricultural process models without having to cut back on the spatial resolution. To serve these needs, creative data science approaches are needed.

Here, two data science approaches were developed involving unsupervised classification to generate a multivariate 3D data product of spatially allocated functional SPUs, each being defined by a multivariate parameter distribution along the depth profile from 0 to 100 cm. The two methods account for differences in variable types and distributions and involve genetic algorithm optimization to identify those SPUs with the lowest internal variability and maximum inter-unit difference with regards to both their soil characteristics and landscape setting.

The high potential of these two approaches was demonstrated by applying them to the agricultural German soil landscape. The resulting data product consists of 20 SPUs that are each described by a multivariate parameter distribution along the depth profile from 0 to 100 cm. It comes along with property- and depth-wise uncertainty estimates. Its spatial resolution in the 2D mapping space is 100 m, and the resolution along the depth profile is 1 cm. It is available in a reduced storage format consisting of two related files, (1) a nationwide raster file with identifiers pointing to (2) the respective multivariate distribution for each functional SPU provided in table format. Each property's distribution is represented by the 5, 25, 50, 75 and 95% quantiles.

The spatial pattern of the nationwide raster shows some similarity with the national soil maps of Germany. The information contained in the spatial units differs, though. Complex SMUs composed of multiple co-occurring SUs of very different characteristics are by no means comparable to spatially allocated SPUs that are each represented by a multivariate parameter distribution. Furthermore, it is interesting that the data product created here is a national-scale representation with significantly fewer SPUs than the SMUs in these soil maps. Additionally, the boundaries of the SPUs differ from those of the SMUs. Why the boundaries differ and whether the number of SPUs would increase if a larger soil profile database is included are two aspects that are valuable to investigate together with colleagues from the soil survey institutes.

The created data product is the second version of such a 3D soil-landscape model for the agricultural landscape of Germany. Compared to Version 1, the number of SPUs increased, and the respective interquartile range of the multivariate distributions and the predictive uncertainty were reduced. Additionally, two further soil properties, TOC and pH, were included. Version 2 of the data product also has a lower uncertainty compared to existing univariate 2D data products while considering the interquartile range of the multivariate distributions. I recommend using them as margins to run agricultural process models. Limitations concerning TOC uncertainty suggest considering all-mineral and partly mineral soils separately in the SPU differentiation. Whether the available data are sufficient to follow such an approach would have to be tested, though.

Funding: This work was funded by the German Federal Ministry of Education and Research (BMBF) in the framework of the funding measure Soil as a Sustainable Resource for the Bioeconomy—BonaRes, project BonaRes (Module B): BonaRes Centre for Soil Research (grant 031B0511). For further information, please visit www.bonares.de.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data product for this study is published on the BonaRes Repository under the following reference: Ließ, M. (2022). 3D soil parameter space of the agricultural landscape [Germany, Version 2]. DOI: https://doi.org/10.20387/bonares-13qm-mw25.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Kapoor, D.; Bhardwaj, S.; Landi, M.; Sharma, A.; Ramakrishnan, M.; Sharma, A. The Impact of Drought in Plant Metabolism: How to Exploit Tolerance Mechanisms to Increase Crop Production. *Appl. Sci.* **2020**, *10*, 5692. [CrossRef]
2. Magombeyi, M.S.; Taigbenu, A.E.; Barron, J. Effectiveness of Agricultural Water Management Technologies on Rainfed Cereals Crop Yield and Runoff in Semi-Arid Catchment: A Meta-Analysis. *Int. J. Agric. Sustain.* **2018**, *16*, 418–441. [CrossRef]
3. Hatfield, J.L.; Dold, C. Water-Use Efficiency: Advances and Challenges in a Changing Climate. *Front. Plant Sci.* **2019**, *10*, 103. [CrossRef] [PubMed]
4. Ließ, M.; Gebauer, A.; Don, A. Machine Learning With GA Optimization to Model the Agricultural Soil-Landscape of Germany: An Approach Involving Soil Functional Types With Their Multivariate Parameter Distributions Along the Depth Profile. *Front. Environ. Sci.* **2021**, *9*, 212. [CrossRef]
5. Searle, R.; McBratney, A.; Grundy, M.; Kidd, D.; Malone, B.; Arrouays, D.; Stockman, U.; Zund, P.; Wilson, P.; Wilford, J.; et al. Digital Soil Mapping and Assessment for Australia and beyond: A Propitious Future. *Geoderma Reg.* **2021**, *24*, e00359. [CrossRef]
6. Mueller, L.; Schindler, U.; Mirschel, W.; Graham Shepherd, T.; Ball, B.C.; Helming, K.; Rogasik, J.; Eulenstein, F.; Wiggering, H. Assessing the Productivity Function of Soils. A Review. *Agron. Sustain. Dev.* **2010**, *30*, 601–614. [CrossRef]
7. Wallach, D.; Palosuo, T.; Thorburn, P.; Mielenz, H.; Buis, S.; Hochman, Z.; Gourdain, E.; Garcia, C.; Andrianasolo, F.; Dumont, B.; et al. Calibration of Crop Phenology Models: Going beyond Recommendations. *bioRxiv* **2022**. [CrossRef]
8. Boeing, F.; Rakovech, O.; Kumar, R.; Samaniego, L.; Schrön, M.; Hildebrandt, A.; Rebmann, C.; Thober, S.; Müller, S.; Zacharias, S.; et al. High-Resolution Drought Simulations and Comparison to Soil Moisture Observations in Germany. *Hydrol. Earth Syst. Sci.* **2022**, *26*, 5137–5161. [CrossRef]
9. Bönecke, E.; Breitsameter, L.; Brüggemann, N.; Chen, T.W.; Feike, T.; Kage, H.; Kersebaum, K.C.; Piepho, H.P.; Stützel, H. Decoupling of Impact Factors Reveals the Response of German Winter Wheat Yields to Climatic Changes. *Glob. Chang. Biol.* **2020**, *26*, 3601–3626. [CrossRef]
10. Webber, H.; Lischeid, G.; Sommer, M.; Finger, R.; Nendel, C.; Gaiser, T.; Ewert, F. No Perfect Storm for Crop Yield Failure in Germany. *Environ. Res. Lett.* **2020**, *15*, 104012. [CrossRef]
11. Drastig, K.; Prochnow, A.; Libra, J.; Koch, H.; Rolinski, S. Irrigation Water Demand of Selected Agricultural Crops in Germany between 1902 and 2010. *Sci. Total Environ.* **2016**, *569–570*, 1299–1314. [CrossRef] [PubMed]
12. Chen, S.; Arrouays, D.; Angers, D.A.; Chenu, C.; Barré, P.; Martin, M.P.; Saby, N.P.A.; Walter, C. National Estimation of Soil Organic Carbon Storage Potential for Arable Soils: A Data-Driven Approach Coupled with Carbon-Landscape Zones. *Sci. Total Environ.* **2019**, *666*, 355–367. [CrossRef] [PubMed]
13. Wiesmeier, M.; von Lützow, M.; Wollschlaeger, U.; Vogel, H.J.; Garcia-Franco, N.; Ließ, M.; Urbanski, L.; Hobley, E.; Lang, B.; Marin-Spiotta, E.; et al. Soil Organic Carbon Storage as a Key Function of Soils—A Review of Drivers and Indicators at Various Scales. *Geoderma* **2019**, *333*, 149–162. [CrossRef]
14. Wang, C.; Amon, B.; Schulz, K.; Mehdi, B. Factors That Influence Nitrous Oxide Emissions from Agricultural Soils as Well as Their Representation in Simulation Models: A Review. *Agronomy* **2021**, *11*, 770. [CrossRef]
15. Bouraoui, F.; Grizzetti, B. Modelling Mitigation Options to Reduce Diffuse Nitrogen Water Pollution from Agriculture. *Sci. Total Environ.* **2014**, *468–469*, 1267–1277. [CrossRef]
16. Sundermann, G.; Wägner, N.; Cullmann, A.; von Hirschhausen, C.R.; Kemfert, C. *Nitrate Pollution of Groundwater Long Exceeding Trigger Value: Fertilization Practices Require More Transparency and Oversight, DIW Weekly Report*; DIW Weekly; Deutsches Institut für Wirtschaftsforschung (DIW): Berlin, Germany, 2020.
17. BGR. *Soil Map of Germany 1:250,000*; Federal Institute for Geosciences and Natural Resources: Hanover, Germany, 2018.
18. Ad-hoc-AG Boden. *Bodenkundliche Kartieranleitung. KA5*, 5th ed.; Bundesanstalt für Geowissenschaften und Rohstoffe in Zusammenarbeit mit den Staatlichen Geologischen Diensten: Stuttgart, Germany, 2005; ISBN 978-3-510-95920-4.
19. Jenny, H. *Factors of Soil Formation. A System of Quantitative Pedology*; Dover Publications: New York, NY, USA, 1941.
20. McBratney, A.B.; Mendonca Santos, M.L.; Minasny, B. On Digital Soil Mapping. *Geoderma* **2003**, *117*, 352. [CrossRef]
21. Padarian, J.; Minasny, B.; McBratney, A.B. Machine Learning and Soil Sciences: A Review Aided by Machine Learning Tools. *Soil* **2020**, *6*, 35–52. [CrossRef]

22. Arrouays, D.; Mulder, V.L.; Richer-de-Forges, A.C. Soil Mapping, Digital Soil Mapping and Soil Monitoring over Large Areas and the Dimensions of Soil Security—A Review. *Soil Secur.* **2021**, *5*, 100018. [CrossRef]
23. Chen, S.; Arrouays, D.; Leatitia Mulder, V.; Poggio, L.; Minasny, B.; Roudier, P.; Libohova, Z.; Lagacherie, P.; Shi, Z.; Hannam, J.; et al. Digital Mapping of GlobalSoilMap Soil Properties at a Broad Scale: A Review. *Geoderma* **2022**, *409*, 115567. [CrossRef]
24. Daniel, Ž.; Minařík, R.; Skála, J.; Beitlerová, H.; Juřicová, A.; Rojas, J.R.; Penížek, V.; Zádorová, T. High-Resolution Agriculture Soil Property Maps from Digital Soil Mapping Methods, Czech Republic. *Catena* **2022**, *212*, 106024. [CrossRef]
25. Gebauer, A.; Sakhaee, A.; Don, A.; Poggio, M.; Ließ, M. Topsoil Texture Regionalization for Agricultural Soils in Germany—An Iterative Approach to Advance Model Interpretation. *Front. Soil Sci.* **2022**, *1*, 25. [CrossRef]
26. Malone, B.; Searle, R. Updating the Australian Digital Soil Texture Mapping (Part 2): Spatial Modelling of Merged Field and Lab Measurements. *Soil Res.* **2021**, *59*, 419–434. [CrossRef]
27. Reddy, N.N.; Chakraborty, P.; Roy, S.; Singh, K.; Minasny, B.; McBratney, A.B.; Biswas, A.; Das, B.S. Legacy Data-Based National-Scale Digital Mapping of Key Soil Properties in India. *Geoderma* **2021**, *381*, 114684. [CrossRef]
28. Padarian, J.; Minasny, B.; McBratney, A.B. Using Deep Learning for Digital Soil Mapping. *Soil* **2019**, *5*, 79–89. [CrossRef]
29. Ma, Y.; Minasny, B.; McBratney, A.; Poggio, L.; Fajardo, M. Predicting Soil Properties in 3D: Should Depth Be a Covariate? *Geoderma* **2021**, *383*, 114794. [CrossRef]
30. Poeplau, C.; Don, A.; Flessa, H.; Heidkamp, A.; Jacobs, A.; Prietz, R. *First German Agricultural Soil Inventory–Core Dataset*; Open Agrar Repositorium: Göttingen, Germany, 2020. [CrossRef]
31. Jacobs, A.; Flessa, H.; Don, A.; Heidkamp, A.; Prietz, R.; Gensior, A.; Poeplau, C.; Riggers, C.; Tiemeyer, B.; Vos, C.; et al. *Landwirtschaftlich Genutzte Böden in Deutschland–Ergebnisse Der Bodenzustandserhebung, Thünen Report 64*; Johann Heinrich von Thünen-Institut: Braunschweig, Germany, 2018; ISBN 9783865761927.
32. DWD. Seasonal Grids of Sum of Precipitation over Germany, Version v1.0. Available online: https://opendata.dwd.de/climate_environment/CDC/grids_germany/seasonal/precipitation/ (accessed on 23 October 2022).
33. DWD. Seasonal Grids of Monthly Averaged Daily Air Temperature (2m) over Germany, Version v1.0. Available online: https://opendata.dwd.de/climate_environment/CDC/grids_germany/seasonal/air_temperature_mean/ (accessed on 23 October 2022).
34. DWD. Seasonal Grids of Sum of Drought Index (de Martonne) over Germany, Version v1.0. Available online: https://opendata.dwd.de/climate_environment/CDC/grids_germany/seasonal/drought_index/ (accessed on 23 October 2022).
35. Swinnen, E.; Van Hoolst, R. Copernicus Global Land Operations "Vegetation and Energy". Issue I1.12, Version 1. Available online: https://land.copernicus.eu/global/sites/cgls.vito.be/files/products/CGLOPS1_ATBD_DMP300m-V1_I1.12.pdf (accessed on 23 October 2022).
36. Swinnen, E.; Dierckx, W.; Toté, C. Gio Global Land Component–Lot I "Operation of the Global Land Component". Quality Assessment Report Proba-V NDVI, VCI and VPI. Issue 1.21. Available online: https://land.copernicus.eu/global/sites/cgls.vito.be/files/products/GIOGL1_QAR_NDVI-VCI-VPI_I1.21.pdf (accessed on 23 October 2022).
37. BGR. *Geomorphographic Map of Germany, GMK1000*; Federal Institute for Geosciences and Natural Resources: Hanover, Germany, 2007.
38. European Environment Agency (EEA). *Copernicus Land Monitoring Service—EU-DEM, European Digital Elevation Model Version 1.1.*; EEA: Copenhagen, Denmark, 2017.
39. BGR; SDG. *Hydrogeological Map of Germany 1:250,000 (HÜK250)*; Federal Institute for Geosciences and Natural Resources (BGR): Hanover, Germany; German State Geological Surveys (SGD): Hanover, Germany, 2019.
40. BGR. *Groups of Soil Parent Material in Germany 1:5,000,000. BAG5000, Version 3.0*; Federal Institute for Geosciences and Natural Resources: Hanover, Germany, 2008.
41. BGR. *Soil Scapes in Germany 1:5,000,000. BGL5000*; Federal Institute for Geosciences and Natural Resources: Hanover, Germany, 2008.
42. INSPIRE Thematic Working Group. *INSPIRE–Infrastructure for Spatial Information in Europe. D2.8.I.2 Data Specification on Geographical Grid Systems–Technical Guidelines*; INSPIRE Thematic Working Group Coordinate Reference Systems & Geographical Grid Systems: Brussels, Belgium, 2014.
43. Conrad, O.; Bechtel, B.; Bock, M.; Dietrich, H.; Fischer, E.; Gerlitz, L.; Wehberg, J.; Wichmann, V.; Böhner, J. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geosci. Model Dev.* **2015**, *8*, 1991–2007. [CrossRef]
44. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1990; ISBN 9780471878766.
45. Ahmad, A.; Khan, S.S. Survey of State-of-the-Art Mixed Data Clustering Algorithms. *IEEE Access* **2019**, *7*, 31883–31902. [CrossRef]
46. Van Mechelen, I.; Boulesteix, A.-L.; Dangl, R.; Dean, N.; Guyon, I.; Hennig, C.; Leisch, F.; Steinley, D. Benchmarking in Cluster Analysis: A White Paper. *arXiv* **2018**, arXiv:1809.10496.
47. Rousseeuw, P.J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
48. Breiman, L. Random Forests. *J. Chem. Inf. Model.* **2001**, *53*, 1689–1699. [CrossRef]
49. Ishwaran, H.; Kogalur, U.B. Package 'RandomForestSRC'. Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). Version 3.1.1. 2022. Available online: https://www.randomforestsrc.org/ (accessed on 23 October 2022).
50. Hothorn, T.; Hornik, K.; Strobl, C.; Zeileis, A. Package 'Party'. A Laboratory for Recursive Partytioning. Version 1.3-11. 2022. Available online: http://party.r-forge.r-project.org/ (accessed on 23 October 2022).
51. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Leaming* **1995**, *20*, 273–297. [CrossRef]

52. Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–39. [CrossRef]
53. Meyer, D. Support Vector Machines—The Interface to Libsvm in Package E1071. *FH Tech. Wien* **2019**, *16*, 130.
54. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning–Data Mining, Inference, and Prediction*, 2nd ed.; Springer Science+Business Media, LLC: New York, NY, USA, 2009; ISBN 978-0-387-84857-0.
55. Affenzeller, M.; Winkler, S.; Wagner, S.; Beham, A. *Genetic Algorithms and Genetic Programming*; Taylor and Francis Group: Boca Raton, FL, USA, 2009; ISBN 978-1-58488-629-7.
56. Batjes, N. *A Taxotransfer Rule Based Approach for Filling Gaps in Measured Soil Data in Primary SOTER Databases (Version 1.1)*; World Soil Information: Wageningen, The Netherlands, 2003.
57. Hugelius, G.; Bockheim, J.G.; Camill, P.; Elberling, B.; Grosse, G.; Harden, J.W.; Johnson, K.; Jorgenson, T.; Koven, C.D.; Kuhry, P.; et al. A New Data Set for Estimating Organic Carbon Storage to 3 m Depth in Soils of the Northern Circumpolar Permafrost Region. *Earth Syst. Sci. Data* **2013**, *5*, 393–402. [CrossRef]
58. Almendra-Martín, L.; Martínez-Fernández, J.; Piles, M.; González-Zamora, Á. Comparison of Gap-Filling Techniques Applied to the CCI Soil Moisture Database in Southern Europe. *Remote Sens. Environ.* **2021**, *258*, 112377. [CrossRef]
59. Wang, Q.; Wang, L.; Zhu, X.; Ge, Y.; Tong, X.; Atkinson, P.M. Remote Sensing Image Gap Filling Based on Spatial-Spectral Random Forests. *Sci. Remote Sens.* **2022**, *5*, 100048. [CrossRef]
60. Taki, R.; Wagner-Riddle, C.; Parkin, G.; Gordon, R.; VanderZaag, A. Comparison of Two Gap-Filling Techniques for Nitrous Oxide Fluxes from Agricultural Soil. *Can. J. Soil Sci.* **2019**, *99*, 12–24. [CrossRef]
61. Kim, Y.; Johnson, M.S.; Knox, S.H.; Black, T.A.; Dalmagro, H.J.; Kang, M.; Kim, J.; Baldocchi, D. Gap-Filling Approaches for Eddy Covariance Methane Fluxes: A Comparison of Three Machine Learning Algorithm Algorithms and Algorithm a Traditional Method with Principal Component Analysis. *Glob. Chang. Biol.* **2020**, *26*, 1499–1518. [CrossRef]
62. Ghanbarian, B.; Pachepsky, Y. Machine Learning in Vadose Zone Hydrology: A Flashback. *Vadose Zo. J.* **2022**, *21*, e20212. [CrossRef]
63. Lamichhane, S.; Kumar, L.; Wilson, B. Digital Soil Mapping Algorithms and Covariates for Soil Organic Carbon Mapping and Their Implications: A Review. *Geoderma* **2019**, *352*, 395–413. [CrossRef]
64. Sakhaee, A.; Gebauer, A.; Ließ, M.; Don, A. Spatial Prediction of Organic Carbon in German Agricultural Topsoil Using Machine Learning Algorithms. *Soil* **2022**, *8*, 587–604. [CrossRef]
65. Ballabio, C.; Panagos, P.; Monatanarella, L. Mapping Topsoil Physical Properties at European Scale Using the LUCAS Database. *Geoderma* **2016**, *261*, 110–123. [CrossRef]
66. Aksoy, E.; Yigini, Y.; Montanarella, L. Combining Soil Databases for Topsoil Organic Carbon Mapping in Europe. *PLoS ONE* **2016**, *11*, 2022. [CrossRef] [PubMed]
67. Poggio, L.; De Sousa, L.M.; Batjes, N.H.; Heuvelink, G.B.M.; Kempen, B.; Ribeiro, E.; Rossiter, D. SoilGrids 2.0: Producing Soil Information for the Globe with Quantified Spatial Uncertainty. *Soil* **2021**, *7*, 217–240. [CrossRef]
68. Van Liedekerke, M.; Panagos, P. Predicted Distribution of SOC Content in Europe (Based on LUCAS, BioSoil and CZO) in the Context of the EU-Funded SoilTrEC Project. *PLoS ONE* **2016**, *11*, e0152098.
69. BGR. *General Geological Map of the Federal Republic of Germany 1:200,000*; Federal Institute for Geosciences and Natural Resources: Hanover, Germany, 2007.
70. Probst, P.; Wright, M.N.; Boulesteix, A.L. Hyperparameters and Tuning Strategies for Random Forest. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1301. [CrossRef]
71. Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics* **2007**, *8*, 25. [CrossRef]
72. BGR. *Soil Map of Germany 1:1,000,000. BÜK1000*; Federal Institute for Geosciences and Natural Resources: Hanover, Germany, 2013.

agriculture

MDPI

Article

Associations of Automatically Recorded Body Condition Scores with Measures of Production, Health, and Reproduction

Ramūnas Antanaitis [1,*], Dovilė Malašauskienė [1], Mindaugas Televičius [1], Mingaudas Urbutis [1], Arūnas Rutkauskas [1], Greta Šertvytytė [1], Lina Anskienė [2] and Walter Baumgartner [3]

[1] Large Animal Clinic, Veterinary Academy, Lithuanian University of Health Sciences, Tilžės Str. 18, LT-47181 Kaunas, Lithuania
[2] Department of Animal Breeding, Veterinary Academy, Lithuanian University of Health Sciences, Tilžės Str. 18, LT-47181 Kaunas, Lithuania
[3] University Clinic for Ruminants, University of Veterinary Medicine, Veterinaerplatz 1, A-1210 Vienna, Austria
* Correspondence: ramunas.antanaitis@lsmuni.lt; Tel.: +37-067-349-064

Abstract: In the present study, we hypothesize that an automated body condition scoring system could be an indicator of health and pregnancy success in cows. Therefore, the objective of this study is to determine the relationship of the automated registered body condition score (BCS) with pregnancy and inline biomarkers such as milk beta-hydroxybutyrate (BHB), milk lactate dehydrogenase (LDH), milk progesterone (mP4), and milk yield (MY) in dairy cows. Indicators from Herd NavigatorTM were grouped into classes based on their arithmetic means. Values were divided into various classes: MY: ≤31 kg/day (first class—67.3% of cows) and >31 kg/day (second class—32.7%); BHB in milk: ≤0.06 mmol/L (first class—80.7% of cows) and >0.06 mmol/L (second class—16.9%); milk LDH activity: ≤27 µmol/min (first class—69.5% of cows) and >27 µmol/min (second class—30.5%); milk progesterone value: ≤15.5 ng/mL (first class—28.8% of cows) and >15.5 ng/mL (second class—71.2%); and BCS: 2.5–3.0 (first class—21.4% of cows), >3.0–3.5 (second class—50.8%), and >3.5–4.0 (third class—27.8%). According to parity, the cows were divided into two groups: 1 lactation (first group—38.9%) and ≥2 lactations (second group—61.1%). Based on our investigated parameters, BCS is associated with pregnancy success because the BCS (+0.29 score) and mP4 (10.93 ng/mL) of the pregnant cows were higher compared to the group of non-pregnant cows. The MY (−5.26 kg, $p < 0.001$) and LDH (3.45 µmol/min) values were lower compared to those in the group of non-pregnant cows ($p < 0.01$). Statistically significant associations of BCS and mP4 with the number of inseminations were detected. The number of inseminations among cows with the highest BCS of >3.5–4.0 was 42.41% higher than that among cows with the lowest BCS of 2.5–3.0 ($p < 0.001$). BCS can also be a health indicator. We found that the LDH content was greatest among cows with the highest BCS of >3.5–4.0; this value was 6.48% higher than that in cows with a BCS of >3.0–3.5 ($p < 0.01$). The highest MY was detected in cows with the lowest BCS of 2.5–3.0, which was 29.55% higher than that in cows with the highest BCS of >3.5–4.0 ($p < 0.001$). BCS was the highest in the group of cows with mastitis (4.96% higher compared to the group of healthy cows), while the highest statistically significant mean differences in body condition score (9.04%) were estimated between the mastitis and metritis groups of cows ($p < 0.001$).

Keywords: precision dairy farming; sensors technology; dairy cows

Citation: Antanaitis, R.; Malašauskienė, D.; Televičius, M.; Urbutis, M.; Rutkauskas, A.; Šertvytytė, G.; Anskienė, L.; Baumgartner, W. Associations of Automatically Recorded Body Condition Scores with Measures of Production, Health, and Reproduction. *Agriculture* **2022**, *12*, 1834. https://doi.org/10.3390/agriculture12111834

Academic Editors: Gniewko Niedbała, Sebastian Kujawa and Milan Shipka

Received: 22 August 2022
Accepted: 31 October 2022
Published: 2 November 2022

1. Introduction

The popularity of autonomous analysis systems on farms has grown as dairy herd sizes have increased. The Herd Navigator (HN, DeLaval International) management tool, which examines many milk constituents automatically during milking, provides daily estimates of milk beta-hydroxybutyrate (BHB) [1]. According to Yu and Maeda [2], the HN system functions autonomously and offers real-time physiological information on lactating

cows while also assisting with farm management decisions. This tool notifies and advises dairy farmers about the condition and health of each cow. To maintain herd health, the system includes unique biological models that take into account measured metrics, cow information, and other risk factors. This process avoids costly treatments and substantial output losses. Significant improvements in reproduction, mastitis, and ketosis have been observed on farms that use this approach [2].

Body condition score (BCS) is a subjective method for calculating the amount of metabolizable energy stored in fat and muscle in a live animal (body reserves) [3]. The body condition score (BCS) is, therefore, a measure for estimating cow body fat reserves [4]. BCS is a helpful method for monitoring the relationships between nutritional management, reproduction, and ketosis and supports farm management decisions [5]. There are some connections between energy status (ES) markers and reproductive features in the first several weeks of lactation. Because the negative effect of ES on fertility is well known, the proposed publicly available milk-based ES indicators are promising management strategies that can assist farmers in identifying cows that may be sensitive to metabolic stress and production disorders as well as determining the appropriateness of feeding techniques and the timing of insemination [1]. The dairy sector places a high value on reproductive performance management, which includes both fertility and calving.

The main objective of commercialized dairy farming is not only to produce one calf per year but also to maintain profitability by changing inputs and output. These inputs need enormous investment in infrastructure, labor, feed, and treatments, among other factors. The most important of these inputs is nutritional control, which is essential for a normal production and reproduction cycle. The connection between nutrition and reproduction was previously established [3]. Loeffler et al. [6] demonstrated that negative energy balance, physical condition decline, and sickness have an impact on fertility, and poor fertility is one of the major limiting factors impacting the dairy industry's economy. Most studies have found that BCS is a useful tool for dairy cow management in predicting energy balance and disease risk variables. Overweight cows were found to be at higher risk of developing metabolic illnesses and had a reduced chance of becoming pregnant during their first breeding. Low BCS at calving causes oestrus to be delayed, follicular development to be delayed, fertility to be diminished, and the inter-calving interval to be extended. The link between nutrition and fertility management is also well understood [3]. It was established that BCS at calving and its variations throughout lactation have an impact on the health and fertility of high-producing dairy cows [7]. A fully automated inline LDH, progesterone (mP4), and BHB analyzer that can be paired with a milking robot is available for purchase on the commercial market to achieve better herd assessment. Using the precision diagnostic methods of this tool can provide greater information on the current parameters influencing dairy cow reproductive physiology. The first commercially accessible 3D BCS system based on image processing technologies was built by the DeLaval corporation [8].

According to Nyman et al. [9], additional research is needed to determine whether the diagnostic qualities of LDH improve with adjustment based on their interactions with other cow variables when used as a diagnostic tool for identifying cows with mastitis. We also discovered a positive relationship between LDH content in milk and milk yield [10]. An automated body condition scoring system would produce more objective and consistent information than observational scoring and result in a more efficient operation that is less stressful for the animal [11]. In terms of research, automated methods would enable the utilization of data collected at various precise time periods, resulting in scores that are unaffected by inter- and intra-evaluator variation [12].

In the present study, we hypothesize that an automatically recorded body condition score has associations with measures of production, health, and reproduction. Therefore, the objective of this study is to determine the associations of automatically recorded body condition scores with measures of production, health, and reproduction (MY, BHB, LDH, and mP4) in dairy cows.

2. Materials and Methods

2.1. Study Design

This research was conducted on a dairy farm in northern Lithuania, in the eastern region of Europe, located at 55.10571, 24.24399. The cows were kept in a loose housing system and fed a total mixed ration (TMR) that was balanced according to their physiological needs. Cows were fed a TMR composed of 30% corn silage, 10% grass silage, 4% grass hay, 50% grain concentrate mash, and 6% mineral mixture. Diets were created to fit or surpass the needs of a 550 kg Holstein cow producing 35 kg milk/d. Every day at 06:00 a.m. and 06:00 p.m., the animals were fed. The cows were milked twice a day, at 05:00 a.m. and 05:00 p.m., using a parlor system. The cows weighed an average of 550 ± 45 kg. In 2021, the average energy-corrected milk yield (4.2% fat, 3.5% protein) per cow and year was 10,500 kg. During the study, contact with animals was kept to a minimum, thus avoiding the impact of the trial on animal welfare. The cows (n = 597) had an average of 2.10 ± 0.05 lactations at 206.52 ± 5.02 days postpartum and were divided into two groups: primiparous cows (n = 232) and multiparous cows (n = 365).

2.2. Measurements

The real-time analyzer Herd Navigator™ (Lattec I/S, Hillerd, Denmark) was applied in conjunction with a DeLaval milking parlor system to collect data on mP4, MY, BHB, and LDH (DeLaval Inc., Tumba, Sweden). During the robot milking operation, an inline sampler automatically took a representative sample of several millilitres of milk from each cow. The material was then loaded into the Herd Navigator™ analyzer for further examination. Three-dimensional BCS cameras were used to measure BCS (DeLaval body condition scoring BCS, DeLaval International AB, Tumba, Sweden). These systems were used to collect daily averages of data on the following biomarkers for each cow from the day of oestrus to 7 days post-oestrus: mP4, MY, BHB, LDH, and BCS.

2.2.1. Measurements of BCS

After each milking, Herd Navigator™Automatic BCS measurements were taken using a commercially available 3D body condition scoring camera system (DeLaval Body Condition Scoring, BCS DeLaval International AB, Tumba, Sweden) with two cameras, including a camera fixed above one of the milking parlor exit races. Cows were identified individually using a radio-frequency identification collar system, allowing for repeated BCS assessments per day. As a result, each cow typically had two visual BCS assessments taken on the same day each week. The camera system reported BCS readings in 0.1 point increments and provided a 1–5 scale [13]. Data from the camera system are given either as a one-day BCS rolling average that removes the lowest and highest 20% of data prior to averaging or as daily (AM and PM) BCS values. Individual daily AM and PM raw BCS data from each camera were accessed via the manufacturer's software (DelPro Farm Manager, DeLaval International AB, Tumba, Sweden) using the pathway Systems > Devices > BCS Camera > BCS CAM as the data were not readily available for download and were downloaded weekly before being automatically overwritten by the system after eight days.

2.2.2. Measurements of BHB, LDH, MY, and mP4

To identify milk BHB concentrations and LDH activities, a real-time analyzer Herd Navigator™ (Lattec I/S, Hillerd, Denmark) was paired with a DeLaval milking robot (DeLaval Inc., Tumba, Sweden). Several millilitres of milk were obtained from each cow during the milking process using an inline sampler to determine the concentrations of the previously listed parameters. The raw data were adjusted using company-specified procedures to account for differences in dry-stick sets and variations in the surrounding humidity. The most extreme outliers were then excluded from the calculations. Data more than 200 mol/min per liter were set to a maximum value of 200, and any negative values were removed from the equation because they did not fall within the typical range of measurements recorded by the Herd Navigator system. This is how data in the Herd

Navigator system were standardized. An optical milk meter was used to measure the milk yield from each cow. The LDH concentration (mol/min) was estimated by dividing the LDH activity by the milk output from the most recent milking activity. The Herd Navigator system was set to automatically collect milk samples and test the mP4 in individual cows using dry-stick biosensor technology and enzyme immunoassays based on a bio-model that establishes the frequency and quantification of mP4 samples. The system adjusted the frequency of assays to an average of six to seven progesterone studies per cycle, depending on the postpartum period and the stage of the oestrus cycle. The data were then provided to a user interface via the analyzer. When the progesterone level in the oestrus cycle of a cow fell below 5 ng/mL, a heat alert was displayed. The mP4 concentration in milk samples began on the first postpartum day and was obtained every 5 days until pregnancy was recognized. Raw (actual) mP4 concentrations were adjusted to smoothed values based on a standardized procedure to correct for outliers predicted in the serial sampling system, as described by Friggens and Chagunda [14], in order to reduce random fluctuation and differences in batches of sticks and reagents.

2.3. Identification of Oestrus and Pregnancy

The cows' oestrus cycles were all synchronized using the OvSynch protocol. When an animal displayed a progesterone alarm (registered by the Herd Navigator system), an increase in cow walking activity (registered by the AMS), or one or more of the oestrus signs described by Van Eerdenburg et al. [15] (standing to be mounted, mucous vaginal discharge, cajoling, restlessness at being mounted but not standing, sniffing the vagina of other cows, resting chin on other cows), rectal palpation was used to assess the cow's uterine tone. The cows were artificially inseminated 12 hours after the start of oestrus (as assessed by the mP4 concentration determined via AMS). At 30–35 days after insemination, the pregnancies were tested with an 'easy scan' ultrasound (IMV imaging, Scotland).

2.4. Identification of Health Status

Out of 850 fresh milking cows (from 1 until 30 days after calving), we randomly chose 483 clinically healthy cows, 21 cows with subclinical ketosis, 26 cows with subclinical mastitis, and 67 cows with metritis.

Healthy group (n = 483). Cows that had no clinical symptoms of disease after calving and BHB values at or below 1.2 mmol/L for the entire 30-day post-calving period were categorized into this group. This group of cows had an average milk F/P of 1.2.

Subclinical ketosis group (SCG) (n = 21). When at least one beta-hydroxybutyrate (BHB) value throughout the 30-day postpartum period was 1.2 mmol/L, the cows were identified as having SCK. For this particular herd of cows, the milk fat/protein ratio (F/P) was recorded as being >1.2. After calving, the cows showed no clinical symptoms of any additional illnesses, including metritis, lameness, mastitis, displaced abomasus, dyspepsia with an average rectal temperature of +38.8 °C, or rumen motility of five to six times every three minutes.

Subclinical mastitis group (n = 26). SCC was used to identify cases that belonged to the subclinical mastitis group (CM). SCM was identified in cows with an SCC of more than 200,000 cells/mL [16]. SCC was assessed once daily during all studies. A general clinical evaluation revealed that none of the cows showed clinical indications indicative of any disease.

Metritis group (n = 67). Every 3 days after calving until day +21, vaginal discharge (VD) was assessed for each cow. A gloved hand was inserted into the vaginal canal up to the cervix to remove any discharge present and allow for visual inspection. Based on the scoring system used by Urton et al. [17], the appearance and smell of the VD were assessed and categorized as follows: putrid (red/brown color, watery, foul-smelling), no mucus or clear mucus = 0, cloudy mucus or mucus with flecks of pus = 1, mucopurulent (50% pus present) and foul-smelling = 2, and mucopurulent (50% pus present) and foul-smelling = 3. All cows had three points.

2.5. Data Analysis and Statistics

Indicators from Herd Navigator™ were grouped into classes based on their arithmetic means. Values were divided into various classes: MY: ≤31 kg/day (first class—67.3% of cows) and >31 kg/day (second class—32.7%); BHB in milk: ≤0.06 mmol/L (first class—80.7% of cows) and >0.06 mmol/L (second class—16.9%); milk LDH activity: ≤27 µmol/min (first class—69.5% of cows) and >27 µmol/min (second class—30.5%); milk progesterone value: ≤15.5 ng/mL (first class—28.8% of cows) and >15.5 ng/mL (second class—71.2%); and BCS: 2.5–3.0 (first class—21.4% of cows), >3.0–3.5 (second class—50.8%), and >3.5–4.0 (third class—27.8%). According to lactation, the cows were divided into two groups: 1 lactation (first group—38.9% of cows) and ≥2 lactations (second group—61.1% of cows). According to their pregnancy status, 1-lactation cows were divided into two groups, non-pregnant (n = 107 or 46.0%) and pregnant cows (n = 125 or 54.0%); similarly, cows with ≥2 lactations were classified as non-pregnant (n = 207 or 57.0%) or pregnant (n = 158 or 43.0%). The average days in milk (DIM) among pregnant cows was 151.60 ± 0.11 days, while that among non-pregnant cows was 151.70 ± 0.08 days. The status of pregnancy among all investigated cows was as follows: pregnant (n = 283) and non-pregnant (n = 314). The statistical analysis of data was performed using the SPSS 25.0 software package. Normal distributions were assessed using the Kolmogorov–Smirnov test. The results from Herd Navigator™ are presented as the mean ± standard error (M ± SE) with a 95% confidence interval (CI). The Pearson correlation (r) was determined to define the linear relationship between BCS and indicators from AMS. Multiple comparisons of group means were calculated using Tukey's test. A probability below 0.05 was considered statistically significant. The chi-square ($\chi 2$) statistic was used to test the relationship between the categorical variable classes of BCS and the indicators investigated using Herd Navigator™, BCS, and the reproductive status of cows.

3. Results

3.1. Associations of Automatically Recorded Body Condition Scores with Measures of Production and Reproduction

For all biomarkers, all differences between groups were significant except for milk β-hydroxybutyrate. The BCS (+0.09 score) and mP4 values of pregnant cows were higher (10.93 ng/mL) compared to those of non-pregnant cows. The MY (−5.26 kg; $p < 0.001$) and LDH values were lower (3.45 µmol/min) compared to those of non-pregnant cows ($p < 0.01$). The data are presented in Table 1. Backward stepwise multivariate logistic regression showed that, of all tested categorical variables (BCS, mP4, LDH, BHB, and MY), only mP4 (OR = 1.197, $p < 0.001$) and MY (OR = 0.886, $p < 0.001$) had a significant relationship with the reproductive status of cows.

The pregnancy status of cows was associated with the BCS assessment ($p < 0.05$). In the class of cows with BCS = 2.5–3.0, 37.5% of cows were pregnant, whereas with BCS > 3.0–3.5 and BCS > 3.5–4.0, 47.9% and 54.2% of cows were pregnant, respectively (Figure 1).

Of all the biomarkers, differences between BCS classes were significant only in LDH and MY. The LDH of cows with the highest BCS of 3.5–4.0 was 6.48% higher than that of the cows with a BCS of 3.0–3.5 ($p < 0.01$). The highest MY was detected in cows with the lowest BCS of 2.5–3.0; it was 29.55% higher compared to that of the cows with the highest BCS of 3.5–4.0 ($p < 0.001$) (Table 2).

Table 1. Means and standard errors of the mean of biomarkers based on the pregnancy status of cows from the day of oestrus to 7 days post-oestrus.

Indicator/Biomarker	Status of Pregnancy	M	SE	95% CI	
				Lower Bound	Upper Bound
BCS, score	Non-pregnant [a]	3.20 ***,[b]	0.019	3.16	3.24
	Pregnant [b]	3.29 ***,[a]	0.019	3.25	3.33
mP4, ng/mL	Non-pregnant	12.89 ***,[b]	0.600	11.71	14.07
	Pregnant	23.82 ***,[a]	0.255	23.31	24.32
LDH, µmol/min	Non-pregnant	25.01 **,[b]	0.962	23.11	26.90
	Pregnant	21.56 **,[a]	0.715	20.15	22.97
BHB, mmol/L	Non-pregnant	0.06	0.001	0.057	0.063
	Pregnant	0.06	0.001	0.056	0.059
MY, kg/day	Non-pregnant	30.54 ***,[b]	0.503	29.55	31.52
	Pregnant	25.28 ***,[a]	0.357	24.58	25.99

Different letters (a and b) indicate statistically significant differences between classes (*** $p < 0.001$, ** $p < 0.01$). BCS—body condition score; mP4—milk progesterone; LDH—milk lactate dehydrogenase; BHB—milk β-hydroxybutyrate; MY—milk yield. M—mean; SEM—standard error of the mean.

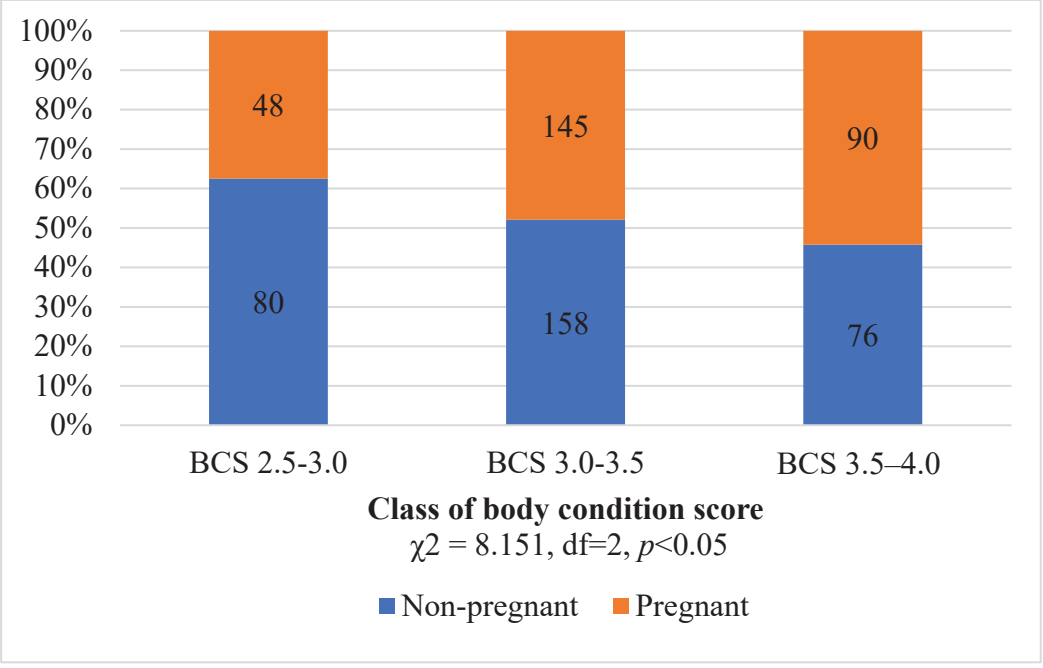


Figure 1. Relation of body condition score with the pregnancy status of cows from the day of oestrus to 7 days post-oestrus.

Table 2. Means and standard errors of the mean of biomarkers based on the body condition score from the day of oestrus to 7 days post-oestrus.

Indicator/Biomarker	Class of BCS	M	SE	95% CI	
				Lower Bound	Upper Bound
mP4, ng/mL	2.5–3.0	16.70	0.865	14.99	18.41
	3.0–3.5	18.05	0.580	16.91	19.19
	3.5–4.0	19.15	0.740	17.69	20.61
LDH, µmol/min	2.5–3.0	27.16	1.356	24.48	29.84
	3.0–3.5	21.81 **,c	0.779	20.27	23.34
	3.5–4.0	23.32 **,b	1.289	20.77	25.86
BHB, mmol/L	2.5–3.0	0.06	0.001	0.06	0.06
	3.0–3.5	0.06	0.002	0.06	0.06
	3.5–4.0	0.06	0.001	0.06	0.06
MY, kg/day	2.5–3.0	32.18 ***,b,c	0.722	30.76	33.61
	3.0–3.5	29.24 ***,a,c	0.419	28.42	30.07
	3.5–4.0	22.67 ***,a,b	0.511	21.66	23.68

Different letters (a, b and c) indicate statistically significant differences between classes ** $p < 0.01$, *** $p < 0.001$. BCS—body condition score; mP4—milk progesterone; LDH—milk lactate dehydrogenase; BHB—milk β-hydroxybutyrate; MY—milk yield. M—mean; SE—standard error of the mean; 95% CI—the 95% confidence interval.

The analysis showed that 51.16% more cows with a BCS of 2.5–3.0 were found in the second mP4 class compared to class 1, 58.41% more cows with a BCS > 3.0–3.5 were found in the second mP4 class, and 67.20% more cows with a BCS of 3 were found in the second mP4 class compared to the first mP4 class (Figure 2). Analysis of cows in the first mP4 class showed that 52.81%–53.93% more cows had a BCS of >3.0–3.5 compared to cows with a BCS of 2.5–3.0 and those with a BCS of 3 ($\chi2 = 26.244$, df = 2, $p < 0.001$). In the second mP4 class, the analysis showed almost the same tendency of cow distribution, with 53.93% more cows having a BCS of 2 and 41.59% more cows having a BCS of >3.5–4.0 ($\chi2 = 60.767$, df = 2, $p < 0.001$) (Figure 2).

The analysis showed that 79.25% more cows with a BCS of 2.5–3.0 were found in the first BHB class compared to the second BHB class; 78.31% more cows with a BCS of 3.0–3.5 and 82.27% more cows with a BCS of >3.5–4.0 were found in the first BHB class compared to the second BHB class. In the first BHB class, there were 56.91% and 47.15% more cows with a BCS of >3.0–3.5 than cows with a BCS of 2.5–3.0 and a BCS of <3.5–4.0, respectively. In the first BHB class, more cows were determined to have a BCS of >3.0–3.5 compared to cows with a BCS of 2.5–3.0 and a BCS of >3.5–4.0 ($\chi2 = 67.214$, df = 2, $p < 0.001$). In the second BHB class, 59.26% more cows had a BCS of >3.0–3.5, and 53.70% more cows had a BCS of >3.5–4.0 ($\chi2 = 18.554$, df = 2, $p < 0.001$) (Figure 2). The distribution of cows according to LDH classes revealed that 33.77% more cows with a BCS of 2.5–3.0 were present in the first LDH class than in the second LDH class. Additionally, 61.36% more cows with a BCS of >3.0–3.5 and 58.47% more cows with a BCS of >3.5–4.0 were found in the first BHB class compared to the second BHB class. In the first LDH class, there were 65.00% and 46.36% more cows with a BCS of >3.0–3.5 compared to cows with a BCS of 2.5–3.0 and a BCS of >3.5–4.0 ($\chi2 = 78.395$, df = 2, $p < 0.001$). In the second LDH class, 40.00% and 42.35% more cows, respectively, had a BCS of >3.0–3.5 and >3.5–4.0 compared to cows with a BCS of 2.5–3.0 and >3.5–4.0 ($\chi2 = 13.276$, df = 2, $p < 0.001$) (Figure 2). The analysis showed that 33.77% more cows with a BCS of 2.5–3.0 were found in the first MY class compared to the second MY class; 48.50% more cows with a BCS of >3.0–3.5 and 90.07% more cows with a BCS of >3.5–4.0 were found in the first MY class compared to the second MY class. In the first MY class, 74.50%–24.50% more cows were determined to have a BCS of >3.0–3.5 compared to cows with a BCS of 2.5–3.0 and >3.5–4.0 ($\chi2 = 86.075$, df = 2, $p < 0.001$). In the second BHB class, 25.24% more cows had a BCS of >3.0–3.5, and 85.44% more cows had a BCS of >3.5–4.0 ($\chi2 = 62.892$, df = 2, $p < 0.001$) (Figure 2).

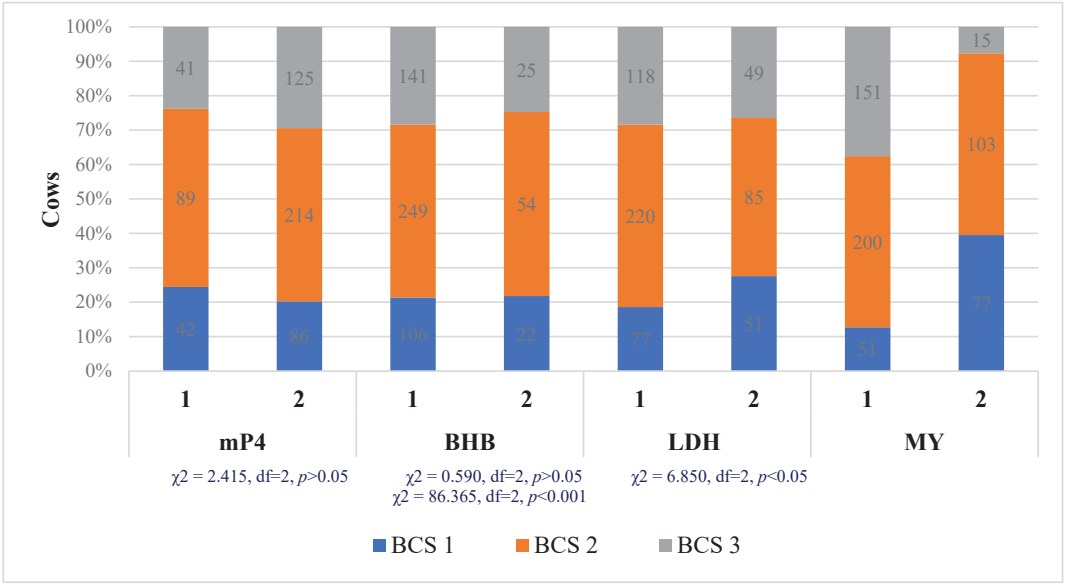


Figure 2. Relationship between the body condition scores of cows with other biomarkers from the Herd Navigator™ system. mP4 1—milk progesterone ≤ 15.5 ng/mL; mP4 2—milk progesterone > 15.5 ng/mL; BHB 1—β-hydroxybutyrate in milk ≤ 0.06 mmol/L; BHB 2—β hydroxybutyrate in milk > 0.06 mmol/L; LDH 1—milk lactate dehydrogenase ≤ 27 µmol/min; LDH 2—milk lactate dehydrogenase > 27 µmol/min; MY 1—milk yield ≤ 31 kg/day; MY 2—milk yield > 31 kg/day; BCS1 = 2.5–3.0; BCS2 ≥ 3.0–3.5; BCS3 ≥ 3.5–4.0.

Differences between classes of biomarkers with the number of inseminations were statistically significant only in BCS and mP4 (Table 3). The number of inseminations among cows with the highest BCS of >3.5–4.0 was 42.41% higher than that among cows with the lowest BCS of 2.5–3.0 ($p < 0.001$). The number of inseminations among cows with lower milk progesterone of ≤15.5 ng/mL (mP4 1) was 13.23% higher than that among cows with a higher concentration of milk progesterone of >15.5 ng/mL ($p < 0.01$). No statistically significant associations were found between BHB and LDH classes and the number of inseminations ($p > 0.05$).

Table 3. Means and standard errors of the mean of biomarkers registered from the day of oestrus to 7 days post-oestrus based on the number of inseminations.

Indicator/Biomarker	Class of Biomarker	M	SE	95% CI	
				Lower Bound	Upper Bound
BCS	2.5–3.0	2.20 ***,c	0.129	1.94	2.46
	3.0–3.5	2.61 ***,c	0.106	2.40	2.82
	3.5–4.0	3.82 ***,a,b	0.195	3.43	4.20
mP4	≤15.5	3.10 **,b	0.221	2.66	3.54
	>15.5	2.69 **,a	0.086	2.52	2.86
BHB	≤0.06	2.76	0.090	2.59	2.94
	>0.06	2.91	0.238	2.43	3.38
LDH	≤27	2.69	0.094	2.50	2.87
	>27	3.04	0.180	2.69	3.40

Different letters (a, b and c) indicate statistically significant differences between classes ** $p < 0.01$, *** $p < 0.001$). BCS—body condition score; mP4—milk progesterone; LDH—milk lactate dehydrogenase; BHB—milk β-hydroxybutyrate.

The BCS was statistically significantly negatively related to the milk yield, lactation ($p < 0.001$), and milk lactate dehydrogenase ($p < 0.05$). It was positively related to the number of inseminations ($p < 0.001$) and milk progesterone concentration ($p < 0.05$).

3.2. Associations of Automatically Recorded Body Condition Scores with Measures of Health

The BCS was highest in the group of cows with mastitis—higher by 4.96% compared to the BCS among the group of healthy cows. The highest statistically significant mean differences in the body condition score (9.04%) were estimated between the mastitis and metritis groups of cows ($p < 0.001$) (Table 4).

Table 4. Means and standard errors of the mean of the body condition score based on the health status of cows.

Biomarker	Disease	Mean	SEM	95% CI	
				Lower Bound	Upper Bound
BCS, score	Healthy [a] $n = 483$	3.26 *,d	0.015	3.23	3.28
	Subclinical ketosis [b] $n = 21$	3.13 *,c	0.075	2.97	3.29
	Subclinical mastitis [c] $n = 26$	3.43 ***,d,*,b	0.070	3.28	3.57
	Metritis [d] $n = 67$	3.12 ***,c,*,a	0.042	3.04	3.21

Different letters (a, b, c and d) indicate statistically significant differences between classes (* $p < 0.05$, *** $p < 0.001$).

The body condition scores in healthy and all diseased groups of cows were statistically significantly negatively related to the milk yield of cows ($p < 0.001$–0.05). The body condition score presented a dependence with the number of inseminations in the opposite direction between groups of cows, showing a positive relationship among healthy cows ($p < 0.001$) and those with mastitis ($p < 0.05$), along with a negative relationship in the metritis group of cows ($p < 0.05$). The body condition score had a positive relationship in healthy cows and a negative relationship in the mastitis group of cows ($p < 0.05$) (Figure 3).

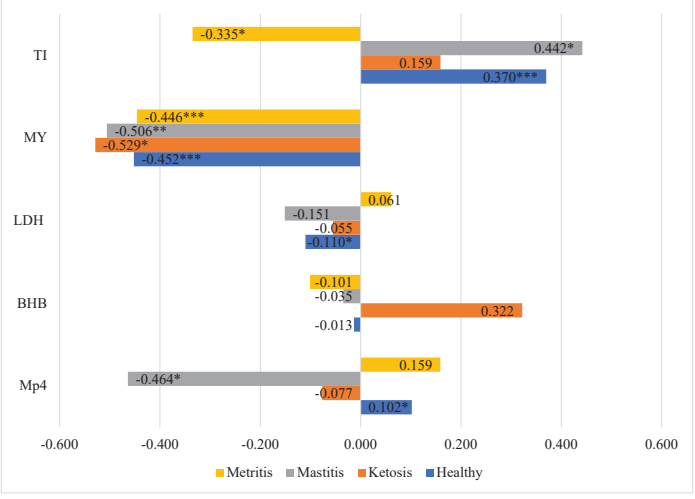


Figure 3. Body condition score correlations with investigated indicators of dairy cows that were eventually diagnosed as healthy or diseased. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. mP4—milk progesterone; BHB—milk β-hydroxybutyrate; LDH—milk lactate dehydrogenase; MY—milk yield; TI—number of inseminations.

4. Discussion

According to our results, the BCS registered from the day of oestrus to 7 days post-oestrus was higher among the pregnant cows (+0.29 score) compared to the BCS in the group of non-pregnant cows. Additionally, the mP4 in pregnant cows was higher (10.93 ng/mL), the MY was lower (−5.26 kg, $p < 0.001$), and the LDH was lower (3.45 µmol/min) compared to the values in the group of non-pregnant cows ($p < 0.01$). Statistically significant associations were detected between BCS, mP4, and the number of inseminations. The number of inseminations of cows with the highest BCS of >3.5–4.0 was 42.41% higher compared to that of the cows with the lowest BCS of 2.5–3.0 ($p < 0.001$). According to Roche et al. [18], the majority of studies on the physiological effects of energy status and energy balance on fertility revealed a positive link between earlier pregnancy attainment, enhanced BCS, and reduced BCS loss during early lactation. However, studies using daily BCS testing in large numbers of cows throughout early lactation remain uncommon. Daily automated BCS evaluations and the resulting enhanced prediction of pregnancy likelihood at AI may influence insemination decisions, such as which type of sperm to utilize [19] and when to discontinue inseminating cows that failed to conceive earlier during lactation [20]. The causes for the lower conception rates among cows with BCS are unknown. According to Britt [21], energy status during the early postpartum period may affect follicular/oocyte quality, resulting in decreased fertility in nursing dairy cows. Reduced functional competence of the ovulated follicles could be due to the development of follicles under negative energy balance or caused by subtle changes in the steroid hormone secretions that regulate gene expression and protein secretion by the endometrium, thereby affecting implantation and pregnancy recognition [22].

Poor postpartum health has a detrimental impact on dairy cow performance, and incidences of uterine, metabolic, and other health issues have been extensively recognized as risk factors for lower subsequent fertility [23]. A loss of body condition, as an indirect measure of energy balance in early lactation, and health, as a measure of metabolic, immunological, and homeostatic functioning, has been shown to produce delayed resumption of ovarian cyclicity postpartum [24–26]. Specific factors affecting the proper resumption of cyclicity include parity [27], changes in recoupling the growth hormone/IGF-1 axis in the liver [28], metabolic and infectious disorders [12,28], insufficient progesterone concentrations [29], and dystocia [30]. These implications may explain the effects of changes in body condition and disease on variables such as P/AI1, supporting the idea of using BCS and health information to predict fertility [12]. The amplitude of the link between milk yield and reproductive success is minor and depends on the herd production level [31]. Buckley et al. [31] discovered that high milk output at first service was associated with an increased risk of being pregnant after 42 days of the breeding season. However, the majority of research has found an antagonistic link between milk production and a variety of reproductive features [22,32]. These findings are consistent with the findings of our study, which showed that pregnant cows had a 5.26 kg/day lower milk output than non-pregnant cows. In previous studies, we found that pregnant cows had a 0.49-point higher body condition score, a 4.36 kg/day lower milk output, and a 6.11 ng/mL higher mP4 concentration than non-pregnant cows. Pregnant cows had a 0.49-point higher body condition score than non-pregnant cows, and cows with a BCS of >3.2 were 22 times more likely to be successful in reproduction than cows with a BCS of 3.2 [33].

We found that the LDH of cows with the highest BCS of >3.5–4.0 was 6.48% higher compared to that of cows with a BCS of >3.0–3.5 ($p < 0.01$). The highest MY was detected in cows with the lowest BCS of 2.5–3.0, which was 29.55% higher than that of the cows with the highest BCS of 3.5–4.0 ($p < 0.001$). The BCS was the highest in the group of cows with mastitis—the score was 4.96% higher compared to that in the group of healthy cows—while the highest statistically significant mean differences in body condition score (9.04%) were estimated between the mastitis and metritis groups of cows ($p < 0.001$). According to various studies, combining multiple sensor data is effective for detecting and differentiating mastitis types in AMS [34,35]. Lactate dehydrogenase (LDH) in dairy milk is

correlated with somatic cell count (SCC) and utilized as a mastitis indication in commercial herd management [9,14]. Cell-damaging mechanisms during mammary inflammation, according to Zank and Schlatterer [36], should be best recognized by monitoring high LDH activity. Yang et al. [27] investigated the variations in milk malondialdehyde levels and enzymatic activity caused by subclinical mastitis in dairy cows. The median value of LDH activity in subclinical mastitis milk was found to be substantially higher than that in normal milk. The authors concluded that measuring this characteristic in milk is an appropriate approach for diagnosing SCM in dairy cows [37]. According to Suriyasathaporn [38], poor body condition reflects a negative energy balance and makes the animal more susceptible to mastitis. Patel et al. [39] found that cows in both groups (under and over the ideal body condition score) were at higher risk of developing subclinical mastitis. In comparison to cows in the high-infection herds, cows in the low-infection herds had considerably lower BCS results throughout the last month prior to calving and the first month of lactation. There were, overall, significant correlations between BCS and the incidence of mastitis infection [40]. A higher incidence of subclinical ketosis in animals with better conditions at calving may be one reason for the increased risk of developing mastitis among fatter cows. Ketosis and mastitis may be positively correlated due to the decreased production of chemoattractants that draw leukocytes to the infected quarter and diminished leukocyte responses when ketone bodies are present [41].

5. Conclusions

According to the aim of our study, to determine the associations of automatically recorded body condition scores with measures of production, health, and reproduction (MY, BHB, LDH, and mP4) in dairy cows, we found that automated registered BCSs can represent an indicator of pregnancy success because the BCS of the pregnant cows was higher (+0.29 score). Moreover, the mP4 was 10.93 ng/mL higher compared to that in the group of non-pregnant cows during insemination. The number of inseminations of cows with the highest BCS of >3.5–4.0 was 42.41% higher compared to that among cows with the lowest BCS of 2.5–3.0.

The automatically recorded BCS in cows with subclinical mastitis was higher by 4.96% compared to that in the group of healthy cows. The BCS was the highest in the group of cows with mastitis, with a 4.96% higher score compared to that in the group of healthy cows. Additionally, the highest statistically significant mean differences in body condition score (9.04%) were estimated between the mastitis and metritis groups of cows.

Author Contributions: R.A.: setup of the field experiment, data collection, and selection and management of the experimental group of animals; L.A.: software and algorithm development, design and setup of field experiments, and data collection and analysis; D.M., M.T., M.U. and G.Š.: setup of the field experiment and data collection; A.R. and W.B.: intensive support in the processing of data in the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: Research Council of Lithuania (project number: S-MIP-22-137).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by Ethics Committee (study approval number: PK016965, 6 June 2017).

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mäntysaari, P.; Juga, J.; Lidauer, M.H.; Häggman, J.; Mehtiö, T.; Christensen, J.M.; Mäntysaari, E.A. The Relationships between Early Lactation Energy Status Indicators and Endocrine Fertility Traits in Dairy Cows. *J. Dairy Sci.* **2022**, *105*, 6833–6844. [CrossRef] [PubMed]
2. Yu, G.-M.; Maeda, T. Inline Progesterone Monitoring in the Dairy Industry. *Trends Biotechnol.* **2017**, *35*, 579–582. [CrossRef] [PubMed]
3. Nazhat, S.A.; Aziz, A.; Zabuli, J.; Rahmati, S. Importance of Body Condition Scoring in Reproductive Performance of Dairy Cows: A Review. *Open J. Vet. Med.* **2021**, *11*, 272–288. [CrossRef]
4. Rodríguez Alvarez, J.; Arroqui, M.; Mangudo, P.; Toloza, J.; Jatip, D.; Rodriguez, J.M.; Teyseyre, A.; Sanz, C.; Zunino, A.; Machado, C.; et al. Estimating Body Condition Score in Dairy Cows From Depth Images Using Convolutional Neural Networks, Transfer Learning and Model Ensembling Techniques. *Agronomy* **2019**, *9*, 90. [CrossRef]
5. Gillund, P.; Reksen, O.; Gröhn, Y.T.; Karlberg, K. Body Condition Related to Ketosis and Reproductive Performance in Norwegian Dairy Cows. *J. Dairy Sci.* **2001**, *84*, 1390–1396. [CrossRef]
6. Loeffler, S.H.; de Vries, M.J.; Schukken, Y.H.; de Zeeuw, A.C.; Dijkhuizen, A.A.; de Graaf, F.M.; Brand, A. Use of Ai Technician Scores for Body Condition, Uterine Tone and Uterine Discharge in a Model with Disease and Milk Production Parameters to Predict Pregnancy Risk at First Ai in Holstein Dairy Cows. *Theriogenology* **1999**, *51*, 1267–1284. [CrossRef]
7. Huang, X.; Hu, Z.; Wang, X.; Yang, X.; Zhang, J.; Shi, D. An Improved Single Shot Multibox Detector Method Applied in Body Condition Score for Dairy Cows. *Animals* **2019**, *9*, 470. [CrossRef]
8. Somers, J.R.; Huxley, J.; Lorenz, I.; Doherty, M.L.; O'Grady, L. The Effect of Lameness before and during the Breeding Season on Fertility in 10 Pasture-Based Irish Dairy Herds. *Ir. Vet. J.* **2015**, *68*, 14. [CrossRef]
9. Nyman, A.-K.; Emanuelson, U.; Waller, K.P. Diagnostic Test Performance of Somatic Cell Count, Lactate Dehydrogenase, and N-Acetyl-β-d-Glucosaminidase for Detecting Dairy Cows with Intramammary Infection. *J. Dairy Sci.* **2016**, *99*, 1440–1448. [CrossRef]
10. Antanaitis, R.; Malašauskienė, D.; Televičius, M.; Juozaitienė, V.; Rutkauskas, A.; Palubinskas, G. Inline changes in lactate dehydrogenase, milk concentration according to the stage and number of lactation periods, including the status of reproduction and milk yield in dairy cows. *Pol. J. Vet. Sci.* **2020**, *23*, 153–156. [CrossRef]
11. Leroy, T.; Aerts, J.M.; Eeman, J.; Maltz, E.; Stojanovski, G.; Berckmans, D. Automatic Determination of Body Condition Score of Cows Based on 2D Images. *Precis. Livest. Farming* **2005**, *5*, 251–255.
12. Pinedo, P.; Manríquez, D.; Azocar, J.; Klug, B.R.; De Vries, A. Dynamics of Automatically Generated Body Condition Scores during Early Lactation and Pregnancy at First Artificial Insemination of Holstein Cows. *J. Dairy Sci.* **2022**, *105*, 4547–4564. [CrossRef] [PubMed]
13. Wildman, E.E.; Jones, G.M.; Wagner, P.E.; Boman, R.L.; Troutt, H.F.; Lesch, T.N. A Dairy Cow Body Condition Scoring System and Its Relationship to Selected Production Characteristics. *J. Dairy Sci.* **1982**, *65*, 495–501. [CrossRef]
14. Friggens, N.C.; Chagunda, M.G.G. Prediction of the Reproductive Status of Cattle on the Basis of Milk Progesterone Measures: Model Description. *Theriogenology* **2005**, *64*, 155–190. [CrossRef]
15. Van Eerdenburg, F.J.C.M.; Loeffler, H.S.H.; van Vliet, J.H. Detection of Oestrus in Dairy Cows: A New Approach to an Old Problem. *Vet. Q.* **1996**, *18*, 52–54. [CrossRef]
16. Nielen, M.; Schukken, Y.H.; Brand, A.; Deluyker, H.A.; Maatje, K. Detection of Subclinical Mastitis from On-Line Milking Parlour Data. *J. Dairy Sci.* **1995**, *78*, 1039–1049. [CrossRef]
17. Urton, G.; von Keyserlingk, M.A.G.; Weary, D.M. Feeding Behavior Identifies Dairy Cows at Risk for Metritis. *J. Dairy Sci.* **2005**, *88*, 2843–2849. [CrossRef]
18. Roche, J.R.; Friggens, N.C.; Kay, J.K.; Fisher, M.W.; Stafford, K.J.; Berry, D.P. Invited Review: Body Condition Score and Its Association with Dairy Cow Productivity, Health, and Welfare. *J. Dairy Sci.* **2009**, *92*, 5769–5801. [CrossRef]
19. Shahinfar, S.; Guenther, J.N.; David Page, C.; Kalantari, A.S.; Cabrera, V.E.; Fricke, P.M.; Weigel, K.A. Optimization of Reproductive Management Programs Using Lift Chart Analysis and Cost-Sensitive Evaluation of Classification Errors. *J. Dairy Sci.* **2015**, *98*, 3717–3728. [CrossRef]
20. Inchaisri, C.; De Vries, A.; Jorritsma, R.; Hogeveen, H. Improved Knowledge About Conception Rates Influences the Decision to Stop Insemination in Dairy Cows. *Reprod. Domest. Anim.* **2012**, *47*, 820–826. [CrossRef]
21. Britt, J.H. Impacts of Early Postpartum Metabolism on Follicular Development and Fertility. *Am. Assoc. Bov. Pract. Conf. Proc.* **1991**, 39–43. [CrossRef]
22. Beam, S.W.; Butler, W.R. Effects of energy balance on follicular development and first ovulation in postpartum. *Reprod Fertil. Suppl.* **1999**, *54*, 411–424. [CrossRef]
23. Carvalho, P.D.; Souza, A.H.; Amundson, M.C.; Hackbart, K.S.; Fuenzalida, M.J.; Herlihy, M.M.; Ayres, H.; Dresch, A.R.; Vieira, L.M.; Guenther, J.N.; et al. Relationships between Fertility and Postpartum Changes in Body Condition and Body Weight in Lactating Dairy Cows. *J. Dairy Sci.* **2014**, *97*, 3666–3683. [CrossRef]
24. Nebel, R.L.; McGilliard, M.L. Interactions of High Milk Yield and Reproductive Performance in Dairy Cows. *J. Dairy Sci.* **1993**, *76*, 3257–3268. [CrossRef]
25. Beever, D.E. The Impact of Controlled Nutrition during the Dry Period on Dairy Cow Health, Fertility and Performance. *Anim. Reprod. Sci.* **2006**, *96*, 212–226. [CrossRef] [PubMed]

26. Stevenson, J.S.; Banuelos, S.; Mendonça, L.G.D. Transition Dairy Cow Health Is Associated with First Postpartum Ovulation Risk, Metabolic Status, Milk Production, Rumination, and Physical Activity. *J. Dairy Sci.* **2020**, *103*, 9573–9586. [CrossRef] [PubMed]
27. Zhang, J.; Deng, L.X.; Zhang, H.L.; Hua, G.H.; Han, L.; Zhu, Y.; Meng, X.J.; Yang, L.G. Effects of Parity on Uterine Involution and Resumption of Ovarian Activities in Postpartum Chinese Holstein Dairy Cows. *J. Dairy Sci.* **2010**, *93*, 1979–1986. [CrossRef]
28. Santos, J.E.P.; Bisinotto, R.S.; Ribeiro, E.S. Mechanisms Underlying Reduced Fertility in Anovular Dairy Cows. *Theriogenology* **2016**, *86*, 254–262. [CrossRef]
29. Wiltbank, M.C.; Souza, A.H.; Carvalho, P.D.; Cunha, A.P.; Giordano, J.O.; Fricke, P.M.; Baez, G.M.; Diskin, M.G. Physiological and practical effects of progesterone on reproduction in dairy cattle. *Animal* **2014**, *8*, 70–81. [CrossRef]
30. Ribeiro, E.S.; Lima, F.S.; Greco, L.F.; Bisinotto, R.S.; Monteiro, A.P.A.; Favoreto, M.; Ayres, H.; Marsola, R.S.; Martinez, N.; Thatcher, W.W.; et al. Prevalence of Periparturient Diseases and Effects on Fertility of Seasonally Calving Grazing Dairy Cows Supplemented with Concentrates. *J. Dairy Sci.* **2013**, *96*, 5682–5697. [CrossRef]
31. Buckley, F.; O'Sullivan, K.; Mee, J.F.; Evans, R.D.; Dillon, P. Relationships Among Milk Yield, Body Condition, Cow Weight, and Reproduction in Spring-Calved Holstein-Friesians. *J. Dairy Sci.* **2003**, *86*, 2308–2319. [CrossRef]
32. Royal, M.D.; Darwash, A.O.; Flint, A.P.F.; Webb, R.; Woolliams, J.A.; Lamming, G.E. Declining Fertility in Dairy Cattle: Changes in Traditional and Endocrine Parameters of Fertility. *Anim. Sci.* **2000**, *70*, 487–501. [CrossRef]
33. Antanaitis, R.; Juozaitienė, V.; Malašauskienė, D.; Televičius, M.; Urbutis, M.; Baumgartner, W. Relation of Automated Body Condition Scoring System and Inline Biomarkers (Milk Yield, β-Hydroxybutyrate, Lactate Dehydrogenase and Progesterone in Milk) with Cow's Pregnancy Success. *Sensors* **2021**, *21*, 1414. [CrossRef] [PubMed]
34. Steeneveld, W.; Vernooij, J.C.M.; Hogeveen, H. Effect of sensor systems for cow management on milk production, somatic cell count, and reproduction. *J. Dairy Sci.* **2015**, *98*, 3896–3905. [CrossRef]
35. Hernández Castellano, L.E.; Wall, S.; Stephan, R.; Corti, S.; Bruckmaier, R. Milk Somatic Cell Count, Lactate Dehydrogenase Activity, and Immunoglobulin G Concentration Associated with Mastitis Caused by Different Pathogens: A Field Study. *Schweiz. Arch. Für Tierheilkd.* **2017**, *159*, 283–290. [CrossRef]
36. Assessment of Subacute Mammary Inflammation by Soluble Biomarkers in Comparison to Somatic Cell Counts in Quarter Milk Samples from Dairy Cows—Zank—1998—Journal of Veterinary Medicine Series A—Wiley Online Library. Available online: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1439-0442.1998.tb00799.x (accessed on 13 August 2022).
37. Klein, R.; Nagy, O.; Tóthová, C.; Chovanová, F. Clinical and Diagnostic Significance of Lactate Dehydrogenase and Its Isoenzymes in Animals. *Vet. Med. Int.* **2020**, *2020*, e5346483. [CrossRef]
38. Suriyasathaporn, W.; Schukken, Y.H.; Nielen, M.; Brand, A. Low Somatic Cell Count: A Risk Factor for Subsequent Clinical Mastitis in a Dairy Herd. *J. Dairy Sci.* **2000**, *83*, 1248–1255. [CrossRef]
39. Patel, Y.; Modi, R.J.; Trivedi, M. Relationship Between Body Condition Score. *Trends Biosci.* **2020**, *13*.
40. Valde, J.P.; Lystad, M.L.; Simensen, E.; Østerås, O. Comparison of Feeding Management and Body Condition of Dairy Cows in Herds with Low and High Mastitis Rates. *J. Dairy Sci.* **2007**, *90*, 4317–4324. [CrossRef]
41. Berry, D.P.; Lee, J.M.; Macdonald, K.A.; Stafford, K.; Matthews, L.; Roche, J.R. Associations Among Body Condition Score, Body Weight, Somatic Cell Count, and Clinical Mastitis in Seasonally Calving Dairy Cattle. *J. Dairy Sci.* **2007**, *90*, 637–648. [CrossRef]

agriculture

MDPI

Article

An Improved Intelligent Control System for Temperature and Humidity in a Pig House

Hua Jin [1,*], Gang Meng [1], Yuanzhi Pan [2,3,4], Xing Zhang [1] and Changda Wang [1]

[1] School of Computer Science and Communication Engineering, Jiangsu University, 301 Xuefu Road, Zhenjiang 212013, China
[2] Artificial Intelligence Lab, Zhenjiang Hongxiang Automation Technology Co., Ltd., Zhenjiang 212000, China
[3] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China
[4] Faculty of Business and Economics, The University of Hong Kong, Hong Kong 999077, China
* Correspondence: jinhua@ujs.edu.cn

Abstract: The temperature and humidity control of a pig house is a complex multivariable control problem. How to keep the temperature and humidity in a pig house within a normal range is the problem to be solved in this paper. The traditional threshold-based environmental control system cannot meet this requirement. In this paper, an intelligent control system of temperature and humidity in a pig house based on machine learning and a fuzzy control algorithm is proposed. We use sensors to collect the temperature and humidity in the pig house and store these data in chronological order. Then, we use these time series data to train the GRU model and then use the GRU model to predict the temperature and humidity change curve in the pig house in the next 24 hours. Finally, the mathematical model of the pig house and related equipment is established, and the output power of the related equipment is calculated based on the prediction results of GRU so as to effectively regulate the indoor temperature and humidity. The experimental results show that compared with the threshold-based environmental control system, our system reduces the abnormal temperature and humidity by about 90%.

Keywords: pig; temperature; humidity; GRU; prediction

Citation: Jin, H.; Meng, G.; Pan, Y.; Zhang, X.; Wang, C. An Improved Intelligent Control System for Temperature and Humidity in a Pig House. *Agriculture* **2022**, *12*, 1987. https://doi.org/10.3390/agriculture12121987

Academic Editors: Gniewko Niedbała and Sebastian Kujawa

Received: 13 October 2022
Accepted: 17 November 2022
Published: 23 November 2022

1. Introduction

Modern confined animal production buildings are typically classified as intensive livestock production houses and are prominently used for the production of milk, eggs, and a variety of meats. Such livestock buildings (LB) are ventilated with natural and/or mechanical ventilation systems. In practice, the continuous release of sensible and latent heat, CO2 from animals, and NH3 released from manure contaminate the animal's environment. The contaminated indoor environment affects animal health and the productivity of the operation [1]. Therefore, a ventilation system (VS), an integral part of livestock buildings, maintains a hospitable environment for the animal.

The traditional individual breeding method has low production efficiency, requires a large investment, and poses a high risk to biological assets, which are difficult to create scale economy effects. Large-scale, remote, less populated, and intelligent breeding plays an important role in improving production efficiency, protecting the environment and reducing both labor costs and the probability of epidemics. Big data, the Internet of Things, and 5G make intelligent management of the breeding industry infinitely possible [2].

1.1. Effects of Temperature on the Growth of Commercial Pigs in Large-Scale Breeding Houses

The appropriate temperature is the precondition to ensure the normal development and reproduction of pigs. Pigs are thermostatic animals that maintain a dynamic balance of heat production and heat dissipation through physiological regulation. When the ambient

temperature gradually increases (heat stress), in order to reduce heat production, pigs reduce activity and food intake and increase water intake, slowing growth and even leading to negative weight gain [3]. When the temperature is low (cold stress), the heat dissipation of pigs increases and feed intake increases in order to maintain heat balance. According to Johnson et al. [4], the feed intake of fattening pigs at the high temperature of 28~35 °C can be reduced by 24.1~29.7% compared with the standard daily feed intake; the daily gain is 6.8–28% lower than the expected daily gain. Raising pigs is a process of biological transformation. The chemical energy of the feed is converted into energy in the animal's body. In terms of net energy, the balance between producing and maintaining energy frequently varies. When the ambient temperature is suitable, maintenance energy decreases, production energy increases, and the feed utilization rate is improved. Feed utilization rates can be expressed by the ratio of net energy produced to total energy intake. When the temperature changes between low and high critical temperatures (T1~T2), the physiological regulation of pigs is weak, and the feed utilization rate is at its highest. When the temperature is higher than T2, the skin blood vessels of pigs expand, increasing the body surface temperature, bringing the body heat to the body surface, increasing epidermis water permeability and accelerating the respiratory rate; pigs can change their breathing mode to improve evaporation and heat dissipation. This process increases the proportion of maintenance energy, thus reducing feed utilization efficiency. When the ambient temperature is lower than T1, there is reduced respiratory frequency in pigs and increased metabolic heat production in the body to compensate for the excessive heat loss by consuming a large amount of feed and, at the same time, the amount of activity is increased; the chemical energy in the body will be converted into maintenance energy through exercise, which leads to a decrease in feeding efficiency. In an environment where there are continuously high temperatures, the resistance of pigs significantly decreases, the body heat balance is destroyed, and the body temperature increases; this can lead to comas, heat radiation diseases, and even death in severe cases. In a low-temperature environment, pigs can contract peripheral blood vessels in order to keep warm, causing local frostbite. Low temperatures can cause respiratory and digestive tract diseases but can also often cause rheumatism, arthritis, and other diseases. Low temperature has a greater impact on piglets, and according to statistics, half of dead piglets either freeze to death or die from cold-related diseases. For newborn piglets in low-temperature environments, incidences of diarrhea and other diseases significantly increase. When the ambient temperature is higher than 33~35 °C, semen quality, sperm count, and the motility of boars decrease. Sows can often develop anestrus and behavioral anestrus (ovulation without estrous symptoms); the estrous cycle is prolonged, and the conception rate is reduced. In a study by Quiniou et al. [5], when the temperature reached 28.4 °C, the semen collection of boars decreased by 24 ml, the sperm motility decreased, and the conception rate during estrus decreased by 5.7%. When the temperature reached 27.7 °C, the litter weight decreased by 1.56 kg, and the fertility rate decreased by 13.02%.

1.2. Limitations of Threshold-Based Controllers

At present, the most popular temperature control system on the market is based on a threshold, which is low-cost and easy to use. For example, Qing Du [6] designed an intelligent monitoring system for chicken coops. The system monitors the indoor temperature change in real time. When the indoor temperature exceeds the threshold, the system dynamically adjusts the output power of the temperature control device based on the difference between the indoor temperature and the threshold. The larger the difference is, the higher the output power is until the indoor temperature restores the target temperature. Yiguang et al. [7] developed an intelligent environmental control system for animal houses based on an ARM M3 single-chip microcomputer. This system consisted of an environmental control box, system software, server, and mobile terminals. The environmental control box was composed of a single-chip microcomputer, ferroelectric, SIM card module, relay, display screen, electric frequency converter, and other electronic

hardware components. This box had three functions, including data collection, device control, and data transmission and communication. This system integrated computer technologies with multi-sensor data fusion technology to collect real-time data pertaining to temperature, relative humidity, light intensity, ammonia, and hydrogen sulfide in animal houses. When the actual measured data exceeded the preset range, the system could automatically control the house environment by switching the fans and other equipment on or off. The preset parameters of the environmental control box can be remotely modified on a PC or cell phone and allow the real-time operation of the system.

Although this control method can restore the indoor temperature to normal, it cannot avoid abnormal temperatures. Because this method takes the abnormal temperature as the condition for the device to be turned on, the device will be turned on only if the temperature is abnormal. Therefore, this control strategy cannot always keep the temperature in a suitable range.

1.3. Related Work

The work mainly involves machine learning, mathematical modeling, and fuzzy control algorithms.

In terms of machine learning, our work draws on the work of Svetozarevic et al. [8]. They proposed a fully black-box, data-driven joint control method for indoor temperature and bidirectional electric vehicle charging. The approach is an end-to-end, data-driven approach that uses historical data to obtain control strategies for multi-output (MIMO) control problems in the architecture–mobile coupling domain. The authors use recurrent neural networks (RNNs) to simulate room temperature and discuss the influence of weather forecasts on model accuracy. A deep deterministic policy gradient (DDPG) algorithm is used to find a continuous MIMO control strategy to control the heating/cooling systems and charge/discharge power of bidirectional electric vehicles. The simulation results show that, while saving energy and cost, this method minimizes the violation of comfort, achieves the desired comfort limit, and provides sufficient energy for the next trip of electric vehicles. Inspired by the use of recurrent neural networks to simulate room temperature by Svetozarevic, we decided to use GRU to simulate the temperature and humidity in pig houses. The GRU network is an improvement of RNN, which solves the problem that traditional neural networks cannot process sequence data, and the data set we use is a time series data set.

In terms of the fuzzy control algorithm, our work draws on the research of Gao and Enriko et al. In the study of Gao et al. [9], error E and error change rate Ec are taken as the inputs of the fuzzy PID controller, and ΔKP, ΔKI, and ΔKD are taken as the outputs of the fuzzy PID controller. According to the experience of field engineers and the theory of experts, the fuzzy subsets of inputs E and Ec and outputs ΔKP, ΔKI, and ΔKD are divided into seven grades: "positive big (PB), positive middle (PM), positive small (PS), zero (ZO), negative small (NS), negative medium (NM), and negative large (NB)". The universe of temperature error E and error change rate Ec is $[-2,2]$, and the quantified grades are $\{-2, -1.5, -0.5, 0, 0.5, 1, 1.5, 2\}$. The universe of humidity error E and error change rate Ec is $[-10,10]$, and the quantified level is $\{-10, -8, -6, -4, -2, 0, 2, 4, 6, 8, 10\}$. According to the structural characteristics of brooder houses and the growth environment of chicks and other parameters, the microclimate simulation model of brooder houses was established using the physical law of energy balance. The simulation results show that the control strategy meets the temperature and humidity control requirements and verify the effectiveness of the control strategy and model. The experimental results can guide the actual environmental control of brooder houses. Enriko et al. [10] developed a chicken coop prototype that focuses on temperature control systems on smart poultry farms via the PID control approach. The sensor utilized is a DHT22 sensor with a calibration accuracy of 96.88 percent. The PID response was found to be satisfactory for the system with $Kp = 10$, $Ki = 0$, and $KD = 0.1$, and the time necessary for the system to reach the specified temperature was 121 s with a 1.03% inaccuracy. In our system, the input of the fuzzy controller is the

difference between the real-time temperature and the target temperature, the fuzzy control rule is the formula derived from the heat balance, and the output is the output power of the equipment.

1.4. Contribution of This Paper

This article has the following key contributions:

(1) We recommend using the GRU network to model the indoor temperature. We train the GRU model with more than 40,000 historical data, each of which includes indoor temperature, outdoor temperature, outdoor humidity, outdoor wind direction, outdoor wind speed, and outdoor air pressure.
(2) According to the prediction results of the GRU model, the controller can start the relevant control equipment in advance before the temperature becomes abnormal so that the indoor temperature can always be kept within a normal range.
(3) In the process of temperature control, due to the influence of many factors, such as the heat exchange of indoor and outdoor air and the heat exchange between the walls of the piggery and the outside world, there may be a large gap between the predicted results of the GRU model and the actual temperature values, resulting in the final control effect not being ideal. In view of this, we use a fuzzy control algorithm to flexibly adjust the output power of the equipment according to the gap between the predicted results and the actual value so as to achieve an ideal control effect.
(4) We have designed and implemented a complete temperature regulation system, which can accurately adjust the temperature in the pigsty and effectively avoid abnormal temperatures.

2. Methodology

Figure 1 shows the adjustment process of temperature and humidity, which mainly includes three steps:

(1) Training GRU model: in order to improve the training efficiency and accuracy of the model, the data set is preprocessed, and the important parameters of the model are adjusted.
(2) Formulate macro-regulation strategy: calculate the output power of related equipment according to the predicted results of the GRU model and the relevant parameters of the equipment.
(3) Making micro-regulation strategy: the prediction results of the GRU model will be affected by the heat exchange of indoor and outdoor air, the heat exchange between the walls of the piggery and the outside world, as well as the heat and moisture produced by the pigs, which leads to the deviation between the predicted results of GRU model and the actual value. At this point, if the temperature and humidity are adjusted according to the calculation results of the second step, the temperature and humidity may not be restored to the target value within a specified period of time. Therefore, we introduce the fuzzy control algorithm. The fuzzy controller will adjust the indoor temperature and humidity according to the values collected by the sensor in real time, so as to achieve the purpose of accurate adjustment.
(4) The balance controller: after formulating the temperature and humidity regulation strategy, the balance controller will adjust the temperature and humidity, respectively, according to the strategy.
(5) After the temperature and humidity are restored to the target value, go back to step (1) and start the next round of temperature and humidity adjustment.

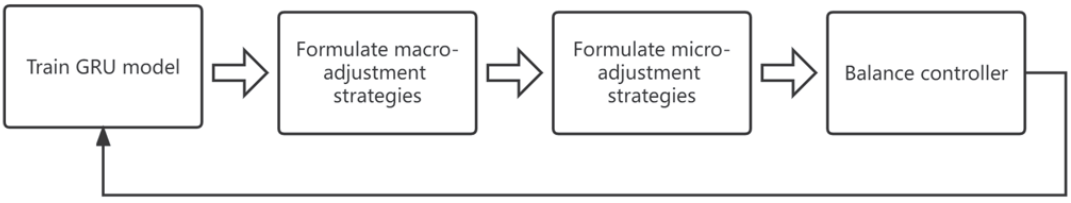


Figure 1. The flow chart of temperature and humidity regulation process includes four steps: training GRU model, formulating macro-adjustment strategies, formulating micro-adjustment strategies, and balance controller.

The dataset used in this section was provided by Zhenjiang Hongxiang Automation Technology Co., Ltd. These data sets were collected by equipment arranged at a pig farm in Liangshan County, Shandong Province, China. The elements of the data set include indoor temperature, indoor humidity, outdoor temperature, outdoor humidity, outdoor wind direction, outdoor wind speed, and outdoor air pressure. The dataset is time series data, and the interval between adjacent data is 5 min.

2.1. Data Preprocessing

Due to sensor damage, temporary failure, and other factors, the measured value of the sensor may be abnormal, which seriously affects the accuracy of the GRU model. Therefore, we preprocessed the dataset before training the model. In addition, in order to improve the convergence speed and accuracy of the GRU model, we also normalized the dataset.

2.1.1. Abnormal Data Detection

The outliers are those that have unreasonable values in the data set. The frequency of abnormal data in the entire data set is very small, and their characteristics are significantly different from normal data. Commonly used methods for detecting abnormal data include outlier detection based on proximity degree [11], density-based outlier detection [12–14], clustering-based outlier detection [15,16], and so on. This paper uses the isolation forest algorithm [17] to detect anomalous data.

The isolation forest algorithm is suitable for the anomaly detection of continuous data. Different from other anomaly detection algorithms that use quantitative indicators, such as distance and density, to describe the degree of alienation between samples, the isolation forest algorithm detects outliers by isolating sample points [18].

2.1.2. Exceptional Data Handling

After detecting the abnormal data, it is necessary to correct the abnormal data. This paper uses the simple moving average method [19] to process abnormal data. The moving average method is a commonly used method to predict one or more future periods of data with a group of recent actual data values. The calculation formula of the simple moving average method is as follows:

$$F_t = \frac{A_{t-1} + A_{t-2} + A_{t-3} + \cdots + A_{t-n}}{n} \quad (1)$$

where F_t denotes the predicted value for the next period; A_{t-1}, A_{t-2}, A_{t-3}, and A_{t-n} represent the actual values of the previous period, the first two periods, the first three periods, and the previous n periods, respectively; n is the number of periods of the moving average.

2.1.3. Data Normalization

Different evaluation indicators often have different dimensions and dimension units, which will affect the results of data analysis. In order to eliminate the dimensional impact between indicators, data standardization is required to solve the comparability between data indicators. After the original data are standardized, all indicators are in the same order of magnitude, which is suitable for comprehensive comparative evaluation. In addition, the data normalization processing also has the advantage of improving the convergence speed and accuracy of the model.

Normalization is to limit the data to a certain range. This paper adopts Min–Max standardization, and the calculation formula is shown as follows:

$$x' = \frac{x - minA}{maxA - minA} \tag{2}$$

where $minA$ and $maxA$ are the minimum and maximum values of attribute A, respectively. Mapping an original value x of A to a value x' between 0 and 1 was performed by max–min normalization.

2.1.4. Processing of Prediction Results of GRU Model

We obtained the weather forecast data for the next 24 hours from the third-party platform. The time interval of these data is 1 hour. Therefore, the time interval of the temperature predicted by the GRU model is also 1 hour. However, the data we want are continuous. Therefore, we use the weighted average method to calculate the temperature *T (t)* at time t. The calculation formula is as follows:

$$T(t) = \frac{(T(t1) + T(t2)) * (t - t1)}{t2 - t1} \tag{3}$$

In the formula, the temperatures of time $t1$ and $t2$ have been predicted by the GRU model; that is, T ($t1$) and T ($t2$) are known, and the temperatures at any other time between $t1$ and $t2$ are not predicted by the GRU model. The time t is between $t1$ and $t2$.

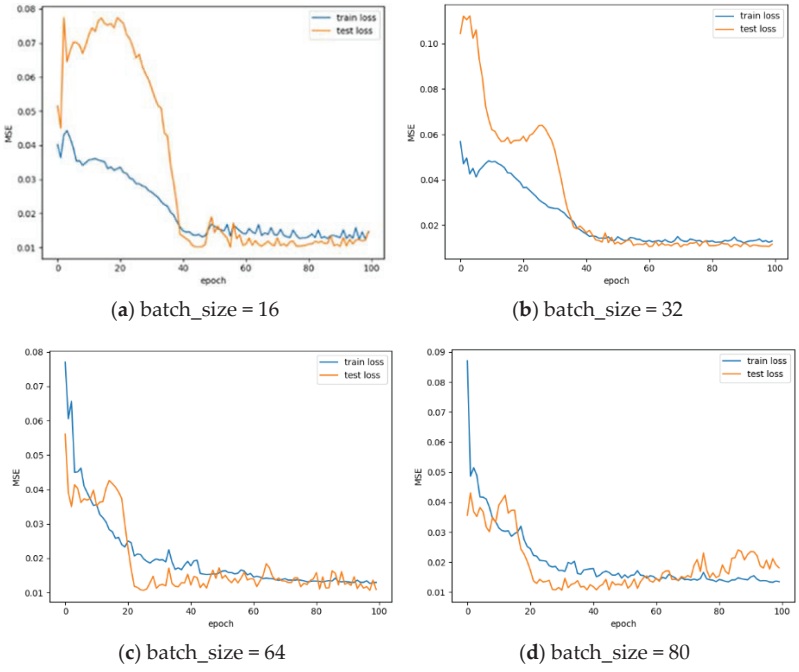
Similarly, humidity is treated in the same way.

2.2. Parameter Adjustment of GRU

We adjusted the important parameters of the GRU model, including the number of samples *batch_size*, the length of parameter time series *Seq* and the number of training rounds *epoch*.

First of all, we carried out an experiment on adjusting the parameter *batch_size*. We set *Seq = 100*, *epoch = 100* and learning rate $\alpha = 0.001$. The curve of the partial loss function is shown in Figure 2. The experimental results show that the MSE values of the training loss function and test loss function are smaller when *batch_size = 32*, and when *epoch* > *40*, MSE tends to be stable, so *batch_size* is set to 32.

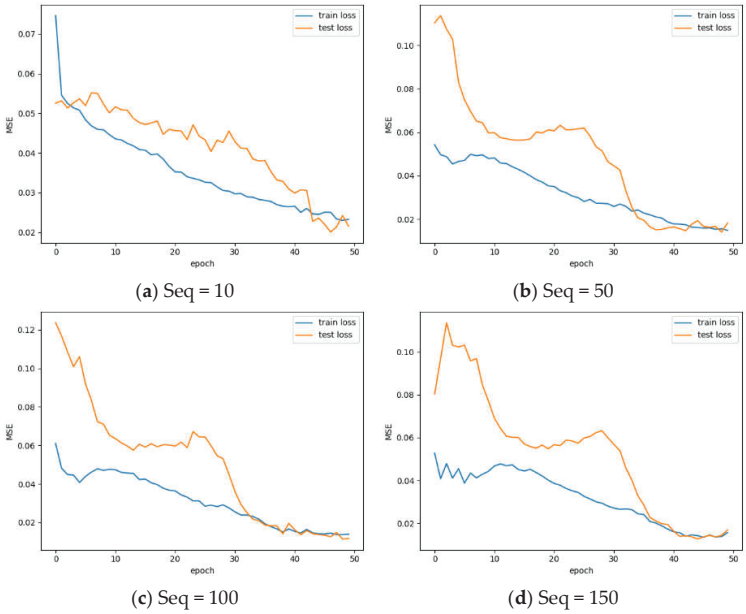
In order to obtain the best value of *Seq*, the *Seq* adjustment experiment was conducted with fixed *epoch = 100*, *batch_size = 32*, and $\alpha = 0.001$. The curve of the partial loss function is shown in Figure 3. The experimental results show that when *Seq = 100*, the MSE values of both the training loss function and the test loss function achieve a small value, and the convergence speed is relatively fast. Therefore, *Seq* is set to 100.



(**a**) batch_size = 16 (**b**) batch_size = 32

(**c**) batch_size = 64 (**d**) batch_size = 80

Figure 2. Parameter *batch_size* adjustment experiment. We set *batch_size* to 16, 32, 64, and 80 for comparative experiments.



(**a**) Seq = 10 (**b**) Seq = 50

(**c**) Seq = 100 (**d**) Seq = 150

Figure 3. Parameter *Seq* adjustment experiment. We set *Seq* to 10, 50, 100, and 150 for comparative experiments.

In the case of *Seq* = *100*, *batch_size* = *32*, and $\alpha = 0.001$, an *epoch* adjustment experiment was conducted. The experimental results are shown in Figure 4. The results show that when *epoch* < *50*, the mean square error of the training set and test set is relatively large; when $50 \leq epoch \leq 110$, the mean square error of the training set and test set reaches the lowest point and tends to be stable; when *epoch* > *110*, the mean square error oscillates or even increases. Considering the computing power and error attenuation of the computer, it is more appropriate to set the epoch value at 50.

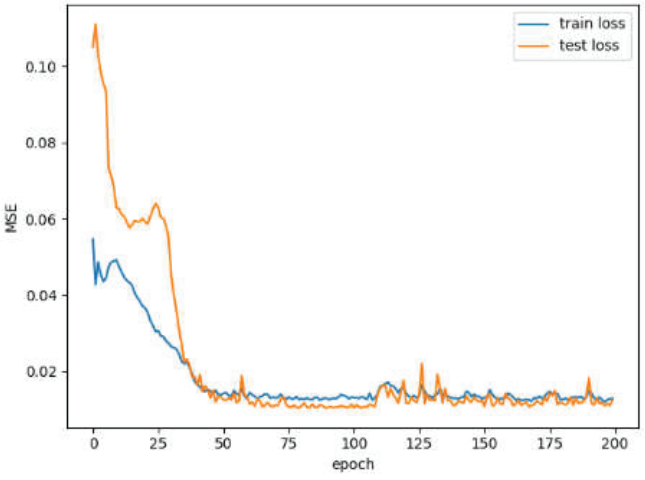


Figure 4. Parameter *epoch* adjustment experiment. The X-axis represents the value of *epoch*, and the Y-axis represents the mean square error.

2.3. Macro-Adjustment Strategy

2.3.1. Formulate the Macroscopic Regulation Strategies of Temperature

The temperature regulation of a pig house is mainly divided into cooling and heating. The cooling equipment mainly includes fans and wet curtains, and the heating equipment mainly includes heaters. In this section, we will derive the output power of each device in different scenarios.

Figure 5 shows the process of macroscopically adjusting the indoor temperature. The specific process is as follows:

(1) First of all, the GRU model will predict the temperature change curve $T(t)$ in the next 24 h.
(2) Compare $T(t)$ with the high-temperature threshold T_{high} and the low-temperature threshold T_{low}. If $T_{low} < T(t) < T_{high}$, it shows that the temperature is not abnormal and directly enters the micro-regulation mode of temperature; if $T(t) \leq T_{low}$, the controller begins to formulate a heating strategy; if $T(t) \geq T_{high}$, the controller begins to formulate a cooling strategy.
(3) If an exception occurs, the controller will run the device in accordance with the policy.
(4) After macro-adjustment, the indoor temperature will not necessarily return to the target temperature, and then it will enter the micro-adjustment mode.
(5) After the micro-adjustment mode, the indoor temperature returns to the target temperature, and the controller enters the next round of regulation.

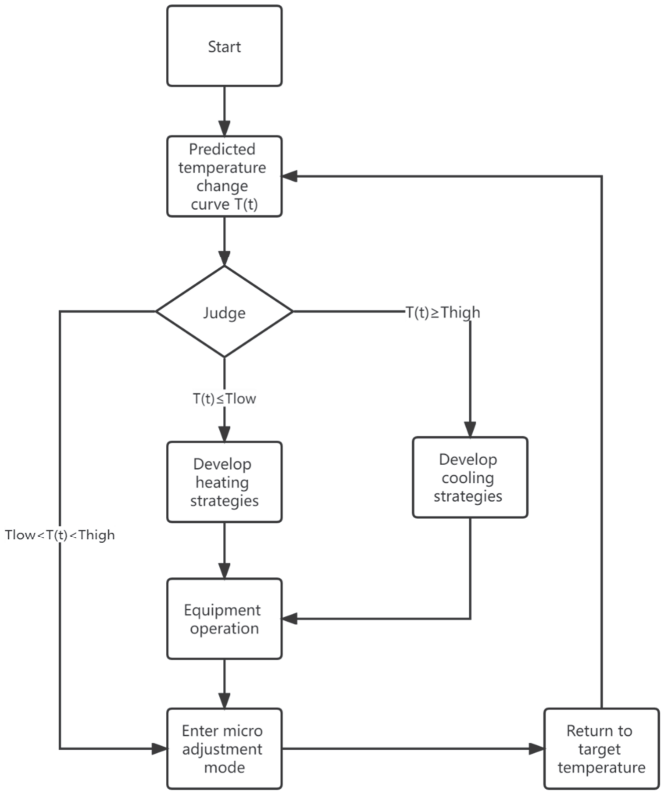


Figure 5. Flow chart of macroscopic regulation of indoor temperature.

The output power of the temperature control device is determined by the severity of the temperature change. Assume that the temperature change curve predicted by the GRU model is $T(t)$. The severity of the temperature change is represented by the absolute value of the derivative of the temperature curve $T(t)$ at time t. The calculation formula is as follows:

$$d = \left|T'(t)\right| \tag{4}$$

The value of d at each time is different, which means that the output power of the temperature control equipment should be changed frequently. For temperature control equipment, if the output power of the equipment is frequently adjusted in a short time, the equipment will be damaged, and the service life of the equipment will be shortened. Therefore, we divide the operating time of the equipment into multiple smaller time periods according to the severity of the temperature change, and the output power of the equipment is the same in each time period. Each time period must satisfy the following conditions: the severity of temperature change $d \leq 0.5$, or $0.5 < d \leq 1$, or $d > 1$.

For the heating process, the minimum output power of the heater is calculated as follows:

$$P\left(t_i, t_j\right) = \frac{C * \rho * V * \left(\left|T\left(t_j\right) - T(t_i)\right| + \left(T_{target} - T_{low}\right) * \frac{t_j - t_i}{DA}\right)}{60 * \left(t_j - t_i\right) * \eta} \tag{5}$$

$P\left(t_i, t_j\right)$ represents the minimum output power of the heating equipment in the time period $\left(t_i, t_j\right)$; C represents the specific heat capacity of the air; ρ represents the density of the air; V represents the volume of the breeding house; T_{target} and T_{low} represent the target

temperature and low-temperature threshold, respectively; DA represents the operating time of the heating equipment; η represents the efficiency of the heating equipment.

During the cooling process, the fan draws the indoor air away so that the outdoor air enters the room through the water curtain. When the flowing air passes through the wet curtain, the water in the wet curtain will absorb the heat in the air and evaporate, taking away a large amount of latent heat so that the temperature of the air passing through the wet curtain is lowered, so as to achieve the purpose of cooling. When the wet curtain is working, the water pump needs to be turned on. Whether the water pump is turned on or not is related to the outdoor temperature and the target temperature. Specifically, when the outdoor temperature is greater than the target temperature, the water pump is turned on; when the outdoor temperature is lower than or equal to the target temperature, the water pump is turned off.

The formula for calculating the minimum output power of the fan is as follows:

$$P(t_i, t_j) = \frac{\left(\left|T(t_j) - T(t_i)\right| + \left(T_{high} - T_{target}\right) * \frac{t_j - t_i}{DA}\right)}{S * T_{out} * (1 - k)} * p \tag{6}$$

T_{high}, T_{target}, and T_{out} represent the high-temperature threshold, target temperature, and outdoor temperature, respectively; S represents the cross-sectional area of the air inlet; k represents the cooling efficiency of the wet curtain (if the water pump is not turned on, k is 0); $P(t_i, t_j)$ represents the minimum output power of the fan in the time period (t_i, t_j); and p represents the corresponding increase in the output power of the fan for every 1m/s increase in the wind speed at the air inlet.

2.3.2. Formulate the Macroscopic Regulation Strategies of Humidity

The humidity regulation process is similar to temperature, but the difference is the calculation formula of equipment output power. Assume that the humidity change curve is $H(t)$. The calculation formula for the severity of humidity change at time t is as follows:

$$d = \left|H'(t)\right| \tag{7}$$

For the humidification process, the minimum output power of the humidifier is calculated as follows:

$$P(t_i, t_j) = p * V * \rho * \alpha * \left(H(t_j) - H(t_i)\right) * ts * C * \left(t_j - t_i\right) / (1000 * DA) \tag{8}$$

In the formula, $P(t_i, t_j)$ represents the minimum output power of the humidifier in the time period (t_i, t_j)); p represents the energy consumption of the humidifier to transport a unit volume of water into the air; V represents the volume of the breeding house; ρ is the density of air; ts is the number of air changes; and C is the loss coefficient.

For the dehumidification process, the minimum output power of the dehumidifier is calculated as follows:

$$P(t_i, t_j) = p * V1 * V2 * \left(H(t_j) - H(t_i)\right) * C * \left(t_j - t_i\right) / DA \tag{9}$$

In the formula, p represents the energy consumption of the dehumidifier to remove the moisture per unit volume in the air; $V1$ represents the volume of the breeding house; $V2$ represents the fresh air volume; and C represents the loss coefficient.

2.4. Micro-Adjustment Strategy

Figure 6 shows the micro-adjustment process based on the fuzzy control algorithm, as described below:

(1) First, compare whether the absolute value of the difference between the actual value $A(t)$ (temperature/humidity) measured by the sensor and the target value *Target* is less than the threshold H. If $|A(t)\text{-}Target| < H$, the room temperature/humidity has

been restored to near the target value, and the current adjustment process is over; otherwise, continue to the next step.

(2) If $A(t)$-$Target > 0$, it is necessary to formulate the cooling/dehumidification strategy; otherwise, the heating/humidification strategy should be established.

(3) The controller operates the equipment according to the policy, and the running time is Duration.

(4) Go back to step (1) and repeat the above steps.

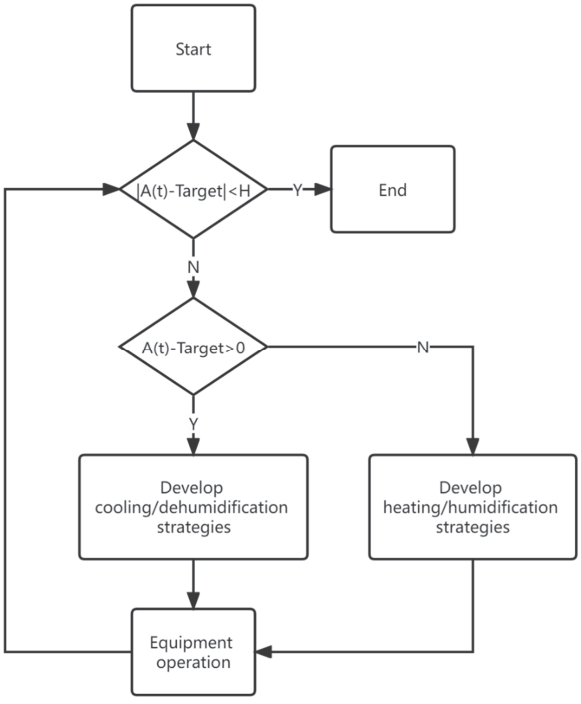


Figure 6. Flow chart of microcosmic regulation of room temperature.

Figure 7 shows the internal structure of the fuzzy controller. The fuzzy controller first receives an input, e. The parameter e is the difference between the value measured by the sensor and the target value. The calculation formula is as follows:

$$e = A(t) - T_{target} \tag{10}$$

The fuzzy interface divides the difference e into three fuzzy sets: positive, zero, and negative. The specific division of the fuzzy set is shown in the following Table 1 (H is a threshold, $|e| < H$ indicates that the actual value is very close to the target value, so there is no need to continue to run the equipment).

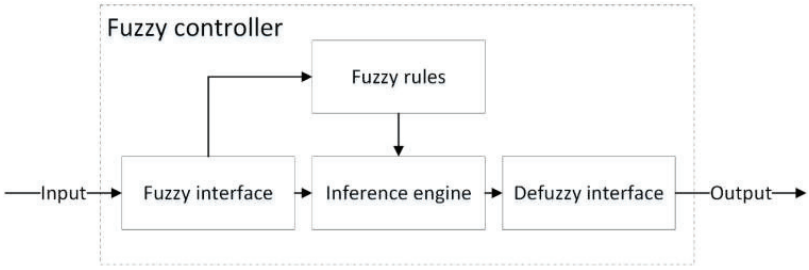


Figure 7. Internal structure of fuzzy controller.

Table 1. Fuzzy sets are divided according to the relationship between *e* and *H*.

Condition	Fuzzy Set
$e < -H$	Positive
$-H \leq e \leq H$	Zero
$e > H$	Negative

In the fuzzy rule and inference engine stage, the controller will do the following: if *e* is in a positive fuzzy set, then a heating/humidification strategy is needed; if *e* is in a negative fuzzy set, then a cooling/dehumidification strategy is needed; otherwise, stop the equipment.

In the defuzzification stage, the controller will calculate the output power of the relevant equipment according to the decision results of the inference engine. If the controller needs to make a heating strategy, the formula for calculating the minimum output power of heaters during the time period (*t*, *t*+*Duration*) is as follows:

$$P(t, t + Duration) = \frac{C * \rho * V * \left| A(t) - T_{target} \right|}{60 * Duration * \eta} \tag{11}$$

If the controller needs to make a cooling strategy, the formula for calculating the minimum output power of fans in the time period (*t*, *t*+*Duration*) is as follows:

$$P(t, t + Duration) = \frac{V * \left| A(t) - T_{target} \right|}{S * T_{out} * (1 - k) * Duration} * p \tag{12}$$

If a humidification strategy is required, the formula for calculating the minimum output power of the humidifiers is as follows:

$$P(t, t + Duration) = p * V * \rho * \alpha * (Target - A(t)) * ts * C * Duration/1000 \tag{13}$$

If a dehumidification strategy is required, the formula for calculating the minimum output power of dehumidifiers is as follows:

$$P(t, t + Duration) = p * V1 * V2 * (A(t) - Target) * C * Duration \tag{14}$$

2.5. Temperature and Humidity Balance Mechanism

In this system, humidity is severely affected during temperature regulation. For example, when the fans are turned on, outdoor air enters the pig house, and the outdoor air is too dry, which will cause the indoor humidity to fall below the threshold. When opening the wet curtain, the humidity in the room may be higher than the threshold due to the

evaporation of water vapor. Therefore, we adjusted the temperature first and adjusted the humidity after the temperature returned to the target value.

3. Experimental Setup

Our experimental site was provided by Zhenjiang Hongxiang Automation Technology Co., Ltd. The experimental site is located in Liangshan County, Jining City, Shandong Province. We conducted an experiment for six months.

3.1. System Framework

Our system needs a lot of data processing and calculation and needs to complete these operations in a short time, so the computing power of the CPU is relatively high. In addition, when we predict indoor temperature, we need to obtain weather forecast data from third-party platforms. Therefore, it is unrealistic to rely solely on the controller to achieve these functions. We divide the system into two parts: cloud platform and terminal.

Figure 8 shows the framework of the system. The cloud platform is responsible for the following modules: data preprocessing, training the GRU model, obtaining weather forecast data, predicting indoor temperature, and formulating macro-adjustment strategies. The terminal device includes a controller and a plurality of acquisition nodes. The controller is responsible for formulating micro-adjustment strategies, controlling the operation of equipment and uploading data. The acquisition node is responsible for collecting data. The cloud platform and the controller communicate through 4G or WIFI, and the controller and collection nodes communicate through the ZigBee network.

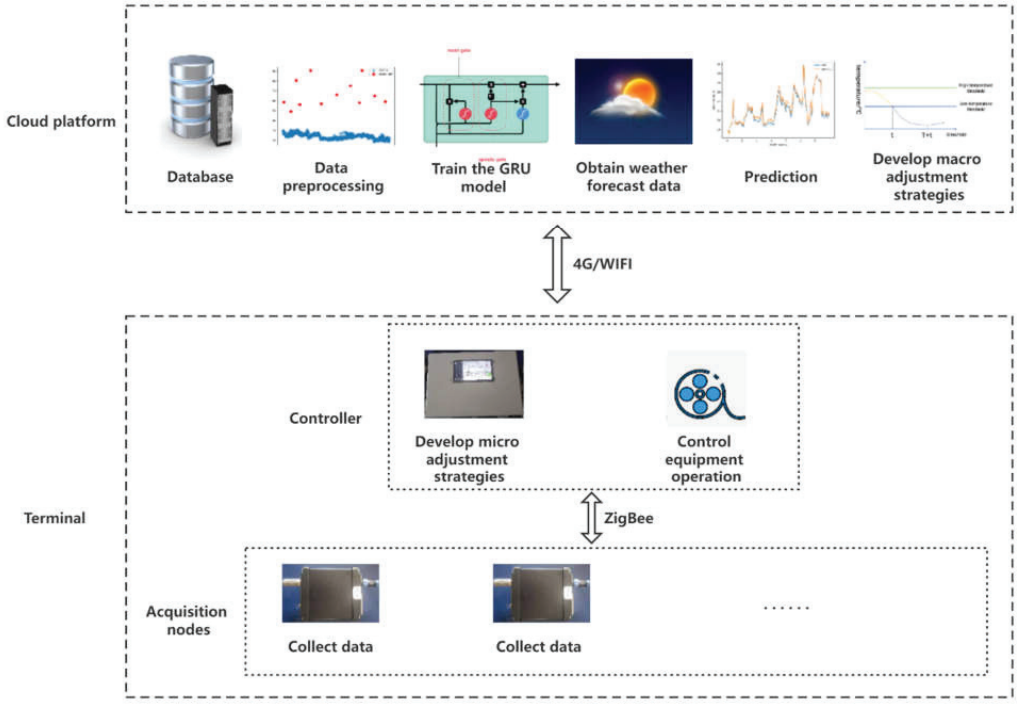
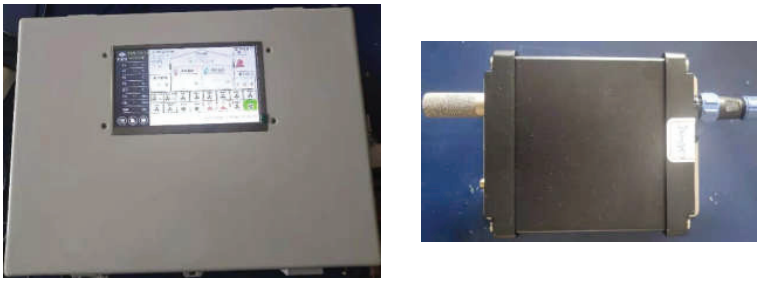


Figure 8. System framework. The framework of the system is mainly divided into two parts: cloud platform and terminal equipment.

3.2. Experimental Equipment

The cloud platform uses the Elastic Compute Service. The terminal equipment includes a controller and two acquisition nodes. The shape of the controller and acquisition node is shown in the following Figure 9:



(**a**) (**b**)

Figure 9. The shape of the controller and the acquisition node. Figure (**a**) is the controller, figure (**b**) is the acquisition node.

The acquisition node includes multiple sensors: temperature, humidity, carbon dioxide, and ammonia. The sensor involved in this paper is a temperature and humidity sensor. The temperature and humidity sensor used in this paper is *TH10S-B-H*, the temperature acquisition range is -40~$120\ ^{\circ}C$, the humidity acquisition range is *0~100% RH*, the temperature accuracy is $0.1\ ^{\circ}C$, the humidity accuracy is *0.1 RH*, the temperature error is $\pm\ 0.2\ ^{\circ}C$, and the humidity error is $\pm 2\%\ RH$.

3.3. Experimental Site Setting

Figure 10 shows the cross-section of the experimental site. The fans and the wet curtains are located on opposite sides of the breeding house. There are 4 heaters, 2 collection nodes, and 2 vents in the farm. The controller is fixed on the outside of the wall of the breeding house. Two humidifiers and dehumidifiers are located near the middle of the breeding house. Below the pigsty is the manure removal area. When the fan is running, the outdoor air enters the room through the wet curtain and is sent to the breeding area through the air inlet. The indoor air is sent to the outside through the fan.

The air intake is located in the ceiling of the pigsty, allowing the air to disperse obliquely into the living area. The study by Hao Li et al. [20] proved that the convective heat transfer coefficient of the pigs in pens with the downward inlet was, on average, 60.4% higher than those with the upward inlet.

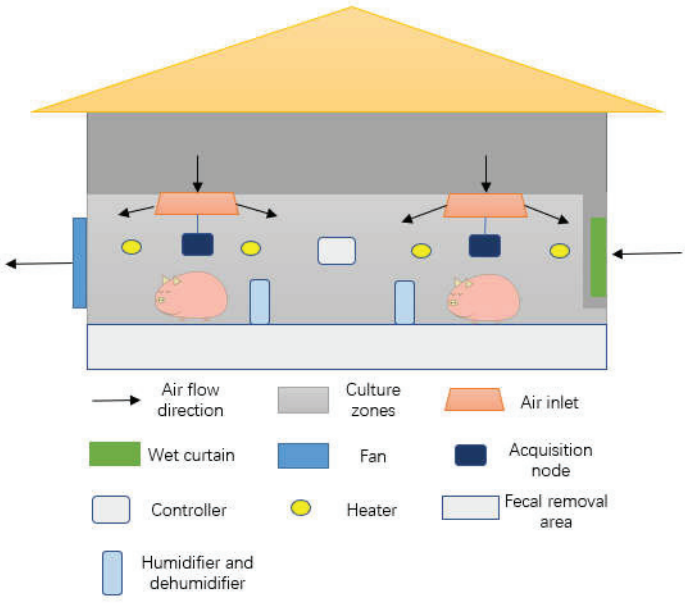


Figure 10. Cross-sectional view of experimental site.

3.4. Explanation of Relevant Experimental Data

The total duration of the experiment was 6 months.

The historical data were mainly collected in the first five months to adjust the super-parameters of the GRU model. These data include indoor temperature, indoor humidity, outdoor temperature, outdoor humidity, outdoor air pressure, outdoor wind direction, and outdoor wind speed. The interval between two adjacent data is 5 min.

When correcting the abnormal data, we set the parameter n in formula (1) to 6 because the temperature of the pigsty will not change much in half an hour.

In the following month, the experiment of adjusting indoor temperature and humidity was carried out. Each prediction of the GRU model predicts the curve of temperature and humidity in the next 24 h (in Section 4.2, we will use experiments to explain why it is 24 h). When calculating the output power of the related equipment, we set the parameter DA in formulas (5), (6), (8), and (9) to 5 min, the *Duration* in formulas (11), (12), (13), and (14) to 10 s, and the threshold H of 2.4 sections to 0.5 °C/1% (temperature/humidity). In other words, in each process of adjusting temperature and humidity, the total time of macro-adjustment is 5 min. The micro-adjustment mode will be entered after 5 min. If the difference between the indoor temperature/humidity and the target temperature/humidity is greater than 0.5/1%, the controller will run for 10 s according to the output power calculated by the fuzzy control algorithm. Then, check whether the indoor temperature/humidity returns to the target temperature/humidity. If the difference between the indoor temperature/humidity and the target temperature/humidity is still greater than 0.5/1%, the controller will recalculate the output power of the equipment according to the fuzzy control algorithm and continue to run for 10 s. Repeat this process until the difference between the indoor temperature/humidity and the target temperature/humidity is less than 0.5/1%. Each time a regulation process is completed, the GRU model re-predicts the temperature/humidity curve in the next 24 h to prepare for the next adjustment.

In addition, the GRU model is not static. Every 24 h, the cloud platform retrains the GRU model based on historical data from the last 3 months.

4. Analysis of Experimental Results

This section provides a detailed analysis of the experimental results of the key steps, including the detection of abnormal data, the prediction results of the GRU model, and the comparison with the adjustment effect of the threshold-based controller.

4.1. Detection of Abnormal Data

We conducted 9000 anomaly detection experiments in the TensorFlow environment. In order to simulate scenarios with different proportions of abnormal data, we inserted different proportions of outliers into the historical data.

Table 2 shows the detection results of the isolation forest algorithm. We can see that when the proportion of outliers is less than or equal to 1%, the detection accuracy of temperature and humidity is almost 100%. When the proportion of abnormal values is less than or equal to 5%, the detection accuracy of temperature and humidity can still remain above 95%. However, in contrast, the accuracy of humidity dropped significantly. When the proportion of outliers is less than or equal to 10%, the anomaly detection accuracy of temperature and humidity drops below 90%. In the actual process of temperature and humidity adjustment in pig houses, the abnormal rate of data collected by sensors is basically below 1%. Therefore, the experimental results of the isolation forest algorithm meet the expectations, and the next steps can be carried out on this basis.

Table 2. The average value of outlier detection results for three different scales, each scale carried out 3000 experiments.

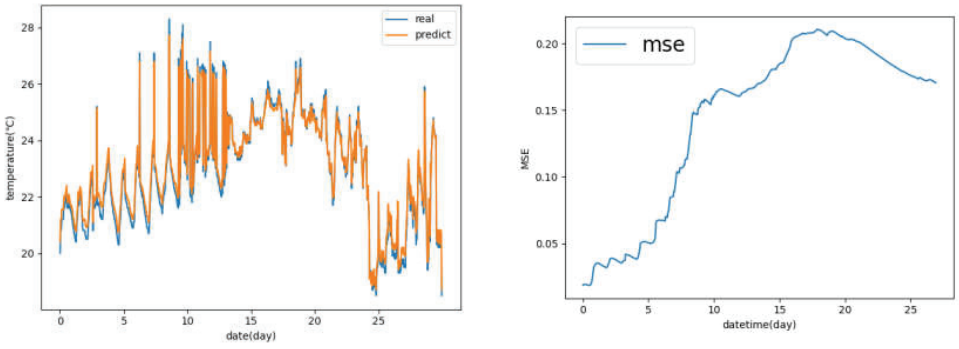
The Proportion of Outliers	Temperature Accuracy	Humidity Accuracy
0%~1%	99.93%	99.91%
1%~5%	97.56%	95.44%
5%~10%	88.36%	81.46%

4.2. Prediction Results of the GRU Model

We trained the GRU model with 3 consecutive months of historical data and used the GRU model to predict the temperature and humidity of the next month.

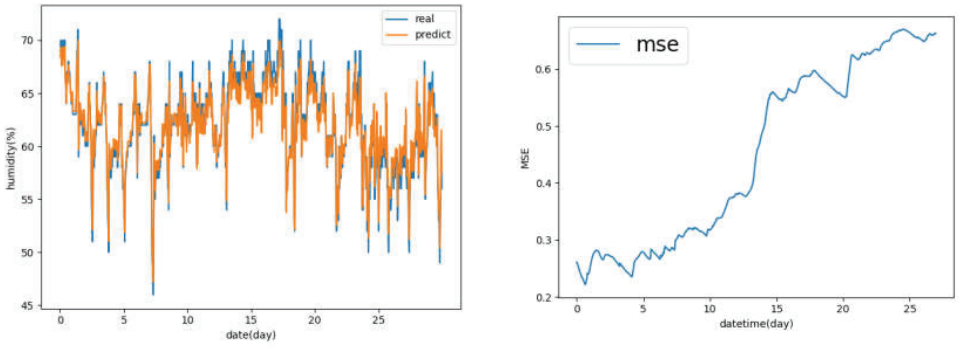
Figures 11 and 12 show the predicted results of temperature and humidity for 30 consecutive days. As can be seen from Figures 11a and 12a, the predicted values of temperature and humidity are basically consistent with the actual values. From Figures 11b and 12b, we find that with the passage of time, the MSE between the predicted value of temperature and humidity and the actual value shows an increasing trend. For the first 1 day, the MSE for temperature was kept below 0.025, and the error for humidity was kept below 0.3. After more than one day, the mean square error of temperature exceeds 0.025. After more than 5 days, the error of temperature and humidity has changed greatly, and the error gradually becomes larger.

Table 3 shows the maximum error of temperature and humidity in different time periods. It can be seen from the table that the maximum error of temperature and humidity gradually increases with the passage of time. On the first day, the maximum error in temperature and humidity was the smallest. From day 1 to day 5, the maximum error in temperature nearly doubled, and humidity remained the same. After the fifth day, the maximum error of temperature and humidity is more than two times that of the first day.



(**a**) (**b**)

Figure 11. Predicted results of temperature for 30 consecutive days. Figure (**a**) is a graph of the actual temperature and the predicted temperature, and Figure (**b**) is the mean square error of the actual temperature and the predicted temperature.



(**a**) (**b**)

Figure 12. Predicted results of humidity for 30 consecutive days. Figure (**a**) is a graph of actual humidity and predicted humidity, and Figure (**b**) is the mean square error of actual humidity and predicted humidity.

Table 3. The maximum error of temperature and humidity in different time periods.

Date	Temperature Max Error (°C)	Humidity Max Error (%)
Day 1	0.4	3
Day 1–Day 5	0.7	3
Day 5–Day 10	0.7	5
Day 10–Day 15	0.9	8
Day 15–Day 30	1.3	12

Based on the analysis of the above figures and tables, we found that the prediction results of the GRU model had the best accuracy on the first day. After more than one day, although the prediction results of the GRU model can still maintain good accuracy, the error increases geometrically compared to the first day. The reason is that for time series data, the data at a certain moment has the strongest correlation with the data at its adjacent moments, and the longer the time distance, the weaker the correlation, and even the correlation between each other can be ignored. From the perspective of time distance,

the first day is closest to the training data set of the GRU model, so the accuracy of the prediction results is the highest. The farther the other time is from the training data set of the GRU model, the lower the accuracy of the prediction results.

Therefore, in order to ensure the accuracy of the prediction results of the GRU model, we recommend retraining the GRU model every 24 h. In this way, the temperature error can be kept below 0.4 °C, and the humidity error can be kept below 3%.

4.3. Comparison with Threshold-Based Controller

We replaced the controller of the system with the threshold-based controller designed by Qing Du et al. [6] and carried out a one-month experiment.

4.3.1. Evaluation Indicators

We use the length of time that the indoor temperature is in an abnormal state every day to measure the system's ability to avoid abnormal temperatures.

$$Total = \sum_{i=1}^{n} (E_i - S_i) \tag{15}$$

where *Total* represents the length of the day when the indoor temperature or humidity is in abnormal states; n represents the total number of temperature or humidity anomalies per day; E_i and S_i represent the end time and start time of the i-th temperature or humidity anomaly, respectively.

4.3.2. Evaluation of the Adjustment Effect

Figure 13 shows the adjustment effect of the two controllers. From the picture, we can see that under the adjustment of the threshold-based controller, the room temperature is abnormal for 20~30 min every day, and our controller reduces it to less than 5 min. However, our expectation is to completely eliminate the temperature anomaly, and the experimental results are not consistent with our expectations. Next, we analyze the regulation process of the controller and find out the reason why the temperature anomaly is not completely eliminated.

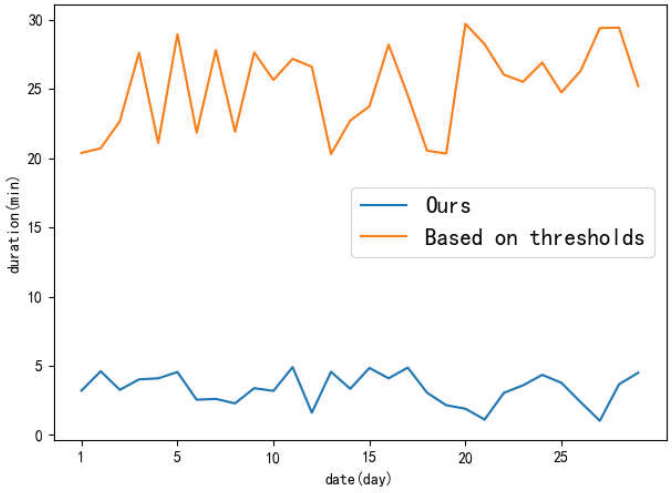


Figure 13. Comparison of the adjustment effect. The X-axis represents the date, and the Y-axis represents the total amount of time that the temperature is abnormal each day.

We find that the temperature may be in an abnormal state at the beginning of each temperature adjustment. Figure 14 shows a complete cooling process. In the previous minute, the room temperature was abnormal. The reason for this phenomenon is that when training the GRU model, we did not consider the effect of the damp and heat produced by the pigs on the indoor temperature. In the process of cooling, pigs will continue to release heat to the outside world, so the indoor temperature is briefly above the high-temperature threshold at the beginning of the operation of the equipment. As a result, the five-minute macro adjustment mode is unable to restore the room temperature to the target temperature. However, in the following micro-adjustment mode, the controller will restore the indoor temperature to an acceptable range; that is, the difference between the indoor temperature and the target temperature is within our preset threshold of 0.5 °C.

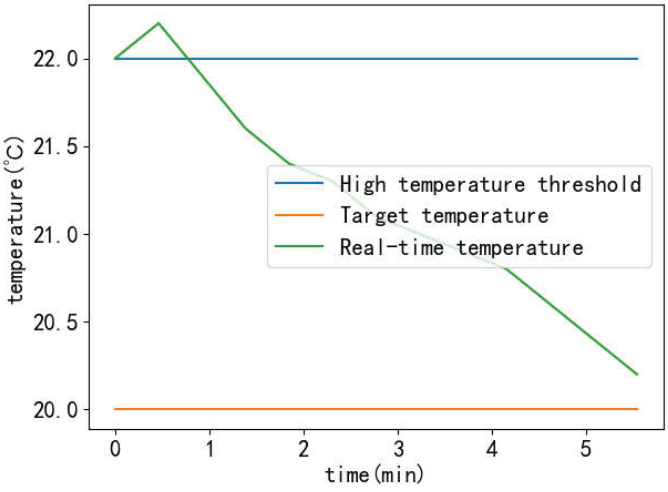


Figure 14. A complete cooling process. The controller reduces the indoor temperature from 22 °C to 20.2 °C in 6 min and stores the indoor temperature every 30 s.

5. Conclusions

This paper designs and implements an improved intelligent control system for the temperature and humidity of a piggery, which mainly includes data preprocessing, training a GRU model, macro-adjustment, micro-adjustment, and other modules. In the data preprocessing module, we use the isolated forest algorithm to detect abnormal data with an accuracy of more than 99%. Before training the GRU model, we adjusted the important parameters of the model, which greatly improved the efficiency and accuracy of the training model. In the terminal controller module, we combined the prediction results of the GRU model with the fuzzy control algorithm to eliminate the influence of humidity and heat generated by the pigs and other factors on the temperature in the piggery and achieved a good regulation effect.

Compared with the threshold-based controller, our controller reduces the abnormal temperature in the pigsty by about 90%. The deficiency is that the zero anomalies of temperature and humidity cannot be realized. Another disadvantage is that the adjustment effect of the system is very dependent on historical data, and the adequacy of historical data directly determines the prediction accuracy of the GRU model.

In a word, the adjustment effect of the system has basically reached our expected effect. It is a useful tool for regulating the temperature and humidity in a piggery.

Author Contributions: Conceptualization, H.J. and G.M.; methodology, H.J.; software, G.M.; validation, H.J., Y.P. and G.M.; formal analysis, C.W.; investigation, H.J., Y.P. and G.M.; resources, Y.P. and G.M.; data curation, Y.P. and G.M.; writing—original draft preparation, G.M.; writing—review and editing, H.J., C.W. and X.Z.; visualization, H.J.; supervision, H.J., C.W. and Y.P.; project administration, H.J. and Y.P.; funding acquisition, Y.P.. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grants 62072217 and 61902156. This research was also supported in part by Zhenjiang Key RD Program (Industrial Prospects and Key Core Technologies) under Grant GY2019015. And The APC was funded by Jiangsu University.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The study did not report any data.

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China under Grants 62072217 and 61902156. This work was also supported in part by Zhenjiang Key RD Program (Industrial Prospects and Key Core Technologies) under Grant GY2019015. In addition, Zhenjiang Hongxiang Automation Co., Ltd. also assisted in the work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Albright, J.L. History and future of animal welfare science. *J. Appl. Anim. Welf. Sci.* **1998**, *1*, 145–166. [CrossRef] [PubMed]
2. Neethirajan, S. The role of sensors, big data and machine learning in modern animal farming. *Sens. Bio-Sens. Res.* **2020**, *29*, 100367. [CrossRef]
3. Johnson, J.S.; Abuajamieh, M.; Fernandez, S.; Seibert, J.T.; Stoakes, S.K.; Nteeba, J.; Baumgard, L. Thermal stress alters postabsorptive metabolism during pre-and postnatal development. In *Climate Change Impact on Livestock: Adaptation and Mitigation*; Springer: New Delhi, India, 2015; pp. 61–79.
4. Quiniou, N.; Dubois, S.; Noblet, J. Voluntary feed intake and feeding behaviour of group-housed growing pigs are affected by ambient temperature and body weight. *Livest. Prod. Sci.* **2000**, *63*, 245–253. [CrossRef]
5. Myer, R.; Bucklin, R. Influence of Hot-Humid Environment on Growth Performance and Reproduction of Swine. 2001. Available online: http://edis.ifas.ufl.edu/AN107 (accessed on 30 May 2007).
6. Du, Q.; Miao, Y.; Zhang, Y. Design of intelligent monitoring system of chicken house environment based on single-chip microcomputer. In *MATEC Web of Conferences*; EDP Sciences: Les Ulis, France, 2018; Volume 227, p. 02008.
7. Zhao, Y.; Nan, X.; Tang, X.; Yang, L.; Xiong, B. Development and application of animal building environmental control system. In Proceedings of the 10th International Livestock Environment Symposium (ILES X), Omaha, NE, USA, 25–27 September 2018; American Society of Agricultural and Biological Engineers: St. Joseph, MI, USA, 2018; p. 1.
8. Svetozarevic, B.; Baumann, C.; Muntwiler, S.; Di Natale, L.; Zeilinger, M.N.; Heer, P. Data-driven control of room temperature and bidirectional EV charging using deep reinforcement learning: Simulations and experiments. *Appl. Energy* **2022**, *307*, 118127. [CrossRef]
9. Gao, L.; Er, M.; Li, L.; Wen, P.; Jia, Y.; Huo, L. Microclimate environment model construction and control strategy of enclosed laying brooder house. *Poult. Sci.* **2022**, *101*, 101843. [CrossRef]
10. Enriko, I.K.A.; Putra, R.A. Automatic Temperature Control System on Smart Poultry Farm Using PID Method. *Green Intell. Syst. Appl.* **2021**, *1*, 37–43. [CrossRef]
11. Aggarwal, C.C. Proximity-Based Outlier Detection. In *Outlier Analysis*; Springer: Cham, Switzerland, 2017; pp. 111–147.
12. Tang, B.; He, H. A local density-based approach for outlier detection. *Neurocomputing* **2017**, *241*, 171–180. [CrossRef]
13. Jayanthi, N.; Hasnabade, M.; Reddy, S.; Deepthi, Y.; Krishna Rao, N.V. Outlier Detection for Data Using Density-Based Technique. In *Energy Systems, Drives and Automations*; Springer: Singapore, 2020; pp. 713–721.
14. Saxena, S.; Rajpoot, D.S. Density-Based Approach for Outlier Detection and Removal. In *Advances in Signal Processing and Communication*; Springer: Singapore, 2019; pp. 281–291.
15. Mahajan, M.; Kumar, S.; Pant, B. A Novel Cluster Based Algorithm for Outlier Detection. In *Computing, Communication and Signal Processing*; Springer: Singapore, 2019; pp. 449–456.
16. Mishra, G.; Agarwal, S.; Jain, P.K.; Pamula, R. Outlier detection using subset formation of clustering based method. In *International Conference on Advanced Computing Networking and Informatics*; Springer: Singapore, 2019; pp. 521–528.
17. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation forest. In Proceedings of the 2008 eighth IEEE international conference on data mining, Pisa, Italy, 15–19 December 2008; pp. 413–422.
18. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation-based anomaly detection. *ACM Trans. Knowl. Discov. Data (TKDD)* **2012**, *6*, 1–39. [CrossRef]

19. Swari, M.H.P.; Qusyairi, M.; Mandyartha, E.P.; Wahanani, H.E. Business Intelligence System using Simple Moving Average Method (Case Study: Sales Medical Equipment at PT. Semangat Sejahtera Bersama). In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2021; Volume 1899, p. 012121.
20. Li, H.; Rong, L.; Zhang, G. Numerical study on the convective heat transfer of fattening pig in groups in a mechanical ventilated pig house. *Comput. Electron. Agric.* **2018**, *149*, 90–100. [CrossRef]

agriculture

MDPI

Article

DFCANet: A Novel Lightweight Convolutional Neural Network Model for Corn Disease Identification

Yang Chen [1], Xiaoyulong Chen [2], Jianwu Lin [1], Renyong Pan [1], Tengbao Cao [1], Jitong Cai [1], Dianzhi Yu [1], Tomislav Cernava [3] and Xin Zhang [1,*]

[1] College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China
[2] College of Tobacco Science, Guizhou University, Guiyang 550025, China
[3] Institute of Environmental Biotechnology, Graz University of Technology, 8010 Graz, Austria
* Correspondence: xzhang1@gzu.edu.cn

Abstract: The identification of corn leaf diseases in a real field environment faces several difficulties, such as complex background disturbances, variations and irregularities in the lesion areas, and large intra-class and small inter-class disparities. Traditional Convolutional Neural Network (CNN) models have a low recognition accuracy and a large number of parameters. In this study, a lightweight corn disease identification model called DFCANet (Double Fusion block with Coordinate Attention Network) is proposed. The DFCANet consists mainly of two components: The dual feature fusion with coordinate attention and the Down-Sampling (DS) modules. The DFCA block contains dual feature fusion and Coordinate Attention (CA) modules. In order to completely fuse the shallow and deep features, these features were fused twice. The CA module suppresses the background noise and focuses on the diseased area. In addition, the DS module is used for down-sampling. It reduces the loss of information by expanding the feature channel dimension and the Depthwise convolution. The results show that DFCANet has an average recognition accuracy of 98.47%. It is more efficient at identifying corn leaf diseases in real scene images, compared with VGG16 (96.63%), ResNet50 (93.27%), EffcientNet-B0 (97.24%), ConvNeXt-B (94.18%), DenseNet121 (95.71%), MobileNet-V2 (95.41%), MobileNetv3-Large (96.33%), and ShuffleNetV2-1.0× (94.80%) methods. Moreover, the model's Params and Flops are 1.91M and 309.1M, respectively, which are lower than heavyweight network models and most lightweight network models. In general, this study provides a novel, lightweight, and efficient convolutional neural network model for corn disease identification.

Keywords: corn leaf disease; real scene; lightweight model; DFCANet

Citation: Chen, Y.; Chen, X.; Lin, J.; Pan, R.; Cao, T.; Cai, J.; Yu, D.; Cernava, T.; Zhang, X. DFCANet: A Novel Lightweight Convolutional Neural Network Model for Corn Disease Identification. *Agriculture* **2022**, *12*, 2047. https://doi.org/10.3390/agriculture12122047

Academic Editors: Gniewko Niedbała and Sebastian Kujawa

Received: 25 October 2022
Accepted: 28 November 2022
Published: 29 November 2022

1. Introduction

Corn is the world's third largest food crop, playing an important role in the agricultural economy [1]. Plant diseases cause significant losses in corn yields [2]. Due to similar disease characteristics, it is hard to distinguish between diseases with the naked eye. Inexperienced growers often misjudge the disease, which leads to the incorrect use of pesticides, affecting the yield and quality of corn and endangering the environment [3]. Having experienced plant pathologists visit the planting site for guidance is costly and difficult to achieve. A lightweight CNN model can be expediently deployed through mobile phones or edge devices. Therefore, automatically identifying leaf diseases through image processing techniques would be of great help to farmers.

With the advancement of digital image processing technology and deep learning methods, humans can automatically identify crop leaf diseases through machine learning algorithms and CNN methods. Increasing studies have been conducted regarding this issue. For instance, Aravind et al. [4] proposed a multi-classification Support Vector Machine (SVM) based on feature bags, which classified leaf spot disease, rust, leaf blight, and healthy leaves in the Plantvillage dataset with an accuracy rate of 83.7%. In addition,

Budiarianto et al. [5] found that RGB features are the most accurate of most classifiers in the commonly used machine learning algorithms.

Using machine learning algorithms to identify crop diseases requires the manual design of features, which is laborious and inconvenient. It is difficult to cope with the identification of multiple crop diseases with different characteristics. Since the CNN method can seamlessly integrate pattern classification while extracting features and improving the efficiency of automatic disease identification, it has been adopted extensively in recent years. By optimizing the LeNet model, Ramar et al. [6] achieved the best accuracy (equal to 97.89%) in classifying three types of diseases in corn in the Plantvillage dataset. Panigrahi et al. [7] proposed a CNN model with both convergence speed and accuracy. The improved CNN model achieved 98.78% of average accuracy. Mohanty et al. [8] obtained 99.35% accuracy in Plantvillage by fine-tuning migration learning on Googlenet. Mishra et al. [9] proposed a CNN model for hardware devices with 88.46% accuracy for the real-time detection of maize diseases in raspberry pi 3.

In the above studies, the crop disease dataset images usually have a simple background. It is significantly different from the real environment. Saleem et al. [10] pointed to the importance of datasets with realistic conditions in plant disease detection and classification. Similarly, in Ferentinos's study, it was demonstrated that the model trained by a simple background image dataset does not work under real conditions [11]. Noise and interference in images collected in natural light make it difficult to distinguish disease features. Therefore, it is difficult to achieve the accuracy of exact disease identification using previous models. Thus, to enhance the details of corn disease characteristics and reduce the complex background noise, Lv et al. [12] proposed an image preprocessing algorithm, called WT-DIR, for a dataset under real conditions. It has an accuracy rate of 98.62% in the DMS-Robust AlexNet network model. Moreover, Zeng et al. [13] improved the ResNet50 method by replacing convolutional kernels, activation functions, and loss functions. The proposed CNN algorithm, called SKPSNet-50, achieved an accuracy of 92.6% in the corn disease dataset taken in real environments.

Furthermore, for complex background images, the attention mechanism can increase the pertinence of the model to focus on the disease area, improving the model's ability to learn the characteristics of diseases. Hence, Akshay et al. [14] proposed an Attention Dense Learning (ADL) mechanism. By stacking five ADLs into a CNN, called DADCNN-5, the simulation achieved a 97.33% accuracy rate in the dataset of complex background images captured by mobile phones. In addition, Zhu et al. [15] achieved an accuracy of 96.58% in a dataset of complex backgrounds by using a transformer-embedded convolutional neural network.

Although the above deep learning models show satisfying accuracies in plant dis-ease identification, they are heavyweight and require abundant computational resources. Therefore, it is necessary to design lightweight neural networks. For instance, Chen et al. [16] simplified the DenseNet algorithm by replacing the standard convolutions with depthwise separable convolutions and proposed a neural network called MS-DNet. The number of parameters in MS-DNet is approximately 0.36M, which is less than the parameter number of existing DenseNet. The proposed model obtained an accuracy rate of 98.32%. Based on MobileNet v2 as the backbone, Chen et al. [17] embedded the attention modules and optimized the loss function. The improved model achieved 98.48% accuracy in rice disease identification under complex background conditions. Meanwhile, the lightweight model named DISE-NET was proposed for the classification of maize small leaf spots by Yin et al. [18]. The dilated inception module and the attention module were designed to enhance the multi-scale feature extraction capability of DISE-NET. Recently, Zeng et al. [19] proposed a lightweight model for mobile deployment called LDSNet, which obtained an accuracy of 95.4% in classifying corn leaf diseases in the field. In addition, Lin et al. [20] presented a lightweight CNN model named GrapeNet based on residual blocks, Residual Feature Fusion Blocks (RFFBs), and Convolution Block Attention Modules (CBAMs)

for grape leaf disease identification. The experiment result showed the GrapeNet model achieved the best classification performance with an accuracy of 86.29%.

In general, large-scale data are necessary to ensure the performance of deep learning models [21]. In routine computer vision tasks, researchers have built large-scale datasets, such as ImageNet [22] and COCO [23]. However, due to the time-consuming nature of collecting and annotating datasets, there are few large-scale and open-access datasets for crop disease. Thus, data augmentation is an efficient way to mitigate data shortfalls. For example, Pan et al. [24] expanded 985 images of corn northern blight and healthy corn leaf to 30,655 through traditional offline data enhancement methods, such as image segmentation, sizing, cropping, and transformation. Richey et al. [25] used various photometric and geometric enhancements to expand the number of images. Nevertheless, offline data augmentation could result in low diversity in crop disease images and lead to overfitting of the model. Alternatively, augmenting data using the Generative Adversarial Network (GAN) method is very efficient. It could enrich the disease characteristics of the dataset [26]. Chen et al. [27] used boost DCGAN and traditional data augmentation methods combined to obtain a better featured-image dataset of corn diseases. However, GAN requires high computational power from the computer and the process is complex. Online data augmentation is a random augmentation of the original data before each training epoch; this method is flexible and simple, and the data are different for each epoch. Traditional online augmentation helps to create many virtual images by randomly rotating, moving, cropping, and flipping the original image. For example, Albarrak, Gulzar, and Hamid et al. [28–30] achieved good results in augmenting datasets, such as seeds and fruits, by traditional online data augmentation. However, the traditional data augmentation method based on geometric transformation loses some feature information about the lesion area. Hence, it is necessary to explore novel approaches for crop disease image online data augmentation.

Table 1 summarizes the main work of the above literature.

Table 1. Comparative analysis of the related work on plant disease identification.

Method	Dataset	Selected Plant/s	Performance Metrics/Accuracy	Ref
SVM	Plantvillage	Corn	83.7%	[4]
SVM	Plantvillage	Corn	83.7%	[5]
Improved LeNet	Plantvillage	Corn	97.89%	[6]
Improved CNN	Plantvillage	Corn	98.78%	[7]
GoogleNet	Plantvillage	38 classes	99.35%	[8]
CNN	Plantvillage	Corn	88.46%	[9]
DMS-Robust AlexNet	Plantvillage, AI challenge, Google web of site and Self-collected diseases	Corn	98.62%	[12]
SKPSNet-50	Own practical database	Corn	92.9%	[13]
DADCNN-5	Own practical database	44 classes	97.33%	[14]
MobileNet-V2 + Transformer	Kaggle datasets	3 classes	96.58%	[15]
MS-DNet	Own practical database	Rice	98.32%	[16]
Mobile-Atten	Self-collected diseases	Rice	98.48%	[17]

Table 1. *Cont.*

Method	Dataset	Selected Plant/s	Performance Metrics/Accuracy	Ref
DISE-NET	Self-collected diseases	Corn	97.12%	[18]
LDSNet	Plantvillage, public website and Self-collected diseases	Corn	95.4%	[19]
GrapeNet	AI challenge	Grape	86.29%	[20]

The identification of corn leaf diseases in a real field environment faces several difficulties, such as complex background disturbances, variations and irregularities in the lesion areas, and large intra-class and small inter-class disparities. In addition, in crop recognition tasks, traditional CNN models with a large number of parameters require more computational resources and are difficult to widely scale up.

To address the above issues, we designed a lightweight CNN model called DFCANet. Inside, the DFCA block improves the feature extraction ability by fusing low-level characteristic information and high-level feature information together. Meanwhile, in order to focus on the lesion area in a complex background, the attention mechanism was applied. Moreover, the DS block can retain useful information better while effectively suppressing noise information by extending the channel dimension and using different down-sampling methods. The main aims of the study are as follows:

(1) Proposing a lightweight convolutional neural network model, called DFCANet, based on DFCA blocks and DS blocks, which are used to identify corn diseases in real environments.
(2) Exploring an online data augmentation method for images of corn leaf diseases.
(3) Comparing the DFCANet with other classical network models to prove the performance advantages of DFCANet and conduct ablation experiments to verify the validity of the different module designs.

2. Materials and Methods

2.1. Data Acquisition and Preprocessing

We acquired corn disease data from four different pathways, including three public datasets and web collections. The three public datasets were CD&S [31], PlantDoc [32], and Corn-Disease (https://github.com/FXD96/Corn-Diseases, accessed on 10 August 2022). In addition, we collected some images from a search engine (https://image.baidu.com/). We obtained images of three types of corn diseases images from the CD&S [31] dataset, namely Northern Leaf Blight (NLB), Gray Leaf Spot (GLS), and Northern Leaf Spot (NLS). The CD&S dataset was acquired under field conditions at Purdue University's Agronomy Center for Research and Education (ACRE) in West Lafayette, Indiana. We obtained images of corn rust leaves under real conditions from the PlantDoc [32] dataset. Additionally, we obtained images of corn leaves infected by the fall armyworm on Corn-Disease. Finally, we crawled the web of healthy corn leaves and a small number of other disease images to balance the data distribution. To summarize, we collected 3271 images, including 537 images of healthy leaves, 688 images of NLB disease, 551 images of NLS disease, 618 images of GLS disease, 445 images of corn rust, and 432 images of corn leaf infected by the fall armyworm. Figure 1 shows a sample corn disease dataset in a real environment.

The data distribution of the dataset in this paper is shown in Figure 2. The training set, the validation set, and the test set were divided in the ratio of 8:1:1. In more detail, the validation set was used to save the model files with the highest accuracy in training, and the test set was used to test the performance of the model.

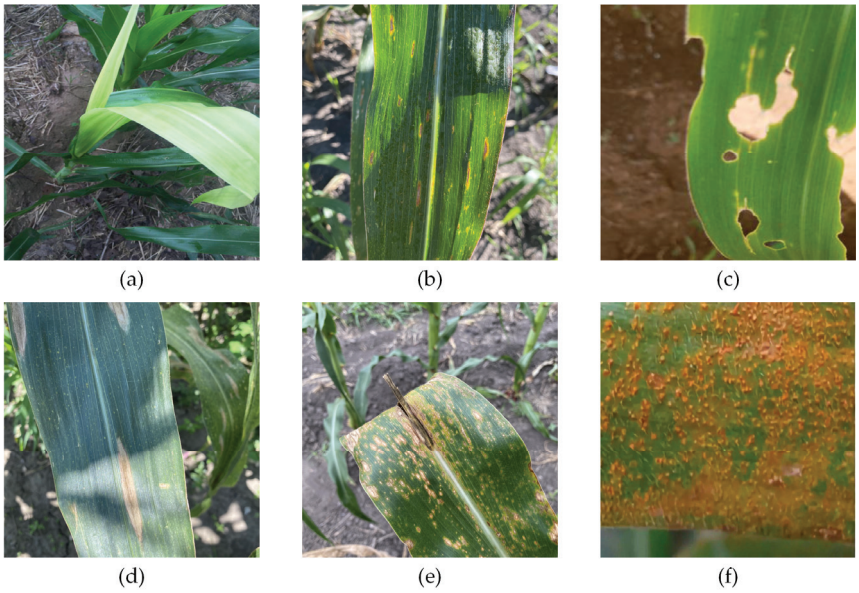


Figure 1. Example of corn leaves. (**a**) Healthy leaf. (**b**) Gray leaf spot leaf. (**c**) Corn leaf infected by fall armyworm. (**d**) Northern leaf blight leaf. (**e**) Northern leaf spot leaf. (**f**) Corn rust leaf.

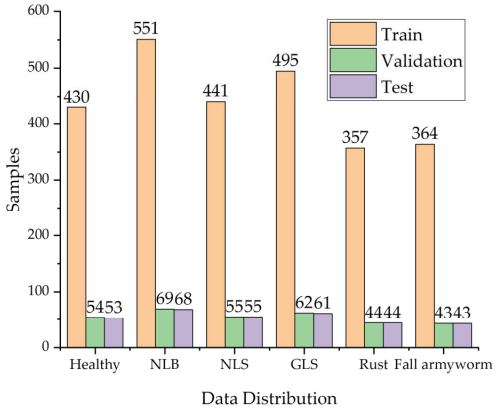


Figure 2. Data distribution of corn leaf disease images.

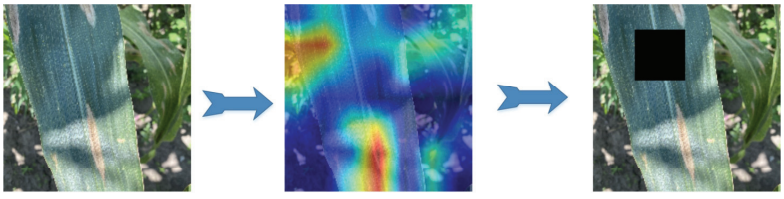
Careful screening and supplementary data make its distribution more balanced, thus avoiding the overfitting of a certain type of disease in model training. The learning of deep learning models requires a large amount of data, but under real-world environmental conditions, the cost of collecting data is high, and the incidence of some crop diseases is low, resulting in a small number of crop disease images collected. Therefore, the data augmentation of the images is necessary. In deep learning, this process can be split into offline data augmentation and online data augmentation. On the one hand, offline data enhancement simply expands the amount of data by manipulating it (e.g., rotating, scaling, and contrasting). However, this method has poor flexibility, requires a vast storage capacity, and is prone to overfitting when the scaling up is too large. On the other hand, online data augmentation is carried out simultaneously in each batch of training, providing high flexibility and enhancing the generalization ability of the model. Figure 2 shows the data distribution of corn leaf disease images.

This article used a data augmentation method called KeepAugment [33], which avoids the disadvantages of traditional Cutout [34] and Random Erasing [35] that might erase diseased areas. Figure 3 illustrates some of the commonly used online data augmentation approaches, which can be seen to potentially mask out diseased areas of corn leaves.

(a) Original Image (b) Cutout (c) Random Erasing (d) GridMask

Figure 3. Classical online data augmentation with (**a**) original image, (**b**) Cutout, (**c**) GridMask [36] and (**d**) Random Erasing.

As shown in Figure 4, the KeepAugment data augmentation method detects important areas through a saliency map and preserves important areas in the image during the augmentation process, avoiding the erasure of disease features. The saliency map region was determined by calculating the backpropagation gradient to obtain the gradient of each pixel value and thus establishing the degree of influence of each pixel value on the category. Additionally, the division of the most important region and the least important region was determined by the sum of all the gradient values of this region being greater or lower than the corresponding threshold value.



Original image Saliency Map Augment

Figure 4. KeepAugment data augmentation. The red area represents the location of interest to the model. KeepAugment data augmentation measures the vital areas of the leaf with the saliency map to avoid masking the area of the lesion.

2.2. DFCANet Model

Figure 5 presents the structure of the DFCANet. It consists of a DFCA block, two different DS blocks, a depthwise convolution layer, an adaptive average pooling layer, and a classifier. All parts will be presented, in detail, in the remainder of this section.

2.2.1. DFCANet

As already shown, the DFCANet mainly consists of two blocks: A Double Fusion with Coordinate Attention [37] (DFCA) and a Down-Sampling block (DS). The complete architecture of DFCANet is shown in Table 2. The DFCA block is used to extract features, taking into consideration that the size of the feature map and the number of channels does not change. The DS block is used for down-sampling, expanding the number of channels and reducing the size of the feature map.

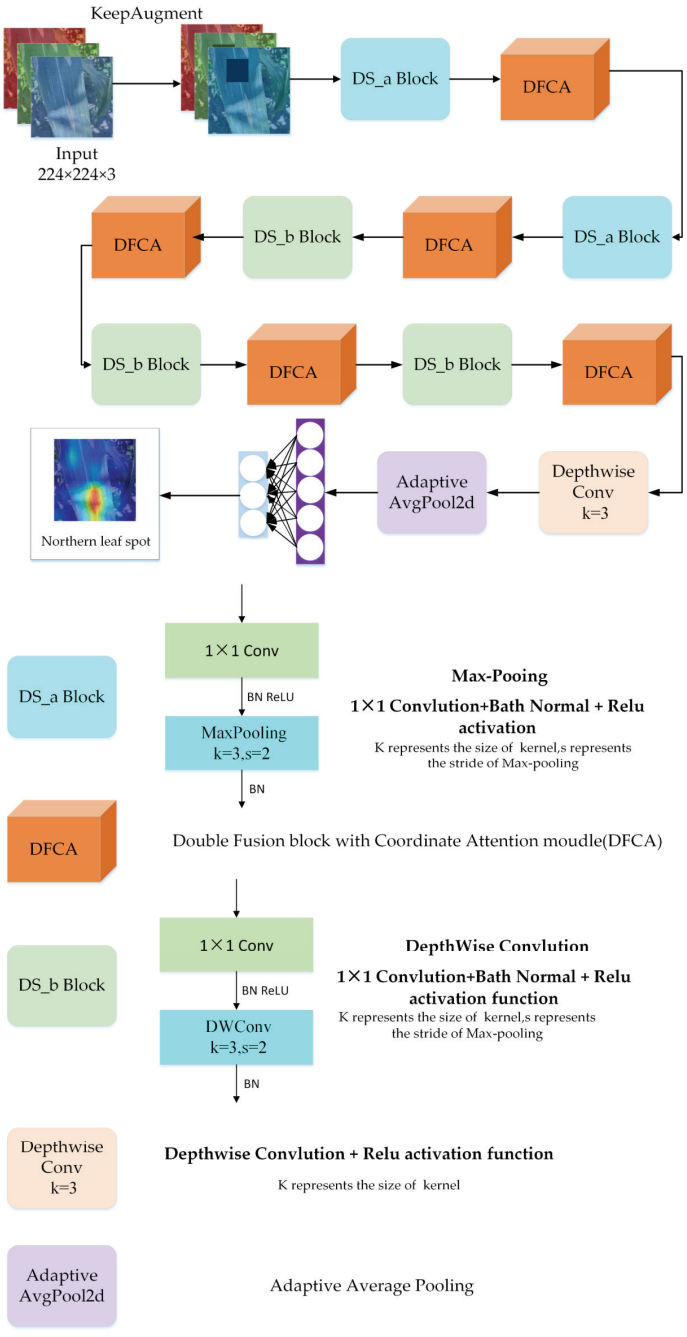


Figure 5. Structure of DFCANet. It is composed of DS_a Blocks, DFCA modules, DS_b Blocks, a Depthwise Conv layer, an Adaptive Average Pooling layer, and a classifier.

Table 2. Architecture of DFCANet.

Input	Operator	Output
$224^2 \times 3$	DS	$112^2 \times 12$
$112^2 \times 12$	DFCA	$112^2 \times 12$
$112^2 \times 12$	DS	$56^2 \times 48$
$56^2 \times 48$	DFCA	$56^2 \times 48$
$56^2 \times 48$	DS	$28^2 \times 96$
$28^2 \times 96$	DFCA	$28^2 \times 96$
$28^2 \times 96$	DS	$14^2 \times 192$
$14^2 \times 192$	DFCA	$14^2 \times 192$
$14^2 \times 192$	DS	$7^2 \times 384$
$7^2 \times 384$	DFCA	$7^2 \times 384$
$7^2 \times 384$	Depthwise Conv	$7^2 \times 384$
$7^2 \times 384$	AdaptiveAvgPool2d	$1^2 \times 384$
$1^2 \times 384$	FC	-

2.2.2. DFCA Block

The DFCA block is shown in Figure 6. It mainly consists of three parts, namely, an inverted bottleneck, coordinate attention, and double fusion. Inspired by ConvNeXt [38], the inverted bottleneck was designed to better extract corn disease characteristics. The coordinate attention module looks for areas of disease characteristics and suppresses noise by recalibrating the channel weights of the input image. The feature information, extracted by CNN in several layers, is different and the feature fusion is to combine low-level extracted features with high-level extracted features to improve the recognition capability of the DFCA block. As shown in Figure 6, the low-level features are input features with a higher resolution and texture information. These features are transformed into mid-level features using feature extraction in the inverted bottleneck and recalibration in the attention module. The mid-level features have information on the location of diseased areas extracted by convolutional feature extraction in order to obtain high-level features with abstract semantic information. ResNet [39] completes the feature fusion by introducing quick identity connections and achieving widespread applications. Unlike ResNet, which only performs feature fusion in one stage, we performed two feature fusions to make it more thorough: The first phase consists of the fusion of low-level features with mid-level features, which can effectively locate the lesion area and ignore the background information, whereas the second phase involves low-level features with high-level features, which greatly enhances the model's ability to extract subtle disease features.

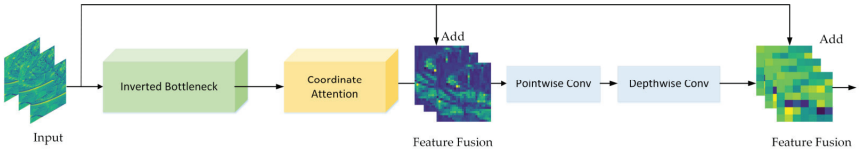


Figure 6. Presentation of the DFCA block.

2.2.3. Depthwise and Pointwise Convolution

This study replaced ordinary convolution with depthwise convolution and pointwise convolution for lightweight. As shown in Figure 7, depthwise convolution convolutes each channel grouped into convolutions, allowing for a better collection of spatial features while significantly reducing the number of parameters. Additionally, point-by-point convolution sets the height and width of the convolutional kernel to one and the depth to the number of input channels. A lower parameter volume is maintained after being cascaded by deep

convolution. The ratio of depthwise convolution and pointwise convolution calculation and ordinary convolutional calculation is as follows:

$$\frac{M \bullet D_k^2 \bullet D_F^2 + M \bullet N \bullet D_F^2}{M \bullet N \bullet D_K^2 \bullet D_F^2} = \frac{1}{N} + \frac{1}{D_k^2} \tag{1}$$

where D_K represents the height and width of the convolutional kernel, D_F represents the height and width of the input feature map, M represents the number of the channel, and N represents the number of channels of the output feature map. From the above equation, it is easy to see that depthwise convolution and pointwise convolution greatly reduce the computational effort of ordinary convolution.

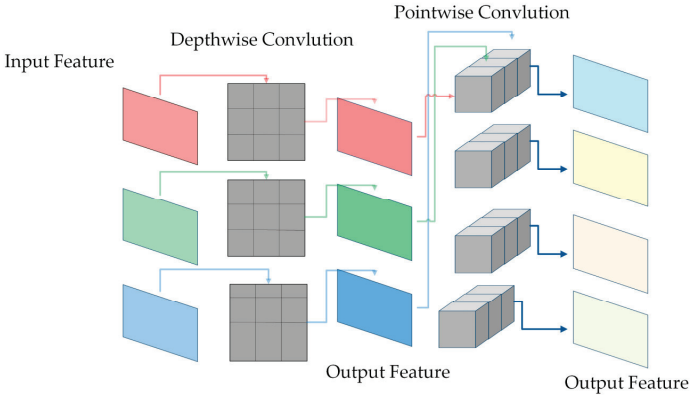


Figure 7. Depthwise convolution and Pointwise convolution.

2.2.4. Inverted Bottleneck

The 1 × 1 pointwise convolution can extend the channel dimension. The inverted bottleneck structure can enrich the feature information, and it has been widely used since it was proposed in MobileNet V2 [40]. Drawing on ConvNeXt, we designed an inverted bottleneck structure as shown in Figure 8, placing the depthwise convolution in front of the pointwise convolution, which saved a considerable amount of computation time.

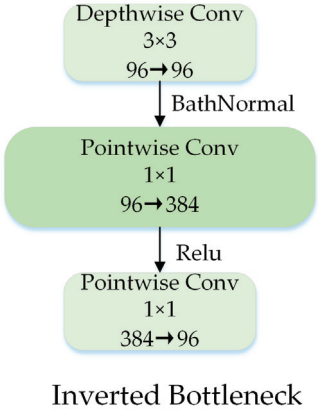


Figure 8. Block designs for inverted bottleneck.

2.2.5. Coordinate Attention Module

Figure 9 shows the details of the CA module.

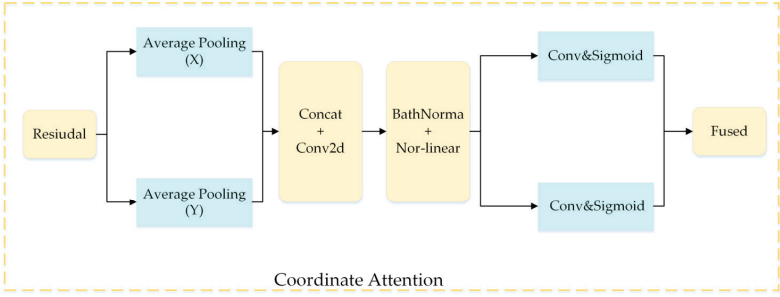


Figure 9. Coordinate attention module.

Specifically, the coordinate attention module is decomposed in the vertical and horizontal directions and is transformed into a pair of one-dimensional feature codes as shown in the equation below:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \tag{2}$$

where x_c denotes the input of the CA module. H and W denote the height and width of the pooling kernel, respectively. z_c represents the c-th channel's output.

The global pooling of the above formula can encode spatial information globally, but it retains location information with difficulty. Thus, the pooling along in both directions is decomposed. After the horizontal decomposition, the output of the first channel with a height h is as follows:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h,i) \tag{3}$$

Equally, after applying the vertical decomposition, the output of the c-th channel with a width of w is as follows:

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j,w) \tag{4}$$

The transformation of the attention module, described above, captures long-term dependencies along one spatial direction while saving position information in the other spatial direction. Finally, the convolution is fed after being spliced together by the feature diagram of the aggregation above:

$$\mathrm{f} = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right) \tag{5}$$

where $\mathrm{f} \in \mathbb{R}^{C/r \times (H+W)}$ represents a feature map encoded in the horizontal and vertical directions, δ denotes a nonlinear activation function, and F_1 represents the 1×1 convolution layer.

Then, we used the 1×1 convolution to compress the channels. After that, we used the sigmoid function for normalization, from which we can obtain two outputs:

$$\mathbf{g}^h = \sigma\left(F_h\left(\mathbf{f}^h\right)\right) \tag{6}$$

$$\mathbf{g}^w = \sigma\left(F_w\left(\mathbf{f}^w\right)\right) \tag{7}$$

where $\mathbf{g}^w$ and $\mathbf{g}^h$ denote the attention weights of the two spatial directions.

The final output can be expressed as:

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{8}$$

As shown in Figure 6, the CA module is added after the inverse bottleneck in the DFCA block. It not only ensures that the network is lightweight, but also makes the allocation of resources more reasonable, and the coordinate attention module can quickly find the area of interest in the disease image, ignoring the background and the noise information. Specifically, the coordinate module can use the sigmoid activation function, described above, to weight the characteristic map of the convolutional network (mainly learning to weight the coefficient) in order to obtain a new salient feature map. This new map is integrated with the original feature map, which can effectively emphasize the disease area and suppress noise and background information, heightening the learning ability of the network.

2.2.6. DS Block

The down-sampling operation, by reducing the size of the feature map, can not only increase the receptive field but also reduce the amount of computation. Usually, the down-sampling operation leads to the loss of some feature information; so, in this paper, the DS block used a 1 × 1 pointwise convolution to extend the channel dimension to reduce the loss of information during the down-sampling process. Two different DS modules were used in this work. Disease images acquired in real scenarios often contain complex backgrounds and noise; thus, they affect the recognition accuracy of CNN. In response to this problem, the Max-Pooling operation remains an important feature to remove distracting information in the initial stages of DFCANet. Depthwise convolution is used to down-sample the feature map, which preserves more useful information than pooling in the later stages. Simultaneously, the characteristics of the disease can be further extracted, and the network performance can be enhanced. Figure 10 illustrates the structure of the two proposed DS blocks.

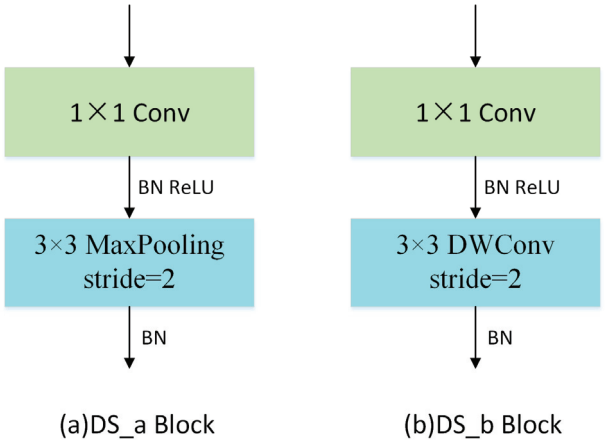


Figure 10. Down-sampling module with (**a**) max-pooling and (**b**) depthwise convolution.

2.3. Experimental Environment and Hyperparameter Setting

The experimental hardware in this study was the Windows 10 Operation System and the Intel(R) Xeon(R) W-2235. The GPU model was NVIDIA GeForce RTX 2080Ti and the software environment adopted was Python 3.8.8 and the Pytorch 1.11.0 framework. The hyperparameters were set as follows: The cross-entropy loss function (CE) was used as the loss function and the Adam optimizer [41] was used to optimize the model. The initial

learning rate and batch size during training were set to 0.002 and 32, respectively. The number of iterations was set to 100.

2.4. Evaluation Indexes

In this study, the accuracy, precision, recall, and F1 score were utilized to perform the evaluation metrics to measure the model's performance.

3. Results

The experiment consisted of three parts: The first was to explore the effects of different data enhancement methods on corn disease identification; the second was to compare different network models; and the third was the ablation experiment of DFCA.

3.1. Impact of Different Data Augmentation Methods on the Model

To explore the online data augmentation methods applicable to corn leaf disease under real environmental conditions, we conducted the experiments as shown in Table 3. The experimental results indicate that most online augmentation methods can increase the diversity of the data, enhance the generalization ability of the DFCANet, and improve the recognition accuracy of the model. The data augmentation strategy using the KeepAugment method, while training DFCANet, eventually led the DFCANet model to reach, respectively, the values of 0.9785, 0.9817, 0.9792, and 0.9803 when computing the average accuracy, precision, recall, and F1 score. Thus, using the KeepAugment data augmentation method leads to an increase in the accuracy by 2.44% compared to the same simulation launched without the use of the data augmentation method. While using the GridMask data augmentation method, we found that the recognition accuracy of the model was reduced by approximately 0.61% because the feature areas of corn diseases might be masked during the application of this data augmentation algorithm.

Table 3. Results of different data augmentation methods when applying the DFCAnet algorithm.

Method	Accuracy	Precision	Recall	F1-Score
Without augmentation	0.9541	0.9535	0.9573	0.9547
Cutout	0.9694	0.9720	0.9720	0.9720
Random Erasing	0.9633	0.9659	0.9671	0.9665
GridMask	0.9480	0.9504	0.9541	0.9516
KeepAugment	0.9785	0.9817	0.9792	0.9803

Note that the random masks of some regions in data augmentation approaches, such as Cutout, Random Erasing, and GridMask, are not necessarily suitable for crop leaf disease datasets.

3.2. Comparative Experiment of Different Network Models

The above results demonstrate the effectiveness of the KeepAugment online augmentation strategy on the corn leaf disease dataset, and the strategy was used by default in all the subsequent experiments.

The proposed DFCANet model was compared with the classical CNN models. As shown in Table 4, DFCANet's average classification accuracy, precision, recall, and F1 score reached 0.9785, 0.9817, 0.9792, and 0.9803, respectively, which are all superior to other CNN methods' performances. The DFCANet model's Params and Flops (Floating points of operations) have far lower values than heavyweight CNN models (VGG16, ResNet50, EffcientNetV2_b0 and ConvNeXt-base). The accuracy of the DFCANet model is 5.19% higher than that of ResNet50. EffcientNet-B0 performs better because its network structure is based on the Neural Architecture Search (NAS) technique to obtain the optimal set of parameters. In addition, EffcientNet has higher accuracy while having lower Flops. The number of model parameters (Params) of DenseNet121 is only 6.96M, but, as this method connects all channels to each other for feature reusing, it helps the model to retain the background noise information in the complex environment of the dataset easily and affects the classification accuracy of the model. Due to the redundant connection mechanism of

DenseNet121, the model has 2.88G Flops. ShuffleNet V2 reduces the complexity of the model by channel splitting and channel disruption, so the Params and Flops of ShuffleNet V2 are lower than those of DFCANet, but the evaluation metrics, such as the accuracy and recall of the model, are lower than those of DFCANet. MobileNet-V2 and MobileNet-V3 have similar network structures, but MobileNet-V3's structure is derived from NAS and has lower Flops. DFCANet's Params are lower than those of MobileNet-V2 and MobileNetV3-large. In general, lightweight network models are better suited to smaller datasets than heavy models. This indicates that lighter models are more effective in small samples.

Table 4. Results of the comparative experiment of corn leaf disease classification performance using the proposed and classic CNN models.

Model	Accuracy	Precision	Recall	F1-Score	Params (M)	Flops
VGG16 [42]	0.9480	0.9483	0.9491	0.9485	134.29	15.50G
ResNet50 [39]	0.9266	0.9314	0.9299	0.9298	23.51	4.12G
EffcientNet-B0 [43]	0.9571	0.9565	0.9591	0.9568	40.09	398.02M
ConvNeXt-B [38]	0.9296	0.9271	0.9336	0.9289	89.00	15.40G
DenseNet121 [44]	0.9357	0.9383	0.9383	0.9383	6.96	2.88G
MobileNet-V2 [40]	0.9480	0.9488	0.9500	0.9485	2.22	318.96M
MobileNetV3-Large [45]	0.9480	0.9453	0.9516	0.9476	4.20	226.43M
ShuffleNetV2-1.0× [46]	0.9449	0.9440	0.9483	0.9460	1.26	149.57M
DFCANet	0.9785	0.9817	0.9792	0.9803	1.91	309.1M

GLS, FA, NLB, H, NLS, and R represent, respectively, corn gray leaf spot, corn leaf infected by fall armyworm, corn northern leaf blight, corn healthy leaf, corn northern leaf spot, and corn rust leaf. Figure 11 shows the confusion matrix for the nine models. DFCANet leads other network models in the number of true positive samples in each category of the test set. This model achieves all correct predictions in three categories (FA, NLB, and NLS). In the test set of NLB, other network models predicted the highest number of errors, e.g., ConvNeXt-B predicted 10 errors, and most of the models incorrectly predicted NLB as GLS, because NLB and GLS have very similar disease characteristics (the symptoms of GLS include multiple greyish-brown and narrow rectangular lesions and the symptoms of NLS include multiple brown spots with circular concentric lesions). In the test set of FA, many models were able to complete all the correct predictions due to the fact that, unlike the complex characteristics of the disease, the insect pest caused the corn leaves to have more distinctive characteristics of mutilation. All obtained results demonstrate the effectiveness of our proposed network model for corn disease identification when dealing with complex backgrounds.

3.3. Ablation Experiments

Table 5 shows a comparison of the results of adding different attention modules (the Squeeze-and-Excitation (SE) module, the CBAM module, and the Coordinate Attention (CA) module). All three attention mechanisms are lightweight, so they do not significantly increase the number of parameters of the model. The accuracy of DFCANet with the CA module is 1.52% and 2.75% higher than that of the SE and CBAM modules, respectively. This is due to the fact that the CA module can capture not only cross-channel information but also direction perception and location perception information. This combination helped the model to accurately locate and identify disease areas and disease features in corn leaves.

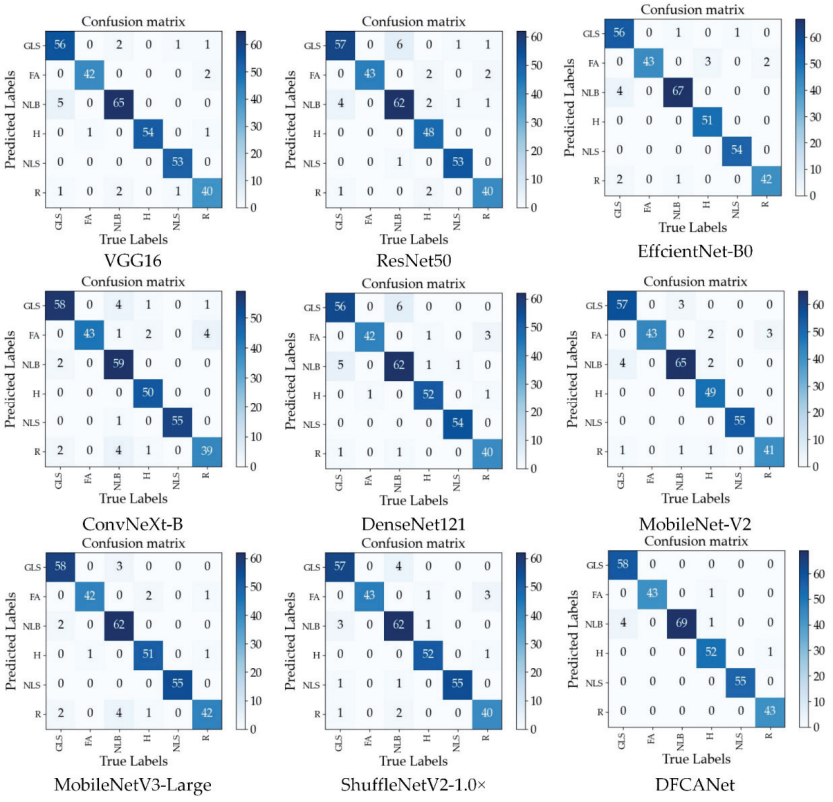


Figure 11. Confusion matrix for the nine models.

Table 5. Comparison of the results of different attention mechanisms.

Attention Mechanism	Accuracy	Precision	Recall	F1-Score	Param (M)	Training Time (Seconds)/Epoch	Test Time (Seconds)/Epoch
SE [47]	0.9633	0.9631	0.9650	0.9638	1.89	70	33
CBAM [48]	0.9510	0.9503	0.9561	0.9526	1.89	68	33
CA	0.9785	0.9817	0.9792	0.9803	1.91	68	33

As shown in Figure 12, the class activation maps of the Grad-cam [49] visualization model vividly demonstrate the regions on which the model focused with different attention mechanisms. It is easy to find that all three attention mechanisms can effectively ignore the background and focus on the lesion area in the images of the real scene; however, compared to the SE and CBAM modules, the CA module can more accurately locate the lesion region of the corn leaves and focus on the disease characteristics more precisely.

To better illustrate the effect of different modules on the model, we designed ablation experiments. First, we simplified DFCANet to a baseline. Specifically, the CA module and the double feature fusion branch were removed from DFCANet, and the DS block was replaced by a common down-sampling module. As shown in Table 6, the accuracy of the model increased by 1.22% after adding the CA module to the baseline network, indicating that the CA attention mechanism can effectively improve the recognition ability of the model. After adding the double-feature fusion branch, the model recognition accuracy increased by 1.53%, indicating that the second double-feature fusion can effectively fuse both the shallow feature information and the deep feature information. Finally, the DS

module designed in this paper improved the model recognition accuracy by 2.44%, which proves that the DS Block can retain effective information and filter interference information at the same time. Compared to the baseline, the subsequent modules do not increase the Params too much, and the increase in Flops is within limits.

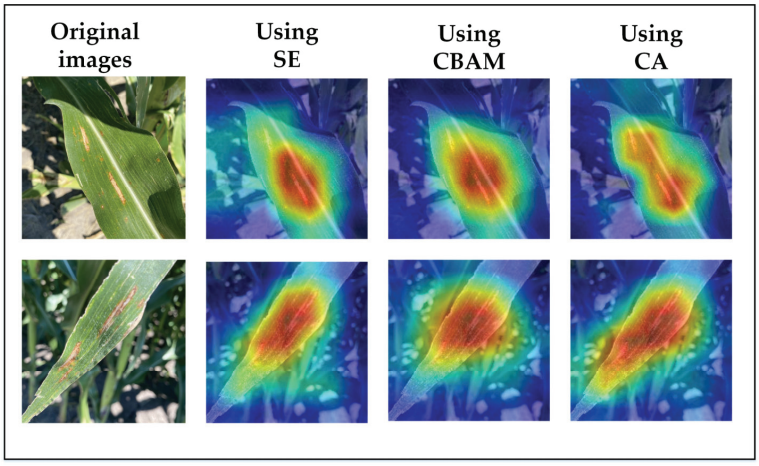


Figure 12. Grad-cam visualization results obtained using different attention mechanisms.

Table 6. Results of the ablation experiments.

	Accuracy	Precision	Recall	F1-Score	Param (M)	Flops
Baseline	0.9266	0.9263	0.9319	0.9281	1.89	302.1M
+CA	0.9388	0.9413	0.9423	0.9413	1.91	304.6M
+Double Fusion	0.9541	0.9605	0.956	0.9576	1.91	304.6M
+DS Block	0.9785	0.9817	0.9792	0.9803	1.91	309.1M

3.4. Simulation of Real Weather Data Augmentation Experiments

To simulate more realistic scenarios, we need to consider the weather conditions of the real environment, which are ignored in the existing datasets. Therefore, we algorithmically added rain, fog, and stronger sunlight effects to the images. As shown in Figure 13, the data augmentation of simulated real weather introduced noise and interference, which has higher requirements on the feature extraction ability of the model and the ability to suppress noise. Data augmentation expanded the number of images to twice the original. In the above experiments, we demonstrated the effectiveness of KeepAugment, so the augmentation of KeepAugment was performed after simulating the data augmentation of real conditions.

Figure 13. Data augmentation by simulating real weather.

Simulating data augmentation under real conditions inevitably causes the image to introduce considerable noise and interference. These disturbances are also commonly

encountered in images taken in real environments, so this affords good robustness to the model. As shown in Table 7, the accuracy, recall, and other evaluation metrics of each model improved after the data enhancement of the simulated real environment. It is worth noting that MobileNetv3-Large performs well in data enhancement with noise interference due to the addition of the SE attention module in MobileNetv3-Large. This shows the effectiveness of the attention module for noise suppression.

Table 7. Comparisons of the recognition accuracy of different models.

Model	Auccary	Precision	Recall	F1-Score	Params (M)	Flops
VGG16 [42]	0.9663	0.9656	0.9655	0.9643	134.29	15.50G
ResNet50 [39]	0.9327	0.9341	0.9359	0.9338	23.51	4.12G
EffcientNet-B0 [43]	0.9724	0.9738	0.9748	0.9736	40.09	398.02M
ConvNeXt-B [38]	0.9418	0.9430	0.9460	0.9440	89.00	15.40G
DenseNet121 [44]	0.9571	0.9560	0.9591	0.9568	6.96	2.88G
MobileNet-V2 [40]	0.9541	0.9580	0.9576	0.9571	2.22	318.96M
MobileNetV3-Large [45]	0.9633	0.9620	0.9665	0.9638	4.20	226.43M
ShuffleNetV2-1.0× [46]	0.9480	0.9501	0.9528	0.9508	1.26	149.57M
DFCANet	0.9847	0.9853	0.9853	0.9853	1.91	309.1M

Figure 14 shows the validation curves in the validation set for different models. As the data augmentation of simulating real environment poses challenges to model training, in order to achieve a better convergence, we increased the training epoch to 200. It can be found that our proposed DCFCANet achieved the most advanced performance compared to other network models. DFCANet's accuracy after curve smoothing averaged approximately 97%, ahead of other networks throughout the training period. Noise and disturbances were introduced in the data enhancement for simulating severe weather, which tested the robustness of the model. From the perspective of the magnitude of the curve fluctuations, the proposed model DFCANet has strong robustness. In addition, DFCANet is substantially ahead of the other models in the initial stage of training, at approximately 73% accuracy (other models are around 45%), which indicates that DFCANet has excellent fitting ability.

3.5. Comparative Experiments of the Public Datasets

To further verify the superiority of the model in this paper, we conducted the experiments in the public dataset Plantvillage (https://github.com/spMohanty/PlantVillage-Dataset, accessed on 10 August 2022), as shown in Table 8. Plantvillage provides the same training and validation sets, so the experimental results in the literature are highly comparable. In recent years, numerous researchers have made significant contributions to the identification of plant diseases. In the task of maize disease identification, Aravind and Budiarianto et al. used machine learning algorithms for feature extraction and classification, so the recognition accuracy was lower than that of CNN models. Ramar and Panigrahi et al. obtained a better performance by improving a CNN model. The CNN model designed by Mishra et al. performed excellently in hardware. The DFCANet designed in this paper achieved the highest recognition accuracy of 99.47% in maize disease identification. In the task of identifying all diseases in Plantvillage, Mohanty, Mohameth, and Huang et al. all achieved high accuracy after fine-tuning by transfer learning. The CNN model designed in this paper was not pre-trained by ImageNet, yet it obtained the highest accuracy, which reflects that DFCANet has excellent feature extraction ability and fitting ability.

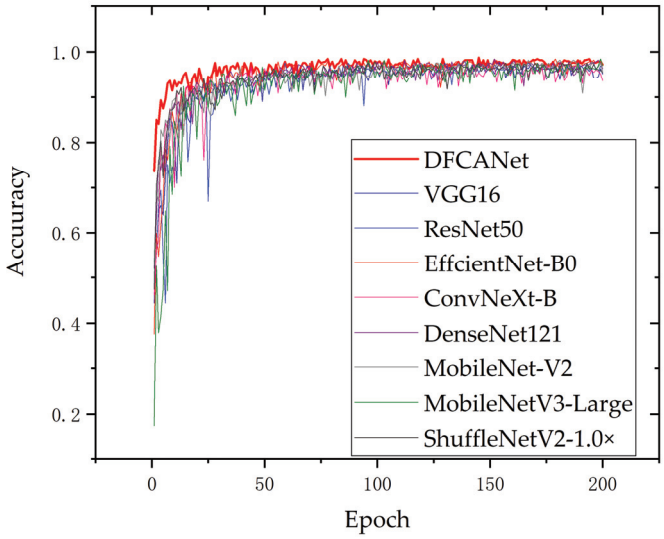


Figure 14. Accuracy variation curves of the different models on the validation set.

Table 8. Comparisons of the recognition accuracy of different models on PlantVillage.

Method	Plant	Accuracy/%	Ref
SVM	Corn	83.7%	Aravind et al. [4]
SVM	Corn	83.7%	Budiarianto et al. [5]
Improved LeNet	Corn	97.89%	Ramar et al. [6]
Improved CNN	Corn	98.78%	Panigrahi et al. [7]
CNN	Corn	88.46%	Mishra et al. [9]
DFCANet	Corn	99.74%	-
GoogleNet	Plantvillage	99.35%	Mohanty et al. [8]
VGG16	Plantvillage	97.82%	Mohameth et al. [50]
NasNet	Plantvillage	99.15%	Huang et al. [51]
DFCANet	Plantvillage	99.58%	-

The images in the Plantvillage dataset were collected under laboratory conditions with simple background images. To further validate the accuracy of DFCANet in field conditions, we performed comparison experiments in the CD&S dataset. The CD&S dataset was the main source of data for this paper and includes images of three types of maize diseases (NLS, GLS, and LB). As with Plantvillage, CD&S provides a fixed training and validation set, which has high comparability. As shown in Table 9, numerous classical models achieved excellent accuracy due to the balanced data distribution of the CD&S dataset. DFCANet achieved the highest accuracy, further verifying its superiority.

Table 9. Comparisons of the recognition accuracy of different models on CD&S dataset.

Model	Accuracy	Precision	Recall	F1-Score	Params (M)	Flops
VGG16 [42]	0.9529	0.9632	0.9528	0.9530	134.29	15.50G
ResNet50 [39]	0.9745	0.9739	0.9746	0.9743	23.51	4.12G
EffcientNet-B0 [43]	0.9847	0.9850	0.9846	0.9846	40.09	398.02M
ConvNeXt-B [38]	0.9477	0.9473	0.9483	0.9476	89.00	15.40G
DenseNet121 [44]	0.9681	0.9680	0.9683	0.9680	6.96	2.88G

Table 9. *Cont.*

Model	Accuracy	Precision	Recall	F1-Score	Params (M)	Flops
MobileNet-V2 [40]	0.9757	0.9756	0.9756	0.9756	2.22	318.96M
MobileNetV3-Large [45]	0.9719	0.9730	0.9716	0.9720	4.20	226.43M
ShuffleNetV2-1.0× [46]	0.9808	0.981	0.9813	0.981	1.26	149.57M
DFCANet	0.9923	0.9926	0.9923	0.9923	1.91	309.1M

4. Discussion

Crop disease recognition is a challenging task in fine-grained classification, mainly because of the small number of crop disease samples, the difficulty of recognizing the type of diseases in complex scenes, and the large intra-class variation and small inter-class variation of crop disease features [52]. In addition, in crop recognition tasks, traditional CNN models with a large number of parameters require more computational resources and are difficult to scale up widely [6,53]. Hence, in this study, we designed a novel model structure to meet the challenges encountered in the real environment when capturing corn leaf photos in order to detect and classify diseases.

To address the problem of a large number of parameters in the traditional CNN model, we replaced ordinary convolution with deep separable convolution. It was demonstrated in Equation (1) that the computation of deep convolution and point convolution is $1/N + 1/D_K^2$ times that of ordinary convolution. The authors of [16] used the same approach to simplify DenseNet, but our custom-designed CNN structure is more specialized and lightweight. Generally, attention mechanisms allow for better resource allocation and effective focus on crop disease areas. The authors of [20,54] introduced CBAM and SE modules to improve the model by 3% and 4.26%, respectively. In this study, we introduced the CA module, and the above-presented ablation experimental results (see Table 5) show that adding the CA modules improved the accuracy of the baseline by 1.22%. Additionally, the visualization results (see Figure 12) show that the CA module focuses better on the lesion area than the SE and CBAM modules. As the complexity of the disease situation increases, a bottleneck emerges in the role of attention mechanisms, which requires models with stronger feature extraction capabilities. The fusion of features is an effective way to enhance the model's ability to extract features. Inspired by the different feature fusion methods of ResNet [39] and DenseNet [44], we discarded the redundant connection method of DenseNet and adopted a deeper fusion method than ResNet. The low-level features were fused to the middle-level and high-level layers, respectively. The experimental results show that our DFCA improved the accuracy of ResNet and DenseNet by 5.17% and 4.28%, respectively. Feature fusion enriches the diversity of features, but the way to retain information in down-sampling has been neglected by researchers [55,56]. Therefore, we designed a DS block to reduce the loss of feature information, and the model accuracy was improved by 2.44% in the ablation experiment. However, most previous studies use traditional offline data augmentation methods or simple online augmentation methods [19,20,28–30]. The experimental results obtained in this study show that the wrong data augmentation method can mask the disease area and reduce the accuracy of the model. For example, the accuracy of using GridMask is 0.61% lower than that of not using data augmentation. Our study validated an online data augmentation method named KeepAugment for crop diseases with a 4.2% higher accuracy than without augmentation. These findings provide implications for future studies on crop disease identification. In addition, despite pixel-based data enhancement methods such as random erasure and masking, we also took full account of what may happen in real environments and performed data enhancement by simulating severe weather. After this data augmentation, DFCANet model was improved to 98.47% accuracy. The experimental results show that data enhancement by simulating real weather not only improves the accuracy of the model but also improves the robustness of the model.

In addition to performing extensive experiments to verify the superiority of each module in DFCANet, we also compared it to the other literature and models in pub-

lic datasets. In the public dataset Plantvillage, for disease identification of corn leaves, DFCANet achieved 99.74% classification accuracy, and the highest accuracy of 98.78% in [4–7,9], which is 0.96% lower than this study. Our method achieved the highest accuracy of 99.58% in the identification of all categories in Plantvillage. The accuracy of our method was 0.23%, 1.76%, and 0.43% higher than that of Mohanty et al. [8], Mohameth et al. [50], and Huang et al. [51], respectively. In addition, DFCANet achieves 99.12% accuracy in CD&S (public dataset with complex background), surpassing other classical CNN models.

Overall, the DFCANet proposed in this study is lightweight and effective for crop disease identification in complex backgrounds. Moreover, our study provided an insightful exploration of the effectiveness and robustness of data enhancement methods for crop disease datasets, which provided a reference for future crop disease data enhancement.

5. Future Work

In the future, our research will focus on the following aspects: (1) Collecting datasets and solving data imbalance problems; (2) exploring more ways to augment data for agricultural imagery; (3) changing the input size of the model; and (4) deploying models to mobile or other edge devices.

The datasets collected in the field environment are critical. However, due to the differences in seasons and regions, the collected datasets showed a long-tail distribution. In this study, we balanced the data volume of each category by artificially supplementing other data sources. In future work, we will focus on solving the problem of the long-tail distribution of data.

For agricultural images, we will explore more data augmentation approaches, such as GAN.

The input image size in this paper was 224 × 224, which can be applied to the general CNN model and lightweight model. However, such an input size ignores some information, so in the follow-up research, we will adopt a larger input size or slice the image into a path. Further, we will adapt the learnable Resizer Model to learn a more appropriate resizing method.

We also need to deploy the model on mobile phones or other edge devices for users to identify disease types.

6. Conclusions

In this study, a lightweight CNN model, called DFCANet, was designed based on the DFCA block and the DS block. DFCA blocks are used to pinpoint disease areas on corn leaves and extract subtle differences to identify and classify different diseases. DS blocks are designed to reduce the loss of disease signature information. The experimental results showed that the accuracy, precision, recall, and F1 score of the proposed model, when used in the classification of corn leaf disease images in real environments, reached 0.9847, 0.9853, 0.9853, and 0.9853, respectively. The accuracy was 1.84%, 5.20%, 1.23%, 4.19%, 2.76%, 3.06%, 2.14%, and 3.67% higher than the VGG16, ResNet50, EffcientNet-B0, Con-vNeXt-B, DenseNet121, MobileNet-V2, Mo-bileNetv3-Large, and ShuffleNetV2-1.0× methods. Furthermore, the Params and Flops of DFCANet were 1.91M and 309.1M, respectively, which are more lightweight than those of other CNN models. We validated the effectiveness of KeepAugment and simulated real-weather data augmentation approaches in crop disease identification. In summary, DFCANet has the advantage of being lightweight and efficient for corn disease identification.

Author Contributions: Conceptualization, X.Z. and Y.C.; methodology, X.Z. and X.C.; software, R.P.; validation, Y.C., X.Z., J.L. and R.P.; formal analysis, J.L. and T.C. (Tengbao Cao); investigation, J.C.; data curation, D.Y. and T.C. (Tengbao Cao); writing—original draft preparation, J.L., R.P. and T.C. (Tomislav Cernava); writing—review and editing, X.Z., T.C. (Tomislav Cernava) and X.C.; visualization, R.P.; supervision, X.Z. and X.C.; project administration, X.Z. and X.C.; funding acquisition, X.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Key Research and Development Plan Key Special Projects, grant number 2021YFE0107700, National Nature Science Foundation of China, grant numbers 61865002 and 31960555, Guizhou Science and Technology Program, grant number 2019-1410, and Outstanding Young Scientist Program of Guizhou Province, grant number KY2021-026. In addition, the study received support by the Program for Introducing Talents to Chinese Universities, 111Program, grant number D20023.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Qiqi, W.; Yinjun, C. Advantages Analysis of Corn Planting in China. *J. Agric. Sci. Technol.* **2018**, *20*, 1.
2. Li, X.; Dong, Y.; Huang, H.; You, H. Effects of fungicides on disease control and yield and quality of silage corn. *Mod. Anim. Husb. Technol.* **2020**, *4*, 7–9.
3. Shiferaw, B.; Prasanna, B.M.; Hellin, J.; Bänziger, M. Crops that feed the world 6. Past successes and future challenges to the role played by corn in global food security. *Food Secur.* **2011**, *3*, 307–327. [CrossRef]
4. Aravind, K.R.; Raja, P.; Mukesh, K.V.; Aniirudh, R.; Ashiwin, R.; Szczepanski, C. Disease Classification in Corn Crop Using Bag of Features and Multiclass Support Vector Machine. In Proceedings of the 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 19–20 January 2018; pp. 1191–1196.
5. Kusumo, B.S.; Heryana, A.; Mahendra, O.; Pardede, H.F. Machine Learning-Based for Automatic Detection of Corn-Plant Diseases Using Image Processing. In Proceedings of the International conference on computer, control, informatics and its applications (IC3INA), IEEE, Tangerang, Indonesia, 1–2 November 2018; pp. 93–97.
6. Ahila Priyadharshini, R.; Arivazhagan, S.; Arun, M.; Mirnalini, A. Maize leaf disease classification using deep convolutional neural networks. *Neural Comput. Appl.* **2019**, *31*, 8887–8895. [CrossRef]
7. Panigrahi, K.P.; Sahoo, A.K.; Das, H. A cnn approach for corn leaves disease detection to support digital agricultural system. In Proceedings of the 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI) (48184), Tirunelveli, India, 15–17 June 2020; pp. 678–683.
8. Mohanty, S.P.; Hughes, D.P.; Salathé, M. Using deep learning for image-based plant disease detection. *Front. Plant Sci.* **2016**, *7*, 1419. [CrossRef]
9. Mishra, S.; Sachan, R.; Rajpal, D. Deep convolutional neural network based detection system for real-time corn plant disease recognition. *Procedia Comput. Sci.* **2020**, *167*, 2003–2010. [CrossRef]
10. Saleem, M.H.; Potgieter, J.; Arif, K.M. Plant disease detection and classification by deep learning. *Plants* **2019**, *8*, 468. [CrossRef]
11. Ferentinos, K.P. Deep learning models for plant disease detection and diagnosis. *Comput. Electron. Agric.* **2018**, *145*, 311–318. [CrossRef]
12. Lv, M.; Zhou, G.; He, M.; Chen, A.; Zhang, W.; Hu, Y. Corn leaf disease identification based on feature enhancement and DMS-robust alexnet. *IEEE Access.* **2020**, *8*, 57952–57966. [CrossRef]
13. Zeng, W.; Li, H.; Hu, G.; Liang, D. Identification of corn leaf diseases by using the SKPSNet-50 convolutional neural network model. *Sustain. Comput. Inform. Syst.* **2020**, *35*, 100695.
14. Pandey, A.; Jain, K. A robust deep attention dense convolutional neural network for plant leaf disease identification and classification from smart phone captured real world images. *Ecol. Inform.* **2022**, *70*, 101725. [CrossRef]
15. Zhu, W.; Sun, J.; Wang, S.; Shen, J.; Yang, K.; Zhou, X. Identifying Field Crop Diseases Using Transformer-Embedded Convolutional Neural Network. *Agriculture* **2022**, *12*, 1083. [CrossRef]
16. Chen, W.; Chen, J.; Duan, R.; Fang, Y.; Ruan, Q.; Zhang, D. MS-DNet: A mobile neural network for plant disease identification. *Comput. Electron. Agric.* **2020**, *199*, 107175. [CrossRef]
17. Chen, J.; Zhang, D.; Zeb, A.; Nanehkaran, Y.A. Identification of rice plant diseases using lightweight attention networks. *Expert Syst. Appl.* **2021**, *169*, 114514. [CrossRef]
18. Yin, C.; Zeng, T.; Zhang, H.; Fu, W.; Wang, L.; Yao, S. Maize small leave spot classification based on improved deep convolutional neural network with multi-scale attention mechanism. *Agronomy* **2022**, *12*, 906. [CrossRef]
19. Zeng, W.; Li, H.; Hu, G.; Liang, D. Lightweight dense-scale network (LDSNet) for corn leaf disease identification. *Comput. Electron. Agri.* **2022**, *197*, 106943. [CrossRef]
20. Lin, J.; Chen, X.; Pan, R.; Cao, T.; Cai, J.; Chen, Y.; Peng, X.; Cernava, T.; Zhang, X. GrapeNet: A Lightweight Convolutional Neural Network Model for Identification of Grape Leaf Diseases. *Agriculture* **2022**, *12*, 887. [CrossRef]
21. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 843–852.
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *60*, 84–90. [CrossRef]

23. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft Coco: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
24. Pan, S.-Q.; Qiao, J.-F.; Wang, R.; Yu, H.-L.; Wang, C.; Taylor, K.; Pan, H.-Y. Intelligent diagnosis of northern corn leaf blight with deep learning model. *J. Integr. Agric.* **2022**, *21*, 1094–1105. [CrossRef]
25. Richey, B.; Shirvaikar, M.V. Deep learning based real-time detection of northern corn leaf blight crop disease using YoloV4. In Proceedings of the Real-Time Image Processing and Deep Learning 2021, Electric Network, 12–16 April 2021; Volume 11736, pp. 39–45.
26. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 139–144.
27. Chen, J.; Wang, W.; Zhang, D.; Zeb, A.; Nanehkaran, Y.A. Attention embedded lightweight network for corn disease recognition. *Plant Pathol.* **2021**, *70*, 630–642. [CrossRef]
28. Albarrak, K.; Gulzar, Y.; Hamid, Y.; Mehmood, A.; Soomro, A.B. A Deep Learning-Based Model for Date Fruit Classification. *Sustainability* **2022**, *14*, 6339. [CrossRef]
29. Gulzar, Y.; Hamid, Y.; Soomro, A.B.; Alwan, A.A.; Journaux, L. A convolution neural network-based seed classification system. *Symmetry* **2020**, *12*, 2018. [CrossRef]
30. Hamid, Y.; Wani, S.; Soomro, A.B.; Alwan, A.A.; Gulzar, Y. Smart seed classification system based on MobileNetV2 architecture. In Proceedings of the 2022 2nd International Conference on Computing and Information Technology (ICCIT), Tabuk, Saudi Arabia, 25–27 January 2022; pp. 217–222.
31. Ahmad, A.; Saraswat, D.; Gamal, A.E.; Johal, G. CD&S Dataset: Handheld Imagery Dataset Acquired Under Field Conditions for Corn Disease Identification and Severity Estimation. *arXiv* **2021**, arXiv:2110.12084.
32. Singh, D.; Jain, N.; Jain, P.; Kayal, P.; Kumawat, S.; Batra, N. PlantDoc: A dataset for visual plant disease detection. In Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, ACM KDD, Hyderabad, India, 5–7 January 2020; pp. 249–253.
33. Gong, C.; Wang, D.; Li, M.; Chandra, V.; Liu, Q. Keepaugment: A simple information-preserving data augmentation approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, 19–25 June 2021; pp. 1055–1064.
34. DeVries, T.; Taylor, G.W. Improved regularization of convolutional neural networks with cutout. *arXiv* **2017**, arXiv:1708.04552.
35. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 13001–13008. [CrossRef]
36. Chen, P.; Liu, S.; Zhao, H.; Jia, J. Gridmask data augmentation. *arXiv* **2020**, arXiv:2001.04086.
37. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, 19–25 June 2021; pp. 13713–13722.
38. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A Convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 21–24 June 2022; pp. 11976–11986.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016; pp. 770–778.
40. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
41. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
42. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
43. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (ICML) PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
44. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
45. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for Mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
46. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. Shufflenet v2: Practical guidelines for efficient Cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
47. Hu, J.; Shen, L.; Sun, G. Squeeze-And-Excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
48. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
49. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-Cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
50. Mohameth, F.; Chen, B.C.; Kane, A.S. Plant disease detection with deep learning and feature extraction using plant village. *J. Computer. Commun.* **2020**, *8*, 10–22. [CrossRef]

51. Huang, J.P.; Chen, J.; Li, K.X.; Li, J.Y.; Liu, H. Identification of multiple plant leaf diseases using neural architecture search. *Trans. Chin. Soc. Agric. Eng.* **2020**, *36*, 166–173.
52. Liu, J.; Wang, X. Plant diseases and pests detection based on deep learning: A review. *Plant Methods* **2021**, *17*, 22. [CrossRef] [PubMed]
53. Subetha, T.; Khilar, R.; Christo, M.S. A comparative analysis on plant pathology classification using deep learning architecture–Resnet and VGG19. *Mater. Today Proc.* 2021, *in press*.
54. Zhao, S.; Peng, Y.; Liu, J.; Wu, S. Tomato Leaf Disease Diagnosis Based on Improved Convolution Neural Network by Attention Module. *Agriculture* **2021**, *11*, 651. [CrossRef]
55. Lin, J.; Chen, Y.; Pan, R.; Cao, T.; Cai, J.; Yu, D.; Chi, X.; Cernava, T.; Zhang, X.; Chen, X. CAMFFNet: A novel convolutional neural network model for tobacco disease image recognition. *Comput. Electron. Agric.* **2022**, *202*, 107390. [CrossRef]
56. Gao, R.; Wang, R.; Feng, L.; Li, Q.; Wu, H. Dual-branch, efficient, channel attention-based crop disease identification. *Comput. Electron. Agric.* **2021**, *190*, 106410. [CrossRef]

agriculture

MDPI

Article

Prediction of Blueberry (*Vaccinium corymbosum* L.) Yield Based on Artificial Intelligence Methods

Gniewko Niedbała [1], Jarosław Kurek [2,*], Bartosz Świderski [2], Tomasz Wojciechowski [1], Izabella Antoniuk [2] and Krzysztof Bobran [3]

[1] Department of Biosystems Engineering, Faculty of Environmental and Mechanical Engineering, Poznań University of Life Sciences, Wojska Polskiego 50, 60-627 Poznań, Poland
[2] Department of Artificial Intelligence, Institute of Information Technology, Warsaw University of Life Sciences, Nowoursynowska 159, 02-776 Warsaw, Poland
[3] Seth Software sp. z o.o., Strefowa 1, 36-060 Głogów Małopolski, Poland
* Correspondence: jaroslaw_kurek@sggw.edu.pl

Abstract: In this paper, we present a high-accuracy model for blueberry yield prediction, trained using structurally innovative data sets. Blueberries are blooming plants, valued for their antioxidant and anti-inflammatory properties. Yield on the plantations depends on several factors, both internal and external. Predicting the accurate amount of harvest is an important aspect in work planning and storage space selection. Machine learning algorithms are commonly used in such prediction tasks, since they are capable of finding correlations between various factors at play. Overall data were collected from years 2016–2021, and included agronomic, climatic and soil data as well satellite-imaging vegetation data. Additionally, growing periods according to BBCH scale and aggregates were taken into account. After extensive data preprocessing and obtaining cumulative features, a total of 11 models were trained and evaluated. Chosen classifiers were selected from state-of-the-art methods in similar applications. To evaluate the results, Mean Absolute Percentage Error was chosen. It is superior to alternatives, since it takes into account absolute values, negating the risk that opposite variables will cancel out, while the final result outlines percentage difference between the actual value and prediction. Regarding the research presented, the best performing solution proved to be Extreme Gradient Boosting algorithm, with MAPE value equal to 12.48%. This result meets the requirements of practical applications, with sufficient accuracy to improve the overall yield management process. Due to the nature of machine learning methodology, the presented solution can be further improved with annually collected data.

Keywords: machine learning; artificial intelligence; yield prediction; blueberry

Citation: Niedbała, G.; Kurek, J.; Świderski, B.; Wojciechowski, T.; Antoniuk, I.; Bobran, K. Prediction of Blueberry (*Vaccinium corymbosum* L.) Yield Based on Artificial Intelligence Methods. *Agriculture* **2022**, *12*, 2089. https://doi.org/10.3390/agriculture12122089

Academic Editor: Yanbo Huang

Received: 27 October 2022
Accepted: 2 December 2022
Published: 6 December 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The blueberry (*Vaccinium corymbosum* L.) is a blooming plant species in the genus Vaccinium within the heather family most commonly found in Eurasia and North America [1]. Blueberries are valued for their antioxidant and anti-inflammatory properties and have neurocognitive benefits [2]. According to FAOSTAT data [3], in 2020, blueberry crop fields in the European Union covered an area of 27,630 hectares, and Poland ranked first with a cultivated area of 9700 hectares and an average yield of 57,010 hg/ha (8th place in the EU).

Increased yield, fruit quality and economic stability in blueberry production requires paying attention to the plantation by identifying key features that affect plant growth and condition throughout the growing season. The yield of blueberry plants depends on several factors, both internal (genetic) and external (growing practices, stimulants, climate) [4]. These factors are most often correlated, but their interaction has not yet been studied in depth.

To predict yields, models are used for estimation during the growing season immediately before harvest [5–8]. The knowledge of the predicted yield for a given year can provide support for decision-making in work planning and the selection of storage space. It can also improve farm profitability and balance the number of inputs used, such as fertilizers, crop protection products or water. Balanced consumption of these products leads to both reduced energy input on the farm and reduced input of human labour. Finally, a plantation can increase profitability due to lower production costs [9,10]. Yield prediction is also used as a tool when theoretical yields need to be estimated in agricultural damage assessments ([11]). It should also be added that in addition to the yield quantity prediction, the yield lost has potential application in practice as shown in research conducted by Khan's team [12], as well as its yield quality, for example, in the form of fruit freshness prediction [13].

In the literature on fruit yield forecasting of orchard crops, some of the most important criteria for dividing yield determination methods include data nature, data type and data source. Thus, the first classification separates them into direct and indirect methods [10]. This classification describes direct methods as yield estimation methods and indirect methods as yield prediction methods, but this is not common nomenclature.

Direct methods are based on data from direct measurements of yield-forming generative organs. These may include measurements of the number of flowers, buds or fruits, their geometric dimensions and/or weight. They are carried out manually or automatically using various types of stationary ground platforms [14–16], mobile ground platforms [17–19] or aerial platforms [20–23].

In the case of indirect methods, yield prediction is performed by developing a predictive model with features indirectly related to yield as inputs. These traits can be divided into several categories. The most important include traits attributed to plants, climate, soil conditions and agrotechnical processes [10]. In the case of data attributed to climate, the most common include meteorological data on historical and current air and soil parameters, such as the amount of natural precipitation and solar activity, i.e., insolation or solar radiation [24]. The data used in yield forecasting, and describing the soil environment, refer both to small time-varying soil parameters, such as the texture of the surface or subsurface layer, and also to medium and short-term variables as soil pH, organic matter (OM) content, salinity (EC), or the content of individual plant nutrients, both macro and micro elements [25,26]. Plant data that are input into prediction models using indirect data most often detail information about the growth status of plants or their organs in successive vegetation phases expressed in the form of vegetation indices, the degree of plant compactness (canopy/biomass), the time and rate of reaching characteristic developmental phases, e.g., flowering, and fruit setting. For the most part, these data come from remote sensing (RS), satellites or UAVs [27–30].

Regardless of their type and nature, the data sources for prediction models can be information from measurements made locally at the site of plant growth and information interpolated from network measurements conducted over larger areas. Local data are increasingly being collected not only from individual measurement stations but also from a grid of sensors mounted on plantations using IoT devices [31]. From the perspective of access to databases for predictive models, it can be noted that the data are private, public and commercial [32]. The abundance of types, natures and sources of data used in predictive models in orchard crops makes it increasingly necessary to employ methods for managing large data sets, Big Data [33,34], in the phase of storing, processing, sharing and analyzing them.

Scientific papers on the yield prediction of blueberries have been published for two decades. However, regading searches performed on the most common literature databases (WoS, SCOPUS) for the phrase "blueberry" and "yield prediction" or "blueberry" and "yield forecast" or "blueberry" and "yield estimation", only 12 and 10 publications from 2002 to 2022, respectively, can be obtained (Figure 1). Thus, it can be concluded that blueberry yield prediction is not the most frequently addressed topic by science, and in

addition, most of the publications are from the United States, Canada, and China, which shows a large knowledge gap for the European region.

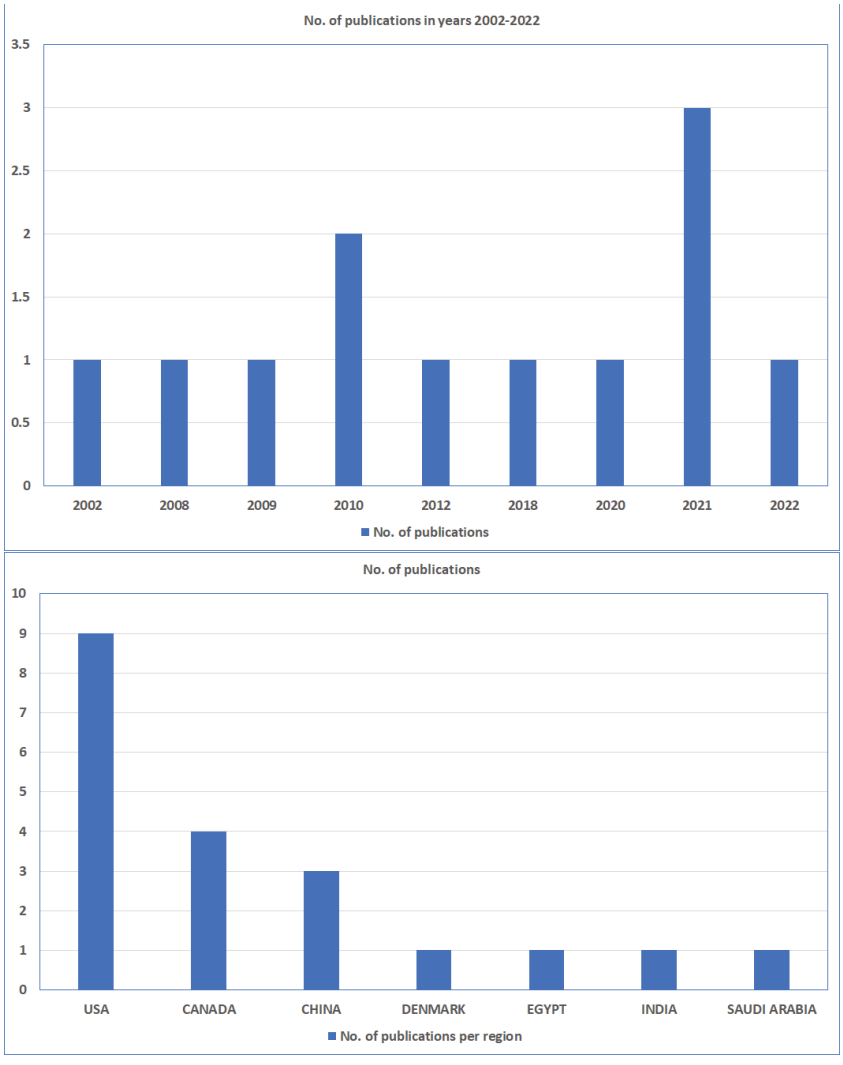


Figure 1. Number of responses as a function of time and region for WOS database search phrases: "blueberry" and "yield prediction" or "blueberry" and "yield forecast" or "blueberry" and "yield estimation" (September 2022).

The available literature on blueberry yield prediction points to several basic topics pursued in this domain. First of all, it should be said that they concern both indirect and direct methods of yield forecasting, although indirect methods are less numerous. In the case of indirect methods, the main data sources for the models are satellite imaging data, ground–based imaging, fruit and leaf spectrometry, and meteorological data [4,11,35–47]. Interestingly, publications using agronomic data from the production process are not found.

Machine learning is the most widely used method for yield modeling due to the highest accuracy of predictions [10,48]. Obsie et al. [11] conducted a study evaluating the importance of bee species composition and weather factors in regulating wild blueberry agroecosystems. This study clarified how bee species composition and weather affect

blueberry yield, and predicted the optimal bee species composition and weather conditions that achieve the best yield. For this purpose, computer simulation and machine learning algorithms were used as predictive tools. These included multiple linear regression (MLR), boosted decision trees (BDT), random forest (RF) and extreme gradient boosting (XGBoost). Seireg et al. [47] employed a set of machine learning techniques to predict wild blueberry yields. For this purpose, they used the method of stack regression (SR) and cascade regression (CR) with a novel combination of machine learning algorithms. The authors used traits that indicated the best regulation for wild blueberry agroecosystems. A total of four feature engineering selection techniques were used, namely inferential variance factor (VIF), sequential forward feature selection (SFFS), sequential backward feature selection (SBEFS) and extreme gradient enhancement based on feature importance (XFI). In this study, Bayesian optimization was applied to popular MLAs to obtain the best hyperparameters for accurate wild blueberry yield prediction. A two-layer structure was used in SR: level-0 containing the light gradient boost machine (LGBM), gradient regression (GBR) and extreme gradient boost (XGBoost), and level-1 providing the output prediction using Ridge. The CR topology is the same MLA used in SR, but in a serial form that takes a new prediction as input into each MLA and removes the previous prediction at each stage. Finally, CR and SR were evaluated with results according to root mean square error (RMSE) and coefficient of determination (R^2). MacEachern et al. [49] conducted research using deep learning convolutional neural networks on fruit maturity stage detection and yield prediction of wild blueberry. In this study, six artificial neural network models based on YOLOv3, YOLOv3-SPP, YOLOv3-Tiny, YOLOv4, YOLOv4-Small and YOLOv4-Tiny architectures were given for analysis. Both 3-class (green berries, red berries, blueberries) and 2-class (immature berries, mature berries) models were developed. The results proved that YOLOv4 performed the best in terms of R^2 accuracy and received the highest F1 score. On the other hand, YOLOv4-Tiny performed the best from the perspective of computational load. Only slight differences were detected in the accuracy of yield prediction models using nonlinear regression, with YOLOv4-Small performing best regarding mean absolute error (RMSE).

The method presented in this paper uses new data sets that are innovative in their structure. To the authors' best knowledge, types of collected data, overall diversity of parameters as well as aggregation of BBCH phases have not been used before in such combination. Taking all of this into account, the data preparation process and use of cumulative features can provide additional insights, resulting in increased model accuracy.

The aim of our research was to develop 11 blueberry yield prediction models using state-of-the-art machine learning algorithms. In addition to selecting the best model for yield prediction, our research had three scientific objectives. First, we analyzed missing and outlier data, and performed normalization of empirical data also used to generate aggregates of predictive data. Second, we performed feature selection, discarding those features that might introduce noise and lead to a deterioration in yield prediction. Third, we determined MAPE prediction errors for each model (algorithm) and identified the best model for yield prediction. Our goal was to determine the most robust model for yield prediction by comparing traditional and state-of-the-art machine learning algorithms while using the minimum number of features to conduct analysis.

2. Materials and Methods

2.1. Dataset Description

The data used in this work came from a marketable blueberry plantation located in southeastern Poland, and the cultivars grown were Chandler, Liberty, Nelson. The data covered six growing seasons in the years spanning from 2016 to 2021 containing several data types, i.e., agronomic data, climatic data, satellite vegetation data and soil data. Data were obtained from both public databases as open data, as well as from private farmer databases and ERP vendor databases. Regardless of the type and source, data were obtained for the subsequent cultivation plots and subplots.

Final structures in the database were prepared and supplemented using obtained dataset. During research, data from cultivation of highbush blueberry (Vaccinium corymbocatfish) were used. The full structure of Highbush blueberry dataset is show in Table 1.

Table 1. Structure of Highbush blueberry data.

Subplot Code	Variety	No of Subplots	Total Area [ha]
102	Nelson	18	26.16
103	Nelson	12	26.4
104	Nelson	12	16.2
105	Nelson	18	26.22
106	Nelson	6	12.48
107	Chandler	18	32.34
108	Chandler	18	29.82
109	Chandler	12	15.9
110	Chandler	12	17.22
111	Chandler	6	13.98
112	Chandler	12	23.46
113	Chandler	12	22.2
114	Liberty	12	17.1
115	Liberty	18	26.82
116	Liberty	12	22.2
117	Liberty	12	23.22
118	Liberty	11	10.07
120	Chandler	5	8.65
121	Chandler	5	8.8
122	Nelson	6	5.82
123	Nelson	6	8.28
	Total	243	393.34

2.1.1. Agronomical Data

Agronomic data was obtained from the Plantator System by Seth Software as well as from Plantator System operating during production in regards to: crop register, harvest registration, registration of hourly work results. The acquired data holds different formats, so in some cases it was necessary to obtain pre–processed dataset, e.g., from soil test results (pdf) or locations of crops (jpg). Agronomic data also comes from private grower databases.

2.1.2. BBCH-Scale

The BBCH-scale is used to identify the phenological development stages of plants. BBCH-scales have been developed for a range of crop species where similar growth stages of each plant are given the same code.

Phenological development stages of plants are used in a number of scientific disciplines (crop physiology, phytopathology, entomology and plant breeding) and in the agriculture industry (risk assessment of pesticides, timing of pesticide application, fertilization, agricultural insurance). The BBCH-scale uses a decimal code system, which is divided into principal and secondary growth stages, and is based on the cereal code system (Zadoks scale) developed by Jan Zadoks.

The abbreviation BBCH derives from the names of the originally participating stakeholders: "Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie". Allegedly, the abbreviation is said to unofficially represent the four companies that initially sponsored its development; Bayer, BASF, Ciba-Geigy, and Hoechst. The phenological development stages obtained from the producer were in following ranges:

(1) BBCH 0–60 (from dormancy to the beginning of flowering)
(2) BBCH 61–70 (from the beginning of flowering to the end of flowering
(3) BBCH > 70 (from the beginning of fruit growth to harvest)

Based on the imported data, the BBCH phase limits in the ranges: 0–60, 61–70, >70 were assigned. They will be later used to calculate aggregated data.

2.1.3. Numerical Features

A total of 89 potential explanatory variables were derived for the target. Target variable was defined as the total yield of harvested crop (harvest) for the blueberries. All numerical data (both explanatory and dependent variables) were aggregated to full year (2016, 2017, 2018, 2019, 2020 or 2021). The target variable was measured in kilograms [kg].

Since after initial tests, the climatic data delivered by the producer were not sufficient, additional information was obtained using the public data available from Institute of Meteorology and Water Management-National Research Institute (IMGW) weather stations. This additional data was stored using following labels:

(1) Air temperature (avg.) [°C]
(2) Air temperature (max.) [°C]
(3) Air temperature (min.) [°C]
(4) Rainfall [mm]
(5) Air relative humidity (avg.) [%]
(6) Air relative humidity (max.) [%]
(7) Air relative humidity (min.) [%]
(8) Dew point temperature (avg.) [°C]
(9) Dew point temperature (max.) [°C]
(10) Dew point temperature (min.) [°C]

In relation to total harvest obtained, two numerical features regarding treatment of the blueberry were stored (treatment features):

(1) Irrigation [kg]
(2) Fertigation [l]

Depending on the season, data regarding nutrient content of fertigation media and soil content were obtained from different certified soil testing laboratories, resulting in variability of analyzed components and different units of measurement. A total of 14 numerical features regarding soil parameters of blueberry has been stored/extracted. The final list for fertigation media is given as follows:

(1) pH
(2) S-SO_4—sulfur [mg/L]
(3) P—phosphorus [mg/L]
(4) K—potassium [mg/L]
(5) C—calcium [mg/L]
(6) Mg—magnesium [mg/L]
(7) Fe—iron [mg/L]
(8) Zn—zinc [mg/L]
(9) Mn—manganese [mg/L]
(10) Cu—copper [mg/L]
(11) B—boron [mg/L]
(12) Cl—chlorine [mg/L]
(13) Na—sodium [mg/L]
(14) N-NO_3—nitrogen [mg/L]

2.1.4. BBCH Soil and Climate Features

Using the BBCH scale, additional potential predictive features in the form of aggregates were determined. After calculating those sets, total of 30 numerical features regarding soil and climate parameters based on BBCH scale of blueberry has been extracted.

(1) Insolation (BBCH 0–60) [W/m^2]
(2) Insolation (BBCH 61–70) [W/m^2]

(3) Insolation (BBCH > 70) [W/m^2]
(4) Rainfall (BBCH 0–60) [mm]
(5) Rainfall (BBCH 61–70) [mm]
(6) Rainfall (BBCH > 70) [mm]
(7) Irrigation (BBCH 0–60) [mm]
(8) Irrigation (BBCH 61–70) [mm]
(9) Irrigation (BBCH > 70) [mm]
(10) Daily air temperature (avg.) (BBCH 0–60) [°C]
(11) Daily air temperature (avg.) (BBCH 61–70) [°C]
(12) Daily air temperature (avg.) (BBCH > 70) [°C]
(13) Daily soil temperature (avg.) (BBCH 0–60) [°C]
(14) Daily soil temperature (avg.) (BBCH 61–70) [°C]
(15) Daily soil temperature (avg.) (BBCH > 70) [°C]
(16) Soil pH (avg.) (BBCH 0–60)
(17) Soil pH (avg.) (BBCH 61–70)
(18) Soil pH (avg.) (BBCH > 70)
(19) Soil humidity (avg.) (BBCH 0–60) [%]
(20) Soil humidity (avg.) (BBCH 61–70) [%]
(21) Soil humidity (avg.) (BBCH > 70) [%]
(22) Soil P—phosphorus (avg.) (BBCH 0–60) [mg/L]
(23) Soil P—phosphorus (avg.) (BBCH 61–70) [mg/L]
(24) Soil P—phosphorus (avg.) (BBCH > 70) [mg/L]
(25) Soil Mg—magnesium (avg.) (BBCH 0–60) [mg/L]
(26) Soil Mg—magnesium (avg.) (BBCH 61–70) [mg/L]
(27) Soil Mg—magnesium (avg.) (BBCH > 70) [mg/L]
(28) Soil K—potassium (avg.) (BBCH 0–60) [mg/L]
(29) Soil K—potassium (avg.) (BBCH 61–70) [mg/L]
(30) Soil K—potassium (avg.) (BBCH > 70) [mg/L]

2.1.5. Vegetation Features

Crop vegetation status data were derived from satellite remote observations. The primary image database used in the imaging was the European Copernicus Sentinel 2 mission's image database, and the Google Earth Engine (GEE) platform was used as the image processing and vegetation index (VI) calculation tool. Those VI were:

(1) EVI—Enhanced Vegetation Index
(2) NDVI—Normalized Difference Vegetation Index
(3) RDVI—Renormalized Difference Vegetation Index
(4) SAVI—Soil-Adjusted Vegetation Index

The above-mentioned VIs are among the most widely used indices in the literature in yield predictions for orchard crops, and were calculated according to Index DataBase [50].

Finally, 16 vegetation features for 4 VIs were calculated, with final division containing Min, Mean, Max, Standard deviation groups. The full list contains the following parameters:

(1) EVI 40 days before harvest (max.)
(2) EVI 40 days before harvest (avg.)
(3) EVI 40 days before harvest (min.)
(4) EVI 40 days before harvest (stddev.)
(5) NDVI 40 days before harvest (max.)
(6) NDVI 40 days before harvest (avg.)
(7) NDVI 40 days before harvest (min.)
(8) NDVI 40 days before harvest (stddev.)
(9) RDVI 40 days before harvest (max.)
(10) RDVI 40 days before harvest (avg.)
(11) RDVI 40 days before harvest (min.)

(12) RDVI 40 days before harvest (stddev.)
(13) SAVI 40 days before harvest (max.)
(14) SAVI 40 days before harvest (avg.)
(15) SAVI 40 days before harvest (min.)
(16) SAVI 40 days before harvest (stddev.)

2.1.6. Selyaninov Hydrothermal Coefficient

The study of climate variability is a topic of interest for many scientists across many diverse fields including hydrologists, meteorologists, farmers, and foresters. All would like to determine as precisely as possible what climatic conditions will prevail in a given area in the future [51–59]. Greater computing power allows us to analyse increasingly complex models while still showing that there are more factors affecting the environment. Therefore, the problem remains unresolved.

In accordance with various scenarios of climate change for Central Europe, the temperature increase will be accompanied by a very small increase in annual rainfall which will be redistributed over the year; an increase in the winter will be accompanied by a decline in the summer [60–63]. In this situation with poor retention possibilities and a simultaneous increase in evaporation, it should be expected that the amount of water that is useful for plants will be reduced during the growing season, with a possible depletion of post-winter stocks. As the authors of the works [64,65] indicate, one must not neglect the effect of the growing variance of precipitation and temperatures, which means that extreme situations that are unfavourable for plant production will occur more frequently.

One element to observe is the assessment of the amount of water present in a given area, especially in extreme values, i.e., floods and droughts. Different types of indicators are used to assess the severity of water shortages. One of these is the Selyaninov hydrothermal coefficient (HTC) [66]. The pattern assesses drought according to the following formula:

$$HTC = \frac{10 \sum_i^n P_i}{\sum_i^n t_i} \quad (1)$$

where:
n—length of the period considered in days,
P_i—rainfall on the i-th day [mm],
t_i—average daily temperature on the i-th day [°C].

Based on the above properties, three aggregated parameters were generated to be used as additional prediction features:

(1) HTC (BBCH 0–60)
(2) HTC (BBCH 61–70)
(3) HTC (BBCH > 70)

2.1.7. GDD Features

In the absence of extreme conditions such as nonseasonal drought or disease, plants grow in a cumulative step-wise manner that is strongly influenced by the ambient temperature. Growing degree days (GDD) takes aspects of local weather into account and allow gardeners to predict (or, in greenhouses, even to control) the plants' pace toward maturity.

Unless stressed by other environmental factors such as soil moisture, the development rate from emergence to maturity for many plants depends upon the daily air temperature. Because many developmental events of plants and insects depend on the accumulation of specific quantities of heat, it is possible to predict when these events should occur during a growing season regardless of differences in temperatures from year to year. Growing degrees days (GDD) is defined as the number of temperature degrees above a certain threshold base temperature, which varies among crop species [67]. The base temperature is that temperature below which plant growth is zero. GDD are calculated each day as maximum temperature plus the minimum temperature divided by 2, minus

the base temperature. GDD are accumulated by adding each day's GDD contribution as the season progresses.

GDD can be used to: assess the suitability of a region for production of a particular crop; estimate the growth-stages of crops, weeds or even life stages of insects; predict maturity and cutting dates of forage crops; predict best timing of fertilizer or pesticide application; estimate the heat stress on crops; plan spacing of planting dates to produce separate harvest dates.

$$GDD = \sum_{i=1}^{n} T_{avg} \quad (2)$$

where:
GDD—Growing Degree Days [°C]
n—length of the period considered in days,
T_{avg}—average daily air temperature ≥ 0 [°C]

Similarly as with HTC, three aggregated prediction features were generated based on this parameter:

(1) GDD (BBCH 0–60)
(2) GDD (BBCH 61–70)
(3) GDD (BBCH > 70)

2.1.8. Aggregates Based on Mineral Fertilization and Fertigation

Depending on the amount of fertilization, structure of used fertilizers and fertigation media, the results can vary. A total of nine parameters based on these features were generated:

(1) Fertilization (BBCH 0–60) [kg]
(2) Fertilization (BBCH 61–70) [kg]
(3) Fertilization (BBCH > 70) [kg]
(4) Fertigation (BBCH 0–60) [l]
(5) Fertigation (BBCH 61–70) [l]
(6) Fertigation (BBCH > 70) [l]
(7) K—potassium-Fertilization (annually) [kg]
(8) N—nitrogen-Fertilization (annually) [kg]
(9) P—phosphorus-Fertilization (annually) [kg]

2.1.9. Harmful Features

Apart from features positively influencing final harvest, two additional ones were derived. Those features concerned randomly occurring harmful conditions that had a significant impact on the obtained crop:

(1) hailstorm percentage of damage [%]
(2) hailstorm cut fruit [%]

2.1.10. Features Summary

A total of 89 potential predictive features were obtained during initial data processing. A summary of all raw data groups is presented in Tables 2 and 3 .

Table 2. Summary of raw data.

Features Group	No. of Raw Data
Treatment features	135,113
Weather features	831,562
Soil features	6929
BBCH soil features	7380
Vegetation features	3936
Selyaninov hydrothermal coefficient	738
GDD features	738
Aggregates based on fertilization and fertigation	9045
Harmful features	110
Total	995,551

Table 3. Summary of all prognostic features

Features Group	No. of Features
Treatment features	2
Weather features	10
Soil features	14
BBCH soil features	30
Vegetation features	16
Sjeljaninow features	3
GDD features	3
Aggregates based on fertilization and fertigation	9
Harmful features	2
Total	89

2.2. Data Preprocessing Methods

Since obtained data were not represented in an easy to use and coherent form, it could not be incorporated in prepared algorithms in original form. In order to prepare the dataset for further use, series of operations were performed in order to adjust it to the requirements of used methodologies. Those operations included data normalization and establishing methods for dealing with missing values.

2.2.1. Data Normalization

Since the original dataset was heterogeneous in nature, before the learning process, the data needed to be normalized.

For the prognostic features in regards to the areadelcared variable, following features needed to be rescaled to 1ha values:

(1) Crop/Harvest
(2) Irrigation
(3) Fertigation

Furthermore, since most AI algorithms work best when specific values are placed in <0, 1> range, all prognostic features were normalized to fit into this range, using following equation:

$$z_i = \frac{(x_i - min(x))}{max(x) - min(x)} \quad (3)$$

where:
z_i: ith normalized value in the feature vector
x_i: ith value in the feature vector
$min(x)$: minimum value in the feature vector
$max(x)$: maximum value in the feature vector

Additionally, the explained variable (crop/harvest) was rescaled using harmful conditions, lowering the values of the crop/harvest variable by % of the occurrence of harmful conditions (if any) such as:

(1) hailstorm percentage of damage [%]
(2) hailstorm cut fruit [%]

2.2.2. Finding and Replacing Missing Values

One of the mandatory analysis procedures in large datasets is finding and handling missing data. It is necessary to search for missing values of each variable. When they occur, such data record is either discarded or the missing value is replaced by a different one. The main problem here is determining the best way to assign missing values in the second case. There are some well-known statistical methods widely used to fill in non-matching data:

(1) Imputation Using (Mean/Median) Values
(2) Imputation Using (Most Frequent) or (Zero/Constant) Values
(3) Stochastic Regression Imputation
(4) Extrapolation and Interpolation
(5) Imputation Using k-NN
(6) Imputation Using XGBoost
(7) Others

Currently, the most effective method of replacing missing data is the prediction for all missing values by XGBoost (Extreme Gradient Boosting) method. The full list of features with missing values in current dataset is given as follows:

(1) S-SO_4—sulfur
(2) Cl—chlorine
(3) Fertigation (BBCH 0–60)
(4) Fertilization (BBCH 0–60)
(5) Hailstorm percentage of damage
(6) Hailstorm cut fruit

A full summary of missing data percentages for each feature is presented in Table 4. In the presented approach, the XGBoost method was used for four of the first features [68]. In case of the remaining two features, due to their different nature, it was determined that the best approach would be replacing the missing values with zero.

Table 4. List of features with missing values.

No.	Feature with Missing Values	% of Missing Values
1	S-SO_4—sulfur	17%
2	Cl—chlorine	33%
3	Irrigation (BBCH 0–60)	46%
4	Fertilization (BBCH 0–60)	47%
5	Hailstorm percentage of damage	67%
6	Hailstorm cut fruit	87%

Extreme Gradient Boosting-XGBoost

XGBoost (Extreme Gradient Boosting) is a model that was first proposed by Tianqi Chen and Carlos Guestrin in 2011 and has been continuously optimized and improved in follow-up studies performed by different scientists (Chen and Guestrin, 2016). The model is a learning framework based on Boosting Tree models.

The traditional Boosting Tree model uses only the first derivative information. When training the nth tree, it is difficult to implement distributed training because the residual of the former n-1 trees is used. XGBoost performs a second-order Taylor expansion on the loss function and it can automatically use the multithreading of the CPU for parallel computing. Additionally, XGBoost uses a variety of methods to avoid overfitting. This approach is

much more precise than simple statistical methods such as using mean, median or modal value to replace missing values. The XGBoost algorithm used in this paper is outlined in Algorithm 1 (Chen and Guestrin, 2016).

Algorithm 1: Algorithm of missing value imputation using XGBoost

procedure FILLMISSINGBYXGBOOST($AllFeatures, feature1$)
$fields \leftarrow AllFeatures - feature1$
$X \leftarrow fields$
$Y \leftarrow feature1$
$indicator \leftarrow isnotnan(Y)$
$X_{train} \leftarrow X[indicator]$
$y_{train} \leftarrow y[indicator]$
$xgbr \leftarrow xgb.XGBRegressor()$
$xgbr.fit(X_{train}, y_{train})$
$y_{est} \leftarrow xgbr.predict(X)$
$y_{est}[indicator] \leftarrow y[indicator]$ ▷ do not touch real value of features
$AllFeatures[feature1] \leftarrow y_{est}$
return $AllFeatures$
end procedure

2.3. Feature Generation Using PCA (Principal Component Analysis) Method

As for Pearson and Chi-square, for the set of all features and for features distinguished by the stepwise fit method, it was checked whether by generating artificial variables (reduction of multidimensionality by the PCA method) the effectiveness of the model would be increased. In that case, PCA was generated for 3, 4, 5, 6, 7, 8, 9 and 10 principal components. In result, the models will have 3 to 10 variables (artificial variables). Each variable is an artificial trait that is a linear combination of the already existing traits.

2.4. Outlier Detection

Outlier observation is an observation of relatively distant values from other elements of the sample [51]. In other words, having an atypical value of the independent (explaining) variable or atypical values of both variables—dependent (explained) and explaining (explaining in multiple regression analysis). This means that the relationship between Xi and Yi for an observation may be different than for the rest of the observations in the dataset.

Outliers may reflect the actual distribution or be the result of chance, but may also indicate a measurement error or a mistake in entering information into the database, etc. A large number of outliers may also be a signal informing that the wrong model was chosen.

Outliers resulting from data errors make the analysis difficult and, in extreme cases, impossible. Methods and coefficients based on the assumption of normal distribution and linear dependencies, such as Pearson's correlation, linear regression, classical correspondence analysis, etc., are particularly sensitive to them. A single outlier can completely change the value and sign of the correlation, even from 0.99 to -0.99

It is therefore necessary to either remove outliers or use robust statistical methods, e.g., rank methods. In our paper, we have decided to eliminate records that are stated as outliers. A comparison of two different methods applied to this problem is shown in Figure 2.

2.4.1. Local Outlier Factor (LOF)

Local Outlier Factor (LOF) is an algorithm that computes a score reflecting the degree of abnormality of the observations. In order to find samples that have significantly lower density than their neighbours, local density deviations are measured for a given point of data.

K-nearest neighbours are used to obtain local density, where LOF score of single point is equal to the ratio of average local density between set of k-neighbouring points and the

initial point. In the case of a normal data point, this density should be similar, while any abnormal data will have smaller local density.

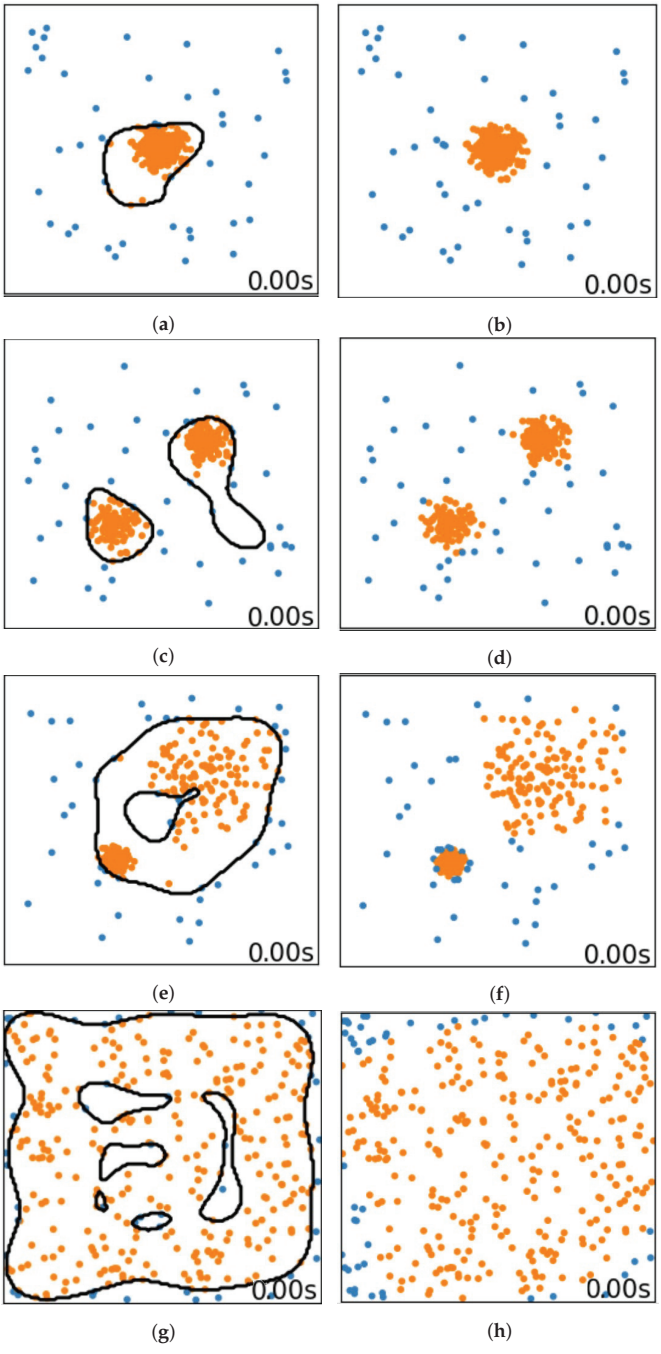


Figure 2. Comparison between two methods of outlier detection. The images used were first presented in [69]. (**a**) SVM one class-case #1; (**b**) LOF-case #1; (**c**) SVM one class-case #2; (**d**) LOF-case #2; (**e**) SVM one class-case #3; (**f**) LOF-case #3; (**g**) SVM one class-case #4; (**h**) LOF-case #4.

One difficulty here is the appropriate choice of k parameter—the number of neighbours to consider. This parameter would usually be either higher than the minimum number of objects a cluster has to contain (in that case objects can be local outliers to that cluster), or smaller than the maximum number of close objects identified as potential outliers. This information usually is not readily available, and might require some testing. In general 20 considered points would work in most cases, while sets with higher outlier numbers (i.e., more than 10%) might require higher values.

Main advantage of this method is that it takes into consideration both local and global dataset properties. It can perform well in datasets, where abnormal samples differ in underlying densities. Instead of simply checking how isolated the sample is, this method will also consider how this sample relates to surrounding samples.

2.4.2. Unsupervised Outlier Detection Based on OneClassSVM

The One-Class SVM was introduced by Schölkopf et al. for that purpose and implemented in the Support Vector Machines module in the svm.OneClassSVM object. It requires the choice of a kernel and a scalar parameter to define a frontier. The RBF kernel is usually chosen although there exists no exact formula or algorithm to set its bandwidth parameter. This is the default in the scikit-learn implementation. The nu parameter, also known as the margin of the One-Class SVM, corresponds to the probability of finding a new, but regular, observation outside the frontier.

All the above methods were used in order to generate best possible feature set. During the research process, the influence of each method was checked, while the final solution will use the best achieved combination.

2.5. Feature Selection

After initial dataset preparation, the next crucial step from the machine learning algorithms point of view was features selection. Finding and using the best possible features that carry the most information can significantly improve overall algorithm performance.

2.5.1. Stepwise Regression

Stepwise regression is a method of adding and removing features from a multi-line model based on their statistical significance. This method begins with the initial model and then takes the steps to modify the model by adding or removing features. At each step, the *p*-value F-statistic is computed to test models with and without the potential feature. If the feature is not currently in the model, the null hypothesis is that the feature would have a factor of zero if it were added to the model. If there is sufficient evidence to reject the null hypothesis, the feature is added to the model. Conversely, if the feature is currently in the model, the null hypothesis is that the feature has a factor of zero. If there is insufficient evidence to reject the null hypothesis, the feature is removed from the model. The method works as follows:

(1) Fit the initial model.
(2) If any features not in the model have a *p*-value less than the input tolerance (e.g., 0.05), add the one with the smallest *p*-value and repeat this step. For example, suppose the initial model is the default model and the input tolerance = 0.05. The algorithm first fits all models consisting of the constant plus the first feature and looks for the next feature that has the smallest p-value, for example feature 4. If feature 4's *p*-value is less than 0.05 then feature 4 is added to the model. Then the algorithm searches among all models consisting of the constant + feature 4 and looks at the next features. If the trait not in the model has a *p*-value less than 0.05, the trait with the smallest *p*-value is added to the model and the process is repeated. When there are no further features to add to the model, the algorithm moves to step 3.
(3) If any features in the model have a *p*-value greater than the output tolerance -premove (e.g., 0.06), remove those with the largest *p*-value and go to step 2; otherwise the algorithm will finish computations and return the resulting feature list.

In each step of the stepwisefit algorithm it uses the least squares method to estimate the model coefficients. After a feature has been added to the model at an earlier stage, the algorithm may later remove it if it is no longer useful in conjunction with later added elements. The method ends when no single step improves the model. However, there is no guarantee that the final model will be optimal (best fit to the data). A different starting model or a different sequence of steps may lead to a better fit. In this sense, step models are locally optimal, but not always best globally.

Total of 90 different values for input and output tolerance were tested (denoted as penter and premove accordingly). In the final set only unique features will be taken into account.

2.5.2. Pearson's Feature Selection Method

In case of large feature sets, there is high possibility that some of them will be correlated. In order to avoid using such elements, the Pearson correlation method was used both for the complete set of all features and for those distinguished by the stepwisefit method. In presented approach, if the Pearson correlation was greater than 0.95, the variable was removed from the set of prognostic features.

2.5.3. Chi-square Feature Selection Method

As for the Pearson correlation approach, the Chi-square statistic was used for the p-value. For the initial set of features, we the variable will be added to the prognostic features if the Chi-squared value for it is greater than 0.05, otherwise the variable is blocked and will not be included in the final set.

2.6. Prediction Methods Applied

In machine learning, there are different prediction methods available. Depending on the specific case, type and amount of used data, as well as various other factors, each method can result in different level of adjustment to the given problem. Therefore, during the research presented in this paper, various approaches were tested in order to extract the best methodology. A summary of the applied classifiers is presented in Table 5.

2.6.1. Linear Regression

In the case of statistical approaches, the linear regression is an approach used for modelling relationship between the result and one or more explanatory variables. In this case ordinary, least-squares linear regression was used. In general, this model fits a linear model with coefficients w = (w1, . . . , wp). The goal here is to minimize the residual sum of squares between predicted target, and an actual target as can be observed in the original dataset [70].

The method used returns the coefficient of determination of the final prediction, defined for R^2 as $(1 - \frac{u}{v})$, where u is the residual sum of squares, with total sum of squares denoted as v. In the worst case, when the input features are disregarded during prediction, the R^2 score would equal to 0.0. In the best case scenario, this score can equal 1.0, or -1.0 because the model can be arbitrarily worse.

In general linear regression obtains good results for problems with goals such as forecasting, prediction, error reduction or explaining variations in goal variables. Because of that, and due to general simplicity it was chosen as one of methods tested in presented approach.

2.6.2. Ridge

Ridge regression can be used to estimate coefficients for multiple-regression models, assuming that various independent variables available in dataset are highly correlated. Since mean square estimators tend to be smaller than the least squares estimators, this provided more precise estimate of ridge parameters. Ridge regression addresses some of the problems of Ordinary Least Squares by imposing a penalty on the size of the coefficients

with l2 regularization. If the l2-norm is given as a regularization, and linear least-squares method is used for loss function, this approach can solve the regression model [71].

2.6.3. Lasso

The Lasso (least absolute shrinkage and selection operator) is a linear model that estimates sparse coefficients with l1 regularization. In general this method amounts to linear regression analysis, performing both variable selection and regularization to enhance obtained results in terms of both prediction accuracy and interpretability [72].

2.6.4. ElasticNet

Elastic-Net is a linear regression model trained with both l1 and l2-norm regularization of the coefficients. In general this model combines linearly the penalties form lasso and ridge methods. Both of those are treated as a special cases in this approach.

Such approach is used to increase overall accuracy. For example, in case of datasets with high-dimensional (p) data with few examples (n), the Lasso method will select maximum of n variables before saturation. Furthermore, if groups of highly correlated variables exist, it tends to select only one variable from such group. In order to overcome this limitations, the elastic net method uses the quadratic part in the penalty (which used alone amounts to ridge regression) [73].

2.6.5. Random Forest Regressor

In general, the random forests or random decision forest is a type of ensemble learning method, that constructs multiple decision trees during training time. It is commonly applied in classification and regression problems. In case of classification, the class selected by most of the trees would be used as an output. In the regression approaches the average prediction value for used trees would be returned. This approach corrects the tendency of decision trees for overfitting to the used training set, and in general it will outperform the decision trees. It is also dependant on data characteristic, which can influence the performance of this methodology.

A random forest regressor is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max_samples parameter if bootstrap = True (default), otherwise the whole dataset is used to build each tree [74].

2.6.6. MLP Regressor

MLP or Multi-layer Perceptron regressor is a model which optimizes the squared error with either LBFGS or stochastic gradient descent method. It is a fully connected feed-forward artificial neural network. In general it will have at least three layers of nodes: input, at least one hidden layer and output. All nodes are neurons using a nonlinear activation function, except from input nodes providing the initial data. Back-propagation is used for training. Main advantage here is that this model can distinguish data that cannot be separated linearly.

Two versions of this model were used in current approach, using 10 and 100 hidden layers, respectively, [75].

2.6.7. SGD Regressor

SGD (Stochastic Gradient Descent) is an iterative method that can be used for stochastic approximation of gradient descent optimization. In this case the actual gradient calculated from the dataset is replaced with its estimate obtained from randomly selected data subset. It is mainly used to reduce computation, especially in high-dimensional optimization problems.

In case of method used in this approach, the linear model was fitted with regularized empirical loss with SGD. For each sample, one at a time the gradient of the loss is estimated,

updating the model with a decreasing learning rate. Penalty added to the loss function is treated as the regularizer, shrinking model parameters towards zero vector. Squared Euclidean norm L2 or the absolute norm L1 or a combination of both (Elastic Net) is used. If because of the regularizer the parameter update crosses 0.0, it will be truncated to this value in order to allow the possibility of learning sparse models, as well as achieving online feature selection [76].

2.6.8. SVR and NuSVR

SVR (Epsilon-Support Vector Regression) is an application of Support Vector Machine to the regression problem. In general, this approach gives the user high flexibility in defining how much error can be accepted in given model. The SVM will then find appropriate line or hyperplane fitting the given set of data.

Objective function of SVR is to minimize the coefficients–L2 norm of the coefficient vector. Second parameter used to define the SVR is the epsilon, or maximum acceptable error. Absolute error will need to be kept less than or equal to this specified margin, and the epsilon parameter can be tuned in order to obtain desired model accuracy.

Sometimes due to the data characteristics, not all points will fall inside specified margin. In such cases additional parameters are needed to account for the possibility of errors that are larger than epsilon. This can be done with slack variables. Those variables are used to denote deviation from the margin, for any value that falls outside of specified epsilon. The C hyperparameter can be additionally tuned, with lower values indicating that the tolerance for values outside epsilon also decreases. The ideal situation in that case would result in simplified version, with no variables falling into that category and C parameter equal to 0.0. One important aspect in this approach is finding the optimal value of C hyperparameter (as low as possible, but at the same time ensuring that all points will be addressed) [77].

Additionally, a modification of this approach was used. NuSVR (Nu Support Vector Regression) is an algorithm used for solving regression problems, applying nu parameter by replacing the epsilon parameter of SVR [78].

Table 5. Summary of applied classifiers.

No.	Classifier
1	Linear regression
2	Ridge
3	Lasso
4	ElasticNet
5	XGB (learning_rate = 0.1, n_estimators = 1000, max_depth = 6)
6	Random Forest (max_depth = 3, n_estimators = 300)
7	MLP (hidden_layer_sizes = 10)
8	MLP (hidden_layer_sizes = 100)
9	SGD
10	NuSVR (nu = 0.2, C = 0.2,kernel = 'rbf', gamma = 0.001)
11	SVR (C = 30,000.0, epsilon = 0.2)

3. Results and Discussion

The dataset was divided into three parts. Sets containing data from years: 2016, 2017, 2018, were used as a training data. Sets containing information from years 2019 and 2020 were used as a validation set. Finally, data from the last year (2021) were used as a test set for all methods. General outline of the numerical experiments performed in this paper are presented in Algorithm 2.

Table 6 presents the best results for each of the tested classifiers. It turns out that each of the best results used stepwise regression. the table shows the penter and premove values that produced the best values in the set. As can be seen, in some cases the reduction of the Pearson and Chi-squared correlation was applied. In many cases, in the best one, the

reduction of multidimensionality with the PCA method was used. For the best result, nine artificial variables were found, obtained from the PCA algorithm. The overall best result was achieved for XGB algorithm and with MAPE equal to 12.48%.

Based on the research presented in this paper, models for blueberry yield prediction have been produced, and the best of these models will be implemented in a dedicated information system. The described research is funded within the framework of a R&D project [79], whose final product will be marketable products in the form of a blueberry yield prediction service.

In this paper, we describe the results of the developed models for data that a farm producer or entrepreneur can obtain simply from their own databases or from public data. The perspective for the development of prediction models is the use of further datasets extended by new types of data (e.g.: terrestrial phenological imaging, soil data from mobile proximal sensors), but also the same type of data but with better quality (e.g.: high- or very-high-resolution satellite imaging data). In addition, the good results of blueberry yield forecasting models at the species level obtained so far are a good prospect for future research on the development of yield prediction models for selected blueberry varieties.

Table 6. Top results of blueberry crop prediction.

Classifier	No. of Cols	Step Wise	P-Enter	P-Remove	Pearson	Chi2	PCA	PCA Comp.	MAPE Val. [%]	MAPE Test [%]
XGB	40	Yes	0.33	0.38	No	No	Yes	8	10.33	12.48
Random Forest	39	Yes	0.24	0.29	No	No	Yes	9	10.20	14.30
Linear Regression	48	Yes	0.44	0.49	No	No	Yes	9	5.70	15.90
SVR	37	Yes	0.20	0.25	No	No	Yes	6	15.09	15.96
Lasso	48	Yes	0.44	0.49	No	No	Yes	8	1.49	17.68
SGD	48	Yes	0.44	0.49	No	No	Yes	9	1.22	17.94
Ridge	48	Yes	0.44	0.49	No	No	Yes	9	0.49	18.64
ElasticNet	37	Yes	0.20	0.25	No	No	Yes	5	13.73	21.84
NuSVR	37	Yes	0.20	0.25	No	No	Yes	4	31.81	34.91
MLP(100)	89	No	0	0	No	No	No	0	97.48	98.25
MLP(10)	46	Yes	0.43	0.48	No	No	No	0	99.70	99.76

Figure 3 shows the effect of each PCA component for the XGBoost model (artificial variable) in explaining the variance of the original variables. Thanks to this, it is possible to determine how many components should be selected in order to achieve a satisfactory percentage of information from the original variables (40 variables after the application of the feature selection and before the application of the PCA transformation). For example, the first component F0 already contains 68.69% of information from the original variables, the second component F1 over 10.69%, the third component F2 6.91%, the fourth F3 3.21%, the fifth F4 2.46%, etc. The overall feature importance, in a descending order is presented at Figure 4.

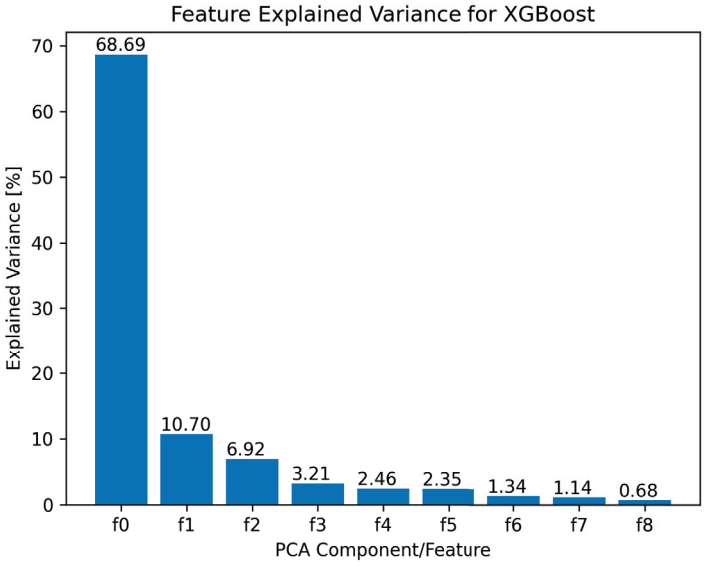


Figure 3. Feature explained variance for XGBoost.

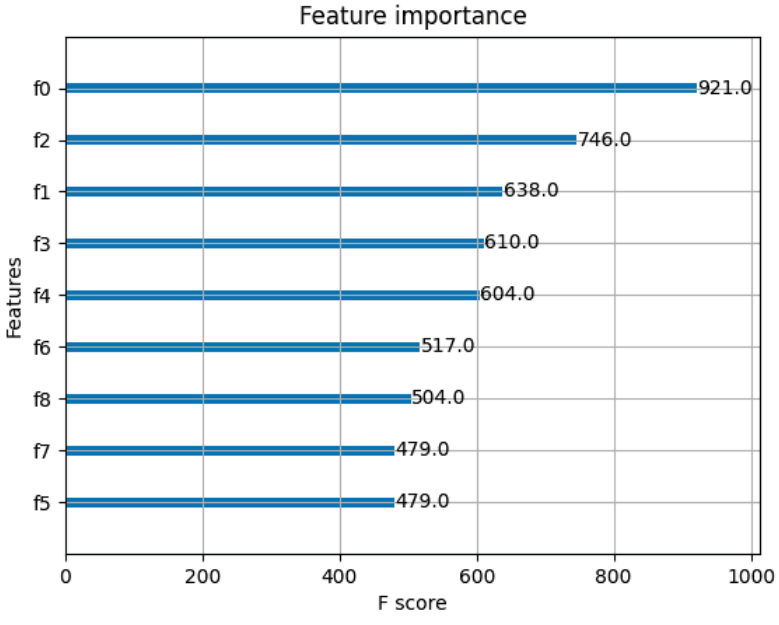


Figure 4. Feature importance of artificial features generated based on PCA (Principal component Analysis).

On the other hand, better approach to selecting the number of PCA components to best estimate how much information process will extract from the original variables is the so-called cumulative percentage of explainable variance. On the graph presented at Figure 5 it can be seen that the first two components explain the original variables to the degree of 79.39%. First three components achieve 86.30%, and, respectively, 89.52% and

91.98 for the first 4 and 5 components. On the other hand, the first 9 PCA components explain as much as 97.49% of the information from the set of 40 original features.

Figure 6 shows the error results for the 11 tested algorithms. The outcomes are presented both in the validation part and the test dataset. The first part of the experiment is used in the training process. After each trained epoch, the result obtained in validation stage are tested to avoid overfitting problems. For this reason, the validation part of the experiment is involved only during the training process and cannot be used as a criterion for selecting the model. On the other hand, the testing part of the experiment is independent of training stage. Based on results obtained during this process, the most effective model should be selected. For this reason, the models were sorted in the increasing order after the MAPE error of the test part is calculated.

The results of the research presented in this article indicate that modelling the yield of blueberries during the growing season is reasonable and brings promising application possibilities.

Predictive models are usually created on the basis of results collected during long-term field and orchard experiments. The amount of empirical data included in the modelling is a very important element because too little can lead to excessive forecast errors. A large amount of data collected from many years is highly desirable regardless of the chosen modelling method. It is widely accepted that when using machine learning methods, the amount of data should be as large as possible [34]. This increases the real chances of obtaining a low forecast error. In addition, the number of features that are considered in modelling with machine learning methods should also be large. Most often, a few or a dozen features up to 30 features are used in crop yield forecasting issues [5,48,80], which do not always fully capture the nature of plant growth and vegetation. However, datasets that are too large (in terms of the amount of data and the number of traits) can lead to inappropriate model operation due to existing cross-correlations between the analyzed features. In such cases, the result can be increased prediction error and long computation time.

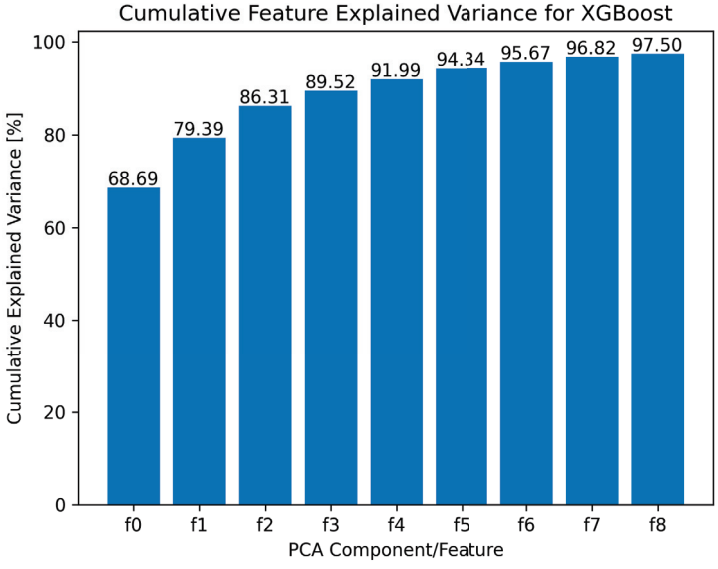


Figure 5. Cumulative Feature explained variance for XGBoost.

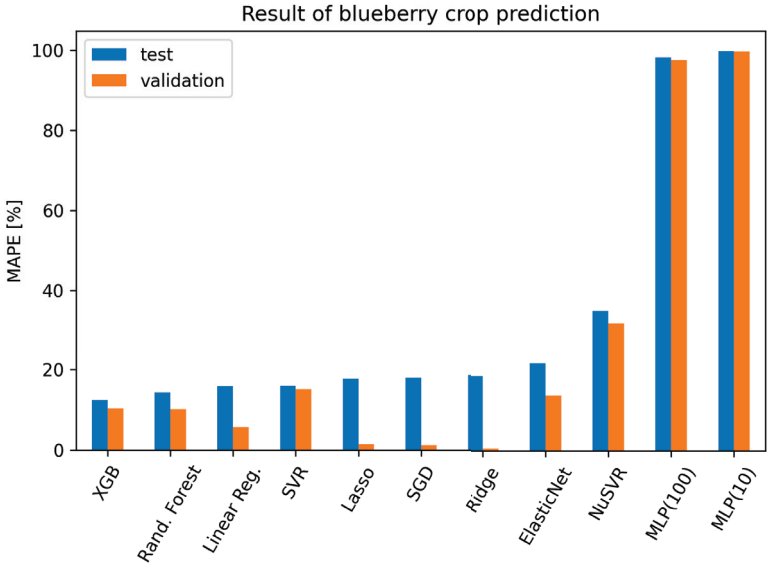


Figure 6. Result of blueberry crop prediction for test and validation.

Algorithm 2: Algorithm of numerical experiments

```
X ← AllFeatures
AllClassifiers ← Classifiers[1, ..., 11]
foreach classifier in AllClassifiers do
    BuildModel(X)
    AllFeaturesPearson ← Perason(X, 0.95)
    BuildModel(AllFeaturesPearson)
    AllFeaturesChi2 ← Chi2(X)
    BuildModel(AllFeaturesChi2)
    /* PCA                                                          */
    foreach n_components=3, ..., 11 do
        PCAAllFeatures ← PCA(X,n_components)
        BuildModel(PCAAllFeatures)
    end
    /* stepwise regression                                          */
    foreach penter=[0.01:0.01:0.90], premove=[0.06:0.01:0.90] do
        StepwiseAllFeatures ← Stepwisefit(X,penter,premove)
        BuildModel(ChosenAllFeatures)
        StepwiseFeaturesPearson ← Perason(StepwiseAllFeatures, 0.95)
        BuildModel(StepwiseFeaturesPearson)
        StepwiseFeaturesChi2 ← Chi2(StepwiseAllFeatures)
        BuildModel(StepwiseFeaturesChi2)
        /* PCA                                                      */
        foreach n_components=3, ..., 11 do
            PCAStepwiseAllFeatures ← PCA(StepwiseAllFeatures,n_components)
            BuildModel(PCAStepwiseAllFeatures)
        end
    end
end
```

In this work, we analyzed three varieties of blueberry (Chandler, Liberty, Nelson), for which data were collected from six growing seasons spanning from 2016 to 2021. Blueberries were grown on an area of 393.34 hectares on 243 subplots. A total of 89

explanatory characteristics were analyzed in the form of meteorological data, irrigation, fertigation and plant fertilization information, soil data, vegetation indices from satellite imagery and time intervals. Some of the data were aggregated into indices in the form of the Sielaninov hydrothermal coefficient and the sum of effective temperatures (GDD). In addition, some of the collected data was aggregated to growing seasons according to the BBCH scale.

Obsie et al. [11] investigated the effects of the spatial distribution of blueberry plants, the species composition of bees in the field, and weather conditions on yield. In their study, they considered three groups of parameters: (1) the average size of blueberry clones within the field; (2) the foraging density of each group of bee taxa; and (3) weather information from 121 to 181 days from the beginning of the calendar year (temperature, precipitation and wind speed). The dataset consisted of 13 features, and 777 records were analyzed, for which 77,700 simulations were performed. MLR, Boosted decision tree (BDT), Random Forest and XGBoost algorithms were used to predict blueberry yields. The XGBoost algorithm achieved the best results for yield prediction, with an R^2 of 0.938 and an RRMSE of 5.444%.

In our study, the best algorithm was XGBoost, for which we obtained a MAPE error of 12.48%. Relating our yield forecast result to the results of Obsie et al. [11], it should be considered that the XGBoost algorithm was found to be the best.

Seireg et al. [47] predicted blueberry yield using mixed machine learning techniques LGBM, GBR, XGBoost and Ridge. The data to create the blueberry yield models came from 30 years and included meteorological information (temperature and precipitation). In total, five datasets (M1-M5) were used for calculations covering a different number of features (7–10) depending on the feature selection method adopted. The best yield prediction results were obtained for seven traits using the stacking technique with a combination of LGBM, GBR, XGBoost and Ridge algorithms. The model achieved an accuracy of 0.984 R^2 and 179.898 RMSE.

MacEachern et al. [49] collected two years of imagery data from 54 points located in 4 blueberry production fields. After their images were taken, all of the blueberry fruit was harvested with a hand rake. This made it possible to determine the yield for each study point. A total of 17,280 images were accepted for analysis, from which 6766 images were randomly selected for labelling. Overall, six deep neural networks (DNNs) YOLOv3 (3) and YOLOv4 (3), in different configurations were used for analysis. The best model was YOLOv4-Small, which achieved 89.67 R^2 and an absolute average error of 24.1%.

The parameter that determines the quality of the forecasts made is the mean absolute percentage error (MAPE). It is most often interpreted as the average percentage deviation between the forecast value and the actual realization. Peng et al. [81] provide threshold values for the correct assessment of the MAPE index. If the error is less than 10%, the degree of goodness of the model is ideal, while a range of 10 to 20% indicates a good fit, and 20 to 30% indicates a level of acceptance. MAPE above 30% indicates low model accuracy and disqualifies the model from practical use. In individual studies, MAPE reached low values.

In our study, we obtained MAPE error of 12.48% for XGBoost. Most of the other algorithms did not exceed 20% MAPE, with the exception of ElasticNet, NuSVR, MLP(100) and MLP(10). This indicates a good model fit. The variables chosen for this model using the stepwise fit method were as follows:

(1) Fertigation
(2) Hailstorm percentage of damage
(3) EVI 40 days before harvest (avg.)
(4) RDVI 40 days before harvest (max.)
(5) Dew point temperature (max.)
(6) NDVI 40 days before harvest (avg.)
(7) SAVI 40 days before harvest (min.)
(8) SAVI 40 days before harvest (stddev.)
(9) Irrigation (BBCH > 70)

(10) Dew point temperature (avg.)
(11) P—phosphorus
(12) Mn—manganese
(13) NDVI 40 days before harvest (avg.)
(14) Fe—iron
(15) RDVI 40 days before harvest (avg.)
(16) Fertigation (BBCH 0–60)
(17) SAVI 40 days before harvest (avg.)
(18) NDVI 40 days before harvest (max.)
(19) pH
(20) B—boron
(21) C—calcium
(22) NDVI 40 days before harvest (min.)
(23) EVI 40 days before harvest (min.)
(24) N-NO3—nitroge
(25) Fertilization (BBCH 61–70)
(26) EVI 40 days before harvest (max.)
(27) Na—sodium
(28) K—potassium
(29) Soil P—phosphorus (avg.) (BBCH 61–70)
(30) SAVI 40 days before harvest (max.)
(31) HTC (BBCH 61–70)
(32) Fertigation
(33) RDVI 40 days before harvest (min.)
(34) HTC (BBCH > 70)
(35) Soil P—phosphorus (avg.) (BBCH > 70)
(36) Irrigation (BBCH 0–60)
(37) K-potassium-Fertilization (annually)
(38) Cu—copper
(39) S-SO_4—sulfur
(40) Rainfall (BBCH 0–60)

On the basis of real features above, PCA has been applied which generated nine artificial features. The nine features importance is depicted in Figure 4.

4. Conclusions

With the growing relevance of blueberry production comes the desirability for better monitoring of crop yields, as well as conscious production management and proper pre-harvest decision-making. The results of the individual analyses presented here indicate that machine learning methods are a very useful tool in the yield prediction of the blueberry varieties Chandler, Liberty and Nelson. Yield prediction models are used to estimate yields during the growing season, immediately before harvest.

The presented models and their quality are strictly dependent on the availability of data, climatic and vegetation parameters of growing areas, cultivated varieties, thus constituting the factors limiting their interoperability. The implementation of the developed models for crops located in regions with significantly different climatic and vegetation conditions than for southeastern Poland should be preceded by additional research work and the implementation of eventual corrections to the developed models.

Overall, predictions made prior to fruit harvest are a valuable source of knowledge before harvesting, selling and storing crops. The presented models were able to achieve sufficient accuracy and error rate, while the presented approach shows great promise for further applications and development.

Author Contributions: K.B., G.N. and T.W. conceived the study design, managed data collection, built the database, and performed the first data analysis. J.K. and B.Ś. carried out all deep data analyses and built models until final results were attained. I.A., G.N., T.W., J.K. and B.Ś. wrote the manuscript with substantial input from K.B. All authors have read and agreed to the published version of the manuscript.

Funding: The project is co-financed by the European Union from the European Regional Development Fund under the Smart Growth Operational Programme. The project is being conducted under the competition of the National Centre for Research and Development, within the 1.1.1 programme for R&D projects of enterprises "Fast track–Agrotech" number: POIR.01.01.01-00-2298/20.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We gratefully acknowledge Łukasz Cypcar and Szymon Margański for facilitating sample collection and organizing data in storage; all the members from Seth Software teams for their assistance and helpful discussions. We thank Joanna Kogut for administrative services without which we could not provide such fruitful results.

Conflicts of Interest: The authors declare that they have no competing interest.

References

1. Qu, H.; Xiang, R.; Obsie, E.Y.; Wei, D.; Drummond, F. Parameterization and Calibration of Wild Blueberry Machine Learning Models to Predict Fruit-Set in the Northeast China Bog Blueberry Agroecosystem. *Agronomy* **2021**, *11*, 1736. [CrossRef]
2. Golovinskaia, O.; Wang, C.K. Review of Functional and Pharmacological Activities of Berries. *Molecules* **2021**, *26*, 3904. [CrossRef] [PubMed]
3. FAOSTAT My Name Is John Doe. Available online: https://www.fao.org/faostat/en/#data/QCL (accessed on 14 September 2022).
4. Salvo, S.; Muñoz, C.; Ávila, J.; Bustos, J.; Ramírez-Valdivia, M.; Silva, C.; Vivallo, G. An estimate of potential blueberry yield using regression models that relate the number of fruits to the number of flower buds and to climatic variables. *Sci. Hortic.* **2012**, *133*, 56–63. [CrossRef]
5. Piekutowska, M.; Niedbała, G.; Piskier, T.; Lenartowicz, T.; Pilarski, K.; Wojciechowski, T.; Pilarska, A.A.; Czechowska-Kosacka, A. The Application of Multiple Linear Regression and Artificial Neural Network Models for Yield Prediction of Very Early Potato Cultivars before Harvest. *Agronomy* **2021**, *11*, 885. [CrossRef]
6. Gorzelany, J.; Belcar, J.; Kuźniar, P.; Niedbała, G.; Pentoś, K. Modelling of Mechanical Properties of Fresh and Stored Fruit of Large Cranberry Using Multiple Linear Regression and Machine Learning. *Agriculture* **2022**, *12*, 200. [CrossRef]
7. Niazian, M.; Sadat-Noori, S.A.; Abdipour, M. Modeling the seed yield of Ajowan (*Trachyspermum ammi* L.) using artificial neural network and multiple linear regression models. *Ind. Crops Prod.* **2018**, *117*, 224–234. [CrossRef]
8. Sabzi-Nojadeh, M.; Niedbała, G.; Younessi-Hamzekhanlu, M.; Aharizad, S.; Esmaeilpour, M.; Abdipour, M.; Kujawa, S.; Niazian, M. Modeling the Essential Oil and Trans-Anethole Yield of Fennel (*Foeniculum vulgare* Mill. var. vulgare) by Application Artificial Neural Network and Multiple Linear Regression Methods. *Agriculture* **2021**, *11*, 1191. [CrossRef]
9. Hara, P.; Piekutowska, M.; Niedbała, G. Selection of Independent Variables for Crop Yield Prediction Using Artificial Neural Network Models with Remote Sensing Data. *Land* **2021**, *10*, 609. [CrossRef]
10. He, L.; Fang, W.; Zhao, G.; Wu, Z.; Fu, L.; Li, R.; Majeed, Y.; Dhupia, J. Fruit yield prediction and estimation in orchards: A state-of-the-art comprehensive review for both direct and indirect methods. *Comput. Electron. Agric.* **2022**, *195*, 106812. [CrossRef]
11. Obsie, E.Y.; Qu, H.; Drummond, F. Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. *Comput. Electron. Agric.* **2020**, *178*, 105778. [CrossRef]
12. Khan, H.; Esau, T.J.; Farooque, A.A.; Abbas, F. Wild blueberry harvesting losses predicted with selective machine learning algorithms. *Agriculture* **2022**, *12*, 1657. [CrossRef]
13. Huang, W.; Wang, X.; Zhang, J.; Xia, J.; Zhang, X. Improvement of blueberry freshness prediction based on machine learning and multi-source sensing in the cold chain logistics. *Food Control* **2022**, *145*, 109496. [CrossRef]
14. Vasconez, J.P.; Delpiano, J.; Vougioukas, S.; Cheein, F.A. Comparison of convolutional neural networks in fruit detection and counting: A comprehensive evaluation. *Comput. Electron. Agric.* **2020**, *173*, 105348. [CrossRef]
15. Häni, N.; Roy, P.; Isler, V. A comparative study of fruit detection and counting methods for yield mapping in apple orchards. *J. Field Robot.* **2020**, *37*, 263–282. [CrossRef]
16. Coviello, L.; Cristoforetti, M.; Jurman, G.; Furlanello, C. GBCNet: In-field grape berries counting for yield estimation by dilated CNNs. *Appl. Sci.* **2020**, *10*, 4870. [CrossRef]
17. Koirala, A.; Walsh, K.; Wang, Z.; McCarthy, C. Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of 'MangoYOLO'. *Precis. Agric.* **2019**, *20*, 1107–1135. [CrossRef]

18. Mekhalfi, M.L.; Nicolò, C.; Ianniello, I.; Calamita, F.; Goller, R.; Barazzuol, M.; Melgani, F. Vision system for automatic on-tree kiwifruit counting and yield estimation. *Sensors* **2020**, *20*, 4214. [CrossRef]
19. Gutiérrez, S.; Wendel, A.; Underwood, J. Ground based hyperspectral imaging for extensive mango yield estimation. *Comput. Electron. Agric.* **2019**, *157*, 126–135. [CrossRef]
20. Kalantar, A.; Edan, Y.; Gur, A.; Klapp, I. A deep learning system for single and overall weight estimation of melons using unmanned aerial vehicle images. *Comput. Electron. Agric.* **2020**, *178*, 105748. [CrossRef]
21. Torres-Sánchez, J.; Mesas-Carrascosa, F.J.; Santesteban, L.G.; Jiménez-Brenes, F.M.; Oneka, O.; Villa-Llop, A.; Loidi, M.; López-Granados, F. Grape cluster detection using UAV photogrammetric point clouds as a low-cost tool for yield forecasting in vineyards. *Sensors* **2021**, *21*, 3083. [CrossRef]
22. Di Gennaro, S.F.; Toscano, P.; Cinat, P.; Berton, A.; Matese, A. A low-cost and unsupervised image recognition methodology for yield estimation in a vineyard. *Front. Plant Sci.* **2019**, *10*, 559. [CrossRef] [PubMed]
23. Apolo-Apolo, O.E.; Pérez-Ruiz, M.; Martínez-Guanter, J.; Valente, J. A cloud-based environment for generating yield estimation maps from apple orchards using UAV imagery and a deep learning technique. *Front. Plant Sci.* **2020**, *11*, 1086. [CrossRef] [PubMed]
24. Khoshnevisan, B.; Rafiee, S.; Mousazadeh, H. Application of multi-layer adaptive neuro-fuzzy inference system for estimation of greenhouse strawberry yield. *Measurement* **2014**, *47*, 903–910. [CrossRef]
25. Papageorgiou, E.; Aggelopoulou, K.; Gemtos, T.; Nanos, G. Yield prediction in apples using Fuzzy Cognitive Map learning approach. *Comput. Electron. Agric.* **2013**, *91*, 19–29. [CrossRef]
26. Wojciechowski, T.; Mazur, A.; Przybylak, A.; Piechowiak, J. Effect of Unitary Soil Tillage Energy on Soil Aggregate Structure and Erosion Vulnerability. *J. Ecol. Eng.* **2020**, *21*, 180–185. [CrossRef]
27. Bai, X.; Li, Z.; Li, W.; Zhao, Y.; Li, M.; Chen, H.; Wei, S.; Jiang, Y.; Yang, G.; Zhu, X. Comparison of machine-learning and casa models for predicting apple fruit yields from time-series planet imageries. *Remote Sens.* **2021**, *13*, 3073. [CrossRef]
28. Van Beek, J.; Tits, L.; Somers, B.; Deckers, T.; Verjans, W.; Bylemans, D.; Janssens, P.; Coppin, P. Temporal dependency of yield and quality estimation through spectral vegetation indices in pear orchards. *Remote Sens.* **2015**, *7*, 9886–9903. [CrossRef]
29. Li, G.; Suo, R.; Zhao, G.; Gao, C.; Fu, L.; Shi, F.; Dhupia, J.; Li, R.; Cui, Y. Real-time detection of kiwifruit flower and bud simultaneously in orchard using YOLOv4 for robotic pollination. *Comput. Electron. Agric.* **2022**, *193*, 106641. [CrossRef]
30. Matese, A.; Di Gennaro, S.F. Beyond the traditional NDVI index as a key factor to mainstream the use of UAV in precision viticulture. *Sci. Rep.* **2021**, *11*, 1–13. [CrossRef]
31. Sinwar, D.; Dhaka, V.S.; Sharma, M.K.; Rani, G. AI-based yield prediction and smart irrigation. In *Internet of Things and Analytics for Agriculture, Volume 2*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 155–180.
32. Engen, M.; Sandø, E.; Sjølander, B.L.O.; Arenberg, S.; Gupta, R.; Goodwin, M. Farm-scale crop yield prediction from multi-temporal data using deep hybrid neural networks. *Agronomy* **2021**, *11*, 2576. [CrossRef]
33. Fukuda, M.; Okuno, T.; Yuki, S. Central Object Segmentation by Deep Learning to Continuously Monitor Fruit Growth through RGB Images. *Sensors* **2021**, *21*, 6999. [CrossRef] [PubMed]
34. Cravero, A.; Pardo, S.; Sepúlveda, S.; Muñoz, L. Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review. *Agronomy* **2022**, *12*, 748. [CrossRef]
35. Angulo-Meza, L.; González-Araya, M.; Iriarte, A.; Rebolledo-Leiva, R.; de Mello, J.C.S. A multiobjective DEA model to assess the eco-efficiency of agricultural practices within the CF+ DEA method. *Comput. Electron. Agric.* **2019**, *161*, 151–161. [CrossRef]
36. Yarborough, D. Development of a crop estimation technique for wild blueberries. In Proceedings of the VII International Symposium on Vaccinium Culture 574, Chillan, Chile, 4–9 December 2000; pp. 409–413.
37. Zaman, Q.; Schumann, A.; Percival, D.; Gordon, R. Estimation of wild blueberry fruit yield using digital color photography. *Trans. ASABE* **2008**, *51*, 1539–1544. [CrossRef]
38. Swain, K.C.; Zaman, Q.U.; Schumann, A.W.; Percival, D.C.; Bochtis, D.D. Computer vision system for wild blueberry fruit yield mapping. *Biosyst. Eng.* **2010**, *106*, 389–394. [CrossRef]
39. Panda, S.S.; Hoogenboom, G.; Paz, J.O. Remote sensing and geospatial technological applications for site-specific management of fruit and nut crops: A review. *Remote Sens.* **2010**, *2*, 1973–1997. [CrossRef]
40. Yang, C.; Lee, W.S.; Williamson, J.G. Classification of blueberry fruit and leaves based on spectral signatures. *Biosyst. Eng.* **2012**, *113*, 351–362. [CrossRef]
41. Tan, K.; Lee, W.S.; Gan, H.; Wang, S. Recognising blueberry fruit of different maturity using histogram oriented gradients and colour features in outdoor scenes. *Biosyst. Eng.* **2018**, *176*, 59–72. [CrossRef]
42. Jafari, F.; Nassar, L.; Karray, F. Time series similarity analysis framework in fresh produce yield forecast domain. In Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Melbourne, Australia, 17–21 October 2021; pp. 2368–2374.
43. Nagaraju, Y.; Hegde, S.U.; Stalin, S. Fine-tuned mobilenet classifier for classification of strawberry and cherry fruit types. In Proceedings of the 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 23–25 June 2021; pp. 1–8.
44. Ni, X.; Li, C.; Jiang, H.; Takeda, F. Three-dimensional photogrammetry with deep learning instance segmentation to extract berry fruit harvestability traits. *ISPRS J. Photogramm. Remote Sens.* **2021**, *171*, 297–309. [CrossRef]

45. Wojciechowski, T.; Niedbala, G.; Czechlowski, M.; Nawrocka, J.R.; Piechnik, L.; Niemann, J. Rapeseed seeds quality classification with usage of VIS-NIR fiber optic probe and artificial neural networks. In Proceedings of the 2016 International Conference on Optoelectronics and Image Processing, ICOIP 2016, Warsaw, Poland, 10–12 June 2016. [CrossRef]
46. Kujawa, S.; Dach, J.; Kozłowski, R.J.; Przybył, K.; Niedbała, G.; Mueller, W.; Tomczak, R.J.; Zaborowicz, M.; Koszela, K. Maturity classification for sewage sludge composted with rapeseed straw using neural image analysis . In Proceedings of the SPIE—The International Society for Optical Engineering, ICOIP 2016, Warsaw, Poland, 10–12 June 2016; Volume 10033, p. 100332H. [CrossRef]
47. Seireg, H.R.; Omar, Y.M.; Abd El-Samie, F.E.; El-Fishawy, A.S.; Elmahalawy, A. Ensemble machine learning techniques using computer simulation data for wild blueberry yield prediction. *IEEE Access* **2022**, *2020*, 3181970. [CrossRef]
48. Niedbała, G. Application of artificial neural networks for multi-criteria yield prediction of winter rapeseed. *Sustainability* **2019**, *11*, 533. [CrossRef]
49. MacEachern, C.B.; Esau, T.J.; Schumann, A.W.; Hennessy, P.J.; Zaman, Q.U. Detection of fruit maturity stage and yield estimation in wild blueberry using deep learning convolutional neural networks. *Smart Agric. Technol.* **2023**, *3*, 100099. [CrossRef]
50. Index DataBase. Available online: https://www.indexdatabase.de/ (accessed on 20 November 2022).
51. Anderberg, M.R. *Cluster Analysis for Applications*; Academic Press: New York, NY, USA, 1983.
52. R Development Core Team . *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2012; ISBN 3-900051-07-0. Available online: http://www.R-project.org (accessed on 20 November 2022).
53. Arabie, P.; Carroll, J.D. MAPCLUS: A mathematical programming approach to fitting the ADCLUS models. *Psychometrika* **1980**, *445*, 211–235. [CrossRef]
54. Tufte, E.R. *Envisioning Information*; Graphics Press: Cheshire, CT, USA, 1990.
55. Tufte, E.R. *The Visual Display of Quantitative Information*; Graphics Press: Cheshire, CT, USA, 1983.
56. Cleveland, W.S. *The Elements of Graphing Data*, revised ed.; Hobart Press: Thousand Oaks, CA, USA, 1994.
57. Cleveland, W.S. *Vizualizing Data*; Hobart Press: Thousand Oaks, CA, USA, 1993.
58. Ball, G.H.; Hall, D.J. *A Novel Method of Data Analysis and Pattern Classification*; Technical Report; Stanford Research Institute: Stanford, CA, USA, 1965.
59. Banfield, J.D.; Raftery, A.E. Model-Based Gaussian and Non–Gaussian Clustering. *Biometrics* **1993**, *49*, 803–821. [CrossRef]
60. Beale, E.M.L. Euclidean cluster analysis. *Bull. Int. Stat. Inst.* **1969**, *43*, 92–94.
61. Bensmail, H.; Meulman, J.J. Model-based clustering with noise: Bayesian inference and estimation. *J. Classif.* **2003**, *20*, 049–076. [CrossRef]
62. Bezdek, J.C. Numerical taxonomy with fuzzy sets. *J. Meth. Biol.* **1974**, *1*, 57–71. [CrossRef]
63. Cox, D.R. Regression models and life tables (with Discussion). *J. R. Stat. Soc.* B **1972**, *34*, 187–220.
64. Heard, N.A.; Holmes, C.C.; Stephens, D.A. A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves. *J. Am. Stat. Assoc.* **2006**, *101*, 18–29. [CrossRef]
65. Fan, J.; Peng, H. Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Stat.* **2004**, *32*, 928–961. [CrossRef]
66. Selyaninov, G. Methods of agricultural climatology. *Agric. Meteorol.* **1930**, *22*, 4–20.
67. Prentice, I.C.; Cramer, W.; Harrison, S.P.; Leemans, R.; Monserud, R.A.; Solomon, A.M. Special Paper: A Global Biome Model Based on Plant Physiology and Dominance, Soil Properties and Climate. *J. Biogeogr.* **1992**, *19*, 117–134. [CrossRef]
68. XGBoost Package. Available online: https://xgboost.readthedocs.io/en/stable/python/python_intro.html (accessed on 20 November 2022).
69. Comparing Anomaly Detection Algorithms for Outlier Detection on Toy Datasets. Available online: https://scikit-learn.org/stable/auto_examples/miscellaneous/plot_anomaly_comparison.html (accessed on 20 November 2022).
70. Least squares Linear Regression. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Linear\Regression.html (accessed on 20 November 2022).
71. Ridge Model. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html (accessed on 20 November 2022).
72. Lasso Model. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html (accessed on 20 November 2022).
73. ElasticNet. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html (accessed on 20 November 2022).
74. Random Forest Regressor. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestReg\ressor.html (accessed on 20 November 2022).
75. Multi-Layer Perceptron Regressor. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLP\Regressor.html (accessed on 20 November 2022).
76. SGD Regressor. Availableonline: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html (accessed on 20 November 2022).
77. Epsilon-Support Vector Regression. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html (accessed on 20 November 2022).

78. Nu Support Vector Regression. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.svm.NuSVR.html (accessed on 20 November 2022).
79. "Pragmatic" Project Webpage. Available online: https://seth.software/zt_portfolio/pragramatic/ (accessed on 20 November 2022).
80. Niedbała, G.; Piekutowska, M.; Weres, J.; Korzeniewicz, R.; Witaszek, K.; Adamski, M.; Pilarski, K.; Czechowska-Kosacka, A.; Krysztofiak-Kaniewska, A. Application of artificial neural networks for yield modeling of winter rapeseed based on combined quantitative and qualitative data. *Agronomy* **2019**, *9*, 781. [CrossRef]
81. Peng, J.; Kim, M.; Kim, Y.; Jo, M.; Kim, B.; Sung, K.; Lv, S. Constructing Italian ryegrass yield prediction model based on climatic data by locations in South Korea. *Grassl. Sci.* **2017**, *63*, 184–195. [CrossRef]

agriculture

MDPI

Article

A Decision-Making Capability Optimization Scheme of Control Combination and PID Controller Parameters for Bivariate Fertilizer Applicator Improved by Using EDEM

Yugong Dang [1], Gang Yang [1], Jun Wang [2,*], Zhigang Zhou [1] and Zhidong Xu [3]

[1] School of Vehicle and Traffic Engineering, Henan University of Science and Technology, Luoyang 471000, China
[2] School of Information Engineering, Henan University of Science and Technology, Luoyang 471000, China
[3] China Petroleum First Construction Co., Ltd., Luoyang 471023, China
* Correspondence: wj@haust.edu.cn

Abstract: The fertilization rate is adjusted through the regulation of opening length and the rotational speed for bivariate fertilizer applicators. It is essential to optimally determine the control combination according to the target fertilization rate and further improve the control performance of fertilization operation in precision agriculture. In this study, a novel decision-making capability optimization scheme of control combination and PID controller parameters is proposed to improve the feasibility and practicability of variable fertilizer applicators. Firstly, EDEM is adopted to acquire the minimum allowable opening length and the proper gap between the spiral blades and the discharge cavity wall, and then calibration experiments are implemented to establish the fitting model of fertilization rate using polynomial fitting. Secondly, the modified sparrow search algorithm (SSA) with chaotic operator and mutation section of the DE algorithm is used to optimize the control combination utilizing the accuracy, uniformity, and adjustment time as the evaluation criteria. Moreover, the tent mapping bat algorithm (TBA) is applied to tune the PID controller parameters for enhancing the accuracy and response speed of the fertilization-rate control system. Compared to the PID controller based on the bat algorithm (BA), traditional PID controller, and fuzzy PID controller, the rise time of the PID controller improved by TBA decreases by 0.018 s, 0.09 s, and 0.038 s, respectively, and the average steady-state deviation of that drops by 0.02 kg ha^{-1}, 1.45 kg ha^{-1}, and 0.19 kg ha^{-1}, respectively. In addition, under the condition of the same controller, compared with SSA, GA, and MOEA/D-DE, the average accuracy of the proposed decision-making algorithm decreases from 1.9%, 2.5%, and 3.5% to 1.8%, the average uniformity drops from 0.52% and 0.48% to 0.47%, and the average adjustment time declines from 0.99 s, 1.48 s, and 1.34 s to 0.5 s. It can be concluded that the method proposed in this study performs better in terms of accuracy and adjustment time but exhibits no apparent effect on the improvement of uniformity.

Keywords: bivariate fertilizer applicator; opening length; rotational speed; control combination determination; PID parameter tuning

Citation: Dang, Y.; Yang, G.; Wang, J.; Zhou, Z.; Xu, Z. A Decision-Making Capability Optimization Scheme of Control Combination and PID Controller Parameters for Bivariate Fertilizer Applicator Improved by Using EDEM. *Agriculture* **2022**, *12*, 2100. https://doi.org/10.3390/agriculture12122100

Academic Editors: Gniewko Niedbała and Sebastian Kujawa

Received: 22 November 2022
Accepted: 6 December 2022
Published: 8 December 2022

1. Introduction

Chemical fertilizer is critical to boosting yield and efficiency in agricultural production. However, a low rate of fertilizer application will not favorably achieve the anticipated target. Moreover, overfertilization unavoidably reduces the utilization rate and affects the content of soil organic matter, leading to environmental issues such as soil hardening, acidification, and destruction of microbial systems. Therefore, various variable-rate fertilization studies have been carried out on quantitative fertilization in specific areas for proper field management [1–3].

Currently, the fertilizer discharge mechanism of variable-rate fertilizer applicators mainly includes fluted-wheel type, centrifugal-disc type, and spiral-shaft type. Compared

with the other two mechanisms, the spiral-shaft type demonstrates more adequate stability and uniformity of fertilization with varying rotational speeds [4–7]. It is necessary to consider the physical properties of fertilizer particles (density, size, shape, elasticity, looseness, etc.) and optimize the variation range of feeding length and the gap between the fertilizer discharge mechanism and the mechanism wall for determining the proper structure of the fertilizer applicator [8]. The mechanism optimization can be conveniently executed using software simulation. Furthermore, as multi-purpose discrete element simulation software, EDEM can be used to analyze the impacts of different parameter combinations on the fertilization rate [9,10]. For instance, Chen et al. adopted EDEM to simulate the fertilization rate under diverse spiral angles and rotational speeds to enhance the accuracy and uniformity of fertilization [11]. Tang et al. used EDEM to optimize parameters, including spiral-shaft dimension, spiral-blade radius, and pitch to achieve a satisfactory coefficient of variation of fertilization rate [12]. However, there are no sufficient studies on the influences of the opening-length range of the fertilizer discharge mechanism and the gap between the spiral blades and the discharge cavity wall on fertilization performance.

Presently, the investigations on variable-rate fertilization focus on univariate fertilization and bivariate fertilization. The univariate fertilization is confronted with the deteriorated accuracy and uniformity of fertilization caused by the improper motor speed and the limited range of fertilization rate due to the single regulating parameter. Hence, researchers focus more on the modification of the fertilization rate by jointly controlling the opening length of the fertilizer outlet and the rotational speed of the fertilizer discharging shaft. Liu et al. developed a bivariate fertilizer applicator for rapid and precise fertilization using two direct current (DC) servo motors [13]. Aaa et al. altered the opening length with pneumatic cylinders and the rotational speed by connecting the driven gear with the motor to expand the adjustment range of the fertilization rate [14]. Although bivariate fertilization achieves a broader regulation scope than univariate fertilization, the nonlinear problem between variates and fertilization rate caused by the increased number of variates complicates the control system.

In addition, accuracy and uniformity are usually used as the assessment metrics of fertilization effectiveness [15]. Currently, the advancement of fertilization performance of variable fertilizer applicators can be implemented by optimizing the decision-making and control systems. Yuan et al. established a fertilization control model based on a genetic algorithm (GA) optimized control combination using Gaussian process regression (GPR), with a mean relative error (MRE) of 0.089 [16]. Nevertheless, the accuracy of fertilization is indirectly affected by the time lag as a result of the transition process of opening length and rotational speed. Thus, the decision-making algorithm should consider the adjustment time during the optimization of the control combination. Zhang et al. built a three-objective problem model with accuracy, uniformity, and adjustment time as the objectives and unraveled the difficulty of the optimal fertilization control decision through the multi-objective evolutionary algorithm based on decomposition (MOEA/D) based on the differential evolution (DE) algorithm to gain a superior MRE of 0.05977 [17]. However, the method using MOEA/D for the solution of the multi-objective optimization model requires a considerably long running time to compute the fertilization decision in the actual scenarios. Moreover, various machine learning algorithms used to improve decision-making performance have different priorities in terms of computational efficiency, accuracy, and convergence speed, resulting in the requirement of global optimization for the whole variation range of variates during the determination of the control combination. The predicament causes many conventional optimization algorithms to drop into local optimum and requires considering the necessities of multiple objectives, failing to gain a reasonable accuracy of optimization results.

Swarm intelligence (SI) algorithms are gradually gaining prominence as more and more high-complexity decision-making problems require solutions within a reasonable time. Xue et al. proposed a new swarm optimization algorithm based on the behavior of group-living sparrows to achieve greater efficiency compared with other SI algorithms,

although it has the shortcoming of easily falling into local extremum [18]. Furthermore, Nguyen et al. recommended a modified sparrow search algorithm (SSA) utilizing the reverse learning strategy to enhance the accuracy and computational efficiency of multi-objective solutions [19]. In addition, a chaotic operator and mutation section in a differential evolution (DE) algorithm are feasible and practical measures for SSA improvement. In this field, Wang et al. suggested an enhanced SSA in terms of global search capability and accuracy using Bernoulli's chaotic maps [20]. Moreover, Kathiroli et al. offered a DE-based SSA to promote computation efficiency and global search capability [21].

Meanwhile, the control system regulates servo motors with closed-loop feedback consisting of a proportional–integral–derivative (PID) controller. The parameter-adaptive PID controller is widely investigated for reducing the inconvenience of empirical debugging and elevating the response speed and stability of the control system [22]. For instance, Zhang et al. designed a variable-rate liquid fertilization control system using an improved PID algorithm and assessed the control performance with rise time and steady-state error [23]. In addition, Bai et al. optimized the PID controller by adopting the integrated time absolute error (ITAE) criterion as a fitness function [24]. However, the optimization or tuning process of these modified PID controllers struggles to balance operation speed and control stability, leading to unbearable computational complexity. The bat algorithm (BA) integrating the advantages of particle swarm, echo, and simulated annealing algorithms is undoubtedly one of the effective solutions to this difficulty [25]. For example, Chaib et al. utilized BA to determine the parameters for optimizing the stability and response speed of a PID controller [26]. Nevertheless, BA also tends to fall into local optimum, and a chaotic operator can be introduced to further facilitate the optimization of PID controller parameters.

To solve the aforementioned difficulty, a control combination decision-making method and control system is proposed to boost the operational performance of a bivariate fertilizer applicator in this study. Firstly, using EDEM software, the gap between the spiral blades and the discharge cavity and the range of opening lengths are optimized, and the fertilization process of the optimized variable-rate fertilizer applicator is precisely modeled. Secondly, utilizing the accuracy, uniformity, and adjustment time of fertilization as assessment criteria, the control combination optimization is determined by the SSA enhanced with a chaotic operator and mutation section of the DE algorithm. Finally, the PID controller optimized based on the improved BA is adopted to regulate the opening length of the fertilizer outlet and the rotational speed of the fertilizer discharging shaft.

The main contributions and innovations of this study are as follows:

1. The mechanical structure of a bivariate fertilizer applicator is optimized by EDEM software to avoid the loss and accumulation of fertilizer particles during the discharge process caused by inappropriate opening length and gap in the actual scenario.
2. The improved SSA using the chaotic operator and mutation section of the DE algorithm is used to promote global searching capability, accuracy, and convergence speed of control combination determination by diversifying the increasing population diversity and avoiding falling into the local optimum.
3. The parameter tuning of the PID controller with an improved BA using tent mapping is applied to reduce the rise time and the steady-state deviation for continuous variation in the control variable's value.

2. Materials and Methods

2.1. Fertilizer Discharge Mechanism

In this study, the designed test bench of the bivariate fertilizer applicator primarily consists of the fertilizer hopper, adjustment device of opening length, and rotational-speed control device. The main structure of the fertilizer applicator is assembled with 45-type steel. As shown in Figure 1, the adjustment device of opening length (L) is constituted by the baffle, connector, lead screw, and servo motor 1 (AKM41H-ANCNC-00, A&S Industry Technology Corporation, Boston, MA, USA). Since the screw pitch is set as 2 mm, the

moving distance of the baffle is 1 mm in case the cylindrical rotor of servo motor 1 rotates half a turn. According to the forward and reverse rotation of servo motor 1, the lead screw is transformed into the linear reciprocating motion of the baffle to change the opening length through the connector. The rotational speed (N) of the fertilizer discharging shaft is regulated directly by servo motor 2 of an identical model connected to the spiral shaft through a coupling, and the motor parameters are R = 1.56 ohm, P_n = 5, J_m = 0.878 kg cm^2, L_s = 5 mH. Therefore, the accurate fertilization rate is obtained by controlling two servo motors to output the expected opening length and rotational speed. The structure optimization of the essential mechanism parameters of the fertilizer applicator is required for acceptable fertilization performance before the test bench is assembled.

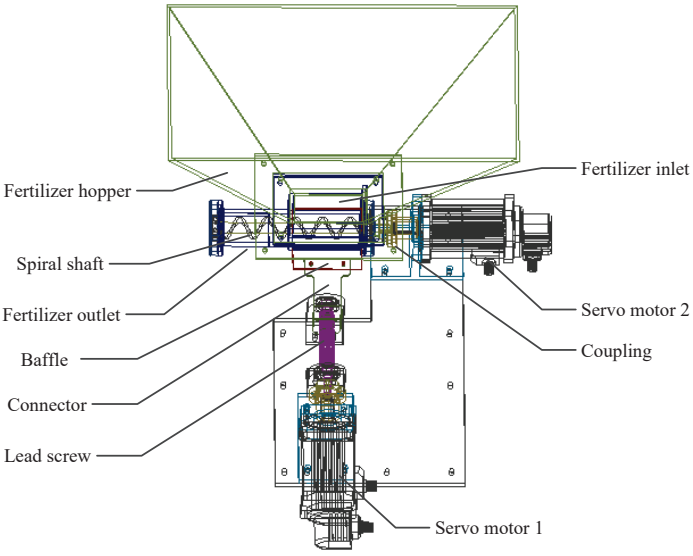


Figure 1. Three-dimensional view of fertilizer applicator.

2.2. *Parameter Optimization Based on EDEM*

The parameter selection of the fertilizer applicator is undoubtedly critical to actual fertilization performance. Once the fertilizer applicator demonstrates significant instability at a small opening length, this leads to lowering accuracy. Meanwhile, the gap between the spiral blades and the discharge cavity can directly induce undesirable accumulation or overhead of fertilizer particles and is related to the regular work of spiral blades. Moreover, an insufficient gap will cause excessive friction while an oversized gap will decrease fertilizer discharge efficiency. As excellent multi-purpose discrete element simulation software, EDEM can be used to analyze the motion law of intricate discrete systems and to detect the impacts of particle size and particle flow on the equipment under study. In our study, EDEM is applied to achieve the minimum allowable opening length and the reasonable gap through simulation.

Prior to the parameter optimization, the essential physical performance parameters of fertilizers must be experimentally obtained or set by manual experience. The compound fertilizer (N-P2O5-K2O) used in this study has nitrogen content of 15%, phosphorus content of 15%, potassium content of 15%, and sulfur content of 10%, with a particle size of 3–5 mm and a slightly irritating odor. One hundred particles are randomly selected, and then the triaxial dimensions are measured by a digital vernier caliper (DL91150, Ningbo Deli Tools Corporation, China) with an accuracy of 0.01 mm. The volume ratio is defined as follows:

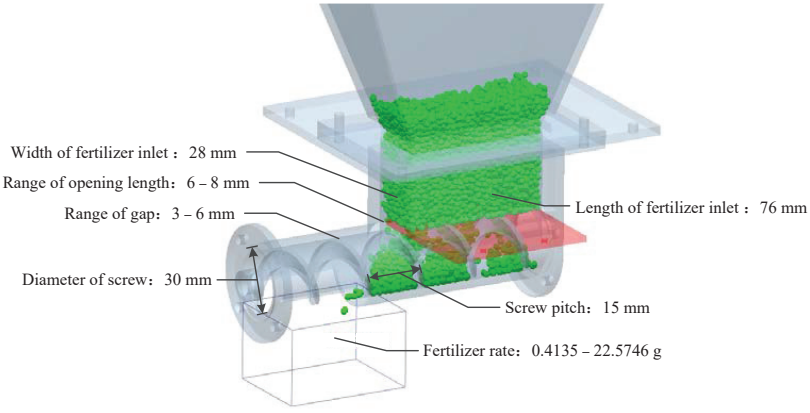
$$\sigma = \frac{V_a}{V_b} \tag{1}$$

where σ depicts the volume ratio of each fertilizer particle, v_a expresses the volume of a fertilizer particle (mm^3), and v_b is the average volume of fertilizer particles (mm^3).

In addition, the density of fertilizer particles is estimated by the drainage method, the average angle of repose of fertilizer particle accumulation is calculated by the injection method, and the oven-drying method is adopted to obtain the moisture content of fertilizer granules. Meanwhile, the static friction coefficient between fertilizer particles and steel plate and the static friction coefficient between fertilizer particles are individually gauged by the plane method. Especially, the rolling friction coefficient is empirically determined as 0.01. Moreover, the collision restitution coefficient between the fertilizer particle and steel plate and the collision restitution coefficient between fertilizer particles are respectively achieved by the free-fall experiment. Furthermore, other parameters are designated according to manual experience, including density of 7890 kg m^3, Poisson's ratio of 0.3 and Young's modulus of 209,000 GP for 45-type steel; Poisson's ratio of 0.25 and Young's modulus of 250 GP for fertilizer particles.

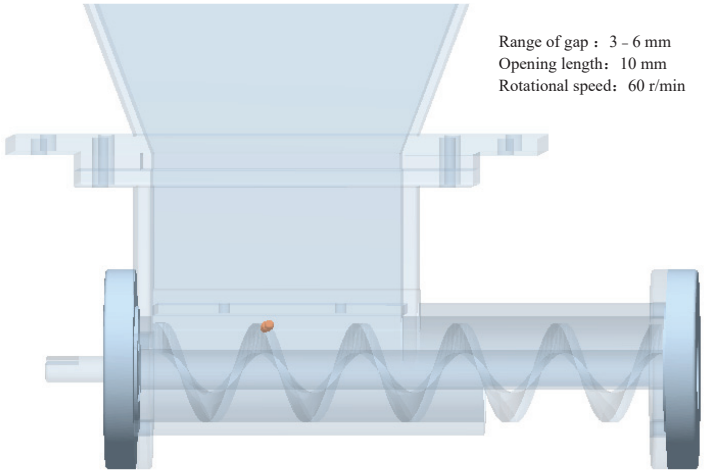
After the acquisition and configuration of the necessary parameters, the established 3D model of the bivariate fertilizer applicator is imported into the EDEM, and 500 g fertilizer is generated in the fertilizer hopper. In addition, the baffle and fertilizer discharging shaft are selected as moving parts. The search range of the minimum acceptable opening length is set from 5 mm to twice the diameter of fertilizer particles with an increment of 1 mm per time. Meanwhile, a virtual mass detector is used to weigh the discharged fertilizer, and the duration per simulation is assigned as 3 s (Figure 2a).

Considering the diameter range of fertilizer particles and the deviation influenced by the coaxiality between the motor shaft and the fertilizer discharging shaft in actual operations, the gap between the spiral blades and the discharge cavity should be greater than 2 mm. Thus, the simulation range of the gap is designated as 3 to 6 mm, and the mixed fertilizer particles of random diameters are produced for the discharge performance test. In addition, the movement process of individual fertilizer particles of 3 to 5 mm is simulated under the selected gap scope for surveying the influence of the gap on the discharge status of the single particle (Figure 2b). Eventually, the optimal range of opening length and the most suitable gap are acquired with the simulation comparisons.



(**a**)

Figure 2. *Cont*.



(**b**)

Figure 2. Illustration of mechanical parameter optimization by EDEM: (**a**) Schematic diagram of opening-length optimization using EDEM; (**b**) schematic diagram of gap determination through EDEM.

2.3. Fertilization Data Acquisition

The test bench of the bivariate fertilizer applicator is built with the parameters optimized through EDEM. To establish the decision-making system, it is essential to measure the fertilization rates in case of various opening lengths and rotational speeds and assemble a fitting model to resolve the optimal combination of control variates. Therefore, indoor calibration experiments are conducted under stationary conditions to obtain fertilization data under different combinations of opening length and rotational speed. Due to the considerable growth variation in the fertilization rate at a small opening length, to accurately acquire the characteristics of the fertilization data, the size step of the opening length is set to 0.5 mm, 1 mm, and 2 mm for the ranges of 8–12 mm, 13–16 mm, and 18–28 mm, respectively. In addition, the variation range of the rotational speed is 10–150 r min^{-1}, and the step size is set to 10 r min^{-1}. The detailed ranges of opening length and rotational speed are listed in Table 1.

Table 1. Change ranges of regulation parameters.

Parameter	Range	Step Size
Opening length	8–12 mm	0.5
	13–16 mm	1
	18–28 mm	2
Rotational speed	10–150 r min^{-1}	10

It is easy to see that the total number of combinations of opening length and rotational speed is 285. Each combination is sampled 20 times, lasting for 30 s each time. The discharged fertilizer for each test is weighed, and the average fertilization rate per combination is recorded (Supplementary Materials). Moreover, 10% of the experimental data are randomly selected as the test set, and the residual of the data is fitted with a polynomial curve fitting function. The data of the test set are compared with the predicted fertilization rate estimated by the fitting model, and the mean relative error (MRE) and coefficient of determination (R^2) are applied to assess the fitting effect. The most satisfactory

fitting model of the fertilization rate is used as the fitness function of the decision-making algorithm.

2.4. Decision-Making Method for Fertilization-Rate Control

For the fitting model, there is a possibility that multiple combinations of opening length and rotational speed correspond to an identical target fertilization rate. At the same time, the control variates must be altered once the target fertilization rate is varied. Therefore, the maximum adjustment time of opening length and rotational speed is one of the principal factors affecting the response speed of the fertilizer applicator. Meanwhile, owing to the continuity of the fertilization process, the shorter the adjustment time, the lower the cumulative error of the fertilization rate. Furthermore, SSA is a novel swarm intelligence optimization method suggested by Xue and Shen. Aiming at the problem that SSA is prone to fall into the local optimum, the modified SSA based on a reverse learning strategy facilitates more efficiency than other machine learning algorithms and can avoid the severe limitation of tending to local extremum. In addition to using the reverse learning strategy, SSA can also be promoted by introducing a chaotic operator and mutation section in the DE algorithm.

Consequently, to overcome the above-mentioned difficulties, the improved SSA with a chaotic operator and mutation section of the DE algorithm is used in this study to optimize the control combination utilizing the adjustment time and fertilization accuracy as the evaluation criteria, and the fitness function is established with the difference between the fitted value of the fertilization-rate model and the expected fertilization rate. Specifically, the solution sets within 5% deviation of the target fertilization rate are probed by iteration. Then, the combination with the minimum discrepancy with the previous control combination is selected as the optimal regulation sequence. In particular, the fitness function of the improved SSA used to determine the control combination is defined as follows:

$$f(x) = \frac{P * 1000 - Q}{Q} \tag{2}$$

where P denotes the predicted value achieved from the fitting model of fertilization rate, and Q expresses the target fertilization rate.

The position coordinates of sparrows in the algorithm are formed by opening length and rotational speed. The position update formula for the predator during the population predation is described as follows [27]:

$$x_i^{t+1} = \begin{cases} x_i^t \cdot e^{(\frac{-i}{\alpha T})}, R_2 < S_T \\ x_i^t + BV, R_2 \geq S_T \end{cases} \tag{3}$$

where x_i^t is the two-dimensional coordinate of the i-th individual of the population for the t-th iteration, T is the maximum number of iterations, α is a random number in the interval of 0 to 1, B is a normally distributed random number, and V is a matrix of $1 \times d$ with each element being 1. In addition, $R_2 \in [0, 1]$ and $S_T \in [0.5, 1]$ indicate the warning value and the safety threshold, respectively.

The position update formula for the intrant is depicted as follows [28]:

$$x_i^{t+1} = \begin{cases} B \cdot e^{(\frac{xw_i^t - x_i^t}{i^2})}, & i > \frac{n}{2} \\ xb_i^{t+1} + \left| x_i^t - xb_i^{t+1} \right| C'(CC')^{-1} \cdot V, & i \leq \frac{n}{2} \end{cases} \tag{4}$$

where C is a matrix of $1 \times d$ with elements randomly being 1 or -1, xw_i^t is the worst position of the population of the t-th iteration, n is the population number, and xb_i^{t+1} is the optimal position of the population of the $t + 1$-th iteration.

The position update formula for the scouter is defined as follows [29]:

$$x_i^{t+1} = \begin{cases} xb_i^t + \beta\left(x_i^t - xb_i^t\right), & f_i \neq f_g \\ x_i^t + K\left(\frac{x_i^t - xw_i^t}{|f_i - f_w| + e}\right), & f_i = f_g \end{cases} \tag{5}$$

where β is a normally distributed random number with a mean value of 0 and a variance of 1, K is a random number between -1 and 1, and e is a minimal constant to prevent the denominator from being 0. Moreover, f_i means the fitness value of the i-th sparrow, f_g and f_g signify the best and worst fitness values of the current sparrow population, respectively.

The introduced chaotic operator and mutation section of the DE algorithm are expressed as follows [30,31]:

$$z_{i+1} = \begin{cases} 2z_i + rand(1) * \frac{1}{pop} & \left(0 \leq z \leq \frac{1}{2}\right) \\ 2(1 - z_i) + rand(1) * \frac{1}{pop} & \left(\frac{1}{2} < z \leq 1\right) \end{cases} \tag{6}$$

$$Tent(x) = x_j + \left(s_j - x_j\right) * z_{i+1} \tag{7}$$

$$x_{new} = x_a^t + \tau\left(x_b^t - x_c^t\right) \tag{8}$$

where z is a random number within [0, 1], pop is the population size, $Tent(x)$ is the value after chaotic mapping, $rand(x)$ is a normally distributed random number with an expectation of 0 and a standard deviation of 1, x_j is the upper bound of the range of individual coordinates, s_j is the lower bound of the range of individual coordinates, x_a^t, x_b^t, and x_c^t are the different individuals of the t-th iteration, $\tau \in [0.5, 1]$ is the scaling factor, and x_{new} is the position of new individuals after mutation operation.

During the search process, the improved SSA randomly generates various solutions from the local optimal solutions of each iteration by tent chaotic mapping and mutation section and computes the corresponding fitness values, and the individual with the best fitness is selected as the ultimate control combination. The parameters of improved SSA are shown in Table 2.

Table 2. Parameter setting of improved SSA.

Algorithm	Parameter	Value
Improved SSA	Population size *pop*	100
	Iteration number T	100
	Function dimension *dim*	2
	Number of predators P_{Num}	20
	Number of intrants J_{Num}	80
	Number of scouters S_{Num}	20
	Upper bound of search scope	(28, 150)
	Lower bound of search scope	(8, 10)

The pseudo-code of the decision-making method of fertilization-rate control based on the improved SSA is illustrated in Algorithm 1.

Algorithm 1 Decision-making method of fertilization-rate control using improved SSA

Input: Target fertilization rate S, population size *pop*, number of predators P_{Num}, number of intrants J_{Num}, number of scouters S_{Num}, function dimension *dim*, iteration number T, previous control combination y.
Output: Optimal control combination (L, N).
1. **for** each z_{i+1}, $i \in [0, 99]$ **do**
2. Compute points z_{i+1} using Equations (6) and (7)
3. **end**
4. $x_{(i,j)} \leftarrow \left((L_s - L_x) * z_i + L_x,\ (N_s - N_x) * z_i + N_x\right)$

Algorithm 1 *Cont.*

5. **for** each $x_{(i,j)}$, $i \in [1, 100]$, $j \in [1, 2]$ do
6. Compute f_i using Equation (2)
7. return min f_i, max f_i, and x_{worse}
8. **end**
9. $x_{new} \leftarrow x_{best}$
10. $T = 1$
11. **while** $T = 100$ **do**
12. **for** each f_i, $i \in [1, 100]$ **do**
13. $SI \leftarrow sort\ from\ min f_i\ to\ max f_i$
14. **end**
15. $st = rand(1)/2 + 0.5$
16. $P_{Num} \leftarrow x_{(i,j)} \leftarrow SI_{1\sim20}$, $J_{Num} \leftarrow x_{(i,j)} \leftarrow SI_{21\sim100}$
17. **for** each P_{Num} **do**
18. **if** $rand(1) < st$ **then**
19. Compute position of predators x_{i+1} using Equation (3)
20. else
21. Compute position of predators x_{i+1} using Equation (3)
22. **end**
23. **end**
24. **for** each J_{Num} **do**
25. **if** $i > 50$ **then**
26. Compute position of intrants x_{i+1} using Equation (4)
27. **else**
28. Compute position of intrants x_{i+1} using Equation (4)
29. **end**
30. **end**
31. **for** each S_{Num} **do**
32. **if** $f_i \neq f_g$ **then**
33. Compute position of intrants x_{i+1} using Equation (5)
34. **else**
35. Compute position of intrants x_{i+1} using Equation (5)
36. **end**
37. **end**
38. **for** each $x_{(i,j)}$, $i \in [1, 100]$, $j \in [1,2]$ **do**
39. Compute f_i using Equation (2)
40. **end**
41. $f_{avg} \leftarrow \sum f_i/100$
42. **for** each f_i **do**
43. **if** $f_i > f_{avg}$ **then**
44. Compute x_{new} using Equations (6) and (7)
45. Compute f_{new} using Equation (2)
46. **end**
47. **if** $f_{new} < f_i$ **then**
48. $f_i \leftarrow f_{new}$
49. $x_i \leftarrow x_{new}$
50. **end**
51. **end**
52. T ← T + 1
53. **end**
54. **for** each $x_{(i,j)}$ **do**
55. $W \leftarrow \left((x_{(i,1)} - y_{(1,1)}\right) * 0.6\ , \left(x_{(i,2)} - y_{(1,2)}\right) * 0.004)$
56. **end**
57. return $(L, N) \leftarrow x_{(i,j)} \leftarrow min\left(W_{(1,1)},\ W_{(1,2)}\right)$
58. **end**

2.5. Design of Fertilization-Rate Control System

The control system of bivariate fertilizer applicators uses the optimal combination of opening length and rotational speed corresponding to the target fertilization rate as the input signal. At the same time, the feedback signals are sampled by a linear displacement sensor (optoNCDT1420, Mirco-Epsilon Measurement Corporation, Ortenburg, Germany) and a rotary encoder (E6C2-CWZ5B, Omron Tateisi Electronics Corporation, Kyoto, Japan), respectively. The difference between the input signal and the feedback signal is input into the PID controller optimized with the improved BA, and then the PID-regulated signal is employed as the q-axis current to import the space vector pulse width modulation (SVPWM) algorithm module after the inverse park transformation. Eventually, the voltage output adjusted by the regulation of the switching sequence and pulse width of the voltage source inverter using the SVPWM module is input to the servo motors for obtaining the desired opening length and rotational speed (Figure 3).

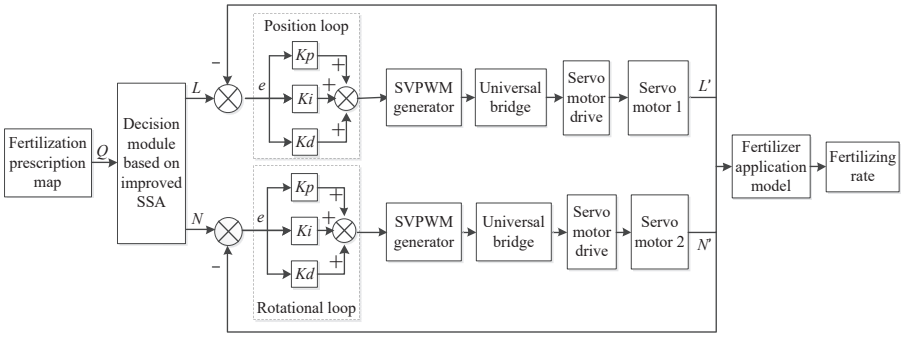


Figure 3. Control block diagram of regulation parameters.

In this study, the chosen servo motors are surface-mounted permanent magnet synchronous motors. Hence, the mathematical model in the synchronous rotating coordinate system *d*–*q* is selected to assemble the servo motor model, including the equations of the stator voltage, flux linkage, electromagnetic torque, and mechanical torque, as shown below [32]:

$$\begin{cases} u_d = Ri_d + \frac{d}{dt}\psi_d - \omega_e\psi_q \\ u_q = Ri_q + \frac{d}{dt}\psi_q + \omega_e\psi_d \end{cases} \tag{9}$$

$$\begin{cases} \psi_d = L_d i_d + \psi_f \\ \psi_q = L_q i_q \end{cases} \tag{10}$$

$$T_e = \frac{3}{2} p_n i_q \left[i_d \left(L_d - L_q \right) + \psi_f \right] \tag{11}$$

$$T_e - T_m = J_m \frac{d\omega_m}{dt} \tag{12}$$

where u_d and u_q are the *d*–*q* axis components of stator voltage, respectively; i_d and i_q are the *d*–*q* axis components of stator current; R indicates the resistance of the stator; T_e and T_m are the electromagnetic torque and the mechanical torque, respectively; J_m is the rotor inertia; ψ_d and ψ_q are the *d*–*q* axis components of stator flux, respectively; ω_e depicts the electromagnetic angular velocity; L_d and L_q are the *d*–*q* axis inductive components, respectively; and ψ_f represents the permanent magnet flux.

In addition, the stator inductance satisfies the case of $L_d = L_q = L_s$. Therefore, the stator voltage can be obtained from Formulas (9) and (10) as follows:

$$\begin{cases} u_d = Ri_d + L_d \frac{d}{dt} i_d - \omega_e L_q i_q \\ u_q = Ri_q + L_q \frac{d}{dt} i_q + \omega_e \left(L_d i_d + \psi_f \right) \end{cases} \tag{13}$$

Since the control variate is directly related to the angle and rotational speed of the motor output, the conversion relationship between electromagnetic angular velocity ω_e, electromagnetic angle θ_e, and motor speed N_r is described as follows [33]:

$$\begin{cases} \omega_e = P_n \omega_m \\ N_r = \frac{30}{\pi} \omega_m \\ \theta_e = \int \omega_e dt \end{cases} \tag{14}$$

where ω_m is the mechanical angular velocity of motor rad s^{-1}, P_n is the number of pole pairs, and N_r is the motor speed r min^{-1}.

Moreover, the baffle position depends on the rotation angle of regulation motor 1, and the rotational speed of the fertilizer discharging shaft is determined by the speed of servo motor 2. Thus, the related transfer functions of the motor control are obtained by executing the Laplace transformation after merging Formulas (11) to (14), as shown below:

$$G_n(s) = \frac{1}{\frac{2L_q J_m \pi}{3P_n \psi_f} s^2 + \frac{2RJ_m \pi}{3P_n \psi_f} s + \frac{4\pi \psi_f}{30}} \tag{15}$$

$$G_\theta(s) = \frac{1}{\frac{2J_m L_q}{3P_n \psi_f} s^3 + \frac{2J_m R}{3P_n \psi_f} s^2 + P_n \psi_f s} \tag{16}$$

Afterwards, the following functions are obtained by substituting the specific motor parameters.

$$\begin{cases} G_n(s) = \frac{1}{1.64e^{-6}s^2 + 5.12e^{-4}s + 0.0468488} \\ G_\theta(s) = \frac{1}{5.27e^{-7}s^3 + 1.63e^{-4}s^2 + 0.5595s} \end{cases} \tag{17}$$

Following the construction of the transfer functions of servomotor control, the tent mapping bat algorithm (TBA) is used to optimize the PID controller by adding a tent chaotic operator in the iteration procedure for promoting the global search capability. Furthermore, the three parameters (proportion K_p, integral K_i, and differential K_d) of the modified PID controller with the best control performance are solved with the minimum value of the ITAE criterion as the evaluation condition [34,35].

The position coordinates of individuals in TBA are composed of the PID controller parameters, and the frequency, position, and velocity of individual transformation are updated as follows [36]:

$$f_a = f_{min} + (f_{max} - f_{min}) z_i, \; f_a \in [0, 1] \tag{18}$$

$$v_i{}^t = v_i{}^{t-1} + \left(x_i{}^t - x_h\right) f_i \tag{19}$$

$$x_i{}^t = x_i{}^{t-1} + v_i{}^t \tag{20}$$

$$z_{i+1} = \begin{cases} 2z_i + rand(1) & \left(0 \le z \le \frac{1}{2}\right) \\ 2(1 - z_i) + rand(1) & \left(\frac{1}{2} < z \le 1\right) \end{cases} \tag{21}$$

where f_a is the frequency of the i-th bat, randomly assigned during initialization. $v_i{}^t$ denotes the velocity of the i-th bat at the t-th iteration, $x_i{}^t$ indicates the position of the i-th bat at the t-th iteration, x_h signifies the current global optimal position, and z_i express the chaotic operator.

In the optimization procedure of the algorithm, individuals perform the random search in the vicinity of the present optimal solution while conducting a local search, and the position is updated according to the equation as follows [37]:

$$x_{xin} = x_{old} + \varepsilon A^t \tag{22}$$

where x_{old} is the current optimal solution, x_{xin} is the updated solution, ε is a random number within $[-1, 1]$, A^t is the mean loudness of the bat population for the t-th iteration.

With the increase in iteration number, the individual loudness A_i gradually decreases, and the pulse emission frequency r constantly grows. The update operations are implemented as follows [38]:

$$A_i{}^{t+1} = \alpha A_i{}^t \tag{23}$$

$$r_i{}^{t+1} = r_i{}^b \left[1 - e^{-\gamma t}\right] \tag{24}$$

where a is the attenuation coefficient of loudness, γ is the increasing coefficient of pulse emission frequency, $r_i{}^b$ is the maximum of pulse emission frequency of the i-th bat.

The parameters of TBA are shown in Table 3. The detailed implementation process is shown in Algorithm 2.

Algorithm 2 PID parameter optimization by improved BA

Input: Variation range of PID parameters, the maximum iterations t_{max}, population number n, initial loudness A, initial pulse emissivity r_0, loudness attenuation coefficient α, increase coefficient of pulse emission frequency γ, population dimension d, initial velocity v, initial frequency f_a, position of individual X, upper bound U_b and lower bound L_b, ITAE criteria evaluation value F, optimal individual position X_{best}.
Output: Optimized PID parameters (k_p, k_i, k_d).
1. **for** i = 1:n **do**
2. $X_i \leftarrow L_b + (U_b - L_b).\times$ rand(1,d)
3. Call sim function to return ITAE value based on the X value $F \leftarrow \int_0^\infty t|e(t)|dt$
4. **end**
5. $X_{best} \leftarrow X_I \leftarrow [F_{min}, I] = \min(F)$
6. $t = 0$
7. **while** $t < t_{max}$ do
8. $i = t + 1$
9. $r_i = r_0 \times (1 - \mathrm{e}(-\gamma \times t))$
10. $A_i = \alpha \times A$
11. **for** j = 1:n **do**
12. Update frequency f_a according to Equations (18) and (21)
13. Update the velocity v according to Formula (19)
14. Update the position of individual X according to Equation (20)
15. **if** rand < r **do**
16. $X = X_{best} + 0.1 \times$ rand(1,d) $\times A_i$
17. **end**
18. $g = X < L_b$
19. $X_{new(g)} = L_{b(g)}$
20. $h = X > U_b$
21. $X_{new(g)} = U_{b(h)}$
22. Call sim function to return ITAE value based on the X_{new} value $F_{new} \leftarrow \int_0^\infty t|e(t)|dt$
23. **if** ($F_{new} < F$) & (rand > A) **do**
24. $X = X_{new}$
25. $F = F_{new}$
26. **end**
27. **if** $F_{new} <= F_{min}$
28. $X_{best} = X_{new}$
29. $F_{min} = F_{new}$
30. **end**
31. $t = t + 1$
32. **end**
33. **return** (k_p, k_i, k_d) $\leftarrow$ ($X_{best(1,1)}$, $X_{best(1,2)}$, $X_{best(1,3)}$)

Table 3. Parameter settings of TBA.

Algorithm	Parameter	Value
TBA	Maximum iterations t_{max}	200
	Population number n	30
	Initial loudness A	1
	Initial pulse emissivity r_0	1
	Loudness attenuation coefficient α	0.97
	Increase coefficient of pulse emission frequency γ	0.1
	Population dimension d	3
	Initial velocity v	0
	Initial frequency f_a	0
	Maximum frequency f_{max}	2
	Minimum frequency f_{min}	1
	Upper bound U_b	(100, 50, 50)
	Lower bound L_b	(0, 0, 0)
	Position coordinates of individuals	(k_p, k_i, k_d)

To test the performance of the PID controller designed for this study, the opening length and the rotational speed provided by the decision-making system are used as input signals and the simulation comparison with the PID control optimized based on BA, the conventional PID control, and fuzzy PID control, respectively, is implemented.

2.6. Assessment Criteria

Under the experimental condition of the same control system, the fertilization performance is verified through the test bench, and the target fertilization rate varies from 350 kg ha^{-1} to 600 kg ha^{-1} with an increment of 50 kg ha^{-1}. Meanwhile, the accuracy and convergence speed of the proposed decision-making algorithm, SSA, GA recommended by Yuan et al., and MOEA/D-DE suggested by Zhang et al. for control combination optimization are compared by the following evaluation criteria. The detailed parameters of GA and MOEA/D-DE are listed in Table 4.

Table 4. Parameter setting of GA and MOEA/D-DE.

Algorithm	Parameter	Value
GA	Size of population	100
	Chromosome number	2
	Maximum iteration	100
	Probability of mutation	0.1
	Probability of crossover	0.8
MOEA/D-DE	Size of population	100
	Number of targets	3
	Number of weight vectors	13
	Maximum iteration	100

(1) Accuracy of fertilization

The accuracy is defined as the relative error (RE) between the measured value and the target value of fertilization rate and is defined explicitly as below:

$$\mathrm{RE} = \frac{|y_s - y_g|}{y_g} \times 100\% \tag{25}$$

where y_s is the fertilization rate obtained from the actual experiment, and y_g is the target fertilization rate.

(2) Uniformity of fertilization

The fertilization rate corresponding to each combination of opening length and rotational speed optimized by these decision-making approaches is measured 10 times for 30 s utilizing the test bench. The uniformity is evaluated through the coefficient of variation (CV) to indicate the dispersion degree of fertilizer particles during the fertilization process as follows:

$$CV = \frac{y_{sd}}{y_m} \times 100\% \tag{26}$$

where y_{sd} is the standard deviation of fertilization rate data, and y_m is the average fertilization rate. In particular, y_{sd} can be calculated by:

$$y_{sd} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_m)^2}{n-1}} \tag{27}$$

where y_i is the fertilization rate sampled at the i-th measurement.

(3) Adjustment time

The maximum time needed to transition from the current fertilization rate to the following fertilization rate for modifying the opening length and rotational speed is used to describe the adjustment time. Obviously, the shorter the adjustment time is, the swifter the response of the fertilizer applicator will be.

To clarify the availability and effectiveness of the PID controller optimization using the improved BA, a series of comparisons with the fuzzy PID controller and the conventional PID controller is carried out by applying the opening length and rotational speed delivered from the decision-making method as the desired value. The specific descriptions of the assessment criteria are as follows.

(1) Accuracy

The accuracy of the controller is characterized as the steady-state deviation between the target value and the stable output value of the controller. The smaller the steady-state error value, the better the controller's control performance.

(2) Response speed

The response speed of the controller is defined as the rise time required for the controller output to increase from 0.1 to 0.9 times the target value. The shorter the rise time, the more rapidly the controller responds.

3. Results and Analysis

The average volume of fertilizer particles is 19.835 mm^3 through the measurement and calculation of stochastically chosen samples, and the distribution proportions of the ratio of relative volume are exhibited in Figure 4. It can be noticed that 72% of the selected fertilizer particles are within the range of 0.5–1.5, and these fertilizer particles are relatively regular and nearly ellipsoid in shape. On the other hand, 19% of the fertilizer particles are in the scope of 1.5–2.5 and much larger than the average volume. This phenomenon is caused by the irregular shape of fertilizer particles with bumps. In addition, 9% of the fertilizer particles are in the span of 0–0.5, and this portion usually has an anomalous cross-section.

The other physical characteristic parameters of fertilizer particles required in EDEM are gauged. Concretely, the density is 1.7163 g cm^{-3}, the average angle of repose is 29.8°, and the moisture content is 5.278%. Meanwhile, the static friction coefficient between fertilizer particles and steel plate is 0.4592, and the static friction coefficient between fertilizer particles is 0.4998. In addition, the collision restitution coefficient between the fertilizer particle and steel plate is 0.427, and the collision restitution coefficient between fertilizer particles is 0.216.

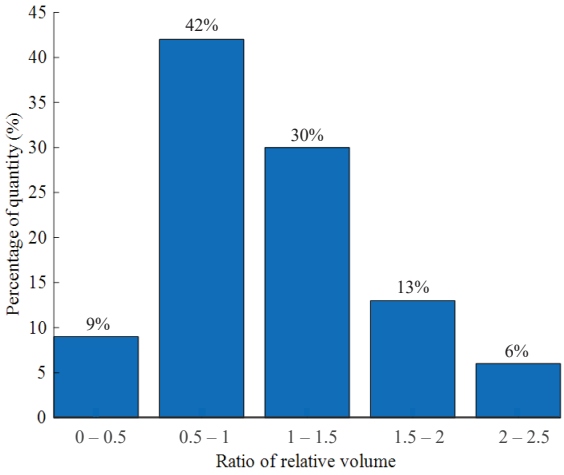


Figure 4. Distribution of ratio of relative volume for fertilizer particles.

The experimental results of fertilization performance using EDEM at small opening lengths are shown in Table 5. It can be observed that the fertilization suspension occurs in different periods of the whole simulation process for the opening length of 5 mm, 6 mm, and 7 mm. In detail, more periods of fertilizer flow interruption emerge in the case of 5 mm, and the undesirable circumstance decreases with the growth of opening length. Contrarily, the fertilization process is smooth without discontinuous fertilizer flow at the opening length of 8 mm. The fertilization suspension in the simulation is induced by the extrusion and accumulation of irregularly shaped fertilizer particles at the position of the fertilizer outlet. This phenomenon can also boost the friction between fertilizer particles, resulting in a decline in the accuracy and uniformity of fertilization. Therefore, the minimum allowable opening length is selected as 8 mm.

Table 5. Fertilization capability for various opening lengths.

Opening Length (mm)	Time (s)	Fertilization Weight (g)
5	0–0.2	0.177
	0.2–1.2	0
	1.2–1.5	0.233
	1.5–3	0
6	0–0.4	0.362
	0.4–1.55	0
	1.55–1.8	0.215
	1.8–2.4	0.147
	2.4–3	0
7	0–0.45	0.609
	0.45–1.05	0
	1.05–1.35	0.296
	1.35–2.75	0
	2.75–3	0.158
8	0–3	22.57

The simulation results of fertilization rate for the variable gap between the spiral blades and the discharge cavity are displayed in Table 6. In the simulation, the deviation of fertilization rate at the gap of 5 mm compared with the gap of 3 mm and 4 mm is 1.22 g and 1.51 g, respectively, the variation of that at the gap of 6 mm compared to the gap of 3 mm and 4 mm is 4.28 g to 4.57 g, and the fertilization-rate difference between the gap of

3 mm and the gap of 4 mm is 0.29 g. It can be seen that the fertilization rate drops with the boost of the gap value, and the fertilization rate is the minimum at the gap of 6 mm. This is because the diameter of fertilizer particles varying from 3 to 5 mm is less than the 6 mm gap value, meaning some fertilizer particles cannot be effectively propelled by the blade of the fertilizer discharging shaft. At the same time, the small gap (3 mm) can also reduce the transportation efficiency of the fertilizer discharging shaft. Thus, the gap value is designed to be 4 mm in the subsequent manufacturing process.

Table 6. Fertilization performance under different gaps between the spiral blades and the discharge cavity.

Gap Value (mm)	Fertilization Rate (g)
3	42.89
4	43.18
5	41.67
6	38.61

The fitting model of fertilization rate obtained by polynomial fitting is illustrated as follows:

$$P = -0.1377 * L^2 - 0.006814 * N^2 + 0.0384 * L * N + 3.984 * L + 4.767 * N - 27.33 \quad (28)$$

where P signifies the fertilization rate for the 30 s test period (g).

A total of 28 sets of data are randomly selected from the experimental fertilization data (Table 7) to verify the accuracy of the fitting model. The opening lengths and rotational speeds of these test samples are imported into the model to achieve the predicted fertilization rates and then compared with the corresponding actual fertilization rates in the test data (Figure 5). It can be found that the experimental fertilization rate under 400 g is remarkably proximate to the forecasted data of the model, while the actual fertilization-rate data above 400 g are slightly distinct from the predicted data of the model. This deviation is due to the manual adjustment of the polynomial second power and the coefficients before the variables for covering the fertilization-rate range suggested by agricultural specialists (150–250 kg ha^{-1}) through enhancing the accuracy of fertilizer application below 400 g [39]. In addition, MRE is 1.704%, R^2 is 0.9977, indicating that the fitting model has tiny prediction variation and noteworthy accuracy.

Table 7. Test set data.

No	Opening Length (mm)	Rotational Speed (r min^{-1})	Fertilization Rate (g 30^{-s})	No	Opening Length (mm)	Rotational Speed (r min^{-1})	Fertilization Rate (g 30^{-s})
1	8	50	230.55	15	18	150	667.55
2	8.5	80	355	16	18	110	507.35
3	9	130	532.85	17	20	10	52.9
4	9.5	20	99.55	18	20	50	255.5
5	10	60	285.7	19	20	100	498.6
6	10.5	90	404.55	20	22	140	659.7
7	11	30	151.2	21	22	20	102.4
8	11.5	70	328.05	22	22	90	425.7
9	12	40	197.45	23	24	40	205.55
10	13	120	554.1	24	24	50	258.35
11	14	40	199	25	26	80	399.6
12	15	60	294.35	26	26	40	206.25
13	16	80	384.1	27	28	100	503.6
14	16	30	151.11	28	28	30	153.3

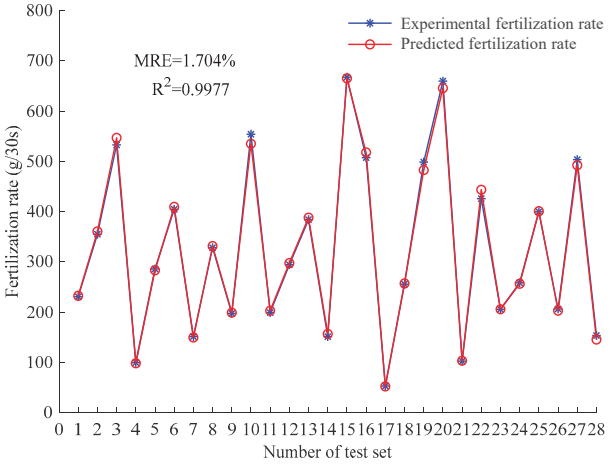


Figure 5. Comparison between test set data and bench test data.

The control combinations of the target fertilization rates optimized by these four optimization algorithms and the corresponding predicted fertilization rates acquired from the fitted model are shown in Table 8. The maximum difference between the forecasted value corresponding to the control combination offered by the improved SSA, SSA, GA, and MOEA/D-DE and the target value is 6.2678 kg ha^{-1}, 9.1940 kg ha^{-1}, 2.0431 kg ha^{-1}, and 16.3877 kg ha^{-1}, respectively. Meanwhile, the minimum difference of that is 1.0756 kg ha^{-1}, 2.0088 kg ha^{-1}, 0.0233 kg ha^{-1}, and 8.9146 kg ha^{-1}, respectively, and the average difference of that is 3.9587 kg ha^{-1}, 5.6964 kg ha^{-1}, 0.6444 kg ha^{-1}, and 12.7799 kg ha^{-1}, respectively. The relative error of the difference relative to the target fertilization rate is 0.86%, 1.27%, 0.16%, and 2.69%. It can be seen that the optimization effect of GA is the best, while that of MOEA/D-DE is the worst. This is due to the diverse operation modes of these algorithms. GA pays more attention to accuracy. Conversely, the method used in this study, SSA, and MOEA/D-DE focus on the response time and weight parameter selection in control combination determination and is beneficial to the feasibility and practicality of application.

The convergence process of these four algorithms is depicted in Figure 6. In the case that the iteration value equals 1, the fitness values of the proposed method, SSA, GA, and MOEA/D-DE are 2.642, 3.786, 6.284, and 34.1, respectively. When the fitness value reaches 0.045, the iteration values of the proposed method, SSA, GA, and MOEA/D-DE are 3, 12, 37, and 68, respectively. It can be noticed that the convergence speed of the proposed method is the fastest, and the MOEA/D-DE algorithm is the slowest. This is caused by the improvement of the SSA using a chaotic operator and mutation section of the DE algorithm that can reduce invalid iterations in the case of falling into the local optimal solution in the global optimization and accelerate the convergence speed.

Taking the opening length and rotational speed offered by the decision-making algorithm when the target fertilization rate increases from 350 kg ha^{-1} to 600 kg ha^{-1} in increments of 50 kg ha^{-1} as the desired values, the average accuracy of opening length of the PID controller optimized by TBA and the PID controller optimized by BA, the traditional PID controller, and the fuzzy PID controller is 0.004 mm, 0.057mm, 0.166 mm, and 0.120 mm, respectively. The average accuracy of the rotational speed of these four controllers is 0 r min^{-1}, 0.193 r min^{-1}, 1.102 r min^{-1}, and 0.403 r min^{-1}, respectively. In particular, for the target opening length of 8.5 mm, the response speed of output boosted from 0.85 mm to 7.65 mm is 0.468 s, 0.497 s, 0.514 s, and 0.648 s, respectively, and the steady-state deviations are 0.01 mm, 0.05 mm, 0.14 mm, and 0.12 mm, respectively. Moreover, for the target rotational speed of 77 r min^{-1}, the response speed of output increased from 7.7 r min^{-1} to 69.3 r min^{-1} is 0.0008 s, 0.00083 s, 0.00088 s, and 0.00088 s, respectively (Figure 7). It can be discovered that the PID controller based on TBA optimization can

accurately follow the shifts of the decision-making method. Therefore, compared with the PID controller based on BA, the traditional PID controller adjusting parameters through manual experience, and the fuzzy PID controller applying fuzzy rules for parameter tuning, the proposed PID controller in this study has a more acceptable control effect, which is slightly better than the controller optimized by BA.

Table 8. Comparisons of four optimization algorithms for control combination.

Target Fertilization Rate (kg ha^{-1})	Algorithm	Opening Length (mm)	Rotational Speed (r min^{-1})	Estimated Fertilization Rate (kg ha^{-1})	Absolute Error (kg ha^{-1})
350	Improved SSA	8.5	77	348.3768	1.6232
	SSA	9.1	79	359.1940	9.1940
	GA	10.3	75	347.9569	2.0431
	MOEA/D-DE	8.2	80	359.0207	9.0207
400	Improved SSA	9.1	88	395.0006	4.9994
	SSA	10.5	87	397.5528	2.4472
	GA	13.2	85	400.3146	0.3146
	MOEA/D-DE	9.2	85	389.6605	10.3395
450	Improved SSA	8.5	105	456.2678	6.2678
	SSA	14.8	97	454.8845	4.8845
	GA	11.6	99	449.603	0.397
	MOEA/D-DE	9.4	95	433.6123	16.3877
500	Improved SSA	9.6	117	505.819	5.819
	SSA	9.9	113	493.2376	6.7624
	GA	15.1	108	499.4116	0.5884
	MOEA/D-DE	16	105	491.0854	8.9146
550	Improved SSA	10.5	127	546.033	3.967
	SSA	12.1	123	541.1185	8.8815
	GA	14.7	122	550.5002	0.5002
	MOEA/D-DE	12.5	115	534.2442	15.7558
600	Improved SSA	11.8	141	601.0756	1.0756
	SSA	20.1	130	602.0088	2.0088
	GA	10.6	143	599.9767	0.0233
	MOEA/D-DE	10.6	135	583.7387	16.2613

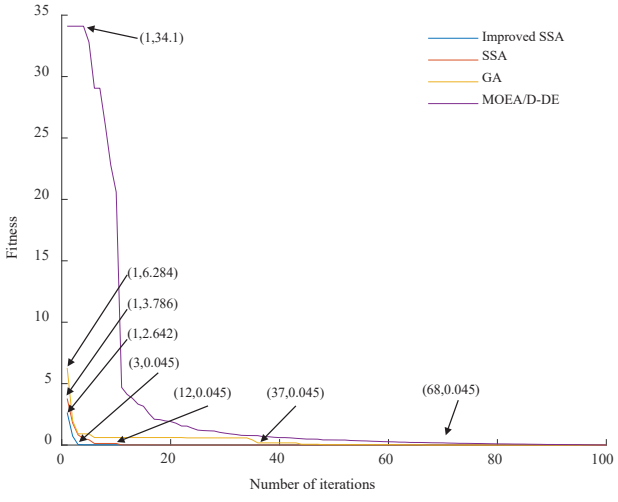
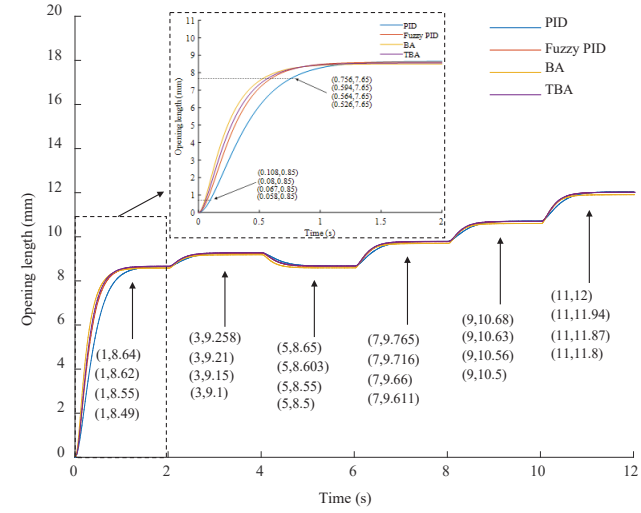
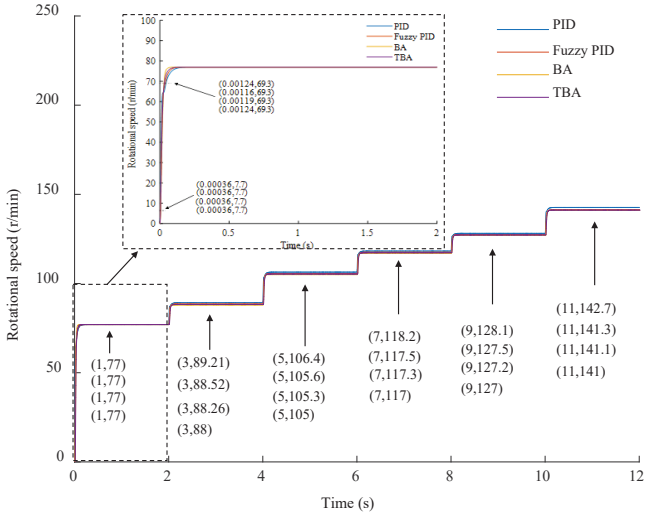


Figure 6. Iteration process of decision-making algorithms.



(**a**)



(**b**)

Figure 7. Comparison of four different PID controllers for fertilization rate varying from 350 kg ha^{-1} to 600 kg ha^{-1}: (**a**) Comparison of continuous opening-length control of PID, fuzzy PID, PID improved by BA, and PID optimized by TBA; (**b**) comparison of continuous rotational-speed control of PID, fuzzy PID, PID improved by BA, and PID optimized by TBA.

The corresponding simulation results of fertilization rate under the same conditions are demonstrated in Figure 8. For the variable target fertilization rate, the average accuracy of the controller designed in this study, the PID controller improved by BA, the PID controller, and the fuzzy PID controller is 4.13 kg ha^{-1}, 4.15 kg ha^{-1}, 5.58 kg ha^{-1}, and 4.32 kg ha^{-1}, respectively. Obviously, the control performance of the PID controller optimized through TBA has a remarkably suitable outcome in terms of accuracy of fertilization and response speed. Especially, the rise time of fertilization rate, increased from 35 kg ha^{-1} (0.1 × 350 kg ha^{-1}) to 315 kg ha^{-1} (0.9 × 350 kg ha^{-1}), of the controller developed in this study, the PID controller improved by BA, the PID controller, and the fuzzy PID controller

is 0.112 s, 0.13 s, 0.202 s, and 0.15 s, respectively. It can be found that the response speed of the proposed control system is optimal. In addition, the fertilization-rate adjustment is achieved by simultaneously regulating the opening length and rotational speed, and the rotational speed changes within the same time. Furthermore, in the short time that the fertilization rate grows from 35 kg ha^{-1} to 300 kg ha^{-1}, the rotational speed can reach the target speed. However, the opening length is far from approximating the goal value, and the time used to boost from 300 kg ha^{-1} to 350 kg ha^{-1} is mainly generated by the opening-length adjustment. Therefore, the influence of the opening length on the equipment adjustment time is much more significant.

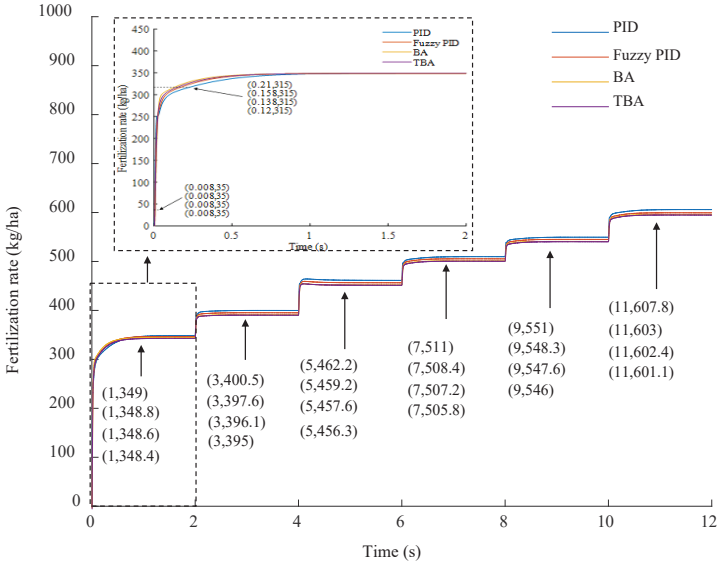


Figure 8. Simulation of fertilization-rate regulation from 350 kg ha^{-1} to 600 kg ha^{-1}.

Under the same control system conditions, the indoor bench test is used to compare the performance of the control combination derived from these four decision-making methods on the bivariate fertilizer applicator. The indoor experimental data are indicated in Table 9.

As shown in Table 6, the average accuracy of the proposed method, SSA, GA, and MOEA/D-DE method is 1.8%, 1.9%, 2.5%, and 3.5%, respectively. The results suggest that the method proposed in this study performs most satisfactorily in terms of accuracy of fertilization. In addition, the average uniformity of the proposed method, SSA, GA, and MOEA/D-DE is 0.47%, 0.47%, 0.52%, and 0.48%, respectively, indicating that the proposed method had better uniformity than the other three methods, and the fluctuation of the fertilization process is slight. Moreover, since the indoor test is carried out with the same equipment, the unit adjustment time of the two variables is identical. The average adjustment time of the proposed method, SSA, GA, and MOEA/D-DE is 0.5 s, 0.99s, 1.48 s, and 1.34 s, respectively. Thus, the adjustment time of the method used in this study is more reasonable.

Table 9. Comparison of indoor tests of four decision-making algorithms.

Target Fertilization Rate (kg ha^{-1})	Methods	Actual Fertilization Rate (kg ha^{-1})	Accuracy of Fertilization	Uniformity of Fertilization	Adjustment Time (s)
350	Improved SSA	355	1.5%	0.53%	0.3
	SSA	358.62	2.5%	0.51%	0.42
	GA	358.93	2.5%	0.5%	1.38
	MOEA/D-DE	358.06	2.3%	0.42%	0.12
400	Improved SSA	388.58	2.9%	0.33%	0.36
	SSA	391.16	2.2%	0.37%	0.32
	GA	386.58	3.4%	0.51%	1.74
	MOEA/D-DE	383.43	4.1%	0.43%	0.6
450	Improved SSA	459.15	2.3%	0.4%	0.36
	SSA	457.06	1.6%	0.42%	0.89
	GA	471.15	4.7%	0.61%	0.96
	MOEA/D-DE	431.8	4.3%	0.43%	0.12
500	Improved SSA	504.84	1%	0.51%	0.66
	SSA	494.17	1.2%	0.49%	1.03
	GA	494.3	1.1%	0.53%	2.1
	MOEA/D-DE	494.63	1.1%	0.54%	3.96
550	Improved SSA	540.55	1.7%	0.6%	0.54
	SSA	538.93	2%	0.54%	0.47
	GA	561.9	2.2%	0.58%	0.24
	MOEA/D-DE	522.25	5%	0.48%	2.1
600	Improved SSA	609.8	1.6%	0.43%	0.78
	SSA	611.24	1.9%	0.49%	2.82
	GA	607.97	1.3%	0.41%	2.46
	MOEA/D-DE	574.3	4.3%	0.55%	1.14

4. Conclusions

In this study, a decision-making capability optimization scheme of control combination and PID controller parameters is proposed for boosting the operation performance of a bivariate fertilizer applicator. To be specific, the gap between the spiral blades and the discharge cavity wall is optimized by EDEM to guarantee the integrity of fertilizer particles and the accuracy of fertilization. Meanwhile, the adjustment of the effective opening-length range guided by EDEM can avoid the blockage of fertilizer particles and is beneficial to the uniformity of fertilization. Moreover, to promote fertilization performance and overcome the deficiency of real-time detection feedback, SSA is improved to strengthen the global search ability by introducing a chaotic operator and mutation section of the DE algorithm for the determination of the optimal control combination. In addition, ITAE used as the evaluation criterion of the optimization effect of PID controller parameters is beneficial to promote the accuracy and response time of the controller significantly. Meanwhile, the PID controller optimized by TBA can effectively cope with the continuous deviation of control variables to enhance comprehensive fertilization performance.

Under the condition of utilizing the same controller, the decision-making method proposed in this study is compared with GA and MOEA/D-DE, and it is found that the accuracy, uniformity, and adjustment time of the control combination acquired by the presented method is the best. The decision-making capability optimization scheme can provide a valuable reference for promoting the working effect of the multi-variable fertilizer applicator.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/agriculture12122100/s1.

Author Contributions: Conceptualization, Y.D.; methodology, G.Y. and J.W.; software, Z.Z.; validation, Z.Z. and Z.X.; formal analysis, G.Y.; investigation, Y.D. and J.W.; resources, J.W.; data curation, Z.X. and G.Y.; writing—original draft preparation, Y.D., G.Y. and J.W.; writing—review and editing, G.Y. and J.W.; visualization, G.Y.; supervision, J.W.; project administration, J.W.; funding acquisition, Y.D. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Program for Science & Technology Innovation Talents in Universities of Henan Province (Grant no. 20HASTIT029), Key Scientific Research Projects in Universities of Henan Province (Grant no. 19A460021), and Key Science and Technology Project of Henan Province (Grant no. 2221022102164, 212102210352).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Qin, J.; Impa, S.; Tang, Q.; Yang, S.; Yang, J.; Tao, Y.; Jagadish, K.S. Integrated nutrient, water and other agronomic options to enhance rice grain yield and N use efficiency in double-season rice crop. *Field Crops Res.* **2013**, *148*, 15–23. [CrossRef]
2. Reidsma, P.; Feng, S.; van Loon, M.; Luo, X.; Kang, C.; Lubbers, M.; Kanellopoulos, A.; Wolf, J.; van Ittersum, M.K.; Qu, F. Integrated assessment of agricultural land use policies on nutrient pollution and sustainable development in Taihu Basin, China. *Environ. Sci. Policy* **2012**, *18*, 66–76. [CrossRef]
3. Shi, Y.; Zhu, Y.; Wang, X.; Sun, X.; Ding, Y.; Cao, W.; Hu, Z. Progress and development on biological information of crop phenotype research applied to real-time variable-rate fertilization. *Plant Methods* **2020**, *16*, 11. [CrossRef] [PubMed]
4. Jafari, M.; Hemmat, A.; Sadeghi, M. Development and performance assessment of a DC electric variable-rate controller for use on grain drills. *Comput. Electron. Agric.* **2010**, *73*, 56–65. [CrossRef]
5. Chen, C.; He, P.; Zhang, J.; Li, X.; Ren, Z.; Zhao, J.; He, J.; Wang, Y.; Liu, H.; Kang, J. A fixed-amount and variable-rate fertilizer applicator based on pulse width modulation. *Comput. Electron. Agric.* **2018**, *148*, 330–336. [CrossRef]
6. Zinkevičienė, R.; Jotautienė, E.; Juostas, A.; Comparetti, A.; Vaiciukevičius, E. Simulation of granular organic fertilizer application by centrifugal spreader. *Agronomy* **2021**, *11*, 247. [CrossRef]
7. Zhu, Q.; Wu, G.; Chen, L.; Zhao, C.; Meng, Z. Influences of structure parameters of straight flute wheel on fertilizing performance of fertilizer apparatus. *Nongye Gongcheng Xuebao/Trans. Chin. Soc. Agric. Eng.* **2018**, *34*, 12–20.
8. Song, L.; Wen, L.; Chen, Y.; Lan, Y.; Zhang, J.; Zhu, J. Variable-rate fertilizer based on a fuzzy PID control algorithm in coastal agricultural area. *J. Coast. Res.* **2020**, *103*, 490–495. [CrossRef]
9. Xu, Y.; Zhang, X.; Wu, S.; Chen, C.; Wang, J.; Yuan, S.; Chen, B.; Li, P.; Xu, R. Numerical simulation of particle motion at cucumber straw grinding process based on EDEM. *Int. J. Agric. Biol. Eng.* **2020**, *13*, 227–235. [CrossRef]
10. Yang, L.; Chen, L.; Zhang, J.; Liu, H.; Sun, Z.; Sun, H.; Zheng, L. Fertilizer sowing simulation of a variable-rate fertilizer applicator based on EDEM. *IFAC Pap.* **2018**, *51*, 418–423. [CrossRef]
11. Chen, H.; Zheng, J.; Lu, S.; Zeng, S.; Wei, S. Design and experiment of vertical pneumatic fertilization system with spiral Geneva mechanism. *Int. J. Agric. Biol. Eng.* **2021**, *14*, 135–144. [CrossRef]
12. Tang, H.; Jiang, Y.; Wang, J.; Wang, J.; Zhou, W. Numerical analysis and performance optimization of a spiral fertilizer distributor in side deep fertilization of a paddy field. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.* **2021**, *235*, 3495–3505. [CrossRef]
13. Liu, C.; Yuan, J.; Liu, J.; Li, C.; Zhou, Z.; Gu, Y. ARM and DSP-based bivariable fertilizing control system design and implementation. *Trans. Chin. Soc. Agric. Mach.* **2010**, *41*, 233–238.
14. Alameen, A.A.; Al-Gaadi, K.A.; Tola, E. Development and performance evaluation of a control system for variable rate granular fertilizer application. *Comput. Electron. Agric.* **2019**, *160*, 31–39. [CrossRef]
15. Fulton, J.P.; Shearer, S.A.; Higgins, S.F.; Darr, M.J.; Stombaugh, T.S. Rate response assessment from various granular VRT applicators. *Trans. ASAE* **2005**, *48*, 2095–2103. [CrossRef]
16. Yuan, J.; Liu, C.-L.; Li, Y.-M.; Zeng, Q.; Zha, X.F. Gaussian processes based bivariate control parameters optimization of variable-rate granular fertilizer applicator. *Comput. Electron. Agric.* **2010**, *70*, 33–41. [CrossRef]
17. Zhang, J.; Liu, G.; Luo, C.; Hu, H.; Huang, J. MOEA/D-DE based bivariate control sequence optimization of a variable-rate fertilizer applicator. *Comput. Electron. Agric.* **2019**, *167*, 105063. [CrossRef]
18. Xue, J.; Shen, B. A novel swarm intelligence optimization approach: Sparrow search algorithm. *Syst. Sci. Control. Eng.* **2020**, *8*, 22–34. [CrossRef]
19. Nguyen, T.-T.; Ngo, T.-G.; Dao, T.-K.; Nguyen, T.-T.-T. Microgrid Operations Planning Based on Improving the Flying Sparrow Search Algorithm. *Symmetry* **2022**, *14*, 168. [CrossRef]
20. Wang, P.; Zhang, Y.; Yang, H. Research on economic optimization of microgrid cluster based on chaos sparrow search algorithm. *Comput. Intell. Neurosci.* **2021**, *2021*, 1–18. [CrossRef]

21. Kathiroli, P.; Selvadurai, K. Energy efficient cluster head selection using improved Sparrow Search Algorithm in Wireless Sensor Networks. *J. King Saud Univ. Comput. Inf. Sci.* **2021**, *34*, 8564–8575. [CrossRef]
22. GirirajKumar, S.; Jayaraj, D.; Kishan, A.R. PSO based tuning of a PID controller for a high performance drilling machine. *Int. J. Comput. Appl.* **2010**, *1*, 12–18. [CrossRef]
23. Zhang, J.; Hou, S.; Wang, R.; Ji, W.; Zheng, P.; Wei, S. Design of variable-rate liquid fertilization control system and its stability analysis. *Int. J. Agric. Biol. Eng.* **2018**, *11*, 109–114. [CrossRef]
24. Bai, J.; Tian, M.; Li, J. Control System of Liquid Fertilizer Variable-Rate Fertilization Based on Beetle Antennae Search Algorithm. *Processes* **2022**, *10*, 357. [CrossRef]
25. Yang, X. A new metaheuristic bat-inspired algorithm. In *Nature Inspired Cooperative Strategies for Optimization (NICSO 2010)*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 65–74.
26. Chaib, L.; Choucha, A.; Arif, S. Optimal design and tuning of novel fractional order PID power system stabilizer using a new metaheuristic Bat algorithm. *Ain Shams Eng. J.* **2017**, *8*, 113–125. [CrossRef]
27. Qinghua, M.; Qiang, Z. Improved sparrow algorithm combining Cauchy mutation and Opposition-based learning. *J. Front. Comput. Sci. Technol.* **2021**, *15*, 1155.
28. Zhu, Y.; Yousefi, N. Optimal parameter identification of PEMFC stacks using adaptive sparrow search algorithm. *Int. J. Hydrog. Energy* **2021**, *46*, 9541–9552. [CrossRef]
29. Gao, B.; Shen, W.; Guan, H.; Zheng, L.; Zhang, W. Research on multistrategy improved evolutionary sparrow search algorithm and its application. *IEEE Access* **2022**, *10*, 62520–62534. [CrossRef]
30. Yi, X. Hash function based on chaotic tent maps. *IEEE Trans. Circuits Syst. II Express Briefs* **2005**, *52*, 354–357.
31. Qin, A.K.; Huang, V.L.; Suganthan, P.N. Differential evolution algorithm with strategy adaptation for global numerical optimization. *IEEE Trans. Evol. Comput.* **2008**, *13*, 398–417. [CrossRef]
32. Feng, Y.; Zheng, J.; Yu, X.; Truong, N.V. Hybrid terminal sliding-mode observer design method for a permanent-magnet synchronous motor control system. *IEEE Trans. Ind. Electron.* **2009**, *56*, 3424–3431. [CrossRef]
33. Li, S.; Liu, Z. Adaptive speed control for permanent-magnet synchronous motor system with variations of load inertia. *IEEE Trans. Ind. Electron.* **2009**, *56*, 3050–3059.
34. Gu, Y.; Yuan, J.; Liu, C. FIS-based method to generate bivariate control parameters regulation sequence for fertilization. *Trans. Chin. Soc. Agric. Eng.* **2011**, *27*, 134–139.
35. Zhang, Y.Z.; Chen, H.T.; Hou, S.Y.; Ji, W.Y.; Ouyang, B.L.; Guo-Qiang, D.; Zhang, J.C. Design and Experiment of Slave Computer Control System for Applying Variable-rate Liquid Fertilizer. *J. Northeast. Agric. Univ.* **2015**, *22*, 73–79.
36. Mishra, S.; Shaw, K.; Mishra, D. A new meta-heuristic bat inspired classification approach for microarray data. *Procedia Technol.* **2012**, *4*, 802–806. [CrossRef]
37. Hasançebi, O.; Teke, T.; Pekcan, O. A bat-inspired algorithm for structural optimization. *Comput. Struct.* **2013**, *128*, 77–90. [CrossRef]
38. Jaddi, N.S.; Abdullah, S.; Hamdan, A.R. Optimization of neural network model using modified bat-inspired algorithm. *Appl. Soft Comput.* **2015**, *37*, 71–86. [CrossRef]
39. Meng Yao, H.; Zhang, L.; Zhi Wen, W.; Dian Lin, Y.; Li Li, W.; Wei Ming, X.; Jian Ning, Z. Estimation of fertilizer usage from main crops in China. *J. Agric. Resour. Environ.* **2017**, *34*, 360.

agriculture

MDPI

Article

Prediction of Protein Content in Pea (*Pisum sativum* L.) Seeds Using Artificial Neural Networks

Patryk Hara [1], Magdalena Piekutowska [2] and Gniewko Niedbała [3,*]

[1] Agrotechnology, Jagiellonów 4, 73-150 Łobez, Poland
[2] Department of Geoecology and Geoinformation, Institute of Biology and Earth Sciences, Pomeranian University in Słupsk, 27 Partyzantów St., 76-200 Słupsk, Poland
[3] Department of Biosystems Engineering, Faculty of Environmental and Mechanical Engineering, Poznań University of Life Sciences, Wojska Polskiego 50, 60-627 Poznań, Poland
* Correspondence: gniewko.niedbala@up.poznan.pl

Abstract: Pea (*Pisum sativum* L.) is a legume valued mainly for its high seed protein content. The protein content of pea is characterized by a high lysine content and low allergenicity. This has made consumers appreciate peas increasingly in recent years, not only for their taste, but also for their nutritional value. An important element of pea cultivation is the ability to predict protein content, even before harvest. The aim of this research was to develop a linear and a non-linear model for predicting the percentage of protein content in pea seeds and to perform a comparative analysis of the effectiveness of these models. The analysis also focused on identifying the variables with the greatest impact on protein content. The research included the method of machine learning (artificial neural networks) and multiple linear regression (MLR). The input parameters of the models were weather, agronomic and phytophenological data from 2016–2020. The predictive properties of the models were verified using six ex-post forecast measures. The neural model (N1) outperformed the multiple regression (RS) model. The N1 model had an RMS error magnitude of 0.838, while the RS model obtained an average error value of 2.696. The MAPE error for the N1 and RS models was 2.721 and 8.852, respectively. The sensitivity analysis performed for the best neural network showed that the independent variables most influencing the protein content of pea seeds were the soil abundance of magnesium, potassium and phosphorus. The results presented in this work can be useful for the study of pea crop management. In addition, they can help preserve the country's protein security.

Keywords: artificial neural networks; multiple linear regression; protein prediction; pea; sensitivity analysis; weather conditions

Citation: Hara, P.; Piekutowska, M.; Niedbała, G. Prediction of Protein Content in Pea (*Pisum sativum* L.) Seeds Using Artificial Neural Networks. *Agriculture* **2022**, *13*, 29. https://doi.org/10.3390/agriculture13010029

Academic Editor: Hongbin Pu

Received: 6 November 2022
Revised: 12 December 2022
Accepted: 19 December 2022
Published: 22 December 2022

1. Introduction

In terms of cultivation, legumes are the world's second largest crop after cereals. They constitute about 30% of the world's plant production [1]. The crop residues of these plants are characterized by a positive organic matter (MO) balance, which makes them a very good forecrop for many crops such as cereals, potatoes and beets. In the years with an uneven distribution of precipitation (temperate climates) or a shortage of precipitation (southern European climatic conditions), when there is a poor uptake of mineral nitrogen, a particularly favorable after-effect of legumes is observed [2,3]. The decomposition of *Fabaceae* crop residues in the soil provides available forms of nitrogen to both successor plants and soil microorganisms.

This process contributes to the intensification of biological nitrogen sorption. The interaction between the legume residues and MO mineralization determines the amount of available N for the next plant [4]. The participation of legumes in crop rotation also contributes to the reduction of weed, pest and disease populations [5]. When grown in rotation with cereals, they counteract soil erosion and improve soil fertility [6] by changing physical and chemical properties. A properly developed root system of legumes is formed

through loosening the soil, which increases soil aeration and becomes a source of a large amount of organic matter rich in nitrogen and other minerals [7]. A significant property of this group of plants is the ability to fix atmospheric nitrogen as a result of symbiosis with *Rhizobium* spp. bacteria. This trait plays an important role in argoecosystems and sustainable crop production, which seeks to reduce the use of mineral fertilizers [8,9]. The symbiosis of legumes with papillary bacteria also reduces inputs and resources by reducing the need for nitrogen fertilizers [10]. The importance of these plants is particularly important in an era of high and volatile mineral fertilizer prices. In many European countries, these fertilizers are periodically becoming a scarce commodity, so legumes are an important element in countering the fertilizer crisis and fitting in with the ideas of sustainable agriculture.

One of the most important plants in the *Fabaceae* family is the pea. More than half of its world production is in Canada, Russia, the United States of America and India [11]. In Poland, peas are the most widely grown legume right after yellow lupin. In 2022, the area of pea cultivation in Poland was more than 105 thousand hectares. For example, the cultivation of soybeans and beans in the same year amounted to 48.20 and 30.70 thousand hectares, respectively. In 2022, peas accounted for 0.68% of the total area in the structure of national sowings [12].

Peas are most valued for their high seed protein content, which can be as high as 31% [13,14]. Pea protein has high nutritional value due to its relatively high content of lysine, an amino acid that limits the nutritional value of cereals. In addition, it is characterized by low allergenicity [15,16]. Despite its significant and obvious advantages, the area under cultivation for this crop is relatively low due to poor profitability as a result of biotic-abiotic factors [17]. However, there is growing interest in plant-based proteins as a substitute for animal-based proteins [18],. The reason for this phenomenon is greater awareness of nutrition, environmental concerns and ethical issues [11]. Therefore, it can be assumed that peas will become increasingly popular among farmers over the next few years, which will be reflected in the increasing area under cultivation.

The demand for protein will continue to grow in the coming years due to the world's expanding human population [1]. This makes it likely that interest in legumes, including peas as a valuable source of protein, will be greater than before. In addition, peas are a very good component in feed production. As a protein-raw material with a satisfactory amino acid composition, it is used in the feeding of slaughter, dairy and laying animals without adversely affecting production and fattening performance [19]. For many years, efforts have been made in European countries, including Poland, to increase the production and use of domestic protein raw materials to replace, or at least supplement, expensive imported post-extraction soybean meal. These measures are also aimed at preserving the country's protein security, as the feed market is largely dependent on imported protein raw materials. Therefore, there is a risk (for now only theoretical) of a shortage of protein feed for animals, and consequently a shortage of food for the population [20]. Therefore, the ability to predict the protein content of pea seeds is very important for the possibility of ongoing decision support, management of national protein resources and risk management [21]. However, prediction of crop quality traits is a very difficult task. During the growing season, plants are exposed to a number of factors that limit both yield and quality, and predicting many of these factors is often an impossible task [22]. The non-uniform course of weather conditions, soil variability or pest pressure causes the growth and development of agricultural crops to proceed differently in each growing season [23]. In addition, the non-linear interaction between environmental factors and plant growth can result in a low-precision predictive model with a large prediction error [24,25]. The large number of factors influencing the quality of yield poses a significant difficulty in the selection of independent variables. Therefore, the construction of predictive models should be supported by very good knowledge of the research object [22]. This knowledge will allow the selection of those input variables that significantly affect yield quality. Among the most commonly used independent variables in the prediction of yield and its chemical and biochemical

characteristics are weather data [26–29]. Average air temperature, total precipitation or total sunshine provide valuable information on plant development conditions. Data on soil mineral abundance, fertilizer application rates, and the course of plant phenological traits are commonly used to build predictive models [30–32].

One effective method for yield quality prediction is machine learning, among which artificial neural networks (ANNs) are of great interest [33–37]. The prototypes of ANNs are the nerve cells that build the human brain, so the operation of artificial neural networks is similar to that of the human brain [38,39]. Each neural network is made up of many simultaneously working and jointly processing elements called neurons. Neurons, due to their function, can be divided into three basic groups: input neurons (they are responsible for inputting the signal into the network), information processing neurons and output neurons ("producing" the results of the network to the outside world) [38,40]. Each of these groups of neurons forms a separate layer, the function of which is the same as the function of the elements from which it is built. Thus, the first is the input layer, which contains a number of neurons equal to the number of independent variables. Its task is to separate the input data into a number of neurons contained in the hidden layer. The hidden layer is built from the n-th number of neurons, the number of which depends on the complexity of the problem being solved by the network. In the structure of a neural network, there can be a different number of hidden layers. The decision on how many hidden layers to use is made by the network developer and is generally an arbitrary decision. The last third layer usually contains only one neuron, responsible for transmitting the result. A neural network, as a layered structure, works by connecting adjacent layers on an "each to each" basis [36,41].

This paper presents the possibility of using ANN to predict the protein content of pea seeds. An extensive analysis of the literature revealed a lack of scientific work of a similar nature. There are no reports on the possibility of predicting the protein content of peas under Polish weather and habitat conditions. In this study, three hypotheses are put forward for verification: (i) artificial neural networks are an effective tool for predicting the protein content of *pea* seeds 20 days before harvest; (ii) it is possible to create a model predicting the protein content of pea seeds based on five-year field trials; and (iii) the ANN model predicts the protein content of pea seeds with greater accuracy compared to the MLR model.

2. Materials and Methods

Experimental data were obtained from a 5-year cycle of field experiments with peas which were conducted in Poland. The results of the experiments were obtained from the field books of the system of the Research Center for Cultivar Testing (COBORU) [42]. Among other things, this institution is engaged in research on distinctiveness, uniformity and durability (DUS) of crop varieties. It is also within the scope of COBORU to conduct field trials for cultivation and use value (VCU). Obtaining positive results from these experiments allows a given variety to be included in the National Variety List. In addition, COBORU supervises the legal protection of varieties entered in the National Register [43]. The field books were created based on data from the official results of experiments under the Program of Registered Varietal Testing (PRVT; in Polish, PDO). PRVT is a system of permanent or periodic testing on the economic value of crop species listed in the National Register or included in the Community Catalogs of Agricultural/Vegetable Varieties (CCA/CCV). PRVT covers both varietal and varietal-agronomic experiments [44].

Field experiments were conducted at the Stations and Experimental Plants for Variety Testing of COBORU located in: Bezek (51°12′6.722″ N 23°16′7.656″ E), Głębokie (52°38′33.18″ N 18°26′16.26″ E), Kawęczyn (52°10′15.157″ N 20°20′49.328″ E), Krzyżewo (53°1′33.535″ N 22°45′28.438″ E), Pawłowice (50°27′14.049″ N 18°29′28.912″ E), Radostowo (53°59′20.566″ N 18°44′41.429″ E) and Sulejów (51°21′8.03″ N 19°52′7.517″ E) (Figure 1). These localities were chosen for their optimal conditions for pea cultivation. These locations are dominated by clay soils of class II-IIIb (polish classification). The experiments

were conducted in accordance with COBORU methodology, which includes a number of agrotechnical recommendations. All studies on selected pea varieties were conducted on plots of 13.86 m^2. The experiments were conducted in an arrangement with variety groups for species in which different morphological types are studied (e.g., traditional and self-terminating varieties, tall, medium-high and low varieties, etc.). In a system with groups of varieties, first the place in repetitions is drawn or determined, and then the order of varieties in groups. The number of repetitions for each variety in each year of the study was 3. A model based on artificial neural networks (N1) and multiple linear regression (RS) was developed for 11 general-purpose pea varieties: Arwena, Astronaute, Batuta, Mecenas, Medyk, Mentor, Olympus, Spot, Starski, Tarchalska and Tytus. These varieties are widely recommended for cultivation in Poland due to their relatively high yield levels and relatively high resistance to biotic factors.

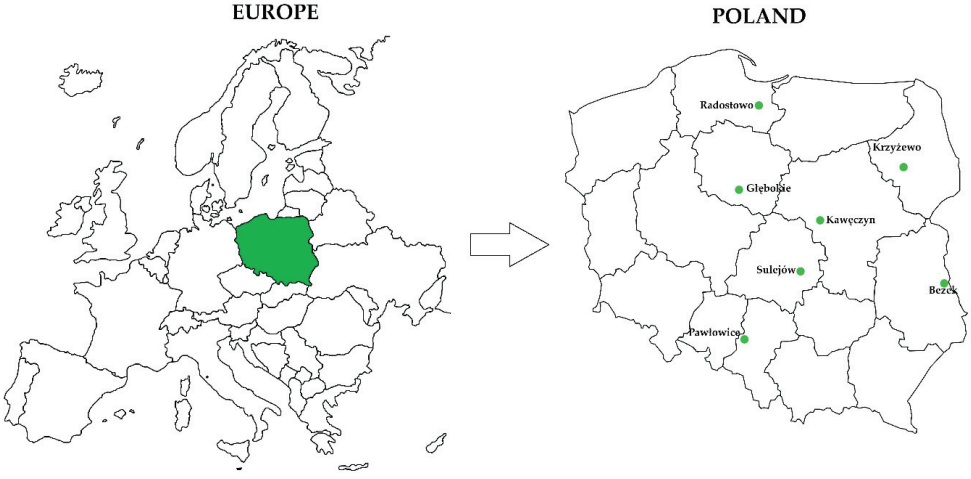


Figure 1. Location of field experiments conducted.

2.1. Data for Model Construction

Two categories of variables were used during the creation of the N1 and RS model. The first group was agronomic data, phytophenological data and results of protein content of *pea* seeds, all of which came from COBORU field books. The second category of variables was meteorological data, which came from a dataset of meteorological phenomena and observations recorded at each COBORU Variety Testing Station and Department. Missing data, such as sunshine totals, were supplemented using historical data from meteorological stations of the Institute of Meteorology and Water Management–National Research Institute. Measurements from meteorological stations that were located closest to the experimental facilities were used [25]. This information was obtained from a public archival database, available electronically [45].

2.2. Construction of the Database

Nineteen independent variables were used to construct the N1 and RS models, as shown in Table 1. The dependent variable was the percentage protein content of pea seeds. Data from a total of 1155 plots were used to construct and verify the N1 and RS model. Each of the analyzed plots constituted one separate case for model construction. The information was grouped into two sets A and B. Set A contained data from 1040 plots and was used to build a neural and regression model. Set B, on the other hand, contained cases from 115 plots and was used to validate the models. Therefore, this set was not used to build the neural network and regression model. It should be noted that the data of set B was selected randomly. However, the determinant representing each case was the

variety. The percentage of protein in the seeds was the dependent variable for the predictive models created.

Table 1. Variables used to build the N1 and RS model.

Symbol	Unit of Measure	Variable Description	Data Range
		Independent Variables	
RAIN	mm	Total rainfall from sowing date to 14 July	96.9–312.4
SUN	h	Total sunshine from sowing date to 14 July	630.5–1051.5
TEMP	°C	Average air temperature from sowing date to 14 July	11.0–17.5
N_F	$kg \cdot ha^{-1}$	Total nitrogen from mineral fertilizers	10–90
P2O5_F	$kg \cdot ha^{-1}$	Total phosphorus from mineral fertilizers	0–80
K2O_F	$kg \cdot ha^{-1}$	Total potassium from mineral fertilizers	0–119
SOWI	days	Number of days from 1 January to sowing date	83–102
P_EMER	days	Number of days from 1 January to the beginning of plant emergence	96–133
HAR	days	Number of days from 1 January to the date of harvesting	184–221
FLOWE	days	Number of days from 1 January to the beginning of flowering	126–169
INI_MA	days	Number of days from 1 January to onset of maturity	167–211
TECH_M	days	Number of days from 1 January to technical maturity	171–216
P_HIG	cm	Plant height	43–156
WEGW	days	Number of plant growing days	87–137
PH	-	Soil pH	5.5–7.5
P2O5_C	Scale from 0 to 4 *	P_2O_5 content in the soil	0–4
K2O_C	Scale from 0 to 4 *	K_2O content in the soil	0–4
MGO_C	Scale from 0 to 4 *	MgO content in the soil	0–4
GEN	feature coded 101 to 111	General variety of peas	-
		Dependent variable	
PROT	%	Percentage of protein in pea seeds	18.56–29.22

* The scale from 0 to 4 refers to the abundance of macronutrients in the soil and is determined as follows: 0—very low, 1—low, 2—medium, 3—high, 4—very high.

The construction of the ANN model was performed on the basis of the predicted date for the calendar year, i.e., 14 July. Based on the analysis of data from all five years of the study (2016–2020), it was shown that pea harvesting was most often performed on 3 August, and the latest on 10 August. The date of 14 July, which is the date of prediction, is the dominant beginning of maturity of the analyzed varieties. This approach to predicting the protein content of pea seeds makes it possible to make predictions 20 days before harvest (based on the dominant harvest date) in the same calendar year.

2.3. Determination of Protein Content in Pea Seeds

Protein determination using the Kjeldahl method is a standard method for determining proteins in plant raw materials [46]. It involves the conversion of protein nitrogen into ammonium sulfate with concentrated sulfuric acid. In the first step, the ground and dried sample is mineralized with sulfuric acid (VI) in the presence of K_2SO_4 and $CuSO_4$ catalysts. To the mineralized and cooled sample, 75 mL of distilled water and 2 of receiving solution are added. The solution thus prepared is distilled for about 4 min. In the final step, the resulting distillate is titrated with standard hydrochloric acid (0.1 $mol \cdot L^{-1}$) until a gray-green color appears. The amount of hydrochloric acid used for titration is the basis for calculating the total protein content of the sample [47].

2.4. ANN Model Development

The construction of the N1 model was performed on the basis of the predicted date for the calendar year, i.e., 14 July. Based on the analysis of data from all five years of the study (2016–2020), it was shown that pea harvesting was most often performed on 3 August, and the latest on 10 August. The date of 14 July, which is the date of prediction, is the dominant beginning of maturity of the analyzed varieties. This approach to predicting the protein content of pea seeds makes it possible to make predictions 20 days before harvest (based on the dominant harvest date) in the same calendar year.

The choice of network type was made by repeatedly building neural networks using an automatic network designer, as well as by reviewing the available literature [21,48–51]. A total of 10,000 networks were tested, which allowed the selection of a multilayer perceptron type network with the following architecture: MLP 19:19-32-1:1 (Figure 2). The selection of the network architecture was guided using the size of the error of the validation, test and learning sets, as well as key network quality parameters, which are shown in Table 2. Achieving the smallest error values by these sets leads to an ANN with high prediction accuracy [52]. An important element in the construction of the prediction model was the division of set A (1040 fields) into three subsets: learning, test and validation. These subsets consisted of 520, 260 and 260 cases (50%–25%–25%), respectively. The entire procedure was performed in the Statistica v7.1 software (TIBCO Software Inc., Palo Alto, CA, USA).

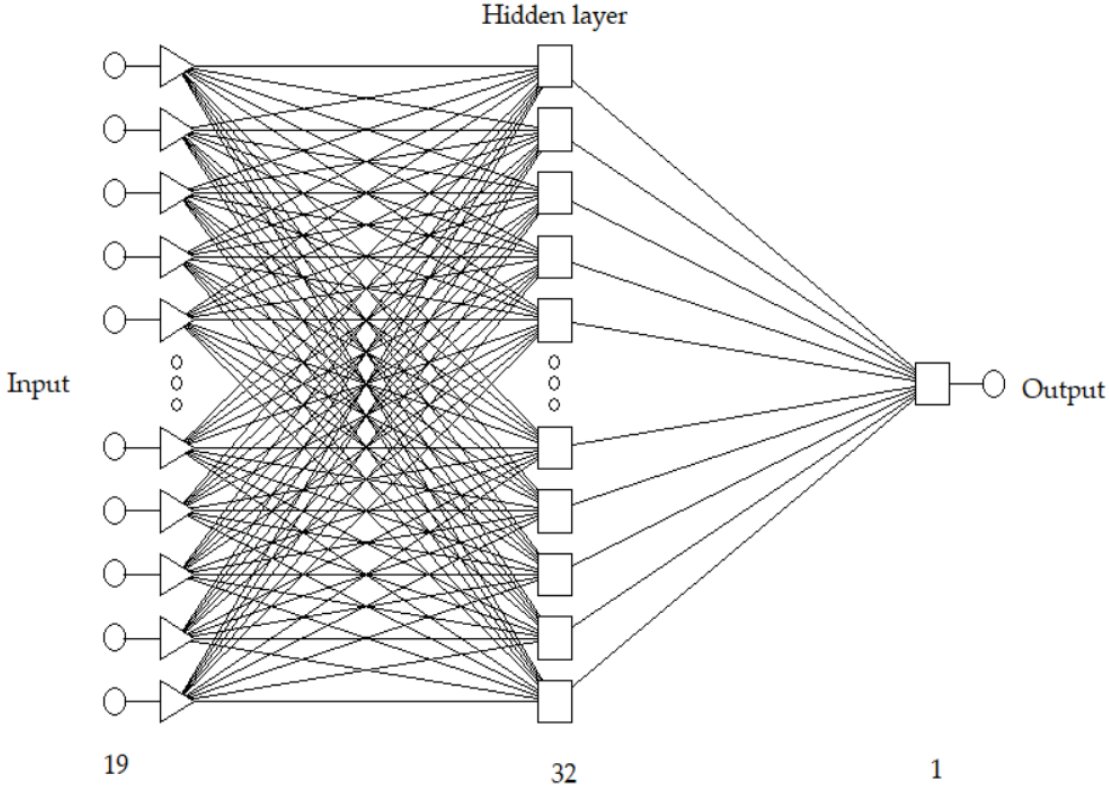


Figure 2. Architecture of neural network with MLP topology.

Table 2. Assessment of the abundance of soils in Poland.

Resources	Phosphorus, mg $P_2O_5 \cdot kg^{-1}$ Soil	Potassium, mg $K_2O \cdot kg^{-1}$ Soil			Magnesium, mg $MgO \cdot kg^{-1}$ Soil		
		Soil Agronomic Category			Soil Agronomic Category		
		Light	Medium	Heavy	Light	Medium	Heavy
very low	up to 50	up to 100	up to 105	up to 170	up to 80	up to 105	up to 120
low	51–80	101–160	106–170	171–260	81–135	106–160	121–220
average	81–115	161–275	171–310	261–350	136–200	161–265	221–330
high	116–185	276–380	311–420	351–510	201–285	266–330	331–460
very high	>185	>380	>420	>510	>285	>330	>460

Due to the coding of the variable P2O5_C, K2O_C and MGO_C, Table 2 presents an assessment of the abundance of Polish soils in phosphorus, potassium and magnesium. Very low abundance of soils in these elements was not recorded in the localities where field experiments were conducted.

2.5. MLR Model Development

Multiple linear regression (MLR) is a statistical tool for detecting interdependencies between independent characteristics and the dependent variable. It makes it possible to determine the strength and type of the detected dependence and to build a functional model that makes it possible to forecast the direction of change in one characteristic on the basis of others [53]. In addition, this method allows you to identify significant and non-significant variables at a given level of probability. This knowledge is particularly important in the further stages of modeling and allows you to identify those invariant variables that do not affect the dependent variable [54]. The general regression formula is shown in Equation (1).

$$Y = b0 + b1 \cdot X1 + b2 \cdot X2 + \ldots + bp \cdot Xp + \varepsilon, \quad (1)$$

where: Y-dependent variable (explained variable), X1, X2 … Xp-independent variables (explanatory variable), b0, b1, b2 … bp-equation parameters, ε-random component (rest of the model).

In this study, a model developed using multiple linear regression (stepwise progressive) was used to predict the percentage protein content of pea seeds. For its construction, the independent variables shown in Table 1 were used. The model was built to compare the effectiveness of the prediction of protein content in peas and was contrasted with a nonlinear model (ANN).

The RS model, like N1, was created based on 1040 cases. The analysis continued through 17 steps, and of the 19 explanatory variables, two of them (number of growing days and number of days from the beginning of the year to emergence) were removed using the model. 115 random observations (set B) were used to predict protein content. Multivariate linear regression analysis was performed using Statistica v7.1 software, and the results are presented graphically and in tabular form.

2.6. Verification of the N1 and RS Models

The obtained N1 and RS models were verified on the basis of measures of predictive properties (Equations (2)–(7)). To calculate them, a set B (115 fields) was used on the basis of which the difference between actual and predicted values was determined. This is example 2 of an equation [22,55]:

$$RAE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y'_i)^2}{\sum_{i=1}^{n} (y_i)^2}}, \quad (2)$$

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y'_i)^2}{n}} \quad (3)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - y'_i| \quad (4)$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y'_i}{y_i}\right| \cdot 100\% \quad (5)$$

$$\mathrm{MAX} = \max_i |y_i - y'_i| \quad (6)$$

$$\mathrm{MAXP} = \max_i \left|\frac{y_i - y'_i}{y_i}\right| \cdot 100\% \quad (7)$$

where: n-number of observations, y_i-actual values, y'_i-predictive values, obtained with the model.

2.7. Sensitivity Analysis of the Neural Network

The final step in creating the N1 model was to identify the independent variables that most influenced the variable explained by the model. To do this, a sensitivity analysis of the network was performed, which allows the rank of each feature to be quantified. The ranks are determined by the size of the deviation quotient, which is the ratio of the error to the error received by all independent variables. The importance of a given feature is greater the higher the deviation quotient achieved by it.

3. Results

3.1. Neural Network Learning and Quality Assessment of Models Predicting Protein Content in Pea Seeds

The obtained multilayer perceptron-type neural network was learned using two methods. The first method of learning-backward error propagation took 100 epochs. The best learning result was obtained by continuing this process with the coupled gradients method and the best result was achieved at 55 epochs. This approach is common in creating predictive models using ANNs [25,28,56–58]. The error for the network did not exceed 0.6 for each of the sets-i.e., learning, validation and test. The results of the N1 predictive model and its basic features are shown in Tables 3 and 4.

Table 3. Subset error size and number of learning epochs of neural networks.

Subsets	Teaching	Validation	Testing
Size of error	0.0551	0.0535	0.0595
Quality	0.3642	0.4145	0.0551
Epochs of learning			
Back propagation method of error	100		
Coupled gradients method	55b *		

* b (best)-the best result in the indicated learning epoch.

Table 4. Quality parameters of the N1 model.

Quality Measures	Value
Average	22.857
Standard deviation	1.895
Average error	0.008
Error deviation	0.744
Average absolute error	0.574
Deviation quotient	0.393
Correlation coefficient r	0.920

The N1 model had a high correlation coefficient (r = 0.920). Satisfactory values were also obtained for the mean error and mean absolute error, which were 0.008 and 0.574, respectively. Low error values and a high correlation coefficient were among the many parameters that determined the selection of the N1 model for further analysis.

An analysis using multiple linear regression showed that the explanatory variables statistically insignificant at the $\alpha = 0.05$ level were the number of days from 1 January to the beginning of flowering (FLOWE), variety (GEN), soil potassium and phosphorus abundance (K2O_C and P2O5_C, respectively) and soil pH (PH).

Based on the results shown in Table 5, the multiple linear regression equation is of the form:

$$\begin{aligned}\text{PROT} = 40.445 + 0.576 \cdot \text{MGO\_C} - 0.005 \cdot \text{RAIN} - 0.021 \cdot \text{K2O\_F} - 0.028 \cdot \text{N\_F} - 0.065 \cdot \text{HAR} + 0.027 \cdot \\ \text{P\_HIGH} - 0.214 \cdot \text{TECH\_M} + 0.152 \cdot \text{INI\_A} + 0.020 \cdot \text{P2O5\_F} - 0.416 \cdot \text{TEMP} + 0.061 \cdot \text{SOWI} + 0.001 \cdot \text{SUN}\end{aligned} \quad (8)$$

Table 5. MLR analysis results.

Factor	MLR: r = 0.6949 R^2 = 0.4829 Standard Error of Estimate = 1.374					
	Beta	**Standard Error Beta**	**b**	**Standard Error b**	***p***	**Significance**
Free Term	-	-	40.445	0.859	0.000000	+
MGO_C	0.323	0.035	0.576	0.062	0.000000	+
RAIN	−0.154	0.034	−0.005	0.001	0.000007	+
K2O_F	−0.298	0.041	−0.021	0.003	0.000000	+
N_F	−0.187	0.031	−0.028	0.005	0.000000	+
HAR	−0.316	0.041	−0.065	0.009	0.000000	+
P_HIG	0.234	0.033	0.027	0.004	0.000000	+
FLOWE	0.062	0.042	0.021	0.015	0.146017	−
TECH_M	−1.063	0.127	−0.214	0.026	0.000000	+
INI_A	0.745	0.138	0.152	0.028	0.000000	+
P2O5_F	0.195	0.043	0.020	0.004	0.000006	+
TEMP	−0.340	0.086	−0.416	0.104	0.000077	+
SOWI	0.182	0.071	0.061	0.024	0.009821	+
GEN	0.0439	0.023	0.026	0.014	0.056315	−
K2O_C	−0.055	0.031	−0.102	0.063	0.083382	−
SUN	0.067	0.032	0.001	0.0006	0.036298	+
P2O5_C	−0.036	0.029	−0.074	0.060	0.211476	−
PH	0.040	0.032	0.158	0.128	0.219669	−

Determination of the level of statistical significance: − non-significant. + significant for $\alpha = 0.05$.

3.2. Sensitivity Analysis of Neural Networks

Verification of the predictive model based on artificial neural networks was carried out using 115 cases (plots). The N1 model with a structure of 19:19-32-1:1 was prepared based on 19 independent variables. The dependent variable was the percentage protein content of pea seeds. The sensitivity analysis performed on crop A showed that the factor with the greatest effect on the protein content of pea seeds was soil magnesium abundance (Table 6). This trait received a rank of one, and removing this variable from the N1 model would result in an increase in the cumulative error value by 2.366 times. The independent variables that received a rank of two and three were the potassium and phosphorus content in the soil. Not including these variables in the model would have increased the error

by 1546 and 1413 times, respectively. Average daily air temperature received a rank of four and, of all the weather variables, had the greatest effect on the protein content of the eleven pea seed varieties.

Table 6. Quality parameters of the N1 model.

Variable	Quotient	Rank
GEN	1.082	18
RAIN	1.158	12
SUN	1.110	16
TEMP	1.396	4
N_F	1.110	15
P2O5_F	1.095	17
K2O_F	1.194	8
SOWI	1.364	5
P_EMER	1.160	11
WEGE	1.257	7
HAR	1.179	9
FLOWE	1.049	19
INI_MA	1.265	6
TECH_M	1.131	13
P_HIG	1.136	13
PH	1.175	10
P2O5_C	1.413	3
K2O_C	1.546	2
MGO_C	2.366	1

The protein content of pea seeds predicted using the N1 model was compared with actual values (Figure 3). A coefficient of determination was obtained at a relatively high level (R^2 = 0.7979), which means that the model's response is very close to the observed values, and that the network has the ability to correctly represent the relationships that are characteristic of the issue being modeled. This procedure was also performed for the RS model (Figure 4). The obtained coefficient of determination of the studied characteristics was 0.3357, so the model has much weaker predictive properties compared to the N1 model.

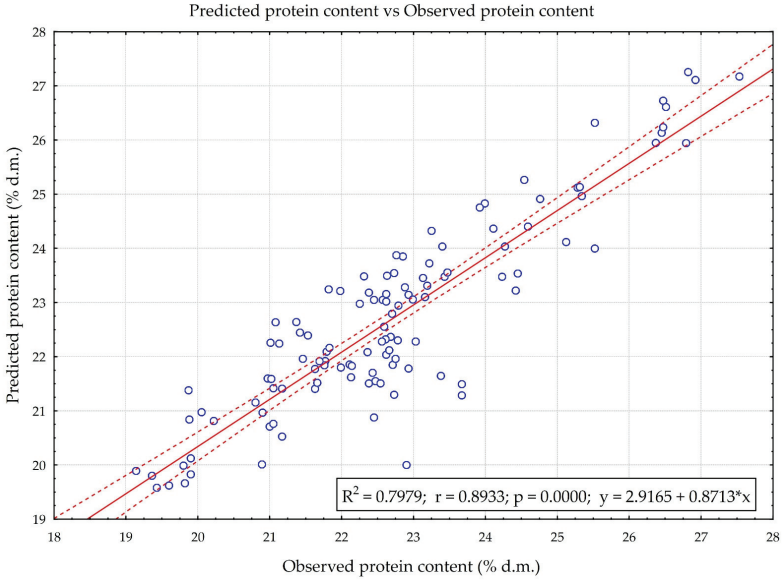


Figure 3. Scatter plot of observed and predicted values for the N1 model.

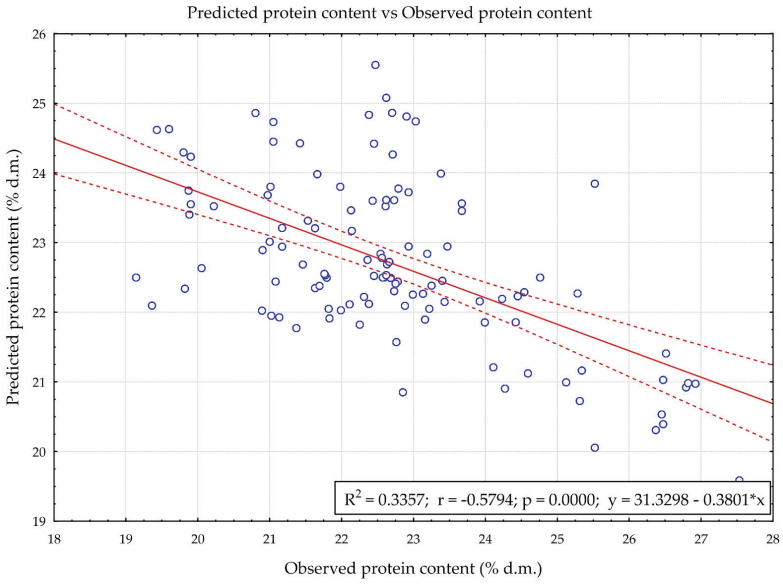


Figure 4. Scatter plot of observed and predicted values for the RS model.

From Figure 5, it can be observed that the protein content of pea seeds changed with the increasing soil magnesium and potassium abundance. An increase in the concentration of magnesium in the soil causes an increase in the percentage of protein content in the analyzed plant. A similar trend is observed for potassium. However, this increase is not as high as in the case of magnesium.

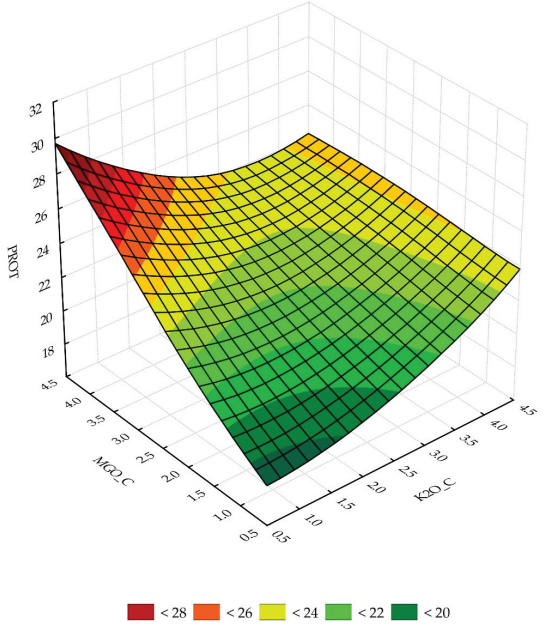


Figure 5. Response surface of the N1 model for percent protein and two dependent variables MGO_C and K2O_C.

Figure 6 shows the relationship of the independent variables (TEPM and P2O5_C) from the sensitivity analysis of the artificial neural network in relation to the dependent variable. From it, it can be observed that an increase in average daily temperatures during the pea growing season promotes the accumulation of protein in its seeds. This relationship also applies to the amount of phosphorus available in the soil. However, as the average daily temperature decreases, the high abundance of phosphorus in the soil does not result in the accumulation of protein in pea seeds.

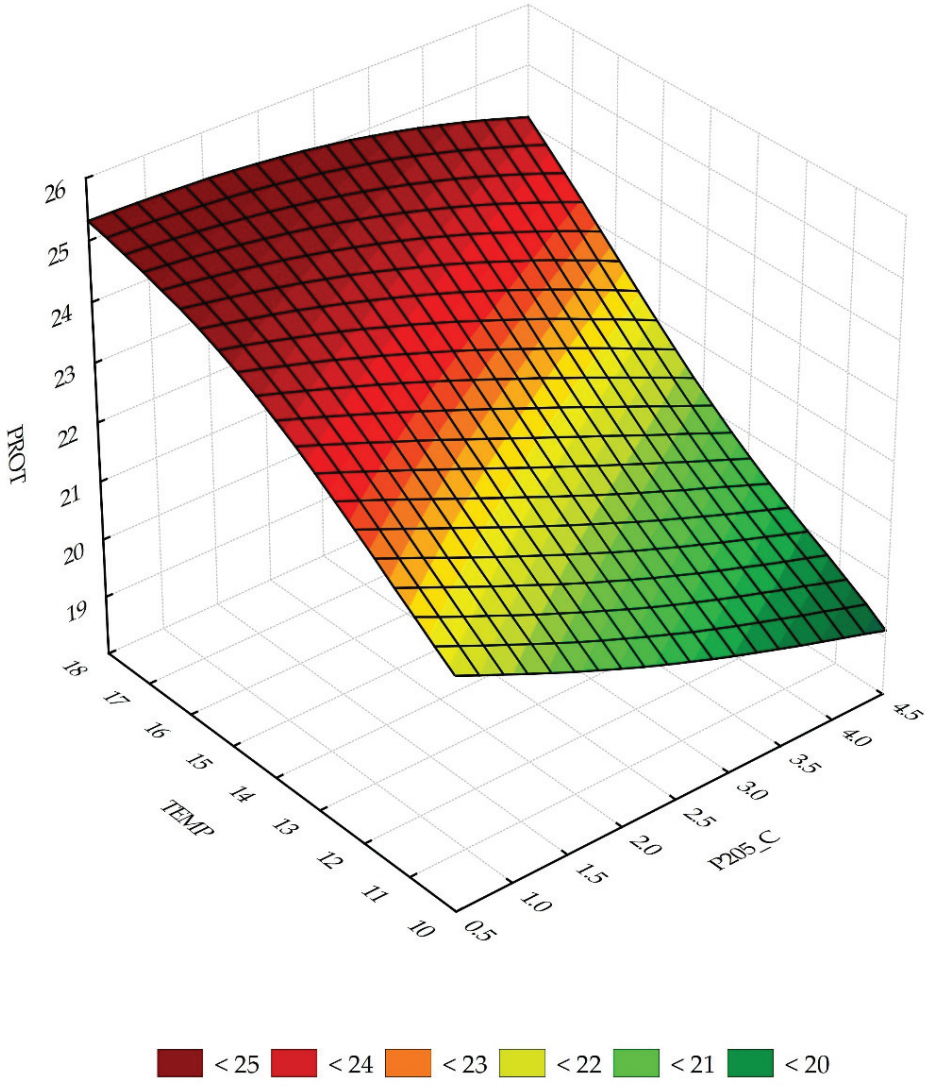


Figure 6. Response surface of model N1 for percent protein and two dependent variables TEMP and P2O5_C.

Changes in protein content as influenced by changes in average daily air temperature (TEMP) and soil magnesium abundance (MGO_C) are shown in Figure 7. An increase in the value of the independent variable TEMP at a low soil magnesium concentration caused the protein content of pea seeds to be less than 21%. A high MgO content in the soil

promotes protein synthesis in seeds only when there are sufficiently high average daily air temperatures during the pea growing season.

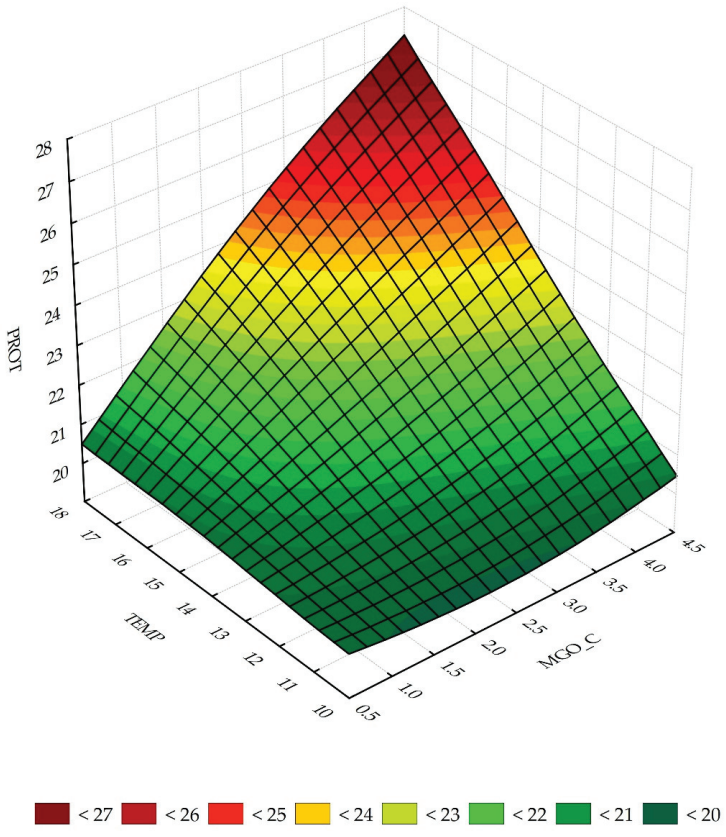


Figure 7. Response surface of the N1 model for protein percentage and two dependent variables TEMP and MGO_C.

3.3. Predictive Properties of the N1 and RS Model

Verification of the correctness of the N1 model and RS model was carried out based on six quality measures: RAE (global relative error of model approximation), RMS (root mean square error), MAE (mean absolute error), MAPE (mean absolute percentage error), MAX (maximum error determined for the whole model) and MAXP (maximum percentage error). Ex-post analysis (Table 7) showed that the N1 model had smaller error values compared to the RS model. The RS model achieved an RMS error of more than three times that of the N1 model. There were similarly large disparities in MAE and MAPE error. The N1 model was also characterized by a more-than-two-times smaller maximum percentage error.

Table 7. Quality assessment of the N1 and RS model.

Error Type	N1 Model	RS Model
RAE	0.037	0.118
RMS	0.838	2.696
MAE	0.617	2.032
MAPE	2.721	8.852
MAX	2.977	7.943
MAXP	13.001	28.853

4. Discussion

An important advantage of artificial neural networks is their ability to model the complex non-linear relationships that occur in agricultural crops [59]. A network with MLP (multi-layer perceptron) topology was used to predict cotton yield. The network was built from four categories of variables: weather data, drought indices, crop vegetation indices and yield. The resulting model had a MAPE error of 1.35%, and the R^2 value was 0.88 [60]. In contrast, Abrougui et al. [32] predicted the yield of potato grown in Chott Meriem (Tunisia). The analysis showed that the best measures of prediction quality (MSE = 0.006, % error = 1.116%) were characterized by a model with a topology of two hidden layers, which consisted of eight neurons in each layer. ANNs have also been successfully used to predict the quality characteristics of plant products such as essential oil content [61], ferulic acid concentration including mycotoxins in wheat grains [62], changes in protein, gluten and water content of stored wheat grains [63] and free radical content of sunflower, palm and rapeseed oil [64]. Niedbała [65] predicted the yield of winter rapeseed (*Brassica napus* L.) using ANN as of 30 June. The study covered fields located in Poland, in the southern part of the Opole Province. The neural network with a MLP topology predicted the explained variable with a MAPE error of 9.43%. The obtained N1 model in the in-house work was also characterized by low prediction errors. For example, the MAPE error was 2.721, which, according to Peng et al. [66] testifies to the model's excellent degree of fit, as the error does not exceed 10%. When MAPE is in the range of 10–20%, the degree of model fit is good. A forecasting model that achieves a MAPE error of more than 30% should be rejected due to poor mapping of predicted values with actual values. A low MAPE error (7.203%) was also obtained in a study by Piekutowska et al. [25], where a neural network with a MLP topology was built to predict the yield of very early potato varieties 40 days before harvest. Prediction of yield and its quality traits before harvesting gains importance in the era of changing climatic conditions. The unstable course of weather during the growing season of crops makes the quality of the crop variable each year. Obtaining models to accurately forecast food production is crucial for policy making and managing national food security plans [67]. Artificial neural networks were also successfully used in wheat yield prediction. The MLP model obtained had an RMS error of 0.4237 [68].

Niazian et al. [30] predicted the essential oil content of ajowan (*Carum copticum* L.) using ANN and MLR. Field studies were conducted from 2014–2015 in central Tehran. Four phenological traits were used as input data. The selection of independent variables was preceded by simple correlation analysis. The study showed that the ANN model with two latent layers predicted essential oil content with a mean squared error of 0.23% and a mean absolute error of 0.14%. In addition, the authors' research showed higher performance of the ANN model compared to the MLR, which had an RMS error of 0.26% and an MAE error of 0.18%. The artificial neural network model also had a higher coefficient of determination value (R^2 = 0.88) compared to the multiple regression model (R^2 = 0.74). The results of our own study also showed the superiority of neural networks over MLR in predicting the protein content of pea seeds. These results are consistent with another study [36], where the performance of MLP and stepwise regression networks was compared in predicting the essential oil content of fennel (*Foeniculum vulgare* Mill.). A total of 11 independent variables were used to build the neural network. The resulting MLP model with a topology of 11:11-9-7-1:1 was characterized by a coefficient of determination of 0.953 and 0.929 for the training and test set, respectively. The stepwise regression model, on the other hand, was characterized by an R^2 magnitude of 0.553. The neural network was additionally characterized by lower prediction errors compared to the MLR. The RMS and MAE error for the ANN were 0.544 and 0.385, respectively, while for the stepwise regression, the magnitudes of these errors were obtained at the level of 0.819 and 0.624. As can be seen from the data in Table 5, the N1 model predicted the protein content of peas with an RMS error of 0.838 and an MAE error of 0.617. The values of these errors are smaller than those obtained using the RS model, demonstrating the greater effectiveness of artificial neural networks over multiple regression in predicting the issue under analysis.

The advantage of artificial neural networks over classical regression modeling is due to the ability of ANNs to approximate non-linear functions [33]. In agricultural crops, many relationships between the analyzed variables have a complex and non-linear course, and MLR models are capable of predicting linear phenomena. Therefore, MLR cannot explain complex nonlinear relationships between independent variables and the dependent variable [69]. In addition to artificial neural networks, random forest regression (RFR) has also been used in agricultural science. Machine learning tools were built to predict the yield of winter rapeseed. The results obtained showed better predictive ability of the RFR model compared to the ANN model [33].

The sensitivity analysis performed for the ANN model showed that the protein content of peas was most affected by soil minerals (Mg, K, P). These variables received rankings one, two and three, respectively (Table 6). Interestingly, the response of pea plants, in terms of protein content, was greater for soil richness in these components than for mineral fertilization. Soil micronutrient and macronutrient abundance critically affects plant growth and development, as well as the quality of the yield obtained [70]. Plants respond less to current mineral fertilization than to high soil nutrient abundance [71]. Yano and Kume [72] noted, in pot experiments, an increased growth of corn root length on soil with a locally elevated phosphorus content. Low root mass increased the efficiency of phosphorus uptake per unit root weight with a low consumption of photosynthetic products.

Magnesium (Mg) received a rank of one in the neural network sensitivity analysis conducted. This means that this element had the greatest impact on the protein content of pea seeds. Mg is one of nine key plant nutrients. It is used in large quantities by plants for proper growth, development and reproduction [73]. Mg has many important physiological functions. It is an essential component of chlorophyll [74], and is involved in CO_2 assimilation reactions in the chloroplast [75]. Magnesium, like potassium (K), is essential for protein biosynthesis and significantly affects the absorption, utilization and metabolism of nitrogen (N) in plant roots [76]. For example, in soybean (*Glycine max* (L.) Merr.), nitrate uptake by the root system was influenced by Mg^{2+} and K^+ ions through the regulation of the NRT2 transporters [77]. In a study conducted by Geng et al. [78], it was shown that rapeseed fertilized with magnesium exhibited increased nitrogen uptake at all fertilization levels (from 0 to 45 kg Mg·ha^{-1}). Much of the Mg contained in leaves appears to be directly or indirectly related to protein synthesis, due to its role in nitrogen metabolism and in the structure and function of ribosomes [76,79,80]. Ribosomes are macromolecular structures that are responsible for protein biosynthesis [81].

A four-year field study conducted in Croatia on six soybean varieties showed that foliar application of magnesium resulted in an increase in protein content in addition to an increase in seed yield. These differences, relative to the control, were statistically significant at the significance level of $\alpha = 0.05$ [82]. In contrast, a study by Sawan et al. [83] investigated the effect of potassium fertilization on cotton protein yield. The study was conducted in Giza, Egypt and covered two growing seasons. The results showed that protein yield significantly increased in plots where potassium was applied (47.4 kg·ha^{-1}) compared to the control, where plants were not fertilized with this element. Similar results were obtained in the present study. A higher Mg content in the soil caused an increase in the protein concentration in pea seeds (Figure 3). A similar relationship was observed for potassium (this variable received a rank of two in the network sensitivity analysis). However, the increase in protein content as a result of higher soil abundance of this element was not as great as in the case of magnesium.

Soil phosphorus content and average daily air temperature were the variables that received ranks three and four, respectively, in the sensitivity analysis (Table 6). Interestingly, TEMP received the highest rank among the weather variables. However, in order for this development not to be disturbed, air temperatures should remain at optimal levels throughout the growing season. Each species has an optimal temperature range for its development, while an excessive decrease or increase can contribute to plant damage [84]. Research conducted by Walter et al. [85] from 2016–2018 show that air temperature was

positively correlated with the protein content of pea and bean (*Vicia faba* L.) seeds grown in Germany. Higher temperatures were associated with higher protein concentrations in both pea and broad bean seeds. These results correlate with the data presented in this paper (Figure 4), where higher protein concentrations were recorded in peas as the average daily temperature increased. The increase in seed protein content was also associated with a higher soil phosphorus content. However, this effect was smaller the lower the daily temperatures were. Low temperatures result in reduced P uptake by plants. When the air temperature is less than 12 °C, the uptake of this element by the root system is largely blocked [86].

Phosphorus is an integral part of cell membranes and nucleic acids and is directly involved in protein synthesis [87]. Therefore, a close relationship between plant P nutrition and various physiological and biochemical traits is frequently and widely reported in the literature [88–91]. The proper nutrition of plants with this component leads to an increase in protein nitrogen and nitrogen of essential amino acids. However, the expected effect can be variable: positive (increase in protein content), neutral or negative (decrease in protein content following the so-called dilution effect). The negative effect results from an increase in usable yield, but in quantitative terms (e.g., for protein yield) it is positive [92].

The results concerning the significance of independent traits generated by using the models do not always coincide. The network sensitivity analysis showed that nitrogen fertilization of pea plants had a negligible effect on seed protein content. However, stepwise regression showed that this variable had a statistically significant effect on the dependent variable under study. Nitrogen is a macronutrient that significantly affects protein synthesis by plants [93]. However, according to a study conducted by Faligowska et al. [17], nitrogen in pea seeds grown in Brody (Poland) was accumulated from three main sources: soil, atmosphere and fertilizers. The largest amount of accumulated nitrogen came from the soil (57.9%), followed by the atmosphere (35.2%) and fertilizers (6.8%). The analysis of the multiple regression results confirmed the well-known state of knowledge about the relationship between nitrogen fertilization and protein content. However, the resulting neural network highlighted less well-known and much more interesting aspects of certain relationships. Relationships between given factors in agricultural production are determined mainly by weather conditions prevailing in a given growing season, habitat conditions and genotypes of crops grown.

5. Conclusions

The developed MLP-type artificial neural network model successfully predicted the protein content of pea seeds 20 days before the harvest date. The model had higher prediction accuracy and lower ex-post error values with respect to the stepwise regression model. The accuracy of MLP networks is mainly dependent on the accuracy of the data at one's disposal and the amount of information fed into the model. In the case of artificial neural networks, an important task is to maintain a balance between the model's ability to approximate and generalize. The analysis conducted showed that the network accurately predicted the dependent variable based on a five-year field study. The results obtained did not allow the rejection of the null hypotheses, thus confirming the validity of the assumptions made at the outset. The N1 model, due to its lower error values and higher R^2 value, more accurately predicted the protein content of peas. For this reason, this model seems to be better in practical application.

This work can serve as a strand for future research on the prediction of pea seed quality such as the content of fat, nitrogen-free compounds, the largest part of which in pea seeds is starch and anti-nutritional compounds. A continuation of the present research will be the optimization of NPK fertilization taking into account the liming needs of general-purpose pea varieties.

Author Contributions: Conceptualization, P.H., M.P. and G.N.; methodology, P.H., M.P. and G.N.; validation, M.P. and G.N.; formal analysis, M.P.; investigation, P.H.; resources, P.H.; data curation, P.H.; writing—original draft preparation, P.H.; writing—review and editing, P.H., M.P. and G.N.;

supervision, G.N.; project administration, M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ANN-artificial neural networks; CCA-Community Catalogs of Agricultural; CCV-Community Catalogs of Varieties of Vegetable; COBORU-Research Center for Cultivar Testing; DUS-distinctiveness, uniformity and durability; FLOWE-number of days from 1 January to the beginning of flowering; GEN-general variety of peas; HAR-number of days from 1 January to the date of harvesting; INI_MA-number of days from 1 January to onset of maturity; K20_C-K2O content in the soil; K2O-potassium oxide; K2O_F-Total potassium from mineral fertilizers; kg-kilogram; MAE-mean absolute error; MAPE-mean absolute percentage error; MAX-maximum error determined for the whole model; MAXP-maximum percentage error; mg-miligram; MgO-magnesium oxide; MGO_C-MgO content in the soil; MLP-multilayer perceptron; MLR-multiple linear regression; MO-organic matter; n-number of observations; N_F-total nitrogen from mineral fertilizers; N1-built its own neural network model; P_EMER-number of days from 1 January to the beginning of plant emergence; P_HIG-plant height; P2O5-phosphorus(V) oxide; P2O5_C-P2O5 content in the soil; P2O5_F-total phosphorus from mineral fertilizers; PH-Soil pH; PROT-Percentage of protein in pea seeds; PRVT-Program of Registered Varietal Experimentation; RAE-global relative error of model approximation; RAIN-total rainfall from sowing date to 14 July; RMS-root mean square error; RS-built its own linear regression model; SOWI-number of days from 1 January to sowing date; SUN-total sunshine from sowing date to 14 July; TECH_M-number of days from 1 January to technical maturity; TEMP-average air temperature from sowing date to 14 July; VCU-trials for cultivation and use value; WEGW-number of plant growing days; y′i-predictive values, obtained with the model; yi-actual values.

References

1. Khatun, M.; Sarkar, S.; Era, F.M.; Islam, A.K.M.M.; Anwar, M.P.; Fahad, S.; Datta, R.; Islam, A.K.M.A. Drought Stress in Grain Legumes: Effects, Tolerance Mechanisms and Management. *Agronomy* **2021**, *11*, 2374. [CrossRef]
2. Atnaf, M.; Tesfaye, K.; Dagne, K. The Importance of Legumes in the Ethiopian Farming System and Overall Economy: An Overview. *Am. J. Exp. Agric.* **2015**, *7*, 347–358. [CrossRef]
3. Graham, P.H.; Vance, C.P. Legumes: Importance and Constraints to Greater Use. *Plant Physiol.* **2003**, *131*, 872–877. [CrossRef] [PubMed]
4. Kalembasa, S.; Szukała, J.; Faligowska, A.; Kalembasa, D.; Symanowicz, B.; Becher, M.; Gebus-Czupyt, B. Quantification of Biologically Fixed Nitrogen by White Lupin (*Lupins albus* L.) and Its Subsequent Uptake by Winter Wheat Using the 15N Isotope Dilution Method. *Agronomy* **2020**, *10*, 1392. [CrossRef]
5. Putra, R.; Powell, J.R.; Hartley, S.E.; Johnson, S.N. Is it time to include legumes in plant silicon research? *Funct. Ecol.* **2020**, *34*, 1142–1157. [CrossRef]
6. Daryanto, S.; Wang, L.; Jacinthe, P.-A. Global Synthesis of Drought Effects on Food Legume Production. *PLoS ONE* **2015**, *10*, e0127401. [CrossRef]
7. Torabian, S.; Farhangi-Abriz, S.; Denton, M.D. Do tillage systems influence nitrogen fixation in legumes? A review. *Soil Tillage Res.* **2019**, *185*, 113–121. [CrossRef]
8. Gentzbittel, L.; Andersen, S.U.; Ben, C.; Rickauer, M.; Stougaard, J.; Young, N.D. Naturally occurring diversity helps to reveal genes of adaptive importance in legumes. *Front. Plant Sci.* **2015**, *6*, 269. [CrossRef]
9. Wang, X.; Yang, Y.; Pei, K.; Zhou, J.; Peixoto, L.; Gunina, A.; Zeng, Z.; Zang, H.; Rasmussen, J.; Kuzyakov, Y. Nitrogen rhizodeposition by legumes and its fate in agroecosystems: A field study and literature review. *Land Degrad. Dev.* **2021**, *32*, 410–419. [CrossRef]
10. Neugschwandtner, R.W.; Bernhuber, A.; Kammlander, S.; Wagentristl, H.; Klimek-Kopyra, A.; Lošák, T.; Zholamanov, K.K.; Kaul, H.-P. Nitrogen Yields and Biological Nitrogen Fixation of Winter Grain Legumes. *Agronomy* **2021**, *11*, 681. [CrossRef]

11. Boukid, F.; Rosell, C.M.; Castellari, M. Pea protein ingredients: A mainstream ingredient to (re)formulate innovative foods and beverages. *Trends Food Sci. Technol.* **2021**, *110*, 729–742. [CrossRef]
12. Powierzchnia Upraw w Gminach. Available online: https://rejestrupraw.arimr.gov.pl/ (accessed on 26 June 2022).
13. Bogahawaththa, D.; Bao Chau, N.H.; Trivedi, J.; Dissanayake, M.; Vasiljevic, T. Impact of selected process parameters on solubility and heat stability of pea protein isolate. *LWT* **2019**, *102*, 246–253. [CrossRef]
14. Pratap, A.; Das, A.; Kumar, S.; Gupta, S. Current Perspectives on Introgression Breeding in Food Legumes. *Front. Plant Sci.* **2021**, *11*, 589189. [CrossRef]
15. Gao, Z.; Shen, P.; Lan, Y.; Cui, L.; Ohm, J.-B.; Chen, B.; Rao, J. Effect of alkaline extraction pH on structure properties, solubility, and beany flavor of yellow pea protein isolate. *Food Res. Int.* **2020**, *131*, 109045. [CrossRef]
16. Chaudhary, A.; Marinangeli, C.; Tremorin, D.; Mathys, A. Nutritional Combined Greenhouse Gas Life Cycle Analysis for Incorporating Canadian Yellow Pea into Cereal-Based Food Products. *Nutrients* **2018**, *10*, 490. [CrossRef] [PubMed]
17. Faligowska, A.; Kalembasa, S.; Kalembasa, D.; Panasiewicz, K.; Szymańska, G.; Ratajczak, K.; Skrzypczak, G. The Nitrogen Fixation and Yielding of Pea in Different Soil Tillage Systems. *Agronomy* **2022**, *12*, 352. [CrossRef]
18. Kornet, C.; Venema, P.; Nijsse, J.; van der Linden, E.; van der Goot, A.J.; Meinders, M. Yellow pea aqueous fractionation increases the specific volume fraction and viscosity of its dispersions. *Food Hydrocoll.* **2020**, *99*, 105332. [CrossRef]
19. Röhe, I.; Göbel, T.W.; Goodarzi Boroojeni, F.; Zentek, J. Effect of feeding soybean meal and differently processed peas on the gut mucosal immune system of broilers. *Poult. Sci.* **2017**, *96*, 2064–2073. [CrossRef]
20. Florek, J. Potential utilization of legumes in feed production in Poland. *Ann. Polish Assoc. Agric. Agribus. Econ.* **2017**, *XIX*, 40–45. [CrossRef]
21. Niedbała, G. Application of artificial neural networks for multi-criteria yield prediction of winter rapeseed. *Sustainability* **2019**, *11*, 533. [CrossRef]
22. Hara, P.; Piekutowska, M.; Niedbała, G. Selection of Independent Variables for Crop Yield Prediction Using Artificial Neural Network Models with Remote Sensing Data. *Land* **2021**, *10*, 609. [CrossRef]
23. Chipanshi, A.; Zhang, Y.; Kouadio, L.; Newlands, N.; Davidson, A.; Hill, H.; Warren, R.; Qian, B.; Daneshfar, B.; Bedard, F.; et al. Evaluation of the Integrated Canadian Crop Yield Forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape. *Agric. For. Meteorol.* **2015**, *206*, 137–150. [CrossRef]
24. Nazir, A.; Ullah, S.; Saqib, Z.A.; Abbas, A.; Ali, A.; Iqbal, M.S.; Hussain, K.; Shakir, M.; Shah, M.; Butt, M.U. Estimation and Forecasting of Rice Yield Using Phenology-Based Algorithm and Linear Regression Model on Sentinel-II Satellite Data. *Agriculture* **2021**, *11*, 1026. [CrossRef]
25. Piekutowska, M.; Niedbała, G.; Piskier, T.; Lenartowicz, T.; Pilarski, K.; Wojciechowski, T.; Pilarska, A.A.; Czechowska-Kosacka, A. The Application of Multiple Linear Regression and Artificial Neural Network Models for Yield Prediction of Very Early Potato Cultivars before Harvest. *Agronomy* **2021**, *11*, 885. [CrossRef]
26. Niedbała, G.; Wróbel, B.; Piekutowska, M.; Zielewicz, W.; Paszkiewicz-Jasińska, A.; Wojciechowski, T.; Niazian, M. Application of Artificial Neural Networks Sensitivity Analysis for the Pre-Identification of Highly Significant Factors Influencing the Yield and Digestibility of Grassland Sward in the Climatic Conditions of Central Poland. *Agronomy* **2022**, *12*, 1133. [CrossRef]
27. Shahhosseini, M.; Hu, G.; Archontoulis, S.V. Forecasting Corn Yield with Machine Learning Ensembles. *Front. Plant Sci.* **2020**, *11*, 1120. [CrossRef]
28. Kakati, N.; Deka, R.L.; Das, P.; Goswami, J.; Khanikar, P.G.; Saikia, H. Forecasting yield of rapeseed and mustard using multiple linear regression and ANN techniques in the Brahmaputra valley of Assam, North East India. *Theor. Appl. Climatol.* **2022**, *150*, 1201–1215. [CrossRef]
29. Pentoś, K.; Mbah, J.T.; Pieczarka, K.; Niedbała, G.; Wojciechowski, T. Evaluation of Multiple Linear Regression and Machine Learning Approaches to Predict Soil Compaction and Shear Stress Based on Electrical Parameters. *Appl. Sci.* **2022**, *12*, 8791. [CrossRef]
30. Niazian, M.; Sadat-Noori, S.A.; Abdipour, M. Artificial neural network and multiple regression analysis models to predict essential oil content of ajowan (*Carum copticum* L.). *J. Appl. Res. Med. Aromat. Plants* **2018**, *9*, 124–131. [CrossRef]
31. van Klompenburg, T.; Kassahun, A.; Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **2020**, *177*, 105709. [CrossRef]
32. Abrougui, K.; Gabsi, K.; Mercatoris, B.; Khemis, C.; Amami, R.; Chehaibi, S. Prediction of organic potato yield using tillage systems and soil properties by artificial neural network (ANN) and multiple linear regressions (MLR). *Soil Tillage Res.* **2019**, *190*, 202–208. [CrossRef]
33. Rajković, D.; Marjanović Jeromela, A.; Pezo, L.; Lončar, B.; Zanetti, F.; Monti, A.; Kondić Špika, A. Yield and Quality Prediction of Winter Rapeseed—Artificial Neural Network and Random Forest Models. *Agronomy* **2021**, *12*, 58. [CrossRef]
34. Abraham, E.R.; Mendes dos Reis, J.G.; Vendrametto, O.; de Oliveira Costa Neto, P.L.; Carlo Toloi, R.; de Souza, A.E.; Oliveira Morais, M. de Time Series Prediction with Artificial Neural Networks: An Analysis Using Brazilian Soybean Production. *Agriculture* **2020**, *10*, 475. [CrossRef]
35. Rathod, S.; Yerram, S.; Arya, P.; Katti, G.; Rani, J.; Padmakumari, A.P.; Somasekhar, N.; Padmavathi, C.; Ondrasek, G.; Amudan, S.; et al. Climate-Based Modeling and Prediction of Rice Gall Midge Populations Using Count Time Series and Machine Learning Approaches. *Agronomy* **2021**, *12*, 22. [CrossRef]

36. Sabzi-Nojadeh, M.; Niedbała, G.; Younessi-Hamzekhanlu, M.; Aharizad, S.; Esmaeilpour, M.; Abdipour, M.; Kujawa, S.; Niazian, M. Modeling the Essential Oil and Trans-Anethole Yield of Fennel (*Foeniculum vulgare* Mill. var. vulgare) by Application Artificial Neural Network and Multiple Linear Regression Methods. *Agriculture* **2021**, *11*, 1191. [CrossRef]
37. Kujawa, S.; Dach, J.; Kozłowski, R.J.; Przybył, K.; Niedbała, G.; Mueller, W.; Tomczak, R.J.; Zaborowicz, M.; Koszela, K. Maturity classification for sewage sludge composted with rapeseed straw using neural image analysis. In Proceedings of the SPIE—The International Society for Optical Engineering, Chengu, China, 29 August 2016; Falco, C.M., Jiang, X., Eds.; SPIE: Bellingham, WA, USA, 2016; p. 100332H. [CrossRef]
38. Maya Gopal, P.S.; Bhargavi, R. A novel approach for efficient crop yield prediction. *Comput. Electron. Agric.* **2019**, *165*, 104968. [CrossRef]
39. Cieniawska, B.; Pentoś, K.; Łuczycka, D. Neural modeling and optimization of the coverage of the sprayed surface. *Bull. Pol. Acad. Sci. Tech. Sci.* **2020**, *68*, 601–608. [CrossRef]
40. Marchant, J.; Onyango, C. Comparison of a Bayesian classifier with a multilayer feed-forward neural network using the example of plant/weed/soil discrimination. *Comput. Electron. Agric.* **2003**, *39*, 3–22. [CrossRef]
41. Pentoś, K.; Łuczycka, D.; Kapłon, T. The identification of relationships between selected honey parameters by extracting the contribution of independent variables in a neural network model. *Eur. Food Res. Technol.* **2015**, *241*, 793–801. [CrossRef]
42. Research Centre for Cultivar Testing (COBORU). Available online: https://coboru.gov.pl/ (accessed on 20 October 2022).
43. Niedbała, G.; Tratwal, A.; Piekutowska, M.; Wojciechowski, T.; Uglis, J. A Framework for Financing Post-Registration Variety Testing System: A Case Study from Poland. *Agronomy* **2022**, *12*, 325. [CrossRef]
44. Porejestrowe Doświadczalnictwo Odmianowe (PDO). Available online: https://coboru.gov.pl/pdo/pdo (accessed on 20 October 2022).
45. Dane Publiczne IMGW. Available online: https://danepubliczne.imgw.pl/ (accessed on 20 October 2022).
46. Mádlíková, M.; Krausová, I.; Mizera, J.; Táborský, J.; Faměra, O.; Chvátil, D. Nitrogen assay in winter wheat by short-time instrumental photon activation analysis and its comparison with the Kjeldahl method. *J. Radioanal. Nucl. Chem.* **2018**, *317*, 479–486. [CrossRef]
47. Simonne, A.H.; Simonne, E.H.; Eitenmiller, R.R.; Mills, H.A.; Cresman, C.P. Could the Dumas Method Replace the Kjeldahl Digestion for Nitrogen and Crude Protein Determinations in Foods? *J. Sci. Food Agric.* **1997**, *73*, 39–45. [CrossRef]
48. Ma, Y.; Zhang, Z.; Kang, Y.; Özdoğan, M. Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. *Remote Sens. Environ.* **2021**, *259*, 112408. [CrossRef]
49. Roy Choudhury, M.; Das, S.; Christopher, J.; Apan, A.; Chapman, S.; Menzies, N.W.; Dang, Y.P. Improving Biomass and Grain Yield Prediction of Wheat Genotypes on Sodic Soil Using Integrated High-Resolution Multispectral, Hyperspectral, 3D Point Cloud, and Machine Learning Techniques. *Remote Sens.* **2021**, *13*, 3482. [CrossRef]
50. Priya, P.K.; Yuvaraj, N. An IoT Based Gradient Descent Approach for Precision Crop Suggestion using MLP. *J. Phys. Conf. Ser.* **2019**, *1362*, 012038. [CrossRef]
51. Bhojani, S.H.; Bhatt, N. Wheat crop yield prediction using new activation functions in neural network. *Neural Comput. Appl.* **2020**, *32*, 13941–13951. [CrossRef]
52. Niedbała, G.; Piekutowska, M.; Weres, J.; Korzeniewicz, R.; Witaszek, K.; Adamski, M.; Pilarski, K.; Czechowska-Kosacka, A.; Krysztofiak-Kaniewska, A. Application of artificial neural networks for yield modeling of winter rapeseed based on combined quantitative and qualitative data. *Agronomy* **2019**, *9*, 781. [CrossRef]
53. Pazhanivelan, S.; Geethalakshmi, V.; Tamilmounika, R.; Sudarmanian, N.S.; Kaliaperumal, R.; Ramalingam, K.; Sivamurugan, A.P.; Mrunalini, K.; Yadav, M.K.; Quicho, E.D. Spatial Rice Yield Estimation Using Multiple Linear Regression Analysis, Semi-Physical Approach and Assimilating SAR Satellite Derived Products with DSSAT Crop Simulation Model. *Agronomy* **2022**, *12*, 2008. [CrossRef]
54. Niedbała, G.; Kurek, J.; Świderski, B.; Wojciechowski, T.; Antoniuk, I.; Bobran, K. Prediction of Blueberry (*Vaccinium corymbosum* L.) Yield Based on Artificial Intelligence Methods. *Agriculture* **2022**, *12*, 2089. [CrossRef]
55. Schwalbert, R.A.; Amado, T.; Corassa, G.; Pott, L.P.; Prasad, P.V.V.; Ciampitti, I.A. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agric. For. Meteorol.* **2020**, *284*, 107886. [CrossRef]
56. Shankar, T.; Malik, G.C.; Banerjee, M.; Dutta, S.; Praharaj, S.; Lalichetti, S.; Mohanty, S.; Bhattacharyay, D.; Maitra, S.; Gaber, A.; et al. Prediction of the Effect of Nutrients on Plant Parameters of Rice by Artificial Neural Network. *Agronomy* **2022**, *12*, 2123. [CrossRef]
57. Niedbała, G.; Kurasiak-Popowska, D.; Piekutowska, M.; Wojciechowski, T.; Kwiatek, M.; Nawracała, J. Application of Artificial Neural Network Sensitivity Analysis to Identify Key Determinants of Harvesting Date and Yield of Soybean (*Glycine max* [L.] Merrill) Cultivar Augusta. *Agriculture* **2022**, *12*, 754. [CrossRef]
58. Wojciechowski, T.; Niedbala, G.; Czechlowski, M.; Nawrocka, J.R.; Piechnik, L.; Niemann, J. Rapeseed seeds quality classification with usage of VIS-NIR fiber optic probe and artificial neural networks. In Proceedings of the 2016 International Conference on Optoelectronics and Image Processing (ICOIP), Warsaw, Poland, 10–12 June 2016; IEEE: Warsaw, Poland, 2016; pp. 44–48. [CrossRef]
59. Manish Lad, A.; Mani Bharathi, K.; Akash Saravanan, B.; Karthik, R. Factors affecting agriculture and estimation of crop yield using supervised learning algorithms. *Mater. Today Proc.* **2022**, *62*, 4629–4634. [CrossRef]

60. Yildirim, T.; Moriasi, D.N.; Starks, P.J.; Chakraborty, D. Using Artificial Neural Network (ANN) for Short-Range Prediction of Cotton Yield in Data-Scarce Regions. *Agronomy* **2022**, *12*, 828. [CrossRef]
61. Akbar, A.; Kuanar, A.; Patnaik, J.; Mishra, A.; Nayak, S. Application of Artificial Neural Network modeling for optimization and prediction of essential oil yield in turmeric (*Curcuma longa* L.). *Comput. Electron. Agric.* **2018**, *148*, 160–178. [CrossRef]
62. Niedbała, G.; Kurasiak-Popowska, D.; Stuper-Szablewska, K.; Nawracała, J. Application of Artificial Neural Networks to Analyze the Concentration of Ferulic Acid, Deoxynivalenol, and Nivalenol in Winter Wheat Grain. *Agriculture* **2020**, *10*, 127. [CrossRef]
63. Szwedziak, K.; Polańczyk, E.; Grzywacz, Ż.; Niedbała, G.; Wojtkiewicz, W. Neural Modeling of the Distribution of Protein, Water and Gluten in Wheat Grains during Storage. *Sustainability* **2020**, *12*, 5050. [CrossRef]
64. Huang, S.; Liu, Y.; Sun, X.; Li, J. Application of Artificial Neural Network Based on Traditional Detection and GC-MS in Prediction of Free Radicals in Thermal Oxidation of Vegetable Oil. *Molecules* **2021**, *26*, 6717. [CrossRef]
65. Niedbała, G. Simple model based on artificial neural network for early prediction and simulation winter rapeseed yield. *J. Integr. Agric.* **2019**, *18*, 54–61. [CrossRef]
66. Peng, J.; Kim, M.; Kim, Y.; Jo, M.; Kim, B.; Sung, K.; Lv, S. Constructing Italian ryegrass yield prediction model based on climatic data by locations in South Korea. *Grassl. Sci.* **2017**, *63*, 184–195. [CrossRef]
67. Nosratabadi, S.; Ardabili, S.; Lakner, Z.; Mako, C.; Mosavi, A. Prediction of Food Production Using Machine Learning Algorithms of Multilayer Perceptron and ANFIS. *Agriculture* **2021**, *11*, 408. [CrossRef]
68. Ahmed, M.U.; Hussain, I. Prediction of Wheat Production Using Machine Learning Algorithms in northern areas of Pakistan. *Telecomm. Policy* **2022**, *46*, 102370. [CrossRef]
69. Meerasri, J.; Sothornvit, R. Artificial neural networks (ANNs) and multiple linear regression (MLR) for prediction of moisture content for coated pineapple cubes. *Case Stud. Therm. Eng.* **2022**, *33*, 101942. [CrossRef]
70. Saba, T.; Liu, W.; Wang, J.; Saleem, F.; Kang, X.; Hui, W.; Gong, W.; Li, H. Effects of organic supplementation to reduced rates of chemical fertilization on soil fertility of Zanthoxylum armatum. *Dendrobiology* **2022**, *87*, 123–136. [CrossRef]
71. Dincă, L.C.; Grenni, P.; Onet, C.; Onet, A. Fertilization and Soil Microbial Community: A Review. *Appl. Sci.* **2022**, *12*, 1198. [CrossRef]
72. Yano, K.; Kume, T. Root Morphological Plasticity for Heterogeneous Phosphorus Supply in *Zea mays* L. *Plant Prod. Sci.* **2005**, *8*, 427–432. [CrossRef]
73. Gransee, A.; Führs, H. Magnesium mobility in soils as a challenge for soil and plant analysis, magnesium fertilization and root uptake under adverse growth conditions. *Plant Soil* **2013**, *368*, 5–21. [CrossRef]
74. Wei, Q.; Guo, Y.; Kuai, B. Isolation and characterization of a chlorophyll degradation regulatory gene from tall fescue. *Plant Cell Rep.* **2011**, *30*, 1201–1207. [CrossRef]
75. Xu, X.-F.; Wang, B.; Lou, Y.; Han, W.-J.; Lu, J.-Y.; Li, D.-D.; Li, L.-G.; Zhu, J.; Yang, Z.-N. Magnesium Transporter 5 plays an important role in Mg transport for male gametophyte development in Arabidopsis. *Plant J.* **2015**, *84*, 925–936. [CrossRef]
76. Xie, K.; Cakmak, I.; Wang, S.; Zhang, F.; Guo, S. Synergistic and antagonistic interactions between potassium and magnesium in higher plants. *Crop J.* **2021**, *9*, 249–256. [CrossRef]
77. Peng, W.T.; Qi, W.L.; Nie, M.M.; Xiao, Y.B.; Liao, H.; Chen, Z.C. Magnesium supports nitrogen uptake through regulating NRT2.1/2.2 in soybean. *Plant Soil* **2020**, *457*, 97–111. [CrossRef]
78. Geng, G.; Cakmak, I.; Ren, T.; Lu, Z.; Lu, J. Effect of magnesium fertilization on seed yield, seed quality, carbon assimilation and nutrient uptake of rapeseed plants. *F. Crop. Res.* **2021**, *264*, 108082. [CrossRef]
79. Chaudhry, A.H.; Nayab, S.; Hussain, S.B.; Ali, M.; Pan, Z. Current Understandings on Magnesium Deficiency and Future Outlooks for Sustainable Agriculture. *Int. J. Mol. Sci.* **2021**, *22*, 1819. [CrossRef] [PubMed]
80. Wang, Z.; Hassan, M.U.; Nadeem, F.; Wu, L.; Zhang, F.; Li, X. Magnesium Fertilization Improves Crop Yield in Most Production Systems: A Meta-Analysis. *Front. Plant Sci.* **2020**, *10*, 1727. [CrossRef]
81. Fischer, E.S.; Lohaus, G.; Heineke, D.; Heldt, H.W. Magnesium deficiency results in accumulation of carbohydrates and amino acids in source and sink leaves of spinach. *Physiol. Plant.* **1998**, *102*, 16–20. [CrossRef] [PubMed]
82. Vrataric, M.; Sudaric, A.; Kovacevic, V.; Duvnjak, T.; Krizmanic, M.; Mijic, A. Response of soybean to foliar fertilization with magnesium sulfate (epsom salt). *Cereal Res. Commun.* **2006**, *34*, 709–712. [CrossRef]
83. Sawan, Z.M.; Hafezb, S.A.; Basyony, A.E.; Alkassas, A.-E.-E.R. Cottonseed: Protein, oil yields, and oil properties as influenced by potassium fertilization and foliar application of zinc and phosphorus. *Grasas Aceites* **2007**, *58*, 40–48. [CrossRef]
84. Skrzyczyńska, J.; Gąsiorowska, B. *Uprawa Roślin*; UPW: Wrocław, Poland, 2020; pp. 49–210.
85. Walter, S.; Zehring, J.; Mink, K.; Quendt, U.; Zocher, K.; Rohn, S. Protein content of peas (*Pisum sativum*) and beans (*Vicia faba*)—Influence of cultivation conditions. *J. Food Compos. Anal.* **2022**, *105*, 104257. [CrossRef]
86. Grzebisz, W. *Nawożenie Roślin Uprawnych 2*; Powszechne Wydawnictwo Rolnicze i Leśne: Poznań, Poland, 2009.
87. Singh, S.K.; Reddy, V.R.; Fleisher, D.H.; Timlin, D.J. Phosphorus Nutrition Affects Temperature Response of Soybean Growth and Canopy Photosynthesis. *Front. Plant Sci.* **2018**, *9*, 1116. [CrossRef]
88. Singh, S.K.; Reddy, V.R. Combined effects of phosphorus nutrition and elevated carbon dioxide concentration on chlorophyll fluorescence, photosynthesis, and nutrient efficiency of cotton. *J. Plant Nutr. Soil Sci.* **2014**, *177*, 892–902. [CrossRef]
89. Singh, S.K.; Reddy, V.R. Response of carbon assimilation and chlorophyll fluorescence to soybean leaf phosphorus across CO_2: Alternative electron sink, nutrient efficiency and critical concentration. *J. Photochem. Photobiol. B Biol.* **2015**, *151*, 276–284. [CrossRef] [PubMed]

90. Taliman, N.A.; Dong, Q.; Echigo, K.; Raboy, V.; Saneoka, H. Effect of Phosphorus Fertilization on the Growth, Photosynthesis, Nitrogen Fixation, Mineral Accumulation, Seed Yield, and Seed Quality of a Soybean Low-Phytate Line. *Plants* **2019**, *8*, 119. [CrossRef] [PubMed]
91. Jin, J.; Wang, G.; Liu, X.; Pan, X.; Herbert, S.J.; Tang, C. Interaction between Phosphorus Nutrition and Drought on Grain Yield, and Assimilation of Phosphorus and Nitrogen in Two Soybean Cultivars Differing in Protein Concentration in Grains. *J. Plant Nutr.* **2006**, *29*, 1433–1449. [CrossRef]
92. Niedbała, G.; Kozłowski, R.J. Application of Artificial Neural Networks for Multi-Criteria Yield Prediction of Winter Wheat. *J. Agric. Sci. Technol.* **2019**, *21*, 51–61.
93. Wu, W.; Ma, B.-L.; Fan, J.-J.; Sun, M.; Yi, Y.; Guo, W.-S.; Voldeng, H.D. Management of nitrogen fertilization to balance reducing lodging risk and increasing yield and protein content in spring wheat. *Field Crop. Res.* **2019**, *241*, 107584. [CrossRef]

agriculture

MDPI

Article

Automatic Disease Detection of Basal Stem Rot Using Deep Learning and Hyperspectral Imaging

Lai Zhi Yong [1], Siti Khairunniza-Bejo [1,2,3,*], Mahirah Jahari [1,2] and Farrah Melissa Muharam [4]

[1] Department of Biological and Agricultural Engineering, Faculty of Engineering, Universiti Putra Malaysia (UPM), Serdang 43400, Selangor, Malaysia
[2] Smart Farming Technology Research Centre, Universiti Putra Malaysia (UPM), Serdang 43400, Selangor, Malaysia
[3] Institute of Plantation Studies, Universiti Putra Malaysia (UPM), Serdang 43400, Selangor, Malaysia
[4] Department of Agriculture Technology, Faculty of Agriculture, Universiti Putra Malaysia (UPM), Serdang 43400, Selangor, Malaysia
* Correspondence: skbejo@upm.edu.my

Abstract: Basal Stem Rot (BSR), a disease caused by *Ganoderma boninense* (*G. boninense*), has posed a significant concern for the oil palm industry, particularly in Southeast Asia, as it has the potential to cause substantial economic losses. The breeding programme is currently searching for *G. boninense*-resistant planting materials, which has necessitated intense manual screening in the nursery to track the progression of disease development in response to different treatments. The combination of hyperspectral image and machine learning approaches has a high detection potential for BSR. However, manual feature selection is still required to construct a detection model. Therefore, the objective of this study is to establish an automatic BSR detection at the seedling stage using a pre-trained deep learning model and hyperspectral images. The aerial view image of an oil palm seedling is divided into three regions in order to determine if there is any substantial spectral change across leaf positions. To investigate if the background images affect the performance of the detection, segmented images of the plant seedling have been automatically generated using a Mask Region-based Convolutional Neural Network (RCNN). Consequently, three models are utilised to detect BSR: a convolutional neural network that is 16 layers deep (VGG16) model trained on a segmented image; and VGG16 and Mask RCNN models both trained on the original images. The results indicate that the VGG16 model trained with the original images at 938 nm wavelength performed the best in terms of accuracy (91.93%), precision (94.32%), recall (89.26%), and F1 score (91.72%). This method revealed that users may detect BSR automatically without having to manually extract image attributes before detection.

Keywords: automatic disease detection; *Ganoderma boninense*; hyperspectral imaging; deep learning

Citation: Yong, L.Z.; Khairunniza-Bejo, S.; Jahari, M.; Muharam, F.M. Automatic Disease Detection of Basal Stem Rot Using Deep Learning and Hyperspectral Imaging. *Agriculture* **2023**, *13*, 69. https://doi.org/10.3390/agriculture13010069

Academic Editors: Gniewko Niedbała and Sebastian Kujawa

Received: 28 November 2022
Revised: 15 December 2022
Accepted: 21 December 2022
Published: 26 December 2022

1. Introduction

Ganoderma boninense (G. boninense) which can cause Basal Stem Rot (BSR) disease has been threatening the oil palm plantations in Southeast Asia for decades and it can cause up to USD 500 million loss annually [1–3]. The young palm trees usually die within 2 years after the first shown symptoms such as yellowing and necrotic leaves, small canopy, stunted growth and unopen spear [4,5]. Proper management and monitoring of the oil palm plantation can help control BSR. One of the approaches is to introduce *Ganoderma* spp. resistant planting materials. The breeding of *Ganoderma* spp. resistant planting materials not only reduces the economic impact of the yield loss, but also helps to create a more sustainable production [6–8]. According to Turnbull et al. [8], if young plants are planted too close to infected palms from the previous generation, the infection might occur much earlier, thus leading to plantation expansion. By introducing *Ganoderma* spp. resistant planting materials, the need for expansion might be avoided, which can help preserve the forest, reducing its environmental impact.

For breeding purposes in a nursery, the worker has to screen the seedlings to differentiate BSR-infected and uninfected seedlings manually. Commonly, for confirmation of the infection, laboratory-based test such as the polymerase chain reaction (PCR), immunofluorescence (IF), fluorescence in situ hybridisation (FISH), enzyme-linked immunosorbent assay (ELISA), flow cytometry (FCM) and gas chromatography–mass spectrometry (GC-MS) may be used, which involves destruction of the samples [9]. Several sensing methods have been developed for non-destructive detection of BSR, including methods involving tomography [10–12], e-nose [13,14], spectroscopy [15–17], thermal imaging [18,19], hyperspectral imaging [1,20–23], lidar [24], terrestrial laser scanning [25], and soil sensing [26].

Li et al. [27] has undertaken a thorough review of various types of plant disease detection using deep learning, and concluded that deep learning is capable of identifying plant leaf diseases with high accuracy. Together with hyperspectral imaging, an early detection of plant disease may be obtained. Furthermore, transfer learning and hyperspectral imaging may be used in rice disease detection [28]. A self-designed CNN model was trained to classify disease with one variety of rice, and the learning was transferred to another three varieties of rice. Fine-tuning, deep Correlation Alignment (CORAL), and deep domain confusion (DDC) were three deep transfer learning approaches that were applied, and they each produced accuracy results of up to 93.33%, 86.67%, and 83.33%, respectively. Although there is currently little research performed on hyperspectral imaging in deep learning, the method has been widely applied to RGB images. Su et al. [29] used Mask RCNN for Fusarium head blight in wheat and it has been shown that the Mask RCNN has high potential for disease detection, achieving precision, recall, F1 score and detection rate at 72.10%, 76.16%, 74.04% and 98.81%, respectively. Feasibility of other state-of-the-art methods such as Faster RCNN, you only look once version 4 (YOLOv4), CenterNet, DetectoRS, Cascade RCNN, Foveabox and Deformabe Detr on the detection of diseased citrus were also studied, where it was indicated that the deep learning-based technique showed good performance in the detection of early stage citrus leaf diseases with CenterNet having the highest accuracy. YOLOv4 had the fastest detection among the models studied [30]. The single shot detector (SSD) was also studied for anthracnose infection on walnut tress where it achieved 87% accuracy on the validation dataset [31]. Aside from all the object detection techniques mentioned, the Mask RCNN was the only method that masked the object rather than only creating bounding boxes around the object, allowing it to be used for image segmentation. Other than Mask RCNN, research studies had been undertaken for VGG16 in plant disease detection. VGG-based transfer learning was able to achieve high average accuracy for cucumber leaf images for seven viral diseases at 93.6% [32]. According to Rahman et al. [33], a fine-tuned VGG16 model achieved a 97.12% accuracy in classifying six classes of pest and disease from rice leaf images. VGG16 was also studied for several disease classifications in tomato plants, where it achieved a net accuracy of 97.23% for seven classes of the diseases studied [34].

Based on the literature review, it can be concluded that hyperspectral imaging has the capacity to detect BSR in the Near-infrared (NIR) region. However, the application of deep learning was not thoroughly studied for BSR detection. Furthermore, no research has been conducted on plant disease identification utilising a pre-trained and widely available deep learning model in the Tensorflow model zoo in conjunction with hyperspectral imagery.

Consequently, the objective of this study is to develop a deep learning model for BSR detection in oil palm plant seedlings that does not rely on human feature extractions using pre-trained models from the Tensorflow model zoo. It contains an analysis of the effect of leaf geometry on wavelength reflectance, an analysis of the effect of image segmentation on model performance, and the identification of the most appropriate model for BSR detection.

2. Materials and Methods

As shown in Figure 1, the overall flowchart of the study includes data pre-processing, wavelength selection for BSR detection and background removal, image generation, and augmentation, as well as model development and comparisons. All the analyses in this

study were undertaken on a machine with Intel core i7th generation CPU with an NVIDIA GeForce RTX 2070.

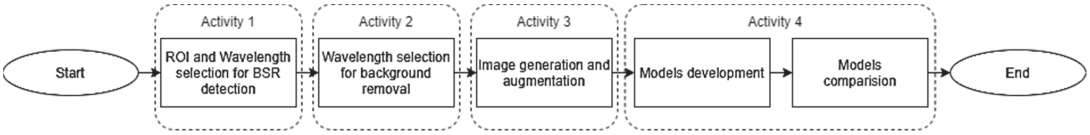


Figure 1. Flowchart of the process involved in this study.

2.1. Dataset

This study used the same hyperspectral images as the study completed in [22], where the images of 10-month-old oil palm seedlings were captured inside a glasshouse from 11.00 am to 2.00 pm on a sunny day using a Cubert FireflEYE S185 (Cubert Gmbh, Ulm, Germany) snapshot hyperspectral camera at 2.6 m constant height from the ground with a black background. All the oil palm seedlings were subjected to the same treatment, in a controlled environment with constant temperature and humidity. Therefore, the difference in the sample reflectance was assumed to be due to the BSR infection. Sixteen bands that demonstrated great performance regardless of the frond number during BSR disease detection using machine learning techniques in [22] were used in this study, i.e., 890 nm, 894 nm, 898 nm, 902 nm, 906 nm, 910 nm, 914 nm, 918 nm, 922 nm, 926 nm, 930 nm, 934 nm, 938 nm, 942 nm, 946 nm, and 950 nm. These bands were extracted from the bands that have shown great separation value between the infected and uninfected plant seedlings. Further, the bands were also tested for any significant difference between the infected and uninfected plant seedlings using a t-test in SPSS statistical software (IBM SPSS Statistics 25, IBM, New York, NY, USA) in which the p-values of the tests were less than 0.05 and obtained good coefficients of variance that were between 5 and 14%.

2.2. Region of Interest (RoI) and Wavelength Selection for BSR Detection

The disease slowly destroys the vascular system and causes symptoms of water and nutrient deficiency. Leaf characteristics display different changing trends under nutritional stress [35]. Therefore, the question if the location of points on the fronds provides significant differences to spectral reflectance, which may thus influence the plant status at early infection, was first investigated. In addition, according to [36], a reflectance spectra is very sensitive to the geometry of the plant. Therefore, to check if there is any effect of the point reflectance at different positions of the leaf structure, an aerial view of an oil palm seedling was divided into three RoIs as shown in Figure 2 and defined as follows:

A: Inner region—2 cm from the centre of the seedling to 5 cm square.
B: Middle region—5 cm from the centre of the seedling to 8 cm square.
C: Outer region—8 cm from centre of the seedling to 11 cm square.

A total of 693 points of reflectance were extracted randomly from 72 hyperspectral images in the *.cub file format using Cube-Pilot software (Cube-Pilot 1.5.8, Cubert Gmbh, Ulm, Germany), where 36 images were obtained from healthy seedlings and 36 images were obtained from infected seedlings. A box plot method was used to remove points that exceeded the lower and upper fences. As a result, only 668 points were left.

Analysis of Variance (ANOVA) was used to check the significant difference between the data groups. However, ANOVA assumes normality. Therefore, before conducting an ANOVA, the modified Wilk–Shapiro test that allows a sample size of more than 30 was used to check for normality [37–40]. First, the range of data was adjusted based on $\mu \pm s$, where μ is the mean, s is the upper and lower limit testing with different values, i.e., s starting from 2σ until 0.5σ with 0.5 decrement where σ is the standard deviation. If there were no wavelengths showing normality until $s = 0.5\sigma$, the data went through the second stage which involved data transformation utilising log and reflection transformation. A summary of the process of the normality check and outlier removal is shown in Figure 3.

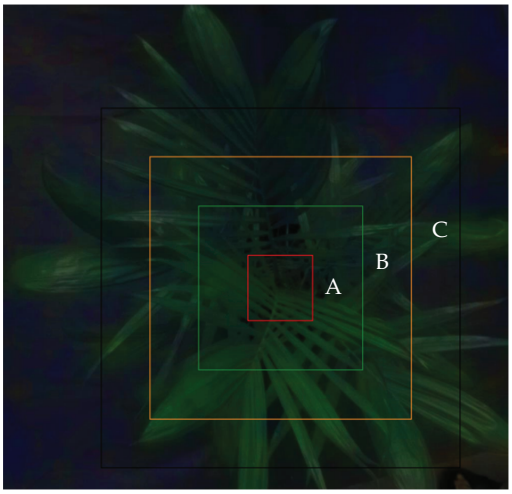


Figure 2. The three regions of interest.

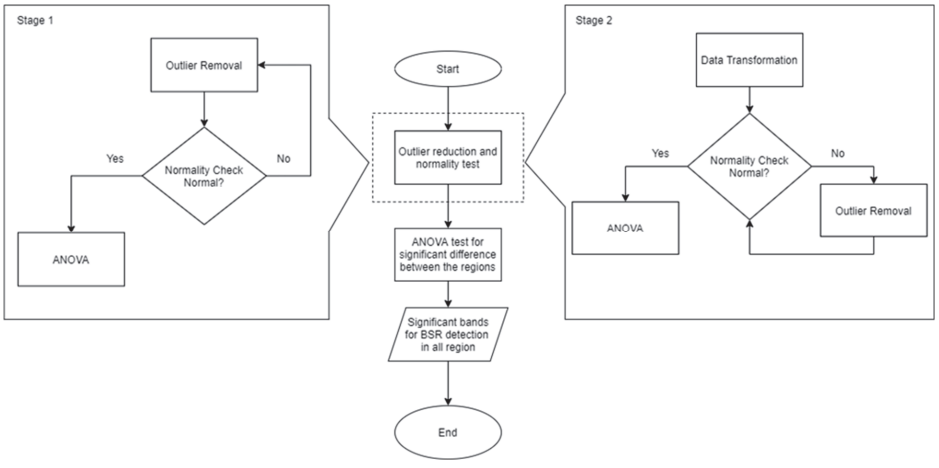


Figure 3. Flowchart of normality check and outlier removal.

After the selection of suitable wavelength(s), the wavelength(s) was (were) used to generate images with 1000 by 1000 pixels as input for the model for BSR detection. The details of image generation are discussed in Section 2.4.1.

2.3. Wavelength Selection for Background Removal

As mentioned earlier, the effect of segmentation on the model performance was studied. Therefore, a suitable wavelength for image segmentation to generate clear images for background removal was determined based on the difference in reflectance between the plant seedling and the background. The separation was calculated as in Equation (1), where $\overline{Ref_{plant}}$ is the average reflectance of the points and $\overline{Ref_{background}}$ is the average reflectance of the background at wavelength i.

$$Separation_i = \left| \overline{Ref_{plant,i}} - \overline{Ref_{background,i}} \right|. \quad (1)$$

The selected wavelengths were used to generate images with 1000 by 1000 pixels for segmentation purposes and the details of image generation are discussed in Section 2.4.2.

2.4. Image Generation and Augmentation

After the wavelengths were selected for both BSR detection and background removal, two sets of images were generated. Each set of images was used for different purposes, and the images were fed to different models. The details are discussed in the model development in Section 2.5.

2.4.1. Image Generation for BSR Detection

As discussed in Section 2.2, the ANOVA was used to identify the significant differences between each RoI. This was done to identify which wavelength had a more consistent performance regardless of the geometry of the leaves. The wavelength(s) which had no significant difference between each RoI was (were) chosen. The images were generated by extracting the chosen wavelength from the hyperspectral images using Python. The images were then augmented using rotation, zoom, horizontal and vertical flip. Each set of images was augmented to a total of 1610 images, where 1127 were used for training and 483 for testing. In this study, these images were defined as original images.

2.4.2. Image Generation for Background Removal

The images generated in Section 2.3 were not clear enough for background removal. Therefore, in order to perform an image segmentation, the wavelength with the greatest separation of reflectance was used to generate the red, green, blue (RGB) images (*.jpg files) using Python. In addition to the RGB images, the grayscale images generated by the Cube Pilot software (Cube-Pilot 1.5.8, Cubert Gmbh, Ulm, Germany) were used as an alpha channel (A), which described the transparency of each pixel, to generate a red–green–blue–alpha (RGBA) image. The additional alpha channel produced more differences between each pixel, resulting in clearer images. The purpose of the RGBA images was solely for segmentation, to be applied as mask images of the original images. The RGBA images were then augmented using rotation, zoom, horizontal and vertical flip. Each set of RGBA images was augmented to a total of 1610 images, where 1127 were used for training and 483 for testing.

2.5. Model Development

This section discussed the architecture of the models as well as the workflow for each of the models developed.

2.5.1. Model Architecture

There were three models used in this study, which were developed from the pre-trained VGG16 and Mask RCNN from the Tensorflow model zoo. This section discusses the general architecture of VGG16 and Mask RCNN.

a. VGG16

The VGG16 model was readily available in Tensorflow and the model was pre-trained with an ImageNet dataset. The architecture of the VGG16 model is shown in Figure 4. VGG16 consisted of five convolutional blocks where each block consisted of convolution layers and a max pooling layer. The five blocks were feature extractors. A dense block was connected to the last convolutional block and the dense layer was a classifier. In this study, the weight of the pre-trained VGG16 model was frozen and was used as a feature extractor and a new dense layer was connected to the convolutional block. The weight of the dense layer was allowed to be trained. This meant that only the pre-trained weight of the convolutional blocks was used and a new classifier was trained with the plant seedling dataset for BSR detection. The model was trained after there was no improvement in the loss for five consecutive epochs. The stopping criterion was set by trial and error. It was

found that after the five consecutive epochs, signs of model overfitting occurred as shown in Figure 5.

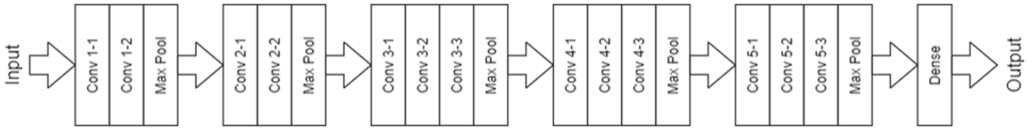


Figure 4. The architecture of VGG16.

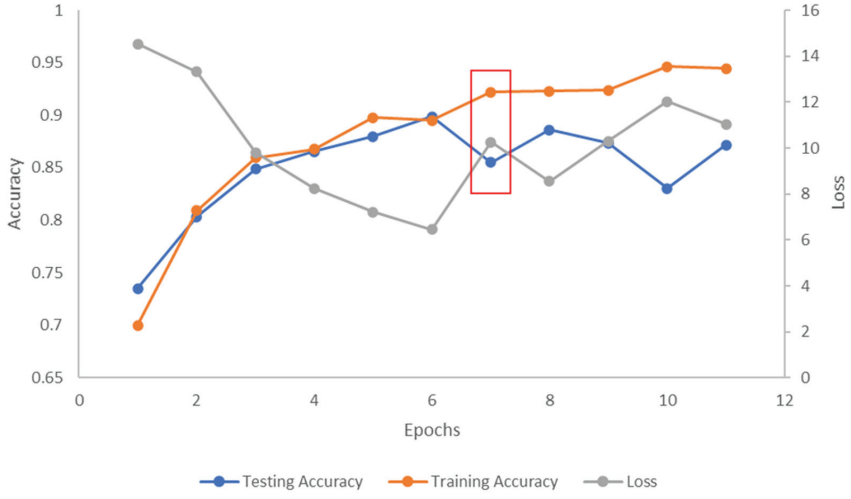


Figure 5. Graph of training accuracy, testing accuracy and loss versus epochs. The red box is the epochs after no improvement in the loss for five consecutive epochs.

b. Mask RCNN

The Mask RCNN is an extension to the Faster RCNN which consisted of two stages. The first stage was a region proposal network (RPN) and the second stage involved class, box offset, and binary mask prediction. The pre-trained Mask RCNN available in the Tensorflow model zoo had a Inception ResNet v2 backbone [41]. The Mask RCNN was an approach for instance segmentation which was based on an instance-first strategy rather than a segmentation-first strategy adopted by other segmentation algorithms [42]. This meant that after the first stage (RPN), in parallel with predicting the class and the box offsets, a mask was also produced. Hence, unlike other algorithms, Mask RCNN does not depend on the mask prediction for classification. Further, most other object detection algorithms utilise RoI pooling for extracting feature maps which compromise the amount of information as quantisation was involved. However, the Mask RCNN used RoIAlign which calculated the value of each sampling points with bilinear interpolation which resulted in the exact value of each sampling point and no quantisation was performed. This allowed the Mask RCNN to predict the mask more accurately. The framework of the Mask RCNN is shown in Figure 6. The Mask RCNN was trained until there was no improvement in the total loss after 3000 consecutive epochs. The stopping criterion was set by trial and error. The epoch with lower loss was taken as the new starting epoch and it was found that by having 3000 consecutive epochs, the lowest total loss could be achieved as shown in Figure 7.

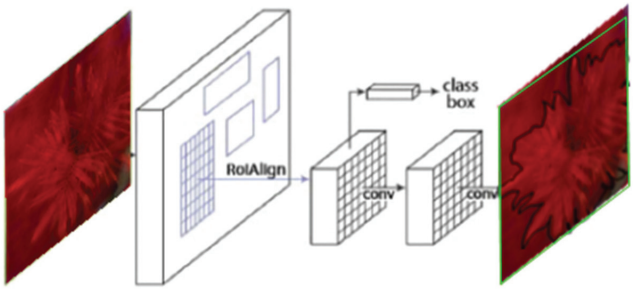


Figure 6. The framework of Mask RCNN, which shows that the classification and localisation did not depend on the mask prediction.

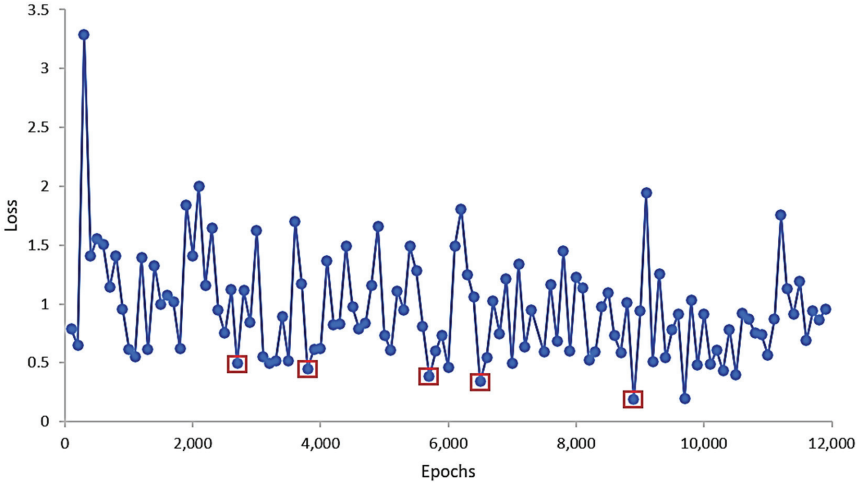


Figure 7. Graph of total loss versus epochs. The red boxes are the lowest loss within each 3000 epochs.

2.5.2. Detection Models

Figure 8 shows the workflow of each of the three models used in this study. Descriptions of each model are as follows:

a. VGG16

The first model used was the VGG16 with the original images generated as in Section 2.4.1 to classify the BSR infected and uninfected seedlings.

b. Mask RCNN + VGG16

The second model consisted of two stages. The first stage involved creating a mask using the Mask RCNN while the second stage involved classifying the BSR-infected and uninfected seedlings using the VGG 16 trained with the segmented images.

Regarding the first stage of the model, before feeding the images to the Mask RCNN, the RGBA images generated in Section 2.4.2 were digitised and labelled using the "labelme" annotation tool as shown in Figure 9. The images were labelled as frond regardless of the infection condition of the seedling as the purpose of the Mask RCNN was solely to provide the mask images for background removal of the original images. Therefore, the output of the Mask RCNN was a segmented image which consisted of the canopy of the oil palm seedling as the RoI. In the second stage of the model, these segmented images were used to train the VGG16 model to classify the BSR-infected and uninfected seedlings.

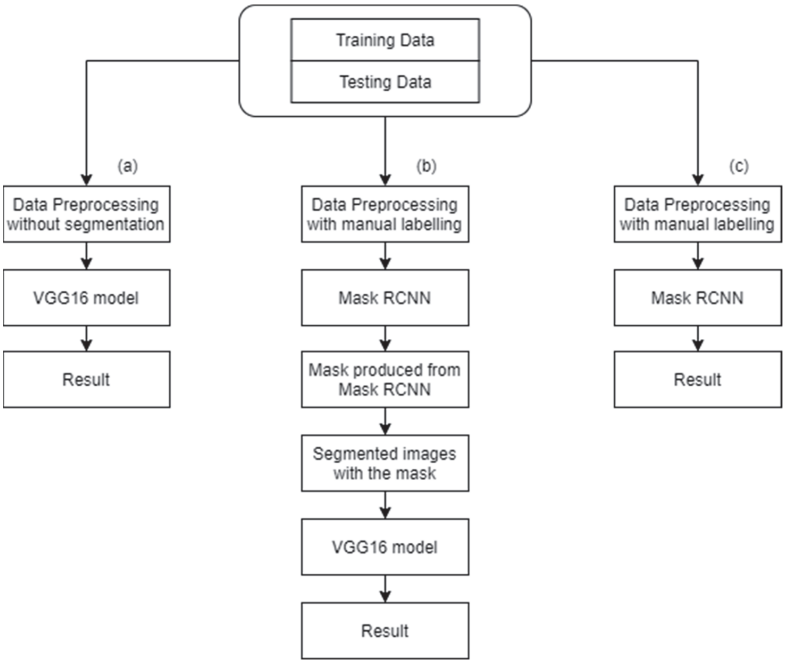


Figure 8. Workflow of each classification model. (**a**) VGG16 model trained with unsegmented images, (**b**) VGG16 trained with segmented images produced by Mask RCNN, (**c**) Mask RCNN trained with manually labelled images.

Figure 9. Example of the digitised labelled image of an oil palm seedling.

c. Mask RCNN

The third model emphasised the use of Mask RCNN for object detection, i.e., infected and uninfected seedlings. The original images were digitised and labelled according to the infection conditions with the "labelme" annotation tool. The output of the model was the class of the infection condition, i.e., infected or uninfected.

2.6. Performance Evaluation of the Models

The intersection of union (IoU) of the mask produced by the mask RCNN and the manually labelled images were calculated as in Equation (2).

$$IoU = \frac{S_i \cap G_i}{S_i \cup G_i} \tag{2}$$

where S_i is segmented image of image I and G_i is the ground truth of image i.

The IoU of the segmented images with the ground truth represented the amount of overlap between the segmented images and the ground truth, where the greater the overlap, the greater the IoU.

Meanwhile, the value of accuracy, precision, recall, specificity and F1 score of each model were calculated as in Equations (3)–(7), respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

$$Precision = \frac{TP}{TP + FP}, \tag{4}$$

$$Recall = \frac{TP}{TP + FN}, \tag{5}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{6}$$

$$F1\ Score = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity}, \tag{7}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. The accuracy described the overall correctness of the model for classifying both infected and uninfected seedlings. The precision indicated the percentage of truly infected plant seedlings from the detected infected plant seedlings. The recall, on the other hand, represented the percentage of truly infected plant seedlings detected correctly from all the truly infected plant seedlings. Meanwhile, specificity illustrated the percentage of truly uninfected plant seedlings detected from all the truly uninfected plant seedlings. Besides precision and recall, the F1 score was calculated to obtain the harmonic mean of the precision and recall.

3. Results

3.1. Identified Wavelength for BSR Detection

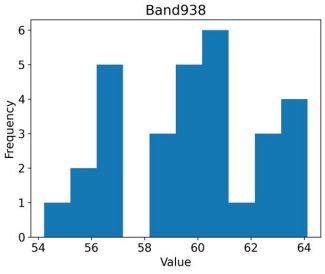
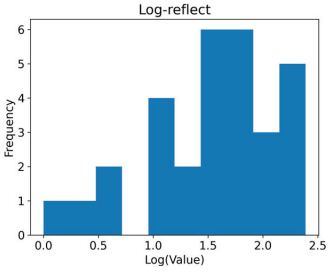
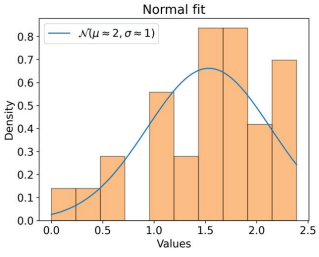
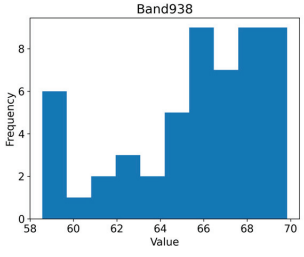
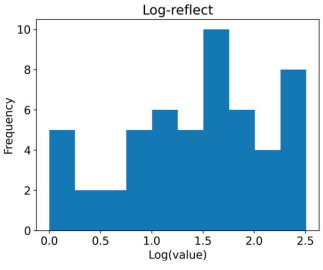
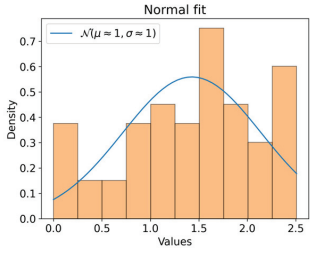
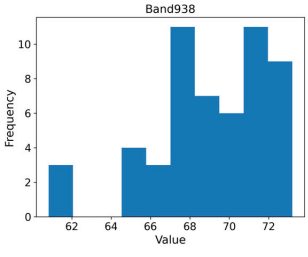
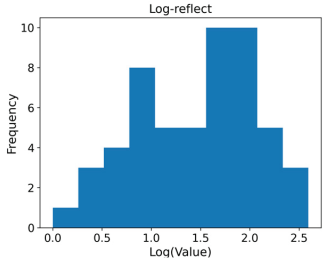
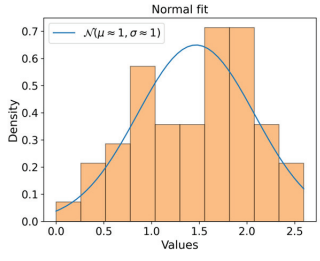
Results of the Wilk–Shapiro test for the infected dataset indicated that after performing the log and reflection transformations, some wavelengths demonstrated normality as the data ranges narrowed. Table 1 illustrates the wavelength and data range at which normality was seen in each region after data transformation. After the data were limited to $\mu \pm 0.5\sigma$, the wavelength at 938 nm was found to be normal in all regions. The distribution of selected data before and after transformation in each region at wavelength 938 nm is presented in Table 2. It was demonstrated that the distribution of the data was skewed, and following transformation, the data were fitted to a normal distribution. The data at wavelength 938 nm were analysed using ANOVA and it was shown that there was no significant difference between all the regions at a confidence level of 95% where the p-value was 0.1629.

Table 1. Wavelengths extracted from infected images that were successfully transformed to a normal distribution after data reduction.

Dataset	Wavelength (nm) with Normal Distribution after Transformation		
	RoI = C	**RoI = B**	**RoI = A**
$\mu \pm 2\sigma$	None	None	None
$\mu \pm 1.5\sigma$	None	None	None
$\mu \pm 1\sigma$	910, 914, 918, 922, 926, 930, 934, 938, 942, 946, 950	890, 894, 898,902	None
$\mu \pm 0.5\sigma$	906, 934, **938**, 942	890, 894, 898, 902, 906, 910, 914, 918, 922, 930, 934, **938**, 942, 946	**938**

Note: A: Inner region—2 cm from the centre of the seedling to 5 cm square, B: Middle region—5 cm from the centre of the seedling to 8 cm square, C: Outer region—8 cm from centre of the seedling to 11 cm square.

Table 2. Data distribution of the original and transformed data for the RoIs A, B and C at wavelength 938 nm at $\mu \pm 0.5\sigma$.

RoI	Original	Transform	Normal Fit
A			
B			
C			

Note: A: Inner region—2 cm from the centre of the seedling to 5 cm square, B: Middle region—5 cm from the centre of the seedling to 8 cm square, C: Outer region—8 cm from centre of the seedling to 11 cm square.

For the uninfected dataset, none of the wavelengths showed normality even after Stage 2 assessment. Therefore, the identified suitable wavelength in the infected dataset, i.e., 938 nm was used to test the significant difference between the data of the three regions

using a Levene's test and a Kruskal–Wallis test. With the results of p-values equal to 0.4234 and 0.3088, respectively, the Levene's test and the Kruskal–Wallis test indicated that there was no significant difference between the three regions. Since there was no effect of leaf geometry on wavelength reflectance at 938nm, the whole canopy images were used to develop the detection models. Further, by reducing the number of bands, the complexity and cost of developing future hardware may be reduced.

3.2. Image Segmentation for Background Removal

3.2.1. Identified Wavelengths for Background Removal

Figure 10 shows a graph of the average reflectance of the background (AVG Back), the average reflectance of the plant seedlings (AVG Frond), and the difference between the two (AVG Delta). Reflectance differed most between the red-edge and NIR spectrum, particularly at wavelengths 766 nm, 762 nm, and 770 nm. However, Figure 11a demonstrates that the images created using these wavelengths were insufficiently clear to distinguish between the background and seedling. Therefore, the wavelength with the greatest difference in each red, green and blue spectrum marked in the red box, i.e., at 750 nm (i.e., 10.08% difference), 554 nm (i.e., 3.54% difference), and 466 nm (i.e., 1.37% difference) were chosen. Although the image became better, as seen in Figure 11b, it was not good enough for manual labelling of the contours of the plant seedlings. As illustrated in Figure 11c, when the grayscale image generated automatically by the Cube Pilot software (Cubert Gmbh, Germany) was added into the alpha channel (A), the contrast between canopy and background images became higher. Therefore, the RGBA image was used as the input image of Mask RCNN for the background removal task.

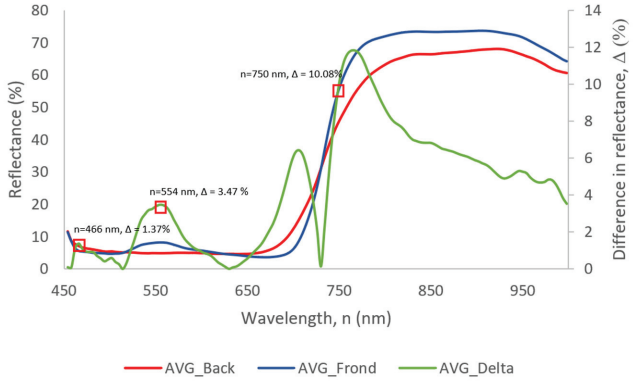
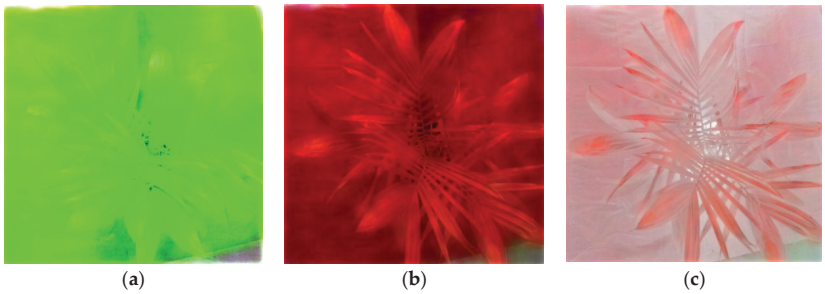


Figure 10. Graph of reflectance of the plant seedling and background against bands wavelength.



(**a**) (**b**) (**c**)

Figure 11. Comparison of images generated using (**a**) wavelengths 766 nm, 762 nm and 770 nm and (**b**) images generated using wavelengths 750 nm, 554 nm and 466 nm without the grayscale layer added in the alpha channel and (**c**) with grayscale layer added in the alpha channel.

3.2.2. Performance of the Mask RCNN for Generating Segmented Images

As shown in Figure 12, the final overall loss of the Mask RCNN was 0.2450, with the mask loss at 0.2409 accounting for the majority of the loss. It was demonstrated that the mask (segmented image) rather than the classification of background and foreground was the primary cause of the loss.

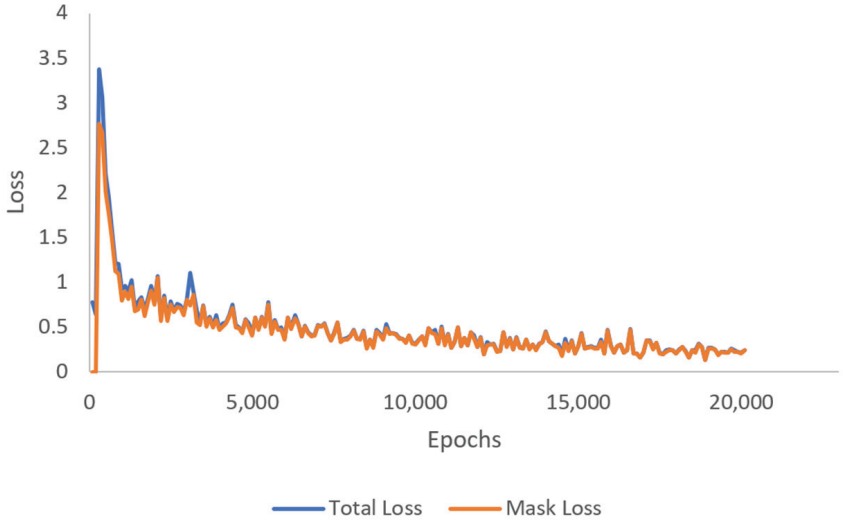


Figure 12. Graph of total loss and mask loss versus epochs of Mask RCNN.

The results also demonstrated that the segmented image had an average IoU of 0.8606. A few samples of the segmented image and the ground truth are presented in Figure 13. The IoU of the segmented image was satisfactory in comparison with the other published research. According to [43], the segmentation of a tree crown was deemed proper if the IoU was larger than or equal to 0.5. In addition, [44] employed semantic segmentation for the detection of apple, peach, and pear flowers, with reported IoU values ranging from 0.001 to 0.811 for several deep learning models. Consequently, this demonstrated that the segmented images generated in this study by using the Mask RCNN were acceptable.

In addition to IoU, the values of accuracy, precision, recall, specificity, and F1 score of the Mask RCNN for image segmentation were computed and tabulated as in Table 3, with plant seedling pixel detected as plant seedling pixel as TP, background pixel detected as background pixel as TN, background pixel detected as plant seedling pixel as FP, and plant seedling pixel detected as background as FN. The results demonstrated that 90.48% of the pixels were accurately identified, and 94.63% of the anticipated seedling pixels were in fact seedlings.

Table 3. Performance of Mask RCNN for image segmentation.

Accuracy	Precision	Recall	Specificity	F1 Score
90.48%	94.63%	90.49%	89.25%	92.51%

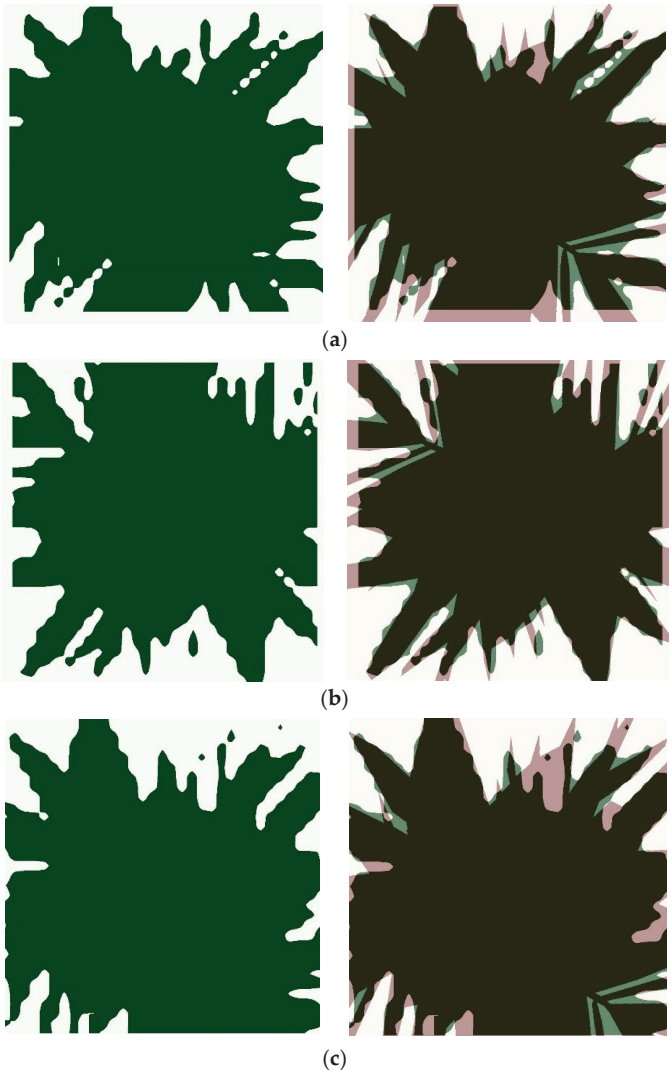


Figure 13. Example of the segmented images at different results of IoU. Figures on the left are the mask predicted using the Mask RCNN, while figures on the right are the predicted mask (green) overlayed on the actual mask (red). (**a**) IoU = 0.8598, (**b**) IoU = 0.8829 and (**c**) IoU = 0.8944.

3.3. Performance of the BSR Detection Models

Tabel 4 shows the results of the precision, recall, specificity, F1 score, and detection time of the models. It has been demonstrated that the VGG16 model trained on the original images had the highest accuracy (91.93%), precision (94.32%), specificity (94.61%), and F1 score (91.72%). Nevertheless, the VGG16 model trained with segmented images exhibited a better recall (95.02%) than the VGG16 model trained with original images (5.76% difference). The Mask RCNN, on the other hand, had the poorest performance, as only 71.42% of the images were properly identified, but it obtained a 100% recall, as all of the actually diseased seedlings were labelled as such. With a specificity value of 42.74%, Mask RCNN was unable to correctly classify uninfected plant seedlings, as they were classified as infected.

In addition, it was demonstrated that segmentation did not enhance the performance of the model. Despite the limitation of comparable research on the identification of plant diseases, similar findings have been observed in a number of other studies. For example, the authors of [45] reported that the segmentation of skin lesions for dermatoscopic image categorisation reduced the performance of EfficientNet. Furthermore, it was revealed that segmentation-free CNN models performed much better than segmentation-dependent models in the diagnosis of breast masses in mammography datasets [46]. Despite the fact that numerous studies have demonstrated that segmenting the RoI improves model performance, the study only considered Support Vector Machine (SVM) and not CNN [47]. Manual segmentation improved the performance of the SVM model for anatomical magnetic resonance imaging, but no other automatic feature selection approach outperformed the unsegmented image, as demonstrated by the study. In addition, research indicated that feature selection may increase model performance. However, this relies on the employed model and dataset [48]. According to [48], NB, ANN, and a multilayer perceptron did not always obtain improved performance after feature selection, as model performance was dependent on the dataset.

Table 4 also indicates the time required to classify each image. It was discovered that VGG 16 could categorise an image in 0.08 s, but Mask RCNN could only do so in 1.59 s. This was consistent with the difficulty of the task for each model type, as the VGG 16 just classified the images without constructing a mask or bounding box, but the Mask RCNN produced a mask and bounding box for each detected image.

Table 4. The performance of the BSR detection models using images extracted from 938 nm wavelength in all regions.

Model	Segmentation	Accuracy	Precision	Recall	Specificity	F1 Score	Average Time for Classification (s/Image)
VGG 16	No	**91.93 %**	**94.32%**	89.26%	**94.61%**	**91.72%**	0.08
VGG 16	Automatic	85.46%	79.79%	95.02%	75.93%	86.74%	0.08
Mask RCNN	Manual labelling	71.43%	63.68%	**100.00%**	42.74%	77.81%	1.59

4. Conclusions

This paper demonstrates the potential of deep learning to automatically detect an early stage of the BSR disease in oil palm seedlings using NIR-hyperspectral imaging. After data transformation and outlier elimination, it was discovered that the entire structure of an aerial view image of a seedling at 938 nm wavelength may be used for detection, as there are no significant differences in any of the RoI. VGG16 trained with the original images accurately classified BSR-infected plant seedlings with an accuracy of 91.93%. Meanwhile, the Mask RCNN trained using RGBA images correctly segmented the aerial view image of the seedling from the background with an average IoU of 0.8606. This study concluded that the best model for BSR identification is the VGG16 model trained using original images, which allows for more automatic BSR detection as reflectance point extraction is not required. However, this study has limitations due to the controlled environment in which the data was collected. In addition, the data collected only accounted for 10-month-old seedlings which might not represent the unseen data from different growth periods in the nursery. In addition, the other wavelengths that did not exhibit normality were not explored in detail. Therefore, further study can be conducted by putting the model to the test in realistic environments, and the non-normal bands can be further investigated.

Author Contributions: Conceptualisation, L.Z.Y. and S.K.-B.; methodology, L.Z.Y., S.K.-B. and F.M.M.; software, L.Z.Y.; validation, L.Z.Y. and S.K.-B.; formal analysis, L.Z.Y. and S.K.-B.; investigation, L.Z.Y. and S.K.-B.; resources, S.K.-B.; data curation, M.J.; writing—original draft preparation,

L.Z.Y.; writing—review and editing, S.K.-B.; visualisation, L.Z.Y.; supervision, S.K.-B., M.J. and F.M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Azmi, A.N.N.; Khairunniza-Bejo, S.; Jahari, M.; Muharram, F.M.; Yule, I. Identification of a Suitable Machine Learning Model for Detection of Asymptomatic Ganoderma Boninense Infection in Oil Palm Seedlings Using Hyperspectral Data. *Appl. Sci.* **2021**, *11*, 11798. [CrossRef]
2. Flood, J.; Hasan, Y.; Foster, H. Ganoderma Diseases of Oil Palm—An Interpretation from Bah Lias Research Station. *Planter* **2002**, *78*, 689–696, 699–706, 709–710.
3. Ishaq, I.; Alias, M.S.; Kadir, J.; Kasawani, I. Detection of Basal Stem Rot Disease at Oil Palm Plantations Using Sonic Tomography. *J. Sustain. Sci. Manag.* **2014**, *9*, 52–57.
4. Castillo, S.Y.; Rodríguez, M.C.; González, L.F.; Zúñiga, L.F.; Mestizo, Y.A.; Medina, H.C.; Montoya, C.; Morales, A.; Romero, H.M.; Sarria, G.A. Ganoderma Zonatum Is the Causal Agent of Basal Stem Rot in Oil Palm in Colombia. *J. Fungi* **2022**, *8*, 230. [CrossRef]
5. Naher, L.; Seri Intan, M.; Noorhazira, S. Trichoderma Harzianum T32 Growth and Antagonistic Performance against Ganoderma Boninense On Different Culture Media. In Proceedings of the 3rd International Conference on Biological, Chemical and Environmental Sciences (BCES-2015), Kuala Lumpur, Malaysia, 21–22 September 2015.
6. Murphy, D.J.; Goggin, K.; Paterson, R.R.M. Oil Palm in the 2020s and beyond: Challenges and Solutions. *CABI Agric. Biosci.* **2021**, *2*, 39. [CrossRef]
7. Naher, L.; Yusuf, U.K.; Tan, S.G.; Ismail, A. Ecological Status of Ganoderma and Basal Stem Rot Disease of Oil Palms (Elaeis Guineensis Jacq.). *Aust. J. Crop Sci.* **2013**, *7*, 1723–1727.
8. Turnbull, N.; de Franqueville, H.; Breton, F.; Jeyen, S.; Syahputra, I.; Cochard, B.; Durand-Gasselin, T. Breeding Methodology to Select Oil Palm Planting Material Partially Resistant to *Ganoderma boninense*. In Proceedings of the 5th Quadrennial International Oil Palm Conference, Bali, Indonesia, 17–19 June 2014.
9. Fang, Y.; Ramasamy, R. Current and Prospective Methods for Plant Disease Detection. *Biosensors* **2015**, *5*, 537–561. [CrossRef]
10. Arango, M.; Martínez, G.; Torres, G. Advances in the Interpretation of Tomographic Images as an Early Detection Method of Oil Palm Affected by Basal Stem Rot in Colombia. *Plant Dis.* **2016**, *100*, 1559–1563. [CrossRef]
11. Idris, A.S.; Mazliham, M.S.; Loonis, P.; Wahid, M.B. *MPOB Information Series*; MPOB: Bandar Baru Bangi, Malaysia, 2010.
12. Su'ud, M.M.; Loonis, P.; Seman, I.A. Towards Automatic Recognition and Grading of Ganoderma Infection Pattern Using Fuzzy Systems. *T. Eng. Comput. Technol.* **2007**, *1*, 6.
13. Markom, M.A.; Shakaff, A.Y.M.; Adom, A.H.; Ahmad, M.N.; Hidayat, W.; Abdullah, A.H.; Fikri, N.A. Intelligent Electronic Nose System for Basal Stem Rot Disease Detection. *Comput. Electron. Agric.* **2009**, *66*, 140–146. [CrossRef]
14. Kresnawaty, I.; Mulyatni, A.S.; Eris, D.D.; Prakoso, H.T.; Tri-Panji; Triyana, K.; Widiastuti, H. Electronic Nose for Early Detection of Basal Stem Rot Caused by Ganoderma in Oil Palm. *IOP Conf. Ser. Earth Environ. Sci.* **2020**, *468*, 012029. [CrossRef]
15. Khaled, A.Y.; Abd Aziz, S.; Bejo, S.K.; Nawi, N.M.; Abu Seman, I. Spectral Features Selection and Classification of Oil Palm Leaves Infected by Basal Stem Rot (BSR) Disease Using Dielectric Spectroscopy. *Comput. Electron. Agric.* **2018**, *144*, 297–309. [CrossRef]
16. Khaled, A.Y.; Abd Aziz, S.; Bejo, S.K.; Mat Nawi, N.; Abu Seman, I. Artificial Intelligence for Spectral Classification to Identify the Basal Stem Rot Disease in Oil Palm Using Dielectric Spectroscopy Measurements. *Trop. Plant Pathol.* **2022**, *47*, 140–151. [CrossRef]
17. Liaghat, S.; Mansor, S.; Ehsani, R.; Shafri, H.Z.M.; Meon, S.; Sankaran, S. Mid-Infrared Spectroscopy for Early Detection of Basal Stem Rot Disease in Oil Palm. *Comput. Electron. Agric.* **2014**, *101*, 48–54. [CrossRef]
18. Johari, S.N.Á.M.; Bejo, S.K.; Lajis, G.A.; DaimDai, L.D.J.; Keat, N.B.; Ci, Y.Y.; Ithnin, N. Detecting BSR-Infected Oil Palm Seedlings Using Thermal Imaging Technique. *Basrah J. Agric. Sci.* **2021**, *34*, 73–80. [CrossRef]
19. Hashim, I.C.; Shariff, A.R.M.; Bejo, S.K.; Muharam, F.M.; Ahmad, K. Classification of Non-Infected and Infected with Basal Stem Rot Disease Using Thermal Images and Imbalanced Data Approach. *Agronomy* **2021**, *11*, 2373. [CrossRef]
20. Kurihara, J.; Koo, V.-C.; Guey, C.W.; Lee, Y.P.; Abidin, H. Early Detection of Basal Stem Rot Disease in Oil Palm Tree Using Unmanned Aerial Vehicle-Based Hyperspectral Imaging. *Remote Sens.* **2022**, *14*, 799. [CrossRef]
21. Liaghat, S.; Ehsani, R.; Mansor, S.; Shafri, H.Z.M.; Meon, S.; Sankaran, S.; Azam, S.H.M.N. Early Detection of Basal Stem Rot Disease (Ganoderma) in Oil Palms Based on Hyperspectral Reflectance Data Using Pattern Recognition Algorithms. *Int. J. Remote Sens.* **2014**, *35*, 3427–3439. [CrossRef]
22. Noor Azmi, A.N.; Bejo, S.K.; Jahari, M.; Muharam, F.M.; Yule, I.; Husin, N.A. Early Detection of Ganoderma Boninense in Oil Palm Seedlings Using Support Vector Machines. *Remote Sens.* **2020**, *12*, 3920. [CrossRef]
23. Khairunniza-Bejo, S.; Shahibullah, M.S.; Azmi, A.N.N.; Jahari, M. Non-Destructive Detection of Asymptomatic Ganoderma Boninense Infection of Oil Palm Seedlings Using NIR-Hyperspectral Data and Support Vector Machine. *Appl. Sci.* **2021**, *11*, 10878. [CrossRef]

24. Husin, N.A.; Khairunniza-Bejo, S.; Abdullah, A.F.; Kassim, M.S.M.; Ahmad, D.; Azmi, A.N.N. Application of Ground-Based LiDAR for Analysing Oil Palm Canopy Properties on the Occurrence of Basal Stem Rot (BSR) Disease. *Sci. Rep.* **2020**, *10*, 6464. [CrossRef] [PubMed]
25. Husin, N.A.; Khairunniza-Bejo, S.; Abdullah, A.F.; Kassim, M.S.M.; Ahmad, D. Multi-Temporal Analysis of Terrestrial Laser Scanning Data to Detect Basal Stem Rot in Oil Palm Trees. *Precis. Agric.* **2022**, *23*, 101–126. [CrossRef]
26. Aziz, M.H.A.; Khairunniza-Bejo, S.; Wayayok, A.; Hashim, F.; Kondo, N.; Azmi, A.N.N. Temporal Changes Analysis of Soil Properties Associated with Ganoderma Boninense Pat. Infection in Oil Palm Seedlings in a Controlled Environment. *Agronomy* **2021**, *11*, 2279. [CrossRef]
27. Li, L.; Zhang, S.; Wang, B. Plant Disease Detection and Classification by Deep Learning—A Review. *IEEE Access* **2021**, *9*, 56683–56698. [CrossRef]
28. Feng, L.; Wu, B.; He, Y.; Zhang, C. Hyperspectral Imaging Combined With Deep Transfer Learning for Rice Disease Detection. *Front. Plant Sci.* **2021**, *12*, 693521. [CrossRef]
29. Su, W.-H.; Zhang, J.; Yang, C.; Page, R.; Szinyei, T.; Hirsch, C.D. Evaluation of Mask RCNN for Learning to Detect Fusarium Head Blight in Wheat Images. In Proceedings of the 2020 ASABE Annual International Virtual Meeting, Virtual, 13–15 July 2020; p. 4.
30. Dananjayan, S. Assessment of State-of-the-Art Deep Learning Based Citrus Disease Detection Techniques Using Annotated Optical Leaf Images. *Comput. Electron. Agric.* **2022**, *14*, 106658. [CrossRef]
31. Anagnostis, A. A Deep Learning Approach for Anthracnose Infected Trees Classification in Walnut Orchards. *Comput. Electron. Agric.* **2021**, *11*, 105998. [CrossRef]
32. Fujita, E.; Uga, H.; Kagiwada, S.; Iyatomi, H. A Practical Plant Diagnosis System for Field Leaf Images and Feature Visualization. *Int. J. Eng. Technol.* **2018**, *7*, 49–54. [CrossRef]
33. Rahman, C.R.; Arko, P.S.; Ali, M.E.; Khan, M.A.I.; Apon, S.H.; Nowrin, F.; Wasif, A. Identification and Recognition of Rice Diseases and Pests Using Convolutional Neural Networks. *Biosyst. Eng.* **2020**, *194*, 112–120. [CrossRef]
34. Rangarajan, A.K.; Purushothaman, R.; Ramesh, A. Tomato Crop Disease Classification Using Pre-Trained Deep Learning Algorithm. *Procedia Comput. Sci.* **2018**, *133*, 1040–1047. [CrossRef]
35. Sun, Y.; Tong, C.; He, S.; Wang, K.; Chen, L. Identification of Nitrogen, Phosphorus, and Potassium Deficiencies Based on Temporal Dynamics of Leaf Morphology and Color. *Sustainability* **2018**, *10*, 762. [CrossRef]
36. Shahrimie, M.A.M.; Mishra, P.; Mertens, S.; Dhondt, S.; Wuyts, N.; Scheunders, P. Modeling Effects of Illumination and Plant Geometry on Leaf Reflectance Spectra in Close-Range Hyperspectral Imaging. In Proceedings of the 2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Los Angeles, CA, USA, 21–24 August 2016; pp. 1–4.
37. Razali, N.M.; Wah, Y.B. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. *J. Stat. Model. Anal.* **2011**, *2*, 21–33.
38. Royston, P. Algorithm AS 181: The W Test for Normality. *Appl. Stat.* **1982**, *31*, 176. [CrossRef]
39. Royston, P. Approximating the Shapiro-Wilk W-Test for Non-Normality. *Stat. Comput.* **1992**, *2*, 117–119. [CrossRef]
40. Royston, P. Remark AS R94: A Remark on Algorithm AS 181: The W-Test for Normality. *Appl. Stat.* **1995**, *44*, 547. [CrossRef]
41. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv* **2016**, arXiv:160207261. [CrossRef]
42. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *arXiv* **2018**, arXiv:170306870.
43. Xi, X.; Xia, K.; Yang, Y.; Du, X.; Feng, H. Evaluation of Dimensionality Reduction Methods for Individual Tree Crown Delineation Using Instance Segmentation Network and UAV Multispectral Imagery in Urban Forest. *Comput. Electron. Agric.* **2021**, *191*, 106506. [CrossRef]
44. Sun, K.; Wang, X.; Liu, S.; Liu, C. Apple, Peach, and Pear Flower Detection Using Semantic Segmentation Network and Shape Constraint Level Set. *Comput. Electron. Agric.* **2021**, *185*, 106150. [CrossRef]
45. Mahbod, A.; Tschandl, P.; Langs, G.; Ecker, R.; Ellinger, I. The Effects of Skin Lesion Segmentation on the Performance of Dermatoscopic Image Classification. *Comput. Methods Programs Biomed.* **2020**, *197*, 105725. [CrossRef]
46. Sawyer Lee, R.; Dunnmon, J.A.; He, A.; Tang, S.; Ré, C.; Rubin, D.L. Comparison of Segmentation-Free and Segmentation-Dependent Computer-Aided Diagnosis of Breast Masses on a Public Mammography Dataset. *J. Biomed. Inform.* **2021**, *113*, 103656. [CrossRef] [PubMed]
47. Chu, C.; Hsu, A.-L.; Chou, K.-H.; Bandettini, P.; Lin, C. Does Feature Selection Improve Classification Accuracy? Impact of Sample Size and Feature Selection on Classification Using Anatomical Magnetic Resonance Images. *NeuroImage* **2012**, *60*, 59–70. [CrossRef] [PubMed]
48. Karabulut, E.M.; Özel, S.A.; İbrikçi, T. A Comparative Study on the Effect of Feature Selection on Classification Accuracy. *Procedia Technol.* **2012**, *1*, 323–327. [CrossRef]

agriculture

MDPI

Article

Automatic Classification of Bagworm, *Metisa plana* (Walker) Instar Stages Using a Transfer Learning-Based Framework

Siti Nurul Afiah Mohd Johari [1], Siti Khairunniza-Bejo [1,2,3,*], Abdul Rashid Mohamed Shariff [1,2,3], Nur Azuan Husin [1,2], Mohamed Mazmira Mohd Masri [4] and Noorhazwani Kamarudin [4]

[1] Department of Biological and Agricultural Engineering, Faculty of Engineering, Universiti Putra Malaysia (UPM), Serdang 43400, Selangor, Malaysia
[2] Smart Farming Technology Research Centre, Universiti Putra Malaysia (UPM), Serdang 43400, Selangor, Malaysia
[3] Institute of Plantation Studies, Universiti Putra Malaysia (UPM), Serdang 43400, Selangor, Malaysia
[4] Malaysian Palm Oil Board (MPOB), No. 6, Persiaran Institusi, Bandar Baru Bangi, Kajang 43000, Selangor, Malaysia
* Correspondence: skbejo@upm.edu.my; Tel.: +60-397694332

Abstract: Bagworms, particularly *Metisa plana* Walker (Lepidoptera: Psychidae), are one of the most destructive leaf-eating pests, especially in oil palm plantations, causing severe defoliation which reduces yield. Due to the delayed control of the bagworm population, it was discovered to be the most widespread oil palm pest in Peninsular Malaysia. Identification and classification of bagworm instar stages are critical for determining the current outbreak and taking appropriate control measures in the infested area. Therefore, this work proposes an automatic classification of bagworm larval instar stage starting from the second (S2) to the fifth (S5) instar stage using a transfer learning-based framework. Five different deep CNN architectures were used i.e., VGG16, ResNet50, ResNet152, DenseNet121 and DenseNet201 to categorize the larval instar stages. All the models were fine-tuned using two different optimizers, i.e., stochastic gradient descent (SGD) with momentum and adaptive moment estimation (Adam). Among the five models used, the DenseNet121 model, which used SGD with momentum (0.9) had the best classification accuracy of 96.18% with a testing time of 0.048 s per sample. Besides, all the instar stages from S2 to S5 can be identified with high value accuracy (94.52–97.57%), precision (89.71–95.87%), sensitivity (87.67–96.65%), specificity (96.51–98.61%) and the F1-score (88.89–96.18%). The presented transfer learning approach yields promising results, demonstrating its ability to classify bagworm instar stages.

Keywords: bagworm; hyperspectral image; deep learning; transfer learning; instar stage

Citation: Johari, S.N.A.M.; Khairunniza-Bejo, S.; Shariff, A.R.M.; Husin, N.A.; Masri, M.M.M.; Kamarudin, N. Automatic Classification of Bagworm, *Metisa plana* (Walker) Instar Stages Using a Transfer Learning-Based Framework. *Agriculture* **2023**, *13*, 442. https://doi.org/10.3390/agriculture13020442

Academic Editors: Gniewko Niedbała and Sebastian Kujawa

Received: 19 December 2022
Revised: 8 February 2023
Accepted: 11 February 2023
Published: 14 February 2023

1. Introduction

Oil palm (*Elaeis guineensis*) is a major agricultural sector in Malaysia, contributing significantly to the country's Gross Domestic Product (GDP). However, the emergence of various pests result in a loss of annual oil palm production [1]. Bagworms, *Metisa plana* Walker (Lepidoptera: Psychidae) are one of the most crucial leaf-eating pests especially in oil palm plantations, causing a yield reduction due to serious defoliation. Furthermore, *Metisa plana* was identified as the pest most widely affecting oil palm in Peninsular Malaysia [2]. Bagworm outbreaks are prevalent in oil palm plantations and have a significant negative economic impact on oil palm yield, causing 10% to 13% leaf defoliation and up to 40% crop losses [3]. As soon as they hatch, the bagworms begin feeding by scraping on the tops of the oil palm leaves. The surface that was scraped dries out and develops a hole. As a result, the lower and central crown sections have a distinct grey appearance due to badly damaged leaves [4]. Chung [5] claims that bagworm-infested palms experience increased foliage damage until all of the fronds are destroyed and blanked. *Metisa plana* has a total life cycle of 103.5 days from egg to adult, with seven larval instar stages. The Malaysian

Palm Oil Board (MPOB) notes that the most critical stage is the active feeding stage, which occurs from the first to the third instar stage and represents the early stage of the bagworm lifecycle. As a result, early detection of bagworm infestation is required to prevent this outbreak from worsening. Detection is critical because it can prevent the infestation from spreading over large areas.

Traditionally, a manual census of freshly damage symptoms was conducted at intervals of two weeks to count the number of larvae per frond and distinguish the instar stage of the bagworm. When there are ten larvae on each frond, early defoliation is detectable at 1% and is considered critical [6]. Identification of the instar stage is essential for effective decision making in pest control management. The physical characteristics and length of the larval case can enable identification of the larval instar stage [7]. The study found rounded leaf pieces attached loosely at the basal end of the case in the second instar stage, and the average length of the case was 4.6 mm. At the third instar stage, there were four to six rectangular leaf pieces attached at the proximal half of the case, which grew to 5.9 mm. The case surface had many loosely attached large round-to-rectangular leaf pieces at the fourth instar stage, and the average length of the case was 9.5 mm. In the fifth instar stage, most of the loose-leaf pieces were glued and formed a smooth surface case with an average length of 11.3 mm. During the manual census, the expert worker referred to these physical characteristics and length; thus, no destructive method was necessary because the stage could be solely recognized by surface morphology. However, manually identifying the instar stage for many larval samples is time-consuming and labor-intensive due to the minor size differences and similar color between the larval stages. As a result, identifying the bagworm instar stages in the infested oil palm would benefit from a quick and dependable remote sensing technique. Making management decisions based on the severity of the infestation is necessary to control insect pests. A worsening bagworm infestation outbreak may also result from inadequate control management knowledge. A constant threat may have terrible consequences because pest populations are growing quickly. Early insect pest detection also reduces production costs and the environmental impact of applying pesticides over a larger area by enabling inputs to be applied in the right quantity and locations [8].

Technology can assist farmers in efficiently detecting destructive insects or pests as well as preventing disease at an early stage [9]. Imaging and computer vision technology are widely used in a variety of fields and have numerous potential applications in contemporary agriculture. Several detection techniques using mechanization and image processing have begun to meet early pest infestation requirements. Kasinathan et al. [10] applied machine learning techniques to classify insect pests based on morphological features. Chiwamba and Nkunika [11] developed an automated system in identifying moths in the field using supervised machine learning. Tageldin et al. [12] implemented machine learning algorithms to predict leafworm infestation in the greenhouse. Machine learning models are typically designed to operate alone and must be redeveloped when attributes and data change. Instead of redeveloping the models, which generally involves a significant amount of effort, transfer learning attempts to regenerate the model and gained knowledge, as well as to significantly reduce model development time and enhance the model performance of the isolated learning model.

Fine-tuning is a transfer learning concept that requires some learning but has been shown to be much faster and more accurate than built models [13]. In fine-tuning, a deep convolutional neural network (CNN) is trained for similar task and the final layers of the model can be fine-tuned to adapt to the new dataset [14]. According to Kamilaris and Prenafeta-Boldú [15], deep learning models based on transfer learning CNN have been widely used in recent years as a powerful class of models for image classification in a variety of agriculture problems such as plant disease recognition [16–18], fruit classification [19–21], weed identification [22–24], and crop pest classification [25–27]. Some researchers use advanced pre-trained CNN models to classify crop pest images and achieve higher accuracy. Rahman et al. [28] used CNN architectures such as VGG16 and InceptionV3 to detect

and recognize rice pests and diseases, achieving 93.33% accuracy. AlexNet was used by Dawei et al. [29] to identify 10 types of pests with an accuracy of 93.84%. The same can be said for Liu et al. [30] who used AlexNet to classify 12 common paddy field pest species and got a mean Accuracy Precision (mAP) of 0.951.

Deep learning employs a highly layered network and a large amount of data, and it is prone to a serious problem known as overfitting [31]. Due to the presence of overfitting, the model performs flawlessly on the training set while fitting poorly on the testing set. To reduce the effect of overfitting, multiple solutions based on different strategies are proposed to inhibit the various triggers i.e., early stopping, data augmentation, and dropout. Tetila et al. [32] applied dropout with a rate of 0.5 and data augmentation to reduce the overfitting in classifying and counting soybean insect pests. Lim et al. [33] examined insect classification performance by applying data augmentation and early stopping to prevent overfitting. In addition, the selection of the optimizers also plays an important role in boosting the performance of the deep learning network model. Table 1 gives a summary of existing studies that applied deep learning approach especially in pest detection.

Ahmad et al. [34] identified both live and dead bagworms *Metisa plana* (first to third instar stage) using a motion tracking technique on oil palm fronds, with high accuracy of 87.5% and 78.8%, respectively. Since classifying bagworm instar stages is essential for early prevention, Mohd Johari et al. [35] used machine learning to identify bagworm instar stages based on spectral properties. It achieved a high level of accuracy (91–95%) and F1-score (0.81–0.91) in classifying bagworm instar stages from instar stage 2 to instar stage 5 using weighted KNN. However, this conventional machine learning approach required users to extract features from an image and then feed those features into the algorithm to perform classification. Therefore, this study aimed to automatically classify the bagworm instar stage, which was done using a transfer learning approach. The convolutional neural network (CNN) with deep architectures was used to execute automatic feature extraction and to learn complex high-level features. There were 5 different deep CNN architectures used i.e., VGG16, ResNet50, ResNet152, DenseNet121 and DenseNet201. These deep CNN models were evaluated by fine-tuning the models using different type of optimizers. This proposed work could be used to identify larval instar stages in oil palm plantations at an early stage to make early decisions regarding controlling bagworm infestations.

Table 1. Summary of existing pest detection models using deep learning approaches.

Model	Type of Pests and Disease	Crop Type	Learning Rate	Method of Reducing Overfitting	Optimizer	Accuracy (%)	References
Proposed CNN	10 beneficial and 10 harmful pests	Various crops	0.001	Image augmentation	Adam	90.00	[26]
Fine-tuned GoogleNet	10 species crop pests	Various crops	0.0001	Image augmentation	Adagrad	98.91	[25]
MatCovNet	Beetle and bugs	Paddy crop	0.01	N/A	SGD	83.08	[27]
Proposed CNN	Brown plant hopper	Paddy crop	0.0001	Dropout (0.3) Image augmentation	Adam	93.30	[28]
AlexNet	10 types of pests	Tea plant	N/A	N/A	N/A	98.92	[29]
AlexNet	12 species of pests	Paddy field	0.01	Dropout (0.7)	SGD (0.8)	95.10	[30]
DenseNet201	Soybean cyst nematode (SCN) eggs	Soybean	0.0001	Dropout (0.5) Image augmentation	SGD (0.9)	94.89	[36]
AlexNet	27 classes of insects	Various crops	0.01	Image augmentation Early stopping	SGD (0.9)	81.82	[33]

2. Materials and Methods

2.1. Overview

Figure 1 depicts the flowchart for this study. It began with data preparation which included the selection of band and cropping. The data was then augmented to increase the number of datasets and split into 70% training and 30% testing. All the datasets were fed into five pre-trained models, i.e., VGG16, ResNet50, ResNet152, DenseNet121 and

DenseNet201. All models were executed using SGD and Adam optimizer. The best model was chosen based on the highest accuracy and short execution time from both optimizers.

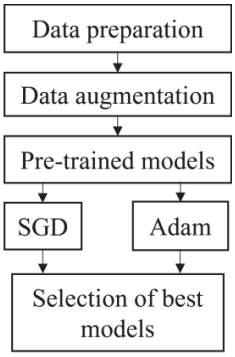


Figure 1. Flowchart of classification models.

2.2. Data Preparation

This study included four instar stages of larva ranging from the second to the fifth instar (i.e., S2, S3, S4 and S5). There was no first instar stage involved because they were too small and challenging to handle. All samples were collected at the Seberang Perak Plantation in Malaysia at the coordinates (4°8.8553′ N, 100°50.357′ E). The samples were brought into the laboratory for image acquisition. The sample was placed on a white background and captured using hyperspectral snapshot camera, FireflEYE S185 (Cubert GmbH, Ulm, Germany). There were 50 larvae per stage in the sample (n = 200). A total of 20 images were captured with 10 larvae in a single shot. The wavelengths of 506 nm and 538 nm were chosen because, according to a prior study by Mohd Johari et al. [35], they were the most important for classifying the four larval instar stages. The size of the captured image was 1000 pixels × 1000 pixels. Since ten larvae were captured in a single image, cropping was necessary to crop a single larva in a single image. All the images were cropped and resized into 224 height (h) × 224 width (w) using Microsoft Paint version 21H1 to get a single larva image as shown in Figure 2. The cropped images of all instar stages are illustrated in Figure 3.

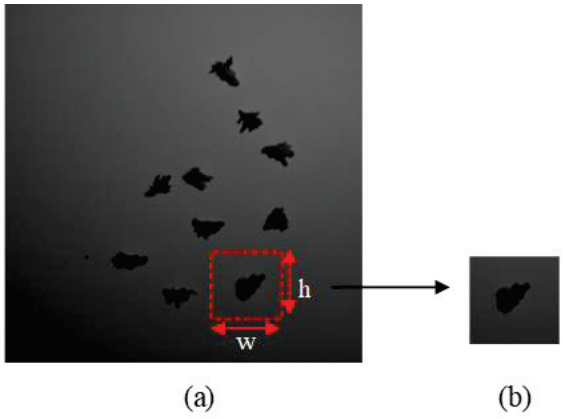


Figure 2. Cropping method, h and w represent height and width of the image, respectively (**a**) input image, (**b**) cropped image.

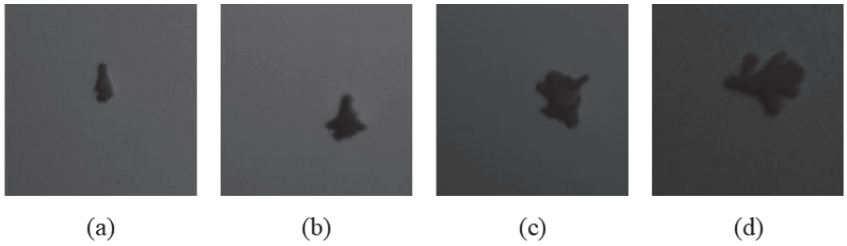


Figure 3. Images of cropped larvae, (**a**) S2, (**b**) S3, (**c**) S4, (**d**) S5.

2.3. Data Augmentation

Data augmentation aims to increase the number of samples in order to minimize the risk of overfitting and boost the accuracy of classification [37]. Therefore, image augmentation process was done, and the techniques were chosen carefully to preserve the actual size of the bagworm according to their instar stage (S2 to S5). The augmentation methods included intensity transformations (i.e., brightness and contrast) as well as random geometrical transformations (i.e., rotation, translation, horizontal flipping, and vertical flipping). The brightness and contrast interval were 0.5. The rotation was done using clockwise rotation along 0° (original images), 45°, 90°, 135°, 180°, 225°, 270°, and 315°. The translation ratio of the image was 0 width and 0.1 height. Meanwhile, the probability of the flipping was 0.5. All these techniques were successfully performed and produced a total of 9000 images.

Training and testing datasets were split from the total dataset, where 70% (6300 images) of the total images were used for training and another 30% (2700 images) were used for testing. During training, 5-fold cross validation was applied to assess the accuracy of the classifiers. Each dataset is divided into five equal folds at random. The classifier is trained on the four remaining folds after one-fold is removed for validation in each repetition. The accuracy of the classifier is then evaluated on the fold that was removed. This process is repeated until all folds have been tested. The major benefit of this approach is that all models are evaluated across all samples in the dataset, with no overlap between the training and testing datasets. The number of training and testing images were balanced for each instar stage as shown in Table 2 to avoid imbalanced data issues, where the accurate result frequently favors a majority class.

Table 2. Number of training and testing images for each instar stage.

Instar Stage	Training	Testing
S2	1575	675
S3	1575	675
S4	1575	675
S5	1575	675
Total	6300	2700

2.4. Transfer Learning CNN Models

The transfer learning approach was used to retrain the prominent technique in deep learning named convolution neural network (CNN). There are two main parts in CNN architecture, i.e., feature extraction and classification. The CNN is made up of three types of layers: convolutional layers, pooling layers, and fully connected (FC) layers, and formed when these layers are stacked together (Figure 4). The convolutional layer is the first layer used to extract various features from the input images. The output of the convolution layer is known as a feature map and usually followed by a pooling layer. The purpose of the pooling layer is to reduce the size of the convolved feature map by reducing the connections

between layers and operating independently on each feature map while maintaining its shape. After the last max-pooling layer, the first fully connected layer flattens all the feature maps, treating this one-dimensional vector (1-D) as a feature representation of the entire image. At this point, the classification operation is initiated.

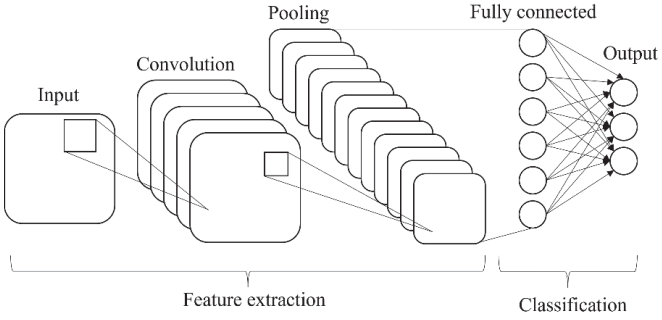


Figure 4. Basic architecture of CNN.

When all the features are connected to the FC layer, the training dataset is prone to overfitting. To address this issue, a dropout layer was used, in which a few neurons are removed randomly from the neural network during the training process, resulting in a smaller model. The most crucial parameter in the CNN model is the activation function. This is used to comprehend and approximate any type of continuous and complex relationship between network variables. In other words, it determines which model information should be executed forward and which should not at the end of the network. There are several types of activation functions such as the rectified linear (ReLU), softmax, tanH and sigmoid. Each of these functions has a specific application. For instances, the sigmoid and softmax functions are preferred for a binary classification CNN model, and to be specific, softmax is generally used for a multi-class classification.

In this study, there were 5 different CNN architectures used i.e., VGG16, ResNet50, ResNet152, DenseNet121 and DenseNet201. These networks performed at the pinnacle of their ability in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [38–40]. Moreover, numerous studies have used these well-known deep learning models for pest detection [41–43]. All pre-trained models were fine-tuned by replacing the last three layers with a fully connected layer, a softmax layer, and an output classification layer. The fully connected layer was set to four classes, which corresponds to four instar stages (S2, S3, S4, and S5). Finally, the new network structure was trained using images of larvae.

2.4.1. VGG16

VGG16 architecture was developed by Simonyan and Zisserman [38] and contains 16 convolution layers. The most distinctive feature of VGG16 is that instead of many hyperparameters, this model focused on having convolution layers of 3×3 filter with stride 1 and always used the same padding and max pooling layer of 2×2 filter with stride 2. This arrangement of convolution and max pooling layers is consistent throughout the architecture. At the end of the architecture, it was fine-tuned and replaced with two fully connected layers and a softmax activation function with four classes. Figure 5 depicts the structure of the modified VGG16 model using transfer learning.

2.4.2. Residual Network (ResNet)

Residual network, also known as ResNet, is one of the famous deep learning models introduced by He et al. [39]. The ResNet network employs a 34-layer plain network architecture based on VGG-19, to which the shortcut connection is added. This shortcut connection is known as the 'skip connection,' and it is at the heart of residual blocks. Because of this skip connection, the output of the layer is no longer the same. The shortcut

connection is used to connect the input x to the output after a few weight layers (Figure 6). The input 'x' is multiplied by the layer weights before a bias term is added as Equation (1).

$$h(x) = f(wx + b) \quad (1)$$

where f () is activation function, w is layer weight, and b is bias.

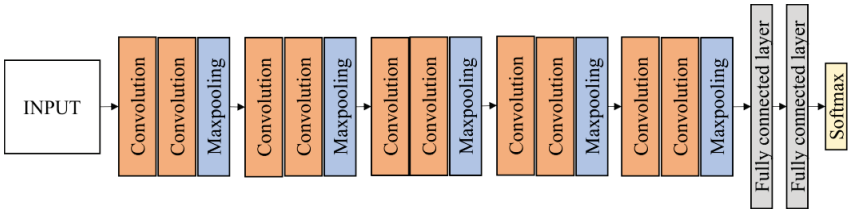


Figure 5. Fine-tuned VGG16 architecture.

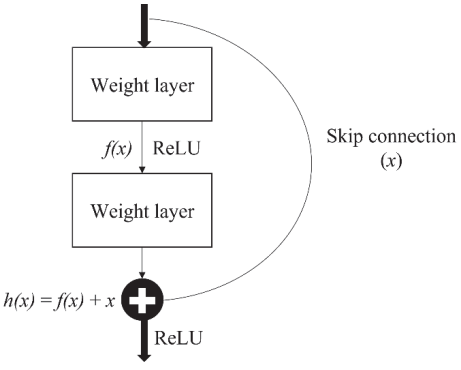


Figure 6. Single residual block.

With the introduction of a new skip connection technique, the output is now *h(x)* as in Equation (2).

$$h(x) = f(x) + x \quad (2)$$

This skip connections throughout ResNet solve the problem of vanishing gradient in deep neural networks by allowing the gradient to flow through an alternate shortcut path. It also helps by allowing the model to learn the identity functions, ensuring that the higher layer performs at least as well as the lower layer, if not better.

In this research, two residual networks of ResNet50 and ResNet152 were evaluated for the bagworm instar stage classification. ResNet50 is a 50-layer deep state of the art convolutional network and ResNet152 is a 152-layer network with recurrent connections using transfer learning. ResNet50 contains a 7×7 convolution layer with 64 kernels, a 3×3 max pooling layers with stride 2, 16 residual building blocks, 7×7 average pooling layers with stride 7 and two fully connected layers before the softmax output layer (Figure 7). The softmax output layer is set to 4 classes. ResNet152 has a layer structure similar to ResNet50, but it has 50 residual blocks instead of 16 residual blocks in ResNet50. The residual blocks reduce the output size while increasing the network depth.

2.4.3. DenseNet121 and DenseNet201

Another CNN architecture named DenseNet which was first explored by Huang et al. [40] was used in this research. It broadens the formulation of residual connection by adding new feature maps of all previous layers in a unit called Dense Block. Each layer in a DenseNet architecture is linked to every other layer, hence the name Densely Connected Convolutional Network. The feature maps from the previous layers are concatenated

(C) and used as inputs in each layer, rather than being summed. As a result, DenseNets require fewer parameters than an equivalent traditional CNN, allowing for feature reuse by discarding redundant feature maps. Essentially, each layer is linked to every other layer within a dense block, while the feature map size remains constant. Dense connectivity in DenseNet architecture can be represented as shown in Equation (3) and details of the dense block are illustrated in Figure 8.

$$x_l = H_l([x_0, x_1, \ldots., x_{l-1}]) \quad (3)$$

where $[x_0, x_1, \ldots, x_{l-1}]$ is the concatenation of the feature-maps, for instance, the output of all the layers preceding l $(0, \ldots, l-1)$. To facilitate implementation, the multiple inputs of Hl are concatenated into a single tensor.

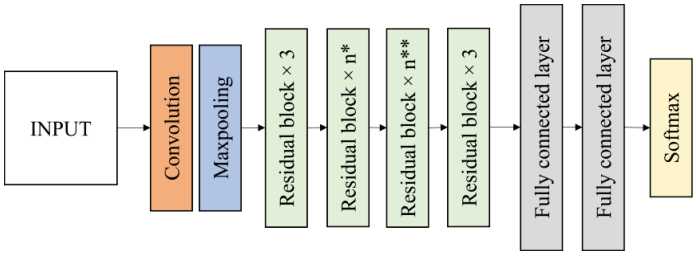


Figure 7. Fine-tuned ResNet50 and ResNet152 architecture (n* = 4 for ResNet50, 8 for ResNet152; n** = 6 for ResNet50, 36 for ResNet152).

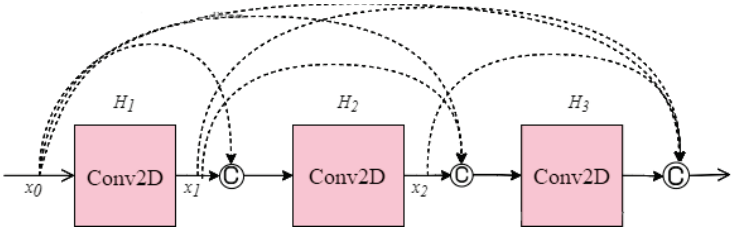


Figure 8. Single dense block.

In this work, DenseNet121 and DenseNet201 were evaluated. Each architecture consists of four dense blocks. The first part of the DenseNet consists of 7 × 7 convolutional layers with stride 2 followed by 3 × 3 max pooling layers with stride 2. The layers between dense blocks are known as transition layers and perform the convolution and pooling. Following the fourth dense block is a classification layer, which accepts feature maps from all layers of the network to perform classification. It consists of two fully connected layers before the softmax output layer which is set to 4 classes. The DenseNet121 and DenseNet201 architecture are illustrated in Figure 9.

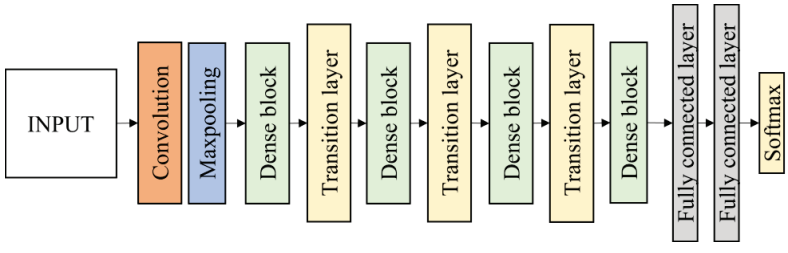


Figure 9. Fine-tuned DenseNet121 and DenseNet201 architecture.

2.5. Optimizer

The improved performance of a deep learning model is significantly influenced by an optimizer. Optimizers can be thought as mathematical functions to adjust the network weights given the gradients and additional information, depending on the conceptualization of the optimizer. Optimizers are built upon the idea of gradient descent, the greedy approach of iteratively minimizing the loss function by following the gradient. For instance, the deep CNN model is trained iteratively by updating the parameters of each layer of the network, and an optimizer is crucial in this process. Categorical cross-entropy is one of the most widely used loss functions for evaluating performance across multiple classes. When the desired output and the predicted output are the same, the cross-entropy value is close to zero, and this is what any optimization technique seeks to achieve.

In this study, all the pretrained models were trained using stochastic gradient descent (SGD) with momentum and adaptive moment estimation (Adam). These optimization methods are described as below.

2.5.1. Stochastic Gradient Descent (SGD)

One of the most widely used and well-liked algorithms for optimizing neural networks at a lower cost is stochastic gradient descent optimization [44]. To reduce error, SGD will update a variable once every epoch. The SGD equation is illustrated as Equation (4). Prior time step variable minus the outcome of learning rate multiple with a gradient vector will be used to update the variable. One of the most widely used methods for accelerating the Gradient Descent algorithm's convergence is momentum.

$$w_{(n+1)} = w_n + \eta \hat{g} \tag{4}$$

where w_n is weight at time n, η is learning rate, and $\hat{g}$ is gradient vector.

2.5.2. Adaptive Moment Estimation (Adam)

Adam optimizer was introduced by Kingma and Ba [45], a first-order gradient with a small memory specification required for efficient stochastic enhancement. Based on an analysis of the first and second moments of the gradients, the optimizer determines discrete versatile learning rates for various parameters. The first moment (mean) and the second moment (variance) are computed as Equations (5) and (6), respectively.

$$m_n = \beta_1 m_{(n-1)} + (1 - \beta_1)\hat{g} \tag{5}$$

$$v_n = \beta_1 v_{(n-1)} + (1 - \beta_2)\hat{g}^2 \tag{6}$$

When decay rates are very low (i.e., β_1 and β_2 are close to zero), the m_t and v_t are biased towards zero. To remedy this predicament, the first and second moments with bias-corrected terms are computed in Equations (7) and (8) as follows:

$$\hat{m}_n = \frac{m_n}{1 - \beta_1^n} \tag{7}$$

$$\hat{v}_n = \frac{v_n}{1 - \beta_2^n} \tag{8}$$

Then, the Adam weight update rule is given by Equation (9):

$$w_{(n+1)} = w_n - \frac{\eta}{\sqrt{\hat{v}_n + \in}} \hat{m}_n \tag{9}$$

2.6. Hyperparameters Configuration

Hyperparameters are variables that determine the network structure, and the process by which the network is trained, i.e., learning rate, mini batch size and number of epochs. Prior to training, hyperparameters are set before optimizing the weights and bias. Models

can have up to ten hyperparameters, and determining the best combination is referred to as a search problem. As a result, choosing the right hyperparameter values can have an impact on the performance of the model [46].

The learning rate describes the learning progress of the proposed model and updates the weight parameters to reduce the loss function of the network. Larger learning rates mean that the weights are changed more every iteration, so that they may reach their optimal value faster, but may also miss the exact optimum. Smaller learning rates mean that the weights are changed less every iteration, so it may take longer time to reach their optimal value, but they are less likely to miss the optima of the loss function. In this study, a well-known good default value for learning rate of 0.001 and categorical cross entropy as loss function were used.

An epoch is the complete training cycle over the entire training dataset, and a subset of the training dataset is referred to as a mini batch for evaluating the gradient descent loss function and updating the weights. In this study, the number of epochs and mini batch size were fixed at 100 and 50, respectively. Furthermore, early stopping was added by monitoring the testing accuracy. The training iteration will be terminated if the testing accuracy reaches its maximum and no improvement is observed after five continuous iterations. Table 3 summarizes the hyperparameters used in this study.

Table 3. Hyperparameters set up.

Hyperparameter	Value
Learning rate	0.001
Epoch	100
Mini-batch size	50

This study used a convolutional neural network of TensorFlow Deep Learning Framework to develop the bagworm classification model. The TensorFlow environment was set up using the Anaconda Individual Edition software with python 3.6.13. The process of training and testing the model was performed by MSI Workstation WE73 8SK with a central processing unit (CPU) of Intel Core i7 powered by Nvidia Quadro P3200 GPU with 6GB GDDR5.

2.7. Model Performance Evaluation

A confusion matrix was used to assess the ability of the classifier to classify the larval instar stage by calculating several performance metrics. The matrix gives rise to four indices: true positive (TP), false positive (FP), false negative (FN), and true negative (TN). The TP and TN correspond to the number of correctly predictions, respectively, while the FP and FN correspond to the number of incorrectly predictions. Additionally, testing time which is defined as the time taken to complete the classification, was also used to assess the model performance. Table 4 lists short descriptions and mathematical formulae of the performance metrics used in this study.

Table 4. The performance metrics with brief descriptions.

Metrics	Equations	Description
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN} \times 100\%$	Percentage of all correctly classified based on the larval instar stages.
Precision	$\frac{TP}{TP+FP}$	Quantifies the number of corrected predicted instar stage.
Sensitivity	$\frac{TP}{TP+FN}$	True positive rate which identifies the proportion of corrected classification.
Specificity	$\frac{TN}{TN+FP}$	True negative rate which identifies the proportion of true negative in the classification.
F-score	$\frac{2\times P\times R}{P+R}$	Weighted average of the true positive rate (sensitivity) and precision.
Testing time	-	Time taken to complete the classification

3. Results

Tables 5 and 6 summarize the classification performance, i.e., accuracy, precision, sensitivity, specificity, F1-score, and the testing time of all the pre-trained models using SGD and Adam optimizers, respectively. Testing time was calculated as the amount of time needed to test the model using 2300 images.

Table 5. Performance of each pre-trained model using SGD optimizer.

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	Testing Time
DenseNet201	**96.97 ± 0.45** [a]	**93.96 ± 0.90**	**97.98 ± 0.89**	**97.98 ± 0.30**	**93.91 ± 0.91** [a]	241.40 ± 48.00 [a]
DenseNet121	96.18 ± 0.33 [a]	92.49 ± 0.61	92.35 ± 0.66	97.45 ± 0.22	92.30 ± 0.67 [a]	130.20 ± 16.30 [b]
ResNet152	95.49 ± 0.51 [a]	91.55 ± 0.87	90.99 ± 1.03	97.00 ± 0.34	90.97 ± 1.04 [a]	306.60 ± 33.82 [a]
ResNet50	93.81 ± 0.67 [b]	88.30 ± 1.09	87.61 ± 1.33	95.87 ± 0.44	87.43 ± 1.50 [b]	**104.00 ± 2.66** [b]
VGG16	94.82 ± 0.24 [b]	89.87 ± 0.55	89.64 ± 0.49	96.55 ± 0.16	89.58 ± 0.52 [b]	235.80 ± 3.20 [a]

Note: Data represents mean (±standard error). The bold font shows the selected best value among other values in the same column. Different letters within the same column indicate statistically difference by the Tukey's HSD test at $p < 0.05$.

Table 6. Performance of each pre-trained models using Adam optimizer.

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	Testing Time (s)
DenseNet201	**96.01 ± 0.28**[a]	92.32 ± 0.50	92.01 ± 0.56	97.34 ± 0.18	**92.02 ± 0.57** [a]	258.40 ± 142.46 [a]
DenseNet121	94.56 ± 0.31 [a]	89.61 ± 0.44	89.12 ± 0.62	96.37 ± 0.21	89.14 ± 0.59 [a]	135.60 ± 46.21 [b]
ResNet152	94.97 ± 0.37 [a]	90.10 ± 0.79	89.93 ± 0.74	96.64 ± 0.25	89.87 ± 0.79 [a]	384.80 ± 185.56 [c]
ResNet50	94.56 ± 0.52 [a]	89.25 ± 1.04	89.10 ± 1.05	96.37 ± 0.34	89.02 ± 1.08 [a]	**133.40 ± 59.49** [b]
VGG16	95.40 ± 0.30 [a]	90.90 ± 0.60	90.81 ± 0.60	96.94 ± 0.20	90.81 ± 0.61 [a]	238.80 ± 11.03 [a]

Note: Data represents mean (± standard error). The bold font shows the selected best value among other values in the same column. Different letters within the same column indicate statistical difference by the Tukey's HSD test at $p < 0.05$.

According to Table 5, the accuracy achieved for all models ranged from 93.81% to 96.97%, while the F1-score ranged from 87.43% to 93.91%. Among the models, DenseNet201 achieved the highest accuracy and F1-score of 96.97% and 92.30%, followed by DenseNet121 which achieved 96.18% accuracy and 92.30% F1-score. ResNet152 was ranked third due to its accuracy of 95.49% and F1-score of more than 90%. It was followed by VGG16 with accuracy (94.82%) and F1-score (89.58%). ResNet50 gained the lowest accuracy and F1-score, 93.81% and 87.43%, respectively. When considering the testing time, DenseNet201 had the longer testing time of 241.40 s, while ResNet50 and DenseNet121 performed the fastest with an average of 104 s and 130.20 s, respectively. Furthermore, there was found to be no significant difference in accuracy and F1-score between DenseNet201 and DenseNet121, which indicate both models can be implemented to achieve high accuracy.

In addition, DenseNet201 achieved the highest performance accuracy and F1-score using the Adam optimizer, with 96.01% and 92.02%, respectively (Table 6). However, it also took a longer testing time (258.40 s), placing it fourth place overall. ResNet50 and DenseNet121 achieved the same accuracy (94.56%) and shortest testing time, i.e., 133.40 s and 135.60 s, respectively. Nevertheless, the F1-score achieved by ResNet50 was the lowest with 89.02%. Following DenseNet201, VGG16 yielded an accuracy of 95.44%, and 90.81% for F1-score. In comparison to SGD, it can be said that the performance of VGG16 and ResNet50 in Adam improved by 0.61% and 0.79%, respectively. Even so, the performance of other models decreased, particularly DenseNet121, which experienced a drop of more than 3% in F1-score. Additionally, there was no significant difference in accuracy and F1-score among them, indicating difficulty in selecting the best model.

Based on the testing time, it was shown that ResNet50 and DenseNet121 had the shortest time in both optimizers. Significant differences in testing time $p < 0.05$ between all the models were observed and no significant difference was identified between ResNet50 and DenseNet121, indicating both models can be initiated in the shortest time. ResNet152 had the longest testing time (>300 s) and was significantly different from ResNet50. This was probably due to the complexity of the architecture which has 152 deep layer state-of-art with residual block. Nonetheless, it still achieved slightly higher accuracy than ResNet50. Similar issues apply to DenseNet201 and DenseNet121, which had a significant difference in testing time but both models achieved high accuracy and an F1-score in comparison to other models. When comparing the two performances in SGD and Adam, it seems that SGD models outperformed Adam especially in testing time, which was quicker and rejected the expectations that it might be slower than Adam.

ResNet50 had the fastest testing time, but the ideal model should take accuracy into consideration since ResNet50 was the least well performing. Therefore, in this study, DenseNet121 was chosen to be the optimal model with shorter testing time and high accuracy. SGD was regarded as the best optimizer since most of the models consistently showed high accuracy and F1-score with less testing time. Thus, DenseNet121 with SGD optimizer was determined to be the best model out of all the models due to its lesser testing time of 130.20 s and obtaining the high accuracy rate of 96.18%.

A confusion matrix was created for DenseNet121 using SGD optimizer, as shown in Table 7. The confusion matrix included performance metrics such as accuracy, precision, sensitivity, specificity, and F1-score for each instar stage.

Table 7. Performance metric of DenseNet121 using SGD optimizer at each instar stage.

Metrics (%)	S2	S3	S4	S5	Average
Accuracy	94.52	94.53	98.08	97.57	96.18
Precision	90.56	89.71	95.87	93.83	92.49
Sensitivity	87.67	88.56	96.50	96.65	92.35
Specificity	96.80	96.51	98.61	97.88	97.45
F1-score	88.89	88.92	96.18	95.21	92.30

Based on Table 7, DenseNet121 performed better in classifying S4 and S5 with accuracy and F1-score more than 95%, compared to S2 and S3. S4 outperformed all instar stages with the highest accuracy and F1-score of 98.08% and 96.18%, respectively. Meanwhile, S5 achieved 97.57% accuracy and 95.21% F1-score. S4 and S5 have distinct physical characteristics that enable the DenseNet121 to learn and classify the stage easily. When compared to S2 and S3, both of which had a smaller size, their classification performance was a little bit lower, since the F1-score was less than 90%. Low sensitivity (<90%) in S2 and S3 indicate that there were more cases of misclassification of correct classes. Nonetheless, S2 and S3 still achieved high accuracy, 94.52% and 94.53%, respectively. Overall, DenseNet121 produced encouraging results and correctly classified every instar stage with a high level of accuracy (>90%).

4. Discussion

In this research, transfer learning was utilized to automatically classify four larval instar stages. Five types of deep CNN models were used, namely, VGG16, ResNet50, ResNet152, DenseNet121, and DenseNet201. Among the models, DenseNet121 using SGD with momentum was selected to be the best model as it demonstrated the best model performance and obtained a quick testing time for the instar stage classification. It achieved 96.18% accuracy, 92.30% F1-score and took 130.20 s to complete the classification process, i.e., 0.048 s per sample. This study shows that all neurons were considered important to identify more detailed properties of larvae as they can achieve high accuracy by not

considering any dropout especially by classifying lower instar levels (S2 and S3) that are very small in size.

Generally, both DenseNet121 and DenseNet201 performed well with good accuracy compared to VGG16, ResNet50 and ResNet152. The main success of DenseNet is due to its dense connection, which maintains the magnitude of gradients during backpropagation. This solves the vanishing gradient problem, which is known to degrade the performance of deep learning algorithms. DenseNet is known to be a modified version of ResNet; ResNet appears to use only one previous feature-map, whereas DenseNet appears to use features from all previous convolutional blocks. ResNet currently uses summation to connect all previous feature maps, whereas DenseNet concatenates them all. DenseNet achieves good results in image recognition and classification due largely to its dense layer connection mode.

Better optimizers are mainly focused on being faster and efficient but are also often known to generalize well (less overfitting) compared to others. Identification of suitable optimizers that improve high accuracy results is very challenging. In this study, SGD outperformed Adam including in execution time, even though Adam converges more quickly. This finding is similar to the work done by Poojary and Pai [47] who compared the performance of two improved CNN models using the SGD, Adam, and RMSProp optimizers and found that SGD performed better. Furthermore, according to Hardt et al. [48], SGD is uniformly stable for strongly convex loss function, and thus might have optimal generalization error. In addition, according to Wilson et al. [49], non-adaptive methods (i.e., SGD) will converge towards a minimum norm solution in a binary least-square classification loss task while adaptive methods (i.e., Adam) can diverge. It often obtained faster initial progress on the training set, but performance fails to generalize on the validation data.

Compared to the previous study by Mohd Johari et al. [35], based on the performance metrics of the model, the proposed method indicated some improvement. In the previous study, weighted KNN was selected as the best model to classify the bagworm instar stage with accuracy and F1-score achieved 95% and 91% (S2), 93% and 85% (S3), 91% and 81% (S4) and 91% and 81% (S5). Meanwhile, in this study, DenseNet121 performed the best to classify the bagworm instar stage with accuracy and high F1-score, achieving 94.52% and 88.89% (S2), 94.53% and 88.92% (S3), 98.08% and 96.18% (S4) and 97.57% and 95.21% (S5), respectively. Previous studies that used spectral properties to classify instar stage found that young instar stages (S2 and S3) could be detected more accurately than adult instar stages (S4 and S5). In the meantime, this study which employed transfer learning with the chosen spectral image, revealed that the model performance improved as the larval size increased, with the adult instar stages (S4 and S5) being categorized better than the young instar stages (S2 and S3). Therefore, it was demonstrated that complex multilayered neural networks provided by deep learning able to self-learn to identify the adult instar stage based on its more distinct physical appearance, not only because of its size, but also because of the changes in its morphological architecture as its skirt develops. However, compared to the earlier study, the S2 result appears to have decreased slightly. According to Wang et al. [50], machine learning has a greater solution effect on small sample datasets, however the deep learning framework has superior accuracy on large sample datasets. This study employed 1,775 datasets for training and 350 datasets for testing, which is five times and four-and-a-half times greater than previous work for training and testing, respectively. It appears that the five times larger datasets utilized in training a deep learning model is insufficient to generalize the model's ability to create high-quality interpretations of tiny size instars (S2) whose morphological architecture is hardly visible. Nevertheless, in general, the transfer learning method clearly outperforms the previous study in classifying all instar stages with an average accuracy achieved more than 94%. It has been clearly demonstrated that transfer learning improves classification. It differs from traditional machine learning in that it involves the use of pre-trained models that were previously used for another task to jumpstart the development process on a new task or problem. Models can gain generalized feature "knowledge" from other datasets with the aid of

transfer learning. This "knowledge" can be used to learn the target dataset, which can significantly boost model performance. Meanwhile, machine learning necessitates starting from scratch by manually extracting features and feeding them into the classification.

The result obtained from this study is considered acceptable and comparable with other similar studies [42,51,52]. Qi et al. [51] used five convolutional neural network structures, i.e., AlexNet, VGG16, ResNet50, DenseNet121 and InceptionV3 to identify peanut-leaf diseases. The optimizer used was SGD with momentum 0.9. It showed that DenseNet121 achieved the best performance with high F1-score slightly lower compared to this study, i.e., 90.50%. Mohsin et al. [42] employed five different deep neural network DNN models, i.e., VGG19, ResNet50, EfficientNetB5, DenseNet121, and InceptionV3 to classify crop-based insect species from a large volume of dataset. DenseNet121 performed the best across all classes with accuracy ranged from 46.31% to 95.36%. Salassa et al. [52] applied DenseNet121 to detect disease in plants by using leaf plant imagery from PlantVillage dataset based on their respective classes and achieved almost similar performance with 96.41% accuracy, which indicates success in detecting plant disease.

5. Conclusions

In this study, a transfer learning approach of deep CNN models was used to classify the four *Metisa plana* larval instar stages using images selected from wavelength 506 nm and 538 nm captured using a hyperspectral snapshot camera in a controlled environment. Five pre-trained models were implemented to categorize the larval instar stages, i.e., VGG16, ResNet5, ResNet152, DenseNet121 and DenseNet201. The findings revealed that among the five pre-trained models, the DenseNet121 (SGD) with 0.9 momentum was identified as the most suitable model due to its minimum processing time (0.048 s per sample) and great accuracy and F1-score of 96.18% and 92.30%, respectively. Monitoring bagworms is essential for early bagworm management and control. Existing research on an automatic technique [34] is limited to the detection of living and dead bagworm larvae but does not distinguish between instar stages. Therefore, this proposed method is crucial for the next phase of developing an autonomous pest detector in which the instar stages can be automatically classified without the need for human intervention because no feature extraction is involved. In future work, a new deep CNN model for the classification of larval instar stages can be constructed completely from scratch and compared to the model that has already been pre-trained. In addition, the accuracy of the results could be improved by using a more extensive dataset, specifically S2.

Author Contributions: Conceptualization, S.N.A.M.J. and S.K.-B.; methodology, S.N.A.M.J.; software, S.K.-B.; formal analysis, S.N.A.M.J. and S.K.-B.; investigation, S.N.A.M.J.; resources, M.M.M.M. and N.K.; data curation, S.N.A.M.J.; writing—original draft preparation, S.N.A.M.J.; writing—review and editing, S.K.-B.; supervision, A.R.M.S., N.A.H. and M.M.M.M.; project administration, S.K.-B.; funding acquisition, S.K.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Higher Education Malaysia (MOHE) under Fundamental Research Grants Scheme (FRGS) (Project number: FRGS/1/2018/TK04/UPM/02/4) and the Graduate Study and Research in Agriculture (SEARCA) under Professorial Chair Award 2022-2023 in the field of Imaging Technology and Remote Sensing.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to restrictions.

Acknowledgments: The authors would like to thank the MPOB for providing sample data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yap, T.H. A review on the management of lepidoptera leaf-eaters in oil palm: Practical implementation of integrated pest management strategies. *Planter* **2005**, *81*, 569–586.
2. Norman, K.; Basri, M.W. Status of common oil palm insect pests in relation to technology adoption. *Planter* **2007**, *83*, 371–385.
3. Benjamin, N. Bagworm Infestation in District Causing Palm Oil Production to Drop. Available online: https://www.thestar.com.my/news/community/2012/11/21/bagworm-infestation-in-district-causing-palm-oil-production-to-drop/ (accessed on 12 June 2020).
4. Corley, R.H.V.; Tinker, P.B. Pests of the Oil Palm. In *The Oil Palm*; Wiley: Hoboken, NJ, USA, 2015; pp. 437–459. [CrossRef]
5. Chung, G.F. *Effect of Pests and Diseases on Oil Palm Yield*; AOCS Press: Urbana, IL, USA, 2012.
6. Kamarudin, N.; Ahmad Ali, S.R.; Mohd Masri, M.M.; Ahmad, M.N.; Che Manan, C.A.H. Controlling *Metisa plana* Walker (Lepidoptera: Psychidae) outbreak using *Bacillus thuringiensis* at an oil palm plantation. *J. Oil Palm Res.* **2017**, *29*, 47–54. [CrossRef]
7. Kok, C.C.; Eng, O.K.; Razak, A.R.; Arshad, A.M. Microstructure and life cycle of *Metisa plana* walker (Lepidoptera: Psychidae). *J. Sustain. Sci. Manag.* **2011**, *6*, 51–59.
8. Tetila, E.C.; Machado, B.B.; Belete, N.A.D.S.; Guimaraes, D.A.; Pistori, H. Identification of Soybean Foliar Diseases Using Unmanned Aerial Vehicle Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2190–2194. [CrossRef]
9. Azfar, S.; Nadeem, A.; Basit, A. Pest detection and control techniques using wireless sensor network: A review. *J. Entomol. Zool. Stud. JEZS* **2015**, *3*, 92–99.
10. Kasinathan, T.; Singaraju, D.; Uyyala, S.R. Insect classification and detection in field crops using modern machine learning techniques. *Inf. Process. Agric.* **2021**, *8*, 446–457. [CrossRef]
11. Chiwamba, S.H.; Phiri, J.; Nkunika, P.O.Y.; Nyirenda, M.; Kabemba, M.M. An application of machine learning algorithms in automated identification and capturing of fall armyworm (FAW) moths in the field. In Proceedings of the ICICT2018, Lusaka, Zambai, 27–30 November 2018; Volume 3, pp. 1–4.
12. Tageldin, A.; Adly, D.; Mostafa, H.; Mohammed, H.S. Applying machine learning technology in the prediction of crop infestation with cotton leafworm in greenhouse. *bioRxiv* **2020**, bioRxiv: 2020.09.17.301168.
13. Mohanty, S.P.; Hughes, D.P.; Salathé, M. Using deep learning for image-based plant disease detection. *Front. Plant Sci.* **2016**, *7*, 1419. [CrossRef]
14. Kaya, A.; Keceli, A.S.; Catal, C.; Yalic, H.Y.; Temucin, H.; Tekinerdogan, B. Analysis of transfer learning for deep neural network based plant classification models. *Comput. Electron. Agric.* **2019**, *158*, 20–29. [CrossRef]
15. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. [CrossRef]
16. Boulent, J.; Foucher, S.; Théau, J. Convolutional Neural Networks for the Automatic Identification of Plant Diseases. *Plant Sci.* **2019**, *10*, 941. [CrossRef] [PubMed]
17. Dinata, M.I.; Mardi Susiki Nugroho, S.; Rachmadi, R.F. Classification of Strawberry Plant Diseases with Leaf Image Using CNN. In Proceedings of the ICAICST 2021—2021 International Conference on Artificial Intelligence and Computer Science Technology, Yogyakarta, Indonesia, 29–30 June 2021; pp. 68–72. [CrossRef]
18. Mique, E.L.; Palaoag, T.D. Rice pest and disease detection using convolutional neural network. In Proceedings of the ACM International Conference Proceeding Series, Jeju, Korea, 27–29 April 2018. [CrossRef]
19. Al-Shawwa, M.O.; Abu-Naser, S.S. Classification of Apple Fruits by Deep Learning. *Int. J. Acad. Eng. Res.* **2019**, *3*, 1–7.
20. Gayathri, S.; Ujwala, T.U.; Vinusha, C.V.; Pauline, N.R.; Tharunika, D.B. Detection of Papaya Ripeness Using Deep Learning Approach. In Proceedings of the Third International Conference on Inventive Research in Computing Applications (ICIRCA-2021), Tamilnadu, 2–4 September 2021; pp. 1755–1758. [CrossRef]
21. Shamim Hossain, M.; Al-Hammadi, M.; Muhammad, G. Automatic Fruit Classification Using Deep Learning for Industrial Applications. *IEEE Trans. Ind. Informatics* **2015**, *15*, 1027–1034. [CrossRef]
22. Huang, S.-C.; Le, T.-H. Convolutional neural network architectures. *Princ. Labs Deep Learn.* **2021**, 201–217. [CrossRef]
23. Jin, X.; Che, J.; Chen, Y. Weed identification using deep learning and image processing in vegetable plantation. *IEEE Access* **2021**, *9*, 10940–10950. [CrossRef]
24. Selvi, C.T.; Sankara Subramanian, R.S.; Ramachandran, R. Weed Detection in Agricultural fields using Deep Learning Process. In Proceedings of the 2021 7th International Conference on Advanced Computing & Communication Systems (ICACCS), India, 19–20 March 2021; pp. 1470–1473. [CrossRef]
25. Li, Y.; Wang, H.; Dang, L.M.; Sadeghi-niaraki, A.; Moon, H. Crop pest recognition in natural scenes using convolutional neural networks. *Comput. Electron. Agric.* **2020**, *169*, 105174. [CrossRef]
26. Malek, M.A.; Reya, S.S.; Hasan, M.Z.; Hossain, S. A Crop Pest Classification Model Using Deep Learning Techniques. In Proceedings of the 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), Dhaka, Bangladesh, 5–7 January 2021; pp. 367–371. [CrossRef]
27. Suthakaran, A.; Premaratne, S. Detection of the affected area and classification of pests using convolutional neural networks from the leaf images. *Int. J. Comput. Sci. Eng.* **2020**, *9*, 1–10.
28. Rahman, C.R.; Arko, P.S.; Ali, M.E.; Iqbal, M.A.; Apon, S.H.; Nowrin, F. Identification and recognition of rice diseases and pests using convolutional neural networks. *Biosyst. Eng.* **2020**, *194*, 112–120. [CrossRef]
29. Dawei, W.; Limiao, D.; Jiangong, N.; Jiyue, G.; Hongfei, Z.; Zhongzhi, H. Recognition pest by image-based transfer learning. *J. Sciene Food Agric.* **2019**, *99*, 4524–4531. [CrossRef]

30. Liu, Z.; Gao, J.; Yang, G.; Zhang, H.; He, Y. Localization and Classification of Paddy Field Pests using a Saliency Map and Deep Convolutional Neural Network. *Sci. Rep.* **2016**, *6*, 20410. [CrossRef] [PubMed]
31. Hara, K.; Saitoh, D.; Shouno, H. Analysis of dropout learning regarded as ensemble learning. In Preceedings of the Artificial Neural Networks Machine Learning (ICANN 2016), Barcelona, Spain, 6–9 September 2016; pp. 72–79. [CrossRef]
32. Tetila, E.C.; MacHado, B.B.; Astolfi, G.; De Souza Belete, N.A.; Amorim, W.P.; Roel, A.R.; Pistori, H. Detection and classification of soybean pests using deep learning with UAV images. *Comput. Electron. Agric.* **2020**, *79*, 105836. [CrossRef]
33. Lim, S.; Kim, S.; Kim, D. Performance Effect Analysis for Insect Classification using Convolutional Neural Network. In Proceedings of the 2017 7th IEEE International Conference on Control System, Computing and Engineering (ICCSCE 2017), Penang, Malaysia, 24–26 November 2017; pp. 210–215.
34. Ahmad, M.N.; Mohamed Shariff, A.R.; Aris, I.; Abdul Halin, I.; Moslim, R. Identification and determination of the spectral reflectance properties of live and dead bagworms, *Metisa plana* Walker (Lepidoptera: Psychidae) using Vis/NIR spectroscopy. *J. Oil Palm Res.* **2020**. [CrossRef]
35. Mohd Johari, S.N.A.; Khairunniza-bejo, S.; Mohamed Shariff, A.R.; Husin, N.A.; Mohd Masri, M.M.; Kamarudin, N. Identification of bagworm (*Metisa plana*) instar stages using hyperspectral imaging and machine learning techniques. *Comput. Electron. Agric.* **2021**, *194*, 106739. [CrossRef]
36. Tetila, E.C.; MacHado, B.B.; Menezes, G.V.; De Souza Belete, N.A.; Astolfi, G.; Pistori, H. A Deep-Learning Approach for Automatic Counting of Soybean Insect Pests. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1837–1841. [CrossRef]
37. Krizhevsky, B.A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2012**, *60*, 84–90. [CrossRef]
38. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
40. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. *arXiv* **2018**, arXiv:1608.06993v5.
41. Swasono, D.I.; Tjandrasa, H.; Fathicah, C. Classification of tobacco leaf pests using VGG16 transfer learning. In Proceedings of the l2th International Conference on Information & Communication Technology and System (lCTS) 2019, Surabaya, Indonesia, 18 July 2019; pp. 176–181. [CrossRef]
42. Mohsin, M.R.; Ramisa, S.A.; Saad, M.; Rabbani, S.H.; Tamkin, S. Classifying Insect Pests from Image Data using Deep Learning. Bachelor Thesis, Brac University, Dhaka, Bangladesh, 2022.
43. Khanramaki, M.; Askari Asli-Ardeh, E.; Kozegar, E. Citrus pests classification using an ensemble of deep learning models. *Comput. Electron. Agric.* **2021**, *186*, 106192. [CrossRef]
44. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747v2.
45. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. *arXiv* **2017**, arXiv:1412.6980v9.
46. Aszemi, N.M.; Dominic, P.D.D. Hyperparameter optimization in convolutional neural network using genetic algorithms. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 269–278. [CrossRef]
47. Poojary, R.; Pai, A. Comparative Study of Model Optimization Techniques in Fine-Tuned CNN Models. In Proceedings of the 2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA), Ras Al Khaimah, United Arab Emirates, 19–21 November 2019; pp. 22–25. [CrossRef]
48. Hardt, M.; Recht, B.; Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 57–78. [CrossRef]
49. Wilson, A.C.; Roelofs, R.; Stern, M.; Srebro, N.; Recht, B. The marginal value of adaptive gradient methods in machine learning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4149–4159.
50. Wang, P.; Fan, E.; Wang, P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognit. Lett.* **2021**, *141*, 61–67. [CrossRef]
51. Qi, H.; Liang, Y.; Ding, Q.; Zou, J. Automatic identification of peanut-leaf diseases based on stack ensemble. *Appl. Sci.* **2021**, *11*, 1950. [CrossRef]
52. Salassa, X.; Al Qarni, W.; Novanza, T.; Diasa, F.G.; Inda, S. Design Plant Disease Detection System Using Deep Learning Convolutional Neural Network. *Khazanah J. Mhs.* **2020**, *12*, 95–96. [CrossRef]

agriculture

MDPI

Article

Real-Time Plant Health Detection Using Deep Convolutional Neural Networks

Mahnoor Khalid [1], Muhammad Shahzad Sarfraz [1], Uzair Iqbal [2], Muhammad Umar Aftab [1], Gniewko Niedbała [3,*] and Hafiz Tayyab Rauf [4,*]

[1] Department of Computer Science, National University of Computer and Emerging Sciences, Islamabad, Chiniot-Faisalabad Campus, Chiniot 35400, Pakistan
[2] Department of Artificial Intelligence and Data Science, National University of Computer and Emerging Sciences (NUCES), Islamabad 35400, Pakistan
[3] Department of Biosystems Engineering, Faculty of Environmental and Mechanical Engineering, Poznań University of Life Sciences, Wojska Polskiego 50, 60-627 Poznań, Poland
[4] Centre for Smart Systems, AI and Cybersecurity, Staffordshire University, Stoke-on-Trent ST4 2DE, UK
* Correspondence: gniewko.niedbala@up.poznan.pl (G.N.); hafiztayyabrauf093@gmail.com (H.T.R.)

Abstract: In the twenty-first century, machine learning is a significant part of daily life for everyone. Today, it is adopted in many different applications, such as object recognition, object classification, and medical purposes. This research aimed to use deep convolutional neural networks for the real-time detection of diseases in plant leaves. Typically, farmers are unaware of diseases on plant leaves and adopt manual disease detection methods. Their production often decreases as the virus spreads. However, due to a lack of essential infrastructure, quick identification needs to be improved in many regions of the world. It is now feasible to diagnose diseases using mobile devices as a result of the increase in mobile phone usage globally and recent advancements in computer vision due to deep learning. To conduct this research, firstly, a dataset was created that contained images of money plant leaves that had been split into two primary categories, specifically (i) healthy and (ii) unhealthy. This research collected thousands of images in a controlled environment and used a public dataset with exact dimensions. The next step was to train a deep model to identify healthy and unhealthy leaves. Our trained YOLOv5 model was applied to determine the spots on the exclusive and public datasets. This research quickly and accurately identified even a small patch of disease with the help of YOLOv5. It captured the entire image in one shot and forecasted adjacent boxes and class certainty. A random dataset image served as the model's input via a cell phone. This research is beneficial for farmers since it allows them to recognize diseased leaves as soon as they noted and take the necessary precautions to halt the disease's spread. This research aimed to provide the best hyper-parameters for classifying and detecting the healthy and unhealthy parts of leaves in exclusive and public datasets. Our trained YOLOv5 model achieves 93 % accuracy on a test set.

Keywords: plant health detection; precision agriculture; deep learning; object detection; YOLOv5

Citation: Khalid, M.; Sarfraz, M.S.; Iqbal, U.; Aftab, M.U.; Niedbała, G.; Rauf, H.T. Real-Time Plant Health Detection Using Deep Convolutional Neural Networks. *Agriculture* **2023**, *13*, 510. https://doi.org/10.3390/agriculture13020510

Academic Editor: Ritaban Dutta

Received: 18 January 2023
Revised: 16 February 2023
Accepted: 17 February 2023
Published: 20 February 2023

1. Introduction

It is challenging to recognize plant diseases by optically analyzing their signs on plant leaves. Skilled agronomists and plant pathologists frequently require help to accurately diagnose certain diseases due to the diverse array of cultivated plants and phyto-pathological issues, resulting in incorrect diagnoses and treatments. Entomologists who are requested to make these diagnoses by visual examination of diseased plant leaves would greatly benefit from the development of an ASO (automated systems operation) to identify and diagnose plant diseases [1]. Humans eat food that comes from plants. Furthermore, because plants create oxygen, they aid in maintaining the oxygen in the air.

Without agriculture, the life we live would not be possible. All the goods we use daily, such as oil, firewood, fiber, pesticides, medicine, and rubber, are extracted from plants.

Plants, crops (fruits, vegetables, etc.), and the natural world are significant to humans. Engaging with nature is crucial for improving an individual's quality of life and delivering various measurable advantages to human beings, including psychological/cognitive advantages [2]. A plant comprises several parts, such as leaves, flowers, stems, and roots. A farmer may cultivate many plants, but diseases can impede their growth. Disease attack is one of the primary reasons that lead to plant loss. Each year 10–16% of plant production is reduced due to disease [3]. In past decades, the health consequences of exposure to nature have been described in detail. However, the role of plants, such as money plants, has received enormously little interest, as compared to the range of crop studies. Urban people spend 80–90% of their lives in houses, offices, schools, etc. Good environments are very important for their health. Indoor plants play an essential role in a good and healthy environment, but their impact on the surroundings and human beings has not yet been quantified [2]. Plants are crucial for removing harmful emissions from the atmosphere and enhancing the ecosystem as well as providing a positive psychological effect, increased health, and a comfortable indoor environment. The above studies have shown that plants benefit humans, so caring for plants is also essential. However, there needs to be more research conducted regarding the money plant.

Currently, several strategies for minimizing plant disease include the removal of damaged plant leaves, mechanical cultivation, and the use of various pesticides. Using the services of an agricultural professional is a simple way to detect plant disease. However, manual disease detection takes a long time and is arduous work. The typical strategy is to use pesticides [4,5]; however, excessive pesticide use may enhance plant growth while harming plant quality. However, spraying more pesticides on plants before even assessing the amount of pesticide required for a specific crop could negatively affect the environment and human health [6].

However, plant disease recognition is more accessible through machine learning. The use of this technique has been identified as a vital advancement and management success for plant disease. The agriculture sector's productivity has grown as a result as well. Additionally, image processing methods have been added to this technology, which has advanced during the last three years to its present state [7,8]. The nation's problems, such as lurgies affecting plants and humans, could be mitigated. Once the unhealthy plants were recognized, they covered a large region.

Machine learning (ML) has been widely employed in the world today. AI, known as ML, enables machines to interact with people and understand their needs. Additionally, it enables machines to perform actions usually performed by people. Several issues impact the reliability and performance of this technology, making it challenging for ML methods to identify specific disorders. Figure 1 shows the traditional method of image processing.

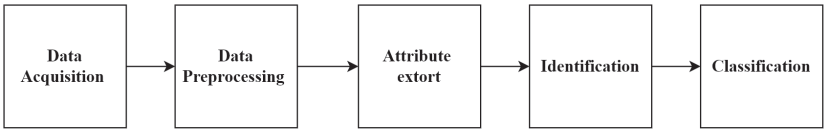


Figure 1. Traditional image pre-processing techniques: The basic method of identifying plant diseases using conventional image recognition processing technologies.

The first problem was the computational time involved with machine learning and deep learning because some methods used to diagnose such diseases must be updated as they rely on obsolete information. Another problem has been segmentation reactivity [9], which refers to the high sensitivity and precision in the relevant field that is required (ROI). A significant amount of resources are required to create and implement the bulk of machine learning and deep learning tasks. Organizations that use this technology for people and plants are frequently supported by non-government organizations, which may affect the development and use of this technology.

To identify diseased leaves, image recognition may be performed. According to background research, by scanning images of infected and healthy leaves, experts in this field have been able to compare them accordingly [10]. Several traditional image processing techniques were used. The image processing had the following steps, e.g., the images were segmented first, then the plant disease features were retrieved, and finally, the disease was categorized. This research developed an attribute image-based method for classifying wheat plant diseases and used an SVM to diagnose the condition [11] successfully. The capability for generalizing new datasets has to be enhanced since attribute information can only be learned superficially. Deep-learning methods, however, are being used in farming research more often, as they can rapidly retrieve deep feature data and are quicker and more accurate than the standard machine learning (ML) algorithms [12,13]. Researchers have created a rider neural network based on the sine-cosine algorithm and discovered that the classifier's identification performance increased significantly [14]. As has been demonstrated, deep-learning models have proliferated over the last decade. Experts have been assisted by the numerous methods used in machine learning (ML) and deep learning (DL) to quickly identify the causes of plant diseases and evaluate their symptoms. In summary, deep learning has shown successful outcomes in the identification of plant diseases. Our research was conducted on categories of plants because both plants and crops have the same importance for the environment and humans. For the sake of world health and well-being, it is essential to identify plant diseases accurately.

Figure 2 below depicts three of the most prevalent diseases that affect plants. Each has distinctive signs and side effects on the leaf; these can be used to differentiate and label the infections visually by the human eye and automatically by algorithms. In today's ever-changing environment, early detection of disease and early prevention is critical to avoid issues that could otherwise arise. Figure 2 shows two types of data depicting the classes in the dataset: the first depicts an unhealthy sample while the second shows healthy sample.Therefore, the main goal was to achieve accurate detection between diseased and healthy, and a deep-learning network-based YOLOv5 model was used in this study. The described model has been useful in plant and crop plantations and agricultural production, according to experiments conducted on images with complicated backdrops. Hence, deep learning (DL) is the most accurate and precise way to identify diseases in plants with the best results. The following is a summary of this work's significant contributions:

- A dataset with several scenarios and sizes was created. There were a variety of sophisticated backdrops, with varying lighting and perspectives, that featured images of damaged leaves. This offered the optimal information to make plant disease research easier.
- This study provided thorough, detailed literature on the methods currently used in plant disease identification. It also reviewed the literature on the datasets used in the research and provided a comparative analysis that identified various studies' advantages and disadvantages.
- This study used different-deep-learning algorithms for the classification of plant anomalies.
- This study established the hyper-parameters for the applied deep algorithm for comparison with state-of-the-art plant disease classification algorithms (standard machine learning (ML) approaches)
- This study evaluated the applied deep-learning algorithm with standard efficiency parameters.
- This research evaluated 4 different target detection techniques, and the results of the trials demonstrated that the suggested approach achieved an mAP of 93.1% on both exclusive and public datasets at 120 frames per second (FPS). Precision planting, visual management, and intelligent decision-making were all features of the YOLOv5 algorithm for economic productivity.

The following represents the paper's essential topics. Before presenting the study objectives, the relevant material and related work are introduced. Next, the model concept and improvement, model training and testing, research object (dataset), and operation

procedure are introduced. Then, the findings are examined to demonstrate the viability and progress of the model used in this study. Finally, the overall research is reviewed, and suggestions for further research are offered.

Figure 2. Data set visual representation.

2. Related Work

This research reviewed the detection of healthy and unhealthy leaves in plants; the different databases used to collect datasets, feature extraction, and feature selection techniques; and the machine-learning and deep-learning models reported in the literature.Furthermore, this research included an overall pipeline strategy employed in previous studies. Some researchers worked on different stress responses in plants using deep-learning models. Plants must handle stress due to various types of environmental factors, such as water stress, dust, and diseases caused by bacteria that affect their health. In the past few years, machine-learning techniques have been used for solving such issues, but deep learning, CNN, and other algorithms have also been used for detection, evaluation, and comparison [15].

Meanwhile, other research has examined water stress in plants using deep-learning techniques. Plant growth is controlled directly by plant water stress and only indirectly by soil water stress, and this has resulted in losses of up to 90% due to water stress and heat stress. Image processing is a directed way to measure the water stress in plants beyond the limitations of traditional image processing. This researcher used deep-learning techniques (Alex Net, Google Net, and Inception V3) on maize (*Zea mays*), okra (*Abelmoschus esculentus*), and soybean (*Glycine max*) crops to identify water stress based on a 1200 images dataset [16].

Maize is the most cultivated crop in the world [17]. Its stable production has a significant influence on food security. In addition, maize is sensitive to drought stress. Many studies have shown that water stress during the agronomy and tasselling stages reduced plant yield and caused a 29–32% final dry. Drought has become a vital factor that lemmatizes maize yield. This research mainly focused on maize drought detection. Many previous studies have explained the traditional detection methods based on power, low-cost, and manual experiments.In recent years, image processing techniques and computer vision technology have become widely used. Image processing is low cost, low power, and convenient for real-time analysis techniques. According to the research, the water supply in the two weeks before and after the pollination period determines the final yield. In this study, they used different directions and wavelengths of Gabor filters. The results of the experiments were 98.84% [18].

Avocado is a tropical fruit with a significant economic value in Florida [19]. The research presented and evaluated an automated detection technique for avocado trees. Remote sensing techniques have been used for detection and evaluation in order to compare

healthy and unhealthy trees. In this study, laurel wilt (LW) disease was the focus, and they differentiated sick and healthy trees (H). The detection of LW during its early stages is challenging because it shows symptoms similar to other stress factors: nutrient deficiency and salt damage. The accuracy of the experiment was 99%. Therefore, low-cost remote techniques can be utilized to differentiate healthy and unhealthy plants.

However, this study focused mainly on plant disease automation in farming. It is a major concern in many countries, as food consumption is growing at a rapid rate due to population growth. Furthermore, modern technologies have improved the efficiency and accuracy of disease detection in plants and animals. The detection procedure is the first step in a workflow designed to combat diseases and limit their spread. The research focused mainly on details about diseases and the implementation of artificial intelligence to detect them quickly. Moreover, machine learning and deep-learning models are used for the automatic detection of plant disease. Various datasets have also been examined for achieving better outcomes for the research community [20].

Furthermore, in other research [21], PlantDoc showed a crop production loss of 35% as a result of plant disease. The early diagnosis of plant infection is still challenging due to a need for more tools and knowledge. This research examined the possibility of using computer vision technologies for low-cost early diagnosis of plant diseases. The main contribution of this research was the PlantDoc dataset that was mentioned in this research.

Previous research [22] has shown the importance of AI in different fields, such as in medical communication, object recognition and detection, etc. This research was only focused on bell pepper.Usually, bell pepper farmers are unaware if their plants are affected by bacterial spot disease. The solution is early detection of infectious spot disease in bell pepper plants. Bacterial spot disease in a bell pepper was detected using YOLOv5-based symptoms on the leaves. They could detect even a small patch of disease faster and more precisely using YOLOv5, and it enabled them to detect diseases during the early stages of development and take appropriate steps to avoid disease spread.This research developed a technique for identifying bacterial spots in bell pepper plants using farm images.

YOLO single-stage real-time object detection has demonstrated the importance of the YOLO principle of object detection [23]. A single-stage network that forecasts the class probability for multiple boxes is known as YOLO. The YOLO network captures the whole image during training. In the aforementioned paper, they discussed the benefits and difficulties of the YOLO algorithm. They contrasted the usual deep-learning methods with YOLO. They noted that YOLO was efficient because it approached object identification as a straightforward regression issue during comparisons. A simple network was optional. They highlighted the core network's 45 FPS (frames per second) speed. More rapid models have been capable of exceeding 150 frames per second. The mean average accuracy was twice as high as other widely used identification techniques. Background inconsistencies were significantly lower, as compared to other deep conventional algorithms, such as faster R-CNN (regions with CNN). There were some drawbacks, as well. Although their method recognized images rapidly, it lacked sufficient accuracy.

In the article YOLOv2: Lighter, quicker, stronger [24], Joseph Redmon proposed a new model that could fix the flaws in the previous version. YOLOv2 aided in resolving the previous version's drawbacks, with its relatively low recall and error analysis in localization. This improved YOLOv2. In YOLOv2, Darknet-19 was implemented. Joint classification and identification, together with hierarchical classification, improved YOLOv2.

A novel YOLO model (YOLOv3) was proposed that advanced the success of YOLOv2 [25]. YOLOv3 had three times the accuracy of conventional approaches, as compared to traditional networks. YOLOv3 yielded good results, as compared to YOLOv2; however, YOLOv3 had limitations, and the component did not operate as intended for linear x y prediction, IOU threshold, and ground-truth assignment. The authors of [26] presented an implementation of the latest iteration, YOLOv4, in YOLOv4: Maximum speed and efficiency of object detection. YOLOv4 increased the accuracy of object recognition. In addition, it raised the FPS and YOLOv3 mean average accuracy by 10% and 12%, respectively.

In another study of deep models that was specifically focused on crops and also relevant to agriculture, the researchers collected data exclusively before deploying several deep and machine learning models to achieve state-of-the-art results [27,28].

3. Comparative Analysis of Selected Study

This study included a comprehensive literature review of previous object detection experiments, as shown in Table 1. The main challenge was to observe all existing models and then compare them with our YOLOv5 model results. After reviewing the existing studies shown in Table 1, we filtered the studies based on our objectives and performed a comparative analysis. A comparative analysis of these studies is shown in Table 2. The different columns show the reference number of the study, the problem under consideration, and the model(s) used to resolve the problem or proposed as a solution. Table 2 shows the analysis and comparisons of the existing problems and the models proposed as solutions.

Table 1. Summary of related work on the identification of plant diseases.

References	Year	Methodology	Dataset Size	Accuracy
[1]	2018	Deep learning	87,848 images	99.53%
[29]	2021	CNN	20,636	98.029%
[30]	2018	Google net Reset	54,306	99.35%
[31]	2018	ANN	Kaggle dataset	80%
[32]	2007	IOT	Custom data	Good Acc.
[33]	2018	3D leaf tracking	12 plants	comparison
[34]	2019	HSI	6 plants	comparison
[35]	2019	Remote sensing	Sentinel-2	Fast, accurate
[36]	2019	Satellite images	Landsat 8	Fast, accurate
[37]	2016	Machine learning	CR262, MTU	comparison
[38]	2019	Alex net	Apples, cherry	Layers convolution
[39]	2018	Alex Net	Tomato leaves	—
[40]	2019	SSD	Banana	RNN Improve

Table 2. Comparative Analysis of Selected Studies.

Ref	Model	Problem	Dataset
[15]	CNN	Plants emotions detection	Kaggle dataset
[16]	Google Net Alex Net Inception V3	Disease detection using deep learning	Custom dataset
[19]	Image processing	Disease Detection in Avocado	Custom dataset
[21]	Mobile Net Faster RCNN	Early plant disease detection	Custom dataset
[22]	YOLOv5	Bell-pepper disease detection	Bell-pepper custom dataset
[25]	YOLOv3	Object detection in images	Kaggle dataset
[25]	YOLOv4	Object detection in images	Kaggle dataset

4. Materials and Methods

The methodology explains and discusses the proposed solution, data collection, pre-processing, model choice, training, and evaluation. The proposed work was based on deep-learning approaches and discussed this approach in detail. A comparative analysis

compared the proposed approach and traditional deep-learning methods. In the section on pre-processing, novel techniques were used for better results. Furthermore, because CNN has been the most frequently used model in image classification, it also assisted in producing accurate results. The proposed models were evaluated by applying techniques and comparing the results. Figure 3 shows a step-by-step procedure that was used in the plant disease detection and classification process. Furthermore, after gathering the data, they were split into two parts, 80/20 training and testing, respectively. Deep-learning (DL) models were then trained, either from scratch or using a learning strategy, and their training plots were obtained to assess the model's relevance. The next phase involved classifying the images using performance matrices, and the last step involved localizing the images using visualization techniques. Several phases were involved in identifying unhealthy plant leaf regions, which are shown in Figure 3.

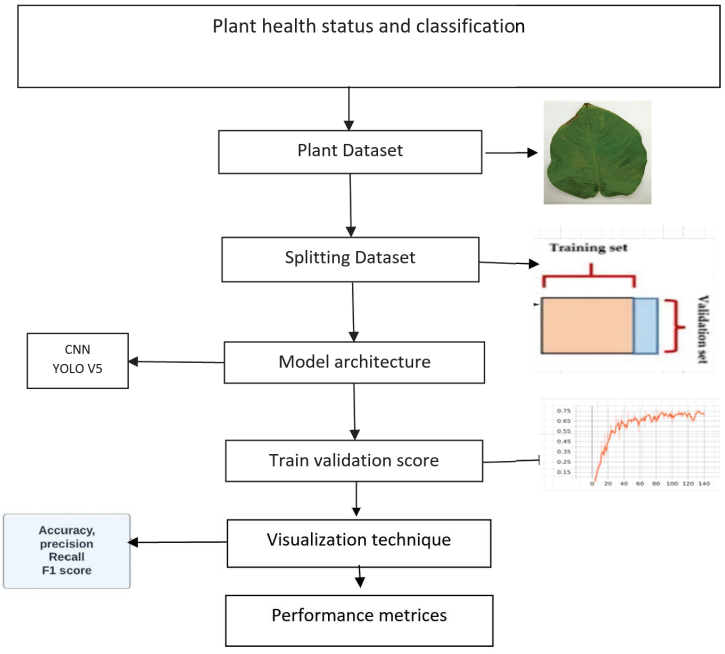


Figure 3. Plant disease identification step-by-step process.

4.1. Experimental Material

The images of the money plant leaves were collected from the University of Agriculture in Faisalabad City, Pakistan, and also from FAST National University of Computer and Emerging Sciences, Faisalabad Chiniot campus, Pakistan. The University of Agriculture is at 31.4300 N, 73.0859 S, and the Fast Chiniot campus is at 31.6076 N, 73.0751 S, with an annual temperature of 15/27 °C. Money plant diseases are typically caused by high humidity and warm temperatures. The dataset was collected in a controlled environment. A HUAWEI/DUAL Lens was used for photography, with an image resolution of 1080 × 2340 pixels and a 19.5:9 aspect ratio. The acquisition of the dataset was one of the key challenges we experienced while working on this research. Using a mobile camera (HUAWEI/DUAL Lens) consumed a great deal of time while capturing images for our exclusive dataset. All the above work was a difficult task for us. Therefore, this research also used the additional dataset offered on the Kaggle website. This research used a public dataset that was collected by Kaggle. The next step was to label the data after obtaining the dataset.

4.2. Preprocessing on Exclusive Data Set

In this research step, the labelIMG tool, as shown in Figure 4, labeled the images. For tool installation, we used Python's pip install labelIMG instruction to install labelIMG. After the installation of labelIMG, the process began with labeling the images. We made bounding boxes around regions of interest (ROI) in the labeling process.

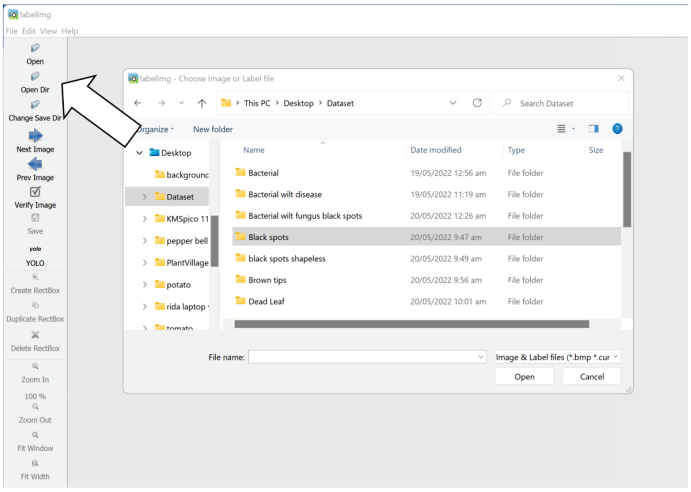


Figure 4. Image selection for labeling.

Briefly, bounding boxes are placed around the unhealthy parts, as shown in Figure 5. We then received a text file as an output after correctly labeling images. Table 3 represents the classes used in our research.

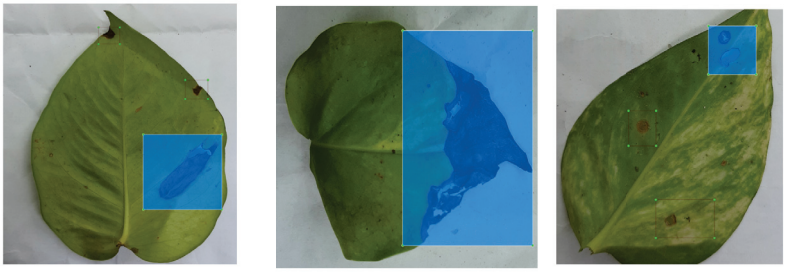


Figure 5. Selected image with unhealthy tag.

Table 3. Dataset Classes.

Class No	Class	Class No	Class
0	Un-Healthy	1	Healthy

One image had multiple bounding boxes depending on the leaves' health condition. The text file had two classes, healthy and unhealthy, with 0 and 1 decimal values. The bounding box class was represented by the first decimal value followed by the centers of the x and y axes, and then the dimensions. The X and Y axes cross at the center point for bounding boxes.These values have been standardized between 0 and 1. We divided the values by the width and height of the image to achieve this. This was conducted rather than using a random integer since it was simpler to estimate values for the network between 0

and 1. The number of bounding boxes drawn determined the number of lines there were in the text file. The dataset was divided into training, testing, and validation. Table 4 shows the total number of training, testing, and validation images.

Table 4. Total Dataset.

Training Sample	Testing Sample	Validation Sample	Total Sample
2000	206	105	2311

After labeling the dataset as healthy or unhealthy, all the above work was conducted with the help of an anaconda prompt. All the above tasks performed in labeling were conducted with labelIMG.

4.3. Preprocessing on Public Data Set

This research work also used a public dataset from the Kaggle site, as shown in Figure 6. The plant village dataset had 28 different classes of plants with 54,309 different images. This research was focused on bell pepper and potato leaves. A total of 1000 images were collected. The images were preprocessed to normalize their dimensions, reduce noise, remove backgrounds, and minimize unwanted distortions. Each dataset image was annotated with the labelIMG tool, and many annotation tools were used, such as coco json, TensorFlow object detector, scale, label box, etc. Therefore, labelIMG tool assisted in making bounding boxes around the leaves in all images. In real life, the images could include many leaves or a mix of infected and healthy leaves. All of the leaves in the images were explicitly labeled with their relevant healthy or unhealthy classes. The complete leave was present in the box while labeling the boxes, and the bounding box area was at least 1/8 (roughly) of the image size. After labeling, all the coordinates of boxes in an image and their related class labels were saved individually in a YOLO file for each image. The resulting pictures were utilized in the next phase as input.

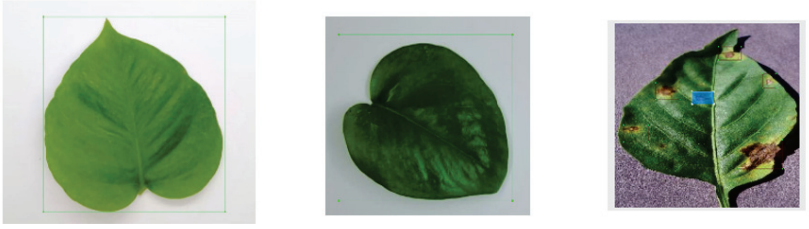


Figure 6. Plant village dataset with bounding boxes.

4.4. Plant Disease Detection and Classification Model Implementations

4.4.1. EfficentDet

Google researchers recently unveiled Efficient Net, which is a convolutional network. The three sections are part of the object detector group known as EfficentDET. Efficient NET serves as the foundation of Google's information collection process. However, it also recycles the milestones from ImageNet's pre-trained network because it employs the same spacing scaling parameters as EfficientNet-B0 through B6. The bi-directional feature network serves as the feature chain for EfficentDET, as shown in Figure 7. BIFPN (bi-directional feature pyramid network) is used for quick and simple 2D feature engineering. The bi-directional feature pyramid uses scales ranging from levels 3 to 7, and the merging process is performed several times. The magnitude levels identify objects in the image that are of various sizes and densities. Image quality and size are directly related to the scaling factor. Consequently, p3 > p4 > p5 > p6 > p7 and p3 > p4 > p5 > p6 > p7 are used to indicate the resolution and quality of objects that were discovered. The bi-directional feature

network is shown in Figure 7. The data flow direction of this method was bi-directional, as shown by the top-down and bottom-up approaches.

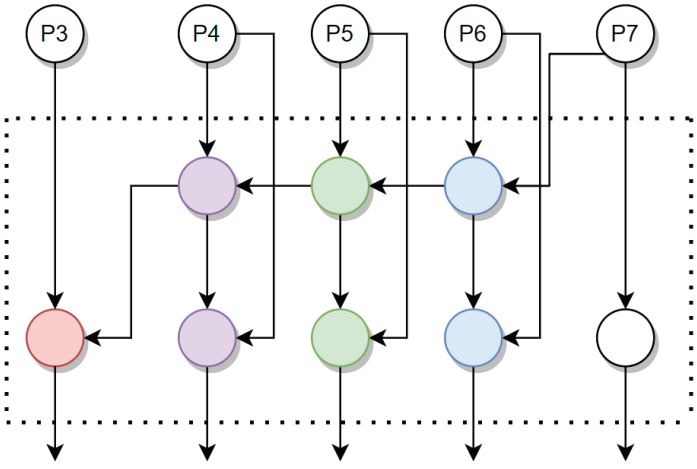


Figure 7. EfficentDet model architecture.

4.4.2. FasterRCNN

Faster R-CNN, the most popular modern variant of the R-CNN series, was released for the first time in 2015. The object proposal method was the only independent network component in the fast regional convolutional network. Faster regional-based CNN and simple RCNN both employ object detection algorithms that are dependent on the hardware capacity. The CPU-based selective search technique, which processes one picture in approximately two seconds, is shown in Figure 8. Moreover, it employs a region proposal network (RPN) to provide object detection recommendations. This improves feature representation overall by reducing the object proposal time per picture from 2 s to 10 milliseconds and enabling the object detection step to share layers with the succeeding detection stages.

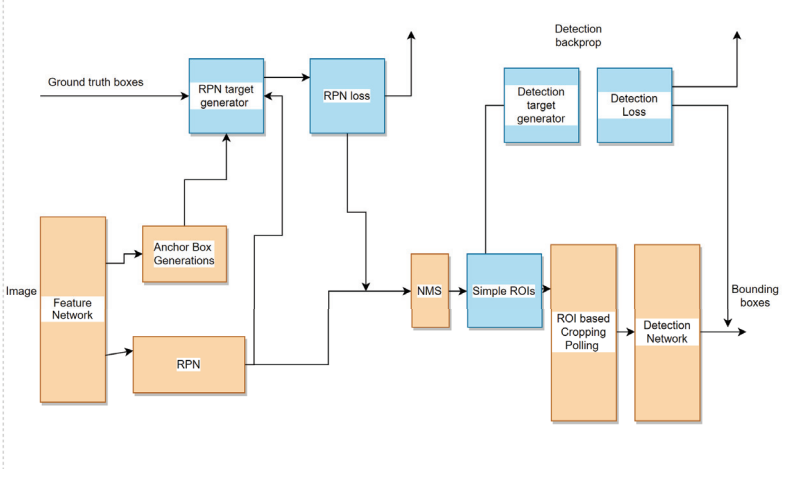


Figure 8. The architecture of Faster R-CNN.

4.4.3. The Principal of YOLOv5 Model

The two-stage object detection methodology led the market before YOLO [25]. It located regions using region of interest-based classifiers and passed those areas on to a stronger classifier. This approach used a lot of resources and necessitates several runs, yet it produced reliable results with high mAP values. The picture is first separated into columns, each of which has an identical SxS-sized dimensional area.

$$S \times S \tag{1}$$

The next step is for each cell to identify and pinpoint the items it contains using the bounding box dimensions, the object name, and the likelihood that the object is present in the columns. Each column "works by itself", processing the grid concurrently while using fewer computer resources as well as training and inference time. In addition, YOLO outperforms other real-time object identification algorithms and produces state-of-the-art results. Furthermore, the output dimensions are

$$S \times S(B \times 5 + C) \tag{2}$$

YOLO Versions

There are six different models introduced to date in the YOLO series (e.g., YOLOv1, YOLOv2, YOLOv3, YOLOv4, YOLOv5, YOLOv6). Our research focused on the YOLOv5 and below series and had better results than the traditional machine-learning model.

YOLOv5 Module

Glenn Jocher presented the one-stage target identification method known as YOLOv5 [31] in 2020. Four network model variations of the YOLOv5 model were distinguished based on differences in network depth and height: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The YOLOv5s network had the highest computational speed but the lowest average processing, whereas the YOLOv5x network exhibited the opposite traits. The YOLOv5 network's model size was around one-tenth that of the YOLOv4 network. Its detection and localization abilities were faster, and its precision was on par with YOLOv4.

4.4.4. YOLOv5 Architecture

The YOLOv5 model in Figure 9 included three crucial components, similar to other single-stage object detectors.

1. Backbone
2. Neck
3. Head

The essential purpose of the model backbone is to extract significant characteristics from an input image. The CSP network was deployed as the backbone to extract important attributes from an input image in YOLOv5. CSPNet demonstrated a considerable reduction in processing time. The primary purpose of the model neck is to produce feature pyramids. Feature pyramids assist the model in making suitable object scaling generalizations. It facilitates recognition of the same item at various sizes and scales. Models perform effectively on unobserved data due to the usage of feature pyramids. Other models such as FPN, BIFPN, and PANet use other feature pyramid methodologies. PANet was utilized in YOLOv5 as a neck to obtain feature pyramids. The last step was the final detection step, which was carried out using the model head. The final output vectors were generated with class probabilities, object scores, and bounding boxes after anchor boxes were applied to the feature.

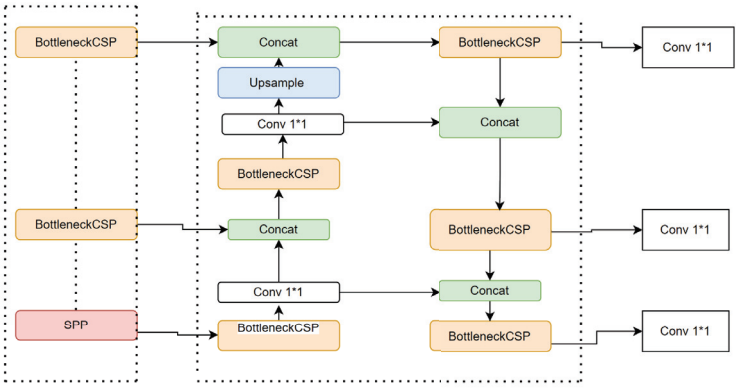


Figure 9. Overview of YOLOv5.

4.4.5. Activation Function

In every deep neural network, the selection of the activation function is of great concern. Many activation functions, such as Leaky ReLU, Mish, and Swish, have recently been developed.

4.4.6. Optimization Function

Optimization was conducted with the assistance of SGD and ADAM.

4.4.7. Cost Function or Loss Function

A compound loss was produced for the YOLO algorithms based on the object score, class probability score, and bounding box regression score. For the loss computation of class probability and object score, ultralytics adopted the binary cross entropy with the logit function from PyTorch. As compared to earlier versions of the YOLO series, version 5 provided superior detection accuracy, a lightweight design, and a quick detection time. Accuracy and effectiveness were important for identifying plant diseases. Therefore, the YOLOv5 model improved disease detection in the money, potato, and bell pepper plants. The model architecture is shown in Figure 9.

4.5. Training Exclusive and Public Datasets on YOLOv5

The following stages were involved in training a custom YOLOv5 model:

1. First step was the environment configuration for YOLO.
2. Obtaining the YOLOv5 repository and installing plugins were the initial steps. As a result, the programming framework was prepared for the execution of instructions for object identification training and inference.
3. We trained a model on the free training environment provided by Google Collab.
4. Google Collab was likely operating on a Tesla P100 GPU.
5. Next, we downloaded the custom data from Roboflow in a YOLOv5 format.
6. After labeling the data, it was then exported to Roboflow. After uploading data into Roboflow, it was converted into one of these formats (VOC XML, coco json, TensorFlow object detection, etc.).
7. After uploading the data, we selected the preprocessing steps and augmentation.
8. Roboflow automatically divided the data into training, testing, and validation sets.
9. After annotating the images, we chose the YOLOv5 pyTorch format.
10. After the format steps, Roboflow provided a key or PIP package, as shown in Figure 10.

```
#from roboflow import Roboflow
#rf = Roboflow(api_key="YOUR API KEY HERE")
#project = rf.workspace().project("YOUR PROJECT")
#dataset = project.version("YOUR VERSION").download("yolov5")
```

Figure 10. Roboflow keys.

The above YOLOv5.YAML file name data.yaml file contained information about the exclusive dataset, as well as the location of the YOLOv5 images. With the data.yaml file, the training process could start immediately:
Img: Image size as the input image
Batch size: determine the length of the batch for training
Epochs: define the training steps
Cfg: model configuration

During training, mAP @ 0.5 minimum average precision was the major concern to determine the detector's performance.

YOLOv5 Evaluation and Validation Metrics

Figure 11 shows verification metrics to determine the training process performance. Once the training process was completed, validation accuracy was performed, as shown in Figure 11.

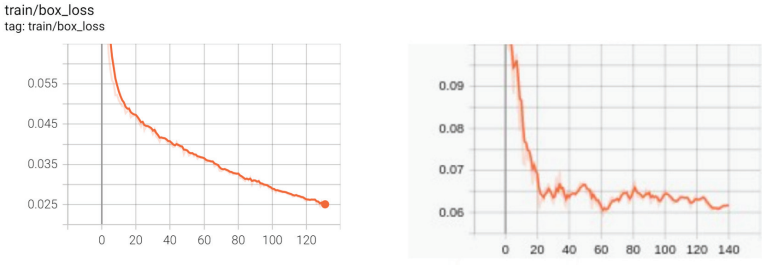


Figure 11. YOLOv5 training and validation graphs.

4.6. YOLOv6

With each iteration, the You Only Look Once model's goal was consistent: to rapidly learn how to predict bounding boxes for specific objects while preserving accuracy. The better a model is, the less hardware that is required to generate and operate it. YOLO models use an image as input and transmit it through a number of fully linked layers in the backbone. You only look at one model after the model uses the neck to represent these backbone elements. After receiving the neck features, the three heads of the YOLO models anticipate objectivity, class, and box reversion. In order to develop an efficient representation and representation-path integral-based convolutional (PAN) neck, YOLOv5 reconstructed the YOLO backbone and neck while accounting for the hardware. It had already been established that YOLOx and YOLOv6 had detachable heads, which implied the network had an additional layer separating these attributes from the ultimate head. Along with structural changes, the YOLOv5 repository also included several enhancements for the training procedure. These enhancements consisted of SIoU box regression loss, SimOTA tag assignment, and anchor-free (not NMS-free) training [33].

5. Results and Analysis

This part of the research discusses the results and observations described in the previous part of the research.

5.1. Dataset Validation Results

After collecting the exclusive dataset, the main task was to perform validation on the dataset and compare it with the public dataset. Here, the validation test was conducted on the exclusive dataset, as shown in Table 5, and then compared with the public datasets, as shown in Table 6.

Table 5. YOLOv5 on exclusive dataset.

Model Name	Accuracy	No. of Images	Precision	Recall	Dataset
YOLOv5	0.61	1k	0.60	0.62	Exclusive Dataset

Table 6. YOLOv5 on public dataset.

Model Name	Accuracy	No. of Images	Precision	Recall	Dataset
YOLOv5	0.60	1k	0.61	0.63	Public Dataset

The comparison chart between the private and public datasets is shown in Figure 12. Additionally, mAP (mean average precision) has established a benchmark for precision and recall in proprietary datasets, requiring the use of 2000 samples as training data.

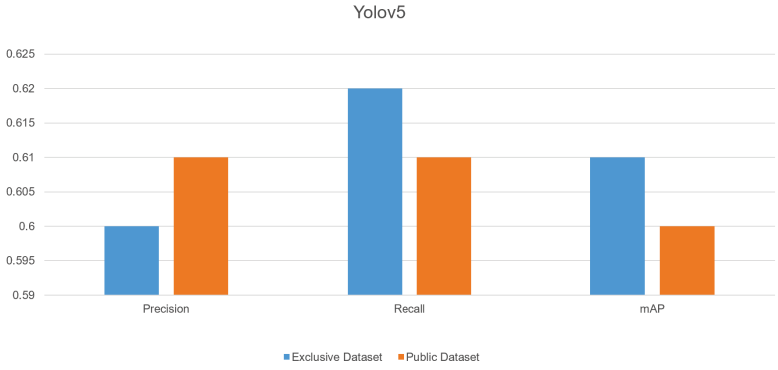


Figure 12. Dataset Validation Comparison Graph.

5.2. FasterRCNN

FasterRCNN has a considerable architecture. FasterRCNN is a fusion of Fast RCNN and region proposal, which makes algorithms very fast and accurate with low computational cost on hardware, such as CPU and GPU. We deployed the exclusive and public datasets on FasterRCNN architecture and achieved results, as shown in Table 7, which could be better than other existing studies.

Table 7. FasterRCNN results.

Model Name	Accuracy	No. of Images	Precision	Recall	Dataset
FasterRCNN	0.45	2k	0.49	0.47	Custom, public dataset

Figure 13 shows the visual results of the model, and Figure 14 shows the precision-recall and mAP, which was 35%. This ratio was not optimal for object detection models.

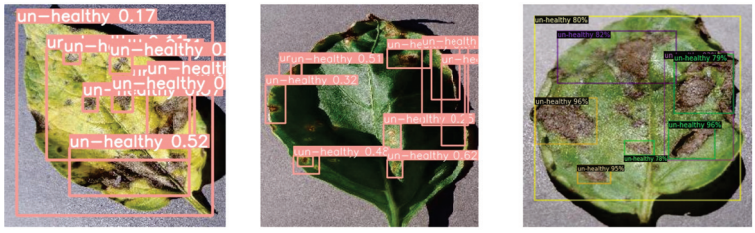


Figure 13. FasterRCNN visual representation.

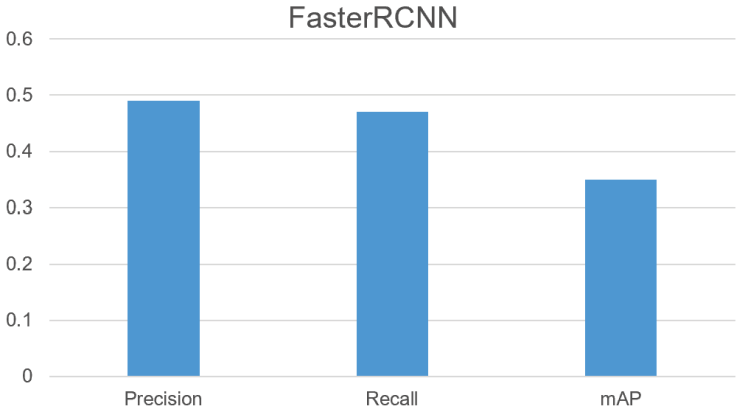


Figure 14. FasterRCNN comparison graph.

5.3. EfficentDET

As discussed above, EfficentDET architecture in implementations presents the visual representations of the results obtained from the EfficentDet code. We added the exclusive and public datasets on EfficentDet architecture with the same hyper-parameters, as shown in Figure 15. This was also not optimal, as compared to other existing studies. Figure 16 shows the visual results of EfficentDET.

```
#smells like some free compute from Colab, nice
gtf.Train_Dataset(root_dir, coco_dir, img_dir, set_dir, batch_size=8, image_size=512, use_gpu=True)
```

Figure 15. EfficentDET parameters.

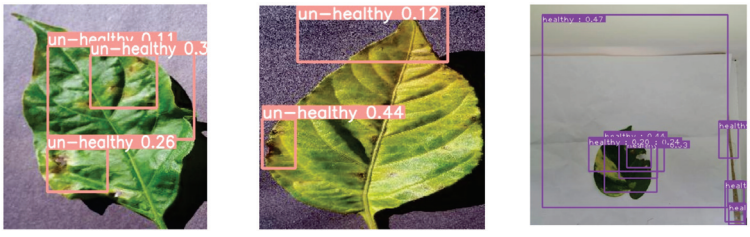


Figure 16. EfficentDET visual results.

Table 8 displays the EfficentDET findings and final results, which showed 35% accuracy with around 39% precision and 49% recall value.

Table 8. EfficentDET results.

Model Name	Accuracy	No. of Images	Precision	Recall	Dataset
EfficentDET	0.35	2k	0.39	0.49	Custom, public dataset

Figure 17 depicts the average accuracy, which was about 60% on average. Though inferior to FasterRCNN, this was sufficient for an object detection model. Farming requires accurate detection while detecting objects. Many additional strategies were conducted, such as color correction and increasing epochs, but the accuracy required improvement. Hence, we switched to the best object detection model, YOLOv5.

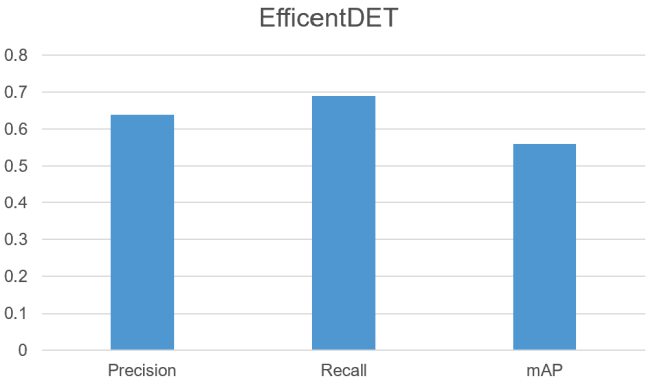


Figure 17. EfficentDET comparison graph.

5.4. YOLOv5 Experiments on Exclusive and Public Datasets

YOLOv5 is a state-of-the-art object recognition model that provides excellent mAP at low-resource demand. Firstly, YOLOv5 was applied to an exclusive dataset, but due to the few available datasets, we had to merge the exclusive dataset with the public to obtain minimum average precision. Figure 18 depicts the model summary, layers epochs, image size, etc.

Model summary: 213 layers, 7015519 parameters, 0 gradients, 15.8 GFLOPs

Class	Images	Labels	P	R	mAP@.5	mAP@.5:.95: 100% 2/2 [00:01<00:00, 1.89it/s]
all	52	82	0.609	0.624	0.617	0.426
healthy	52	8	0.65	0.875	0.895	0.713
un-healthy	52	74	0.568	0.374	0.338	0.139

Accuracy 0.69%

Figure 18. YOLOv5 hyperparameters.

In the second experiment, we examined the validation loss, mAP, precision, and recall. Figure 19's graphs show the minimum average precision (mAP) at 0.5 and the minimum average precision in the range of 0.5 to 0.95. The criteria for accuracy, recall, and intersection over union (IoU) was used to plot the graphs.

$$Precision(P) = \frac{TP}{TP + FP} \tag{3}$$

$$Recall(R) = \frac{TP}{TP + FN} \tag{4}$$

- *TP* (True Positive) = How many instances were accurately detected
- *FP* (False Positive) = Number of incorrectly identified cases between healthy and unhealthy leaves
- *FN* (False Negative) = Number of unidentified cases between healthy and unhealthy leave
- IOU = Intersection over union
- K = threshold

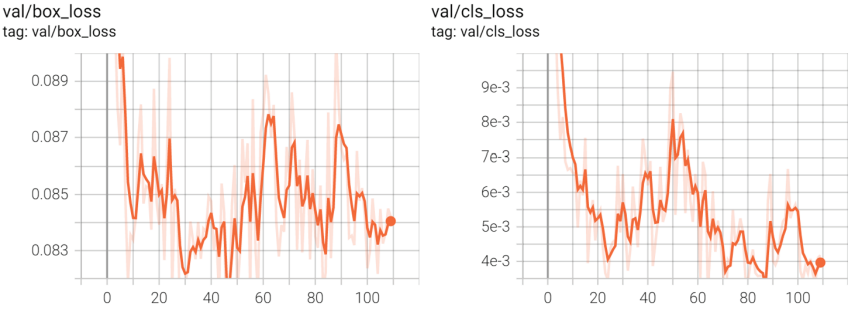


Figure 19. Validation loss representing a fluctuation in results due to insufficient datasets.

Graphs of mAP (mean average precision), precision, and recall for training data of 1000 samples are displayed in Figures 20 and 21. Graphs are an excellent tool for exploratory data analysis, as they provide an overview of the relationships throughout the whole dataset.

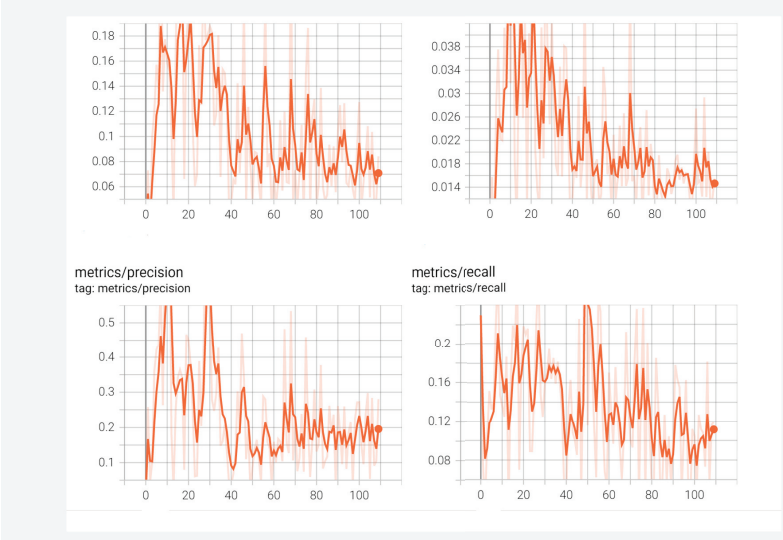


Figure 20. YOLOv5: mAP precision, and recall for training data of 1000 samples.

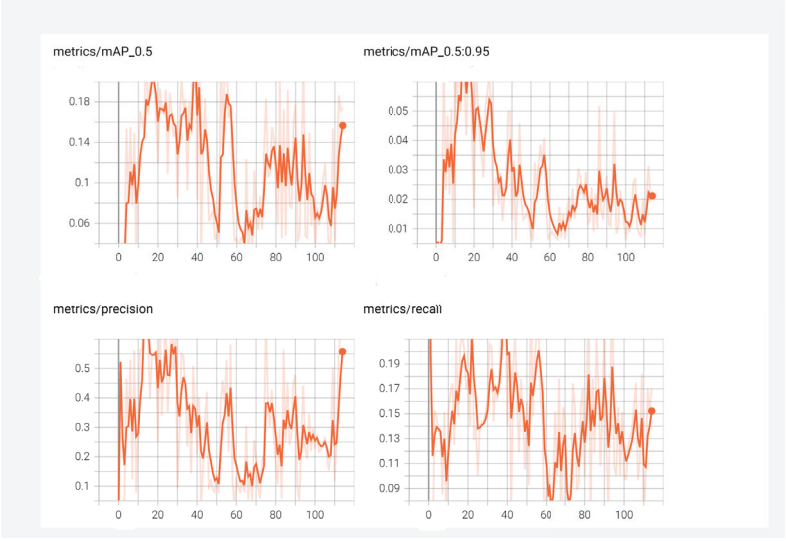


Figure 21. mAP Graphs mAP, precision, recall, is increasing or decreasing with changes in hyperparameters.

5.5. YOLOv5 Results

YOLO environment configuration with complete and public datasets is described below. To begin training, YOLOv5 YAML required two files. The first YAML defined the locations of the test and training data as well as the number of classes of objects being detected and the names of the items that belong to each class. The tuning parameters for training and testing are shown in Figure 22. The overall findings of our model are shown in Table 9, which were 93%, with a precision of 75% and a recall of 95%. The major advantage of YOLOv5 was that, as compared to FasterRcnn, YOLOv5 operates 2.5 times faster and managed better performance and detection of even small objects.

Table 9. YOLOv5 testing results.

Model Name	Accuracy	No. of Images	Precision	Recall	Dataset
YOLOv5	0.93	2k	0.75	0.95	Custom, public dataset

```
!python train.py --img 416 --batch 16 --epochs 150 --data {dataset.location}/data.yaml --weights yolov5s.pt --cache
```

Figure 22. YOLOv5 experiment 2.

Figure 23's graphs show the mAP at 0.5 and the mAP in the range of 0.5 to 0.95. The mAP graph was increasing incrementally with increasing epochs. The criteria for accuracy, recall, and intersection over union (IoU) were used to plot the graphs. Figure 23 depicts a precision graph that increases up to 25 epochs before fluctuating.

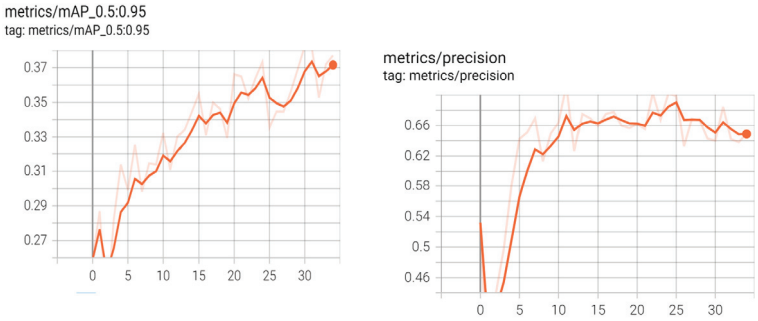


Figure 23. YOLOv5 training and validation graphs.

Figure 24 depicts the final recall metric, which was increasing with each epoch.

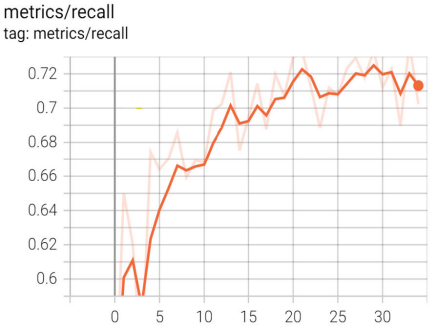


Figure 24. Recall graph YOLOv5: The recall is gradually increasing with the increment of epochs.

Figure 25 shows that the validation loss of YOLOv5 decreased significantly until epoch 20. After that, the validation loss declined and stopped at 0.06 and 0.05, at epoch 30.

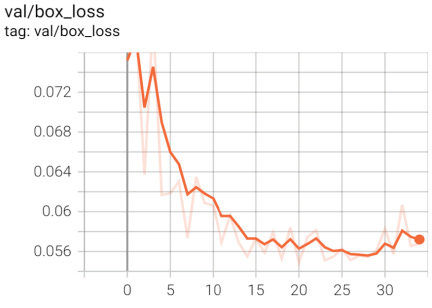


Figure 25. Validation loss graph YOLOv5: Validation loss was decreasing gradually, indicating the model prediction was significant.

Figure 26 shows the curve that indicated the confidence via F1 score.In the insight figure, the orange line shows the healthy part of detection, and the green color shows the unhealthy part of detection. The F1 curve indicated the PPV (precision) and TPR (recall) collectively as one visualization for every threshold.

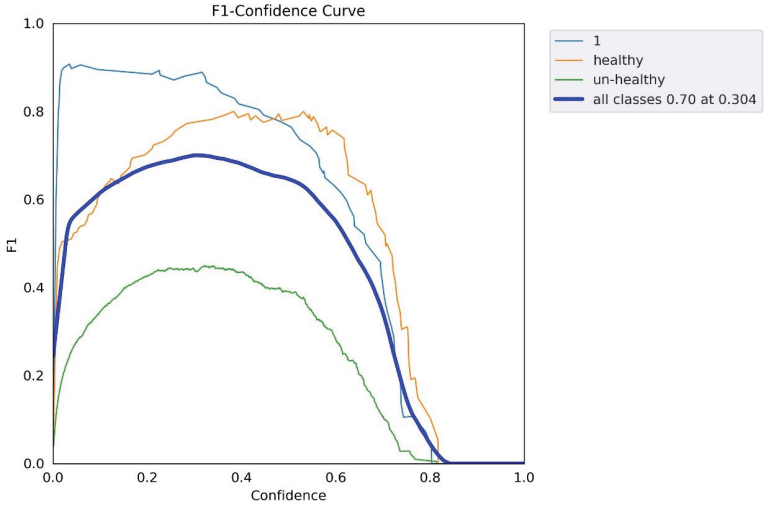


Figure 26. F1 confidence curve: Comparison of progression of mean F1 score across all experiments, grouped by training mechanism.

Figures 27 and 28 depict the confidence via recall and precision curves, respectively, using the R curve.

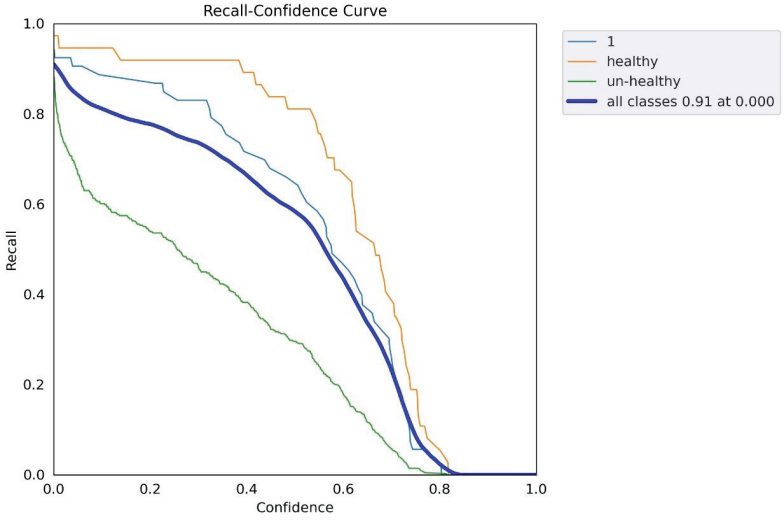


Figure 27. Confidence according to R Graph: Across all experiments, recall growth was compared and classified by training technique. It provided a substitute for the precision–recall curve.

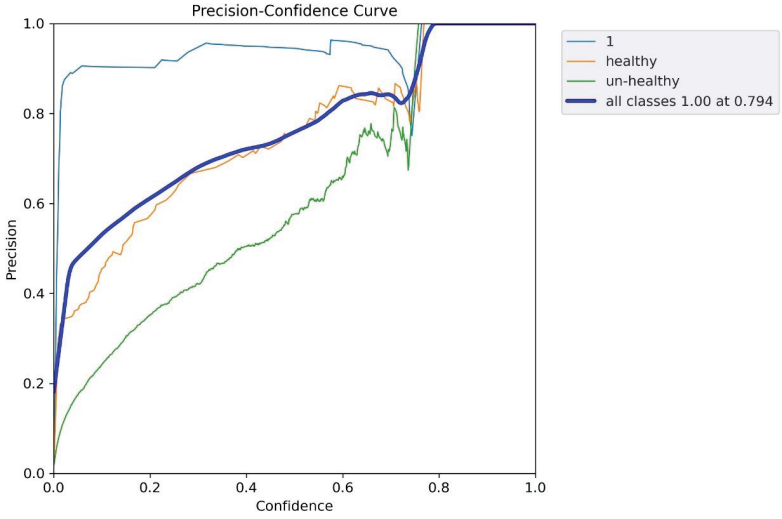


Figure 28. Confidence according to P curve graph: Comparison of the progression of precision across all experiments, grouped by training mechanism.

Figure 29 shows the compromise between recall and precision at different thresholds.

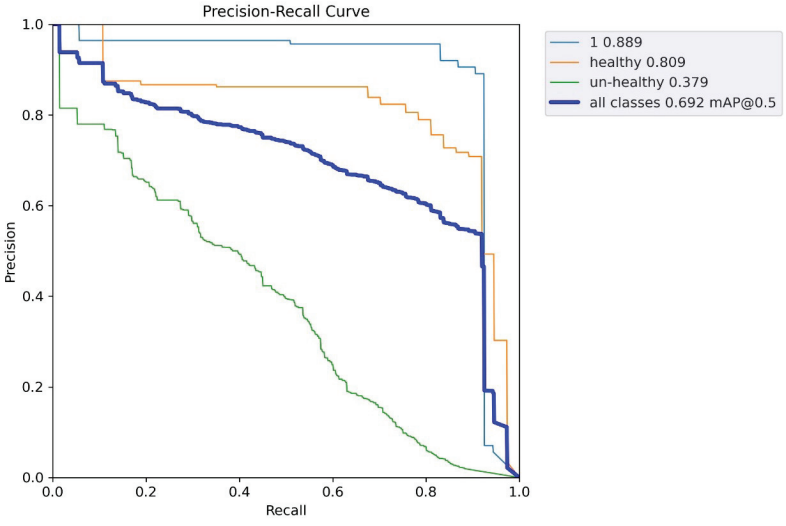


Figure 29. Precision according to recall graph:The precision via recall graph represents both high recall and high precision.

Confusion Matrix

A confusion matrix shows the variations between real and expected values, as shown in Figure 30. It assessed the effectiveness of our machine learning classification model using a table-like structure.

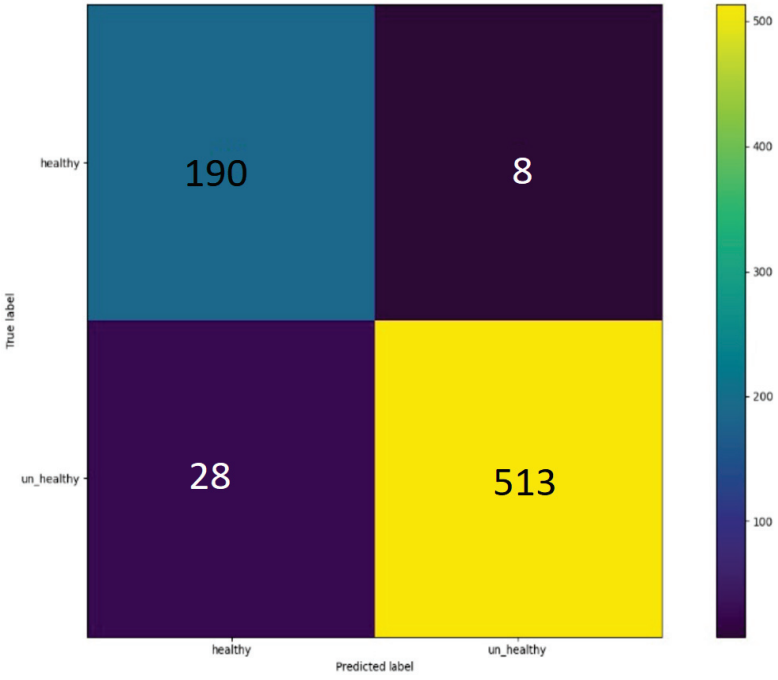


Figure 30. Confusion matrix for YOLOv5: healthy images (190) identified, 513 images were unhealthy and correctly identified.

A total of 190 images were true positives in this detection. Eight images in this selection were false positives, and 28 were false negatives. Furthermore, the last 513 images identified as unhealthy were correctly identified.

5.6. Experiments on YOLOv6

YOLOv6 was introduced recently with multiple changes. The same database was applied on YOLOv6, and results were obtained. As shown in Table 10, we deployed both datasets (exclusive and public datasets) on YOLOv6 and achieved 32% accuracy.

Table 10. YOLOv6 results.

Model Name	Accuracy	No. of Images	Precision	Recall	Dataset
YOLOv6	0.32	2k	0.35	0.58	Exclusive, Public dataset

Figure 31 shows a comparison graph that depicts precision 30%, recall, and mAP of the YOLOv6 model.

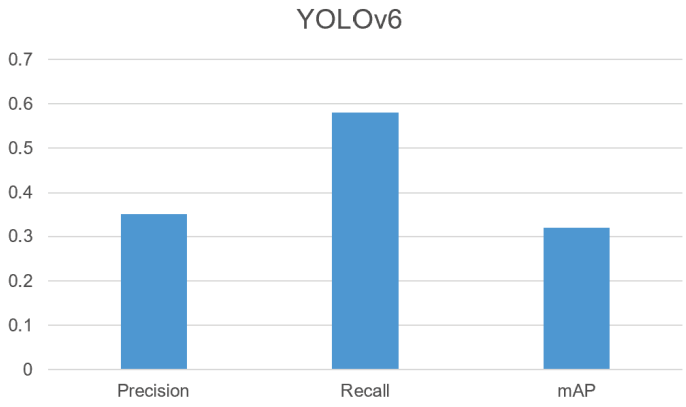


Figure 31. Comparison chart: Comparison between precision, recall, and mAP.

5.7. Comparative Analysis

Figure 32 shows a comparative analysis between different deep-learning algorithms (EfficentDet, FasterRCNN, YOLOv5, and YOLOv6). The graph displays the mAp score that was produced by FasterRCNN, YOLOv5 (our method), EfficentDet, and YOLOv6.

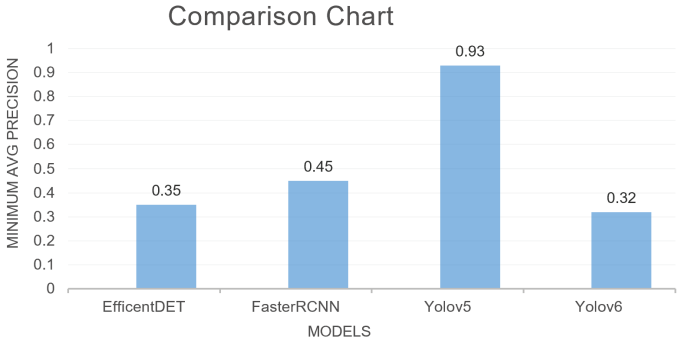


Figure 32. Comparison Chart: Comparative analysis between EfficentDET, FasterRCNN, YOLOv5, and YOLOv6.

6. Conclusions

This study used deep-learning techniques to classify healthy and unhealthy leaves. The identification and recognition of plant diseases in the ecological world are crucial for controlling plant diseases. In this research, a step-by-step procedure was performed. The first step of this research was gathering data; two types of datasets were included. These were the exclusive and public datasets, and then we performed the preprocessing steps on the datasets. Labeling was the most important step in preprocessing because this research had to follow a format acceptable to the selective neural network used for object detection and localization on a region of interest. Therefore, this research provided an acceptable format. Following preprocessing, this research employed augmentation techniques to increase the quantity and quality of the datasets. This research empirically compared four deep neural models to determine the best hyper-parameters and exclusive data validation. Based on FasterRCNN, the model had 0.49 precision with 0.47 recall, and the accuracy was 0.35, which was very low, as compared to the state-of-the-art models. Then, the dataset was deployed on the EfficentDET neural network, and the results were improved, as compared to FasterRCNN but not as good as state-of-the-art models. In this research, validation of the public dataset with the comparison of the exclusive dataset was also performed extensively.

Furthermore, after performing the validation test, the YOLOv5 model was trained on the public and exclusive datasets. Initially, the YOLOv5 model was trained on the pre-trained hyper-parameters, after which we adjusted the hyper-parameters, as shown in the results section, so the the mAP (0.5) was significant, and the final result was 93%. The approach presented in this research outperformed earlier iterations of YOLO in speed and accuracy. It may increase crop productivity by detecting and classifying plant disease.

7. Limitations

Our research had some limitations. There were still instances of missing or incorrect detection. In order to increase the model's detection precision, the model's mechanism needs to be further refined. Furthermore, using a high-resolution lenses for image capture could improve accuracy further.

Future Work

After completing the work described above (e.g., dataset collection, preprocessing, data annotation, data validation, and empirical investigation), we showed that fast detection is possible, but it still requires specific hardware designs. Certain gaps need to be filled by future research. The accuracy, execution time, and minimum average precision of the models should be higher if the dataset is gathered using high-quality lenses (ultra-wide angle) and a large team (7–8 members). YOLOv5 architecture will be improved in the future and deployed as an android application, so it can be used for real-time object recognition with the assistance of YOLOv5's improved architecture.

Author Contributions: Conceptualization, M.K., U.I., M.U.A., M.S.S., G.N. and H.T.R.; methodology, M.K.; validation, U.I., M.U.A., M.S.S., G.N. and H.T.R.; formal analysis, U.I. and M.U.A.; investigation, U.I. and M.U.A.; data curation, M.K. and M.S.S.; writing—original draft preparation, M.K.; writing—review and editing, M.S.S., U.I., M.U.A., G.N. and H.T.R.; supervision, M.S.S., G.N. and H.T.R.; Resources, M.S.S., G.N. and H.T.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research has not received any funding.

Data Availability Statement: The dataset shall be available through declaration from all authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ferentinos, K.P. Deep learning models for plant disease detection and diagnosis. *Comput. Electron. Agric.* **2018**, *145*, 311–318. [CrossRef]
2. Deng, L.; Deng, Q. The basic roles of indoor plants in human health and comfort. *Environ. Sci. Pollut. Res.* **2018**, *25*, 36087–36101. [CrossRef]
3. Balasundram, S.K.; Golhani, K.; Shamshiri, R.R.; Vadamalai, G. Precision agriculture technologies for management of plant diseases. In *Plant Disease Management Strategies for Sustainable Agriculture through Traditional and Modern Approaches*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 259–278.
4. Trivedi, P.; Leach, J.E.; Tringe, S.G.; Sa, T.; Singh, B.K. Plant–microbiome interactions: From community assembly to plant health. *Nat. Rev. Microbiol.* **2020**, *18*, 607–621. [CrossRef] [PubMed]
5. Wang, X.; Yang, W.; Wheaton, A.; Cooley, N.; Moran, B. Efficient registration of optical and IR images for automatic plant water stress assessment. *Comput. Electron. Agric.* **2010**, *74*, 230–237. [CrossRef]
6. Khan, S.; Narvekar, M.; Hasan, M.; Charolia, A.; Khan, A. Image processing based application of thermal imaging for monitoring stress detection in tomato plants. In Proceedings of the 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 27–29 November 2019; IEEE: New York, NY, USA, 2019; pp. 1111–1116.
7. Hasan, R.I.; Yusuf, S.M.; Alzubaidi, L. Review of the state of the art of deep learning for plant diseases: A broad analysis and discussion. *Plants* **2020**, *9*, 1302. [CrossRef] [PubMed]
8. Arivazhagan, S.; Shebiah, R.N.; Ananthi, S.; Varthini, S.V. Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features. *Agric. Eng. Int. CIGR J.* **2013**, *15*, 211–217.
9. Lin, K.; Gong, L.; Huang, Y.; Liu, C.; Pan, J. Deep learning-based segmentation and quantification of cucumber powdery mildew using convolutional neural network. *Front. Plant Sci.* **2019**, *10*, 155. [CrossRef]

10. Blumenthal, J.; Megherbi, D.B.; Lussier, R. Supervised machine learning via Hidden Markov Models for accurate classification of plant stress levels & types based on imaged Chlorophyll fluorescence profiles & their rate of change in time. In Proceedings of the 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), Annecy, France, 26–28 June 2017; IEEE: New York, NY, USA, 2017; pp. 211–216.
11. Shrivastava, V.K.; Pradhan, M.K. Rice plant disease classification using color features: A machine learning paradigm. *J. Plant Pathol.* **2021**, *103*, 17–26. [CrossRef]
12. Chen, M.; Tang, Y.; Zou, X.; Huang, K.; Huang, Z.; Zhou, H.; Wang, C.; Lian, G. Three-dimensional perception of orchard banana central stock enhanced by adaptive multi-vision technology. *Comput. Electron. Agric.* **2020**, *174*, 105508. [CrossRef]
13. Li, Q.; Jia, W.; Sun, M.; Hou, S.; Zheng, Y. A novel green apple segmentation algorithm based on ensemble U-Net under complex orchard environment. *Comput. Electron. Agric.* **2021**, *180*, 105900. [CrossRef]
14. Mishra, M.; Choudhury, P.; Pati, B. Modified ride-NN optimizer for the IoT based plant disease detection. *J. Ambient Intell. Humaniz. Comput.* **2021**, *12*, 691–703. [CrossRef]
15. Abisha, A.; Bharathi, N. Review on Plant health and Stress with various AI techniques and Big data. In Proceedings of the 2021 International Conference on System, Computation, Automation and Networking (ICSCAN), Puducherry, India, 30–31 July 2021; IEEE: New York, NY, USA, 2021; pp. 1–6.
16. Chandel, N.S.; Chakraborty, S.K.; Rajwade, Y.A.; Dubey, K.; Tiwari, M.K.; Jat, D. Identifying crop water stress using deep-learning models. *Neural Comput. Appl.* **2021**, *33*, 5353–5367. [CrossRef]
17. Jiang, B.; Wang, P.; Zhuang, S.; Li, M.; Gong, Z. Drought stress detection in the middle growth stage of maize based on gabor filter and deep learning. In Proceedings of the 2019 Chinese Control Conference (CCC), Guangzhou, China, 27–30 July 2019; IEEE: New York, NY, USA, 2019; pp. 7751–7756.
18. Ahmed, F.; Al-Mamun, H.A.; Bari, A.H.; Hossain, E.; Kwan, P. Classification of crops and weeds from digital images: A support vector machine approach. *Crop Prot.* **2012**, *40*, 98–104. [CrossRef]
19. Abdulridha, J.; Ehsani, R.; Abd-Elrahman, A.; Ampatzidis, Y. A remote sensing technique for detecting laurel wilt disease in avocado in presence of other biotic and abiotic stresses. *Comput. Electron. Agric.* **2019**, *156*, 549–557. [CrossRef]
20. Khan, R.U.; Khan, K.; Albattah, W.; Qamar, A.M. Image-based detection of plant diseases: From classical machine learning to deep learning journey. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 5541859. [CrossRef]
21. Singh, D.; Jain, N.; Jain, P.; Kayal, P.; Kumawat, S.; Batra, N. PlantDoc: A dataset for visual plant disease detection. In Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, Hyderabad, India, 5–7 January 2020; pp. 249–253.
22. Mathew, M.P.; Mahesh, T.Y. Leaf-based disease detection in bell pepper plant using YOLOv5. *Signal Image Video Process.* **2022**, *16*, 841–847. [CrossRef]
23. Khan, S.; Tufail, M.; Khan, M.T.; Khan, Z.A.; Anwar, S. Deep learning-based identification system of weeds and crops in strawberry and pea fields for a precision agriculture sprayer. *Precis. Agric.* **2021**, *22*, 1711–1727. [CrossRef]
24. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. Scaled-YOLOv4: Scaling cross stage partial network. In Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13029–13038.
25. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
26. Paez, A.; Gebre, G.M.; Gonzalez, M.E.; Tschaplinski, T.J. Growth, soluble carbohydrates, and aloin concentration of Aloe vera plants exposed to three irradiance levels. *Environ. Exp. Bot.* **2000**, *44*, 133–139. [CrossRef] [PubMed]
27. Jajja, A.I.; Abbas, A.; Khattak, H.A.; Niedbała, G.; Khalid, A.; Rauf, H.T.; Kujawa, S. Compact Convolutional Transformer (CCT)-Based Approach for Whitefly Attack Detection in Cotton Crops. *Agriculture* **2022**, *12*, 1529. [CrossRef]
28. Niedbała, G.; Kurek, J.; Świderski, B.; Wojciechowski, T.; Antoniuk, I.; Bobran, K. Prediction of Blueberry (*Vaccinium corymbosum* L.) Yield Based on Artificial Intelligence Methods. *Agriculture* **2022**, *12*, 2089. [CrossRef]
29. Jasim, M.A.; Al-Tuwaijari, J.M. Plant leaf diseases detection and classification using image processing and deep-learning techniques. In Proceedings of the 2020 International Conference on Computer Science and Software Engineering (CSASE), Duhok, Iraq, 16–18 April 2020; IEEE: New York, NY, USA, 2020; pp. 259–265.
30. Swain, S.; Nayak, S.K.; Barik, S.S. A review on plant leaf diseases detection and classification based on machine learning models. *Mukt Shabd* **2020**, *9*, 5195–5205.
31. Ranjan, M.; Weginwar, M.R.; Joshi, N.; Ingole, A. Detection and classification of leaf disease using artificial neural network. *Int. J. Tech. Res. Appl.* **2015**, *3*, 331–333.
32. Bolliger, P.; Ostermaier, B. Koubachi: A mobile phone widget to enable affective communication with indoor plants. In Proceedings of the Mobile Interaction with the Real World (MIRW 2007), Singapore, 9 September 2007; p. 63.
33. Gélard, W.; Herbulot, A.; Devy, M.; Casadebaig, P. 3D leaf tracking for plant growth monitoring. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; IEEE: New York, NY, USA, 2018; pp. 3663–3667.
34. Mishra, P.; Feller, T.; Schmuck, M.; Nicol, A.; Nordon, A. Early detection of drought stress in Arabidopsis thaliana utilsing a portable hyperspectral imaging setup. In Proceedings of the 2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), Amsterdam, The Netherlands, 24–26 September 2019; IEEE: New York, NY, USA, 2019; pp. 1–5.

35. Alexandridis, T.K.; Moshou, D.; Pantazi, X.E.; Tamouridou, A.A.; Kozhukh, D.; Castef, F.; Lagopodi, A.; Zartaloudis, Z.; Mourelatos, S.; de Santos, F.J.N.; et al. Olive trees stress detection using Sentinel-2 images. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; IEEE: New York, NY, USA, 2019; pp. 7220–7223.
36. Ciężkowski, W.; Kleniewska, M.; Chormański, J. Using Landsat 8 Images for The Wetland Water Stress Calculation: Upper Biebrza Case Study. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; IEEE: New York, NY, USA, 2019; pp. 6867–6870.
37. Bhugra, S.; Chaudhury, S.; Lall, B. Use of leaf colour for drought stress analysis in rice. In Proceedings of the 2015 Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), Patna, India, 16–19 December 2015; IEEE: New York, NY, USA, 2015; pp. 1–4.
38. Ahmed, K.; Shahidi, T.R.; Alam, S.M.I.; Momen, S. Rice leaf disease detection using machine learning techniques. In Proceedings of the 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 24–25 December 2019; IEEE: New York, NY, USA, 2019; pp. 1–5.
39. Zhang, K.; Wu, Q.; Liu, A.; Meng, X. Can deep learning identify tomato leaf disease? *Adv. Multimed.* **2018**, *2018*, 6710865. [CrossRef]
40. Selvaraj, M.G.; Vergara, A.; Ruiz, H.; Safari, N.; Elayabalan, S.; Ocimati, W.; Blomme, G. AI-powered banana diseases and pest detection. *Plant Methods* **2019**, *15*, 92. [CrossRef]

agriculture

MDPI

Article

Multi-Node Path Planning of Electric Tractor Based on Improved Whale Optimization Algorithm and Ant Colony Algorithm

Chuandong Liang, Kui Pan, Mi Zhao and Min Lu *

College of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832000, China
* Correspondence: lm_mac@shzu.edu.cn

Abstract: Under the "Double Carbon" background, the development of green agricultural machinery is very fast. An important factor that determines the performance of electric farm machinery is the endurance capacity, which is directly related to the running path of farm machinery. The optimized driving path can reduce the operating loss and extend the mileage of agricultural machinery, then multi-node path planning helps to improve the working efficiency of electric tractors. Ant Colony Optimization (ACO) is often used to solve multi-node path planning problems. However, ACO has some problems, such as poor global search ability, few initial pheromones, poor convergence, and weak optimization ability, which is not conducive to obtaining the optimal path. This paper proposes a multi-node path planning algorithm based on Improved Whale Optimized ACO, named IWOA-ACO. The algorithm first introduces reverse learning strategy, nonlinear convergence factor, and adaptive inertia weight factor to improve the global and local convergence ability. Then, an appropriate evaluation function is designed to evaluate the solving process and obtain the best fitting parameters of ACO. Finally, the optimal objective function, fast convergence, and stable operation requirements are achieved through the best fitting parameters to obtain the global path optimization. The simulation results show that in flat environment, the length and energy consumption of IWOA-ACO planned path are the same as those of PSO-ACO, and are 0.61% less than those of WOA-ACO. In addition, in bump environment, the length and energy consumption of IWOA-ACO planned path are 1.91% and 4.32% less than those of PSO-ACO, and are 1.95% and 1.25% less than those of WOA-ACO. Therefore, it is helpful to improve the operating efficiency along with the endurance of electric tractors, which has practical application value.

Keywords: path planning; ACO; IWOA; electric tractor

Citation: Liang, C.; Pan, K.; Zhao, M.; Lu, M. Multi-Node Path Planning of Electric Tractor Based on Improved Whale Optimization Algorithm and Ant Colony Algorithm. *Agriculture* **2023**, *13*, 586. https://doi.org/10.3390/agriculture13030586

Academic Editors: Gniewko Niedbała, Sebastian Kujawa and Jin He

Received: 31 December 2022
Revised: 23 February 2023
Accepted: 24 February 2023
Published: 28 February 2023

1. Introduction

Based on the needs of large agricultural bases in Xinjiang and the development of facility agriculture, agricultural machinery has been widely used. As a new type of agricultural machinery, the electric tractor has the advantages of low pollution, low noise, high efficiency, easy operation, etc. [1]. In recent years, it has been gradually applied to agricultural production [2,3]. From the reduction in loss and the improvement in endurance, we need to plan the optimal path to improve efficiency of electric tractors.

Multi-node path planning refers to the path after traversing all nodes in a certain area, starting from a node, with the set of optimal operation rules. This is an increasingly important area in automated production includes plant protection and farmland leveling [4,5], as well as tractor path planning [6]. Ref. [7] proposed a path planning and tracking control method, which was helpful for vehicle collision-free driving. Ref. [8] proposed a tractor path planning method applied in complex environment, which improved the tractor operation efficiency and coverage. In recent years, intelligent algorithm is a major area of interest within the field of multi-node path planning [9–11], such as ACO algorithm, genetic algorithm (GA), Particle Swarm Optimization (PSO), etc. Ref. [12] proposed an

improved ant colony algorithm based on the adaptive volatile coefficient for the traveling salesman problems. Ref. [13] proposed a parallel ant colony algorithm for multi-node path planning of facility greenhouse robots. Ref. [14] proposed ant colony optimization algorithm variants to increase the probability of the algorithm to find the object. Ref. [15] proposed an improved ant colony optimization algorithm to solve the traveling salesman problem. Ref. [16] proposed an improved particle swarm optimization for multi-node path planning. Ref. [17] proposed GA for the traveling salesman problems. As the bionic ant colony foraging behavior adopted by ACO is similar to path planning, ACO is a commonly used algorithm to solve the multi-node path planning problem. When applying ACO to solving specific problems, it is possible to set the iteration rules of the algorithm parameters in a targeted way. However, ACO requires many parameters, and it is difficult to determine the optimal fitting combination of parameters. Therefore, it is easy for the algorithm to fall into local optimal solution, which is not conducive to obtaining the optimal path.

The Whale Optimization Algorithm (WOA) is an intelligent algorithm proposed by Australian scholars [18] which simulates the foraging behavior of whales to solve the objective function. WOA has the advantages of small number of setting parameters and strong convergence performance; however, it easily falls into local optimum and has low convergence accuracy. In recent years, a number of researchers have sought to improve WOA. References proposed improved whale optimization algorithms based on elite backward learning [19], the crossover and mutation operations [20], nonlinear convergence factor [21–23], and adaptive weighting factor [22,23], respectively, which balance between global and local convergence capability and enhance the diversity of the initial solution. Therefore, WOA has the possibility to further optimize population initialization and iteration rules to improve algorithm performance. However, a major problem with those is the complexity of optimization algorithm logic, which reduces the operational efficiency of the algorithm.

The main contributions of this paper are as follows:

1. This paper proposes a fusion improved Whale Optimization Algorithm and Ant Colony algorithm, named IWOA-ACO, to plan the multi-node path of the electric tractor by optimizing the parameters of ACO. At the same time, IWOA introduces reverse learning strategy, nonlinear convergence factor and adaptive inertia weighting factor to balance between global and local convergence capability of it and enhance the diversity of the initial solution. Then, IWOA-ACO improves the evaluation function to ensure accurate evaluation of ACO performance during iteration. The block diagram of IWOA-ACO is shown in Figure 1.

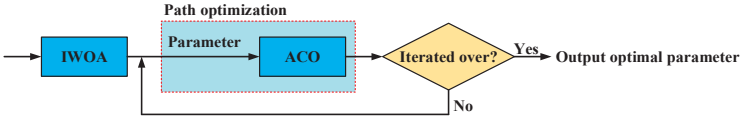


Figure 1. The block diagram of IWOA-ACO.

2. The operation node data of cultivated land environment in Xinjiang is taken and as an example, the experimental results show that IWOA-ACO algorithm can optimize ACO setting parameters to plan the optimal path of electric tractor, which is better than WOA-ACO and PSO-ACO in path length and energy consumption. Moreover, considering the flat terrain of Xinjiang, this paper constructs a bump environment model and conducts simulation experiments to reflect the universality of IWOA-ACO algorithm. In addition, the experimental results show that the length and energy consumption of the planned path of the algorithm is still better than those of WOA-ACO and PSO-ACO, reflecting the good value of the algorithm.

In a word, we propose an IWOA-ACO algorithm to plan a path with better path length and energy consumption. Based on the experimental analysis of complex nodes

and experimental fields in Xinjiang, the path length and energy consumption of electric tractors planned by IWOA-ACO are better than those of WOA-ACO and PSO-ACO, thus improving the operating efficiency and endurance level of electric tractors.

2. Methods

2.1. Analysis of ACO

This section briefly introduces the operation principle of ACO algorithm and the impact of the main parameters on the algorithm performance. Aiming at the iteration redundancy problem that may be caused by setting an excessively high maximum iteration number of the algorithm, the Iteration Early Termination Strategy (IETS) is introduced to enhance the operational efficiency of the algorithm.

2.1.1. The Introduction of ACO

Ant foraging is a group behavior. In the process of foraging, ants will release volatile pheromones and perceive the residual pheromone concentration released by after other ones. Ants use roulette strategy when choosing the path, and will prioritize the path with short path and high pheromone concentration, which constitutes a positive feedback mechanism. After a period of time, the ant colony will choose an optimal path. The state transition function is shown by

$$P_{ij}^{t_{\mathrm{ACO}}} = \begin{cases} \dfrac{Ta_{ij}(t_{\mathrm{ACO}})^{\alpha} Et_{ij}(t_{\mathrm{ACO}})^{\beta}}{\sum\limits_{t_{\mathrm{ACO}} \in allow(t_{\mathrm{ACO}},i)} Ta_{ij}(t_{\mathrm{ACO}})^{\alpha} Et_{ij}(t_{\mathrm{ACO}})^{\beta}}, j \in allow(t_{\mathrm{ACO}}, i) \\ 0, j \notin allow(t_{\mathrm{ACO}}, i) \end{cases} \tag{1}$$

where t_{ACO} is the current iteration, $P_{ij}^{t_{\mathrm{ACO}}}$ is the state transition function, β is the heuristic function importance factor, α is the pheromone importance factor, i and j are the adjacent nodes, and $allow(t_{\mathrm{ACO}},i)$ is the node that has not been accessed. $Ta_{ij}(t_{\mathrm{ACO}})$ is pheromone concentration, as presented in Equation (2). $Et_{ij}(t_{\mathrm{ACO}})$ is heuristic function, as presented in Equation (3).

$$Ta_{ij}(t_{\mathrm{ACO}}) = (1 - rh)Ta_{ij}(t_{\mathrm{ACO}} - 1) + \Delta Ta_{ij}(t_{\mathrm{ACO}}) \tag{2}$$

where rh is the pheromone concentration volatilization factor, and $\Delta Ta_{ij}(t_{\mathrm{ACO}})$ is the new increment of pheromone on the path, which is related to whether ants pass through the path.

$$Et_{ij}(t_{\mathrm{ACO}}) = \frac{1}{Dis_{ij}} \tag{3}$$

where Dis_{ij} is the European distance through the path.

When researchers apply the ant colony algorithm to specific problems, the main parameters to be set are α, β, rh and the ant quantity, such as m. This paper analyzes the impact of four parameters on the performance of ACO algorithm as follows [24]:

1. α and β influence $P_{ij}^{t_{\mathrm{ACO}}}$ together. α reflects the importance of the ant colony to the existing pheromone when searching the path, and β reflects the degree to which the ant colony pays attention to the local shortest path when searching the path. When the setting values of α and β are large, the local convergence ability of the algorithm is strong. On the contrary, the algorithm has strong global convergence.
2. rh affects the pheromone concentration level, thus affecting $P_{ij}^{t_{\mathrm{ACO}}}$. When the setting value of rh is large, the pheromone concentration on the path is low, and the positive feedback effect is weakened, so the algorithm has strong global convergence ability, but the convergence speed of the algorithm is slow. On the contrary, the pheromone concentration on the path is high, and the positive feedback effect is enhanced, so the algorithm converges quickly, but it easily falls into the local optimal solution.
3. m affects the convergence ability and running speed of the algorithm. When the setting value of m is large, the global convergence of the algorithm is good, but the

running speed is slow. On the contrary, the running speed of the algorithm is fast, but the convergence performance is poor, and it easily falls into the local optimal solution.

2.1.2. The Introduction of IETS

In ACO algorithm, when the set number of iteration terminations is larger than the solution problem, multiple iterations will have the same value at the end of the iteration, reducing the efficiency of the algorithm. This paper introduces IETS: when the iterative solution reaches the set value t_{set} with the same number of consecutive times t_0 and the iterative solution is less than the set value x_{set}, the iteration is terminated. Otherwise, the iteration will continue until the maximum iteration period t_{max} is met. The flowchart of IETS is shown in Figure 2.

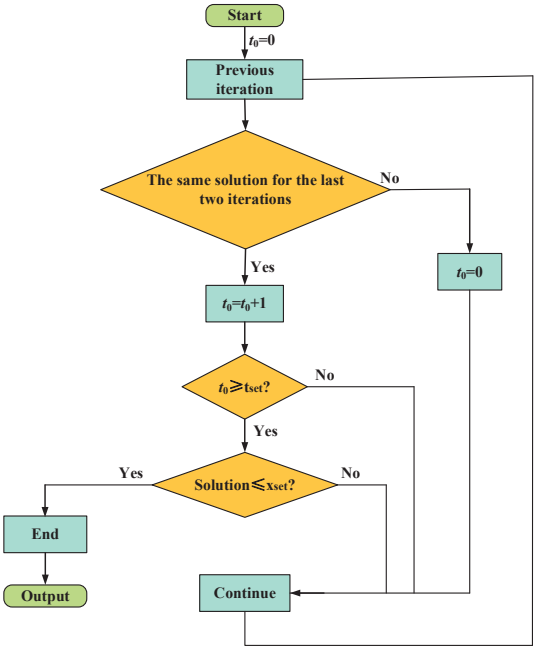


Figure 2. The flowchart of IETS.

2.2. Analysis of IWOA

This section introduces the operating principle of IWOA. Test functions are used to verify the operation performance of IWOA.

2.2.1. The Introduction of WOA

The operation logic of WOA originates from the foraging behavior of whales, including three search mechanisms: local optimization consisting of encirclement predation mechanism, spiral mechanism, and global optimization consisting of random search mechanism.

1. Encirclement predation mechanism

The encirclement predation mechanism originates from the behavior of whales to identify and encircle their prey. The location of whales closest to their prey can be regarded as a local optimal solution, and other whales converge to the local optimal solution, as presented in Equations (4) and (5).

$$D = |CX^*(t) - X(t)| \quad (4)$$

$$X(t+1) = X^*(t) - AD \tag{5}$$

where X is the current whale position, X^* is the best whale location. A and C are coefficient vectors, as presented in Equations (6) and (7).

$$A = 2ar - a = a(2r - 1) \tag{6}$$

$$C = 2r \tag{7}$$

where r is a random variable belonging to [0, 1], a is the linearly decreasing convergence factor from 2 to 0.

2. Spiral mechanism

The spiral mechanism is based on the fact that whales emit bubbles and move to their prey in a spiral motion in the process of foraging, as presented in Equation (8).

$$X(t+1) = De^{bl}\cos(2\pi l) + X^*(t) \tag{8}$$

where b is a constant, l is a random variable belonging to [0, 1].

3. Random search mechanism

The random search mechanism is based on the fact that whales not only approach the nearest whale location from their prey, but also randomly select other whale locations in the population to search, as presented in Equations (9) and (10).

$$D^* = |CX_{rand}(t) - X(t)| \tag{9}$$

$$X(t+1) = X_{rand}(t) - AD^* \tag{10}$$

where X_{rand} is the position of whales randomly selected in the population.

The operation logic of the three search mechanisms is as follows:

```
p = rand(0,1)
if p < 0.5
    if |A| < 1
        Encirclement predation, (5)
    else
        Random search, (10)
    end
else
    Spiral, (8)
end
```

2.2.2. Analysis of IWOA

In order to improve the convergence performance of WOA, this paper optimizes WOA in three aspects to obtain IWOA.

1. Population initialization with reverse learning strategy (PIRL)

The quality of the initial population affects the convergence speed and accuracy of the algorithm, and high-quality initial population is conducive to the rapid convergence of the algorithm. In WOA, the initialization of the population is completely random, which has the advantage of ensuring the diversity of the initial population, but the quality of the initial population generated by this scheme is not high. To ensure the diversity of the initial population and improve the quality of the initial population, PIRL is introduced in this paper. The strategy steps are as follows:

- Establish a random initial population, and analyze the position of individuals in the random initial population in turn.
- Set a random variable p belonging to [0, 1]. Adopt reverse learning strategy if $p \geq 0.3$. Assume that the position of individual r in the d-dimensional space is $X_r(1, 2, \ldots, d)$,

then the corresponding reverse individual is $\widetilde{X}_r(1,2,\ldots,d)$, as presented in Equation (11).

$$\widetilde{X}_r(k) = L(k) + U(k) - X_r(k) \tag{11}$$

where $L(k)$ and $U(k)$ are the boundaries of population space. The fitness values of individuals $X_r(1,2,\ldots,d)$ and $\widetilde{X}_r(1,2,\ldots,d)$ are calculated, respectively, and the individual with better fitness is retained as the final initial population.

- Do not adopt reverse learning strategy if $p < 0.3$. The individual $X_r(1,2,\ldots,d)$ is retained as the final initial population.

2. Nonlinear convergence factor.

After analyzing the operation logic of the three search mechanisms, the paper concludes that the size of $|A|$ determines the global and local search of the algorithm. According to Equation (6), the size of convergence factor a determines the size of $|A|$. In WOA, convergence factor a is linearly decreasing. Therefore, when the number of iterations is greater than half of the maximum number of iterations, $a < 1$. In the middle of iteration, $|A|$ drops to a low value too early, so that WOA changes from global search to local search too early, increasing the possibility of the algorithm falling into the local optimal solution.

In order to better balance the ability of global search and local search of the algorithm and make the algorithm turn to local search after full global search, exponential nonlinear convergence factor a^* is introduced in this paper, as presented in Equation (12).

$$a^* = \frac{-e^{\frac{t}{T_{max}}\ln 51} + 51}{25} \tag{12}$$

where a^* is the nonlinear convergence factor, T_{max} is the maximum number of iterations.

Assuming T_{max} = 50, the curve of a^* is drawn as shown in Figure 3.

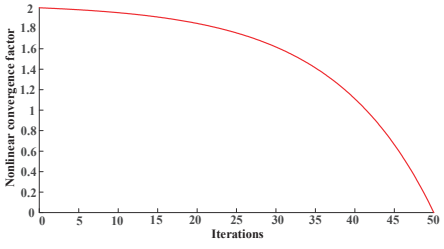


Figure 3. The curve of nonlinear convergence factor.

In the early and middle stages of the iteration process, the value of a^* is kept at a high level to ensure the global search capability of the algorithm, and it decreases rapidly in the late iteration period to ensure that the algorithm focuses on local search.

3. Adaptive inertia weighting factor

The idea of inertia weight factor is derived from PSO [25]. Individuals consider the impact of the current position when they move to the local optimal solution. In WOA, the value of inertia weight factor is always 1. In the early of iteration process, a large inertia weight factor is helpful to the global search of WOA. However, in the late iteration period, it causes the individual to pay too much attention to the current optimal solution position, so the algorithm easily falls into local optimization.

Therefore, this paper introduces adaptive inertial weighting factor w, as presented in Equation (13). In the early of iteration process, the global search ability of IWOA can be enhanced by the large value of w. In the late iteration period, the smaller value of w can

reduce the dependence of individuals on the current optimal solution and avoid IWOA falling into the local optimal solution.

$$w = 0.4\left(\frac{t}{\mathrm{T_{max}}}\right)^2 - 0.8\frac{t}{\mathrm{T_{max}}} + 0.9 \tag{13}$$

In IWOA, Equations (5) and (8) should be changed to Equations (14) and (15).

$$X(t+1) = wX^*(t) - AD \tag{14}$$

$$X(t+1) = De^{bl}\cos(2\pi l) + wX^*(t) \tag{15}$$

2.2.3. Operation Steps of IWOA

IWOA is optimized in terms of population initialization and iteration parameters. The operation steps of IWOA are as follows:

1. Set algorithm parameters, such as variable dimension, number of population individuals, and maximum number of iterations.
2. Randomly initialize the population within the range of variable values.
3. Establish initial population by PIRL and record the individual position with better fitness as the optimal position.
4. The algorithm updates the location of individuals based on different search mechanisms. Set a random variable as p = rand (0,1). If $p < 0.5$ and $|A| < 1$, the algorithm individually updates their positions according to the encirclement predation mechanism, as in Equation (14). If $p < 0.5$ and $|A| \geq 1$, individually update position according to the random search mechanism, as presented in Equation (10). If $p \geq 0.5$, individually update position according to the spiral mechanism, as in Equation (15).
5. The algorithm restricts the range of the updated position of the individuals, calculates the fitness values of those, and updates the optimal position.
6. The algorithm judges whether the maximum number of iterations is reached. If so, it exits the iteration and outputs the optimal location and fitness value. If not, it returns to step (4) to continue iteration.

2.2.4. Performance Testing for IWOA

This paper selects four test functions to verify the performance of IWOA [26], as shown in Table 1. Then, this paper sets the population number as 40 and the maximum number of iterations as 100, and compares the performance of the three algorithms, including IWOA, WOA and PSO.

Table 1. Introduction to test functions.

Name	Expression	Dimension	Domain of Definition	Theoretical Optimal Value
Sphere	$F = \sum_{i=1}^{n} x_i^2$	30	[−100, 100]	0
Quartic	$F = rand(0,1) + \sum_{i=1}^{n} ix_i^4$	30	[−1.28, 1.28]	0
Schwefel 2.26	$F = \sum_{i=1}^{n} -x_i \sin\sqrt{\lvert x_i \rvert}$	30	[−500, 500]	−12,569
Rastrigin	$F = \sum_{i=1}^{n} \left[x_i^2 - 10\cos(2\pi x_i) + 10\right]$	30	[−5.12, 5.12]	0

Sphere and Quartic are unimodal functions used to test the local search ability of those algorithms, and Schwefel 2.26 and Rastigin are multimodal functions used to test the global search ability of those algorithms. And the convergence curves of test functions are shown in Figure 4.

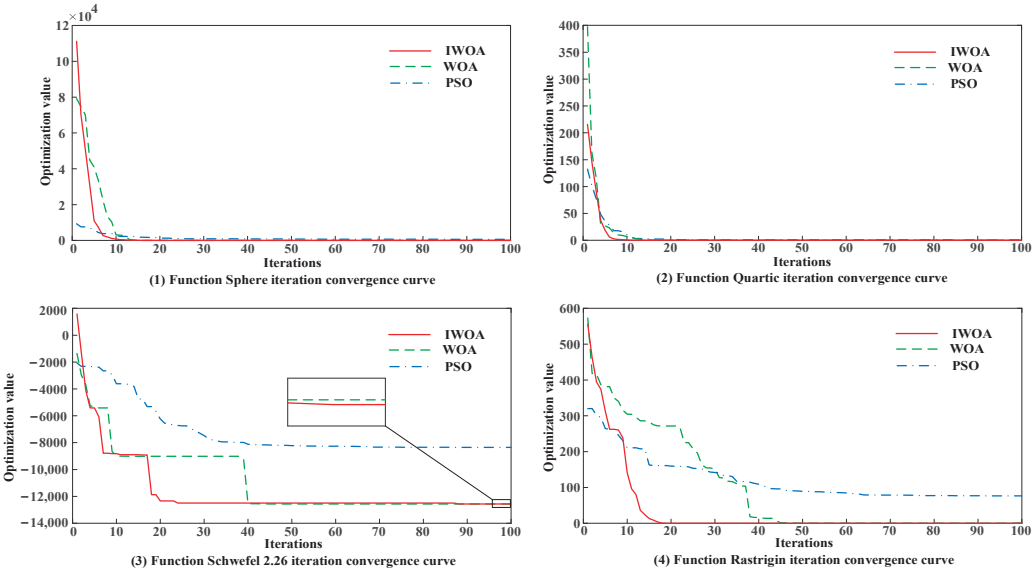


Figure 4. Convergence curves of test functions.

IWOA has converged to the optimal value in the 10th to 20th iterations, so IWOA is superior to WOA and PSO in convergence speed.

In order to better compare and analyze the performance of IWOA, this paper uses those algorithms to run each test function 30 times, and records the optimal values, the worst values and the standard deviation of the optimization results. The analysis of optimization results is shown in Table 2.

Table 2. The analysis of optimization results.

Name	Statistic	IWOA	WOA	PSO	Optimal Algorithm
Sphere	Optimal value	6.66×10^{-40}	4.83×10^{-17}	478.21	IWOA
	Worst value	7.01×10^{-29}	5.89×10^{-11}	2296.15	
	Standard deviation	1.30×10^{-29}	1.06×10^{-11}	407.98	
Quartic	Optimal value	8.41×10^{-6}	7.01×10^{-4}	0.29	IWOA
	Worst value	2.22×10^{-3}	5.81×10^{-2}	1.85	
	Standard deviation	5.02×10^{-4}	1.15×10^{-2}	0.45	
Schwefel 2.26	Optimal value	−12,569.24	−12,318.02	−8677.97	IWOA
	Worst value	−10,600.90	−7422.42	−5034.17	
	Standard deviation	453.58	1378.79	908.84	
Rastrigin	Optimal value	0	0	73.14	IWOA
	Worst value	0	1.30×10^{-6}	182.02	
	Standard deviation	0	2.33×10^{-7}	23.79	

The optimal and the worst values of the function obtained by IWOA convergence are better than those of WOA and PSO, and the standard deviation of the data obtained by IWOA in 30 groups of tests is lower, so the performance of IWOA is more stable. In general, IWOA performs better than WOA and PSO in unimodal and multimodal test functions. Therefore, IWOA performs better in local search and global search, verifying the optimization effect.

2.3. Multi-Node Path Planning Problem of Electric Tractor

2.3.1. Analyze the Application Scenario

Multi-node path planning problem refers to the optimal path planning for multiple nodes in a certain area. The legal path is the path that starts from a node and traverses all other nodes. In agricultural production, the nodes are the marshal point of the harvested crops, and the electric tractor needs to traverse all the nodes in the region to collect all the harvested crops. In this paper, we employ a cultivated area, including node data and operating parameters of electric tractor, as the experimental object to verify the effectiveness of IWOA-ACO.

In order to exclude the influence of irrelevant factors on the experiment, we make the following assumptions:

1. The path between nodes is a segment.
2. Neglect the turning action of electric tractor at nodes.
3. Neglect the air resistance of the electric tractor.
4. Focus on the path length and energy consumption of electric tractors between nodes.

We choose the cultivated area located near 87.4 E and 44.3 N. The positions of 26 nodes in space are shown in Figure 5.

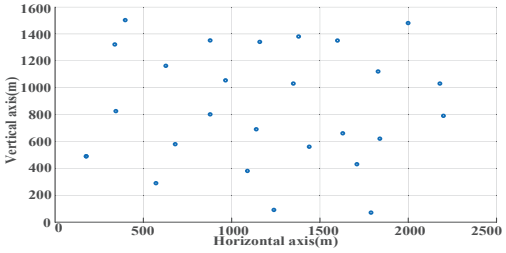


Figure 5. Spatial location map of 26 nodes.

2.3.2. Kinematics Model of Electric Tractor

When the electric tractor travels between different operating points, it often travels at a low speed and at a constant speed. The energy loss in the operation of the electric tractor is mainly the energy consumed during travel. The force analysis of the electric tractor under different road conditions (flat ground and slope) [27] is shown in Figure 6.

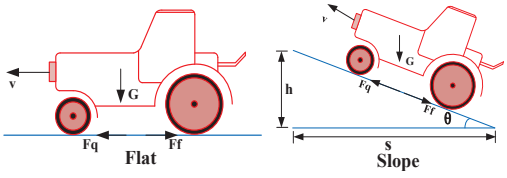


Figure 6. Force analysis of electric tractor.

When the electric tractor runs on the slope, the slope angle $\theta_{i,j}$ is shown in Equation (16). The path length of electric tractor traveling between nodes i and j is shown in Equation (17).

$$\theta_{i,j} = \arctan\left(\frac{h_{i,j}}{s_{i,j}}\right) \tag{16}$$

$$\gamma_{i,j} = \sqrt{{s_{i,j}}^2 + {h_{i,j}}^2} \tag{17}$$

where $\gamma_{i,j}$ is the path length between nodes, $s_{i,j}$ is the horizontal distance between nodes, $h_{i,j}$ is the height difference between nodes.

The mechanical expression of the electric tractor in uniform running is shown in Equation (18).

$$F_{\mathrm{q}} = \mu \mathrm{G} \cos(\theta_{i,j}) + \mathrm{G} \sin(\theta_{i,j}) \tag{18}$$

where F_{q} is the driving force of power motor of electric tractor, μ is the friction coefficient of traveling road surface, G is the gravity of electric tractor.

The energy consumed by the motor during operation is shown in Equation (19).

$$Q_{i,j} = U q_{i,j} \tag{19}$$

where U is the input voltage of motor, $q_{i,j}$ is the power consumed by the motor.

The driving force of the electric tractor during traveling is shown in Equation (20).

$$F_{\mathrm{q}} = \frac{T_{\mathrm{x}} i_{\mathrm{g}} i_0 \eta}{\mathrm{R}} = \frac{T_{\mathrm{t}} i_{\mathrm{g}} i_0 \eta}{\mathrm{R}} \tag{20}$$

where T_{x} is the output torque of planet carrier, i_{g} is the transmission speed ratio of electric tractor, i_0 is the differential speed ratio, η is the transmission efficiency; R is the wheel radius of electric tractor, T_{t} is the torque of motor, as shown in Equation (21).

$$T_{\mathrm{t}} = \frac{9550P}{n} \tag{21}$$

where n is the motor speed, P is the output power of motor.

The relationship between the traveling speed of the electric tractor v and the motor speed is shown in Equation (22).

$$v = \frac{0.377 \mathrm{R} n}{i_{\mathrm{g}} i_0} \tag{22}$$

From Equations (16)–(22), the energy consumed by electric tractor $Q_{i,j}$ when traveling between nodes i and j is shown in Equation (23).

$$Q_{i,j} = \frac{(\mu \mathrm{G} \cos(\theta_{i,j}) + \mathrm{G} \sin(\theta_{i,j})) s_{i,j}}{3.6 \eta \cos(\theta_{i,j})} \tag{23}$$

Regarding balance, from Equations (17) and (23), the kinematic function model of electric tractor is shown in Equation (24).

$$F_{\mathrm{kin}} = f(\gamma_{i,j}, Q_{i,j}) \tag{24}$$

2.4. IWOA-ACO

The basic idea of IWOA-ACO is to use ACO to solve the optimal path of multiple nodes, and then use IWOA to optimize the operation parameters of ACO. There are two key problems when fusing IWOA and ACO algorithm. On the one hand, IWOA-ACO needs to set appropriate evaluation function to evaluate the solution process and results of ACO algorithm. On the other hand, IWOA-ACO needs to put ACO algorithm into IWOA iteration to solve the multi-node problem, and in IWOA iteration, input the four setting parameters mentioned in Section 2.1 to ACO algorithm.

An appropriate evaluation function is crucial for IWOA-ACO to optimize the parameters of ACO. The evaluation function should reflect the optimality, fast convergence and algorithm stability of the objective function of ACO for solving multi-node path planning. The evaluation function is shown in Equation (25).

$$y = k_1 f_1 + k_2 f_2 + k_3 f_3 + k_4 f_4 \tag{25}$$

where f_1 is the difference between the path length value obtained by ACO and the empirical optimal value of path length, as shown in Equation (26). f_2 is the difference between the energy consumption value obtained by ACO and the empirical optimal value of energy

consumption, as shown in Equation (27). f_1 and f_2 represent the optimality of the algorithm for solving the objective function. f_3 is the iteration number of ACO, representing the fast convergence of the algorithm, as shown in Equation (28). f_4 is the standard deviation of the iterative data of ACO, representing the stability of the iterative data of the algorithm, as shown in Equation (29). k_1, k_2, k_3 and k_4 are the weight coefficients.

$$f_1 = length - length\_\min \tag{26}$$

$$f_2 = energy - energy\_\min \tag{27}$$

$$f_3 = \sum length \tag{28}$$

$$f_4 = \delta(length) \tag{29}$$

The flow chart of IWOA-ACO is shown in Figure 7.

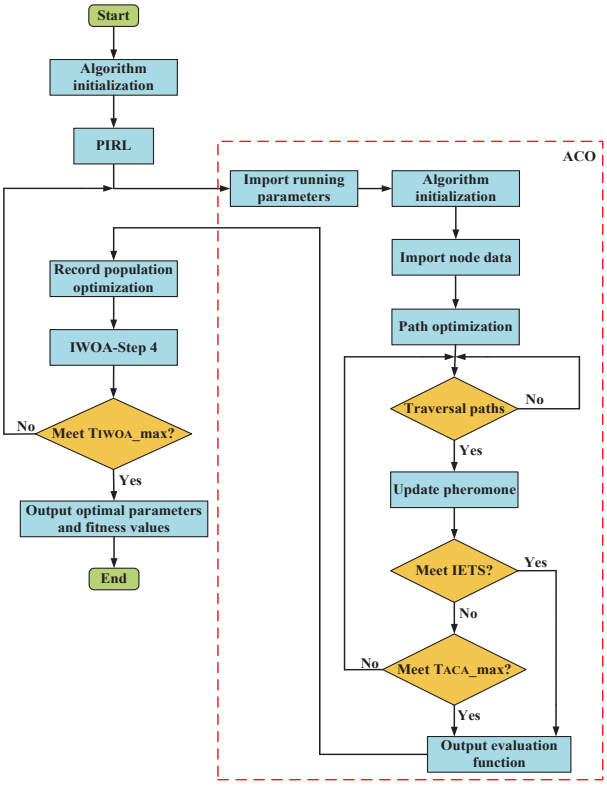


Figure 7. The flow chart of IWOA-ACO.

3. Results and Discussion

This section applies IWOA-ACO to solve the problem and analyzes the simulation results of flat and bump environment by the three algorithms so as to verify the performance of IWOA-ACO.

3.1. Simulation of Flat Environment

First, according to the analysis in Section 2.3.1, we need to extract the node data from the test field, and the experiment needs a planned optimal path for the electric tractor to traverse all nodes. Second, we need to set the relevant parameters of the electric tractor as follows: the friction coefficient of traveling road surface is 0.07, the gravity of electric tractor

is 10,700, and the transmission efficiency is 0.85. Then, we need to unify the evaluation functions of the three algorithms into Equation (25), and unify the setting parameters as follows: the variable dimension is 4, the number of individuals in the population is 30, the maximum number of iterations is 50. Finally, in order to eliminate the impact of algorithm simulation environment on algorithm performance, we unify the simulation environment of the three algorithms as follows: Windows10 (64 bit), Core (TM) i7-8550U, CPU 1.80 GHz, 16 GB, MatlabR2017a.

In order to better analyze the operational performance of IWOA-ACO, this paper solves the node path planning problem as shown in Figure 5 with IWOA-ACO, WOA-ACO and PSO-ACO, compares the iteration curves of the evaluation functions of the three algorithms, and records the operation parameters of ACO algorithm, respectively.

The convergence curve of the evaluation function corresponding to the three algorithms is shown in Figure 8. The parameter values of ACO obtained by convergence of three functions are shown in Table 3. The convergence value of the evaluation function of IWOA-ACO is better than that of WOA-ACO and PSO-ACO.

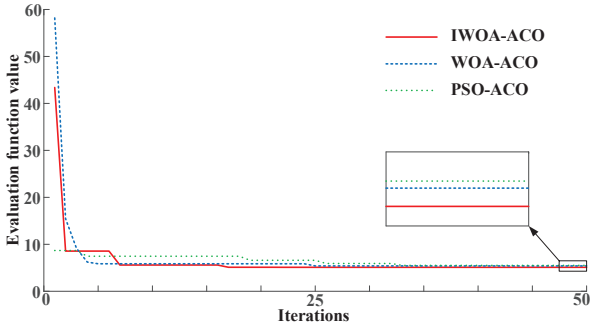


Figure 8. The convergence curve of the evaluation function in the simulation of flat environment.

Table 3. The parameter values of ACO in the simulation of flat environment.

Algorithm	m	rh	β	α
PSO-ACO	80	0.80	4.39	0.50
WOA-ACO	80	0.80	3.39	2.00
IWOA-ACO	80	0.80	5.00	0.50

Ref. [24] referred to the scheme of determining the parameter values of standard ACO algorithm by empirical method. In order to better verify the convergence performance of the IWOA-ACO algorithm, we employ the optimization scheme of the standard ACO algorithm as the control group in the comparison simulation. The parameter values of the algorithm are set according to Ref. [24]: $m = 50$, $\alpha = 1$, $\beta = 7$ and $rh = 0.3$.

ACO plans the 26 node paths of the electric tractor according to the operation parameters in Table 3 and experience parameters. The generated path planning diagram is shown in Figure 9. This paper uses the iteration path length convergence curve to compare the convergence performance of ACO under different parameters, as shown in Figure 10.

3.2. Simulation of Bump Environment

The cultivated land environment in Xinjiang is characterized by flat terrain [28,29]. Therefore, the node data in the cultivated area selected in this paper is approximately 2D. In order to further explore the adaptability of IWOA-ACO in the diversified cultivated land environment, we introduce the data of 31 nodes in the bump environment. The experimental steps and the parameters are set as shown in Section 3.1, and this paper compares the simulation results of the three algorithms. The positions of 31 nodes in space are shown in Figure 11.

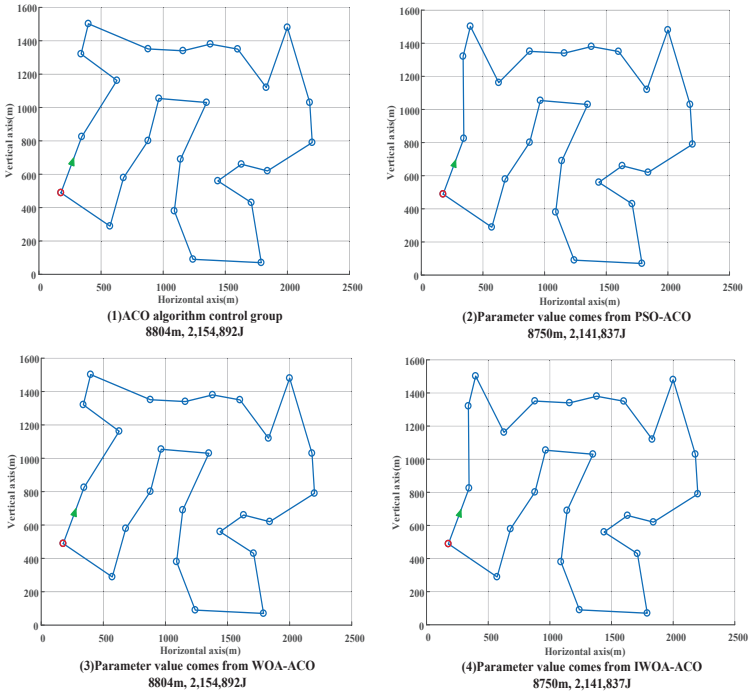


Figure 9. Path planning diagram in the simulation of flat environment. The red node is the starting point, and the green arrow is the travel direction of the electric tractor.

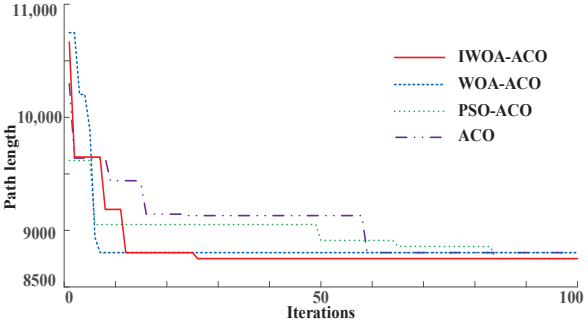


Figure 10. The iterative path length convergence curve in the simulation of flat environment.

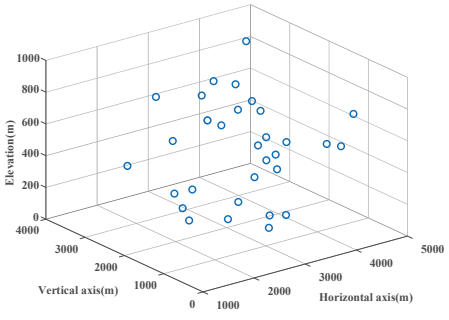


Figure 11. Spatial location map of 31 nodes.

The convergence curve of the evaluation function corresponding to the three algorithms is shown in Figure 12. The parameter values of ACO obtained by convergence of three functions are shown in Table 4. The convergence value of the evaluation function of IWOA-ACO is better than that of WOA-ACO and PSO-ACO. Moreover, the evaluation function value of IWOA-ACO can converge to the optimal value after five iterations, while WOA-ACO requires 20 iterations and PSO-ACO requires 15 iterations. It can be seen that the convergence performance of IWOA-ACO is better than that of WOA-ACO and PSO-ACO in solving the multi-node path planning problem.

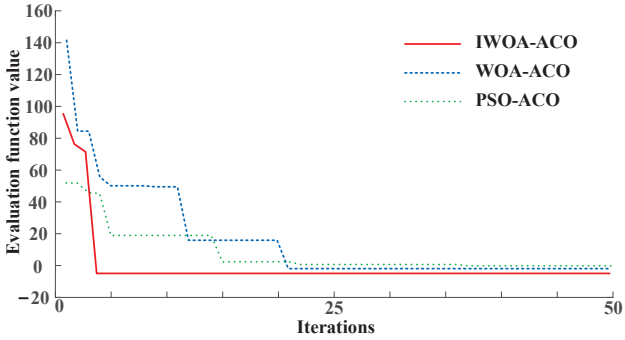


Figure 12. The convergence curve of the evaluation function in the simulation of bump environment.

Table 4. The parameter values of ACO in the simulation of bump environment.

Algorithm	m	rh	β	α
PSO-ACO	47	0.2	4.69	0.55
WOA-ACO	76	0.32	3.68	2
IWOA-ACO	72	0.78	2	0.83

As in Section 3.1, in order to better verify the convergence performance of the IWOA-ACO algorithm, we employ the optimization scheme of the standard ACO algorithm as the control group in the comparison simulation. The parameter values of the algorithm are $m = 50$, $\alpha = 1$, $\beta = 7$ and $rh = 0.3$.

ACO plans the 31-node path of electric tractor based on the operating parameters obtained from the above three algorithms and experience parameters, and the resulting path planning diagram is shown in Figure 13. This paper uses the iterative path length convergence curve to compare the convergence performance of ACO under different parameters, as shown in Figure 14.

3.3. Discussion for Flat Environment Results

Based on Figures 9 and 10, this paper analyzes the impact of ACO algorithm control group and the three parameter combinations shown in Table 3 on the performance of ACO as follows:

On the one hand, as far as the convergence speed of ACO algorithm is concerned, IWOA-ACO is equivalent to WOA-ACO and faster than PSO-ACO.

On the other hand, the path length planned by PSO-ACO is 8750 (m), and the energy consumed by electric tractor is 2,141,837 (J). The length and energy consumption of IWOA-ACO planned path are the same as those of PSO-ACO, and are 0.61% less than those of WOA-ACO and ACO algorithm control group (the value of the path length is 8804 m, and the value of the energy is 2,154,892 J). In the simulation of flat environment, the path length and energy consumption data of electric tractor are shown in Table 5. It is worth mentioning that since the nodes are approximately distributed in 2D space, the energy consumed by the electric tractor is proportional to the path length.

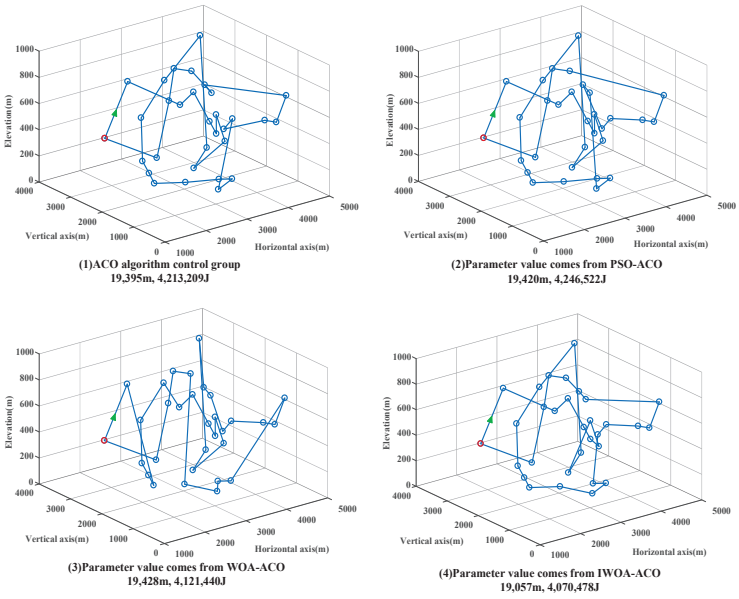


Figure 13. Path planning diagram in the simulation of bump environment. The red node is the starting point, and the green arrow is the travel direction of the electric tractor.

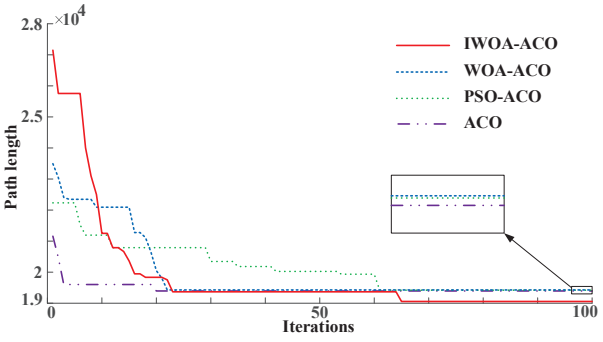


Figure 14. The iterative path length convergence curve in the simulation of bump environment.

Table 5. The path length and energy consumption data in the simulation of flat environment.

Algorithm	ACO Control Group	PSO-ACO	WOA-ACO	IWOA-ACO
The path length (m)	8804 0%	8750 −0.61%	8804 0%	8750 −0.61%
The energy (J)	2,154,892 0%	2,141,837 −0.61%	2,154,892 0%	2,141,837 −0.61%

In general, the path planned by IWOA-ACO for the electric tractor has the advantages of fast convergence speed of WOA-ACO and strong convergence ability of PSO-ACO, which is helpful for efficient operation of the electric tractor.

3.4. Discussion for Bump Environment Results

In the simulation of bump environment, the path length and energy consumption data of electric tractor are shown in Table 6.

Table 6. The path length and energy consumption data in the simulation of bump environment.

Algorithm	ACO Control Group	PSO-ACO	WOA-ACO	IWOA-ACO
The path length (m)	19,395 0%	19,420 +0.13%	19,428 +0.17%	19,057 −1.74%
The energy (J)	4,213,209 0%	4,246,522 +0.79%	4,121,440 −2.18%	4,070,478 −3.39%

Based on Figures 13 and 14, this paper analyzes the impact of the three parameter combinations shown in Table 4 on the performance of ACO as follows:

First of all, ACO, according to the parameters obtained from PSO-ACO, requires approximately 60 iterations to converge to the optimal value. The path length planned by the algorithm is 19,420 (m), and the energy consumed by electric tractor is 4,246,522 (J). The reason is that a small quantity of ants (the value is 47) leads to the slow convergence speed of the algorithm, and a small number of the pheromone concentration volatilization factor (the value is 0.2) leads to the local optimal solution of the algorithm.

In the second place, ACO, according to the parameters obtained from WOA-ACO, requires approximately 20 iterations to converge to the optimal value. The path length planned by the algorithm is 19,428 (m), and the energy consumed by electric tractor is 4,121,440 (J). The reason is that a large quantity of ants (the value is 76) leads to the fast convergence speed of the algorithm, but a small number of the pheromone concentration volatilization factor (the value is 0.32) leads to the local optimal solution of the algorithm.

Once more, ACO, according to the parameters obtained from IWOA-ACO, requires approximately 20 iterations to converge to the optimal value. The algorithm converges faster than the ACO algorithm with the parameters obtained from PSO-ACO, and approximates to the ACO algorithm with the parameters obtained from WOA-ACO. The path length planned by the algorithm is 19,057 (m), which is 1.91% less than that planned by the ACO algorithm with the parameters obtained from PSO-ACO and 1.95% less than that planned by the ACO algorithm with the parameters obtained from WOA-ACO. The energy consumed by electric tractor is 4,070,478 (J), which is 4.32% less than that optimized by the ACO algorithm with the parameters obtained from PSO-ACO and 1.25% less than that optimized by the ACO algorithm with the parameters obtained from WOA-ACO. In addition, the length and energy consumption of IWOA-ACO planned path are 1.74% and 3.39% less than those of ACO algorithm control group.

The reasons for the above results are as follows. On the one hand, a large quantity of ants (the value is 72) leads to the fast convergence speed of the algorithm and a large value of pheromone concentration volatilization factor (the value is 0.78) leads to good global convergence of the algorithm. On the other hand, the difference between the pheromone importance factor (the value is 0.83) and the heuristic function importance factor (the value is 2) is small, so that the algorithm can fully consider the pheromone concentration and heuristic function in the iterative process. Therefore, the algorithm can balance global and local searches.

However, IWOA-ACO has some limitations in practical application. On the one hand, IWOA-ACO can only obtain a set of set value parameters of ACO algorithm with good matching, but the ideal ACO parameter should be an adaptive function. On the other hand, affected by the fluctuation of ACO convergence results, the reliability of IWOA-ACO evaluation function has a negative correlation with the optimization time of the algorithm. We need to adjust the weight of evaluation function reliability and optimization time according to specific conditions.

4. Conclusions and Future Research

4.1. Conclusions

This paper proposes IWOA-ACO to plan the operation path of the electric tractor. IWOA introduces reverse learning strategy, nonlinear convergence factor and adaptive

inertia weighting factor to balance between global and local convergence capability of it and enhance the diversity of the initial solution. At the same time, IWOA-ACO improves evaluation function to ensure accurate evaluation of ACO performance during iteration.

First of all, taking a cultivated land environment in Xinjiang as an example, IWOA-ACO is used to plan the optimal path for the electric tractor to traverse the crop concentration points. The simulation results show that the algorithm has the advantages of fast convergence speed and good global convergence performance, which is helpful to improve the working efficiency of the electric tractor. Furthermore, taking the complex nodes in a concave–convex environment as an example, the length and energy consumption of IWOA-ACO planned path are 1.91% and 4.32% less than those of PSO-ACO, and are 1.95% and 1.25% less than those of WOA-ACO. This verifies the strong adaptability of IWOA-ACO to various environments.

In conclusion, IWOA-ACO can reduce the length and energy consumption of the planned path, which improves the operational efficiency and endurance of the electric tractor and assists the development of green agricultural machinery.

4.2. Suggestions for Future Work

There is room for further progress in the research on obtaining optimization parameters of ACO algorithm. This paper makes a brief analysis of them to provide research ideas for follow-up researchers.

1. As analyzed in Section 2.1.1, the ideal ACO parameter should not be a fixed value, but an adaptive function that is an iterative rule. In the research scheme proposed in this paper, IWOA-ACO algorithm can only obtain a set of set value parameters of ACO algorithm with good matching, but the step of obtaining the function from the set value parameters still needs to be completed by researchers. We propose that IWOA-ACO cannot directly derive the parameter iteration rules of ACO algorithm because of the limitations of its performance and dimensions. The analysis is as follows:

Assuming that the number of parameters to be determined by ACO algorithm is n_x, and the maximum number of iterations set by ACO algorithm when solving the path planning is m_x, the calculation dimension C_d of IWOA-ACO is as shown in Equation (30).

$$C_d = n_x m_x \tag{30}$$

Taking the simulation in Section 3 as an example, the maximum number of iterations is $m_x = 100$. In ACO algorithm iteration, the ant quantity should be set as a constant. Therefore, the value of C_d should be $C_d = 1 + 300 = 301$. However, the value of C_d with the scheme adopted in this paper is $C_d = 4$.

In a word, this scheme, obtaining the adaptive functions directly from IWOA-ACO, has a large calculation dimension, which greatly increases the computational complexity of the algorithm. Therefore, the algorithm easily falls into the local optimal solution, and the operation effect may not be as good as that of the scheme adopted in this paper. In further research, it might be possible to use a better performance algorithm to achieve this scheme.

2. As described in Section 2.4, in IWOA-ACO, the evaluation function is calculated by running the ACO algorithm only once under a set of parameter values. However, ACO algorithm, like GA, PSO, and WOA, belongs to intelligent algorithm, which has a high probability of obtaining the optimal value, but it cannot guarantee that every time it is the optimal value. If we aim to obtain a more reliable evaluation function, we need to allow ACO to run n_t times under each set of parameters to calculate the evaluation function by integrating the operation results. However, this will increase the running time of IWOA-ACO algorithm, as shown in Equation (31).

$$N_T = t_{IWOA-ACO} n_t \tag{31}$$

where $t_{\text{IWOA-ACO}}$ is the time required for the scheme adopted in this paper which lets ACO run once under each set of parameters to calculate the evaluation function.

In a word, we realize that researchers can set the running times of the ACO algorithm n_t on the basis of measuring the running time of the IWOA-ACO algorithm N_T and the reliability of the evaluation function.

Author Contributions: Methodology, C.L. and M.L.; Validation, C.L. and M.Z.; Writing—Original Draft Preparation, C.L. and K.P.; Writing—Review & Editing, M.L., M.Z. and K.P.; Visualization, C.L.; Funding Acquisition, M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [the International Cooperation Project of Shihezi University] grant number [GJHZ202003]. The APC was also funded by the same grant.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: When requested, the authors will make available all data used in this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, T.; Yan, G.; Wen, L.; Liao, Y. Current situation and development of electric agricultural machinery in China. *J. Agric. Mech. Res.* **2012**, *34*, 236–240.
2. Bing, K.; Liu, Y.; Song, F.; Zhang, S.; Xu, B.; Yu, S. Current status of electric tractor research in China. *Agric. Dev. Equip.* **2021**, *07*, 22–23.
3. Ueka, Y.; Yamashita, J.; Sato, K.; Doi, Y. Study on the development of the electric tractor: Specifications and traveling and tilling performance of a prototype tractor. *Eng. Agric. Environ. Food* **2013**, *6*, 160–164. [CrossRef]
4. Wang, Y.; Wang, W.; Xu, F.; Wang, J.; Chen, H. Path planning approach based on improved ant colony optimization for sprayer UAV. *Trans. Chin. Soc. Agric. Mach.* **2020**, *51*, 92, 103–112.
5. Jing, Y.; Jin, Z.; Liu, G. Three dimensional path planning method for navigation of farmland leveling based on improved ant colony algorithm. *Trans. Chin. Soc. Agric. Mach.* **2020**, *51*, 333–339.
6. Shang, G.; Chen, P.; Han, J.; Xia, C. Research on path planning algorithms for multi-operating points of electric tractors based on optimal energy consumption. *J. Chongqing Univ. Technol. Nat. Sci.* **2020**, *34*, 50–57.
7. Ljungqvist, O.; Evestedt, N.; Axehill, D.; Cirillo, M.; Pettersson, H. A path planning and path-following control framework for a general 2-trailer with a car-like tractor. *J. Field Robot.* **2019**, *36*, 1345–1377. [CrossRef]
8. Parsons, T.; Hanafi Sheikhha, F.; Ahmadi Khiyavi, O.; Seo, J.; Kim, W.; Lee, S. Optimal Path Generation with Obstacle Avoidance and Subfield Connection for an Autonomous Tractor. *Agriculture* **2022**, *13*, 56. [CrossRef]
9. Yang, J. Research on Ant Colony Algorithm and Its Application. Ph.D. Thesis, Zhejiang University, Hangzhou, China, 2007.
10. Ellefsen, K.O.; Lepikson, H.A.; Albiez, J.C. Multi objective coverage path planning: Enabling automated inspection of complex, real-world structures. *Appl. Soft Comput.* **2017**, *61*, 264–273. [CrossRef]
11. Contreras-Cruz, M.A.; Ayala-Ramirez, V.; Hernandez-Belmonte, U.H. Mobile robot path planning using artificial bee colony and evolutionary programming. *Appl. Soft Comput.* **2015**, *30*, 319–326. [CrossRef]
12. Jiang, K.; Li, M.; Zhang, H. Improved ant colony algorithm for travelling salesman problem. *J. Comput. Appl.* **2015**, *35*, 114–117.
13. Wang, H.; Fu, Y.; Yue, Y.; Zhao, H. Research on multi-point path planning of greenhouse robot based on parallel ant colony algorithm. *Jiangsu Agric. Sci.* **2019**, *47*, 237–241.
14. Morin, M.; Abi-Zeid, I.; Quimper, C.G. Ant colony optimization for path planning in search and rescue operations. *Eur. J. Oper. Res.* **2023**, *305*, 53–63. [CrossRef]
15. Stodola, P.; Otřísal, P.; Hasilová, K. Adaptive ant Colony optimization with node clustering applied to the travelling salesman problem. *Swarm Evol. Comput.* **2022**, *70*, 101056. [CrossRef]
16. Yuan, B.; Wang, W.; Wang, H. Multi-objective point path planning based on improved PSO algorithm. *J. Zhejiang Univ. Sci. Technol.* **2022**, *34*, 225–232, 284.
17. Yang, P. Multi Objective Point Path Planning for Mobile Robot. Master's Thesis, China West Normal University, Nanchong, China, 2021.
18. Mirjalili, S.; Lewis, A. The whale optimization algorithm. *Adv. Eng. Softw.* **2016**, *95*, 51–67. [CrossRef]
19. Meng, X.; Cai, C. Whale optimization algorithm based on elite reverse learning and Lévy flight. *Electron. Meas. Technol.* **2021**, *44*, 82–87.
20. Li, X.; Yang, D.; Li, X.; Wu, R. Flexible job shop AGV fusion scheduling method based on HGWOA. *China Mech. Eng.* **2021**, *32*, 938–950, 986.
21. Li, A.; Liu, S. Multi-strategy improved whale optimization algorithm. *Appl. Res. Comput.* **2022**, *39*, 1415–1421.

22. Huang, Y.; Zhang, L. Improved whale optimization algorithm and its application. *Comput. Eng. Appl.* **2019**, *55*, 220–226, 270.
23. Yang, B.; Li, C.; Li, Z.; Zhang, J. Improved whale optimization algorithm and application in path planning. *Comput. Meas. Control* **2021**, *29*, 187–193, 201.
24. Lei, C.; Zhao, H.; Jiang, N. Robot path planning based on particle swarm optimization and ant colony fusion algorithm. *J. Chongqing Univ. Technol. Nat. Sci.* **2020**, *34*, 235–241.
25. Lv, B.; Guo, Z.; Zhao, W.; Zhang, F. A review on optimization methods of standard particle swarm optimization. *Sci. Technol. Innov.* **2021**, *28*, 33–37.
26. He, X.; Zhang, G.; Chen, Y.; Yang, S. Multi-class algorithm of WOA-SVM using Lévy flight and elite opposition-based learning. *Appl. Res. Comput.* **2021**, *38*, 3640–3645.
27. Chen, P. Research on Path Planning Algorithms for Horticultural Electric Tractor Autonomous Operation. Master's Thesis, Jiangsu University, Zhenjiang, China, 2019.
28. Qi, X.; Li, W. Study on countermeasures of sustainable agricultural technology change in xinjiang production and construction corps. *Forum Sci. Technol. China* **2009**, 123–128.
29. Cai, X. Evaluation of agricultural regional economy of Xinjiang Corps. *Xinjiang State Farms Econ.* **2012**, 29–35.

agriculture

MDPI

Article

Development and Evaluation of a Deep Learning Based System to Predict District-Level Maize Yields in Tanzania

Isakwisa Gaddy Tende [1], Kentaro Aburada [2,*], Hisaaki Yamaba [2], Tetsuro Katayama [2] and Naonobu Okazaki [2]

[1] Department of Computer Studies, Dar es Salaam Institute of Technology, Dar es Salaam P.O. Box 2958, Tanzania
[2] Faculty of Engineering, University of Miyazaki, Miyazaki 889-2192, Japan
* Correspondence: aburada@cs.miyazaki-u.ac.jp

Abstract: Prediction of crop yields is very helpful in ensuring food security, planning harvest management (storage, transport, and labor), and performing market planning. However, in Tanzania, where a majority of the population depends on crop farming as a primary economic activity, the digital tools for predicting crop yields are not yet available, especially at the grass-roots level. In this study, we developed and evaluated Maize Yield Prediction System (MYPS) that uses a short message service (SMS) and the Web to allow rural farmers (via SMS on mobile phones) and government officials (via Web browsers) to predict district-level end-of-season maize yields in Tanzania. The system uses LSTM (Long Short-Term Memory) deep learning models to forecast district-level season-end maize yields from remote sensing data (NDVI on the Terra MODIS satellite) and climate data [maximum temperature, minimum temperature, soil moisture, and precipitation (rainfall)]. The key findings reveal that our unimodal and bimodal deep learning models are very effective in predicting crop yields, achieving mean absolute percentage error (MAPE) scores of 3.656% and 6.648%, respectively, on test (unseen) data. This system will help rural farmers and the government in Tanzania make critical decisions to prevent hunger and plan better harvesting and marketing of crops.

Keywords: electronic-agriculture; digital farming; machine learning; yield prediction; remote sensing; short message service (SMS); Web

Citation: Tende, I.G.; Aburada, K.; Hisaaki, Y.; Katayama, T.; Okazaki, N. Development and Evaluation of a Deep Learning Based System to Predict District-Level Maize Yields in Tanzania. *Agriculture* **2023**, *13*, 627. https://doi.org/10.3390/agriculture13030627

Academic Editors: Sebastian Kujawa and Gniewko Niedbała

Received: 14 September 2022
Revised: 20 February 2023
Accepted: 2 March 2023
Published: 6 March 2023

1. Introduction

In Tanzania, majority of the work force are engaged in agricultural activities [1]. The majority of these Tanzanians are rural small-scale farmers who depend on crop farming as their primary source of income [2,3], with maize (corn) being the most cultivated food crop [4]. Because majority of Tanzanians economically depend on agriculture, prediction of crop yields is of great importance, as it can help the central government to ensure food security and prevent hunger (for example, by importing food from abroad when low yields are predicted), while the rural farmers, farmer cooperative union officers, and local governments can plan better harvest management (storage, transport, and labor) and marketing of crops. However, digital tools for predicting crop yields are not yet available, especially at the grass-roots level in Tanzania (e.g., at the district level among rural farmers, farmer cooperative union officers, and local governments), making it hard to plan food assurance, harvesting, and marketing of crops. To address this issue, digital tools to predict crop yields at the grass-roots level in Tanzania are needed.

Machine learning is known to be effective for regression and classification problems, and it has been used in several studies to predict crop yields. For instance, Chen et al. [5] used a convolutional neural network (CNN) deep learning model based on faster region to predict the strawberry yield in the United States by using high-resolution aerial orthoimages, Fernandez-Beltran et al. [6] used a 3D CNN model to predict the rice yield in Nepal using Sentinel-2 satellite images, and Danilevicz et al. [7] used a CNN model to predict the maize yield using multispectral images, Wang et al. [8] used DNN (Deep Neural Network)

and other machine learning models such as SVM (Support Vector Machine), RF (Random Forest) and AdaBoost (Adaptive Boosting) to forecast winter wheat yield at county level by using climatic data, satellite images, and soil maps data) in the Conterminous United States (CONUS), and Wang et al. [9] proposed an algorithm based on decision tree which integrates topographical data and phenology characteristics of rice to map rice field patches in Chongqing, China. In addition, several studies have used RNN (recurrent neural network) models, including LSTM (Long Short-Term Memory) [10] and GRU (gated recurrent units) deep learning models to predict crop yields such as Alibabaei et al. [11] who proposed LSTM and GRU deep learning models to forecast potato and tomato yields using irrigation scheduling data and climatic data in Portugal, Haider et al. [12] who proposed an LSTM deep learning model to forecast wheat production in Pakistan using time-series wheat production data, Cho et al. [13] who proposed LSTM model which uses attention mechanism to forecast tomato yields in South Korea using time-series environmental variables, and Zhang et al. [14] who proposed an LSTM model to predict maize yields at county level in China using environmental data, optical data, fluorescence data, and thermal satellite image data. In contrast, several studies have used models that combine both CNNs and RNNs to predict crop yields by combining input data temporal features and spatial features such as Nevavuori et al. [15] who proposed CNN-LSTM and convolutional LSTM deep learning models to predict yields of oats, barley, and wheat in Finland by using weather and aerial drone images and Sun et al. [16] who proposed a deep CNN-LSTM model by using remote sensing and weather data.

Although the results from these studies reveal that deep learning is very effective for predicting crop yields, two issues make the deep learning models developed by these studies ineffective in Tanzania. First, most of these studies stop after developing the deep learning models and did not develop easy-to-use information systems, making them difficult to use for common people, such as rural farmers and farmer cooperative union officers in Tanzania. Second, although several studies [14–16] showed that using deep learning models to predict yields by combining remote sensing and climate data can lead to high accuracy in yield prediction, an information gap remains on the performance of deep learning models that use that data combination to predict crop yields at the grassroots level, especially in developing African countries like Tanzania, which is heavily rain-dependent and has two unique crop seasons depending on which rainfall modality the district belongs to, namely districts with unimodal (one) rain season or districts with bimodal (two) rain seasons.

To address these issues, this study had three objectives. The first objective was to develop two deep learning models (unimodal deep learning model for unimodal districts and bimodal deep learning model for bimodal rainfall districts) that use Tanzania district time-series remote sensing data (NDVI (Normalized Difference Vegetation Index)) and climate data [maximum temperature, minimum temperature, soil moisture, and precipitation (rainfall)] to predict end-of-season district-level maize yields. The second objective was to develop Maize Yield Prediction System (MYPS) based on the developed deep learning models that uses the short message service (SMS) of mobile phones (a viable medium because 86% of Tanzanians have mobile cellular subscriptions [17]) and a Web system to allow rural farmers, farmer cooperative union officers, and government officers in Tanzania to predict end-of-season district-level maize yields. The third objective was to evaluate the performance of the developed deep learning models in predicting end-of-season district-level maize yields and hence filling the existing information gap. As far as we know, no previous work has attempted to predict district-level end-of-season maize yields in Tanzania by using deep learning and the combination of NDVI, maximum temperature, minimum temperature, soil moisture, and precipitation data.

Due to the current COVID-19 pandemic, it was not possible to travel to Tanzania and deliver MYPS to farmers for testing and collecting system improvement feedback. Due to this reason, the scope of this work was limited to just developing deep learning system prototype that allows users in Tanzania to predict district-level maize yields and evaluating

its performance in predicting the correct yields. This study aimed to provide answers to the following research questions. First question, what deep learning model designs can accurately predict district-level maize yields in Tanzania? Second question, to what extent are the developed deep learning models accurate in predicting district-level maize yields in Tanzania? Third question, what design of the system can be used to automatically process and give response to queries in Swahili (the national language of Tanzania) for maize yield prediction via SMS and the Web?

2. Materials and Methods

2.1. Collection of Data

It is important to indicate that the first author (I.G.T.) is of Tanzanian nationality. Functional and non functional requirements as well as information needs from rural farmers and government officers in the ward of Kyimo, Rungwe district, Mbeya region, Tanzania, were collected before system development started. Mbeya region was selected as a case study area because it is one of the leading regions in cultivating maize in Tanzania. Structured questionnaire guide was prepared by the first author in Swahili language in order to collect data from rural farmers and farmer cooperative union officers. Prepared questionnaire guide consisted of questions for collecting respondents' primary data such as personal information (e.g., age and gender), functional and non functional requirements, as well as information needs. Due to the COVID-19 pandemic, it was not possible to travel to Tanzania. Due to this reason, we asked one research assistant to conduct a survey in the households with 30 maize farmers and 5 farmer cooperative union officers in the ward of Kyimo, who were purposively selected. Functional and non functional requirements as well as information needs of government officers (ministry and district officers) were not collected, instead, we adopted user requirements we previously collected from government officers in our previous study [18]. We obtained research permits from the local government of Kyimo ward. The research assistant was given an allowance of 50 US dollars to facilitate his transportation. The research assistant conducted the survey for two weeks from 6 June 2022 to 20 June 2022. Afterwards, finishing the data collection activity in Tanzania, we analyzed the collected respondents' information needs and user requirements by using statistical tools like cross-tabulation.

2.2. Requirements Analysis

The key information need from respondents was the prediction of end-of-season maize yield at the district level. User requirements (system features) collected from survey respondents include ability to access the system via text SMS (requested by farmers) and via a Web system (assumed for district and ministry officers based on requirements collected in our previous study [18]), system availability, system security and a very short response time.

We made several key design decisions in order to meet requirements of users. For example, to allow farmers and farmer cooperative union officers to predict maize yields via SMS, we designed a function for requesting yield prediction via SMS queries which are based on keyword and which restrict users to use a certain format to write SMS queries by typing first a keyword and then two single words which are separated by single spaces for representing the district and maize season for which they are requesting the yield. To meet the need to predict yields, we included LSTM-based deep learning models. The LSTM deep learning models were trained and tested with Tanzania district time-series data (NDVI, maximum temperature, minimum temperature, soil moisture, and precipitation) together with historical district maize yields to train the network to correctly predict end-of-season maize yields. The LSTM network was chosen because of its high ability to process sequential time-series data [19–21] and its high performance in predicting other crop yields [11–14].

We used UML (Unified Modeling Language) diagrams were used for analyzing the requirements from users. The use case diagram shown in Figure 1 shows functions of different users of the system. For instance, the role of the district officer is to prepare remote sensing and climate data for his/her district in a particular crop season in comma-separated

values (CSV) format in Microsoft Excel and upload that data into the system's SQL database, which can then be used to forecast the end-of-season maize yield by the deep learning models for that particular district and crop season, after which farmers, farmer cooperative union officers, ministry officers, and district officers themselves can request prediction of the end-of-season district maize yield and use the predicted yield results to make informed decisions on planning food assurance, harvest management, and maize marketing.

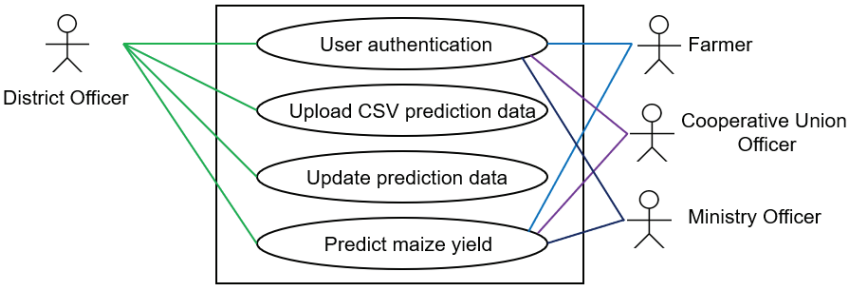


Figure 1. Use case diagram describing users' functions.

The sequence diagram in Figure 2 shows the steps of how a mobile user (farmer or cooperative union officer) uses SMS to request an end-of-season maize yield prediction from the system. First, the registered user must compose an SMS query with a keyword, district name, and crop season separated by single spaces and then send the SMS to the system's phone number. After the SMS gateway receives the SMS from a registered user, extraction and inspection of the SMS content is done by a query controller. It is important to note that each SMS query has to start with the keyword which is the first word in the SMS in order to activate the different functions of the system. For instance, the keyword used when a farmer signs up to the system is different from the keyword used when the farmer requests a yield prediction. For instance, to request the end-of-season maize yield prediction for Liwale district and the 2021/2022 crop season, the mobile user would compose the following SMS: "TABIRIMAVUNO Liwale 20212022" and send the SMS to the system's phone number. After extracting SMS content, comparison between the SMS keyword and stored keywords in the system is done by the query controller. If a match is found between the two keywords, then SQL stored procedure for that particular keyword is executed by the query controller to insert SMS data (district and crop season) into a database prediction requests table. Whenever it receives new values, the prediction requests table immediately activates a trigger to call the deep learning controller with SMS data as parameters. The deep learning controller retrieves the rain modality of the district in the SMS data from the database, as well as the prediction data [maximum temperature, minimum temperature, soil moisture, precipitation (rainfall), and NDVI] for the district and crop season in the SMS query. The deep learning controller then calls either the unimodal or bimodal pretrained deep learning model with prediction data as parameters to request a maize yield prediction. The already trained (pretrained) deep learning model predicts the maize yield based on the received parameters (prediction data). It is important to note that, the deep learning model is only trained once, and it does not need retraining every time a new prediction request is sent by users. If there is no match between the two keywords, a message of keyword error is retrieved from the SQL database. Finally, the predicted result is saved into the database (for future reference when prediction for the same district and crop season is requested) and an SMS is sent back to the user to provide the predicted maize yield or an error message.

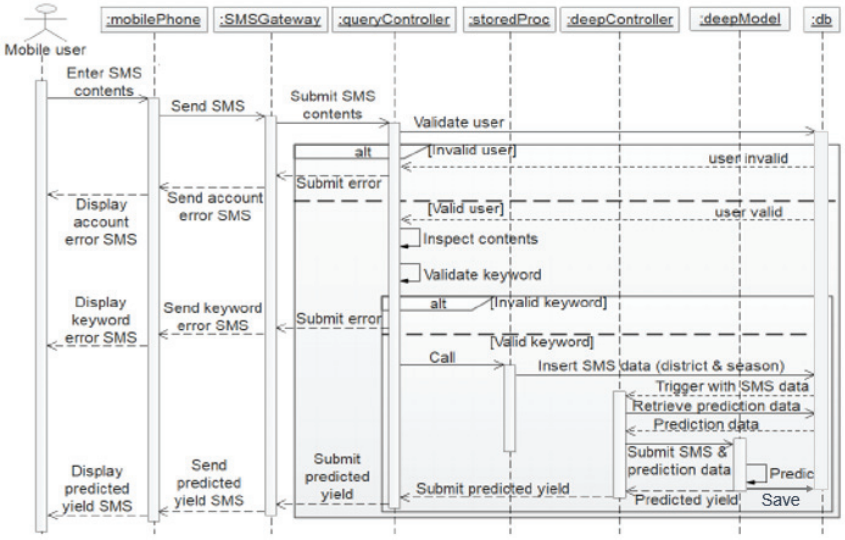


Figure 2. Sequence diagram for first request of maize yield prediction via SMS.

2.3. Tanzania Districts and Crop Seasons

Tanzania is a country located in East Africa and is subdivided into regions at the first administrative level and into districts at the second administrative level. In Tanzania, maize cultivation depends on rainfall, and also the calendar of maize cultivation is different from district to district based on the rainfall modality of the district [22]. The districts in the northern, eastern and western highlands receive bimodal rainfall (two rainy seasons per year) while the remaining districts in the central and southern regions receive unimodal rainfall (one rainy season per year) [22,23]. Phenology of maize crop in Tanzania is shown through the maize cultivation calendar in Figure 3. Table 1 shows Tanzania mainland districts whose data were used in this study to train and test the deep learning models. A previous study [23] was used as guidance for determine each district's rainfall modality. Note that some regions and districts were not involved in this study due to missing historical district maize yield data, which are required to train the deep learning models. For unimodal districts, we used data (NDVI, maximum temperature, minimum temperature, soil moisture, and precipitation) from the beginning of rainy season (November) to the end of rainy season (May) to train and test the unimodal deep learning model. For bimodal districts, we used the data of both rainy seasons (start and end of each, from September to January and from March to June) to train and test the bimodal deep learning model. Figure 4 shows map of the study area (the districts of Tanzania).

2.4. Software Development Process

MYPS was developed by using Waterfall software development model as shown in Figure 5. First, feasibility study through survey was conducted to determine feasibility of the proposed system and collect information needs and user requirements, then the user requirements were analyzed and specified. Next, the system was designed based on user requirements. Next, computer programs were written and individul modules were tested. Afterwards integration testing of the modules was conducted and the whole system was tested. In future the system will be delivered to Tanzanian farmers and maintained to improve it based on users feedback and correct any errors.

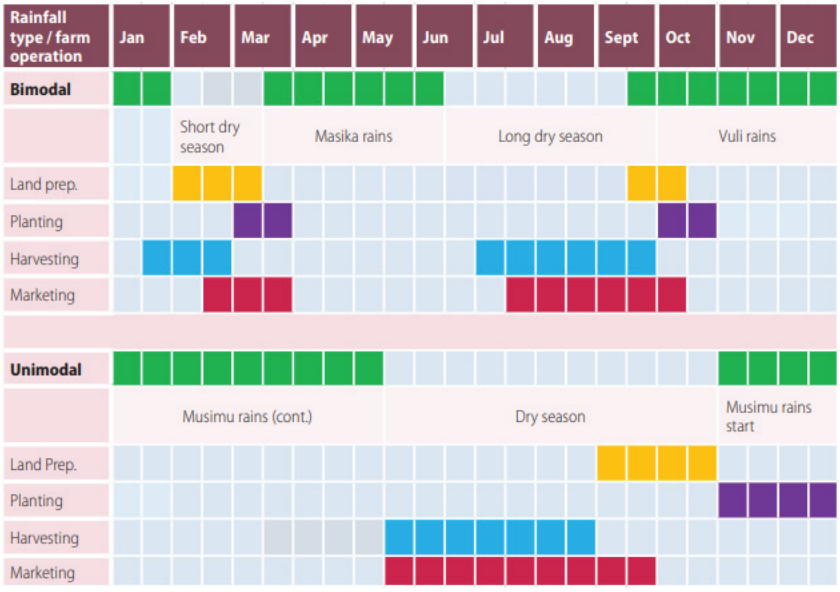


Figure 3. Maize cultivation calendar in Tanzania (Source: [22]).

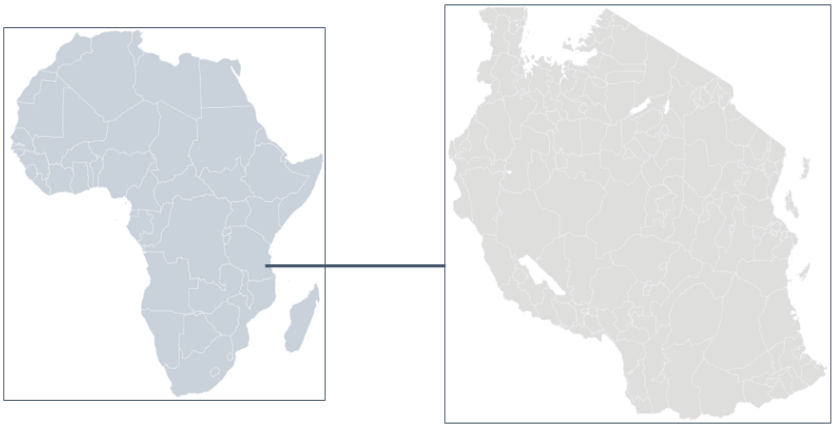


Figure 4. Map of Tanzania in Africa (**left**) and map of Tanzania districts (**right**).

Table 1. Tanzania districts whose data were used to train and test deep learning models.

Rainfall Modal	Region	Districts
Unimodal	Dodoma	Kondoa and Mpwapwa
	Iringa	Ludewa, Makete, and Mufindi
	Lindi	Kilwa, Lindi Rural, Lindi Urban, Liwale, Nachingwea, and Ruangwa
	Mbeya	Chunya, Kyela, Mbarali, Mbeya Rural, and Mbeya Urban
	Songwe	Ileje and Mbozi
	Morogoro	Kilombero, Kilosa, Morogoro Rural, Mvomero, and Ulanga
	Mtwara	Masasi, Mtwara Rural, Mtwara Urban, Newala, and Tandahimba

Table 1. *Cont.*

Rainfall Modal	Region	Districts
	Rukwa	Nkasi, Sumbawanga Rural, and Sumbawang Urban
	Ruvuma	Mbinga, Namtumbo, Songea Rural, Songea Urban, and Tunduru
	Simiyu	Bariadi, Maswa, and Meatu
	Shinyanga	Kahama, Kishapu, Shinyanga Rural, and Shinyanga Urban
	Geita	Bukombe
	Singida	Iramba, Manyoni, and Singida Rural
	Tabora	Igunga, Nzega, Sikonge, Tabora Urban, Urambo, and Uyui
Bimodal	Arusha	Karatu, Ngorongoro, and Monduli
	Dar es Salaam	Ilala, Temeke, and Kinondoni
	Kagera	Muleba
	Kigoma	Kasulu, Kibondo, and Kigoma Rural
	Kilimanjaro	Hai, Moshi Rural, Mwanga, Rombo, and Same
	Manyara	Hanang, Kiteto, and Simanjiro
	Mara	Bunda and Serengeti
	Mwanza	Ilemela, Kwimba, Magu, Misungwi, Sengerema, and Ukerewe
	Pwani	Bagamoyo, Kisarawe, Mafia, Mkuranga, and Rufiji
	Tanga	Handeni, Lushoto, Pangani, and Tanga

2.5. Deep Learning Data

2.5.1. Terra MODIS Satellite NDVI

NDVI [shown in Equation (1)] is an important vegetation index and is normally used to indicate the health of vegetation [24]. NDVI is computed from the red (visible) and NIR (near-infrared) lights which are reflected by vegetation and captured by satellite images in spectral bands. Normally, healthy vegetation (high NDVI value) has ability to absorb large amount of the red light, and has ability to reflect a large amount of the near-infrared light, while, in contrast, vegetation which is unhealthy (low NDVI value) has ability to reflect larger amount of red light and less amount of near-infrared light. In this study, the NDVI data were used to indicate the health of maize grown in Tanzania districts. The 8-day (collected every eight days) time series (from 2002 to 2010) mean (average) unsmoothed NDVI from NASA (National Aeronautics and Space Administration) Terra MODIS (Moderate Resolution Imaging Spectroradiometer) satellite [25] for each individual district in Table 1 were downloaded from the NASA GIMMS Global Agricultural Monitoring application [26] in CSV format. IFPRI SPAM 2010 v1 maize crop masks for Tanzania districts were used to ensure that the NDVI data came from district areas that grow only maize. IFPRI SPAM 2010 v1 crop masks are part of the NASA GIMMS Global Agricultural Monitoring application. Monthly NDVI values for each district were computed by averaging the 8-day NDVI values. We downloaded and processed NDVI data for every district involved in this study in the month of April 2022.

$$NDVI = \frac{NIR - red}{NIR + red}. \quad (1)$$

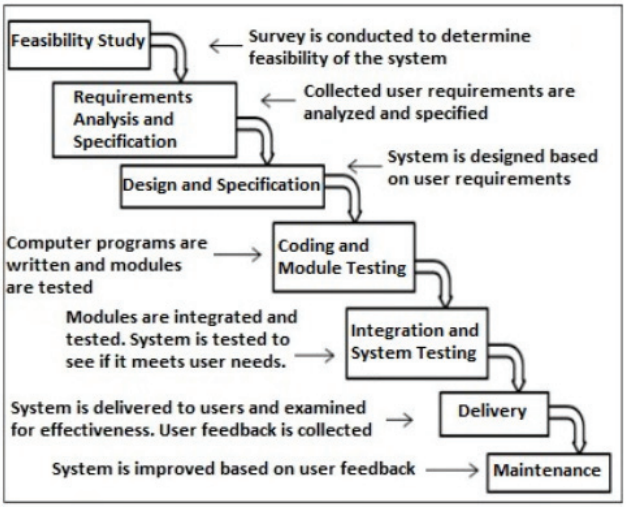


Figure 5. Waterfall software development model.

2.5.2. Climate Data

Time-series (from 2002 to 2010) monthly mean (average) climate data [maximum temperature, minimum temperature, soil moisture, and precipitation (rainfall)] for each district in Table 1 were downloaded in CSV format from the TerraClimate-Monthly dataset [27] in the Google Earth Engine (GEE) [28] through the ClimateEngine research application [29]. ClimateEngine offers easy-to-use interfaces to download time series data from the GEE in CSV format instead of directly interacting with the GEE through computer programs. Global Administrative Unit Layers (GAUL) Administration 2 [30] (which are part of ClimateEngine) were used to download climate data individually for each district in Tanzania. We downloaded and processed climate data for two months (May and June 2022) for every district involved in this study.

2.5.3. Historical Maize Yield Data

Historical maize yield data for the 2002/2003, 2003/2004, 2004/2005, 2005/2006, 2006/2007, 2007/2008, 2008/2009, and 2009/2010 maize seasons for all involved districts (in PDF or Microsoft Excel format) were downloaded from the website of the Tanzania Ministry of Agriculture [31]. Because almost all districts showed no strong trend in maize yields (increasing or decreasing) over a period of 8 years (from 2002/2003 to 2009/2010), we decided not to perform any detrending for district maize yields.

2.5.4. Input Data

To train the deep learning models, input data (NDVI, maximum temperature, minimum temperature, soil moisture, and precipitation) features were combined to form time-series input data points. Each data point has several timesteps of input data (seven data timesteps for a unimodal district representing 7 months of a unimodal maize season or nine data timesteps for a bimodal district representing 9 months of a bimodal maize season). Each data point was labeled with one numerical value to represent the maize yield at the end of that season. Each district had eight data points representing each season from 2002/2003 to 2009/2010, with data points from the 2002/2003 to 2008/2009 seasons used for training the deep learning models and data points from the 2009/2010 season used for testing the deep learning models on their effectiveness to correctly predict maize yields using data they had never seen before.

Table 2 shows an example of a data point (2002/2003 season) for a unimodal district (Liwale district) and another example of a data point (2002/2003 season) for a bimodal

district (Kiteto district). Figure 6 shows historical maize yields for Mufindi district over a period of 8 years, from the 2002/2003 season to the 2009/2010 season, while Figure 7 shows the variation of 8 years of time-series data from the 2002/2003 season to the 2009/2010 season for Mufindi district (unimodal district). All districts showed variations in data values over the duration of each maize season.

Table 2. Liwale and Kiteto district single data points for 2002/2003 maize season.

District	Date	Max-Temp (°C)	Min-Temp (°C)	Precipitation (mm)	Soil-Moisture (mm)	NDVI	Yield (Tonne/Hectare)
Liwale	11/1/2002	31.02	22.31	107.5	30.68	0.48	
	12/1/2002	30.92	22.28	140.84	50.02	0.57	
	1/1/2003	29.66	22.66	170.39	99.03	0.7	
	2/1/2003	30.28	22.38	132	119.28	0.76	
	3/1/2003	30.47	22.34	109.86	115.9	0.78	
	4/1/2003	29.6	21.39	82.24	103.64	0.78	
	5/1/2003	28.3	20.1	37.03	81.24	0.75	0.32
Kiteto	9/1/2002	27.07	13.57	15.2	14.67	0.31	
	10/1/2002	28.78	15.16	48.02	12.87	0.29	
	11/1/2002	29.5	16.54	36.59	11.48	0.31	
	12/1/2002	29.07	16.81	142.17	41.15	0.38	
	1/1/2003	29.13	16.4	50.58	27.17	0.52	
	3/1/2003	29.6	16.41	56.66	17.35	0.61	
	4/1/2003	28.82	17.06	46.35	14.83	0.65	
	5/1/2003	25.53	15.52	83.55	18.73	0.63	
	6/1/2003	25.39	14.11	7.25	14.7	0.52	0.53

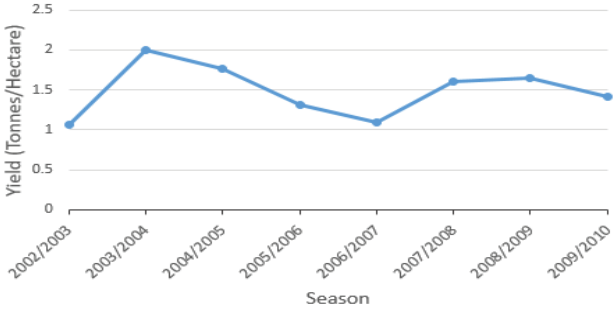


Figure 6. Mufindi district maize yields from 2002/2003 to 2009/2010 seasons.

2.5.5. Correlation of Input Data

As strongly correlated data can affect the training of machine learning models, we used the Pearson correlation coefficient (Pearson's r) [32] to analyze the correlation between the input data features climate data [maximum temperature, minimum temperature, soil moisture, precipitation (rainfall), and NDVI] for the 8 years of maize seasons (2002/2003 to 2009/2010). On average, correlation heatmaps of all districts showed that all five of the input data features are suitable for use in training deep learning models, with no parameters showing overly strong correlation. Figure 8 shows the Mufindi district correlation heatmap.

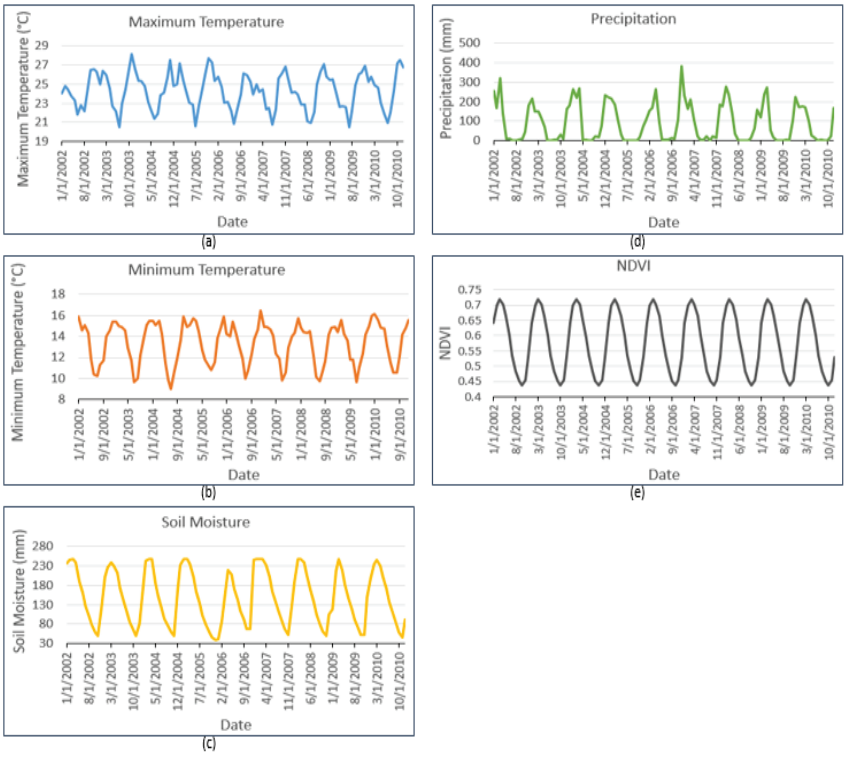


Figure 7. Mufindi district data features (Maximum Temperature, Minimum Temperature, Soil Moisture, Precipitation and NDVI) from 2002/2003 to 2009/2010 maize seasons.

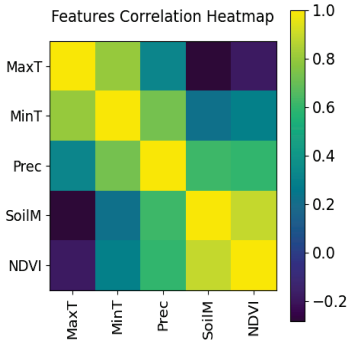


Figure 8. Mufindi district correlation heatmap based on data from 2002/2003 to 2009/2010 maize seasons.

2.6. Deep Learning Models

2.6.1. Architecture of Deep Learning Models

The purpose of using deep learning model is for pattern recognition in the timesteps of individual data points in input data and correctly predict the maize yield. We decided to use the same architecture (shown in Figure 9) for both unimodal and bimodal deep learning models. The only difference between the two deep learning models is the data points they process (data points with seven timesteps are used to train and test the unimodal deep learning model, while data points with nine timesteps are used to train and test the bimodal deep learning model). The architecture of the deep learning models is explained as follows.

- Preprocessing: Both the unimodal and bimodal deep learning models need to be trained with training data to correctly predict the maize yield and tested with test data to evaluate their effectiveness in predicting maize yields. In order to train and test the models, each data point has to be labelled. In this study, the labels are maize yields at the end of crop seasons. We labeled all district data points with their corresponding end-of-season maize yields as shown in Table 1. Because deep learning models only process numerical data as input; therefore, we transformed all data points into the float32 format. The unimodal training set contained 371 data points (2597 timesteps of input data), while the unimodal test set contained 53 data points (371 timesteps of input data). The bimodal training set contained 245 data points (2205 timesteps of input data), while the bimodal test set contained 35 data points (315 timesteps of input data). All of the datasets are provided in the Supplementary Materials.
- LSTM layers: To learn the patterns in the input data points and correctly predict maize yields, we chose to use an LSTM network (a type of RNN) because it performs highly in processing data which have sequences such as time series data and data containing text. The proposed LSTM network is used to process the data points, timestep by timestep, by looping over the timesteps of the input data points at the same time keeping a memory (state) of the timestep data that it has processed. While doing this, the LSTM network saves the information with the purpose of using it later for preventing older signals from gradually vanishing (vanishing gradients) which results into better input data understanding. LSTM layer has several arguments one of which is output-dimensionality (number of units), indicating LSTM layer dimensionality for the output space. Usually, sequence batches are processed by the LSTM layer which takes the 3D tensor with the shape (batch-size, timesteps, input-features) as input and then returns a 3D tensor with the shape (batch-size, timesteps, output-features), like in the first LSTM layer in our deep learning model architecture or a 2D tensor with the shape (batch-size, output-features), like in the second LSTM layer in our deep learning model architecture]. The batch-size is used to indicate the amount of samples that has to be processed in every batch, input-features is used to indicate the input feature space dimensionality, timesteps indicates the length of sequence, and output-featuresis used to indicate the output feature space dimensionality.
- Dense layer: Trained to output a single numerical value as the predicted maize yield.

Figure 9. Architecture of unimodal and bimodal deep learning models.

2.6.2. Deep Learning Loss Function (MSE)

While training the deep learning models, we used loss function of Mean squared error (MSE) [refer to Equation (2)]. MSE measures the average squared difference between the forecasted (predicted) maize yield value x_i and the true (actual) maize yield value y_i. During training, the Adam optimizer [33] minimizes this loss to ensure the model learns appropriate weights to help predicting maize yield values which are close to the true maize yield values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2. \quad (2)$$

2.6.3. Deep Learning Evaluation Metric (MAPE)

The mean absolute percentage error (MAPE) [refer to Equation (3)] was used as an evaluation metric while training and testing the deep learning models. It measures the

prediction accuracy between the predicted maize yield value x_i and actual maize yield value y_i.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}|\frac{y_i - x_i}{y_i}|. \quad (3)$$

2.7. K-Fold Cross-Validation

To build an effective deep learning model, it is important that deep learning model hyperparameters are tuned. While training of both the unimodal and bimodal deep learning models, we continuously adjusted several parameters, like number of LSTM layers, the size of LSTM layers, and amount of epochs, by looking at how the models perform on the validation datasets. During the process of training our deep learning models, K-fold cross-validation [34] was used. In K-fold cross-validataion, the training dataset is partitioned into K partitions, and then for every partition, the model is trained on the rest $K-1$ partitions and then evaluated on that specific partition. Then the model's validation MSE score is the average of the K validation MSE scores. For our research, we used Four-fold cross-validation, choosing four as the value of K. Unimodal and bimodal training sets were separately used in K-fold cross-validation experiments for the unimodal and bimodal deep learning models, respectively.

2.8. Design of System

2.8.1. Architecture of System

MYPS design uses three-tier architecture which is shown in Figure 10. In case of an SMS query, the Ozeki NG SMS Gateway [35] together with the MySQL stored procedures are used for interpreting the keyword, aunthenticating the user (cooperative union officer or farmer), and executing SQL queries to insert prediction data into trigger tables, which in turn pass the prediction data to Python programming language scripts of the deep learning models. The deep learning model (unimodal or bimodal) in turn predicts the district maize yield, which is then sent back to the user via SMS. Authenticating users assures security, while automatically responding to users' requests assures short response time and availability. In case of the Web system, a district officer interacts with the Web system by using Web browser, gets authenticated and then registers or updates district prediction data and requests district maize yields prediction. In addition, a ministry officer uses a Web browser to interact with the Web system, and after authentication, he/she can request a district maize yield prediction. To request a district maize yield prediction, both users use HTML forms in the Pug (Jade) template engine to interact with Express.js and then Node.js, which interacts with the Python deep learning models through a Node.js child process spawn functionality [36], after which the predicted district maize yield is displayed. HTTPS (Hypertext Transfer Protocol Secure) in Web requests, users' authentication, and sessions in Node.js programming language assure security. Automatic processing and response of Web requests by the Node.js server, the Express.js server, as well as MySQL assure availability.

2.8.2. Database Design

The MySQL database management system was used to create the system's database, which is relational and contains several tables related to each other. The entity relationship diagram (ERD) in Figure 11 shows the relational database design. The ERD was generated using MySQL Workbench [37].

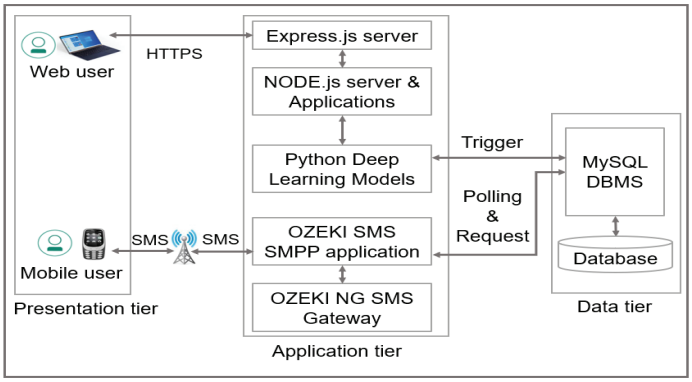


Figure 10. System design using three-tier architecture.

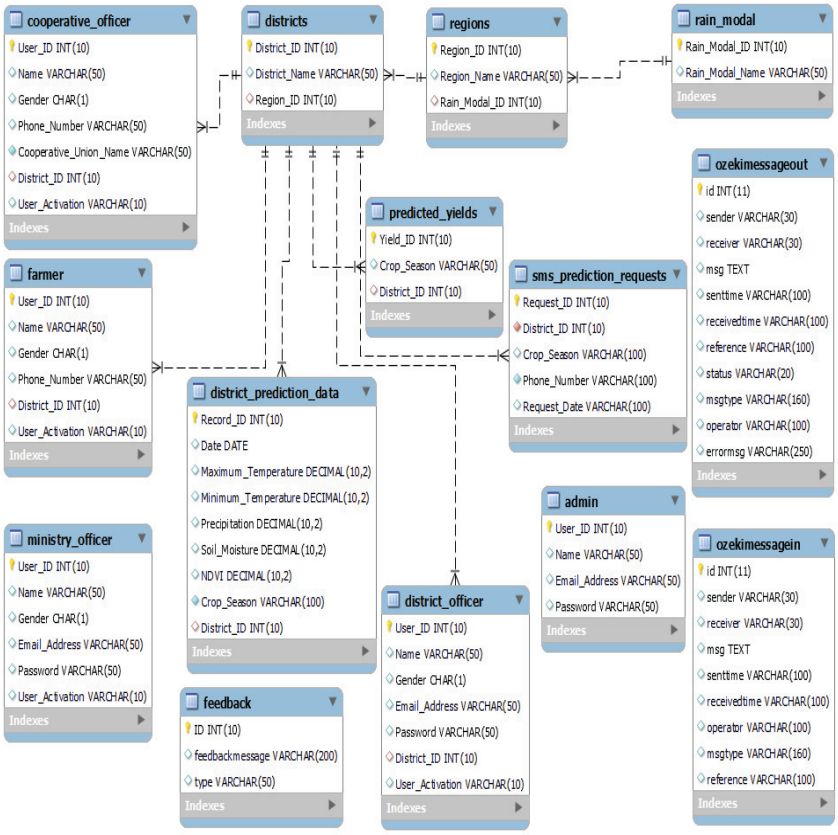


Figure 11. Entity relationship diagram with database design.

2.9. Prototype Implementation and Testing

We implemented MYPS prototype with all requested functions. It was not possible to travel to Tanzania due to the restrictions in travelling around the world, hence MYPS was not delivered to Tanzanian users. In order to test the functions in the system, the first author took the role of district officer and prepared and uploaded two data points in CSV format, each for a different district, Liwale (unimodal district) and Kiteto (bimodal district).

A fellow student was also asked by the first author to take farmer's role and then via SMS, request a prediction of maize yield from the system. Figure 12 shows how a ministry officer can request and receive Liwale district end-of-season maize yield predictions in the Web system, as well as how a farmer can sign up and request a Liwale district end-of-season maize yield prediction via SMS.

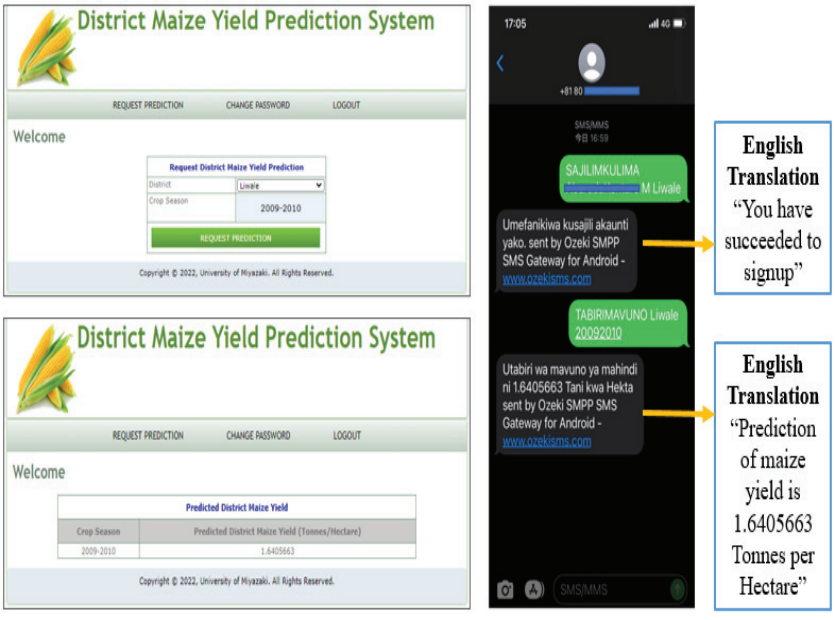


Figure 12. Top left: Ministry officer requests Liwale district end-of-season maize yield in the Web system. **Bottom left**: Requested maize yields displayed in the Web system. **Right**: Farmer signs up and requests and receives Liwale district end-of-season maize yields in Swahili.

3. Results

3.1. Hyperparameter Tuning Experiments

While training the deep learning models, we tuned the models by varying different hyperparameters. Four-fold cross-validation was used in evaluating models' performances on validation datasets by observing MSE scores. While doing this, modification to the hyperparameters was done accordingly to improve the models' validation accuracies. This activity was repeated a number of times to get the models' best hyperparameters. For example, Figures 13 and 14 indicate the average validation accuracy per each epoch in the cross-validation experiment while training the unimodal and bimodal deep learning models, respectively. At the end, we got the following hyperparameters for both deep learning models: 2 layers of LSTM, output-dimensionality of 100 and 200 for the first and second LSTM layers respectively, batch-size of 16, learning rate of 0.001 for Adam optimizer which is used to minimize the loss, and 4500 training epochs.

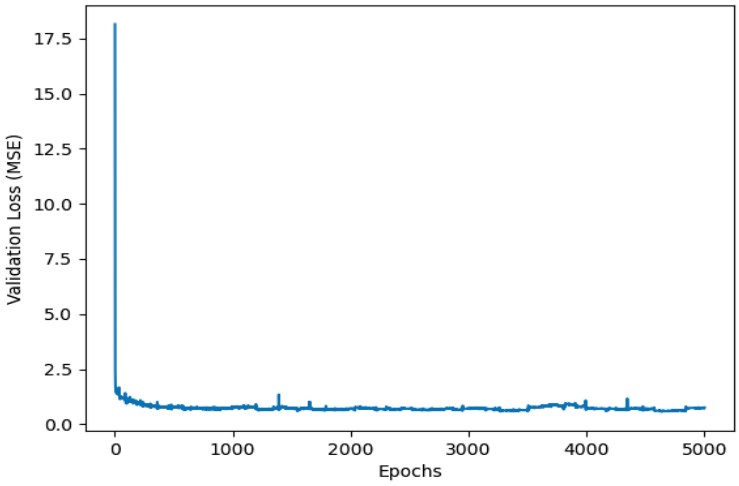


Figure 13. Average validation MSE in Four-fold cross-validation experiment for unimodal deep model.

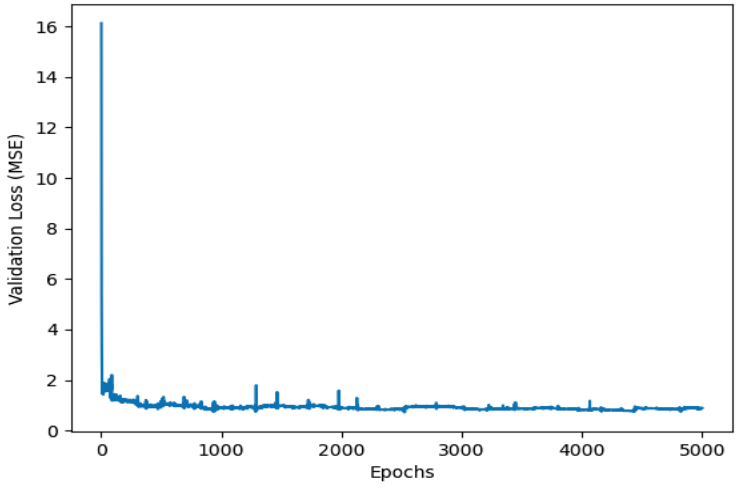


Figure 14. Average validation MSE in Four-fold cross-validation experiment for bimodal deep model.

3.2. Final Training Experiments

After completing tuning hyperparameters, the deep learning models were configured according to the hyperparameters we got and afterwards final experiments were conducted by training the unimodal and bimodal deep learning models on unimodal and bimodal training sets, respectively. Figures 15 and 16 show the MAPE scores of the final training of the models. We conducted experiments in Keras version 2.3.1 deep learning library and TensorFlow version 2.0.0 deep learning backend on a desktop computer with Windows 10 Operating System (OS), 3.60-GHz Intel (R) Core (TM) i7 processor and 16 GB RAM (Random Access Memory).

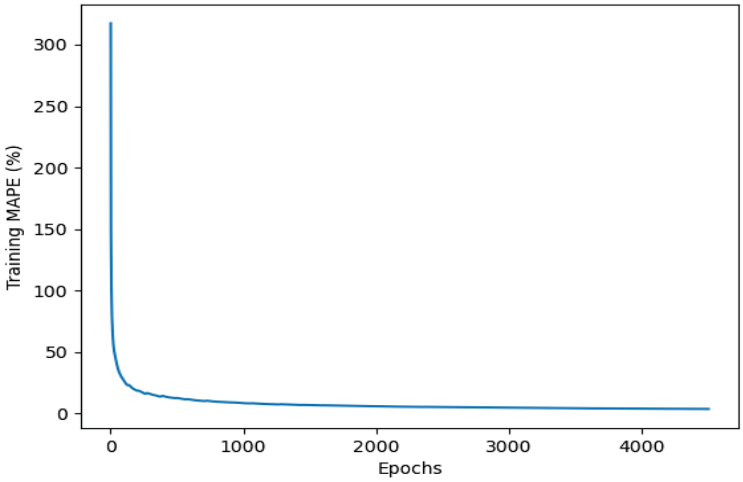


Figure 15. Final training MAPE of unimodal deep learning model.

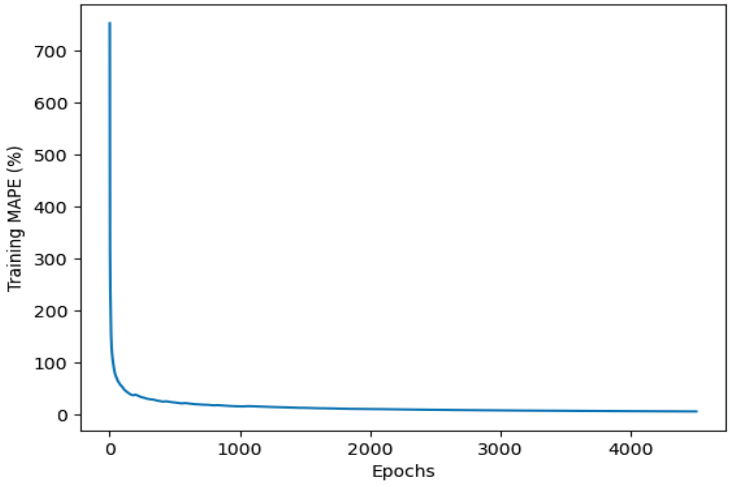


Figure 16. Final training MAPE of bimodal deep learning model.

3.3. Prediction Accuracy on Test Sets and Comparisons

For evaluating prediction accuracies of the deep learning models, their MAPE scores were evaluated on datasets that they had never seen before (test sets). Our unimodal and bimodal deep learning models obtained MAPE scores of 3.656% and 6.648%, respectively. These results indicate that, the deep learning models that we have proposed are highly effective and have high prediction accuracies for the end-of-season district maize yields in Tanzania.

In contrast, the evaluation of the proposed deep learning model in Nevavuori et al. [15], which also used an LSTM network to predict crop yields, revealed MAPE scores of 7.17% and 5.51% for crop yield predictions. These results imply that, on top of proposing easy to use Swahili based system, our proposed deep learning models have sufficient effectiveness because they have yield prediction accuracies which are comparable to those of existing deep learning models that have shown effectiveness in predicting crop yields in other countries.

4. Discussion and Conclusions

4.1. Discussion

4.1.1. Improved Accessibility

The inclusion of SMS in our proposed system allows grass-roots users, such as rural farmers and cooperative union officers, to request and receive end-of-season maize district yields even if they do not have Internet access, computers, or the knowhow to directly interact with deep learning models. This maize yield prediction will help these rural farmers, as well as the government, to make better and critical plans for food assurance, harvest management, and crop marketing.

4.1.2. Deep Learning Models Impact on Processing Combined NDVI and Climate Data

This study's results reveal that, our LSTM deep learning models in MYPS can use combined input data of NDVI, maximum temperature, minimum temperature, soil moisture, and precipitation to predict end-of-season Tanzania district-specific maize yields with great effectiveness. The findings help in filling an existing information gap on the impact of deep learning models in predicting crop yields using this data combination, especially in Tanzania.

4.1.3. Study Limitations

Because some districts have missing historical maize yield data, the deep learning models did not have equal representation of data points during training, and some districts in Tanzania were not included at all. This might limit generalization of the developed system when deployed for all districts in Tanzania.

4.1.4. Major Contributions

Major contributions of this study include the following:

- Completed and ready-to-use deep learning-based information system that allows farmers, cooperative union officers, district officers, and ministry officers in Tanzania to forecast end-of-season district maize yields via SMS and Web system.
- Novel method of using SMS to query deep learning models using MySQL triggers.
- Deep learning architectures that can also be adopted and used by other researchers.
- Deep learning datasets with prediction data from almost all districts in Tanzania, which can be used by other researchers.
- Performance evaluation findings that fill the existing information gap on effectiveness of deep learning models in predicting crop yields in Tanzania.

4.2. Conclusions

In this work, we have developed MYPS which is based on deep learning and which is accessed by SMS and the Web for predicting end-of-season maize yields for Tanzania districts. The key finding is that our deep learning networks are effective in predicting end-of-season Tanzania district-specific maize yields. As part of the implementation policy, we recommend to the government of Tanzania to invite investors who are able to implement deep learning based solutions to predict crop yields via SMS technology in affordable low end mobile phones in Tanzania. Future work will involve delivering MYPS to Tanzanian farmers and later evaluating its usability (how easy is to learn and use it) to the farmers through System Usability Scale (SUS).

Supplementary Materials: The following datasets are available online at https://www.mdpi.com/article/10.3390/agriculture13030627/s1: Data S1, Time-series datasets with maximum temperature, minimum temperature, soil moisture, precipitation, NDVI, and historical maize yield data for Tanzanian districts from 2002/2003 to 2009/2010 maize seasons for the purpose of training, validating, and testing the deep learning models.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, visualization, writing—original draft preparation, I.G.T.; resources, data curation, writing—review and editing, I.G.T., K.A., H.Y., T.K. and N.O.; supervision, project administration, funding acquisition, K.A., H.Y., T.K. and N.O. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Japan Society for the Promotion of Science, KAKENHI Grant Numbers JP18K11268 and JP21K11849.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the article and Supplementary Materials.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. URT [United Republic of Tanzania]: National Agriculture Policy. Available online: http://extwprlegs1.fao.org/docs/pdf/tan141074.pdf (accessed on 30 August 2022).
2. The World Bank: Rural Population. Available online: https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS?locations=TZ (accessed on 30 August 2022).
3. URT [United Republic of Tanzania]: Household Budget Survey (HBS), 2011/12—Key Findings Report. Available online: https://www.nbs.go.tz/index.php/en/census-surveys/poverty-indicators-statistics/household-budget-survey-hbs/148-household-budget-survey-hbs-2011-12-key-findings-report (accessed on 30 August 2022).
4. USDA [United States Department of Agriculture]: Tanzania—United Republic of, Grain and Feed Annual 2019 Tanzania Corn, Wheat and Rice Report. Available online: https://apps.fas.usda.gov/newgainapi/api/report/downloadreportbyfilename?filename=Grain%20and%20Feed%20Annual_Dar%20es%20Salaam_Tanzania%20-%20United%20Republic%20of_4-9-2019.pdf (accessed on 30 August 2022).
5. Chen, Y.; Lee, W.S.; Gan, H.; Peres, N.; Fraisse, C.; Zhang, Y.; He, Y. Strawberry Yield Prediction Based on a Deep Neural Network Using High-Resolution Aerial Orthoimages. *Remote Sens.* **2019**, *11*, 1584. [CrossRef]
6. Fernandez-Beltran, R.; Baidar, T.; Kang, J.; Pla, F. Rice-Yield Prediction with Multi-Temporal Sentinel-2 Data and 3D CNN: A Case Study in Nepal. *Remote Sens.* **2021**, *13*, 1391. [CrossRef]
7. Danilevicz, M.F.; Bayer, P.E.; Boussaid, F.; Bennamoun, M.; Edwards, D. Maize Yield Prediction at an Early Developmental Stage Using Multispectral Images and Genotype Data for Preliminary Hybrid Selection. *Remote Sens.* **2021**, *13*, 3976. [CrossRef]
8. Wang, Y.; Zhang, Z.; Feng, L.; Du, Q.; Runge, T. Combining Multi-Source Data and Machine Learning Approaches to Predict Winter Wheat Yield in the Conterminous United States. *Remote Sens.* **2020**, *12*, 1232. [CrossRef]
9. Wang, L.; Ma, H.; Li, J.; Gao, Y.; Fan, L.; Yang, Z.; Yang, Y.; Wang, C. An automated extraction of small- and middle-sized rice fields under complex terrain based on SAR time series: A case study of Chongqing. *Comput. Electron. Agric.* **2022**, *200*, 107232. [CrossRef]
10. Sepp, H.; Jürgen, S. Long short-term memory. *J. Neural Comput.* **1997**, *9*, 1735–1780.
11. Alibabaei, K.; Gaspar, P.D.; Lima, T.M. Crop Yield Estimation Using Deep Learning Based on Climate Big Data and Irrigation Scheduling. *Energies* **2021**, *14*, 3004. [CrossRef]
12. Haider, S.A.; Naqvi, S.R.; Akram, T.; Umar, G.A.; Shahzad, A.; Sial, M.R.; Khaliq, S.; Kamran, M. LSTM Neural Network Based Forecasting Model for Wheat Production in Pakistan. *Agronomy* **2019**, *9*, 72. [CrossRef]
13. Cho, W.; Kim, S.; Na, M.; Na, I. Forecasting of Tomato Yields Using Attention-Based LSTM Network and ARMA Model. *Electronics* **2021**, *10*, 1576. [CrossRef]
14. Zhang, L.; Zhang, Z.; Luo, Y.; Cao, J.; Tao, F. Combining Optical, Fluorescence, Thermal Satellite, and Environmental Data to Predict County-Level Maize Yield in China Using Machine Learning Approaches. *Remote Sens.* **2020**, *12*, 21. [CrossRef]
15. Nevavuori, P.; Narra, N.; Linna, P.; Lipping, T. Crop Yield Prediction Using Multitemporal UAV Data and Spatio-Temporal Deep Learning Models. *Remote Sens.* **2020**, *12*, 4000. [CrossRef]
16. Sun, J.; Di, L.; Sun, Z.; Shen, Y.; Lai, Z. County-Level Soybean Yield Prediction Using Deep CNN-LSTM Model. *Sensors* **2019**, *19*, 4363. [CrossRef] [PubMed]
17. The World Bank: Mobile Cellular Subscriptions (per 100 People)—Tanzania. Available online: https://data.worldbank.org/indicator/IT.CEL.SETS.P2?locations=TZ (accessed on 30 August 2022).
18. Tende, I.G.; Kubota, S.; Yamaba, H.; Aburada, K.; Okazaki, N. Evaluation of farmers market information system to connect with some social stakeholders. *J. Inf. Process.* **2018**, *26*, 247–256.
19. Wang, J.-H.; Liu, T.-W.; Luo, X. Combining Post Sentiments and User Participation for Extracting Public Stances from Twitter. *Appl. Sci.* **2020**, *10*, 8035. [CrossRef]

20. Al-Laith, A.; Shahbaz, M.; Alaskar, H.F.; Rehmat, A. AraSenCorpus: A Semi-Supervised Approach for Sentiment Annotation of a Large Arabic Text Corpus. *Appl. Sci.* **2021**, *11*, 2434. [CrossRef]
21. Yasar, H.; Kilimci, Z.H. US Dollar/Turkish Lira Exchange Rate Forecasting Model Based on Deep Learning Methodologies and Time Series Analysis. *Symmetry* **2021**, *12*, 1553. [CrossRef]
22. Feed the Future (U.S. Government's Global Hunger and Food Security Initiative): Maize Production Manual for Smallholder Farmers in Tanzania. Available online: https://cgspace.cgiar.org/bitstream/handle/10568/109806/Maize%20production%20manual%20for%20smallholder%20farmers%20in%20Tanzania.pdf?sequence=1 (accessed on 30 August 2022).
23. United Nations World Food Programme: Special Report, FAO/WFP Crop and Food Supply Assessment Mission to the United Republic of Tanzania. Available online: https://www.fao.org/3/w7958e/w7958e00.htm (accessed on 30 August 2022).
24. NASA Eath Observatory: Normalized Difference Vegetation Index (NDVI). Available online: https://earthobservatory.nasa.gov/features/MeasuringVegetation/measuring_vegetation_2.php#:~:text=Normalized%20Difference%20Vegetation%20Index%20(NDVI,up%20the%20spectrum%20of%20sunlight (accessed on 30 August 2022).
25. NASA TERRA: Moderate Resolution Imaging Spectroradiometer. Available online: https://terra.nasa.gov/about/terra-instruments/modis (accessed on 30 August 2022).
26. NASA: GIMMS Global Agricultural Monitoring. Available online: https://glam1.gsfc.nasa.gov/ (accessed on 30 August 2022).
27. Abatzoglou, J.; Dobrowski, S.; Parks, S.; Hegewisch, K.C. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015. *Sci. Data* **2018**, *5*, 170191. [CrossRef]
28. Google: Google Earth Engine. Available online: https://earthengine.google.com/ (accessed on 30 August 2022).
29. Climate Engine: Research App. Available online: https://app.climateengine.com/climateEngine (accessed on 30 August 2022).
30. FAO [The Food and Agriculture Organization]: Global Administrative Unit Layers (GAUL). Available online: https://data.review.fao.org/map/catalog/srv/api/records/9c35ba10-5649-41c8-bdfc-eb78e9e65654 (accessed on 30 August 2022).
31. URT[United Republic of Tanzania], Ministry of Agriculture: TAKWIMU. Available online: https://www.kilimo.go.tz/resources/category/takwimu (accessed on 30 August 2022).
32. Wikipedia: Pearson Correlation Coefficient. Available online: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient (accessed on 30 August 2022).
33. Keras: Adam Optimizer. Available online: https://keras.io/api/optimizers/adam/ (accessed on 30 August 2022).
34. Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-Validation. In *Encyclopedia of Database Systems*; Liu, L., Özsu, M.T., Eds.; Springer: Boston, MA, USA, 2016.
35. OZEKI NG SMS Gateway. Available online: https://ozekisms.com/ (accessed on 30 August 2022).
36. Node.js Organization: Node.js v18.8.0 Documentation, Child Process. Available online: https://nodejs.org/api/child_process.html#child_processspawncommand-args-options (accessed on 30 August 2022).
37. MySQL Workbench. Available online: https://www.mysql.com/products/workbench/ (accessed on 30 August 2022).

agriculture

MDPI

Article

Prediction of Pea (*Pisum sativum* L.) Seeds Yield Using Artificial Neural Networks

Patryk Hara [1], Magdalena Piekutowska [2] and Gniewko Niedbała [3,*]

[1] Agrotechnology, Jagiellonów 4, 73-150 Łobez, Poland
[2] Department of Geoecology and Geoinformation, Institute of Biology and Earth Sciences, Pomeranian University in Słupsk, 27 Partyzantów St., 76-200 Słupsk, Poland
[3] Department of Biosystems Engineering, Faculty of Environmental and Mechanical Engineering, Poznań University of Life Sciences, Wojska Polskiego 50, 60-627 Poznań, Poland
* Correspondence: gniewko.niedbala@up.poznan.pl

Abstract: A sufficiently early and accurate prediction can help to steer crop yields more consciously, resulting in food security, especially with an expanding world population. Additionally, prediction related to the possibility of reducing agricultural chemistry is very important in an era of climate change. This study analyzes the performance of pea (*Pisum sativum* L.) seed yield prediction by a linear (MLR) and non-linear (ANN) model. The study used meteorological, agronomic and phytophysical data from 2016–2020. The neural model (N2) generated highly accurate predictions of pea seed yield—the correlation coefficient was 0.936, and the RMS and MAPE errors were 0.443 and 7.976, respectively. The model significantly outperformed the multiple linear regression model (RS2), which had an RMS error of 6.401 and an MAPE error of 148.585. The sensitivity analysis carried out for the neural network showed that the characteristics with the greatest influence on the yield of pea seeds were the date of onset of maturity, the date of harvest, the total amount of rainfall and the mean air temperature.

Keywords: pea; seeds yield prediction; ANN; MLR; sensitivity analysis

Citation: Hara, P.; Piekutowska, M.; Niedbała, G. Prediction of Pea (*Pisum sativum* L.) Seeds Yield Using Artificial Neural Networks. *Agriculture* **2023**, *13*, 661. https://doi.org/10.3390/agriculture13030661

Academic Editors: Raul Morais dos Santos and Bing Liu

Received: 14 February 2023
Revised: 26 February 2023
Accepted: 10 March 2023
Published: 12 March 2023

1. Introduction

There are many challenges facing modern agriculture. The priority is to increase food production while minimizing environmental impact [1,2]. This task seems particularly difficult in the face of climate change and the occurrence of increasingly frequent extreme weather events, which pose a serious threat to crop yields [3]. It is assumed that about 67% of crop variability is governed by the weather conditions that prevail throughout the crop growing season, with 33% governed by other factors, such as agrotechnology or habitat conditions [4]. Therefore, early and accurate forecasting of crop yields is becoming increasingly important [5]. Being able to estimate yields a few weeks before harvest allows an appropriate strategy to be taken for the pricing of agricultural products. Yield prediction can also be a useful tool for decision makers in regulating both exports in case of surpluses and imports in times of agricultural commodity shortages [6]. In addition, early information on yields can help farmers in terms of work planning and storage space selection [7]. This knowledge can also help to improve the profitability of agricultural production by optimizing the number of crop protection and/or fertilization treatments. Lower usage of these products leads to a reduction in total labor inputs and on-farm energy inputs [8,9]. In the final balance, these factors contribute to increased labor productivity, conservation of natural resources and increased farm profitability through lower production costs [10]. Accurate and early prediction of crop yields also plays an important role in global food security by providing valuable information to various stakeholders (farm owners, agronomists, etc.) [11,12].

The need for accurate and timely predictive models for agricultural crop yield has led to a growing interest in this topic from the scientific community [13–18]. However,

developing a predictive model is not an easy task [19]. The issue is extremely complex due to the multitude of factors affecting crop yield. The most commonly cited determinants include genotype and weather conditions, including rainfall, sunshine, minimum and maximum air temperature, habitat conditions (soil pH, soil nutrient abundance, etc.), agronomics and the interactions of these factors [20–23]. Many different approaches are used in yield forecasting, and each method has strengths as well as limitations [24]. One such method is multiple linear regression (MLR). MLR, as a statistical tool, is able to predict yield based on, among other things, agronomic data. However, the effectiveness of this method is often questioned due to its low prediction accuracy [25]. The major disadvantage of MLR models is that they are not appropriate for explaining non-linear and complex relationships between yield and the factors that influence yield [26].

With the development of information technology, modern mathematical algorithm techniques such as machine learning (ML) have begun to be applied. The possibility of using models based on artificial intelligence has contributed to an increase in accurate forecasting of random and non-linear issues [27]. This feature has made machine learning the method most commonly used in yield modeling [28,29]. In addition to its high prediction quality, ML is able to identify patterns in datasets and reveal complex relationships between independent variables [30]. Additionally, the advantage of ML over traditional linear regression methods lies in its ability to use, as explanatory variables, two or more spectral variables from satellite imagery [31]. The inclusion of these data in yield modeling is becoming an increasingly common practice due to further improvements in prediction quality and the prospect of capturing new correlations between these factors and crop yield [32]. Furthermore, in machine learning-based models, it is possible to use linguistic variables without having to code them in advance, as is the case with regression models [9,33]. The accuracy of yield estimation that is achieved by machine learning methods means that these models require large datasets from a variety of sources. In the case of a small number of predictors, the proper calculation of yield variability by ML usually suffers from a large prediction error [34]. Other limitations in the use of ML are that some methods require computationally powerful equipment and that analysis time is much longer than it is for multivariate linear regression [35,36]. Machine learning models are also sensitive to significant correlations between independent characteristics. For this reason, the dataset that is fed into the model often requires prior preparation, and additional statistical analyses may need to be performed to capture these correlations [37]. Some of the most successful machine learning techniques are support vector machines (SVMs), convolutional neural networks (CNNs), random forest (RF), k-nearest neighbors (kNNs) and artificial neural networks (ANNs) [38–41].

ANNs are a mathematical tool that can create a non-linear representation of the connections between the explained variable and the input variables [42]. ANNs are, to some extent, inspired by the functioning of parts of the real (biological) nervous system [43]. However, the connection patterns of neurons in artificial neural networks are chosen arbitrarily and are not a model of actual neural structures. ANNs as a computer tool are distinguished by their ability to solve practical problems in a computerized manner without prior mathematical formalization [44]. Another advantage is that it is not necessary to refer to any theoretical assumptions about the problem being solved when working with neural networks. Even the assumption of causal relationships between exploratory and explanatory features need not be enforced [45]. The computations performed by the ANN are performed in parallel. The artificial neurons that make up the network perform their computational tasks simultaneously. This makes the network capable of solving the problem under analysis in a short period of time. However, the more complex the problem the neural network investigates, the more time it takes to find the right solution [46]. The most characteristic feature of artificial neural networks is the ability to learn from examples and the ability to self-generalize the acquired knowledge (generalization) [47]. The threat to generalization is overlearning. An overlearned network excessively adapts the acquired knowledge to irrelevant details of specific learning cases [48].

One commonly used ANN model is the multilayer perceptron (MLP) [17,18,49–51]. It is a fully connected unidirectional neural network [7] that typically consists of three layers: an input layer, at least one hidden layer consisting of sigmoidal neurons and an output layer consisting of sigmoidal or linear neurons. The back-propagation method is the most commonly used technique for learning MLP networks [52]. This method is based on the concept of correcting, at each stage of learning, the values of the weights based on the evaluation of the error made by each neuron during the learning of the network [53].

The present work is a continuation of the authors' previous research [54], which aimed to determine the effectiveness of linear (MLR) and non-linear (MLP) models in predicting the protein content of *Pisum sativum* L. pea seeds. The current study focused on the possibility of predicting the seed yield of general pea seeds using ANNs, and MLR was used as a comparative model. In addition, the study aimed to test three hypotheses: (i) the artificial neural networks model is an effective tool in predicting pea seed yield 20 days before harvest; (ii) five-year field trials under a variety of experimental conditions allow for the construction of a model predicting pea yield; and (iii) neural networks can predict yield more accurately than the MLR model.

2. Materials and Methods

This research was carried out between 2016 and 2020 at the Stations and Experimental Plants of the Research Center for Cultivar Testing (COBORU). The mission of COBORU is to stimulate innovation in plant breeding and seed science and to support the implementation of diverse progress into agricultural practice [55]. The work of this unit is focused, among other things, on research into the distinctiveness, uniformity and stability of crop varieties in Poland. In addition, COBORU is involved in conducting field research on the assessment of the cultivation and use value of agricultural crops. These studies are conducted under conditions as close as possible to production conditions. The results of the conducted experiments make it possible to determine whether a given variety can be entered into the National List of Varieties [56].

The experimental plots were located in Poland at the following locations: Bezek (N 51°12′6.722″ E 23°16′7.656″), Głębokie (N 52°38′33.18″ E 18°26′16.26″), Kawęczyn (N 52°10′15.157″ E 20°20′49.328″), Krzyżewo (N 53°1′33. 535″ E 22°45′28.438″), Pawłowice (N 50°27′14.049″ E 18°29′28.912″), Radostowo (N 53°59′20.566″ E 18°44′41.429″) and Sulejów (N 51°21′8.03″ E 19°52′7.517″). The experiments were situated in locations that are optimal for pea cultivation in terms of habitat. These localities are characterized by a temperate warm climate, with average monthly air temperatures ranging from −5.0 to −2.0 °C in January and 16 to 18 °C in July. Average annual precipitation is in the range of 550–800 mm [57]. According to Polish soil classification, clay soils of classes II-IIIb prevail in these localities. The data for the construction of the models are official data, coming from a variety of COBORU tests, and are acceptable to all authorities related to agricultural production in Poland. The nature of the experiments and the way in which they were carried out are recorded in the methodology [58], which is a set of experimental concepts and guidelines. This ensures that all research assumptions are met. The research was conducted in the same way at each COBORU point. The meteorological data were obtained from the archive database of the Institute of Meteorology and Water Management at the National Research Institute. A detailed description of the conduct of the experiments, the acquisition of the dataset and the sources of these data were previously described by the authors of this paper [54]. The construction of an ANN (N2) and MLR (RS2) model was performed based on 11 general purpose pea cultivars: Arwena, Astronaute, Batuta, Mecenas, Medyk, Mentor, Olimp, Spot, Starski, Tarchalska and Tytus.

2.1. Construction of the Database

The first and most important step in the construction of linear and non-linear models is the appropriate selection of input variables. The importance of this step is due to the fact that the chosen input parameters directly affect the performance of the resulting models [37].

The input variables shown in Table 1 were used to build the N2 and RS2 model. The output variable was pea seed yield expressed in t·ha^{-1}. The dataset consisted of 1155 cases/plots. Each plot was a separate case for model building. All cases that formed the dataset were divided into two sets: A and B. Data from 1040 plots were assigned to set A, while set B was created from the remaining 115 cases and was used for model validation.

Table 1. The structure and scope of the independent variables used in the construction and verification of the N2 and RS2 models.

Symbol	Unit of Measure	Description of the Variable	The Scope of Data
		Independent Variables	
RAIN	mm	Rainfall in the period from sowing to 14 July	96.9–312.4
SUN	h	Sum of insolation that occurred in the period from sowing to 14 July	630.5–1051.5
TEMP	°C	Average daily air temperature in the period from sowing to 14 July	11.0–17.5
N_F	kg/ha	Amount of nitrogen introduced into the soil with mineral fertilizers	10–90
P2O5_F	kg/ha	Amount of phosphorus incorporated into the soil with mineral fertilizers	0–80
K2O_F	kg/ha	Amount of potassium introduced into the soil with mineral fertilizers	0–119
SOWI	days	Date of sowing of field peas—defined as number of days since the beginning of the year	83–102
P_EMER	days	Pea crop emergence—defined as number of days since the beginning of the year	96–133
HAR	days	Date of harvesting of field pea plants—defined as the number of days from 1 January	184–221
FLOWE	days	Flowering onset date—number of days from the beginning of the year	126–169
INI_MA	days	Maturity onset date—defined as the number of days from 1 January	167–211
TECH_M	days	Technical maturity date—number of days since the beginning of the year	171–216
P_HIG	cm	Height of plants	43–156
WEGW	days	Number of plant growing days	87–137
PH	-	Soil reaction (pH)	5.5–7.5
P2O5_C	Range from 0 to 4 *	Phosphorus (V) oxide content of the soil	0–4
K2O_C	Range from 0 to 4 *	Potassium oxide content of the soil	0–4
MGO_C	Range from 0 to 4 *	Magnesium oxide content of the soil	0–4
GEN	Feature coded 101 to 111	Variety of peas	–
		Dependent Variable	
YIELD	t·ha^{-1}	Pea seed yield	2.30–8.02

* The range from 0 to 4 refers to the nutrient abundance of the soil. A value of 0 indicates very low abundance, 1 indicates low abundance, 2 indicates medium abundance, 3 indicates high abundance and 4 indicates very high abundance. Range from 0 to 4.

2.2. Construction of the N2 Model

In the present study, it was assumed that the forecast of pea seed yield would be made before harvesting [59], i.e., 14 July. The forecast date was selected based on the dominance of the onset of maturity of the pea varieties included in the dataset. The analysis of the dataset showed that the harvesting of peas of general use varieties was most often performed on 3 August. Therefore, the obtained linear and non-linear model predicted the yield 20 days before the harvest of peas grown under experimental conditions.

The construction of N2 models consisted of input variables being repeatedly provided to the network [60]. A total of 10,000 neural networks were tested using an automatic network designer. Different ANN model structures were analyzed, including variations in the number of neurons in the hidden layer. This method selected a model with an MLP architecture of 19:19-24-1:1 (Figure 1). The multilayer perceptron is a type of ANN widely recommended in works on similar topics due to its high potential for non-linear function estimation [18,61–63]. The main advantage of MLPs is the ability to discriminate data that cannot be linearly separated [7].

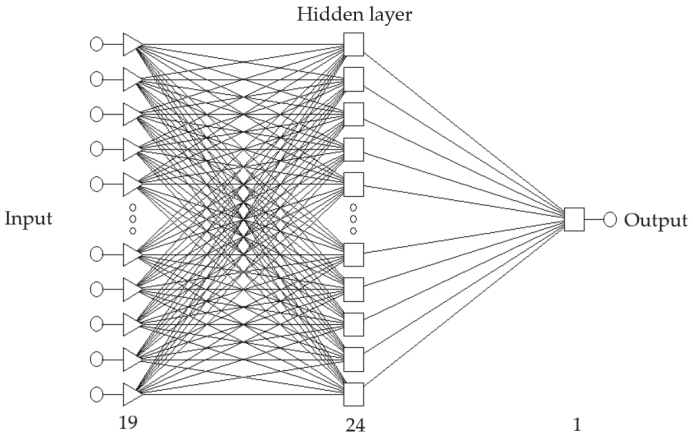


Figure 1. Network structure for the N2 model.

The optimization of the artificial neural network structure was performed by obtaining the minimum validation error. The selection of the neural network model was also guided by the size of the training and test set errors and other important quality parameters, as shown in Table 2. In this research, set A was divided into three subsets: learning, test and validation. This division is common in the development of predictive models using ANNs [5,59,64]. In the present study, 50% of the records (or 520 cases) were assigned to the learning subset. The test and validation subsets consisted of the same number of objects, i.e., 260, each representing 25% of the cases from the entire A set. The construction of the N2 model was carried out using Statistica v7.1 (TIBCO Software Inc., Palo Alto, CA, USA).

Table 2. Quality parameters and error rates of subsets and the number of learning epochs of neural networks.

Subsets	Teaching	Validation	Testing
Size of error	0.0556	0.0590	0.0679
Quality	0.3576	0.3645	0.4311
Epochs of learning			
Back-propagation method of error		100	
Coupled gradients method		110b *	

* b (best)—the best result in the indicated learning epoch.

2.3. Construction of the RS2 Model

Due to their simplicity, multivariate linear regression models are commonly used in the prediction of agricultural crop yields [65]. MLR models the combination of a dependent trait and two or more independent traits by creating a linear equation to the observed data [66]. The value of the explanatory variable (Y) is related to the value of the explanatory variables (X) according to Equation (1) [62]:

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + ... + b_p \cdot X_p + \varepsilon, \quad (1)$$

where Y is the dependent variable (explained variable), X_1, X_2 ... X_p represents the independent variables (explanatory variable), b_0, b_1, b_2 ... b_p represents equation parameters and ε denotes the random component (rest of the model).

For the purpose of this work, an MLR model (stepwise progressive) was built based on the explanatory variables presented in Table 1. The procedure for building the RS2 model was similar to that for the N2 model. The computational analysis took eighteen steps. All the steps involved in building and verifying the RS2 model were performed, as with the N2 models, in Statistica v7.1 (TIBCO Software Inc., Palo Alto, CA, USA).

2.4. Evaluation Criteria for the N2 and RS2 Models

Six performance criteria (global relative error of model approximation (RAE), root mean square error (RMS), mean absolute error (MAE), mean absolute percentage error (MAPE), maximum error determined for the whole model (MAX) and maximum percentage error (MAXP)) were used to evaluate the resulting predictive models [54]. In order to calculate the values of these errors, a set B was required, which was used to determine the difference between the predicted and observed values. The magnitudes of these errors were calculated from the equations below:

$$\mathrm{RAE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y'_i)^2}{\sum_{i=1}^{n}(y_i)^2}}, \tag{2}$$

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y'_i)^2}{n}} \tag{3}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - y'_i| \tag{4}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y'_i}{y_i}\right| \cdot 100\% \tag{5}$$

$$\mathrm{MAX} = \max_i \cdot |y_i - y'_i| \tag{6}$$

$$\mathrm{MAXP} = \max_i \left|\frac{y_i - y'_i}{y_i}\right| \cdot 100\% \tag{7}$$

where n is the number of observations, y_i is the actual values and y'_i is the predictive values obtained with the model.

2.5. Sensitivity Analysis of the Neural Network

The final stage in the construction of predictive models based on ANNs is sensitivity analysis of the neural network. This stage consists of differentiating the independent variables in terms of their influence on the dependent variable. The method of calculating and interpreting the results obtained from the sensitivity analysis has been discussed in previous works by the authors of this paper [17,18,54].

3. Results

3.1. Overall Assessment of the Predictive Quality of the N2 and RS2 Models

Building predictive models based on artificial neural networks requires partitioning of the dataset. In this paper, the dataset was divided into three subsets: learning, validation and testing. The learning set contains both input and output data as patterns of valid signals. Based on these, the learning algorithm confronts the actual behavior of the network. The validation set is used indirectly in the learning process. Its task is to participate in periodic validation during the learning of the model, which prevents the occurrence of network overfitting [67]. The test set, on the other hand, is intended for one-time control after the training is completed. In general, this procedure is aimed at checking whether there was any loss of network generalization ability during training that may have resulted from coincidence, despite cyclical internal validation. Table 2 shows the error sizes and the quality of each subset. From this, we can observe that the learning set had the smallest error size (0.0556). The test set, on the other hand, was characterized by the largest error among the analyzed subsets. The error value for this subset was 0.0679. A different relationship can be observed in the case of subset quality. The test set was characterized by the highest value of this feature (0.4311), and the learning set by the lowest (0.3576). The validation set took an intermediate position (0.3645).

The N2 model was learned using two methods: back-propagation of the error and the coupled gradients method. The point at which there is an increase in error for the

validation set is the signal to stop training the neural network and recover the best weight from the epoch that preceded the start of the error increase. In the case at hand, the first learning method lasted for 100 epochs, and by continuing the learning process with the coupled gradients method, it was possible to obtain the best result, which was achieved at 110 epochs.

The obtained N2 model predicting pea seed yield was characterized by a relatively low mean error value, which was about 0.015 (Table 3). The model also obtained a small mean absolute error, which did not exceed the value of 0.305. In turn, the correlation coefficient reached a relatively high value (0.936). In developing predictive models, it is important that the model built is characterized by low error magnitudes and a high correlation coefficient value, as only such a model will be able to accurately predict the dependent variable.

Table 3. Qualitative measures of the N2 model.

Quality Parameter	Value
Average	4.504
Standard deviation	1.106
Average error	0.015
Error deviation	0.389
Average absolute error	0.305
Deviation quotient	0.352
Correlation coefficient r	0.936

Multivariate linear regression analysis showed that the input variables that were not statistically significant ($\alpha = 0.05$) were the date of harvest (HAR), the date of plant technical maturity (TECH_M) and the dose of potassium brought into the soil with mineral fertilizer (K2O_F).

Based on the results in Table 4, the form of the MLR equation was determined:

$$\begin{aligned} YIELD = {} & 0.215 \times P2O5\_C - 0.23 \times N\_F - 0.089 \times P\_EMER + 0.123\, INI\_MA - 0.007 \times TEMP - 0.007 \times RAIN - \\ & 0.003 \times SUN + 0.022 \times P2O5\_F + 0.481 \times PH + 0.169 \times K2O\_C - 0.040 \times FLOWE - 0.150 \times MGO\_C - 0.031 \times \\ & GEN + 0.005 \times P\_HIG + 0.034 \times WEGW \end{aligned} \quad (8)$$

3.2. Evaluation of Neural Network Sensitivity Analysis

The purpose of sensitivity analysis is to identify the independent variables that most influenced the dependent trait, pea seed yield. Based on the study, it can be observed that the onset of pea plant maturity (INI_MA) influenced the yield to the greatest extent (Table 5). This feature was ranked 1, and not including it in the N2 models would increase the cumulative error by a factor of 2378. The feature that received a rank of 2 was harvest date (HAR). Not including this variable in the model would increase the cumulative error by about 1.677 times. The variables ranked 3 and 4 were rainfall (RAIN) and mean air temperature (TEMP) calculated from sowing to 14 July. The absence of these variables in the N2 models would increase the cumulative error by 1.575 and 1.471 times, respectively.

Figure 2 shows a scatter plot of observed versus predicted values. From it, it can be concluded that the N2 model was characterized by a good level of prediction of pea seed yield, as evidenced by the relatively high value of the coefficient of determination (R^2), which was about 0.84. A much lower value of this indicator was obtained for the RS2 model (Figure 3). The R^2 coefficient did not exceed a value of 0.58, indicating that the response of the model is strongly discrepant with the observed values. The resulting model has virtually no ability to adequately represent the relationships characteristic of the issue under consideration.

Table 4. Results of multiple linear regression (MLR) analysis.

Factor	MLR: r = 0.7656 R^2 = 0.5788 Standard Error of Estimate = 0.7184					
	Beta	**Standard Error Beta**	**b**	**Standard Error b**	***p***	**Significance**
Free Term	–	–	−2.207	2.018	0.274282	–
HAR	−0.086	0.136	−0.010	0.017	0.526117	–
P2O5_C	0.177	0.026	0.215	0.031	0.000000	+
N_F	−0.1166	0.030	−0.012	0.003	0.000108	+
P_EMER	−0.480	0.046	−0.089	0.009	0.000000	+
INI_MA	1.027	0.124	0.123	0.015	0.000000	+
TEMP	0.398	0.080	0.283	0.057	0.000001	+
RAIN	−0.343	0.030	−0.007	0.001	0.000000	+
SUN	−0.225	0.029	−0.003	0.000	0.000000	+
P2O5_F	0.370	0.040	0.022	0.002	0.000000	+
PH	0.209	0.029	0.481	0.068	0.000000	+
K2O_C	0.143	0.029	0.169	0.035	0.000001	+
FLOWE	−0.199	0.040	−0.040	0.008	0.000001	+
MGO_C	−0.144	0.031	−0.150	0.032	0.000004	+
GEN	−0.089	0.021	−0.031	0.007	0.000017	+
P_HIG	0.077	0.030	0.005	0.002	0.010286	+
TECH_M	−0.200	0.114	−0.024	0.013	0.078386	–
WEGW	0.358	0.167	0.034	0.0158	0.032558	+
K2O_F	0.067	0.038	0.003	0.002	0.081382	–

Determination of the level of statistical significance: – non-significant; + significant for $\alpha = 0.05$.

Table 5. Results of the neural network sensitivity analysis.

Variable	Quotient	Rank
INI_MA	2.378	1
HAR	1.677	2
RAIN	1.575	3
TEMP	1.471	4
P_EMER	1.468	5
MGO_C	1.395	6
SOWI	1.387	7
K2O_C	1.356	8
P2O5_F	1.333	9
WEGE	1.261	10
P_HIG	1.170	11
PH	1.136	12
TECH_M	1.129	13
K2O_F	1.112	14
P2O5_C	1.110	15
GEN	1.079	16
SUN	1.052	17
FLOWE	1.052	18
N-F	1.045	19

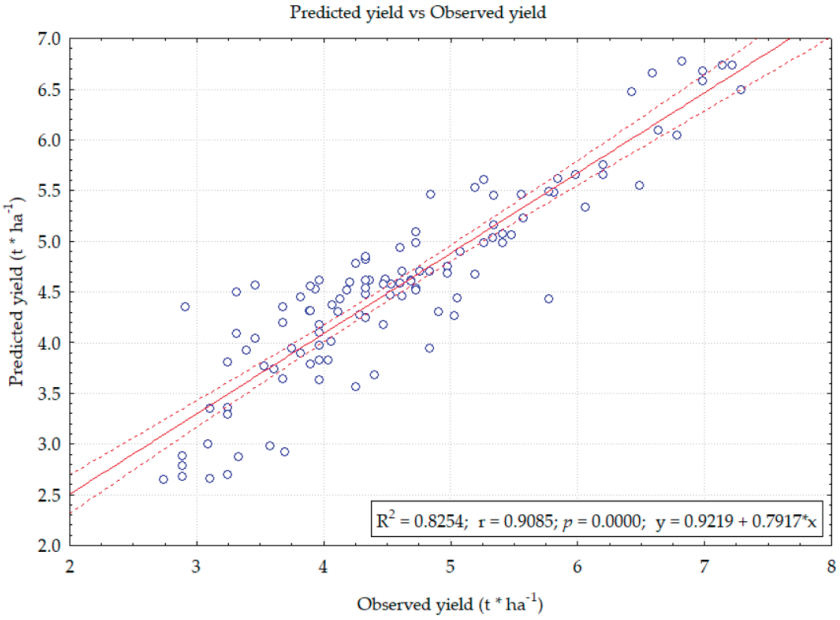


Figure 2. Correlation diagram of observed values against predicted values for the N2 model.

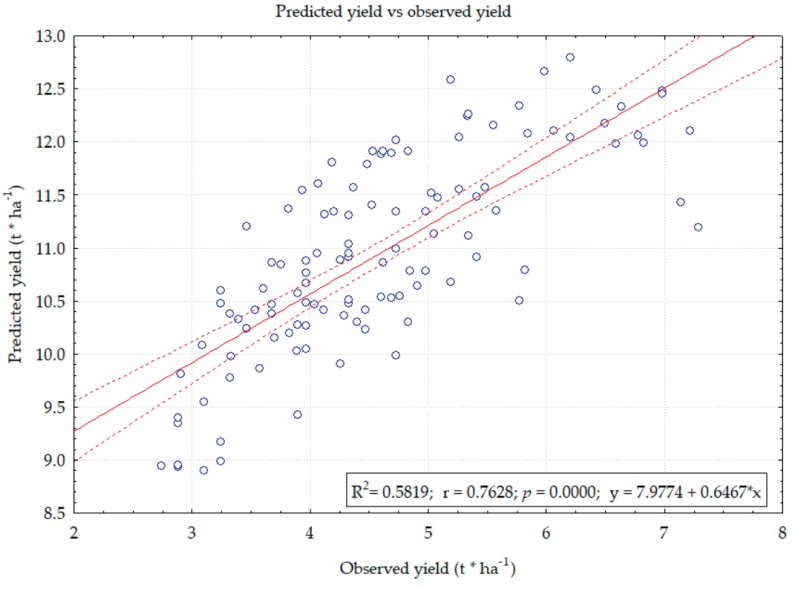


Figure 3. Correlation diagram of observed values against predicted values for the RS2 model.

The sensitivity analysis of the N2 model shows that some of the most important variables affecting pea yield were the date of onset of plant maturity (INI_MA) and the date of harvest (HAR). The relationship of these variables is shown in Figure 4. The plants reaching the maturity stage results in low seed yield. The same is true for early harvesting, which can also result in low yields. Higher yields were achievable when plants reached the onset of maturity later and when the harvest date was later.

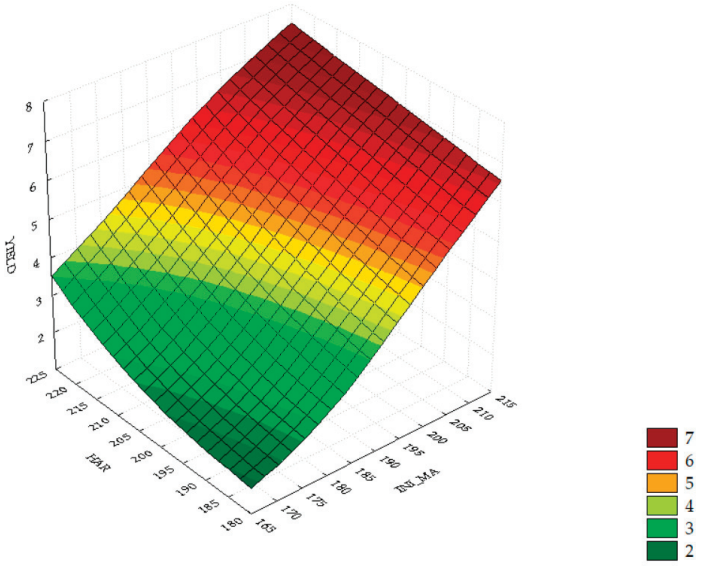


Figure 4. Response surface for yield size and two variables: HAR (harvest date) and INI_MA (start of crop maturity).

The relationship between pea yield and seed harvest date (HAR) and mean daily air temperature (TEMP) is shown in Figure 5, from which we can observe that the yield increased with increasing mean daily air temperature. A later harvest date also contributed to higher yields. At low TEMP, harvesting too early resulted in low crop efficiency. In addition, it can be observed from Figure 5 that temperature was the characteristic that most determined harvest date. An average daily air temperature of 18 °C allowed the seeds to be harvested around 29 June (180 days from the beginning of the year).

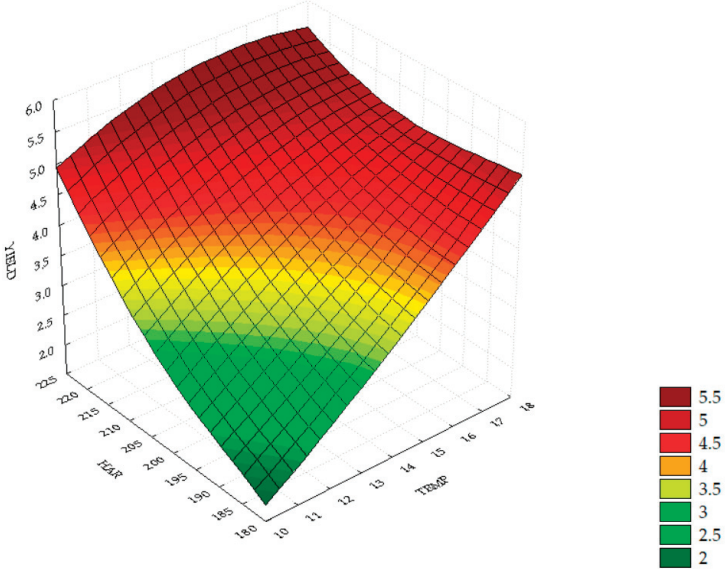


Figure 5. Response surface for yield size and two variables: HAR (harvest date) and TEMP (mean daily air temperature).

Based on the sensitivity analysis of the neural network, it was possible to determine the relationship of the independent variables of onset of maturity (INI_MA) and mean daily temperature (TEMP) in relation to seed yield (Figure 6). Plants yielded best when they reached the onset of maturity stage later (215 days) and when the average air temperature exceeded 17 °C. When peas reached the onset of maturity stage at 165 days (counted from the beginning of the year), plants were characterized by low yields (about 2 t·ha^{-1}). An increase in the TEMP trait contributed to an increase in plant yield. However, when the onset of maturity was reached early, this increase was insignificant and pea seed yield did not exceed 3 t·ha^{-1}.

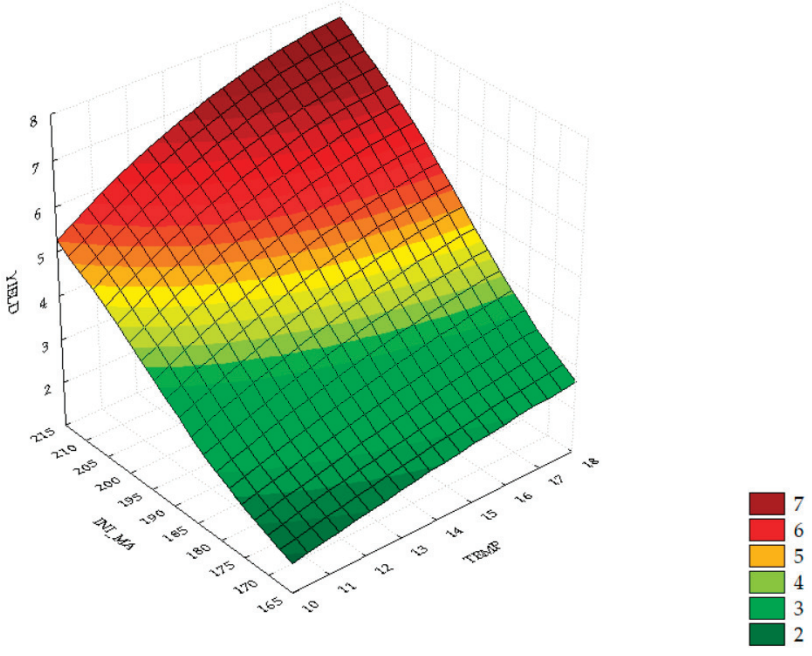


Figure 6. Response surface for yield size and two variables: INI_MA (start of crop maturity) and TEMP (mean daily air temperature).

3.3. Comparative Analysis of N2 and RS2 Models Based on Model Evaluation Criteria

The constructed models were verified for their validity. For this purpose, basic quality criteria were used, the values of which are shown in Table 6. The N2 model achieved a root mean square error (RAE) of 0.094 and the maximum percentage error (MAPE) was around 7.98. Much larger error values were obtained for the RS2 model. The RMS error was determined to be 6.401 and the MAPE reached a value of 148.585. Such high errors make the multiple linear regression method an unsuitable tool for forecasting pea seed yield. The strongly non-linear relationships affecting pea yield cause the linear method to mispredict the dependent variable.

Table 6. The quality of the generated neural models.

Error Type	N2 Model	RS2 Model
RAE	0.094	1.361
RMS	0.443	6.401
MAE	0.347	6.361
MAPE	7.976	148.585
MAX	1.398	7.739
MAXP	48.050	237.384

4. Discussion

Yield prediction methods have been used extensively in a number of works and have built models predicting the yields of maize, potato, winter wheat and orchard fruit, among others [68–73]. Specialized equipment, such as drones equipped with multispectral cameras, has been used to build some models in order to obtain information on crop characteristics. However, these devices can often be very expensive, and it is additionally necessary to have adequate knowledge of their operation. Consequently, models based on field imaging can be very difficult for agricultural producers and their application in agricultural practice may be limited. The proposed N2 and RS2 models were built using weather, agronomic and phytophenological data. Analysis of the prediction quality of the models showed that multiple linear regression was ineffective in estimating pea seed yield under the experimental conditions. The obtained MAPE error of 148.585 significantly exceeded the prediction accuracy threshold, disqualifying the obtained model as a suitable tool in yield prediction. According to Peng et al. [15], a predictive model that achieves an MAPE error greater than 30% is characterized by a poor representation of the predicted and observed values of the compound and should therefore be discarded. When MAPE is <10%, the model exhibits an excellent degree of fit. Such a low mean absolute percentage error was obtained for the N2 model based on ANNs. This model had an MAPE error of 7.976 (Table 6). It should be noted, however, that there is no appropriate comparative method for ANNs, and such models are therefore mainly compared with classical regression analyses worldwide. Such comparisons have been made, among others, by Kumari et al. [74], who predicted the yield of Indian nickel (*Cajanus cajan*—a bean crop) using a two-layer feed-forward neural network and an MLR model. The study was conducted in the Varanasi region (India) and the input data used were for the 1985–1986 and 2011–2012 periods. Five weather characteristics were used to build the models, i.e., minimum and maximum temperature, rainfall, and maximum and minimum relative humidity. The study conducted showed that the ANN model outperformed the MLR model in the prediction of Indian nickel yield. The RMS error for the ANN model was 299.93 kg·ha^{-1}, while the MLR model had an error magnitude of 884.02 kg·ha^{-1}. Artificial neural networks and multiple regression models were also used for the prediction of crescent beans (*Phaseolus lunatus* L.) [75]. The study was conducted in the northeastern part of Brazil, and independent variables used included the date of flowering onset, the date of pod maturity onset and pod length. The analysis showed that the MLP model forecasted yield more accurately compared to the MLR model, as evidenced by the MAPE, RMS and MAE error values. These error values were 1.701, 0.565 and 0.425 for the MLP model and 6.458, 0.828 and 0.690 for the MLR model, respectively. The effectiveness of feed-forward neural networks in predicting the yield of oilseed rape and mustard grown in northeast India was also demonstrated by Kakati et al. [5]. The ANN model predicted crop yield for the Dhubri region with an RMS error of 11.3 and an R^2 value of 0.976, while the stepwise multiple linear regression (SMLR) model had an R^2 value of 0.756 and an RMS error of 65.4 Ang et al. [76] investigated the feasibility of using different models, including MLR models and DNNs (deep neural networks), to predict the yield of oil palm grown in the state of Pahang (Malaysia). The deep neural network consisted of three hidden layers: the first layer contained 256 neurons, the second layer contained 480 neurons and the third layer contained 256 neurons. The model accurately predicted yield with an RMS error of 2.92 and an MAPE error of 0.09.

In contrast, the MLR model had RMS and MAPE errors of 6.20 and 0.7, respectively. In addition, the R^2 coefficient of determination was 0.91 for the DNN model and 0.49 for the MLR model.

Other machine learning techniques are also effective in crop yield prediction, surpassing the quality of predictions made by classical MLR models. This is confirmed by the study by Sun et al. [77]. The authors used the random forest (RF) method to predict the yield of winter wheat grown in China. The study covered the years 2014–2018, and the integration of satellite, weather and geographical data was used to build the model. The average RMS error for the MLR model was 1229.97 and the coefficient of determination was 0.73. These results are significantly different from those obtained by the RF model. The mean RMS error for this model was more than 2.5 times smaller than that of the MLR model (465.32), and the R^2 coefficient was equal to 0.85. Zhao et al. [78] obtained similar results by investigating the ability of RF and MLR models in estimating the yield of winter wheat grown in the North China Plain. The researchers analyzed the applicability of these models at different periods of plant development. The results showed that the RF and MLR models obtained the best results for the period from the beginning of grain filling to the milk stage. However, the MLR model had a larger RMS error (778.0) and r-ratio (0.79) compared to the RF model, for which these parameters were 683.0 and 0.86, respectively.

A comparison of actual and predicted values (Figures 2 and 3) shows that the coefficient of determination for the N2 model (R^2 = 0.8254) was at a higher level than it was for the RS2 model (R^2 = 0.5819). These results show that the RS2 model had much weaker predictive properties with respect to the N2 model. However, multiple linear regression models, as already mentioned, are commonly used in yield prediction. This method has many limitations, such as the assumption of a linear relationship between the exploratory variable and the explanatory variable [17]. If the relationship between these variables is non-linear, the regression model will tend to perform poorly. In addition, linear regression assumes that the input variables are not correlated with each other. If there is multicollinearity in the dataset, then this assumption is violated, and the performance of the regression model will be reduced. Additionally, linear models are assumed to have a constant variance under error conditions (homoskedasticity), which is often not true. Another problem that hinders proper prediction using MLR models is the presence of outlier points, which significantly affect its performance [65]. On the other hand, ANN models, including those with MLP architecture, are capable of predicting agricultural crop yields even in the case of strong non-linear relationships between the independent variables and the dependent variable. In addition, the main function of neural networks is to identify hidden patterns and features in the dataset. This activity is made possible by the two most important parts of the network, i.e., the activation function and the weighting parameters [5]. From the research carried out, all the variables tested in the study are characterized by non-linear patterns. Therefore, the RS2 model could not properly estimate the yield. Our research shows that the choice of method for creating the model is a kind of compromise that requires its creator to have a very thorough knowledge of the test object. This ensures that accurate yield predictions are obtained. However, ANNs are also not free from certain limitations. One of the biggest is that neural network models require a lot of, sometimes very specific, input data to train [79]. Acquiring such data can often be cumbersome, and for regions where observational records are lacking, obtaining short-term predictions is significantly difficult [80]. In addition, the appropriate selection of independent characteristics must be supported by extensive knowledge of the issue being modeled [9]. In the present study, three categories of independent variables were used: weather data, agronomic information and phytophenological data. These variables are publicly available and the results of analyses involving these data are easy to interpret.

The inclusion of climatic conditions when modeling agronomic issues is an important element when seeking to obtain a high-performance model. The inclusion of information related to air temperature, sunshine and precipitation during the growing season is reasonable, as these factors strongly determine plant growth and development [81]. Plant

productivity is significantly affected by the temperature distribution during the growing season. However, the influence of this factor is reduced when there is an adequate water supply to the plants [82]. This assumption was fulfilled in this work because data from typical years, without weather anomalies that affect the quality of the models, were selected for analysis. According to a study by Aubakirov et al. [19], the amount of precipitation and temperature had the greatest impact on the multiplicity of yield of wheat grown in the North Kazakhstan region. Consideration of these data by the authors made it possible to build a back-propagation artificial neural network model that predicted wheat yield multiplicity with an MAPE error of 12.02 and an RMS error of 3.368. Similar observations were reached by Nedbała et al. [59], who identified key meteorological factors affecting soybean (*Glycine max* [L.] Merrill) yield and harvest date. A sensitivity analysis of the MLP network found that the variable that most influenced soybean seed yield was air temperature in the second ten days of May. In contrast, the variables that most influenced soybean harvest date were rainfall totals in the first ten days of June and the first ten days of August. The inclusion of environmental variables in modeling is also highlighted by Vojnov et al. [60], who attribute significant effects on plant parameters and on the performance of ANN models to these data.

Many scientific disciplines and the agricultural industry commonly use phytophenological periods of plant development. Among other things, they are helpful in determining when to apply inputs. These phases have been standardized for ease of communication between agronomists, naturalists, breeders of new agricultural varieties, etc. [7]. In modeling agricultural crop yields, this information is exploited to enhance the efficiency of the models built. As reported by Shamsabadi et al. [20], the inclusion of data such as number of days to emergence, days to maturity and number of days to flowering in the model significantly affected the performance of the MLP model. The model predicted the seed yield of hybrid wheat that was grown in the northern part of Iran. In the present study, empirical data in the form of phytophenological periods were also used, which allowed for the construction of an N2 model with an MAPE error of 7.976 (Table 6) and a correlation coefficient of 0.936 (Table 3).

A sensitivity analysis of the N2 model showed that the independent variables with the greatest impact on pea seed yield under the conditions tested were the date of onset of maturity and the harvest date. These traits received a rank of 1 and 2, respectively. From Figures 4 and 6, we can observe that the later occurrence of onset of the maturity phase resulted in an increase in yield. The rate of transition of plants from one phenological phase to another depends on weather conditions [83]. Peas at the onset of maturity tolerate lower temperatures than those at the flowering stage. The length of this period is determined by average and minimum daily temperatures. Lower temperatures during this period favor the accumulation of starch in the seeds, thereby increasing yield [84]. Pea harvesting should be optimized based on weather conditions and seed moisture content. Figures 4 and 5 show that harvesting at a later date has a positive effect on yield. Pea varieties grown in Poland are characterized by uneven maturation. At the beginning, pods located at the lower part of the plant ripen, and pods located in the higher parts of the plant ripen at the end [85]. Harvesting too early may result in the upper pods not reaching the stage of technical maturity and not accumulating enough starch, proteins or other assimilates; thus, the weight of 1000 seeds may be lower than that of the seeds placed in the lower pods. It should be remembered, however, that harvesting peas from the field too late may result in a decrease in yield due to lodging of the plants and pod breakage [86].

The sum of precipitation and mean air temperature are the variables ranked 3 and 4 in the sensitivity analysis of the network. Weather conditions during the growing season of plants are one of the most important environmental factors affecting plant growth and development. Temperature and the amount of rainfall vary the yield of peas from one crop year to another [87]. A study conducted by Pandey et al. [88] proved that water deficiency in pea cultivation reduces the photosynthetic efficiency of plants, disrupts nutrient transport and affects structural changes in leaves due to the presence of reactive oxygen species.

These changes ultimately lead to a decrease in plant yield. Therefore, the optimum rainfall over the growing season of peas should be 280 mm on light soil, 250 mm on medium soil and 22 mm on compact soils [85].

The results obtained from the study show that the N2 model can be a promising information tool due to its accurate prediction of seed yield in pea. Our study further confirmed the hypothesis that the right approach to independent trait selection supports the process of identifying the most important variables affecting yield [76]. Such models may be of interest to breeders of new pea varieties. Knowing the variables with the greatest impact on yield, it is possible to improve new varieties by optimizing certain time intervals in the phenology so as to obtain a high final yield. An opportunity for the advancement of predictive models is the possibility of using new types of data, such as ground-based phenological imaging, or the use of the same dataset but of higher quality, such as high- or very high-resolution spectral data [7].

5. Conclusions

With climate change and an increasing global population, there is a growing need to better predict the yield of agricultural crops, as well as the correct way for farmers to grow their crops. The analyses conducted show that an artificial neural network model is a useful tool in predicting pea yield 20 days before harvest. The N2 model accurately predicted the independent variable with a correlation coefficient $r > 0.9$ and MAPE and RMS values of 7.976 and 0.443, respectively. At the same time, it was proven that the RS2 model is not able to accurately estimate pea yield. The model had an MAPE error of 148.585. Therefore, the potential practical application of this model in pea production is not possible. The choice of modeling technique is crucial in accurately estimating yield. Furthermore, modeling pea yields a few weeks before harvest carries promising possibilities for practical application.

Pre-harvest yield forecasting is a valuable source of information that is particularly relevant for farmers, agronomists and decision makers. Further research will focus on comparative analysis of the ANN model against other machine learning techniques such as RBF.

Author Contributions: Conceptualization, P.H., M.P. and G.N.; methodology, P.H., M.P. and G.N.; validation, M.P. and G.N.; formal analysis, M.P.; investigation, P.H.; resources, P.H.; data curation, P.H.; writing—original draft preparation, P.H.; writing—review and editing, P.H., M.P. and G.N.; supervision, G.N.; project administration, M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable. Data Availability Statement The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ANN—artificial neural networks; COBORU—Research Center for Cultivar Testing; CNN—convolutional neural networks; DNN—deep neural network; FLOWE—number of days from 1 January to the beginning of flowering; GEN—general variety of peas; HAR—number of days from 1 January to the date of harvesting; INI_MA—number of days from 1 January to onset of maturity; kNN—k nearest neighbors; K20_C—K_2O content in the soil; K_2O—potassium oxide; K2O_F—Total potassium from mineral fertilizers; kg—kilogram; MAE—mean absolute error; MAPE—mean absolute percentage error; MAX—maximum error determined for the whole model; MAXP—maximum percentage error; MgO—magnesium oxide; MGO_C—MgO content in the soil; ML—machine learning; MLP—multilayer perceptron; MLR—multiple linear regression; n—number of observations; N_F—total nitrogen from mineral fertilizers; N2—built its own neural network model; P_EMER—number of days from 1 January

to the beginning of plant emergence; P_HIG—plant height; P_2O_5—phosphorus(V) oxide; P2O5_C—P_2O_5 content in the soil; P2O5_F—total phosphorus from mineral fertilizers; PH—Soil pH; PROT—Percentage of protein in pea seeds; RAE—global relative error of model approximation; RAIN—total rainfall from sowing date to July 14; RF—random forest; RMS—root mean square error; RS2—built its own linear regression model; SOWI—number of days from 1 January to sowing date; SUN—total sunshine from sowing date to 14 July; SVM—support vector machines; TECH_M—number of days from 1 January to technical maturity; TEMP—average air temperature from sowing date to July 14; WEGW—number of plant growing days; y'_i—predictive values, obtained with the model; y_i—actual values.

References

1. Szparaga, A.; Kuboń, M.; Kocira, S.; Czerwińska, E.; Pawłowska, A.; Hara, P.; Kobus, Z.; Kwaśniewski, D. Towards sustainable agriculture-agronomic and economic effects of biostimulant use in common bean cultivation. *Sustainability* **2019**, *11*, 4575. [CrossRef]
2. Rokhafrouz, M.; Latifi, H.; Abkar, A.A.; Wojciechowski, T.; Czechlowski, M.; Naieni, A.S.; Maghsoudi, Y.; Niedbała, G. Simplified and Hybrid Remote Sensing-Based Delineation of Management Zones for Nitrogen Variable Rate Application in Wheat. *Agriculture* **2021**, *11*, 1104. [CrossRef]
3. Kukal, M.S.; Irmak, S. Climate-Driven Crop Yield and Yield Variability and Climate Change Impacts on the U.S. Great Plains Agricultural Production. *Sci. Rep.* **2018**, *8*, 3450. [CrossRef] [PubMed]
4. Khavse, R.; Singh, R.; Manikandan, N.; Chaudhary, J. Influence of Temperature on Rapeseed-Mustard Yield at Selected Locations in Chhattisgarh State. *Curr. World Environ.* **2014**, *9*, 1034–1036. [CrossRef]
5. Kakati, N.; Deka, R.L.; Das, P.; Goswami, J.; Khanikar, P.G.; Saikia, H. Forecasting yield of rapeseed and mustard using multiple linear regression and ANN techniques in the Brahmaputra valley of Assam, North East India. *Theor. Appl. Climatol.* **2022**, *150*, 1201–1215. [CrossRef]
6. Chergui, N.; Kechadi, M.-T.; McDonnell, M. The Impact of Data Analytics in Digital Agriculture: A Review. In Proceedings of the 2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies" (OCTA), Tunis, Tunisia, 6–8 February 2020; IEEE: New York, NY, USA, 2020; pp. 1–13.
7. Niedbała, G.; Kurek, J.; Świderski, B.; Wojciechowski, T.; Antoniuk, I.; Bobran, K. Prediction of Blueberry (*Vaccinium corymbosum* L.) Yield Based on Artificial Intelligence Methods. *Agriculture* **2022**, *12*, 2089. [CrossRef]
8. He, L.; Fang, W.; Zhao, G.; Wu, Z.; Fu, L.; Li, R.; Majeed, Y.; Dhupia, J. Fruit yield prediction and estimation in orchards: A state-of-the-art comprehensive review for both direct and indirect methods. *Comput. Electron. Agric.* **2022**, *195*, 106812. [CrossRef]
9. Hara, P.; Piekutowska, M.; Niedbała, G. Selection of Independent Variables for Crop Yield Prediction Using Artificial Neural Network Models with Remote Sensing Data. *Land* **2021**, *10*, 609. [CrossRef]
10. Yildirim, T.; Moriasi, D.N.; Starks, P.J.; Chakraborty, D. Using Artificial Neural Network (ANN) for Short-Range Prediction of Cotton Yield in Data-Scarce Regions. *Agronomy* **2022**, *12*, 828. [CrossRef]
11. Ali, A.; Rondelli, V.; Martelli, R.; Falsone, G.; Lupia, F.; Barbanti, L. Management Zones Delineation through Clustering Techniques Based on Soils Traits, NDVI Data, and Multiple Year Crop Yields. *Agriculture* **2022**, *12*, 231. [CrossRef]
12. Wang, J.; Si, H.; Gao, Z.; Shi, L. Winter Wheat Yield Prediction Using an LSTM Model from MODIS LAI Products. *Agriculture* **2022**, *12*, 1707. [CrossRef]
13. Johnson, M.D.; Hsieh, W.W.; Cannon, A.J.; Davidson, A.; Bédard, F. Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. *Agric. For. Meteorol.* **2016**, *218–219*, 74–84. [CrossRef]
14. Gonzalez-Sanchez, A.; Frausto-Solis, J.; Ojeda-Bustamante, W. Attribute Selection Impact on Linear and Nonlinear Regression Models for Crop Yield Prediction. *Sci. World J.* **2014**, *2014*, 509429. [CrossRef] [PubMed]
15. Peng, J.; Kim, M.; Kim, Y.; Jo, M.; Kim, B.; Sung, K.; Lv, S. Constructing Italian ryegrass yield prediction model based on climatic data by locations in South Korea. *Grassl. Sci.* **2017**, *63*, 184–195. [CrossRef]
16. Niedbala, G.; Kozlowski, R.J. Application of artificial neural networks for multi-criteria yield prediction of winter wheat. *J. Agric. Sci. Technol.* **2019**, *21*, 51–61.
17. Piekutowska, M.; Niedbała, G.; Piskier, T.; Lenartowicz, T.; Pilarski, K.; Wojciechowski, T.; Pilarska, A.A.; Czechowska-Kosacka, A. The Application of Multiple Linear Regression and Artificial Neural Network Models for Yield Prediction of Very Early Potato Cultivars before Harvest. *Agronomy* **2021**, *11*, 885. [CrossRef]
18. Niedbała, G.; Wróbel, B.; Piekutowska, M.; Zielewicz, W.; Paszkiewicz-Jasińska, A.; Wojciechowski, T.; Niazian, M. Application of Artificial Neural Networks Sensitivity Analysis for the Pre-Identification of Highly Significant Factors Influencing the Yield and Digestibility of Grassland Sward in the Climatic Conditions of Central Poland. *Agronomy* **2022**, *12*, 1133. [CrossRef]
19. Aubakirova, G.; Ivel, V.; Gerassimova, Y.; Moldakhmetov, S.; Petrov, P. Application of artificial neural network for wheat yield forecasting. *East. Eur. J. Enterp. Technol.* **2022**, *3*, 31–39. [CrossRef]
20. Shamsabadi, E.E.; Sabouri, H.; Soughi, H.; Sajadi, S.J. Using of Molecular Markers in Prediction of Wheat (*Triticum aestivum* L.) Hybrid Grain Yield Based on Artificial Intelligence Methods and Multivariate Statistics. *Russ. J. Genet.* **2022**, *58*, 603–611. [CrossRef]

21. Khaki, S.; Wang, L.; Archontoulis, S.V. A CNN-RNN Framework for Crop Yield Prediction. *Front. Plant Sci.* **2020**, *10*, 1750. [CrossRef]
22. Sabatino, L.; D'Anna, F.; Iapichino, G.; Moncada, A.; D'Anna, E.; De Pasquale, C. Interactive Effects of Genotype and Molybdenum Supply on Yield and Overall Fruit Quality of Tomato. *Front. Plant Sci.* **2019**, *9*, 1922. [CrossRef] [PubMed]
23. Awad, M. Toward Precision in Crop Yield Estimation Using Remote Sensing and Optimization Techniques. *Agriculture* **2019**, *9*, 54. [CrossRef]
24. Nazir, A.; Ullah, S.; Saqib, Z.A.; Abbas, A.; Ali, A.; Iqbal, M.S.; Hussain, K.; Shakir, M.; Shah, M.; Butt, M.U. Estimation and Forecasting of Rice Yield Using Phenology-Based Algorithm and Linear Regression Model on Sentinel-II Satellite Data. *Agriculture* **2021**, *11*, 1026. [CrossRef]
25. Feizi, H.; Sattari, M.T.; Prasad, R.; Apaydin, H. Comparative analysis of deep and machine learning approaches for daily carbon monoxide pollutant concentration estimation. *Int. J. Environ. Sci. Technol.* **2023**, *20*, 1753–1768. [CrossRef]
26. Meerasri, J.; Sothornvit, R. Artificial neural networks (ANNs) and multiple linear regression (MLR) for prediction of moisture content for coated pineapple cubes. *Case Stud. Therm. Eng.* **2022**, *33*, 101942. [CrossRef]
27. Ge, J.; Zhao, L.; Yu, Z.; Liu, H.; Zhang, L.; Gong, X.; Sun, H. Prediction of Greenhouse Tomato Crop Evapotranspiration Using XGBoost Machine Learning Model. *Plants* **2022**, *11*, 1923. [CrossRef]
28. Niedbała, G. Application of artificial neural networks for multi-criteria yield prediction ofwinter rapeseed. *Sustainability* **2019**, *11*, 533. [CrossRef]
29. Wojciechowski, T.; Niedbala, G.; Czechlowski, M.; Nawrocka, J.R.; Piechnik, L.; Niemann, J. Rapeseed seeds quality classification with usage of VIS-NIR fiber optic probe and artificial neural networks. In Proceedings of the 2016 International Conference on Optoelectronics and Image Processing (ICOIP), Warsaw, Poland, 10–12 June 2016; IEEE: Warsaw, Poland, 2016; pp. 44–48.
30. van Klompenburg, T.; Kassahun, A.; Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **2020**, *177*, 105709. [CrossRef]
31. Ballesteros, R.; Intrigliolo, D.S.; Ortega, J.F.; Ramírez-Cuesta, J.M.; Buesa, I.; Moreno, M.A. Vineyard yield estimation by combining remote sensing, computer vision and artificial neural network techniques. *Precis. Agric.* **2020**, *21*, 1242–1262. [CrossRef]
32. Crusiol, L.G.T.; Sun, L.; Sibaldelli, R.N.R.; Junior, V.F.; Furlaneti, W.X.; Chen, R.; Sun, Z.; Wuyun, D.; Chen, Z.; Nanni, M.R.; et al. Strategies for monitoring within-field soybean yield using Sentinel-2 Vis-NIR-SWIR spectral bands and machine learning regression methods. *Precis. Agric.* **2022**, *23*, 1093–1123. [CrossRef]
33. Niedbała, G.; Kurasiak-Popowska, D.; Stuper-Szablewska, K.; Nawracała, J. Application of Artificial Neural Networks to Analyze the Concentration of Ferulic Acid, Deoxynivalenol, and Nivalenol in Winter Wheat Grain. *Agriculture* **2020**, *10*, 127. [CrossRef]
34. Chergui, N. Durum wheat yield forecasting using machine learning. *Artif. Intell. Agric.* **2022**, *6*, 156–166. [CrossRef]
35. Phan, P.; Chen, N.; Xu, L.; Dao, D.M.; Dang, D. NDVI Variation and Yield Prediction in Growing Season: A Case Study with Tea in Tanuyen Vietnam. *Atmosphere* **2021**, *12*, 962. [CrossRef]
36. Bouras, E.H.; Jarlan, L.; Er-Raki, S.; Balaghi, R.; Amazirh, A.; Richard, B.; Khabba, S. Cereal Yield Forecasting with Satellite Drought-Based Indices, Weather Data and Regional Climate Indices Using Machine Learning in Morocco. *Remote Sens.* **2021**, *13*, 3101. [CrossRef]
37. Taşan, S.; Cemek, B.; Taşan, M.; Cantürk, A. Estimation of eggplant yield with machine learning methods using spectral vegetation indices. *Comput. Electron. Agric.* **2022**, *202*, 107367. [CrossRef]
38. Jeevaganesh, R.; Harish, D.; Priya, B. A Machine Learning-based Approach for Crop Yield Prediction and Fertilizer Recommendation. In Proceedings of the 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 28–30 April 2022; IEEE: New York, NY, USA, 2022; pp. 1330–1334.
39. Tugrul, B.; Elfatimi, E.; Eryigit, R. Convolutional Neural Networks in Detection of Plant Leaf Diseases: A Review. *Agriculture* **2022**, *12*, 1192. [CrossRef]
40. Dayal, M.; Gupta, M.; Gupta, M.; Bara, A.R.; Chaubey, C. Introduction to Machine Learning Methods With Application in Agriculture. In *Applying Drone Technologies and Robotics for Agricultural Sustainability*; IGI Global: Hershey, PA, USA, 2023; pp. 184–203.
41. Dhillon, M.S.; Dahms, T.; Kuebert-Flock, C.; Rummler, T.; Arnault, J.; Steffan-Dewenter, I.; Ullmann, T. Integrating random forest and crop modeling improves the crop yield prediction of winter wheat and oil seed rape. *Front. Remote Sens.* **2023**, *3*, 1010978. [CrossRef]
42. Khalifani, S.; Darvishzadeh, R.; Azad, N.; Seyed Rahmani, R. Prediction of sunflower grain yield under normal and salinity stress by RBF, MLP and, CNN models. *Ind. Crops Prod.* **2022**, *189*, 115762. [CrossRef]
43. Maya Gopal, P.S.; Bhargavi, R. A novel approach for efficient crop yield prediction. *Comput. Electron. Agric.* **2019**, *165*, 104968. [CrossRef]
44. Tadeusiewicz, R. *Elementarne Wprowadzenie Do Techniki Sieci Neuronowych z Przykładowymi Programami*; Akademicka Oficyna Wydawnicza PLJ: Warsaw, Poland, 1998.
45. Li, X.; Hu, T.; Gong, P.; Du, S.; Chen, B.; Li, X.; Dai, Q. Mapping Essential Urban Land Use Categories in Beijing with a Fast Area of Interest (AOI)-Based Method. *Remote Sens.* **2021**, *13*, 477. [CrossRef]
46. Sabzi-Nojadeh, M.; Niedbała, G.; Younessi-Hamzekhanlu, M.; Aharizad, S.; Esmaeilpour, M.; Abdipour, M.; Kujawa, S.; Niazian, M. Modeling the Essential Oil and Trans-Anethole Yield of Fennel (*Foeniculum vulgare* Mill. var. vulgare) by Application Artificial Neural Network and Multiple Linear Regression Methods. *Agriculture* **2021**, *11*, 1191. [CrossRef]

47. Abrosimov, M.; Brovko, A. High Generalization Capability Artificial Neural Network Architecture Based on RBF-Network. In Proceedings of the ICIT 2019: Recent Research in Control Engineering and Decision Making, Saratov, Russia, 7–8 February 2019; Springer: Cham, Switzerland, 2019; pp. 67–78.
48. Liakos, K.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine Learning in Agriculture: A Review. *Sensors* **2018**, *18*, 2674. [CrossRef] [PubMed]
49. Bhojani, S.H.; Bhatt, N. Wheat crop yield prediction using new activation functions in neural network. *Neural Comput. Appl.* **2020**, *32*, 13941–13951. [CrossRef]
50. Bazrafshan, O.; Ehteram, M.; Dashti Latif, S.; Feng Huang, Y.; Yenn Teo, F.; Najah Ahmed, A.; El-Shafie, A. Predicting crop yields using a new robust Bayesian averaging model based on multiple hybrid ANFIS and MLP models. *Ain Shams Eng. J.* **2022**, *13*, 101724. [CrossRef]
51. Torsoni, G.B.; de Oliveira Aparecido, L.E.; dos Santos, G.M.; Chiquitto, A.G.; da Silva Cabral Moraes, J.R.; de Souza Rolim, G. Soybean yield prediction by machine learning and climate. *Theor. Appl. Climatol.* **2023**, *151*, 1709–1725. [CrossRef]
52. Soroush, F.; Ehteram, M.; Seifi, A. Uncertainty and spatial analysis in wheat yield prediction based on robust inclusive multiple models. *Environ. Sci. Pollut. Res.* **2022**, *30*, 20887–20906. [CrossRef] [PubMed]
53. Heng, S.Y.; Ridwan, W.M.; Kumar, P.; Ahmed, A.N.; Fai, C.M.; Birima, A.H.; El-Shafie, A. Artificial neural network model with different backpropagation algorithms and meteorological data for solar radiation prediction. *Sci. Rep.* **2022**, *12*, 10457. [CrossRef]
54. Hara, P.; Piekutowska, M.; Niedbała, G. Prediction of Protein Content in Pea (*Pisum sativum* L.) Seeds Using Artificial Neural Networks. *Agriculture* **2022**, *13*, 29. [CrossRef]
55. Research Centre for Cultivar Testing (COBORU). Available online: https://coboru.gov.pl/ (accessed on 20 October 2022).
56. Niedbała, G.; Tratwal, A.; Piekutowska, M.; Wojciechowski, T.; Uglis, J. A Framework for Financing Post-Registration Variety Testing System: A Case Study from Poland. *Agronomy* **2022**, *12*, 325. [CrossRef]
57. Zintegrowana Platforma Edukacyjna. Available online: https://zpe.gov.pl/a/cechy-klimatu-polski/DbdxuNIhI (accessed on 10 January 2023).
58. Wiatr, K. Rośliny rolnicze. In *Metodyka Badania Wartości Gospodarczej Odmian (WGO) Roślin Uprawnych*; Centralny Ośrodek Badania Odmian Roślin Uprawnych: Słupia Wielka, Poland, 1998.
59. Niedbała, G.; Kurasiak-Popowska, D.; Piekutowska, M.; Wojciechowski, T.; Kwiatek, M.; Nawracała, J. Application of Artificial Neural Network Sensitivity Analysis to Identify Key Determinants of Harvesting Date and Yield of Soybean (*Glycine max* [L.] Merrill) Cultivar Augusta. *Agriculture* **2022**, *12*, 754. [CrossRef]
60. Vojnov, B.; Jaćimović, G.; Šeremešić, S.; Pezo, L.; Lončar, B.; Krstić, Đ.; Vujić, S.; Ćupina, B. The Effects of Winter Cover Crops on Maize Yield and Crop Performance in Semiarid Conditions—Artificial Neural Network Approach. *Agronomy* **2022**, *12*, 2670. [CrossRef]
61. Geetha, M.C.S.; Elizabeth Shanthi, I. Forecasting the Crop Yield Production in Trichy District Using Fuzzy C-Means Algorithm and Multilayer Perceptron (MLP). *Int. J. Knowl. Syst. Sci.* **2020**, *11*, 83–98. [CrossRef]
62. Pentoś, K.; Mbah, J.T.; Pieczarka, K.; Niedbała, G.; Wojciechowski, T. Evaluation of Multiple Linear Regression and Machine Learning Approaches to Predict Soil Compaction and Shear Stress Based on Electrical Parameters. *Appl. Sci.* **2022**, *12*, 8791. [CrossRef]
63. Gorzelany, J.; Belcar, J.; Kuźniar, P.; Niedbała, G.; Pentoś, K. Modelling of Mechanical Properties of Fresh and Stored Fruit of Large Cranberry Using Multiple Linear Regression and Machine Learning. *Agriculture* **2022**, *12*, 200. [CrossRef]
64. Shankar, T.; Malik, G.C.; Banerjee, M.; Dutta, S.; Praharaj, S.; Lalichetti, S.; Mohanty, S.; Bhattacharyay, D.; Maitra, S.; Gaber, A.; et al. Prediction of the Effect of Nutrients on Plant Parameters of Rice by Artificial Neural Network. *Agronomy* **2022**, *12*, 2123. [CrossRef]
65. Khan, S.N.; Li, D.; Maimaitijiang, M. A Geographically Weighted Random Forest Approach to Predict Corn Yield in the US Corn Belt. *Remote Sens.* **2022**, *14*, 2843. [CrossRef]
66. Arroyo, Á.; Cambra, C.; Basurto, N.; Rad, C.; Navarro, M.; Herrero, Á. Regression Techniques to Predict the Growth of Potato Tubers. In Proceedings of the 17th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2022), Salamanca, Spain, 5–7 September 2022; Springer: Cham, Switzerland, 2023; pp. 217–225.
67. Tadeusiewicz, R.; Szaleniec, M. *Leksykon Sieci Neuronowych*; Fundacja na Rzecz Promocji Nauki Polskiej: Wrocław, Poland, 2015.
68. Zhang, Q.; Wang, K.; Han, Y.; Liu, Z.; Yang, F.; Wang, S.; Zhao, X.; Zhao, C. A crop variety yield prediction system based on variety yield data compensation. *Comput. Electron. Agric.* **2022**, *203*, 107460. [CrossRef]
69. Piekutowska, M.; Adamski, M.; Czechowska-Kosacka, A.; Wójcik Oliveira, K.; Niedbała, G.; Wojciechowski, T.; Czechlowski, M. Modeling methods of predicting potato yield—Examples and possibilities of application. *J. Res. Appl. Agric. Eng.* **2018**, *63*, 176.
70. Elbeltagi, A.; Zhang, L.; Deng, J.; Juma, A.; Wang, K. Modeling monthly crop coefficients of maize based on limited meteorological data: A case study in Nile Delta, Egypt. *Comput. Electron. Agric.* **2020**, *173*, 105368. [CrossRef]
71. Gené-Mola, J.; Gregorio, E.; Auat Cheein, F.; Guevara, J.; Llorens, J.; Sanz-Cortiella, R.; Escolà, A.; Rosell-Polo, J.R. Fruit detection, yield prediction and canopy geometric characterization using LiDAR with forced air flow. *Comput. Electron. Agric.* **2020**, *168*, 105121. [CrossRef]
72. Ronchetti, G.; Manfron, G.; Weissteiner, C.J.; Seguini, L.; Nisini Scacchiafichi, L.; Panarello, L.; Baruth, B. Remote sensing crop group-specific indicators to support regional yield forecasting in Europe. *Comput. Electron. Agric.* **2023**, *205*, 107633. [CrossRef]
73. Atamanyuk, I.; Havrysh, V.; Nitsenko, V.; Diachenko, O.; Tepliuk, M.; Chebakova, T.; Trofimova, H. Forecasting of Winter Wheat Yield: A Mathematical Model and Field Experiments. *Agriculture* **2022**, *13*, 41. [CrossRef]

74. Kumari, P.; Mishra, G.C.; Srivastava, C.P. Statistical models for forecasting pigeonpea yield in Varanasi region. *J. Agrometeorol.* **2016**, *18*, 306–310. [CrossRef]
75. Sousa, A.M.d.C.B.d.; Silva, V.B.d.; Lopes, Â.C.d.A.; Gomes, R.L.F.; Carvalho, L.C.B. Prediction of grain yield, adaptability, and stability in landrace varieties of lima bean (*Phaseolus lunatus* L.). *Crop Breed. Appl. Biotechnol.* **2020**, *20*, 1–7. [CrossRef]
76. Ang, Y.; Shafri, H.Z.M.; Lee, Y.P.; Bakar, S.A.; Abidin, H.; Mohd Junaidi, M.U.U.; Hashim, S.J.; Che'Ya, N.N.; Hassan, M.R.; Lim, H.S.; et al. Oil palm yield prediction across blocks from multi-source data using machine learning and deep learning. *Earth Sci. Inform.* **2022**, *15*, 2349–2367. [CrossRef]
77. Sun, Y.; Zhang, S.; Tao, F.; Aboelenein, R.; Amer, A. Improving Winter Wheat Yield Forecasting Based on Multi-Source Data and Machine Learning. *Agriculture* **2022**, *12*, 571. [CrossRef]
78. Zhao, Y.; Xiao, D.; Bai, H.; Tang, J.; Liu, D.L.; Qi, Y.; Shen, Y. The Prediction of Wheat Yield in the North China Plain by Coupling Crop Model with Machine Learning Algorithms. *Agriculture* **2022**, *13*, 99. [CrossRef]
79. Li, Y.; Guan, K.; Yu, A.; Peng, B.; Zhao, L.; Li, B.; Peng, J. Toward building a transparent statistical model for improving crop yield prediction: Modeling rainfed corn in the U.S. *Field Crops Res.* **2019**, *234*, 55–65. [CrossRef]
80. Filippi, P.; Whelan, B.M.; Vervoort, R.W.; Bishop, T.F.A. Mid-season empirical cotton yield forecasts at fine resolutions using large yield mapping datasets and diverse spatial covariates. *Agric. Syst.* **2020**, *184*, 102894. [CrossRef]
81. Ruan, G.; Li, X.; Yuan, F.; Cammarano, D.; Ata-UI-Karim, S.T.; Liu, X.; Tian, Y.; Zhu, Y.; Cao, W.; Cao, Q. Improving wheat yield prediction integrating proximal sensing and weather data with machine learning. *Comput. Electron. Agric.* **2022**, *195*, 106852. [CrossRef]
82. Skrzyczyńska, J.; Gąsiorowska, B. *Uprawa Roślin*; UPW: Wrocław, Poland, 2020; pp. 49–210.
83. Lamichaney, A.; Parihar, A.K.; Hazra, K.K.; Dixit, G.P.; Katiyar, P.K.; Singh, D.; Singh, A.K.; Kumar, N.; Singh, N.P. Untangling the Influence of Heat Stress on Crop Phenology, Seed Set, Seed Weight, and Germination in Field Pea (*Pisum sativum* L.). *Front. Plant Sci.* **2021**, *12*, 635868. [CrossRef] [PubMed]
84. Grzebisz, W. *Technologia Nawożenia Roślin Uprawnych—Fizjologia Plonowania*; Tom 1 Olei.; Powszechne Wydawnictwo Rolnicze i Lesne: Poznań, Poland, 2021.
85. Kotecki, A. *Uprawa Roślin Tom III.*; Wydawnictwo Uniwersytetu Przyrodniczego we Wrocławiu: Wrocław, Poland, 2020.
86. Singh, A.K.; Srivastava, C.P. Effect of plant types on grain yield and lodging resistance in pea (*Pisum sativum* L.). *Indian J. Genet. Plant Breed.* **2015**, *75*, 69. [CrossRef]
87. Wysokinski, A.; Lozak, I. The Dynamic of Nitrogen Uptake from Different Sources by Pea (*Pisum sativum* L.). *Agriculture* **2021**, *11*, 81. [CrossRef]
88. Pandey, J.; Devadasu, E.; Saini, D.; Dhokne, K.; Marriboina, S.; Raghavendra, A.S.; Subramanyam, R. Reversible changes in structure and function of photosynthetic apparatus of pea (*Pisum sativum*) leaves under drought stress. *Plant J.* **2023**, *113*, 60–74. [CrossRef] [PubMed]

agriculture

MDPI

Perspective

Neural Modelling from the Perspective of Selected Statistical Methods on Examples of Agricultural Applications

Piotr Boniecki [1], Agnieszka Sujak [1,*], Gniewko Niedbała [1], Hanna Piekarska-Boniecka [2], Agnieszka Wawrzyniak [1] and Andrzej Przybylak [1]

[1] Department of Biosystems Engineering, Poznań University of Life Sciences, 50 Wojska Polskiego Str., 60-637 Poznań, Poland
[2] Department of Entomology and Environmental Protection, Poznań University of Life Sciences, 159 Dąbrowskiego Str., 60-594 Poznań, Poland
* Correspondence: agnieszka.sujak@up.poznan.pl; Tel.: +48-61-846-6074

Abstract: Modelling plays an important role in identifying and solving problems that arise in a number of scientific issues including agriculture. Research in the natural environment is often costly, labour demanding, and, in some cases, impossible to carry out. Hence, there is a need to create and use specific "substitutes" for originals, known in a broad sense as models. Owing to the dynamic development of computer techniques, simulation models, in the form of information technology (IT) systems that support cognitive processes (of various types), are acquiring significant importance. Models primarily serve to provide a better understanding of studied empirical systems, and for efficient design of new systems as well as their rapid (and also inexpensive) improvement. Empirical mathematical models that are based on artificial neural networks and mathematical statistical methods have many similarities. In practice, scientific methodologies all use different terminology, which is mainly due to historical factors. Unfortunately, this distorts an overview of their mutual correlations, and therefore, fundamentally hinders an adequate comparative analysis of the methods. Using neural modelling terminology, statisticians are primarily concerned with the process of generalisation that involves analysing previously acquired noisy empirical data. Indeed, the objects of analyses, whether statistical or neural, are generally the results of experiments that, by their nature, are subject to various types of errors, including measurement errors. In this overview, we identify and highlight areas of correlation and interfacing between several selected neural network models and relevant, commonly used statistical methods that are frequently applied in agriculture. Examples are provided on the assessment of the quality of plant and animal production, pest risks, and the quality of agricultural environments.

Keywords: artificial neural networks; empirical data analysis; statistical methods; agriculture

Citation: Boniecki, P.; Sujak, A.; Niedbała, G.; Piekarska-Boniecka, H.; Wawrzyniak, A.; Przybylak, A. Neural Modelling from the Perspective of Selected Statistical Methods on Examples of Agricultural Applications. *Agriculture* **2023**, *13*, 762. https://doi.org/10.3390/agriculture13040762

Academic Editor: Dengpan Xiao

Received: 17 February 2023
Revised: 22 March 2023
Accepted: 23 March 2023
Published: 25 March 2023

1. Introduction

In scientific research, a model represents the studied object that may differ from that concept which exists in reality. In the above context, the original is always a primary concept in relation to the substitute that is its model. Thus, anything that is similar in some respect to something else, which is referred to as the original, can be regarded as a model. A model can be a static or dynamic structure that represents the original in a satisfactory (in some respects) way. Thus, a model creates the possibility of indirectly examining existing originals to better understand them or to improve them.

Researchers' aspirations to understand and to explain the laws governing nature more fully and scientifically have led to the importance of searching for new methods that most effectively support cognitive processes [1–3]. These methods undoubtedly include artificial intelligence techniques, among which neural network models occupy a special place. Neural network models are created deductively on sets of assumptions, resulting

from current scientific knowledge concerning the structure and mechanisms of the brain. The constructed models aim to mimic, in particular, those characteristics of biological neural systems that may be of technical use, which primarily include the resilience of biological systems to damage even to a significant part of their components, the extraordinary capacity for learning, and the high speed of information processing. These models, which are essentially probabilistic structural models of the studied systems, allow for effective and fast support and enrichment of the cognitive process, both at the stage of empirical research and at the analytical stage. In agricultural sciences, often, the cause-and-effect relationship of the examined phenomena and processes is unknown due to the complex structure of the investigated problems. Therefore, an obvious consequence is that the mathematical descriptions of the investigated problems are also generally unknown. If only random empirical data or, possibly, the results of computer simulations are available, it seems important to search for alternative methods of building models of empirical systems that describe investigated dependencies in agriculture. Such a possibility is offered by methods based on modern techniques that use neural network models.

In recent years, neural network models have played a special role because they are able to solve a range of issues, including scientific problems defined as unstructured (not susceptible to algorithmising). They can also effectively solve problems for which there is insufficient scientific knowledge or a lack of representative empirical data [4–6]. Neural network models are the general name for mathematical formulas and their software (or hardware) structures that implement signal processing through a network of interconnected elements (neurons) performing elementary mathematical operations. The morphology presented above is inspired by the architecture and functioning of natural neural systems, in particular, the structure and operation of the brain [7,8].

In a mathematical context, neural network models represent weighted graphs that form network topologies in which empirical systems are analogous to processes occurring in biological neuronal structures [9–11].

Neural modelling is the process of creating, verifying, and exploiting generated artificial neural networks. Therefore, the term "neural modelling" can be interpreted either as the process of producing structural models of the brain (generation of hardware and software neural network models) or as the process of reproducing the studied reality (both material and non-material) into an existing (hardware or software) neural (cellular) structure.

Neural network models and traditional statistical methods largely overlap and are complementary. In fact, they are tools for processing (analysing) and extracting, for example, knowledge contained in empirical data, among other things, obtained as a result of conducted experimental research. Therefore, the results of using neural network models and classical statistics methods are often similar, despite the fact that their techniques are fundamentally different.

An additional complication that can overshadow the mutual similarities and interrelationships is the fact that each of these scientific methodologies uses a different set of characteristic terminology. On the one hand, the nomenclature of statistics has been developed over the years and is primarily from the theory of probability. In the terminology used in neural network models, on the other hand, investigators use statistics and are generally seen as researchers primarily concerned with the generalisation process carried out on the basis of noisy data acquired through the design and subsequent execution of an experiment. It is worth noting that, in this context, artificial neural networks creatively complement and even expand traditional statistical models [12–17].

The aim of this perspective paper was to identify and highlight areas of correlation and interfacing between several selected neural network models and relevant commonly used statistical methods. Our work is divided into subsections showing different approaches in the context of models and mathematics. At the end of each chapter, some examples are described on issues in agriculture, agronomy, and agricultural engineering.

2. Overview of Trends and Methods Used by Researchers

An alternative approach to the classical description and analysis of empirical systems is to use methods of artificial intelligence in the broadest sense, and so-called neural network techniques in particular. Artificial neural networks are an intensively evolving field of knowledge increasingly used in many scientific approaches. They are also often of practical importance, where they are used in various types of artificial intelligence applications, for example, expert systems, autonomous systems, etc. Neural networks are versatile approximation systems that represent multidimensional (e.g., empirical) datasets. They have the ability to learn and adapt to changing environmental conditions. They also support the generalisation of acquired knowledge gained through learning. Unlike traditional methods of information processing, offered by cyclic calculating machines (computers) executing a prewritten programme, artificial neural networks are based on optimisation algorithms that enable the design of an appropriate network morphology, and then the selection of parameters of this structure that are adapted to the problem being solved.

2.1. Linear Neural Networks versus Linear Regression Models

The popularity of artificial neural networks is largely due to the possibility of modelling nonlinear issues relatively easily, i.e., practically solving problems described by bias regression models. However, this does not mean that structurally simple linear models should be neglected in the process of analysing empirical data. The general rule used in scientific research is that when there is a choice between a simple and a more complex model, the simpler (less complex) model should always be preferred, unless of course the complex model fits the data significantly better [3,18]. It is worth noting that the simplest model approximating the observed empirical relationship (described mathematically) is the linear model. Its intuitive simplicity is undoubtedly the reason for its great popularity (also in the empirical sciences), in particular, because of the computational possibilities offered by discrete mathematics and, especially, matrix calculus [14,19,20].

A linear model is represented by an artificial neural network with no hidden layers which means that it is one-layer structure (except for the special case of three-layer autoassociative networks implementing dimension reduction of the data vector by means of a linear principal components analysis (PCA) transformation). The neurons in the output layer are fully linear [21,22], i.e., they are neurons in which the total excitation is determined as a linear combination of the input values and which have a linear activation function. Of course, only the second layer processes the information, while the role of the first layer is to introduce the information (signal) into the network. It is important that it has both a linear postsynaptic function and a linear activation function. The training of linear neural networks is performed using supervised methods (with a teacher). During the training process of linear networks, the pseudo-inversion method works best. This technique optimises the output layer of the network, that is, it directly provides optimal coefficient values for all neurons of the entire linear layer of the network.

The statistical equivalents for standard linear neural network models are logistic regression for classification issues and linear (least squares) regression for regression analysis methods [5,22].

These methods can be equivalent to some simple forms of selected artificial neural network topologies. On the one hand, it is well known that logistic regression is based on the assumption of normality of the class distribution and homogeneity of the covariance matrix. Linear regression, on the other hand, requires a linear relationship between output and input variables and assumes a Gaussian perturbation of the output variables. When these assumptions are not met, which is unfortunately a common case, neural networks give better predictions compared to classical regression analysis [2,23,24].

This fact makes it possible to consider artificial neural networks as a more excellent and complementary stochastic tool with a wide range of applications. It is worth emphasising that they are useful, among other things, for analysing empirical data, both numerical

(numerical) and linguistic (nominal). This also means that artificial neural networks are a versatile instrument, characterised by a large spectrum of potential applications of a utilitarian nature and significant ease during their practical operation. Due to their low-complexity structure and ease of learning, as well as their large range of applications, networks with a linear transition function (activation function) are among those frequently used in engineering practice.

However, it is important to note that the aforementioned neural network models can only solve tasks that involve finding a linear (i.e., representable by a non-personal transformation matrix) reproduction of a set of input signals to a set of output signals.

An exemplary structure of a linear neural network is shown in Figure 1.

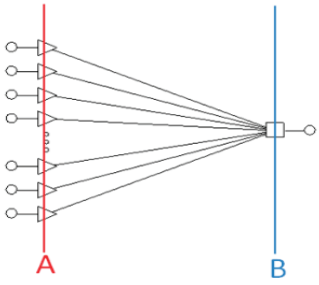


Figure 1. Structure of a linear network with single output. A—input layer and B—output layer.

Boniecki et al., 2009 [25] conducted a neural network analysis of digital images to identify the degree of maturity of compost produced from sewage sludge. In [21], the authors proposed a methodology for creating the learning sets necessary to generate artificial neural networks (ANN) for analysing graphically encoded information on the quality of pork half carcasses. Publication [22] deals with the determination of a selected family of butterflies presented in digital images by means of a generated ANN classifier, and an additional study [24] describes the application of neural network image analysis methods in quality assessment of greenhouse tomatoes using linear ANNs.

2.2. Nonlinear Layered Networks Trained with Supervision as a Subclass of Nonlinear Regression Models

The concept of regression in mathematical statistics is primarily associated with the process of empirically determining a stochastic relationship between correlated variables, where one most often attempts to model the problem using linear regression. However, often, the enormous complexity of the empirical systems under study does not allow such a significant simplification, and then it is necessary to build a nonlinear regression model [26].

In general, to solve a regression problem using a neural network model, the primary objective is to estimate the unknown value of the continuous output variable when the values of the input signals are known. Regression problems can be solved using different types of neural networks. The most commonly used network topologies are: multilayer perceptron (MLP); radial basis function (RBF), i.e., radial neural networks; generalised regression neural network GRNN), i.e., networks implementing generalised regression; and, in special cases, linear neural networks.

In regression problems, the purpose of an artificial neural network is to acquire (during the learning process) the ability to translate data representing input variables into continuous output variables, which are then the response of the network in its exploitation phase (after the learning process).

Unidirectional multilayer perceptron-type networks, i.e., MLP networks, are among the best studied and most widely used network topologies in practice. A multilayer perceptron network represents the so-called parametric class of neural network models (the number of neurons constituting its structure is significantly smaller than the size of the learning set).

The basic properties of MLP networks include the following:

- MLPs are one-way networks;
- MLPs are trained using a supervised technique (i.e., algorithms modify weights and threshold values using learning sets containing both input and set output values);
- An MLP network has a multilayer architecture, i.e., there is an input layer, a hidden layer and an output layer;
- Connections only allow communication between neurons in adjacent layers;
- Neurons in the network aggregate the input data by determining the weighted sum of the inputs (using a linear aggregation formula);
- The activation function of the input neurons is linear, that of the hidden neurons is nonlinear, and that of the output neurons, in general, is nonlinear;
- Due to the saturation level in sigmoidal activation functions, the data processed by the network require appropriate rescaling (so-called pre- and post-processing of the data).

MLP-type networks can be trained by using a supervised technique (based on learner, validation, and test sets) following several efficient iterative methods. Training techniques have their roots in both classical optimisation methods and heuristic methods. The choice of algorithm depends on the type and nature of the investigated problem.

An illustrative structure of a nonlinear neural network using MLP as an example is shown in Figure 2.

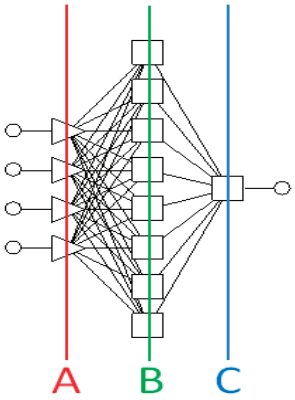


Figure 2. Structure of a multilayer perceptron (MLP) network with 1 input layer, 1 output layer, and 1 hidden layer: A—input layer; B—hidden layer; C—output layer.

Specialised neural networks of the GRNN type provide a neural representation of a statistical function approximation method. These networks have exactly four layers [24,27]. They consist of an input layer, a radial layer (storing the so-called centres), a regression layer, and an output layer. The neurons of the radial layer represent the cluster centres present in the learning data that are generally derived from an experiment. An appropriate clustering procedure such as the resampling algorithms k-means or Kohonen's is used to teach this layer. This allows the number of neurons in the radial layer to be smaller than the number of learning instances. The regression layer must have exactly one more neuron than the output layer and it is made of linear neurons. This layer contains two types of neurons. Type A neurons calculate the conditional regression for each output variable and a single type B neuron is used to calculate the probability density. It is known from practice that regression networks of the GRNN type learn in a very short time, but unfortunately tend to reach large sizes. This fact makes their performance during normal operation rather slow.

Hinton and Osindero, 2006 [27] presented a novel fast-learning algorithm dedicated to deep neural networks for the identification of graphical information used mostly in agriculture. Moody and Darkin, 1988 [26], in turn, dealt with fast learning implemented

in adaptive ANNs built from locally tuned processing units. Wawrzyniak et al., 2023 [28] described a neural network model that was used to study the potential of biomethane from cattle manure. Boniecki et al., 2021 [29] compared the affectivity of MLP and RBF neural network models on graphical classification and qualitative identification of compost. Pentoś et al., 2022 [30] used an ANN and MLR to investigate the relationship between electrical and selected soil mechanical properties. Construction of MLP ANN and multiple regression (MLR) models for tuber yield prediction of three very early potato varieties Arielle, Riviera, and Viviana was the aim of the work by Piekutowska et al., 2021 [31]. Sabzi-Nojadeh et al., 2021 [32] conducted a study to compare the performance of an ANN and MLR in predicting essential oil yield (EOY%) and trans-anethole yield (TAY%) of fennel populations. In contrast, Gorzelany et al., 2022 [33] studied the selected mechanical properties of fresh and stored large fruit cranberry fruit using an ANN and MLR. An example of a multilayer perceptron neural network (MLP-NN) designed and applied to model the internal temperature and relative humidity in a greenhouse was included in work by [34]. For the specific NN backpropagation as a training algorithm, the input variables were the external temperature and relative humidity, wind speed, solar irradiance, as well as the internal temperature and relative humidity, up to three timesteps before the modelled time step. The maximum errors of the modelled temperature and relative humidity were 0.877 K and 2.838%, respectively, whereas the coefficients of determination were 0.999 for both parameters. The intention of the authors was to provide the right parameters for the greenhouse control system.

Crop diseases are critical factors that hamper agricultural development, severely affecting yields and crop security. The existing crop disease identification models make it difficult to focus on a disease area. In addition, crops with similar disease traits are easily misidentified [35–37]. To address the above problems, accurate and efficient disease identification models have been proposed that incorporate local and global image features into the trait analysis and improve the separation between similar diseases. Frequently, the traditional methods of agricultural pest detection cannot satisfy the needs of agricultural production because of low efficiency and accuracy. In their work, Yao et al., 2017 [38] proposed a method for classification of high-resolution agricultural remote sensing images based on convolution neural networks (CNNs). By training and testing the CNNs (with a large number of high-resolution images), the crop classification achieved a rate of 99.66% after optimising the network parameters. Li et al., 2022 [39] proposed a model that applied convolutional neural networks for pest identification based on transfer learning. A dataset of agricultural pest images was adopted as an experimental dataset. The effect of data augmentation on the classification performance of different samples was compared. The results showed that the classification effect of the model based on transfer learning was generally better than that based on new learning. Compared to new learning, transfer learning significantly improved the recognition ability of the model and significantly reduced the training time to achieve the same classification accuracy. In [40], an anchor-free region convolutional neural network (AF-RCNN) for precision recognition and classification of 24 classes of pests was proposed. Interestingly, a running time of 0.07 s per image was achieved to perform real-time detection. In addition, a transformer encoder was introduced in the model as a convolution operation, and therefore, the model could establish the relationship between long-distance traits and extract global disease image traits, and centre loss was applied as a penalty term to expand the inter-class difference of crop disease traits and reduce their intra-class gap. An identification accuracy of more than 96% was achieved, even with a complex background [41].

2.3. Discriminant Analysis with Kernel versus Normalised Neural Networks with Radial Symmetry

A discriminant function analysis is most commonly used to obtain information on deciding which variables best divide an empirical dataset into naturally occurring groups. The main idea behind a discriminant function analysis is to decide whether the groups

differ due to the mean of a certain variable, and then use this variable to predict group membership (e.g., new cases).

Networks with radial basis functions are examples of neural networks characterised by circular symmetry. In 1988, the radial basis function (RBF) network was proposed by Broomhead and Lowe, 1988 [42] and, independently, in 1989 by Moody and Darkin [26]. RBF-type radial networks implement a distinctive way of transforming an input set into an output set. Such a representation involves fitting an approximating function of multiple variables to the desired values. Thus, it involves spanning over the learning set and fiting a multidimensional hypersurface to the desired values.

RBF networks usually require more neurons than perceptron-type networks, but they are much faster to learn. They typically consist of three layers: an input layer (usually linear), a hidden layer with neurons with radial activation functions, and an output layer containing linear neurons. The basic properties of RBF-type networks include the following:

- RBFs are unidirectional networks;
- RBFs are trained using a hybrid technique, i.e., the hidden radial layer is trained using a nonsupervised method (weights and threshold values are modified using learning sets containing only input values, for these algorithms, output values in the dataset are not required and are ignored if they occur), while the linear output layer is trained using a supervised technique, for example, using a pseudo-inversion method;
- RBFs have a three-layer architecture, i.e., there is an input layer, a hidden (radial) layer, and an output layer;
- Connections only allow communication between neurons in adjacent layers;
- The activation function of the input neurons is linear and that of the hidden neurons is nonlinear (radial), and that of the output neurons is fully linear.

RBF networks learn relatively quickly and have the advantage of never extrapolating functions over too large a distance from known data. These networks can have definite advantages, especially when modelling (with them) physical phenomena known to have circular symmetry. However, they are generally much larger than MLP networks solving the same tasks, making them more time-consuming to run on a computer simulating the network. However, these networks can have definite advantages when the network is implemented as a dedicated electronic circuit.

Despite the similarity, there are important differences between the two types of networks:

- A radial RBF-type network has a determined structure (with one radial hidden layer and a linear output layer), while an MLP-type network can have a different number of hidden layers and the output neurons can be either linear or nonlinear;
- When radial functions are used, there is more variation in the choice of their shape.

The most common equivalent of the sigmoid function is the basis function.

The structure of a nonlinear neural network is shown in Figure 3 using the RBF model as an example.

Boniecki et al., 2012 [11] described a neural network image analysis process for estimating the aerobic and anaerobic decomposition of organic matter using the example of straw degradation. Piekarska-Boniecka et al. (2008) [43] identified parasitic wasps of the subfamily pimplinae (Hymenoptera, ichneumonidae) using RBF-type ANNs. The authors of [44] addressed the problem of dimension reduction of digital image descriptors in the process of neural network identification of damaged grains of malting barley. In [45], RBF-type neural networks were analysed as a dedicated tool for the study of selected empirical systems derived from agricultural engineering. In the study by [46], the effectiveness of multiple regression techniques combined with different types of artificial neural networks (ANNs) including RBF were investigated using regression results as input variables to estimate sunflower grain yield under normal and salinity conditions. A very interesting group of studies are those that predict the possible quality of water. In the study by Hong et al., 2020 [47], a radial basis function artificial neural network (RBF-ANN) as well as a hybrid method of RBF-ANN and grey relational analysis (GRA)

were proposed to predict trihalomethane (THM) levels in drinking water. The results indicated the high capability of RBF-ANN to learn the complex nonlinear relationships involved in THM formation (regression coefficients (rp = 0.760–0.925) and prediction accuracy (N25 = 92–98%). It was found that RBF-ANN using fewer water quality parameters based on GRA achieved excellent performance in THM prediction (rp = 0.760–0.946 and N25 = 92–98%). In the work by Deng et al., 2021 [48], occurrence of haloketones (including dichloropropanone, trichloropropanone, and total HKs) was studied with means of linear, log-linear regression models, back propagation (BP), as well as radial basis function (RBF) artificial neural networks (ANNs) in real water supply systems. The results showed that the overall prediction ability of RBF and BP ANNs was better than linear/log-linear models. Though the BP ANN showed excellent prediction performance in internal validation (N_{25} = 98–100%, R^2 = 0.99–1.00), it could not effectively predict HK occurrence in external validation (N_{25} = 62–69%, R^2 = 0.202–0.848). The prediction ability of RBF-ANN in external validation (N_{25} = 85%, R^2 = 0.692–0.909) was quite good, which was comparable to that in internal validation (N_{25} = 74–88%, R^2 = 0.799–0.870).

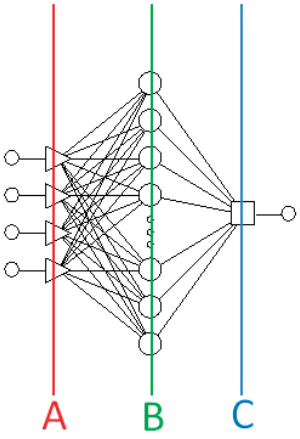


Figure 3. Structure of a radial basis function (RBF) with 1 input layer, 1 output layer, and 1 radial hidden layer: A—input layer, B—hidden layer, C—output layer.

2.4. Principal Component Analysis Method vs. Unsupervised Neural Networks

Principal components analysis (PCA) is a linear dimension reduction technique to identify mutually orthogonal axes (dimensions) with the greatest variance within the original empirical data. This technique realises a projection of the data into a space of fewer dimensions, formed from a subset of the components with the greatest variance. Therefore, principal component analysis allows the orthogonal directions of the new coordinate system to be determined in such a way that they correspond to the directions of maximum variance of the original data. Dimension reduction is achieved by projecting the data from the original signal space into the aforementioned new space characterised by a lower dimension and formed by the subset of components with the greatest variance.

The PCA-based learning algorithm implements a linear transformation by rotating the data included in the learning set to a new coordinate system, formed by the eigenvectors of the autocorrelation matrix determined for the learning data. The eigenvalues, corresponding to each eigenvector, determine how much of the "variability" present in the empirical data is represented by the corresponding eigenvectors. The PCA method often makes it possible to solve problems that would otherwise simply have too many variables to build an effective neural network. Principal component analysis can be easily implemented in an artificial neural network optimised in nonsupervised mode (trained using Hebb's rule) [8]. It transforms the input cases of the network (which are points in a "larger" space to a "smaller" space) in such a way that the variability in the resulting

dataset is maximised. A special group of neural networks with Hebbian learning are auto-associative networks, which use the aforementioned algorithm in the learning process. This means that the auto-associative network, during the learning process, acquires the ability to reduce the dimension of the input data (thus performing a type of data compression). By definition, auto-associative neural networks are those that reproduce input values at their outputs. Such an operation is justified because the considered auto-associative network has a significantly smaller number of neurons in the middle (hidden) layer than in the input (and output) layer. During the process of feeding the learning vector into the ANN, the data must pass through a bottleneck between the input and output. In order to fulfil its task, which in this case is to reproduce the input information in the output (as defined by an auto-associative network), the network must first learn to represent the input data with a smaller number of signals generated by the neurons of the hidden layer.

The neural network can only master the ability to reconstruct the full input data from the previously compressed information encoded inside (in the hidden layer) of the network in the next stage. This means that the auto-associative neural network, during the learning process, acquires the ability to reduce the dimension of the input data packed into the hidden (middle) layer.

In general, an auto-associative neural network designed for data compression should consist of a minimum of three layers:

- An input layer (with the number of neurons corresponding to the input number of data);
- An output layer (of the same size);
- A hidden layer (with a much smaller number of neurons).

An auto-associative neural network should be trained in such a way that it accurately reproduces the input data on its outputs. This is why it has exactly the same number of inputs as outputs, and the variables used to teach it are assigned a particular character of so-called input/output variables. As indicated above, the idea behind the operation of such a network is that the number of hidden neurons is much smaller than the number of inputs or outputs, effectively forcing the information to be "squeezed" through a representation of a smaller (compressed) dimension. An auto-associative (three-layer) network constructed as above performs a transformation of the input data to the hidden layer (of reduced dimension), and then another transformation back to the output layer. It can be shown that when the neurons used in the hidden layer and in the output layer have linear (or quasi-linear) characteristics, such a network actually learns to approximate the standard PCA algorithm. Thus, in the present case, it is a substitute for the linear PCA method presented above [49].

One substantial problem with the PCA method is its linear nature. For this reason, it cannot be used to reduce the dimension of the input data in any case. It can only be used to identify exclusively linear transformations that optimise the condensation of the information contained in the considered variables, based on the directional search for the maximum variance.

An alternative approach, free of the limitation indicated above, is to use a particular auto-associative neural network topology that realises a nonlinear version of PCA. The idea of realising a nonlinear dimension reduction is, in fact, to use an analogous neural network topology, but built with nonlinear neurons, i.e., having in their structure a nonlinear membrane potential of the postsynaptic function. To fully exploit the possibilities offered by a nonlinear network, however, more than one layer of neurons is needed for each of the two transformations implemented (for both compression and decompression). For this reason, when creating a network for nonlinear data compression, a neural network topology with five layers should be generated [17,50–52].

The hidden middle layer is the layer that reduces the dimension of the input signal, while the layer between it and the input layer performs precisely the required nonlinear compression of the input data. The corresponding two-layer network structure, located between the hidden layer and the output layer, performs the inverse transformation by

decompressing the previously compressed signal. Figure 4 illustrates the structure of a nonlinear neural network based on the example of a PCA model.

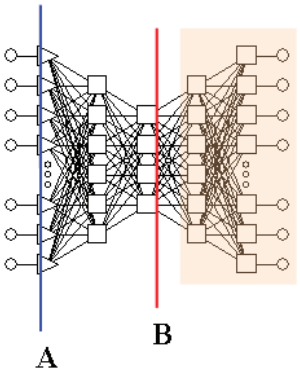


Figure 4. A neural auto-associative network implementing a nonlinear PCA transformation: A—input layer of the 5-layer auto-associative network; B—hidden layer of the 5-layer auto-associative network (compressing in a non-inear manner).

Studies by [44,52] presented a method using neural network reduction of graphical descriptors to determine the quality and extent of mechanical damage in malting barley acquired from digital images. A novel approach for analysing a graphical image was developed to map disturbances and to evaluate the ecological status of mangroves based on multispectral WorldView-2 and Sentinel-2 and remote sensing techniques [53]. The spatial disturbance index (SDI) was calculated based on landscape metrics and principal component analysis (PCA) in order to detect degraded mangroves. The first three components of the PCA explained in total 87% (47%, 27%, and 13%, respectively) of the variance of information distribution, demonstrating PCA effectiveness.

Bourland and Kamp, 1988 [49] demonstrated an example of an auto-associative process using MLP-type ANNs and matrix singular value decomposition techniques. The authors of [54] dealt with nonlinear principal component analysis using selected auto-associative neural network topologies.

Krevh et al., 2023 [55] examined long-term data on soil water regime and nitrate dynamics to better understand the implications of N management in relation to groundwater pollution. The principal component analysis (PCA) method was performed to identify the relationships among all soil properties and environmental characteristics. The results showed the complex interaction of soil hydraulic properties, precipitation patterns, plant uptake, and N application.

A vegetation ecosystem model was constructed to provide a simulation method for improving water use efficiency capacity as well as for predicting the future response of water use efficiency (WUE) to climate change [56]. The structural relationship and degree of influence among factors was determined by using a structural equation model (SEM) that was transformed into ANN topology, where the PCA method was employed to reduce spatial dimensionality. The results showed that different influencing factors on WUE presented a diversity with different levels. The ANN structure optimised by using SEM fit better, and the PCA-SEM-ANN model had a very high explanatory and precision for environmental control of the ecosystem as well as WUE simulation.

2.5. Cluster Analysis by K-Means vs. Topological Kohonen Maps

A k-means cluster analysis is based on a statistical procedure that has been known for a relatively long time. It is an algorithm for determining k centers that optimally represent the structure of N points (k < N). In the initial phase of the algorithm, randomly selected cases from the empirical dataset are taken as centres. Each case is then assigned to the

cluster represented by the centre nearest the case. Once all cases have been assigned, new centres are determined as the focus of each cluster. The process of assigning cases to classes is repeated until equilibrium is reached.

With reference to neural network models, the k-means algorithm determines radial centres, which are stored in radial neurons located in the first hidden layer of the network. The k-means method assigns each learning event to one of the k clusters (where k is the number of radial neurons). Each of the assigned clusters is represented by the centre of gravity of the cases belonging to it, and the distance of each case from the centre of gravity of the cluster assigned to it is smaller than the distance between that case and the centre of gravity of any other cluster (this is the basic criterion for including a particular case in a particular cluster). The centres of gravity of the clusters are copied to the radial neurons.

An important group of unsupervised networks trained with methods based on the k-means algorithm are artificial neural networks also known as Kohonen maps. They have a two-layer structure (without hidden layers), i.e., they only have an input layer and an output layer made of radial neurons. The learning algorithm proposed by TuevoKohonen is, in fact, an adapted and modified (for artificial neural networks) method of the well-known k-means method. This algorithm determines cluster centres for the output radial layer, which is usually a two-dimensional topological map. During its operation, the input pattern is repeatedly presented and the weights of the winner neuron and the bespoke neurons immediately adjacent to the winner are modified. The correction carried out is intended to make the "tuned" weights similar to the input pattern. Thus, the aim of learning with the Kohonen method is to identify a set of cluster centres that best reflect the actual distribution of learning cases. As a result of learning, the neural network acquires the ability to activate neighbouring neurons (when given similar input data) which results in a set of output signals from the network that can be interpreted as a map representing the topological relationships between the input data in the space of weights and inputs. The neighbourhood plays a key role in the learning of Kohonen networks. By modifying the neurons surrounding the winning neuron, rather than just the neuron itself, similar data will be close in the topological map. The neighbourhood size is reduced as the learning process progresses, and the learning rate is similarly reduced. Thus, at the beginning of learning, a coarse mapping is obtained, with large clusters of neurons corresponding to similar cases. Later, smaller details of the map are obtained, as individual neurons in the clusters begin to respond to more subtle differences between similar cases [21,57].

Kohonen-type neural networks are typically used in the classification process. They are also sometimes used to generate signals useful for vector coding of input data. The "output" of the network is then a vector of weights of the winning neuron of the output layer of the Kohonen network, representing a vector code, suitable, for example, for signal compression using a so-called codebook. Figure 5 shows an example of a Kohonen-type neural network structure.

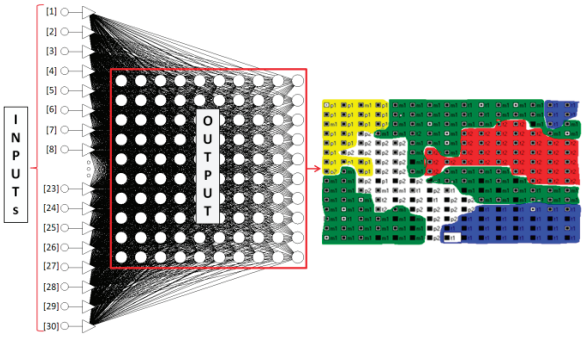


Figure 5. An example of the structure of a Kohonen-type neural network applied as a colour (classificator) separator.

The Kohonen self-organising map neural network (KSOM-NN) has attracted significant attention from researchers because of the quality of data visualisation and interpretation [18]. In [58], a comprehensive review is presented of the usefulness of the Kohonen neural network model for solving selected agricultural classification problems. Some applications of the cluster analysis by using k-means or Kohonen maps are described below. In [59], it was found that classical machine learning methods such as K-nearest neighbours could achieve good classification of agricultural images collected using Sentinel-1 radar. The accuracy was as high as 86%.

In [13], the application of a Kohonen-type neural network was demonstrated in the process of nonparametric qualitative classification of tomatoes. In turn, [18,60] performed the classification of selected orchard pests using a neural network of the self-organising feature maps (SOFM) type. The same neural network topology was used in [61] for neural network classification of compost maturity levels. A Kohonen's neural network map was used to construct a multi-objective genetic local search algorithm. In the new algorithm, for each generation, a population of neurons was trained using elite solutions of the genetic algorithms in the current population. The training rule of neurons was developed using the concepts of the learning rule of self-organising map and variable neighborhood search algorithm to improve the local and global search [62].

2.6. Bayesian Statistics versus Probabilistic Neural Networks

Techniques for estimating probability density functions from extracted data have a long history in statistics. Bayesian estimation is a well-known method of statistical analysis based on Bayes theorem, which is the foundation of probability theory [50,51,63]. This theorem states that the a posteriori probability of a parameter "p" is proportional to the a priori probability "p" multiplied by the probability "p" determined from the empirical data. Bayesian statistics can be applied for the probability density function estimation of model parameters using available data, and as a result of this analysis, the model is selected whose parameters maximise the aforementioned probability density function, thus reducing the global error.

In recent years, an approach to probability density function estimation based on so-called kernel approximation has become important. In kernel approximation, simple functions (e.g., a Gaussian function) are located at the point of occurrence of each available case, and then these are added to obtain an estimator of the joint probability density function. The operation of the aforementioned method can be described in that the presence of a case at a point in the input space implies a high probability density at that point. Therefore, the clustering of cases that are close to each other indicates an area of high density. If enough empirical data are available, a relatively good approximation of the true probability density function is obtained [8,64].

The kernel approach to approximating probability density functions is very similar to the use of neural networks with radial basis functions. This inspired the creation of a new category of neural network models called probabilistic neural networks (PNNs). Probabilistic neural networks, which are a subset of Bayesian networks, are essentially implementation of the so-called kernel function approximation method and are used exclusively for classification. They typically have three or four layers. These distinguish between an input layer, a layer of radial neurons, and a layer of linear classifier neurons. Optionally, a fourth layer, which is also linear and contains a squared cost matrix, can be included. Since the attached cost matrix is squared, the third and fourth layers must have the same number of neurons. The basic properties of PNN-type networks include the following:

- PNNs are unidirectional networks;
- PNNs are trained using a nonsupervised technique, i.e., weights and threshold values are modified using learning sets containing only input values (for these algorithms, output values in the dataset are not required and are ignored if they occur);

- PNNs have a three or four-layer architecture, distinguishing between an input layer, a hidden (radial) layer, and a linear output classifier layer (optionally, a fourth layer may be included, which is also linear and contains a quadratic cost matrix and has the same number of neurons as the third layer),
- Connections only allow communication between neurons located in neighbouring layers;
- The activation function of the input neurons is linear, that of the hidden neurons is nonlinear (radial), and that of the output neurons is fully linear.

Among the greatest advantages of PNNs is the generation of probability values (or, more precisely, their estimates) at the outputs, rather than just the "raw decisions" themselves, which greatly facilitates the evaluation and interpretation of the results. Another advantage of PNN-type networks is the high speed of their learning. The biggest disadvantage of PNN-type networks is their size, due to the fact that the structure of networks of this type must contain neurons corresponding to the individual examples considered, which means that the entire learning set is mapped in the structure of a PNN-type neural network. This causes, among other things, very high memory requirements for the creation and operation of such networks and this is the reason why the time required to run such a network is rather long. PNN-type neural networks are particularly useful during experiments aimed at defining network prototypes (e.g., when decisions are made regarding the choice of input variables), since the short learning time of these networks makes it possible to perform a large number of tests (e.g., with different input datasets) in a short period of time. The structure of a nonlinear neural network using the PNN model as an example is shown in Figure 6.

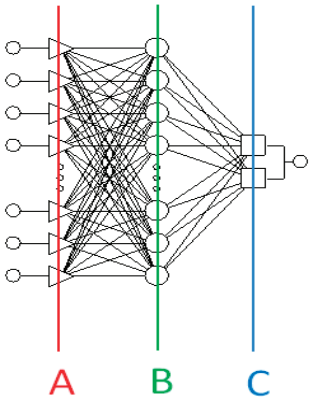


Figure 6. Probabilistic neural network (PNN) with 1 input layer, 1 output layer, and 1 radial hidden layer: A—input layer, number of neurons equal to the number of features; B—hidden layer built of radial neurons also called the pattern layer, the number of neurons is equal to the number of learning instances; C—output layer built of linear classifying neurons (summation layer), the number of neurons is equal to the number of classes.

Examples of application of Bayesian statistics and probabilistic neural networks are described below. First, the newest tutorial by Krause and Bokinala demonstrated how Bayesian networks can be built from a wide range of sources of expert knowledge and data [65]. Drury et al., 2017 [66] applied Bayesian networks in agricultural examples. Studies by [51] and [67] provided a comprehensive compendium of knowledge on probabilistic neural networks. In [68], Sujak et al., 2017 conducted an environmental bioindication study using a Bayesian network implementation with the grey heron as a model species. The work was aimed at identifying the quality of the agricultural environment. Interesting applications of Bayesian networks include their use to model agricultural water demand, groundwater intake, surface water supply, and climate which show complex, nonlinear relationships with groundwater storage in agricultural regions. A novel machine

learning-based approach to model a groundwater ensemble has been used in combination with a Bayesian model averaging approach to predict groundwater storage. The results showed that the machine learning model had remarkable predictive ability without loss of accuracy, but with higher computational efficiency [69]. An exemplary application was the development of a Bayesian network (BN) model to predict nitrogen loading to surface water at the farm level, using specific site characteristics such as landscape, soil, cropping system, and to compare different conventional and conservation agricultural practices. A sensitivity analysis showed the significance of weather factors on surface runoff and topographic information for soil erosion, while agricultural treatments were found to be less important. Bayesian networks have been shown to be a useful and flexible tool for data and knowledge assimilation and a practical approach for testing and comparing the effects of different agricultural interventions on agricultural system emissions [70].

Wu et al., 2023 [71] proposed a novel agricultural drought prediction model for long lead times by integrating vine copulas with Bayesian model averaging (hereafter, the BVC model) which improved agricultural drought management, food security assessment, and early drought warning. An interesting application of Bayesian networks was the project by Lachaud et al., 2022 [72] in which Bayesian networks were applied to study food security. There are many other applications of Bayesian networks, and it is impossible to mention it all in one single publication.

2.7. Analysis of Time Series and Their Neural Network Representation

A time series is used to refer to a set of data values (most often empirical) that is in a determined order over time. A time series data analysis implies the study of sequences of measurement data that are characterised by a non-random (determined) ordering. Unlike classical analyses performed on random samples, a time series analysis is based on the assumption that successive values in a dataset represent successive measurements, taken at equal time intervals [8].

A time series analysis aims to detect the nature of a phenomenon represented by a sequence of realised observations and to forecast the future values of an identified time series. The above-mentioned priorities require identifying and then describing the characteristic elements of the time series in question. In the course of a time series analysis, it is most often the case that subsequent values of a series are projected on the basis of previous values of the same variable or on the basis of values of other variables. Regardless of the accuracy of the theoretical justification of the form of the generated model, it is always possible to predict future values of a time series based on extrapolation methods.

Artificial neural networks are well suited to analysis time series predictions. Usually, this is because a continuous variable is being predicted; therefore, time series prediction is essentially a specialised form of regression. Typically, the next value of a time series is forecast based on a certain number of preceding values. In this case, the aim is to determine a value that is "one step ahead" in time. However, there is nothing to prevent the process of long-term forecasting from being considered in a similar way. It should be noted that any type of artificial neural network can be used for time series forecasting. However, the choice of network topology (MLP, RBF, PNN, GRNN, Kohonen network, etc.) must be adequate to realise the appropriate form of input signal processing. This form is determined by the type and nature of the problem studied. Indeed, first, it must be decided whether it is a regression or classification problem. Then, the selected neural network is adapted for time series forecasting by setting values for characteristic parameters such as "row" and "forecast horizon". The "row" parameter specifies how many instances (observations of previous values of the variable under consideration) must be introduced into the inputs of the network so that, based on them, the network can determine the output value (the projected next value of this variable). The "forecast horizon" parameter, in turn, tells how far away the value of the variable to be predicted (forecasted) is from the end of the series of retrospective observations given at the input.

Time series analyses have broad applications. Examples of using time series with neural network representation applications are described below. Bishop (1995) [8] presented an overview of applications of time series interpreted in the form of artificial neural networks. Boniecki and Weres, 2001 [45] demonstrated selected applications of neural network time series for empirical modelling of processes occurring in agricultural systems. In turn, [73] described techniques for forecasting yields of selected agricultural crops using a neural network time series model.

An analysis of the literature has shown that, compared with traditional models, deep neural networks can enhance data structure mining and overall information simulation capabilities through innovative and efficient structures. This means that it is also possible to extend the range of environmental parameter selection for agricultural facilities and to achieve environmental prediction end-to-end optimisation through an intelligent time series model based on deep neural networks [74].

One of the exemplary studies aimed at analyzing time-series trends and variability analysis of observed rainfall and temperature records is that of Tofu et al. (2023) [75]. The study focused on smallholder farmers' perceptions, including the analysis of their response strategies to both observed and perceived climate variability and its determinants. The impacts of abnormal trends or shifts in observed rainfall and temperature patterns, along with socioeconomic, technological, and behavioural factors, call for policymakers to set strategies that will enable smallholder farmers to improve their livelihoods, ensure food security, and build resilience in the face of climate extremes.

3. Conclusions and Recommendations

The growing interest in artificial neural networks is not the result of chance but due to the need to improve data analysis methods by making them faster and more automated. The primary distinguishing factor of neural network models, compared to symbolic machine learning methods (e.g., rules or decision trees), is the completely different manner in which knowledge is acquired by the system during the learning process. This post-symbolic representation is a direct consequence of the applied method of computation and the adopted topology of the neural network model.

Neural network models have simply proven to be convenient as well as effective instruments that are useful for a wide variety of practical tasks. In fact, they have been successfully applied to solve all sorts of problems, not only in agriculture but also in fields as diverse as finance, medicine, engineering in the broadest sense, geology, and physics. In fact, there may be many more applications, since artificial neural networks can be used wherever problems arise in data processing and analysis (also in the form of multidimensional random variables), prediction, classification, or control. They should also include so-called unstructured types of problems, i.e., problems that are difficult or even impossible to algorithmise, or for which there is insufficient scientific knowledge or limited available empirical data.

Neural networks are a sophisticated modelling technique capable of representing extremely complex functions. In particular, ANNs are nonlinear in nature, which significantly enriches their application possibilities, and therefore gives them an advantage over commonly used statistical methods. Neural networks can also control the complex problem of multidimensionality, which, when using other methods (e.g., statistical methods), significantly hinders the process of modelling nonlinear functions with a large number of independent variables. It is worth highlighting that calculations based on neural network analyses can be carried out on computers with a neuroprocessor central unit that performs its tasks in a significantly different way from currently used microprocessors, in which processing is carried out according to the concept proposed in the 1940s (previous century) by John von Neuman (cyclic machine). The neuroprocessor was inspired by the results of many years of neuroscientific research on the brain. For technical reasons, numerical simulators are currently most commonly used instead of hardware neuroprocessors, occurring in the form of the artificial neural networks discussed above.

These applications have emerged and have been developed as a result of work carried out in the field of distributed processing techniques (in particular parallel information processing) and research in the area of artificial intelligence. Scientific work that has concerned the construction of models of the basic neural structures found in the brain has been of fundamental importance.

The most important and desirable characteristics of neural network models include the following:

- High processing power;
- Reliability and resistance to interruptions;
- Ease of use;
- Simple structure;
- Ability to generalise the acquired knowledge;
- Biological inspiration.

The creation of neural network models (as well as stochastic models) requires an adequate set of empirical data, which is needed to construct the training, validation, and test sets that are the basis for creating the weights matrix, which is the learning set structure necessary for generating ANNs. For numerical reasons, it is important that the number of learning vectors (experimental cases) is as large as possible. It is standard practice to have 10 learning cases per variable. The structure of the learning set is generally determined by the procedures implemented in the individual ANN software simulators such as Statistica and MatLab. This also applies to the methods of validation and verification of the generated neural network models.

The studies in the literature described in this paper indicate the usefulness of the developed neural network models, among others, as important supports for decision-making processes occurring during agricultural production (broadly defined). In particular, the generated ANNs can be implemented in modern information systems to support a variety of processes in agricultural systems.

Author Contributions: Conceptualization, P.B., A.S. and G.N.; Investigation P.B., A.S., G.N., H.P.-B., A.P. and A.W.; Resources, P.B., A.S., A.W. and G.N.; Writing—original draft preparation, P.B., A.S. and G.N.; Writing—review and editing, P.B., A.S. and G.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: This study includes no data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Méndez, L.J.; Lira, A.; Lasa, R.; Cerdeira, S. Delineation of site-specific management zones for pest control purposes: Exploring precision agriculture and species distribution modeling approaches. *Comput. Electron. Agric.* **2019**, *167*, 105101. [CrossRef]
2. Haan, N.L.; Zhang, Y.; Landis, D.A. Predicting Landscape Configuration Effects on Agricultural Pest Suppression. *Trends Ecol. Evol.* **2020**, *35*, 175–186. [CrossRef] [PubMed]
3. Castañeda, A.; Castaño, V.M. Smart frost measurement for anti-disaster intelligent control in greenhouses via embedding IoT and hybrid AI methods. *Measurement* **2020**, *164*, 108043. [CrossRef]
4. Nowakowski, K.; Raba, B.; Tomczak, R.J.; Boniecki, P.; Kujawa, S.; Nowak, P.J.; Matz, R. Identification of Physical Parameters of Cereal Grain using Computer image Analysis and Neural Models. In Proceedings of the International Conference on Digital Image Processing (ICDIP 2013), Beijing, China, 21–22 April 2013. [CrossRef]
5. Badgujar, C.; Das, S.; Figueroa, D.M.; Flippo, D. Application of Computational Intelligence Methods in Agricultural Soil–Machine Interaction: A Review. *Agriculture* **2023**, *13*, 357. [CrossRef]
6. Kujawa, S.; Niedbała, G. Artificial Neural Networks in Agriculture. *Agriculture* **2021**, *11*, 497. [CrossRef]
7. Fausett, L. *Fundamentals of Neural Networks*; Prentice Hall: New York, NY, USA, 1994.
8. Bishop, C. *Neural Networks for Pattern Recognition*; University Press: Oxford, UK, 1995.

9. Boniecki, P.; Nowakowski, K.; Tomczak, R.Ł. Neural networks type MLP in the process of identification chosen varieties of maize. In Proceedings of the International Conference on Digital Image Processing (ICDIP 2011), Chengdu, China, 15–17 April 2011. [CrossRef]
10. Boniecki, P.; Koszela, K.; Piekarska-Boniecka, H.; Nowakowski, K.; Przybył, J.; Zaborowicz, M.; Raba, B.; Dach, J. Identification of selected apple pests, based on selected graphical parameters. In Proceedings of the International Conference on Digital Image Processing (ICDIP 2013), Beijing, China, 21–22 April 2013. [CrossRef]
11. Boniecki, P.; Nowakowski, K.; Ślósarz, P.; Dach, J.; Pilarski, K. Neural image analysis for estimating aerobic and anaerobic decomposition of organic matter based on the example of straw decomposition. In Proceedings of the International Conference on Digital Image Processing (ICDIP 2012), Kuala Lumpur, Malaysia, 8–9 June 2012. [CrossRef]
12. Nowakowski, K.; Boniecki, P.; Dach, J. The identification of mechanical damages of kernels basis on neural image analysis. In Proceedings of the International Conference on Digital Image Processing (ICDIP 2009), Bangkok, Thailand, 7–9 March 2009.
13. Boniecki, P.; Nowakowski, K.; Tomczak, R.J.; Kujawa, S.; Piekarska-Boniecka, H. The application of the Kohonen neural network in the non-parametric quality-based classification of tomatoes. In Proceedings of the International Conference on Digital Image Processing (ICDIP 2012), Kuala Lumpur, Malaysia, 8–9 June 2012. [CrossRef]
14. Boniecki, P.; Piekarska-Boniecka, H.; Świerczyński, K.; Koszela, K.; Zaborowicz, M.; Przybył, J. Detection of the granary weevil based on X-ray images of damaged wheat kernels. *J. Stored Prod. Res.* **2014**, *56*, 38–42. [CrossRef]
15. Nowakowski, K.; Boniecki, P.; Tomczak, R.J. Identification process of corn and barley kernels damages using neural image analysis. In Proceedings of the International Conference on Digital Image Processing (ICDIP 2011), Chengdu, China, 15–17 April 2011. [CrossRef]
16. Sujak, A.; Jakubas, D.; Kitowski, I.; Boniecki, P. Identification of Factors Affecting Environmental Contamination Represented by Post-Hatching Eggshells of a Common Colonial Waterbird with Usage of Artificial Neural Networks. *Sensors* **2022**, *22*, 3723. [CrossRef]
17. Świetlicka, I.; Sujak, A.; Muszyński, S.; Świetlicki, M. The application of artificial neural networks to the problem of reservoir classification and land use determination on the basis of water sediment composition. *Ecol. Indic.* **2017**, *72*, 759–765. [CrossRef]
18. Pilarski, K.; Boniecki, P.; Ślósarz, P.; Dach, J.; Piekarska-Boniecka, H.; Koszela, K. Classification of chosen orchard pests with using the SOFM neural network. *AJAR* **2012**, *7*, 6357–6362.
19. Deepa, S.; Alli, A.; Sheetac; Gokila, S. Machine learning regression model for material synthesis prices prediction in agriculture. *Mater. Today Proc.* 2021, *in press*. [CrossRef]
20. Maes, W.H.; Steppe, K. Perspectives for Remote Sensing with Unmanned Aerial Vehicles in Precision Agriculture. *Trends Plant Sci.* **2019**, *24*, 152–164. [CrossRef] [PubMed]
21. Zaborowicz, M.; Foujd, A.; Boniecki, P.; Przybył, K.; Gierz, Ł.; Koszela, K.; Ślósarz, P.; Lisiak, D.; Przybył, J. Methodology of data processing in the process of neural image analysis of pork half carcasses. In Proceedings of the Tenth International Conference on Digital Image Processing (ICDIP 2018), Shanghai, China, 11–14 May 2018. [CrossRef]
22. Zaborowicz, M.; Boniecki, P.; Piekarska-Boniecka, H.; Koszela, K.; Mueller, W.; Górna, K.; Okoń, P. Neural classification of the selected family of butterflies. In Proceedings of the International Conference on Digital Image Processing (ICDIP 2017), Hong Kong, China, 19–22 May 2017. [CrossRef]
23. Elbeltagi, A.; Nagy, A.; Mohammed, S.; Pande, C.B.; Kumar, M.; Bhat, S.A.; Zsembeli, J.; Huzsvai, L.; Tamás, J.; Kovács, E.; et al. Combination of Limited Meteorological Data for Predicting Reference Crop Evapotranspiration Using Artificial Neural Network Method. *Agronomy* **2022**, *12*, 516. [CrossRef]
24. Zaborowicz, M.; Boniecki, P.; Koszela, K.; Przybył, J. Application of neural image analysis in evaluating the quality of greenhouse tomatoes. *Sci. Hortic.* **2017**, *218*, 222–229. [CrossRef]
25. Boniecki, P.; Dach, J.; Nowakowski, K.; Jakubek, A. Neural image analysis of maturity stage during composting of sewage sludge. In Proceedings of the International Conference on Digital Image Processing (ICDIP 2009), Bangkok, Thailand, 7–9 March 2009. [CrossRef]
26. Moody, J.; Darkin, C.J. Fast learning in networks of locally-tuned processing units. *Neural Comput.* **1989**, *1*, 281–294. [CrossRef]
27. Hinton, G.; Osindero, S.; Tech, Y.-W. A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef]
28. Wawrzyniak, A.; Przybylak, A.; Sujak, A.; Boniecki, P. Neural Modelling in the Exploration of the Biomethane Potential from Cattle Manure: A Case Study on Herds Structure from Wielkopolskie, Podlaskie, and Mazowieckie Voivodeships in Poland. *Sensors* **2023**, *23*, 164. [CrossRef]
29. Boniecki, P.; Zaborowicz, M.; Sujak, A. Comparison of MLP and RBF neural models on the example graphical classification. In Proceedings of the International Conference on Digital Image Processing (ICDIP 2021), Singapore, 20–21 June 2021. [CrossRef]
30. Pentoś, K.; Mbah, J.T.; Pieczarka, K.; Niedbała, G.; Wojciechowski, T. Evaluation of Multiple Linear Regression and Machine Learning Approaches to Predict Soil Compaction and Shear Stress Based on Electrical Parameters. *Appl. Sci.* **2022**, *12*, 8791. [CrossRef]
31. Piekutowska, M.; Niedbała, G.; Piskier, T.; Lenartowicz, T.; Pilarski, K.; Wojciechowski, T.; Pilarska, A.A.; Czechowska-Kosacka, A. The Application of Multiple Linear Regression and Artificial Neural Network Models for Yield Prediction of Very Early Potato Cultivars before Harvest. *Agronomy* **2021**, *11*, 885. [CrossRef]

32. Sabzi-Nojadeh, M.; Niedbała, G.; Younessi-Hamzekhanlu, M.; Aharizad, S.; Esmaeilpour, M.; Abdipour, M.; Kujawa, S.; Niazian, M. Modeling the Essential Oil and Trans-Anethole Yield of Fennel (Foeniculum vulgare Mill. var. vulgare) by Application Artificial Neural Network and Multiple Linear Regression Methods. *Agriculture* **2021**, *11*, 1191. [CrossRef]
33. Gorzelany, J.; Belcar, J.; Kuźniar, P.; Niedbała, G.; Pentoś, K. Modelling of Mechanical Properties of Fresh and Stored Fruit of Large Cranberry Using Multiple Linear Regression and Machine Learning. *Agriculture* **2022**, *12*, 200. [CrossRef]
34. Petrakis, T.; Kavga, A.; Thomopoulos, V.; Argiriou, A.A. Neural Network Model for Greenhouse Microclimate Predictions. *Agriculture* **2022**, *12*, 780. [CrossRef]
35. Altalak, M.; Ammaduddin, M.; Alajmi, A.; Rizg, A. Smart Agriculture Applications Using Deep Learning Technologies: A Survey. *Appl. Sci.* **2022**, *12*, 5919. [CrossRef]
36. Abdullahi, H.S.; Sheriff, R.E.; Mahieddine, F. Convolution neural network in precision agriculture for plant image recognition and classification. In Proceedings of the 2017 Seventh International Conference on Innovative Computing Technology (INTECH), Luton, UK, 16–18 August 2017; pp. 1–3. [CrossRef]
37. Escamilla-García, A.; Soto-Zarazúa, G.M.; Toledano-Ayala, M.; Rivas-Araiza, E.; Gastélum-Barrios, A. Applications of Artificial Neural Networks in Greenhouse Technology and Overview for Smart Agriculture Development. *Appl. Sci.* **2020**, *10*, 3835. [CrossRef]
38. Yao, C.; Zhang, Y.; Zhang, Y.; Liu, H. Application of Convolutional Neural Network in Classification of High Resolution Agricultural Remote Sensing Images. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *42*, 989–992. [CrossRef]
39. Li, C.; Zhen, T.; Li, Z. Image Classification of Pests with Residual Neural Network Based on Transfer Learning. *Appl. Sci.* **2022**, *12*, 4356. [CrossRef]
40. Jiao, L.; Dong, S.; Zhang, S.; Xie, C.; Wang, H. AF-RCNN: An anchor-free convolutional neural network for multi-categories agricultural pest detection. *Comput. Electron. Agric.* **2020**, *174*, 105522. [CrossRef]
41. Zhu, W.; Sun, J.; Wang, S.; Shen, J.; Yang, K.; Zhou, X. Identifying Field Crop Diseases Using Transformer-Embedded Convolutional Neural Network. *Agriculture* **2022**, *12*, 1083. [CrossRef]
42. Broomhead, D.S.; Lowe, D. Multivariable functional interpolation and adaptive networks. *Complex Syst.* **1988**, *2*, 321–355.
43. Piekarska-Boniecka, H.; Wilkaniec, B.; Dolańska-Niedbała, E. Parasitic wasps of the pimplinae subfamily (hymenoptera, ichneumonidae) of agricultural landscape refugium habitats in central Wielkopolska. *Acta Sci. Pol. Hortorum Cultus* **2008**, *7*, 23–29.
44. Boniecki, P.; Sujak, A.; Pilarska, A.A.; Piekarska-Boniecka, H.; Wawrzyniak, A.; Raba, B. Dimension Reduction of Digital Image Descriptors in Neural Identification of Damaged Malting Barley Grains. *Sensors* **2022**, *22*, 6578. [CrossRef]
45. Boniecki, P.; Weres, J. Neural Networks as a Tool in the Analysis of Agricultural Engineering Empirical Systems. *J. Res. Appl. Agric. Eng.* **2001**, *46*, 73–76.
46. Khalifani, S.; Darvishzadeh, R.; Azad, N.; SeyedRahmani, R. Prediction of sunflower grain yield under normal and salinity stress by RBF, MLP and, CNN Models. *Ind. Crops Prod.* **2022**, *189*, 115762. [CrossRef]
47. Hong, H.; Zhang, Z.; Guo, A.; Shen, L.; Sun, H.; Liang, Y.; Wu, F.; Lin, H. Radial Basis Function Artificial Neural Network (RBF Ann) as well as the hybrid method of RBF Ann and Grey Relational Analysis able to well predict trihalomethanes levels in tap water. *J. Hydrol.* **2020**, *591*, 125574. [CrossRef]
48. Deng, Y.; Zhou, X.; Shen, J.; Xiao, G.; Hong, H.; Lin, H.; Wu, F.; Liao, B.-Q. New methods based on back propagation (BP) and radial basis function (RBF) Artificial Neural Networks (Anns) for predicting the occurrence of haloketones in tap water. *Sci. Total Environ.* **2021**, *772*, 145534. [CrossRef]
49. Bourland, H.; Kamp, Y. Auto-association by multilayer perceptrons and singular value decomposition. *Biol. Cybern.* **1988**, *59*, 204–291. [CrossRef]
50. Dach, J.; Czekała, W.; Boniecki, P.; Lewicki, A.; Piechota, T. Specialised internet tool for biogas plant modelling and marked analysing. In Proceedings of the International Conference on Digital Image Processing (ICDIP 2014), Singapore, 9–10 February 2014. [CrossRef]
51. Speckt, D.F. Probabilistic Neural Networks. *Neural Netw.* **1990**, *3*, 109–118. [CrossRef]
52. Boniecki, P.; Raba, B.A.; Pilarska, A.; Sujak, A.; Zaborowicz, M.; Pilarski, K.; Wojcieszak, D. Neural Reduction of Image Data in Order to Determine the Quality of Malting Barley. *Sensors* **2021**, *21*, 5696. [CrossRef]
53. Toosi, N.B.; Soffianian, A.R.; Fakheran, S.; Waser, L.T. Mapping Disturbance in Mangrove Ecosystems: Incorporating Landscape Metrics and PCA-Based Spatial Analysis. *Ecol. Indic.* **2022**, *136*, 108718. [CrossRef]
54. Kramer, M.A. Nonlinear principal components analysis using autoassociative neural networks. *AIChe J.* **1991**, *37*, 233–243. [CrossRef]
55. Krevh, V.; Filipović, L.; Petošić, D.; Mustać, I.; Bogunović, I.; Butorac, J.; Kisić, I.; Defterdarović, J.; Nakić, Z.; Kovač, Z.; et al. Long-Term Analysis of Soil Water Regime and Nitrate Dynamics at Agricultural Experimental Site: Field-Scale Monitoring and Numerical Modeling Using HYDRUS-1D. *Agric Water Manag.* **2023**, *275*, 108039. [CrossRef]
56. Lu, N.; Niu, J.; Kang, S.; Singh, S.K.; Du, T. A hybrid PCA-Sem-Ann Model for the prediction of water use efficiency. *Ecol. Model.* **2021**, *460*, 109754. [CrossRef]
57. Kebonye, N.M.; Eze, P.N.; Agyeman, P.C.; John, K.; Ahado, S.K. Efficiency of the T-Distribution Stochastic Neighbor Embedding Technique for Detailed Visualization and Modeling Interactions between Agricultural Soil Quality Indicators. *Biosyst Eng.* **2021**, *210*, 282–298. [CrossRef]

58. Boniecki, P. The Kohonen Neural Network in Solving Classification Problems in Agricultural Engineering. *J. Res. Appl. Agric Eng.* **2005**, *50*, 37–41.
59. Ndikumana, E.; Ho Tong Minh, D.; Baghdadi, N.; Courault, D.; Hossard, L. Deep Recurrent Neural Network for Agricultural Classification using multitemporal SAR Sentinel-1 for Camargue, France. *Remote Sens.* **2018**, *10*, 1217. [CrossRef]
60. Boniecki, P.; Piekarska-Boniecka, H. The SOFM Neural network in the process of identification of selected orchard pests. *J. Res. Appl. Agric. Eng.* **2004**, *49*, 5–10.
61. Boniecki, P.; Idzior-Haufa, M.; Pilarska, A.; Pilarski, K.; Kolasa-Wiecek, A. Neural classification of compost maturity using artificial neural network type Self-Organizing Feature Map and algorithm. *IJERPH* **2019**, *16*, 3294. [CrossRef]
62. Hakimi-Asiabar, M.; Ghodsypour, S.H.; Kerachian, R. Multi-Objective Genetic Local Search Algorithm Using Kohonen's Neural Map. *Comput. Ind. Eng.* **2009**, *56*, 1566–1576. [CrossRef]
63. Van de Schoot, R.; Depaoli, S.; King, R.; Kramer, B.; Märtens, K.; Tadesse, M.G.; Vannucci, M.; Gelman, A.; Veen, D.; Willemsen, J.; et al. Bayesian statistics and modelling. *Nat. Rev. Methods Prim.* **2021**, *1*, 1. [CrossRef]
64. Imholz, M.; Vandepitte, D.; Moens, D. Bayesian estimation of interval bounds based on limited data. In Proceedings of the ISMA 2018-USD 2018, Leuven, Belgium, 17–19 September 2018; KU Leuven, Department of Mechanical Engineering: Leuven, Belgium, 2019; pp. 5207–5214.
65. Krause, P.J.; Bokinala, V. A Tutorial on Data Mining for Bayesian Networks, with a Specific Focus on IoT for Digital Agriculture. *Internet Things* **2023**, *22*, 100738. [CrossRef]
66. Drury, B.; Valverde-Rebaza, J.; Moura, M.-F.; de Andrade Lopes, A. A survey of the applications of Bayesian networks in agriculture. *Eng. Appl. Artif. Intell.* **2017**, *65*, 29–42. [CrossRef]
67. Patterson, D. *Artificial Neural Networks*; Prentice Hall: Singapore, 1996.
68. Sujak, A.; Kusz, A.; Rymarz, M.; Kitowski, I. Environmental Bioindication Studies by Bayesian Network with Use of Grey Heron as Model Species. *Environ. Model. Assess.* **2017**, *22*, 103–113. [CrossRef]
69. Yin, J.; Medellín-Azuara, J.; Escriva-Bou, A.; Liu, Z. Bayesian machine learning ensemble approach to quantify model uncertainty in predicting groundwater storage change. *Sci. Total Environ.* **2021**, *769*, 144715. [CrossRef] [PubMed]
70. Radomyski, A.; Ashauer, R. A Site-Specific Indicator of Nitrogen Loads into Surface Waters from Conventional and Conservation Agriculture Practices: Bayesian Network Model. *Ecol. Indic.* **2022**, *145*, 109641. [CrossRef]
71. Wu, H.; Su, X.; Singh, V.P.; AghaKouchak, A.; Liu, Z. Bayesian vine copulas improve agricultural drought prediction for long lead times. *Agric. For. Meteorol.* **2023**, *331*, 109326. [CrossRef]
72. Lachaud, M.A.; Bravo-Ureta, B.E. A bayesian statistical analysis of return to agricultural R&D investment in Latin America: Implications for food security. *Technol. Soc.* **2022**, *70*, 102054.
73. Boniecki, P.; Mueller, W. Expectation crops of chosen agricultural fetuses with the help of neural model by time series. *J. Res. Appl. Agric Eng.* **2006**, *51*, 40–44.
74. Liu, G.; Zhong, K.; Li, H.; Chen, T.; Wang, Y. A State of Art Review on Time Series Forecasting with Machine Learning for Environmental Parameters in Agricultural Greenhouses. *Inf. Process. Agric.* 2022, *in press*. [CrossRef]
75. Tofu, D.A.; Mengistu, M. Observed time series trend analysis of climate variability and smallholder adoption of new agricultural technologies in west Shewa, Ethiopia. *Sci. Afr.* **2023**, *19*, e01448. [CrossRef]

agriculture

MDPI

Article

Machine-Learning Approach to Non-Destructive Biomass and Relative Growth Rate Estimation in Aeroponic Cultivation

Oskar Åström [1], Henrik Hedlund [2] and Alexandros Sopasakis [1,*]

[1] Department of Mathematics, Faculty of Science, Lund University, 221 00 Lund, Sweden
[2] Alovivum AB, Göingegatan 6, 222 41 Lund, Sweden
* Correspondence: alexandros.sopasakis@math.lth.se

Abstract: We train and compare the performance of two machine learning methods, a multi-variate regression network and a ResNet-50-based neural network, to learn and forecast plant biomass as well as the relative growth rate from a short sequence of temporal images from plants in aeroponic cultivation. The training dataset consists of images of 57 plants taken from two different angles every hour during a 5-day period. The results show that images taken from a top-down perspective produce better results for the multi-variate regression network, while images taken from the side are better for the ResNet-50 neural network. In addition, using images from both cameras improves the biomass estimates from the ResNet-50 network, but not those from the multi-variatemultivariate regression. However, all relative growth rate estimates were improved by using images from both cameras. We found that the best biomass estimates are produced from the multi-variate regression model trained on top camera images using a moving average filter resulting in a root mean square error of 0.0466 g. The best relative growth rate estimates were produced from the ResNet-50 network training on images from both cameras resulting in a root mean square error of 0.1767 g/(g·day).

Keywords: machine learning; aeroponics; hydroculture; neural network; regression; biomass; fresh weight; relative growth rate; image analysis

Citation: Åström, O.; Hedlund, H.; Sopasakis, A. Machine-Learning Approach to Non-Destructive Biomass and Relative Growth Rate Estimation in Aeroponic Cultivation. *Agriculture* **2023**, *13*, 801. https://doi.org/10.3390/agriculture13040801

Academic Editors: Sebastian Kujawa and Gniewko Niedbała

Received: 17 February 2023
Revised: 16 March 2023
Accepted: 21 March 2023
Published: 30 March 2023

1. Introduction

In conventional farming, the soil is used to transfer nutrients to the plant. Although this is possibly the cheapest approach, it is not necessarily the most effective for plant growth or even the most environmental. Large amounts of water, for instance can be lost into the surroundings due to evaporation and dispersion into groundwater [1]. Furthermore, fertilizers used in field farming are also transported away from their intended targets to the environment while leading to nutrient waste and water eutrophication [2]. In addition, farming on fields leaves the plants exposed to a number of harmful factors, such as droughts, soil-borne diseases, bad weather, pest attacks and floods [1,3].

An alternative to soil-based farming is hydroculture, where the nutrients are carried to the plants using water. In particular, hydroponic farming has seen a large increase in commercial use in the last few years [4] with companies, such as Nordic Harvest, Swegreen and Gronska, being some recent Nordic examples.

Aeroponic cultivation is another soil-less farming method that also uses water as the nutrient carrier. Unlike regular hydroculture, however, which uses flowing water, aeroponics uses an aerosol to deliver the nutrients. The first direct benefit of this approach is reduced water usage, which is of great interest for several climate affected regions finding themselves under drought conditions. As a comparison, the water usage of 1 kg of tomatoes is 200–400 L in conventional soil farming, 70 L in hydroponics and 20 L in aeroponics [5].

Aeroponics, however, has an advantage over hydroponics in terms of plant growth rates. A limiting factor of hydroponic farming is that the oxygen content in the water must

be a maximum of 8 ppm. There are no such constraints, however, in aeroponics and it has been shown [6] that the aeration of the plant roots is much better leading to a higher growth rate.

The current emergence of hydrocultural practices, such as aeroponics, is one of many possible paths toward securing sustainable food production in the future. The ability to have controlled growth conditions, which such practices entail, makes it possible to optimize the cost-to-yield in any number of different metrics of importance: water, nutrition, land-use, emissions, or finances. However, in order to make these optimizations, it is necessary to determine the effects that parameters, such as temperature, pH, lighting, space, CO_2, and nutrient concentration, have on the development of plants. These experiments often require destructive measurements of the plant's characteristics by harvesting them. This prevents the plant from growing further and only results in one data point per plant. This method is not only labor intensive but results in poor estimates on an individual level and can only contribute to an estimate of the population as a whole. In order to estimate the effects on the scale of individual plants, continuous and non-destructive biomass estimates are the preferable choice.

The goal of this work is to evaluate two different image-based machine-learning methods for estimating plant growth in aeroponic farming. These two methods are: multivariate regression (MVR) and neural networks using a pre-trained ResNet-50 (R-50) network. The aim is for these methods to estimate the biomass (g) of a given plant as well as the Relative Growth Rate (RGR) measured in $g/(g \cdot day)$. With a sufficiently accurate estimate of RGR, a virtual sensor could be developed to measure the RGR continuously. This would allow for real-time optimization of plant growth, which could decrease the amount of unnecessary resources spent, including water, nutrients, lighting, electricity and labor.

A high RGR is desirable both for industry and research [7] and non-invasive methods that can achieve that result are desirable. It is well-known, however that [8] RGR varies between plant individuals. As a result, estimating the RGR requires multiple time data points per plant. Measurements based on removing parts of the plant can impact or even stop growth. Destructive methods for measuring RGR are therefore not useful for practical industrial applications and may be the reason behind the lack of studies in this area. Non-destructive machine learning approaches such as those we propose here are therefore needed.

We begin in Section 2 with the state of the art in this field. Both of the methods used, the multi-variate regression network and the ResNet-50-based neural network, as well as the image collection practices used are presented in Section 3. We present our results in Section 4. We discuss these results in Section 5 and end with our conclusions in Section 6.

2. State of the Art and Challenges

While the modern incarnations of aeroponics have existed since the mid-1900s [9], the use of aeroponic platforms in conjunction with machine learning (ML) is still in its infancy [10]. As recently as 2022, in the field of deep learning, there were only "14 publications on hydroponic agriculture, one in aquaponics and none on aeroponics for soil-less growing media" as reported in [10]. There are for instance, some studies [11,12] that employed image-to-biomass estimation in hydroponic cultivation. In that respect, it is now possible to find public datasets for training machine-learning models although still from a single (top-down) viewpoint [13]. There exist studies in aeroponics that apply machine-learning methods in order to predict yield from manually measured features, such as the number of leaves and stem diameter [6]. However, there do not exist any studies investigating the potential of image analysis as a tool for biomass prediction in aeroponic cultivation [10,14].

In studies analyzing plant growth using machine learning, it is common to use a classification output where the plant growth is lumped into a number of stages of progression [7,15,16], corresponding to different visual and biological processes in the plants. This makes the learning easier since the targets are limited to discrete values, and usually one is

only interested in a rough estimate of the plant's state of growth. However, for the purpose of estimating the relative growth rate of biomass, we need to be as precise as possible, and therefore a numeric value of the biomass is necessary.

The success or failure of ML approaches depends heavily on the available data. For some applications, collecting the needed data may result in gradually destroying the source of the data at the same time. This is the case in several cell-staining applications where, in order to understand the structures in the cell, one must apply chemicals that illuminate these structures and also destroy them at the same time. These issues also occur while collecting the data needed to measure biomass in plant cultivation studies. Classical approaches typically require data to be collected through weighing by removing leaves. This is a highly manual task, which naturally also results in a low frequency of data points.

In these cases, reported collection frequencies range from hree to four times a week [11] to once at the end of the growth period [12]. This results in a low number of data points and provides less information about growth patterns. In order to achieve a higher time resolution, we exploit a known fact in plant physiology. It has been shown [17] that, in controlled environments where nutrients are provided with free access, the relative growth rate (RGR) is constant. This enables the interpolation of biomass between data points, reducing the need for frequent data gathering and increasing the accuracy of biomass estimation methods.

Another challenge in most approaches dealing with estimating plant growth from images is occlusion [11,12]. Plants tend to obscure other plants as they grow, making individual plant growth more difficult to estimate exactly at the moment it is needed the most. Occlusion, which is thought of as an image data-collection problem, indirectly implies that plants may, in fact, be affecting the growth rate of other plants. For this reason, we consider and analyze the effects of having multiple viewpoints as input.

3. Data Processing and Methods

The main goal of this paper is to compare machine-learning methods for estimating the biomass of a plant given a set of images. The first method that we consider is Multi-Variate Regression (MVR), which employs linear regression on a set of manually curated non-structural features. This is similar to methods employed by Jung, Dae-Hyun, et al. in [12] for estimating biomass in hydroponically grown lettuce. The second approach we consider is based on a convolutional neural network (CNN) architecture, where we furthermore employ transfer learning with a pre-trained ResNet-50 neural network. We refer to this method as R-50. This method is similar to methods employed by N. Buxbaum et al. for hydroponically grown lettuce in [11].

Furthermore, two different models are generated and trained for each method, one for the top and one for the angled camera. The average of the estimates produced by these models is then used to construct a third estimate, called the Dual View. The resulting biomass estimates generated by these three views on the test set are then compared in three tasks:

- Task: Single Image. Biomass estimation from a single time point.
- Task: Moving Average. Biomass estimation from three consecutive time points.
- Task: RGR. Relative growth rate estimation from three randomly sampled time points.

The root mean square error (RMSE) of the results on these tasks is used to compare the different methods.

3.1. Experiment Setup

The setup used for plant growth and development consists of four growth beds and one reservoir. The reservoir contains nutrient-rich water, along with sensors for temperature, pH, and electrical conductivity and a heater to keep the water at a preferred set temperature. The water from the reservoir is then pumped to the four growth beds.

Each growth bed consists of a container with a removable lid. Inside the container, there are two sonicators, which contain membranes that vibrate at an ultrasonic frequency

that agitate the water into an aerosol, in the form of dense fog. The lid of the container has 24 holes with small baskets (plant holders), in which the plants are placed such that the roots hang down and are immersed in the aerosol, continuously replenishing the supply of water and nutrients. The bottom of the container has a drainage pipe that returns the water back to the reservoir, as shown in Figure 1. The plants were grown over a 5-day period, with 57 plants completing the full cycle.

The camera rig consists of 8 cameras (2 per bed) of the model PlexGear WC-800 (manufactured by PlexGear, Malmö, Skane, Sweden) set up as shown in Figure 1. Four of the cameras capture images from a top view, placed on the long edge of each bed. These cameras are referred to as the top cameras. The other four cameras are placed further down and over the adjacent bed so that they capture images from a lower angle. These cameras are referred to as the angled cameras.

The cameras are connected to a computer next to the rig, which runs a script to capture images with each camera at 1 h intervals. These images are then stored locally and sent directly to a cloud-sharing service so that the camera status and plant development can be monitored remotely.

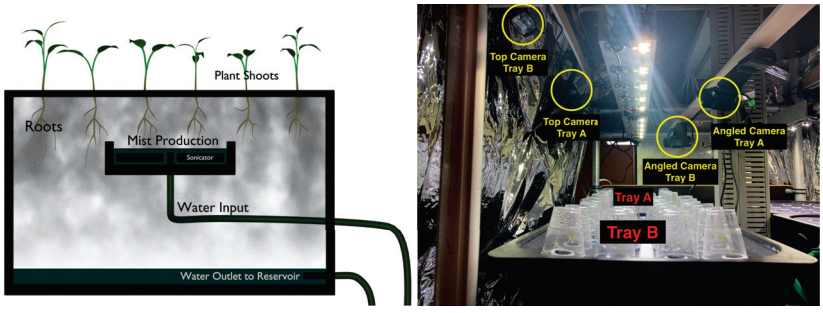


Figure 1. Growth bed schematic from the side (**left**). Camera rig for image capture (**right**).

3.2. Plant Physiology

The research conducted at the Swedish University of Agricultural Sciences (SLU) during a long period from the 1970s to the millennium shift has created a broad and solid scientific framework for controlled plant growth and development and, thereby, highly efficient and productive plant cultivation. This research is well-documented in scientific papers, dissertations and plant development databases [8,17–21] and provides a solid ground for our hypotheses and assumptions to be tested within the scope of this work.

One of the main assumptions in this work is that there exists a strong correlation and coherence between the relative rates of growth for biomass and leaf area. This has been verified in different experiments and is documented in [8,20], where biomass growth rates have been determined by weighing plants, roots and leaves, and leaf area growth rates have been determined by measuring the projected area via a scanner/copier. This correlation provides a strong indication that image analysis should be able to estimate biomass well.

Another assumption is that the RGR is constant for the plants grown for the data collection. A paper by O. Hellgren and T. Ingestad [17] demonstrated that, in controlled cultivation experiments with a constant relative addition rate of nutrients, plants have a constant relative growth rate. To achieve this, nutrients and water are supplied with free access, which implies non-limiting addition rates. This should ensure that the plant's RGR is close to constant, meaning that each gram of plant increases its mass by a constant amount per day.

3.3. Target Data

The ground target data is the biomass for a given plant at a given time point. This is gathered by weighing the plants at two time points. At each measurement, every plant is

weighed in its plant holder. The measured weight of the plant holder is then subtracted. These measurements then need to be extrapolated to every time point in the dataset. This is conducted through the assumption of a constant RGR, which is defined as

$$RGR = \dot{w}/w = a = const \Rightarrow \dot{w} = a \times w,$$

where $w(t)$ is the biomass. Since the growth rate is assumed to be linearly dependent on the biomass, then the biomass follows an exponential function. The log biomass is, therefore, in theory at least, a first-order polynomial with respect to time. Fitting this line to our measurements gives us an approximation of the log-biomass curve, which can be extrapolated to every time point. Note that the assumption of a constant RGR in the training set does not necessitate that future estimates on other plants require a constant RGR. This assumption only facilitates more efficient data gathering.

In reality, however, the plants exhibit an environmental shift when they are introduced to the cultivation platform. This means that the RGR could be lower early in the growth period before stabilizing later on to a constant value. The target data might, therefore initially differ slightly from the true biomass; however, they should eventually produce a reasonable approximation.

3.4. Input Images

The input data to the models consists of images of a specific plant at a given time during growth. These images were taken at 1 h intervals. At each such time point, a total of eight images are captured based on our two camera angles and the four growth beds. Each image covers an entire growth bed. These images are then transformed and grid-aligned through a projective transformation, shown in Figure 2, and then divided into segments, capturing a square around each plant. These images were resized to 64×64 pixels.

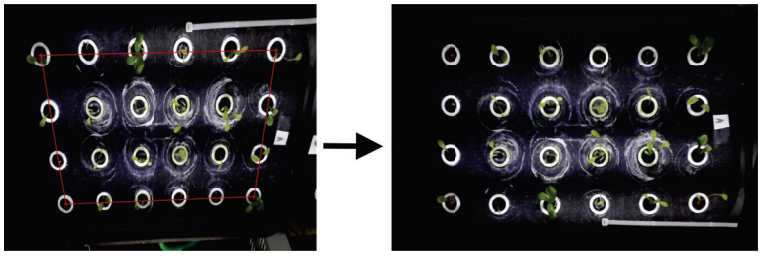


Figure 2. Full-tray image transformation and alignment.

In total , the dataset contains images of 57 plant individuals taken over a 5 day period, captured at 1 h intervals. This resulted in a total of 10,197 individual plant images. This dataset was split into training, validation and test sets. This split was performed based on individual plants, such that six randomly chosen individuals were placed in the validation and test sets, while the remaining 45 individuals were placed in the training set. This resulted in a [77.4%, 11.3% and 11.3%] split of the dataset. The dataset was further split into images from the top and angled camera respectively, resulting in the two datasets needed to train the two models for each ML method.

The dataset was pre-processed in a number of ways. First, the large degree of redundancy in the images increases the risk of over-training. This was combated through image augmentation by rotating each image by a random amount. A number of color spaces [22,23] were also however, the normal RGB color space was found to be optimal likely due to the pre-trained network having been trained on RGB images. The target biomass was also pre-processed through log transformation as this transformed the exponential biomass growth to a linear correlation. This led to the biomass being more evenly distributed. In addition, the log-biomass targets were normalized to the interval [0, 1]

based on the training set. The effect of these actions on the biomass distribution can be seen in Figure 3.

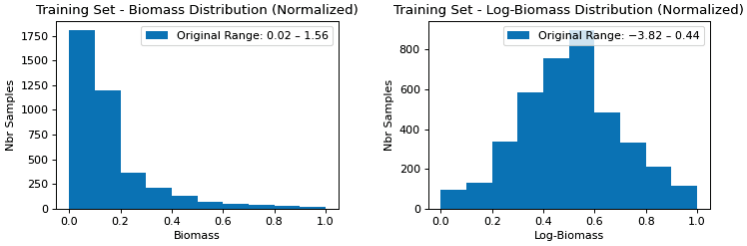


Figure 3. Effect of log-transform and normalization on biomass distribution for target data.

3.5. Method 1: Multi-Variate Regression

The first method used linear multi-variate regression on a number of features. The features used were all pixel-wise features summed over the image, meaning they all have the form

$$F_i(I) = \sum_{x=1}^{64} \sum_{y=1}^{64} f_i(I_{x,y}), \qquad f_i(p) : \mathbb{R}^3 \to \mathbb{R},$$

where I is an image and $I_{x,y}$ is a pixel in that image. The features used were inspired in part by features commonly used in plant segmentation [24,25]. The investigated pixel-wise features for the function f_i are shown in Table 1.

Table 1. Description of the 13 investigated features for the MVR model.

Set Name	Features	Description
Color	3	Amount of red, green or blue in pixel
Ratio	3	Ratio between R/G, G/B, B/R in pixel
Hue	3	Number of pixels with hue between R-G, G-B, B-R
Highlight	3	Same as 'Hue', but weighed by saturation
Vegetative Index [26]	1	$\text{Green} \times \text{Red}^{-2/3} \times \text{Blue}^{-1/3}$

To identify useful features, an iterative step-wise construction was employed. This was conducted by starting with an empty model containing only the intercept. All neighboring models, reached by adding or removing a feature, were evaluated, and the best model was chosen. This continued until the current model outperformed all of its neighbors. The models were compared using either the AIC (Akaike Information Criterion) or the Bayesian Information Criterion (BIC). Models were also created with respect to either targeting the biomass or the log biomass. This resulted in four models per camera view with respect to evaluation metric and a target unit. The models with the lowest RMSE on the validation set were chosen as the final model for each camera.

3.6. Method 2: R-50-Based Neural Network

The second method used is a convolutional neural network (CNN), inspired by previous research in biomass prediction [11]. The network utilizes a pre-trained image recognition network called ResNet-50 [27], trained on data from ImageNet [28], as a base followed by a custom-made regression head.

The R-50-based network relies on a residual block, which, in contrast to regular NN layers, in the left side of Figure 4, aims to predict the residual between the input to an imagined desired output. This is performed by adding the input to the residual block to the output as can be seen in the right side of Figure 4. This architecture allows for a very

deep NN without exhibiting the accuracy degradation that has been previously observed in such networks [27].

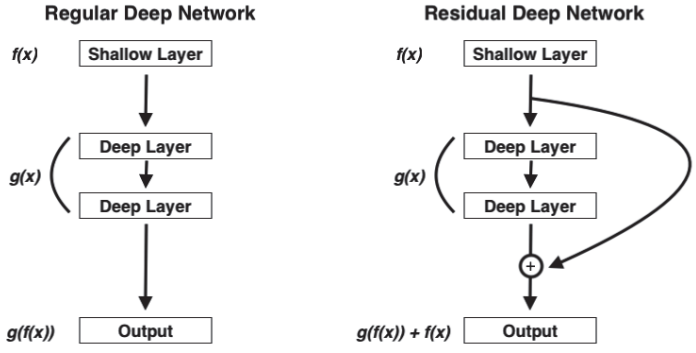


Figure 4. Architecture of a regular deep network (**left**) versus a residual deep network (**right**).

More specifically the R-50-based network includes a ResNet-50 network, which returns a feature vector of size 1000. This is then fed into a regression head, consisting of three densely connected layers of sizes 512, 128, and 1 respectively. The first two use ReLU activation, while the final output layer uses linear activation. Each network was trained for 20 epochs with a learning rate of 0.0005 and L2 regularization with a factor of 0.1. The pre-trained weights of the R-50-based network were also set to be trainable, so as to be fine-tuned for our problem. The architecture of the full network can be seen in Figure 5. Two models were created using images from the top and angled cameras respectively.

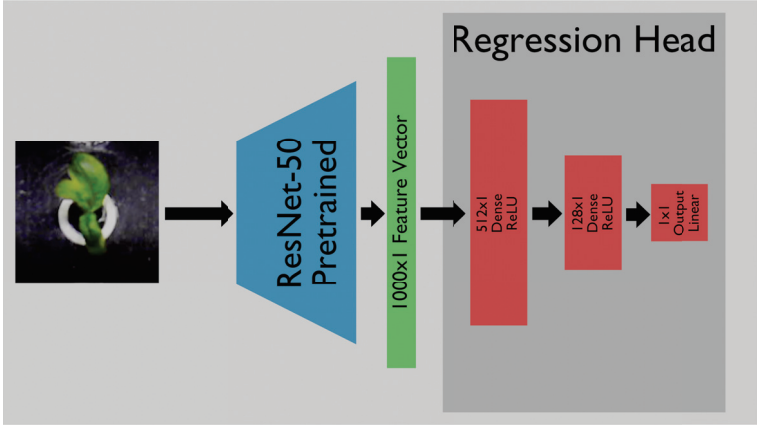


Figure 5. Graph of the architecture used for the R-50 network, including layer sizes, connection types, and activation function.

4. Results

In this section, we provide a summary of our main findings, which include training and testing results from the MVR and R-50-based network models as well as error comparisons of the different tasks depending on the number and type of image views undertaken.

4.1. Model Creation

It was found that AIC outperformed BIC on the validation set when constructing the MVR models. In addition, targeting the log biomass resulted in a higher accuracy on the validation set when using images from the top camera. Additionally, targeting

untransformed biomass was found to be superior for the angled camera. The parameter weights used in the final models for the top and angled camera are shown in Table 2.

Table 2. Features included in the two final MVR models.

Feature	Top View	Angled View
Intercept	0.1427	0.4567
R	−30.8498	−4.0938
G	29.9778	3.9349
B	-	-
R/G	−4.4577	−0.8282
G/B	−0.4435	1.0791
B/R	−4.3437	−0.7736
Hue_{RG}	0.8625	0.6988
Hue_{GB}	1.1628	0.4656
Hue_{BR}	-	-
HL_{RG}	−1.3134	-
HL_{GB}	−2.0212	-
HL_{BR}	6.2662	1.0448
VEG	-	−0.6336

Figure 6 shows the training process of the R-50-based network in terms of loss (MSE) for the top and angled cameras respectively.

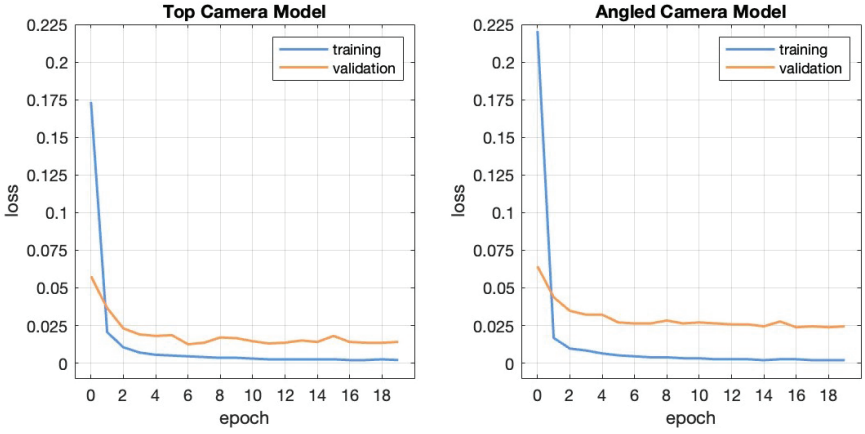


Figure 6. Training Loss (MSE) for the R-50 based network trained on images from the top (**left**) and angled (**right**) cameras respectively.

4.2. Method Comparison

Table 3 shows the RMSE on the test set for both the MVR network and R-50-based network. This table shows the quality of the method on a single image (SI) biomass prediction, the average (MA) of three chronologically consecutive biomass predictions, and RGR prediction using three random data points. Each of these tasks was performed using the model for the top view, angled view, and the average between their outputs (dual view). In addition, the RMSE is also presented in that table for the full test set as well as for some individual plants randomly chosen from the test set. These values are in the unit of g for biomass estimates and g/(g· day) for RGR estimates. For comparison, the true biomasses in the test set range up to 0.35 g and the true RGR range up to 0.17 g/(g· day).

Similarly, confidence intervals of the MSE on the test set with respect to camera view (Top, Angled, and Dual) and method (MVR and R-50 based network) are presented in Figure 7.

Table 3. RMSE on test set as a whole and a few, randomly picked, individual plants from the MVR network (above) and R-50 based network (below). Results are presented for all tasks: SI = Single Image, MA = Moving Average, RGR = Relative Growth Rate. The best method for a given task for each individual is highlighted in bold.

Multi-Variate Regression Method							
Task—View	All	#24	#33	#42	#64	#65	#78
SI—Top	**0.0466**	**0.0316**	**0.0260**	**0.0448**	**0.0573**	**0.0643**	0.0447
SI—Angled	0.1227	0.0924	0.0640	0.1347	0.2173	0.0960	0.0585
SI—Dual	0.0734	0.0521	0.0406	0.0882	0.1151	0.0751	**0.0363**
MA—Top	**0.0391**	**0.0253**	**0.0200**	**0.0426**	**0.0382**	**0.0614**	0.0332
MA—Angled	0.1178	0.0875	0.0598	0.1336	0.2089	0.0922	0.0479
MA—Dual	0.0703	0.0491	0.0380	0.0875	0.1096	0.0729	**0.0287**
RGR—Top	0.2268	0.2088	0.2038	0.1911	0.2413	0.2573	0.2500
RGR—Angled	0.3669	0.2816	0.2684	0.2261	0.2889	0.5410	0.4788
RGR—Dual	**0.1984**	**0.2049**	**0.1736**	**0.1658**	**0.2019**	**0.2383**	**0.1974**
ResNet-50 Method							
Task—View	All	#24	#33	#42	#64	#65	#78
SI—Top	0.0862	0.0339	0.0301	0.0261	0.1833	**0.0553**	**0.0744**
SI—Angled	0.0653	0.0265	0.0332	0.0284	0.0823	0.0809	0.0992
SI—Dual	**0.0550**	**0.0260**	**0.0228**	**0.0233**	**0.0735**	0.0639	0.0840
MA—Top	0.0812	0.0291	0.0278	**0.0166**	0.1752	**0.0502**	**0.0689**
MA—Angled	0.0637	**0.0214**	0.0304	0.0271	0.0804	0.0792	0.0985
MA—Dual	**0.0523**	0.0230	**0.0204**	0.0205	**0.0669**	0.0620	0.0829
RGR—Top	0.2647	0.2195	0.1982	0.3260	0.2444	0.2483	0.3243
RGR—Angled	0.1837	0.1910	0.1490	**0.1314**	**0.1818**	0.2294	0.2022
RGR—Dual	**0.1767**	**0.1434**	**0.1427**	0.1687	0.1867	**0.2078**	**0.2000**

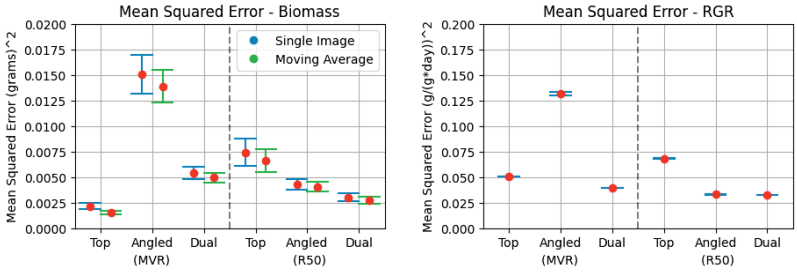


Figure 7. Confidence intervals (95%) for mean square errors in prediction on the test set using MVR or R-50 based network, with respect to camera view (top, angled, or dual). (**Left**): Biomass predictions in tasks Single Image and Moving Average. (**Right**): RGR predictions in task RGR.

5. Discussion

We trained a multi-variate regression and a R-50-based convolutional neural network, on a combination of images from different viewpoints and observed their performance towards estimating plant biomass and relative growth rate.

5.1. Multi-Variate Regression

It was found that the Akaike Information Criterion (AIC) metric outperformed the Bayesian Information Criterion (BIC) metric when generating MVR models. In [29], a comparison was made between AIC and BIC and they concluded that BIC had an advantage over AIC if the model used to generate the dataset was included in the set of possible

models. They also noteed, however, that, since data from the real world are too complicated, "the primary foundations of the BIC criteria do not apply in the biological sciences and medicine and the other 'noisy' sciences" [29]. This also holds true for our case, since it is reasonable to believe that the biomass of the plants was not generated using the pixel values of the captured images. This could be the reason that the AIC performs much better.

Overall, MVR performed substantially better on data from the top camera compared to the angled camera data as can be seen in Figure 7. This holds true on an individual level as can be seen in rows 1–2 in Table 3. This is reasonable since the top camera is a better representation of the leaf area and is thus a better representation of the biomass. In addition, the nature of the angled camera causes parts of other plants to be included in the images. Since MVR has no way of distinguishing between these pixels, there is an inherent flaw with applying this non-structural method to the images from the angled camera.

Similar to the biomass estimate, the RGR estimate was significantly better for the top camera compared to the angled camera. However, for the RGR estimates, the dual view was found to be superior. This shows that a poor estimate of biomass did not necessarily lead to a poor RGR estimate, even on an individual level. The selection of the three random samples could, on the other hand, be affected by outliers. This might cause the variance-reducing effect of the dual view to be more prominent in the RGR estimate than in the biomass estimate, thus leading to an improvement in the former, but not the latter.

5.2. The ResNet-50-Based Convolutional Neural Network

For the R-50-based network, the top camera underperformed than the angled camera as seen in Figure 7. However, this only holds true when looking at the overall RMSE. When investigating the RMSE on the individual plants in the lower section of Table 3, the top camera performed better on four out of six plants. Plant #64 had a surprisingly large RMSE in the top view, which inflates the overall RMSE.

As opposed to the results from the MVR, the dual view performed better than both the top and angled views. This indicates that having two perspectives is beneficial for the R-50-based network. The reason for this could be that the difference in MAE between the top and angled view is much smaller, making the variance-reducing effect of their average more prominent.

Another pattern that was found in the R-50-based network results is that the RGR estimate was better when using the dual view compared to the top or angled view.

5.3. Method Comparison

Comparing the RMSE between the two methods shows that the MVR performed best on the top camera, while the R-50-based network performed the best on the angled camera as can be seen in Figure 7. As mentioned, the reason that the MVR has a poor result on the angled view could be that a smaller fraction of the image consists of plant pixels, and that other plants can show up and distort the estimate. However, the additional structural information in the angled view, such as plant height, could be used by the R-50-based model which might be why it performs better.

The dual view inhibits some interesting behaviors. Since the dual view is created from the average of the top and angled camera, the performance is somewhat close to their average. However, it is always slightly lower. The reason for this is likely that the averaging also has a variance-reducing effect, leading to better estimates. We can see these effects in the left graph of Figure 7. This means that the dual view is superior when the difference between the top and angled cameras is small, such as in the R-50-based method. But if the difference is large, such as in the MVR method, the variance-reducing effect is not large enough to overcome the benefits of using only the superior camera.

In every case, the moving average estimate had a lower RMSE than the corresponding RMSE for the single-image task. This shows that there is a benefit to capturing multiple images at a high frequency, even though the biomass increase is minimal. This benefit likely comes from the fact that the flickering of the LED lights was noticeable in the images.

The noise from this flickering likely leads to some variation in the estimates, which is reduced by the moving average filter.

In the case of estimating the RGR, we found that, in general, the dual view was found to be superior for both methods.

5.4. Prediction Quality

The best result for the MVR model produced an RMSE of around 0.0391 g using the moving average filter and top camera images. With biomasses up to 0.35 g in our test set, this represents a relative RMSE of 11.2%, which is comparable to previous studies. In the paper by Wenjian Liu et al. [24], for instance, they achieved an RMSE of 0.32 g on fresh biomass samples up to 3 g, which corresponds to a relative RMSE of 10.7%. Other papers have achieved even better accuracy for datasets with larger plant biomass weights. The paper by N. Buxbaum et al. [11], for instance, used images of 3888 individuals and obtained an RMSE of between 1 and 2 g for plant biomass up to 40 g, corresponding to a relative RMSE of 2.5–5%.

We believe that a reason for this value of relative RMSE could be attributed to the large variance in the estimates of time-adjacent images due to the flicker of the LED lights creating visible variations between the images. Having a setup where the collected images are not affected by such external factors should therefore improve both the biomass as well as the RGR estimates.

There do not exist any aeroponic studies to compare these results against, but in general, the RGR estimates were not sufficiently accurate. Detecting relative changes in biomass could be easier for larger plants, as the visual difference is larger the further into the growth period the plants are. A dataset containing more individuals would likely also increase the accuracy of the predictions. It should be noted that the RGR estimates were made using three random points from the entire growth period. The resulting estimate could be considered to measure the 'average' RGR. Since our data were assumed to have a constant RGR this does not matter. However, for conditions where the RGR varies over time (for example, if the conditions change during growth), the RGR should instead be constructed from a time series with a shorter time span depending on the time resolution desired.

6. Conclusions

In this work, we trained two different machine-learning models, a multi-variate regression model (MVR) and a ResNet-50 (R-50) neural network, to discover growth patterns in plants based solely on camera images. We then compared the abilitys of those models to forecast plant biomass and plant relative growth rate (RGR). Our proposed approach to estimate plant development, therefore, relies on non-destructive methods. As a result, it can be possible for farmers to intervene at an early stage if needed, in order to influence and improve growth even at the individual plant level.

We note also that, in general for any type of plant growing environment, as long as data is recorded consistently the resulting model will be able to learn. In our study, we only required a short series of images in order to estimate biomass. In that respect, changes in atmospheric or soil moisture did not influence the imaging of the plants which in turn are responsible for the accuracy of our estimates.

Based on the results in Section 5 above, we see that the biomass estimate can be improved greatly for both models when applying the moving average filter over the neighboring time points. Having multiple cameras does not seem to improve the estimates from the multi-variate regression model, but can improve the estimates based on the neural network model.

We also see that the best RGR estimates can be produced when images from both cameras are used to train the neural network model. More generally however, using images from both cameras improves the RGR estimate for both the multi-variate regression as well as the R-50 based neural network.

Although the quality of the resulting biomass predictions is comparable to other studies, the way these predictions are produced is significantly different. They are based on non-destructive data collection, and as a result, they can only be improved as more data become available for the study. There currently do not exist any other similar results investigating the ability of machine learning to predict RGR or biomass from images in aeroponics.

The findings in this study highlight promising patterns in camera and model behavior, such as the effects of moving average filters and multiple camera angles. Such approaches also outline the potential for future research into virtual sensors of RGR. Such sensors would contain the cameras as well as a computer analyzing the images using the models and transmitting the estimated biomass and RGR. Further research in this field, using datasets of more individuals over long periods of time and using more real measurements, is vital in order to verify these results.

Author Contributions: Conceptualization, O.Å. and H.H.; Formal analysis, O.Å., H.H. and A.S.; Funding acquisition, H.H. and A.S.; Investigation, O.Å., H.H. and A.S.; Methodology, O.Å., H.H. and A.S.; Project administration, O.Å., H.H. and A.S.; Resources, O.Å., H.H. and A.S.; Software, O.Å. and A.S.; Supervision, H.H. and A.S.; Validation, O.Å., H.H. and A.S.; Visualization, O.Å.; Writing—original draft, O.Å., H.H. and A.S.; Writing—review and editing, O.Å. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research is partially supported by grants from eSSENCE (number 138227), Vinnova (number 2020-03375), Formas 2022-00757, and the Swedish National Space Board.

Data Availability Statement: Data available on request due to restrictions eg privacy or ethical. The data presented in this study are available on request from the corresponding author. The data are not publicly available due to persons in the background of the side images.

Acknowledgments: The training and data handling was enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

Conflicts of Interest: The authors declare no conflict of interest. Some of the data in this study were provided by Alovivum AB. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine learning
RGR	Relative growth rate
MVR	Multi-variate regression
R-50	Neural network with ResNet-50 base
NN	Neural network
CNN	Convolutional neural network
MSE	Mean square error
RMSE	Residual mean square error

References

1. Olympios, C.M. Overview of soilless culture: Advantages, constraints and perspectives for its use in Mediterranean countries. *Cah. Options Méditerranéennes* **1999**, *1*, 307–324.
2. Ghorbel, R.; Chakchak, J.; Malayoğlu, H.; Çetin, N. Hydroponics "Soilless Farming": The Future of Food and Agriculture—A Review. In Proceedings of the 5th International Students Science CongressProceedings, Rome, Italy, 20–22 October 2021.
3. Sheikh, B.A. Hydroponics: Key to sustain agriculture in water stressed and urban environment. *Pak. J. Agric. Agric. Eng. Vet. Sci.* **2006**, *22*, 53–57.
4. Tunio, M.H.; Gao, J.; Shaikh, S.A.; Lakhiar, I.A.; Qureshi, W.A.; Solangi1, K.A.; Chandio, F.A. Potato production in aeroponics: An emerging food growing system in sustainable agriculture forfood security. *Chil. J. Agric. Res.* **2020**, *80*, 118–132. [CrossRef]
5. Ziegler, R. *The Vertical Aeroponic Growing System*; Synergy International Inc.: Sausalito, CA, USA, 2015.

6. Mokhtar, A.; El-Ssawy, W.; He, H.; Al-Anasari, N.; Sammen, S.S.; Gyasi-Agyei, Y.; Abuarab, M. Using Machine Learning Models to Predict Hydroponically Grown Lettuce Yield. *Front. Plant Sci.* **2022**, *13*, 706042. [CrossRef]
7. Gao, D.; Qiao, L.; An, L.; Zhao, R.; Sun, H.; Li, M.; Tang, W.; Wang, N. Estimation of spectral responses and chlorophyll based on growth stage effects explored by machine learning methods. *Crop J.* **2022**, *10*, 1292–1302.
8. Hedlund, H. *Temperature Distribution and Plant Responses of Birch (Betula Pendula Roth.) at Constant Growth*; Acta Universitatis Agriculturae Sueciae Agraria, Swedish University of Agricultural Sciences: Uppsala, Sweden, 1999.
9. Carter, W.A. A method of growing plants in water vapor to facilitate examination of roots. *Phytopathology* **1942**, *732*, 623–625.
10. Ojo, M.O.; Zahid, A. Deep Learning in Controlled Environment Agriculture: A Review of Recent Advancements, Challenges and Prospects. *Sensors* **2022**, *22*, 7965. [CrossRef]
11. Buxbaum, N.; Lieth, J.H.; Earles, M. Non-destructive Plant Biomass Monitoring With High Spatio-Temporal Resolution via Proximal RGB-D Imagery and End-to-End Deep Learning. *Front. Plant Sci.* **2022**, *13*, 758818. [CrossRef]
12. Jung, D.H.E.A. Image Processing Methods for Measurement of Lettuce Fresh Weight. *J. Biosyst. Eng.* **2015**, *40*, 89–93. [CrossRef]
13. Beck, M.A.; Liu, C.; Bidinosti, C.P.; Henry, C.J.; Godee, C.M.; Ajmani, M. Presenting an extensive lab- and field-image dataset of crops and weeds for computer vision tasks in agriculture. *arXiv* **2021**, arXiv:2108.05789.
14. Mehra, M.; Saxena, S.; Sankaranarayanan, S.; Tom, R.J.; Veeramanikandan, M. IoT based hydroponics system using Deep Neural Networks. *Comput. Electron. Agric.* **2018**, *155*, 473–486. [CrossRef]
15. Broms, C.; Nilsson, M.; Oxenstierna, A.; Sopasakis, A.; Åström, K. Combined analysis of satellite and ground data for winter wheat yield forecasting. *Smart Agric. Technol.* **2023**, *3*, 100107. [CrossRef]
16. Kumar, P.; Prasad, R.; Gupta, D.K.; Mishra, V.N.; Vishwakarma, A.K.; Yadav, V.P.; Bala, R.; Choudhary, A.; Avtar, R. Estimation of winter wheat crop growth parameters using time series Sentinel-1A SAR data. *Geocarto Int.* **2018**, *33*, 942–956.
17. Hellgren, O.; Ingestad, T. A comparison between methods used to control nutrient supply. *J. Exp. Bot.* **1996**, *47*, 117–122. [CrossRef]
18. Ingestad, T.; Hellgren, O.; Lund Ingestad, A. *Data Base for Birch Plants at Steady State*; Technical Report 75; Sveriges Lantbruksuniversitet Rapporter: Uppsala, Sweden, 1994 .
19. Hellgren, O.; Ingestad, T.; Lund Ingestad, A. *Data Base for Tomato Plants at Steady-State—Methods and Performance of Tomato Plants (Lycopersicon esculentum Mill cv Solentos) under Non-Limiting Conditions and under Limitation of Nitrogen and Light*; Technical Report 74; Institutionen foer Ekologi och Miljoevaard (Sweden): Uppsala, Sweden, 1994 .
20. Hellgren, O.; Ingestad, T. *Responses of Birch (Betula Pendula Roth) and Tomato Plants (Lycopersicon Esculentum Mill cv Solentos) to CO_2 Concentration and to Limiting and Non-Limiting Supply of CO_2*; Technical Report 3; Biotron, Swedish University of Agricultural Sciences: Uppsala, Sweden, 1996.
21. McDonald, A.J.S.; Lohammar, T.; Ingestad, T. Net assimilation rate and shoot area development in birch (Betula pendula Roth.) at different steady-state values of nutrition and photon flux density. *Trees* **1992**, *6*, 1–6. [CrossRef]
22. Praveen Kumar, J.; Domnic, S. Image based leaf segmentation and counting in rosette plants. *Inf. Process. Agric.* **2019**, *6*, 233–246. [CrossRef]
23. Yang, W.; Wang, S.; Zhao, X.; Zhang, J.; Feng, J. Greenness identification based on HSV decision tree. *Inf. Process. Agric.* **2015**, *2*, 149–160. [CrossRef]
24. Liu, W.; Li, Y.; Liu, J.; Jiang, J. Estimation of Plant Height and Aboveground Biomass of Toona sinensis under Drought Stress Using RGB-D Imaging. *Forests* **2021**, *12*, 1747. [CrossRef]
25. Lati, R.N.; Filin, S.; Eizenberg, H. Robust Methods for Measurement of Leaf-Cover Area and Biomass from Image Data. *Weed Sci.* **2011**, *59*, 276–284. [CrossRef]
26. Hague, T.; Tillett, N.D.; Wheeler, H.C. Automated Crop and Weed Monitoring in Widely Spaced Cereals. *Precis. Agric.* **2006**, *7*, 21–32. [CrossRef]
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)* **2016**, *322*, 770–778. [CrossRef]
28. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
29. Burnham, K.; Anderson, D. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*; Springer: Berlin/Heidelberg, Germany, 2002.

agriculture

MDPI

Article

Vegetation Indices for Predicting the Growth and Harvest Rate of Lettuce

Ana Luisa Alves Ribeiro [1], Gabriel Mascarenhas Maciel [2,*], Ana Carolina Silva Siquieroli [3], José Magno Queiroz Luz [4], Rodrigo Bezerra de Araujo Gallis [5], Pablo Henrique de Souza Assis [6], Hugo César Rodrigues Moreira Catão [4] and Rickey Yoshio Yada [7]

[1] Postgraduate Program in Agronomy, Institute of Agrarian Sciences, Federal University of Uberlândia, Uberlândia 38410-337, Brazil; analuisaribeiro@ufu.br
[2] Institute of Agrarian Sciences, Federal University of Uberlândia, Monte Carmelo 38500-000, Brazil
[3] Institute of Biotechnology, Federal University of Uberlândia, Monte Carmelo 38500-000, Brazil; carol@ufu.br
[4] Institute of Agrarian Sciences, Federal University of Uberlândia, Uberlândia 38410-337, Brazil; jmagno@ufu.br (J.M.Q.L.); hugo.catao@ufu.br (H.C.R.M.C.)
[5] Institute of Geography, Federal University of Uberlândia, Monte Carmelo 38500-000, Brazil; rodrigogallis@ufu.br
[6] Postgraduate Program in Agriculture and Geospatial Information, Institute of Agrarian Sciences, Federal University of Uberlândia, Monte Carmelo 38500-000, Brazil; pablohnrqsa@gmail.com
[7] Faculty of Land and Food Systems, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; r.yada@ubc.c
* Correspondence: gabrielmaciel@ufu.br

Abstract: Urbanization has provided greater demand for food, and the search for strategies capable of reducing waste is essential to ensure food security. Lettuce (*Lactuca sativa* L.) culture has a short life cycle and its harvest point is determined visually, causing waste and important losses. Using vegetation indices could be an important alternative to reduce errors during harvest definition. The objective of this study was to evaluate different vegetation indices to predict the growth rate and harvest point of lettuce. Twenty-five genotypes of biofortified green lettuce were evaluated. The Green Leaf Index (GLI), Normalized Green Red Difference Index (NGRDI), Spectral Slope Saturation Index (SI), and Overall Hue Index (HUE) were calculated from images captured at 1, 8, 18, 24, and 36 days after transplanting (vegetative state). The diameter and average leaf area of plants were measured using QGIS software. Green mass, number of leaves, and plant and stem diameter were measured in the field. The means were compared using the Scott–Knott test ($p \leq 0.05$) and simple linear regression models were generated to monitor the growth rate, obtaining R^2 values ranging from 62% to 99%. Genetic dissimilarity was confirmed by the multivariate analysis presenting a cophenetic correlation coefficient of 88.49%. Furthermore, validation between data collected in the field versus data obtained by imaging was performed using Pearson's correlations and showed moderate to high values. Overall, the vegetation indices SI, GLI, and NGRDI were efficient for monitoring the growth rate and determining the harvest point of different green lettuce genotypes, in attempts to reduce waste and losses. It is suggested that the definition of the harvest point based on vegetation indices are specific for each genotype.

Keywords: food safety; image phenotyping; *Lactuca sativa* L.; vegetables

Citation: Ribeiro, A.L.A.; Maciel, G.M.; Siquieroli, A.C.S.; Luz, J.M.Q.; Gallis, R.B.d.A.; Assis, P.H.d.S.; Catão, H.C.R.M.; Yada, R.Y. Vegetation Indices for Predicting the Growth and Harvest Rate of Lettuce. *Agriculture* **2023**, *13*, 1091. https://doi.org/10.3390/agriculture13051091

Academic Editors: Gniewko Niedbała and Sebastian Kujawa

Received: 20 March 2023
Revised: 25 April 2023
Accepted: 27 April 2023
Published: 19 May 2023

1. Introduction

Every strategy to improve food security is of fundamental importance. The interest and importance of these actions have increased, including in the scientific area. It is now a consensus opinion to say that food waste reflects directly on the lack of food and consequently on the price, causing hunger. It is estimated that food production should increase by 30% by 2030 due to population growth [1].

Among the main vegetables, lettuce (*Lactuca sativa* L.) stands out and is present daily in food. Lettuce, which belongs to the Asteraceae family, is considered an annual and herbaceous plant and is among the most popular and consumed vegetables in Brazil and worldwide. More than 1.5 million tons of this crop are produced in Brazil, and its activity is concentrated near the large centers called "green belts" [2,3].

With the country facing the search for a healthier diet, especially post-COVID-19, lettuce cultivation in Brazil has increased considerably. In this context, producers have increased cultivation areas. Despite all the benefits, growing lettuce presents great difficulty in defining the harvesting point, causing significant losses and waste [4], mainly due to the plant presenting a short cycle and early bolting [5].

The main parameter that defines the development of the lettuce crop is the number of leaves [6], and in small and large areas of cultivation, the growth rate and harvest point are identified visually. However, in large plantation areas, producers face difficulties in performing this monitoring, causing significant waste and losses [7]. The use of digital images collected by UAVs can be useful and assist in decision-making. Thus, new strategies to define the harvest point in lettuce are needed.

Image phenotyping has been used to assist in the selection and characterization of quantitative and qualitative variables in specific individuals through non-destructive analyses [8]. Unmanned aerial vehicles (UAVs) with attached cameras and sensors perform analyses and follow the stages of crop development, from the visible electromagnetic spectrum to the infrared spectrum [9]. The cost, time, and labor to obtain information in the field and laboratory are reduced when remote sensing is employed.

Vegetation indices are based on reflectance, and their values vary according to the vegetation cover and its biophysical characteristics [10]. There have been reports of the potential use of images in several plant species [11–16]. In lettuce, vegetation indices are being used to differentiate pigment levels present in the leaves and to estimate leaf area indices using infrared images [17–19]. For eucalyptus cultures, remote sensing has proven efficient at monitoring plantations using vegetation indices. Overall, remote sensing is a low-cost technique and can be applied in large extensions [20]. Studies of the prediction of growth rate and harvest point of vegetables from images have been insufficient.

In this context, the objective of this study was to evaluate different vegetation indices to predict the growth rate and harvest point of lettuce.

2. Materials and Methods

2.1. Genetic Material and Place of Experiment

The experiment was conducted at the Experimental Vegetable Station (18°42′43.19″ S and 47°29′55.8″ O, 873-m altitude) of the Federal University of Uberlândia (UFU), Monte Carmelo campus, Minas Gerais, Brazil.

Twenty-five genotypes were evaluated (Figure 1), with two commercial controls (cv. Grand Rapids and Uberlândia 10,000) and 23 biofortified tropicalized green lettuce lines belonging to the UFU germplasm bank registered with BG α BIOFORT Software [21].

The genotypes employed in this study were derived from seven successive self-fertilizations between the cultivars PIRA 72 and Uberlândia 10,000 from 2013 to 2018. The seedlings were produced in expanded polyethylene trays with 200 cells filled with coconut fiber commercial substrate. Transplanting was performed when the lettuce plants presented four definitive leaves. The seedlings were transferred to 1.3-m beds fabricated with a rotary bed former.

The following physical and chemical characteristics of the soil were assessed: clayey texture (>50%); pH in $CaCl_2$ = 4.9; Ca = 3.3 $cmol_c$ dm^{-3}; Mg = 1.3 $cmol_c$ dm^{-3}; H + Al = 4.9 $cmol_c$ dm^{-3}; SB = 4.90 $cmol_c$ dm^{-3}; SOM = 3.9 dag kg^{-1}; P (rem) = 79.1 mg dm^{-3}; K = 0.29 $cmol_c$ dm^{-3}, CEC = 9.80 $cmol_c$ dm^{-3}; and BS% = 50. Cultivation was performed as recommended for lettuce culture [22]. Climate conditions were monitored daily during the experiment.

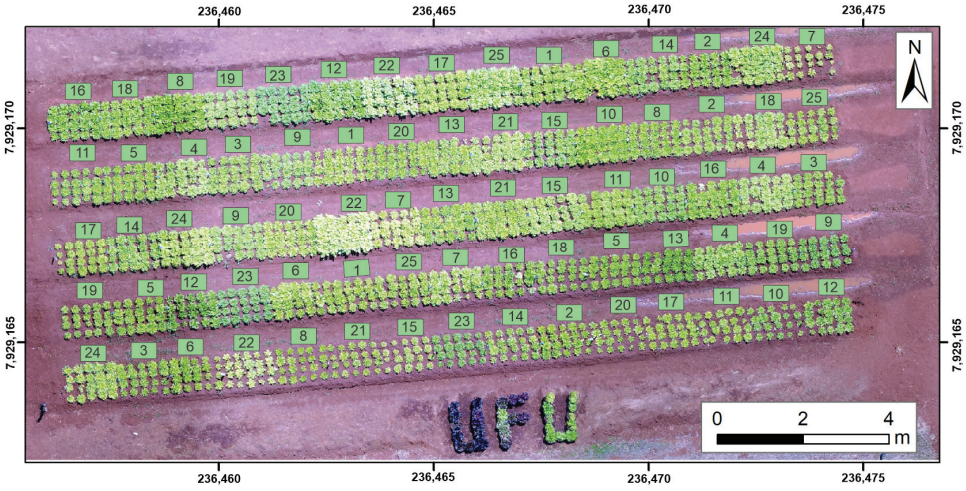


Figure 1. Distribution of green lettuce genotypes in the field. 1: UFU-206#1#6#1; 2: UFU BIOFORT189E8; 3: UFU-197#3#1#1; 4: UFU-125#1#1#1; 5: UFU-7#1#2#1; 6: UFU BIOFORT155E12; 7: UFU BIOFORT120E21; 8: UFU BIOFORT189E22; 9: UFU-197#2#1#1; 10: UFU-199#3#1#1; 11: UFU-206#1#1#1; 12: UFU BIOFORT206E32; 13: UFU BIOFORT197E34; 14: UFU-197#2#2#1; 15: UFU BIOFORT155E39; 16: UFU BIOFORT189E43; 17: UFU-206#1#4#1; 18: UFU-125#2#2#1; 19: UFU-206#1#2#1; 20: UFU BIOFORT189E48; 21: UFU-206#1#5#1; 22: UFU-040#5#5#1; 23: UFU MC BIOFORT; 24: Grand Rapids; and 25: Uberlândia 10,000.

The experiment was conducted in a randomized block design with three repetitions, totaling 75 plots. The plots consisted of 20 plants, with spacing of 0.25 × 0.25 m between plants.

2.2. *Acquisition and Processing of Aerial Images*

During the execution of the experiment, five flights were performed on different days after transplanting (DAT) (1, 8, 18, 24, and 36 DAT). The aerial images were captured using a Phantom 4 Advanced drone model, with a visible camera (RGB) that had a resolution of 20 megapixels.

Using DroneDeploy software, the flights were performed following the parameters of 20 m in height, 80% longitudinal overlap, and 75% lateral overlap. The orthoimage was generated using Pix4d software. The calculation of the vegetation indices (Table 1) and the image reclassification were performed using R software, version 3.6.3 [23], and the R package FieldImageR [24].

Table 1. Vegetation indices used in the experiment.

Vegetation Indices	Equation	Reference
SI—Spectral Slope Saturation Index	$\frac{R-B}{R+B}$	[25]
HUE—Overall Hue Index	$\text{atan}\left(2 \times \frac{(B-G-R)}{30.5 \times (G-R)}\right)$	[25]
GLI—Green Leaf Index	$\frac{(2 \times G - R - B)}{(2 \times G + R + B)}$	[26]
NGRDI—Normalized Green Red Difference Index	$\frac{G-R}{G+R}$	[27]

G = green band; R = red band; B = blue band

The Overall Hue Index (HUE) was calculated and used to form the mask layer and reclassify the RGB image, excluding the soil. After calculating the Green Leaf Index (GLI), Normalized Green Red Difference Index (NGRDI), and Spectral Slope Saturation Index (SI), the average index for each plot was obtained for all flights.

Growth rate monitoring was performed with non-destructive methods using imagery. In addition to vegetation indices, leaf area in software (LAS) and plant diameter in software (PDS) were extracted using QGIS software, version 3.0, for all flights. PDS was measured using the Raster Calculator tool with six central plants measured.

To obtain LAS, the pixel values in the green band were extracted from the RGB image. Using QGIS software, version 3.0, and the function r. recode, a classification from 1 to 0 could be assigned for the soil and plant, respectively. Thus, the contour of the plants was measured, enabling the calculation of the average leaf area in the respective plots.

2.3. Evaluation of Agronomic Data in the Field

At the commercial point (36 DAT), in addition to capturing images, the green mass (GM) was measured in the field by weighing the leaves, counting the number of leaves (NL), and determining plant diameter (PD) and stem diameter (SD). Six central plants from each plot were used for the evaluations.

2.4. Experimental Flowchart

The methodological steps, including image processing and data analysis, are presented in the experimental flowchart (Figure 2).

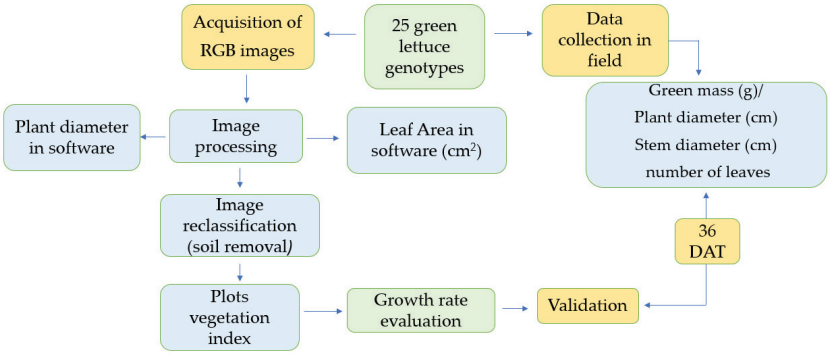


Figure 2. Flowchart of the experiment steps: data collection in the field and information obtained using images.

2.5. Statistical Analysis

The results measured in the field and the data from images extracted in the last flight (36 DAT) were subjected to analysis of variance using the F test ($p \leq 0.05$). Furthermore, means were compared using the Scott–Knott test ($p \leq 0.05$). A dendrogram was obtained through multivariate analysis using the Unweighted Pair-Group Method Using Arithmetic Averages (UPGMA). This method was applied to prove the genetic diversity among the treatments of the experiment. To validate vegetation indices, experiments with demonstrably dissimilar treatments (greater number of branches of the dendrogram) were necessary, increasing the spectrum of efficiency of the indices evaluated [17].

Pearson's correlation between the variables collected in the field and from the images was determined individually between the genotypes (per se), and the significance of the coefficients was verified. Simple linear regression models were generated after the observation of the correlations to monitor the growth rate and the ideal harvest point for the lettuce culture. Graphs were generated for the response variables LAS and PDS and the vegetation indices GLI, NGRDI, and SI at different DAT. Statistical analyses were performed using R [23] and Genes version 1990.2019.91 [28] software.

3. Results

During the experiment, maximum temperatures ranged from 18.8 °C to 32 °C and minimum temperatures ranged from 5.7 °C to 20.9 °C (Figure 3).

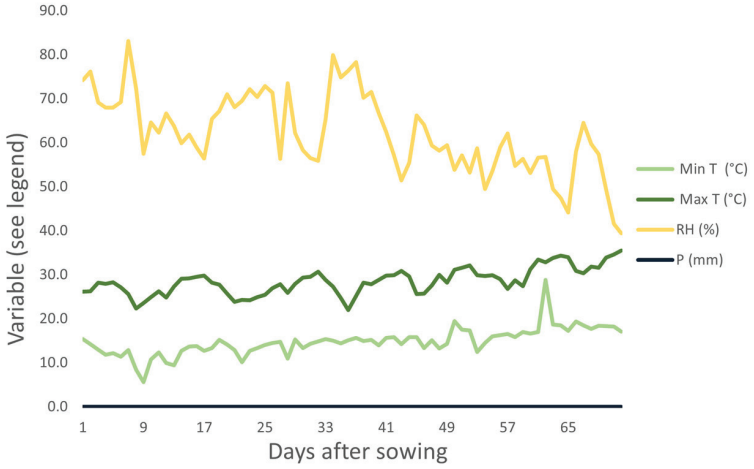


Figure 3. Climate conditions during the execution of the experiment (August and September 2019) in Monte Carmelo, Minas Gerais, Brazil. Humidity (RH), minimum temperature (Min Temp), maximum temperature (Max Temp), and precipitation (P).

3.1. Germplasm Evaluation

The lettuce genotypes differed from one another for the vegetation indices GLI, SI, and NGRDI and the variable NL (Table 2). There were no differences in the characteristics of GM, PD, and SD.

NL was highlighted for the genotypes UFU BIOFORT189E8, UFU-125#1#1#1, UFU-7#1#2#1, UFU BIOFORT155E12, UFU BIOFORT189E22, UFU-197#2#1#1, UFU-199#3#1#1, UFU-206#1#1#1, UFU BIOFORT206E32, UFU BIOFORT197E34, UFU BIOFORT189E43, UFU-125#2#2#1, UFU BIOFORT189E48, UFU-040#5#5#1, and Uberlândia 10,000, the values of which were superior to those of the other genotypes and the commercial control cv. Grand Rapids (Table 2). The increase in NL, compared with that in cv. Grand Rapids, was 45.4% for UFU BIOFORT189E22, 44.9% for UFU BIOFORT189E43, and 42.35% for UFUBIOFORT197E34.

The genotypes showed different behaviors ($p \leq 0.05$) among the vegetation indices SI, GLI, and NGRDI (Table 3). UFU-206#1#6#1, UFU-125#1#1#1, UFU BIOFORT120E21, UFU-197#2#1#1, UFU BIOFORT155E39, UFU BIOFORT189E48, UFU-206#1#5#1, and UFU-040#5#5#1 did not differ from genotype Uberlândia 10,000 or the commercial cultivar Grand Rapids in the vegetation index SI.

For GLI, the genotypes UFU BIOFORT189E8, UFU-197#3#1#1, UFU-7#1#2#1, UFU BIOFORT155E12, UFU BIOFORT189E22, UFU-199#3#1#1, UFU BIOFORT206E32, UFU BIOFORT197E34, and UFU BIOFORT189E43 were superior to the others. Regarding NGRDI, the genotypes UFU-7#1#2#1, UFU BIOFORT189E22, UFU-199#3#1#1, UFU BIOFORT206E32, UFU BIOFORT197E34, UFU-197#2#2#1, UFU BIOFORT189E43, and UFU MC BIOFORT1 showed average increments of 24%, 14.1%, 20%, 35.8%, 28.2%, 11.17%, 28.8%, and 26.47%, respectively, which were all higher than the readout for the commercial cv. Grand Rapids (Table 3).

Table 2. Means for the data collected in the field. Green mass (GM), plant diameter (PD), stem diameter (SD), and number of leaves (NL) in the green lettuce genotypes.

Genotype	MV (g)		DP (cm)		DH (cm)		NF
	$\bar{x}$	σ	$\bar{x}$	Ơ	$\bar{x}$	σ	$\bar{x}$
UFU-206#1#6#1	178.8	±84.0	25.3	±3.11	1.84	±0.22	23.3 b
UFU BIOFORT189E8	163.9	±31.8	24.5	±1.02	2.12	±0.18	26.5 a
UFU-197#3#1#1	197.1	±22.3	26.3	±0.86	1.96	±0.11	22.2 b
UFU-125#1#1#1	195.8	±8.17	28.3	±0.71	2.12	±0.08	27.7 a
UFU-7#1#2#1	154.3	±34.8	25.5	±1.42	2.12	±0.11	29.3 a
UFU BIOFORT155E12	154.3	±38.1	24.5	±2.14	2.00	±0.16	28.6 a
UFU BIOFORT120E21	157.5	±60.3	24.8	±4.60	1.86	±0.51	17.8 b
UFU BIOFORT189E22	181.8	±86.8	27.6	±3.33	2.00	±0.20	33.3 a
UFU-197#2#1#1	168.6	±32.2	25.1	±2.35	1.63	±0.34	29.5 a
UFU-199#3#1#1	132.2	±45.0	25.4	±1.68	1.94	±0.40	28.5 a
UFU-206#1#1#1	180.8	±76.4	24.6	±1.72	1.65	±0.27	27.7 a
UFU BIOFORT206E32	114.9	±51.5	24.5	±2.62	1.97	±0.38	29.4 a
UFU BIOFORT197E34	120.8	±53.8	25.3	±1.79	2.09	±0.25	32.6 a
UFU-197#2#2#1	167.3	±57.2	25.0	±3.80	1.99	±0.49	25.5 b
UFU BIOFORT155E39	141.2	±66.9	24.0	±3.68	1.50	±0.08	25.5 b
UFU BIOFORT189E43	140.5	±36.5	25.4	±1.96	1.87	±0.24	33.2 a
UFU-206#1#4#1	180.2	±60.0	27.0	±2.54	2.16	±0.22	23.0 b
UFU-125#2#2#1	191.1	±67.2	26.9	±3.43	2.09	±0.18	28.0 a
UFU-206#1#2#1	134.3	±45.5	25.3	±1.36	1.78	±0.08	22.0 b
UFU BIOFORT189E48	161.3	±57.2	28.7	±2.45	2.30	±0.39	27.2 a
UFU-206#1#5#1	160.9	±64.4	25.1	±3.70	1.78	±0.27	23.5 b
UFU-040#5#5#1	139.3	±41.1	26.4	±0.49	1.91	±0.12	27.0 a
UFU MC BIOFORT1	105.8	±32.8	22.7	±2.03	1.77	±0.32	20.0 b
Grand Rapids	189.7	±7.14	26.9	±1.16	1.95	±0.12	22.9 b
Uberlândia 10000	140.8	±59.8	25.1	±2.12	2.14	±0.04	29.6 a
Overall Average	160.9		25.3		1.967		27.25

$\bar{x}$: mean; σ: standard deviation. Averages followed by different letters in the column differ from one another by the Scott–Knott test ($p \leq 0.05$).

Table 3. Means for the data obtained from image analysis. Plant diameter in software (PDS), leaf area in software (LAS), Spectral Slope Saturation Index (SI), Normalized Green Red Difference Index (NGRDI), and Green Leaf Index (GLI) of the green lettuce strains.

Genotype	DPS (cm)		AFS (cm^2)		SI	GLI	NGRDI
	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	$\bar{x}$	$\bar{x}$
UFU-206#1#6#1	22.93	±2.70	476.2	±63.52	147.9 a	0.262 b	0.162 b
UFU BIOFORT189E8	20.31	±2.73	433.2	±49.73	135.2 b	0.286 a	0.183 b
UFU-197#3#1#1	20.78	±0.31	432.9	±71.69	132.8 b	0.279 a	0.183 b
UFU-125#1#1#1	18.46	±4.55	487.3	±17.16	151.2 a	0.254 b	0.172 b
UFU-7#1#2#1	18.58	±5.83	462.1	±60.48	122.7 b	0.325 a	0.212 a
UFU BIOFORT155E12	22.81	±1.41	477.6	±88.73	138.5 b	0.286 a	0.175 b
UFU BIOFORT120E21	20.23	±4.12	391.1	±150.2	155.2 a	0.206 b	0.118 b
UFU BIOFORT189E22	20.10	±4.08	453.1	±131.6	134.2 b	0.307 a	0.194 a
UFU-197#2#1#1	20.53	±5.53	445.0	±58.51	146.7 a	0.243 b	0.181 b
UFU-199#3#1#1	20.97	±5.06	448.5	±136.2	132.9 b	0.302 a	0.204 a
UFU-206#1#1#1	20.47	±4.56	424.1	±100.3	134.9 b	0.270 b	0.169 b
UFU BIOFORT206E32	20.24	±4.88	485.3	±93.04	123.5 b	0.309 a	0.231 a
UFU BIOFORT197E34	20.91	±7.13	492.4	±62.31	135.5 b	0.319 a	0.218 a
UFU-197#2#2#1	20.47	±1.88	443.3	±90.66	138.6 b	0.272 b	0.189 a
UFU BIOFORT155E39	20.63	±2.99	431.9	±137.4	148.4 a	0.235 b	0.156 b
UFU BIOFORT189E43	22.33	±2.44	482.1	±69.73	131.3 b	0.317 a	0.219 a
UFU-206#1#4#1	20.35	±4.10	435.6	±108.9	138.7 b	0.263 b	0.156 b

Table 3. *Cont.*

Genotype	DPS (cm)		AFS (cm^2)		SI	GLI	NGRDI
UFU-125#2#2#1	18.88	±6.25	459.9	±98.53	139.5 b	0.272 b	0.177 b
UFU-206#1#2#1	18.49	±4.75	427.8	±78.82	132.4 b	0.257 b	0.180 b
UFU BIOFORT189E48	21.67	±4.26	439.0	±130.5	144.5 a	0.263 b	0.161 b
UFU-206#1#5#1	22.37	±3.10	458.5	±154.6	148.5 a	0.255 b	0.158 b
UFU-040#5#5#1	23.98	±1.73	510.8	±140.2	168.2 a	0.196 b	0.128 b
UFU MC BIOFORT1	20.90	±1.40	442.2	±77.55	136.2 b	0.245 b	0.215 a
Grand Rapids	24.61	±1.08	531.8	±39.36	145.0 a	0.267 b	0.170 b
Uberlândia 10000	23.41	±3.10	456.1	±118.0	144.4 a	0.256 b	0.162 b
Overall Average	21.82		477.56			0.267	0.177

$\bar{x}$: mean; σ: standard deviation. Averages followed by different letters in the column differed from each other by the Scott-Knott test ($p \leq 0.05$).

The indices GLI and NGRDI had approximately 67% similarity in the choice of genotypes with the largest estimates of vegetative development. Both indices selected the genotypes UFU-7#1#2#1, UFU BIOFORT189E22, UFU-199#3#1#1, UFU BIOFORT206E32, UFU BIOFORT197E34, and UFU BIOFORT189E43.

3.2. Genetic Dissimilarity

The dendrogram (UPGMA) obtained by the generalized Mahalanobis distance confirmed the existence of genetic dissimilarity among the genotypes evaluated. The cophenetic correlation coefficient was 88.49%. A cut-off line was drawn at 31.23% dissimilarity, and the formation of four groups was identified (Figure 4).

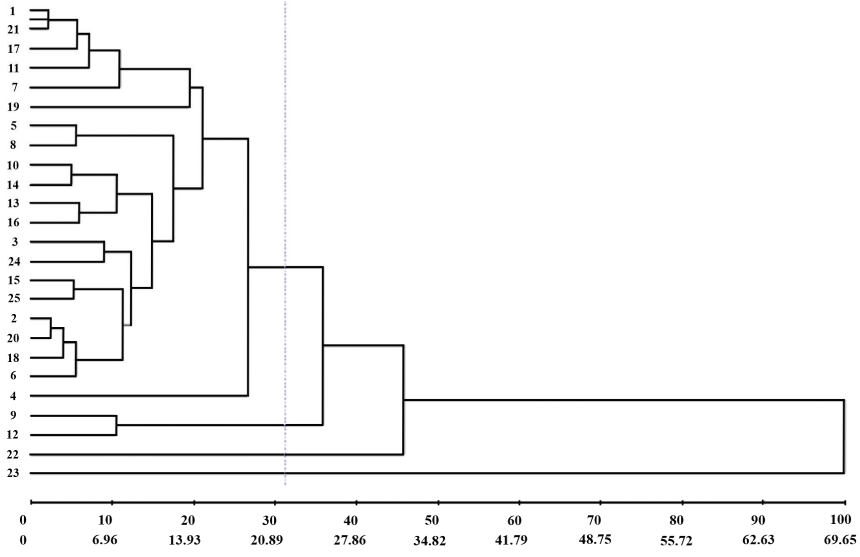


Figure 4. UPGMA dendrogram obtained by a generalized Mahalanobis distance of 25 green lettuce genotypes. 1: UFU-206#1#6#1; 2: UFU BIOFORT189E8; 3: UFU-197#3#1#1; 4: UFU-125#1#1#1; 5: UFU-7#1#2#1; 6: UFU BIOFORT155E12; 7: UFU BIOFORT120E21; 8: UFU BIOFORT189E22; 9: UFU-197#2#1#1; 10: UFU-199#3#1#1; 11: UFU-206#1#1#1; 12: UFU BIOFORT206E32; 13: UFU BIOFORT197E34; 14: UFU-197#2#2#1; 15: UFU BIOFORT155E39; 16: UFU BIOFORT189E43; 17: UFU-206#1#4#1; 18: UFU-125#2#2#1; 19: UFU-206#1#2#1; 20: UFU BIOFORT189E48; 21: UFU-206#1#5#1; 22: UFU-040#5#5#1; 23: UFU MC BIOFORT; 24: Grand Rapids; and 25: Uberlândia 10,000. Purple line: cut-off line at 31.23% dissimilarity.

Group I consisted of 84% of the genotypes analyzed; Group II comprised the genotypes UFU-197#2#1#1 and UFU BIOFORT206E32; and Groups III and IV comprised only one genotype each, UFU-040#5#5#1 and UFU MC BIOFORT1, respectively.

3.3. Monitoring Growth Rate

The acquisition of RGB images using UAVs enabled the monitoring of the growth rate in the lettuce crop over DAT (Figure 5).

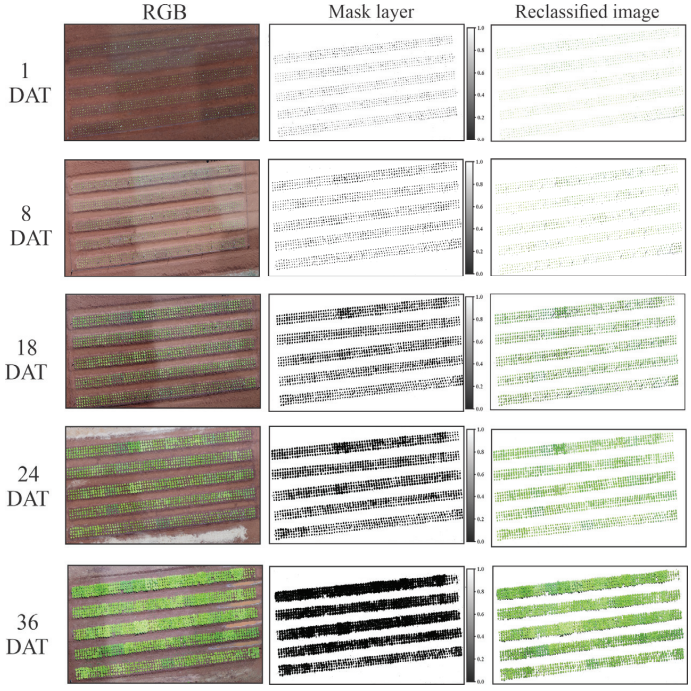


Figure 5. RGB image, mask layer, and reclassified image of the flights obtained by UAVs during the experiment on different days after transplanting (1, 8, 18, 24, and 36 DAT).

The images were reclassified, or the soil was removed in an automated manner using the vegetation indices and the RGB image. This step was performed to obtain better reproducibility and noise reduction of the image (Figure 6). After analysis of the pixel values in the histogram obtained by the vegetation indices, the HUE index enabled the discrimination of the plant and soil, using a cut-off value of 1.5 for the formation of the mask layer.

The evaluation of the PDS and LAS extracted using the images revealed growth over the DAT. The diameter obtained by the QGIS software showed a variation among the genotypes of 5.6 to 9.21 cm in the first flight (DAT 1); 7.86 to 13 cm in the second flight; and 10.6 to 18.56 cm, 11.1 to 21.5 cm, and 11 to 24.6 cm in the third, fourth, and fifth flights, respectively. For LAS, the lowest values were observed for the first flight (DAT 1) and the highest values were found for the fifth flight (DAT 5), which varied from 6.74 cm^2 to 531.8 cm^2 among the genotypes.

Obtaining the LAS and PDS values from the images was efficient. The regression equations presented values of the determination coefficient (R^2) between 78% and 99% for the genotypes evaluated. The vegetation indices SI, GLI, and NGRDI with the PDS and LAS values were coherent for lettuce growth rate, with an increase in the values observed over the flights for the genotypes (Figures 6–8).

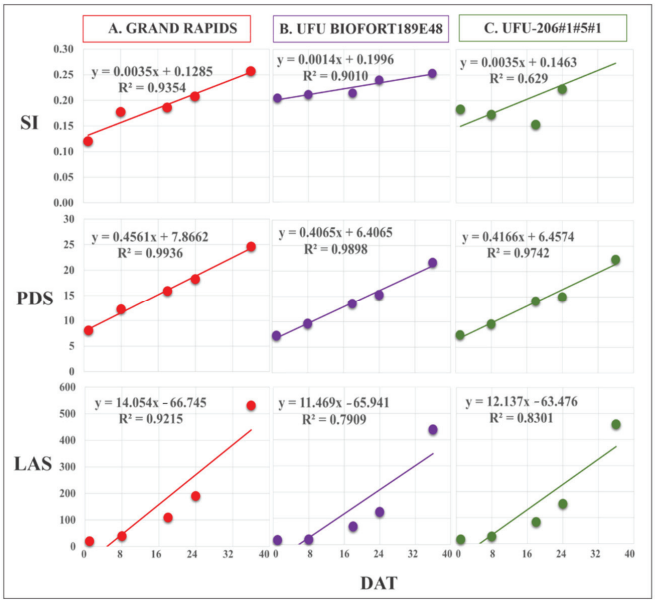


Figure 6. Regression equations of Spectral Slope Saturation Index (SI), leaf area in software (LAS) (cm^2), and plant diameter (PDS) obtained by imaging on different days after transplanting (DAT). The presented genotypes were selected based on their performance for growth rate monitoring using the vegetation index. (**A**) superior (Grand Rapids), (**B**) intermediate (UFU BIOFORT189E4), and (**C**) inferior (UFU-206#1#5#1).

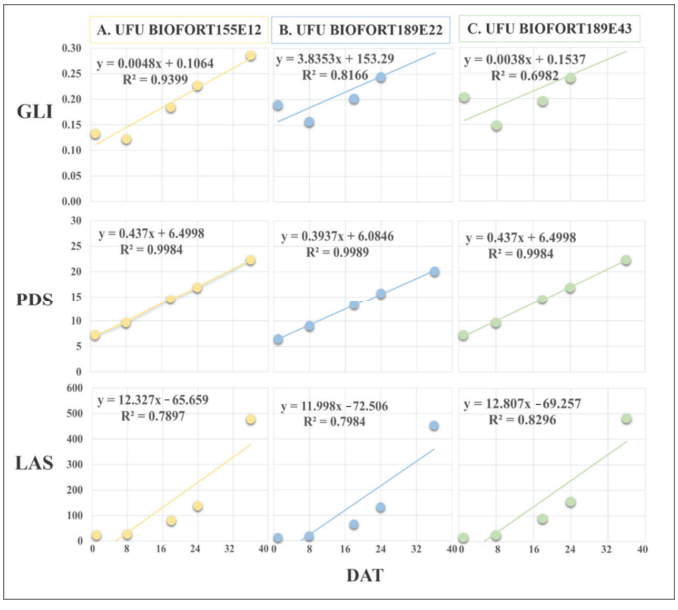


Figure 7. Regression equations of Green Leaf Index (GLI), leaf area in software (LAS) (cm^2), and plant diameter (PDS) obtained by imaging on different days after transplanting (DAT). The presented genotypes were selected based on their performance for growth rate monitoring using the vegetation index. (**A**) superior (UFU BIOFORT155E12), (**B**) intermediate (UFU BIOFORT189E22), and (**C**) inferior (UFU BIOFORT189E43).

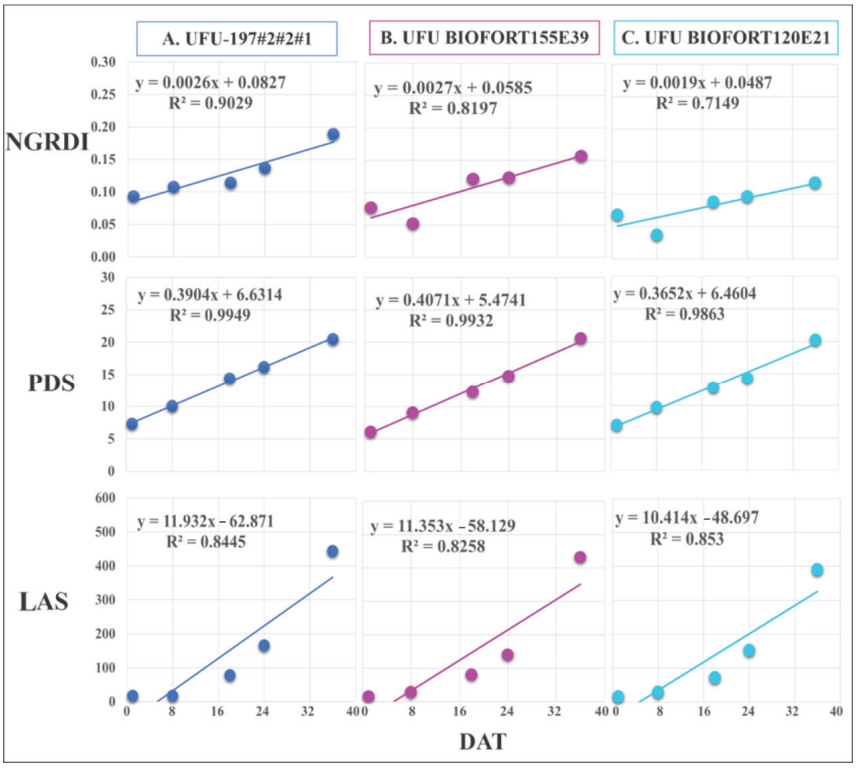


Figure 8. Regression equations of Normalized Green Red Difference Index (NGRDI), leaf area in software (LAS) (cm^2), and plant diameter (PDS) obtained by imaging on different days after transplanting (DAT). The presented genotypes were selected based on their performance for growth rate monitoring using the vegetation index. (**A**) superior (UFU-197#2#2#1), (**B**) intermediate (UFU BIOFORT155E39), and (**C**) inferior (UFU BIOFORT120E21).

Grand Rapids and the genotype UFUBIOFORT189E48 had good adjustments in the regression, with R^2 = 93.5% and 90%, respectively. The genotype UFU-206#1#5#1 had an R^2 = 62.9%, indicating that the linear regression did not explain the genotype behavior throughout its cycle (Figure 6).

GLI was efficient at determining the plant cover, along with the development and harvest point of the genotypes UFU BIOFORT155E12 (R^2 = 94%) and UFU BIOFORT189E22 (R^2 = 81%). The regression equation of genotype UFU BIOFORT189E43 referring to GLI had an R^2 = 69.8%.

NGRDI best estimated the development of genotypes UFU BIOFORT155E39 and UFU 197#2#2#1 (R^2 = 81.9% and 90.2%, respectively) in lettuce crop monitoring using RGB images. Compared with the genotypes above, UFU BIOFORT155E21 had a regression equation with an R^2 = 71.4%.

3.4. Validation of the Image Phenotyping Technique

The vegetation indices showed different behaviors for the genotypes under study and enabled the analysis of each genotype individually (Figure 9).

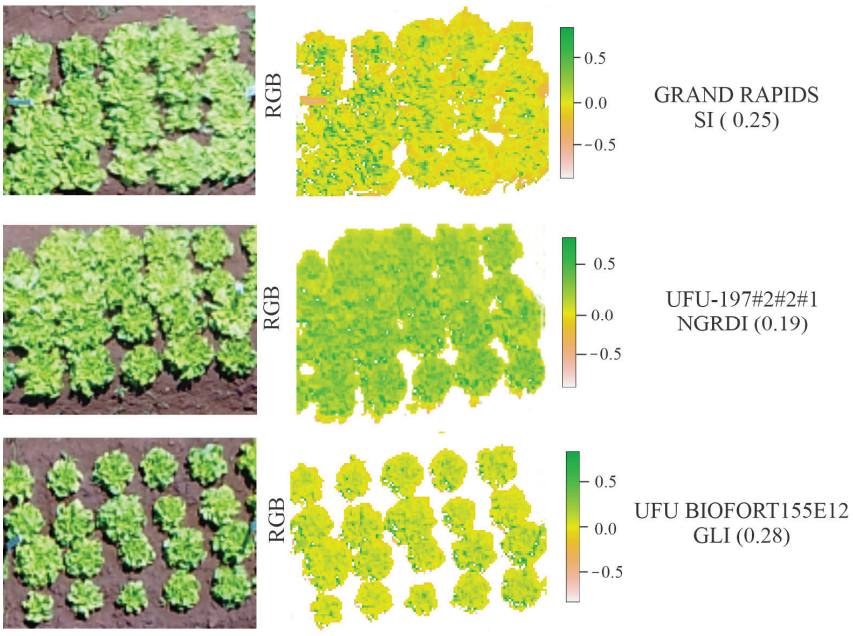


Figure 9. Representation of the behaviors of the vegetation indices: Spectral Slope Saturation Index (SI), Normalized Green Red Difference Index (NGRDI), and Green Leaf Index (GLI) for lettuce genotypes Grand Rapids, UFU-197 #2#2#1, and UFU BIOFORT 155E12.

Commercial cv. Grand Rapids and the genotypes UFU BIOFORT189E48, UFU BIOFORT189E43, UFU BIOFORT155E39, UFU BIOFORT155E12, UFU-206#1#5#1, and UFU BIOFORT189E22 had high correlations between GM values collected in the field and LAS from image analysis (0.78, 0.81, 0.82, 0.88, 0.88, 0.99, and 1.00, respectively). A high and positive correlation was found between PDS from image analysis and PD in the field. This behavior was observed for the genotypes UFU BIOFORT155E48, UFU-206#1#5#1, UFU BIOFORT120E21, UFU BIOFORT155E43, UFU BIOFORT155E39, and UFU BIOFORT189E22 (0.83, 0.85, 0.92, 0.98, 1, and 1, respectively) (Table 4).

A positive correlation was found of the vegetation indices SI, GLI, and NGRDI with the variables determined in the field: GM, PD, SD, and NL. The values presented a variation of 0.55–1.0 for the genotypes UFU BIOFORT155E12, UFU BIOFORT120E21, and UFU-206#1#5#1 in relation to SI. A positive correlation was verified of GLI and NGRDI with all the characteristics analyzed in the field for the genotypes UFU BIOFORT189E22 and UFU BIOFORT155E39. Furthermore, the highest correlation values were observed in the association of these indices with GM in UFU BIOFORT189E22 (0.99 and 1.00, respectively) and UFU BIOFORT155E39 (0.99 and 0.97, respectively). This second genotype also showed the highest correlation values with NL (0.97 for NGRDI and 0.99 for GLI) (Table 4).

PDS and LAS positively correlated with GM, PP, SD, and NL. This result was observed for the genotypes UFU BIOFORT155E12, UFU BIOFORT120E21, UFU BIOFORT189E22, UFU BIOFORT189E48, and UFU-206#1#5#1. In Grand Rapids, GLI and NGRDI positively correlated with PD and SD. In addition, PDS and LAS had high correlations with GM (0.97 and 0.78, respectively) and NL (0.94 and 1.0, respectively) (Table 4).

Table 4. Correlations per se between data collected in the field and data obtained by imaging at 36 days after transplanting (DAT) of nine green lettuce genotypes. Plant diameter in software (PDS), leaf area in software (LAS), Spectral Slope Saturation Index (SI), Normalized Green Red Difference Index (NGRDI), Green Leaf Index (GLI), green mass (GM), plant diameter (PD), stem diameter (SD), and number of leaves (NL).

		GM	PD	SD	NL
UFU BIOFORT155E12	SI	0.88 **	0.90 **	0.89 **	0.84 **
	GLI	0.64 *	0.60 *	0.62 *	0.69 *
	NGRDI	0.77 **	0.74 **	0.75 **	0.81 **
	PDS	0.48 *	0.52 *	0.50 *	0.42 *
	LAS	0.88 **	0.90 **	0.89 **	0.85 **
UFU BIOFORT120E21	SI	0.89 **	1.00 *	0.67 *	0.93 **
	GLI	1.00 *	0.88 **	0.96 **	0.98 **
	NGRDI	0.99 **	0.87 **	0.97 **	0.98 **
	PDS	0.68 *	0.92 **	0.37 *	0.75 **
	LAS	0.68 *	0.92 **	0.37 *	0.75 **
UFU BIOFORT189E22	SI	−0.35 ^{ns}	−0.43 ^{ns}	0.00 ^{ns}	−0.94 ^{ns}
	GLI	0.99 **	0.98 **	0.97 **	0.55 *
	NGRDI	1.00 *	0.99 **	0.95 **	0.61 *
	PDS	1.00 *	1.00 *	0.93 **	0.66 *
	LAS	1.00 *	1.00 *	0.90 **	0.71 **
UFU—197#2#1#1	SI	−0.94 ^{ns}	−0.93 ^{ns}	−0.74 ^{ns}	−0.92 ^{ns}
	GLI	−0.72 ^{ns}	−0.07 ^{ns}	−0.93 ^{ns}	−0.75 ^{ns}
	NGRDI	−0.70 ^{ns}	−0.04 ^{ns}	−0.92 ^{ns}	−0.74 ^{ns}
	PDS	0.71 **	0.04 *	0.92 **	0.74 **
	LAS	0.59 *	−0.11 ^{ns}	0.85 **	0.63 *
UFU BIOFORT155E39	SI	−0.56 ^{ns}	−0.64 ^{ns}	0.46 *	−0.56 ^{ns}
	GLI	0.99 **	0.97 **	0.60 *	0.99 **
	NGRDI	0.97 **	0.94 **	0.69 *	0.97 **
	PDS	0.99 **	1.00 *	0.39 *	0.99 **
	LAS	0.88 **	0.92 **	0.00 ^{ns}	0.88 **
UFU BIOFORT189E43	SI	−0.54 ^{ns}	−0.71 ^{ns}	0.82 **	0.53 *
	GLI	0.99 **	1.00 *	−0.97 ^{ns}	−0.99 ^{ns}
	NGRDI	0.98 **	0.91 **	−0.83 ^{ns}	−0.98 ^{ns}
	PDS	0.91 **	0.98 **	−1.00 ^{ns}	−0.90 ^{ns}
	LAS	0.82 **	0.92 **	−0.98 ^{ns}	−0.81 ^{ns}
UFU BIOFORT189E48	SI	0.98 **	1.00 *	1.00 *	0.89 **
	GLI	−0.22 ^{ns}	−0.09 ^{ns}	0.07 *	−0.45 ^{ns}
	NGRDI	−0.13 ^{ns}	−0.01 ^{ns}	0.16 *	−0.37 ^{ns}
	PDS	0.76 **	0.83 **	0.91 **	0.57 *
	LAS	0.81 **	0.87 **	0.94 **	0.64 *
UFU—206#1#5#1	SI	0.87 **	0.89 **	0.76 **	0.83 **
	GLI	0.59 *	0.55 *	0.74 **	0.65 *
	NGRDI	0.79 **	0.76 **	0.90 **	0.84 **
	PDS	0.88 **	0.85 **	0.95 **	0.91 **
	LAS	0.99 **	0.99 **	1.00 *	1.00 *
GRAND RAPIDS	SI	−0.33 ^{ns}	−0.37 ^{ns}	−0.79 ^{ns}	0.24 *
	GLI	−0.11 ^{ns}	0.73 **	0.98 **	−0.64 ^{ns}
	NGRDI	0.03 *	0.63 *	0.94 **	−0.53 ^{ns}
	PDS	0.97 **	−0.89 ^{ns}	−0.53 ^{ns}	0.94 **
	LAS	0.78 **	−1.00 ^{ns}	−0.84 ^{ns}	1.00 *

*: significant at 5% or less probability; **: significant at 1% or less probability, both by Student's *t* test; ^{ns}: not significant.

4. Discussion

In the present study, the climate was not very favorable, as the optimal temperature for the development of the lettuce culture is approximately 18 °C [29], but despite the irregular weather conditions, the plants did not suffer alterations.

The agronomic characteristics presented in the lettuce crop could be related to climate, genetic factors, and photoperiod [30,31]. These characteristics are defined and used as tools for product selection. Lettuce with greater GM is often selected by the consumer [31].

A greater NL on the lettuce plant positively impacts its commercialization. This characteristic can be used as a parameter to define climatic adaptations of the genotypes [32]. Other studies have revealed genotypes with better performance in NL compared to the commercial cv. Grand Rapids [33–35]. This result highlights the efficiency of genetic improvement in the lettuce crop, generating products of higher quality than those available on the market.

Remote sensing has become a tool with the potential to assist in monitoring and decision-making regarding crops. The evaluation of plant development by images is linked to records over time [36]. Therefore, the vegetation indices are an important tool in the evaluation of the vegetative development and identification of the harvest point in the lettuce crop.

Data analysis revealed the existence of genetic variability among the characterized genotypes. This information validates the use of phenotyping using images [17,18]. Furthermore, knowing the genetic variability among genotypes is essential for the selection of the best genotypes in breeding programs [37].

Variability was identified in vegetative development among the evaluated strains. The high correlation values highlight greater reliability in the clustering generated, and the closer these values are to one, the better the representativeness and quality of the cluster [38].

Similarity in the vegetative development of the strains was observed for the GLI and NGRDI. Similar behavior of the indices that use RGB (red, blue, and green) can be explained by their having similar detection characteristics for vegetation. GLI can be used to evaluate vegetation and has a good correlation with the chlorophyll content present in plants [39,40]. NGRDI has a strong relationship with chlorophyll content at different times of crop development, in addition to presenting strong potential to estimate the biomass of vegetation [41,42]. HUE has been used in different vegetation covers to differentiate between vegetation and non-vegetation pixels [43].

The vegetation indices analyzed by means of images were expressed differently among the lettuce genotypes. This difference occurred, for instance, when there were different levels of carotenoids [17,18]. When studying wheat crops, researchers found that the NDVI values among cultivars were influenced by the phenological stages of the crop and the amount of nitrogen present in the soil [44,45].

Plant phenotyping using images is more consistent than that using the conventional phenotyping method and can be useful in breeding programs [46,47]. In a study conducted with lettuce, researchers found a correlation of 0.68 between the anthocyanin contents quantified in the laboratory and the vegetation indices CIG, CIV, GNDVI, and NDVI [17].

Research with information extracted from images has revealed high correlations between the indices and different phenotypic characteristics of some crops. In brachiaria grass, a correlation of 0.92 was observed between control (%) and NDVI values extracted using images [48]. In corn, digital images were used to evaluate crop performance [46].

Studies have shown that the measurements of the leaf area of lettuce crops are performed via the traditional approach of counting leaves and electronic meters, which may or may not be destructible [49]. However, in large plantation areas, producers face difficulties in performing this monitoring, causing significant waste and losses [7]. In this context, the evaluation of leaf area and plant diameter obtained through images becomes a fast, effective, and low-cost tool for the use of plant phenotyping. The results presented in this work suggest that the methodology of collecting information through images adequately monitors the development of lettuce plants over time.

The monitoring of leaf area in different years in other crops, such as eucalyptus, using NDVI, SRI, and SAVI revealed equations with R^2 values ranging from 6.1% to 67.2% [20]. This study obtained R^2 values ranging from 78% to 92% for leaf area obtained through images. This result highlights the efficiency in phenotyping by imaging in the characterization of the development of lettuce plants. Regression models were generated for the respective vegetation indices and genotypes during winter. However, it is suggested to use the methodology during another season of the year and for other lettuce segments.

The information obtained in the present work indicates that phenotyping technology using RGB images to analyze and obtain information regarding vegetation indices, leaf area, and lettuce plant diameter has great potential. These results could facilitate the monitoring of the growth rate of lettuce plants and enable the determination of their harvest point. Image phenotyping is a low-cost technology and tool using RGB sensors, which can assist in decision-making and reduce the labor and costs associated with the existing crops. Image phenotyping is also a useful tool in genetic improvement, facilitating the characterization and selection of plants.

5. Conclusions

The vegetation indices SI, GLI, and NGRDI with the PDS and LAS values were coherent for lettuce growth rate, with an increase in the values observed over the flights for each genotype.

The correlations between data collected in the field and data obtained by imaging ranged from moderate to strong. Overall, the vegetation indices SI, GLI, and NGRDI were efficient for monitoring the growth rate and determining the harvest point of different green lettuce genotypes, in attempts to reduce waste and losses.

It is suggested that the definition of the harvest point based on vegetation indices be specific for each genotype.

Author Contributions: Conceptualization, G.M.M., A.C.S.S., R.B.d.A.G. and J.M.Q.L.; methodology, A.L.A.R. and P.H.d.S.A.; software, A.L.A.R., R.B.d.A.G. and P.H.d.S.A.; validation, A.L.A.R. and P.H.d.S.A.; formal analysis, G.M.M., A.C.S.S., J.M.Q.L. and H.C.R.M.C.; investigation, A.L.A.R. and G.M.M.; resources, G.M.M., A.C.S.S., J.M.Q.L. and H.C.R.M.C.; data curation, G.M.M. and R.B.d.A.G.; writing—original draft preparation, A.L.A.R.; writing—review and editing, G.M.M., A.C.S.S., J.M.Q.L., H.C.R.M.C. and R.Y.Y.; visualization, A.C.S.S. and R.Y.Y.; supervision, G.M.M.; project administration, G.M.M.; funding acquisition, G.M.M. and A.C.S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Brazilian National Council for Scientific and Technological Development (CNPq) Grant No 308824/2020-2, the Minas Gerais Research Foundation (FAPEMIG), the Coordination for the Improvement of Higher Education Personnel (CAPES), and the Federal University of Uberlândia (UFU).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Maggio, A.; Scapolo, F.; Crekinge, T.V.; Serraj, R. Global drivers and megatrends in agri-food systems. In *Agriculture & Food Systems to 2050—Global Trends, Challenges and Opportunities*; Serraj, R., Pingali, P., Eds.; Food and Agriculture Organization of the United Nations: Rome, Italy; Cornell University: Ithaca, NY, USA, 2018; Volume 2, pp. 47–83. [CrossRef]
2. Camara, G.R.; Busato, L.M.; Almeida, B.F.; Moraes, W.B. Elaboration and validation of diagrammatic scale for lettuce powdery mildew. *Summa Phytopathol.* **2018**, *44*, 116–121. [CrossRef]
3. ABCSEM. Associação Brasileira do Comércio de Sementes e Mudas. Available online: http://www.abcsem.com.br/dados-do-setor (accessed on 20 November 2021).
4. Carvalho-Filho, J.L.S.; Gomes, L.A.A.; Biguzzi, F.A.; Maluf, W.R.; Ferreira, S. F4 families of crispleaf lettuce with tolerance to early bolting and homozygous for resistance to Meloidogyne incognita race 1. *Hortic. Bras.* **2009**, *27*, 335–339. [CrossRef]
5. Sala, F.C.; Costa, C.P. Retrospective and trends of Brazilian lettuce crop. *Hortic. Bras.* **2012**, *30*, 187–194. [CrossRef]

6. Sediyama, M.A.N.; Pedrosa, M.W.; Salgado, L.T.; Pereira, P.C. Summer and winter performance of lettuce cultivars grown in a hydroponic system. *Científica* **2019**, *37*, 98–106. [CrossRef]
7. Aliotte, J.T.B.; Filassi, M.; Oliveira, A.L.R. Characterization of fruit and vegetable distribution logistics of Campinas Supply Center/SP. *Rev. Econ. Social. Rural* **2022**, *60*, e252673. [CrossRef]
8. Dhondt, S.; Wuyts, N.; Inzé, D. Cell to whole-plant phenotyping: The best is yet to come. *Trends Plant Sci.* **2013**, *18*, 428–439. [CrossRef]
9. Sousa, C.A.F.; Cunha, B.A.D.B.; Martins, P.K.M.; Molinari, H.B.C.; Kobayashi, A.K.; Souza, M.T., Jr. New approach for plant phenotyping: Concepts, current tools and perspectives. *Rev. Bras. Geogr. Fís.* **2015**, *8*, 660–672. [CrossRef]
10. Ponzoni, F.J.; Shimabukuro, Y.E.; Kuplich, T.M. *Sensoriamento Remoto da Vegetação*, 2nd ed.; Oficina de Textos: São Paulo, Brazil, 2012; 176p.
11. Zhang, J.; Naik, H.S.; Assefa, T.; Sarkar, S.; Reddy, R.V.C.; Singh, A.; Ganapathysubramanian, B.; Singh, A.K. Computer vision and machine learning for robust phenotyping in genome-wide studies. *Sci. Rep.* **2017**, *7*, 44048. [CrossRef]
12. Fernandez-Gallego, J.A.; Kefauver, S.C.; Gutiérrez, N.A.; Nietotaladriz, M.T.; Araus, J.L. Wheat ear counting in-field conditions: High throughput and low-cost approach using RGB images. *Plant Methods* **2018**, *14*, 22. [CrossRef]
13. Makanza, R.; Zaman-Allah, M.; Cairns, J.E.; Magorokosho, C.; Tarekegne, A.; Olsen, M.; Prasanna, B.M. High-throughput phenotyping of canopy cover and senescence in maize field trials using aerial digital canopy imaging. *Remote Sens.* **2018**, *10*, 330. [CrossRef]
14. Beloti, I.F.; Maciel, G.M.; Gallis, R.B.A.; Finzi, R.R.; Clemente, A.A.; Siquieroli, A.C.S.; Juliatti, F.C. Low-altitude, high-resolution aerial imaging for field crop phenotyping in Cucurbita pepo. *Genet. Mol. Res.* **2020**, *19*, 18598. [CrossRef]
15. Silva, M.F.; Maciel, G.M.; Gallis, R.; Barbosa, R.L.; Carneiro, V.Q.; Rezende, W.S.; Siquieroli, A.C.S. High-throughput phenotyping by RGB and multispectral imaging analysis of genotypes in sweet corn. *Hortic. Bras.* **2022**, *40*, 92–98. [CrossRef]
16. Elangovan, A.; Duc, N.T.; Raju, D.; Kumar, S.; Singh, B.; Vishwakarma, C.; Gopala Krishnan, S.; Ellur, R.K.; Dalal, M.; Swain, P.; et al. Imaging Sensor-Based High-Throughput Measurement of Biomass Using Machine Learning Models in Rice. *Agriculture* **2023**, *13*, 852. [CrossRef]
17. Clemente, A.A.; Maciel, G.M.; Siquieroli, A.C.S.; Gallis, R.B.A.; Medeiros, L.M.; Duarte, J.G. High-throughput phenotyping to detect anthocyanins, chlorophylls, and carotenoids in red lettuce germplasm. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102533. [CrossRef]
18. Maciel, G.M.; Gallis, R.B.A.; Barbosa, R.L.; Pereira, L.M.; Siquieroli, A.C.S.; Peixoto, J.V.M. Image phenotyping of inbred red lettuce lines with genetic diversity regarding carotenoid levels. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *81*, 154–160. [CrossRef]
19. Maciel, G.M.; Gallis, R.B.A.; Barbosa, R.L.; Pereira, L.M.; Siquieroli, A.C.S.; Peixoto, J.V.M. Image phenotyping of lettuce germplasm with genetically diverse carotenoid levels. *Bragantia* **2020**, *79*, 224–235. [CrossRef]
20. Berger, R.; Silva, J.A.A.; Ferreira, R.L.C.; Candeias, A.L.B.; Rubilar, R. Vegetation indices for the leaf area index estimation in clonal plantations of *Eucalyptus saligna* Smith. *Ciênc. Florest.* **2019**, *29*, 885–899. [CrossRef]
21. Maciel, G.M.; Siquieroli, A.C.S.; Gallis, R.B.A.; Pereira, L.M.; Sales, V.F. Programa de computador BG α Biofort. Depositor: Federal University of Uberlândia. BR512019002403-6. Deposit: 1 February 2019. Concession: 23 October 2019. Available online: https://busca.inpi.gov.br/pePI/servlet/ProgramaServletController (accessed on 10 March 2023).
22. Filgueira, F.A.R. *Novo Manual de Olericultura: Agrotecnologia Moderna na Produção e Comercialização de Hortaliças*, 3rd ed.; Editora UFV: Viçosa, Brazil, 2013; 421p.
23. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020; Available online: https://www.r-project.org/ (accessed on 20 January 2023).
24. Matias, F.I.; Caraza-Harter, M.V.; Endelman, J.B. FIELDimageR: An R package to analyze orthomosaic images from agricultural field trials. *Plant Phenome J.* **2020**, *3*, e20005. [CrossRef]
25. Escadafal, R.; Belghith, A.; Bem, M.H. Indices spectraux pour la télédétection de la dégradation des milieux naturels en Tunisie aride. In Proceedings of the Actes du Sixième Symposium International. Mesures Physiques et Signatures Spectrales en Télédétection, Val d'Isèr, France, 17–24 January 1994.
26. Louhaichi, M.; Borman, M.M.; Johnson, D.E. Spatially located platform and aerial photography for documentation of grazing impacts on wheat. *Geocarto Int.* **2001**, *16*, 65–70. [CrossRef]
27. Tucker, C. Red and Photographic Infrared Linear Combinations for Monitoring Vegetation. *Remote Sens. Environ.* **1979**, *8*, 127–150. [CrossRef]
28. Cruz, C.D. Genes: A software package for analysis in experimental statistics and quantitative genetics. *Acta Sci. Agron.* **2013**, *35*, 271–276. [CrossRef]
29. Fontes, P.C.R.; Nick, C. *Olericultura Teoria e Prática*, 2nd ed.; Editora UFV: Viçosa, Brazil, 2019; 632p.
30. Queiroz, J.P.S.; Costa, A.J.M.; Neves, L.G.N.; Seabra Junior, S.; Barelli, M.A.A. Phenotypic stability of the lettuce in different periods and cropping environments. *Rev. Ciênc. Agron.* **2014**, *45*, 276–283. [CrossRef]
31. Oliveira, A.C.B.; Sediyama, M.A.N.; Pedrosa, M.W.; Garcia, N.C.P.; Garcia, S.L.R. Genetic divergence and discard of variables in lettuce cultivated under hydroponic system. *Acta Sci. Agron.* **2003**, *26*, 211–217.
32. Diamante, M.S.; Seabra, S., Jr.; Inagaki, A.M.; Silva, M.B.; Dallacort, R. Production and resistance to bolting of loose-leaf lettuce grown in different environments. *Rev. Ciênc. Agron.* **2013**, *44*, 133–140. [CrossRef]

33. Medeiros, D.C.; Freitas, K.C.S.; Veras, F.S.; Anjos, R.S.B.; Borges, R.D.; Cavalcante, N.J.G.; Nunes, G.H.S.; Ferreira, H.A. Quality of lettuce seedlings depending on substrates with and without biofertilizer addition. *Hortic. Bras.* **2008**, *26*, 186–189. [CrossRef]
34. Ferreira, L.L.; Aniceto, R.R.; Montenegro, I.N.A.; Ribeiro, T.S.; Almeida, D.G.; Porto, V.C.N. Adaptability and development of cultivars of lettuce in the Brejo microregion, Paraiba. *Sci. Plena* **2013**, *9*, 040202-1.
35. Mendes, F.T.C.; Freitas, A.S.; Alcantra, E.; Marques, R.F.P.V.; Oliveira, A.S.; Barbosa, R.A.; Padua, M.C.; Junqueira, R.R. Agronomic performance of lettuce cultivars in aquaponics. *Res. Soc. Dev.* **2021**, *10*, 2525–3409. [CrossRef]
36. Pandit, S.; Tsuyuki, S.; Dube, T. Estimating above-ground biomass in sub-tropical buffer zone community Forests, Nepal, using Sentinel 2 data. *Remote Sens.* **2018**, *10*, 601. [CrossRef]
37. Araujo, J.C.; Telhado, S.F.P.; Sakai, R.H.; Ledo, C.A.S.; Melo, P.C.T. Univariate and multivariate procedures for agronomic evaluation of organically grown tomato cultivars. *Hortic. Bras.* **2016**, *34*, 374–380. [CrossRef]
38. Cruz, C.D.; Regazzi, A.J.; Carneiro, P.C.S. *Modelos Biométricos Aplicados ao Melhoramento Genético, 3rd ed*; Editora UFV: Viçosa, Brazil, 2014; 668p.
39. Hunt, E.R.; Hively, W.D.; McCarty, G.W.; Daughtry, C.S.T.; Forrestal, P.J.; Kratochvil, R.J.; Carr, J.L.; Allen, N.F.; Fox-Rabinovitz, J.R.; Miller, C.D. NIR-Green-Blue high-resolution digital images for assessment of winter cover crop biomass. *GIsci. Remote Sens.* **2011**, *48*, 86–98. [CrossRef]
40. Ballesteros, R.; Ortega, J.F.; Hernandez, D.; Campo, A.D.; Moreno, M.A. Combined use of agro-climatic and very high-resolution remote sensing information for crop monitoring. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *72*, 66–75. [CrossRef]
41. Hunt, E.R.; Cavigelli, M.; Daughtry, C.S.T.; McMurtrey, J.E., III; Walthall, C.L. Evaluation of digital photography from model aircraft for remote sensing of crop biomass and nitrogen status. *Precis. Agric.* **2005**, *6*, 359–378. [CrossRef]
42. Poley, L.G.; Mcdermid, G.J. A systematic review of the factors influencing the estimation of vegetation aboveground biomass using unmanned aerial systems. *Remote Sens.* **2020**, *12*, 1052. [CrossRef]
43. Beniaich, A.; Silva, M.L.N.; Avalos, F.A.P.; Menezes, M.D.; Cândido, B.M. Determination of vegetation cover index under different soil management systems of cover plants by using an unmanned aerial vehicle with an onboard digital photographic camera. *Semin. Ciênc. Agrár.* **2019**, *40*, 49–66. [CrossRef]
44. Reznick, J.P.K.; Pauletti, V.; Barth, G. Field estimate with NDVI of grain yield and quality of wheat flour. *Rev. Bras. Eng. Agríc. Ambient.* **2021**, *25*, 801–806. [CrossRef]
45. Rissini, A.L.L.; Kawakami, J.; Genu, A.M. Normalized difference vegetation index and yield of wheat cultivars under different application rates of nitrogen. *Rev. Bras. Ciênc. Solo* **2015**, *39*, 1703–1713. [CrossRef]
46. Makanza, R.; Zaman-allah, M.; Cairns, J.E.; Eyre, J.; Burgueno, J.; Pacheco, A.; Diepenbrock, C.; Magorokossho, C.; Terekegne, A.; Olsen, M.; et al. High-throughput method for ear phenotyping and kernel weight estimation in maize using ear digital imaging. *Plant Methods* **2018**, *14*, 49. [CrossRef]
47. Walter, A.; Liebisch, F.; Hund, A. Plant phenotyping: From bean weighing to image analysis. *Plant Methods* **2015**, *11*, 14. [CrossRef]
48. Alvarenga, C.B.; Mundim, G.S.M.; Santos, E.A.; Gallis, R.B.A.; Zampiroli, R.; Rinaldi, P.C.N.; Prado, J.R. Normalized difference vegetation index for desiccation evaluation with glyphosate + 2,4-D in magnetized spray solution. *Braz. J. Biol.* **2023**, *83*, e246579. [CrossRef]
49. Zuffo, A.M.; Juffo Júnior, J.M.; Silva, L.M.A.; Silva, R.L.; Menezes, K.O. Growth analysis in lettuce cultivars in southern Piauí. *Rev. Ceres* **2016**, *63*, 145–153. [CrossRef]

agriculture

MDPI

Article

Deep CNN-Based Planthopper Classification Using a High-Density Image Dataset

Mohd Firdaus Ibrahim [1,2], Siti Khairunniza-Bejo [1,3,4,*], Marsyita Hanafi [5], Mahirah Jahari [1,3], Fathinul Syahir Ahmad Saad [6] and Mohammad Aufa Mhd Bookeri [7]

[1] Department of Biological and Agricultural Engineering, Faculty of Engineering, Universiti Putra Malaysia, Serdang 43400, Malaysia; gs58903@student.upm.edu.my (M.F.I.); jmahirah@upm.edu.my (M.J.)
[2] Faculty of Mechanical Engineering and Technology, Universiti Malaysia Perlis, Arau 02600, Malaysia
[3] Smart Farming Technology Research Centre, Universiti Putra Malaysia, Serdang 43400, Malaysia
[4] Institute of Plantation Studies, Universiti Putra Malaysia, Serdang 43400, Malaysia
[5] Department of Computer and Communication Systems Engineering, Faculty of Engineering, Universiti Putra Malaysia, Serdang 43400, Malaysia; marsyita@upm.edu.my
[6] Faculty of Electrical Engineering and Technology, Universiti Malaysia Perlis, Arau 02600, Malaysia; fathinul@unimap.edu.my
[7] Engineering Research Centre, Malaysian Agriculture Research and Development Institute, Seberang Perai 13200, Malaysia; aufa@mardi.gov.my
* Correspondence: skbejo@upm.edu.my; Tel.: +60-397694332

Abstract: Rice serves as the primary food source for nearly half of the global population, with Asia accounting for approximately 90% of rice production worldwide. However, rice farming faces significant losses due to pest attacks. To prevent pest infestations, it is crucial to apply appropriate pesticides specific to the type of pest in the field. Traditionally, pest identification and counting have been performed manually using sticky light traps, but this process is time-consuming. In this study, a machine vision system was developed using a dataset of 7328 high-density images (1229 pixels per centimetre) of planthoppers collected in the field using sticky light traps. The dataset included four planthopper classes: brown planthopper (BPH), green leafhopper (GLH), white-backed planthopper (WBPH), and zigzag leafhopper (ZIGZAG). Five deep CNN models—ResNet-50, ResNet-101, ResNet-152, VGG-16, and VGG-19—were applied and tuned to classify the planthopper species. The experimental results indicated that the ResNet-50 model performed the best overall, achieving average values of 97.28% for accuracy, 92.05% for precision, 94.47% for recall, and 93.07% for the F1-score. In conclusion, this study successfully classified planthopper classes with excellent performance by utilising deep CNN architectures on a high-density image dataset. This capability has the potential to serve as a tool for classifying and counting planthopper samples collected using light traps.

Keywords: planthoppers; convolutional neural network; machine vision; paddy cultivation

Citation: Ibrahim, M.F.; Khairunniza-Bejo, S.; Hanafi, M.; Jahari, M.; Ahmad Saad, F.S.; Mhd Bookeri, M.A. Deep CNN-Based Planthopper Classification Using a High-Density Image Dataset. *Agriculture* **2023**, *13*, 1155. https://doi.org/10.3390/agriculture13061155

Academic Editors: Gniewko Niedbała and Sebastian Kujawa

Received: 31 March 2023
Revised: 26 May 2023
Accepted: 26 May 2023
Published: 30 May 2023

1. Introduction

Since 1910, rice has been a staple in the daily diet of Malaysians, either consumed directly as cooked rice or in the form of flour derived from the milling process [1]. Malaysians heavily rely on rice as their primary source of nutrition, consuming an average of 80 kg per person, which contributes to 26% of their daily caloric intake [2]. This indicates national consumption of approximately 2.7 million tons of rice per year. Despite local production, the quantity falls short of meeting demand, leading the Malaysian government to import approximately 30% of its rice from countries such as Thailand, India, Vietnam, and Pakistan [3]. Revision of policies and implementation of measures are necessary to ensure food security concerning rice.

Threats posed by planthopper insects have been a significant issue affecting rice production in Malaysia [4]. The Malaysian Agricultural Research and Development Institute (MARDI) has identified four major planthopper species in the country: the brown planthopper (*Nilaparvata lugens* (*Stal*)), the green leafhopper (*Nephotettix malayanus*), the white-backed planthopper (*Sogatella furcifera* (*Harvath*)), and the zigzag leafhopper (*Recilia dorsalis* (*Motschulsky*)). To prevent pest outbreaks, pesticides are commonly applied in the fields. However, excessive pesticide use can have negative impacts on both plants and the environment. These effects include the development of pesticide resistance by the pests and the presence of high pesticide residue concentrations in rivers [5]. Therefore, effective pesticide management strategies need to be implemented, with early detection of pest outbreaks playing a crucial role in the process.

In pest control, monitoring the occurrence pattern of pests plays a crucial role [6]. MARDI has developed a manual identification and counting process conducted by trained experts. To facilitate this process, they have designed a solar-powered light trap system specifically for capturing pests during nighttime [7]. The system consists of an LED light shielded by a transparent plastic sheet, approximately the size of A3 paper. The plastic sheet is coated with sticky glue to trap the pests. Since flying insects are attracted to light due to positive phototaxis, they are drawn toward the light source. The sticky sheet is enclosed within a transparent box that contains multiple small holes with a diameter of 5 mm. These holes prevent larger flying insects from becoming trapped in the sticky sheet. The collection of the sticky light trap is carried out the following day, and the trapped insects are manually counted by the experts. However, this counting process can be time-consuming, taking up to 6 h for a single light trap. Furthermore, the accuracy and efficiency of the manual counting process may be affected by factors such as fatigue and emotional aspects of the inspector. Consequently, applying this manual process on a large scale is challenging due to limitations in the counting procedure.

Several research studies have focused on utilising machine vision for pest identification and classification. One technique proposed for image analysis incorporates scene interpretation [8]. An automatic detection system for harmful insects inside the greenhouse has been developed using three feature extraction methods: the pyramidal histogram of gradient, Gabor filter, and colour data. The system utilised a support vector machine (SVM) for accurate prediction of whiteflies (1283 samples) with a 98.5% accuracy rate, as well as greenflies (49 samples) with a 91.8% accuracy rate. To capture the images of pests, a pan-tilt camera was employed as the acquisition device [9]. The study employed a centralised server to process and analyse the recorded field video. Images were extracted from the video on a frame-by-frame basis, and the SVM classification was performed individually for each frame.

Convolutional neural networks (CNNs) have recently been applied in pest classification. CNNs are a class of artificial neural networks (ANNs) that employ deep learning architecture and are commonly used for visual image classification [10–18] and detection [19–24]. The architecture of a CNN consists of an input layer, hidden layers, and an output layer as the final layer. The hidden layers include various types of layers, such as convolutional layers, pooling layers, and dense (also known as fully connected) layers. During the convolutional phase, the image is transformed into a feature map, also known as an activation map, with a specific shape. Convolutional layers combine the input with the output before passing it to the next layer. Global pooling layers may be incorporated to reduce the dimensionality of the data throughout the convolutional phase. The output of a cluster of neurons is aggregated at one layer and then transmitted to a single neuron in the subsequent layer. The dense layers establish connections between every neuron in one layer and every neuron in the next layer.

Although the application of CNN for planthopper classification has not been widely explored, its performance in classifying other pests has been proven. For instance, CNN was used to detect wheat sawfly, wheat mite, and wheat aphid in the paddy fields of Anhui province, China, achieving an accuracy of up to 90.88% for wheat sawfly detection and

a minimum accuracy of 70.2% for wheat mite detection [25]. In another study, VGG-19 was employed to train on 4800 images of 24 types of insects, achieving a mean average precision (mAP) of 0.8922 with a training time of 70 h [26]. Among other examples, moth [19,25], oilseed rape pest [26], bark beetles [20], forest pest [27], citrus pest [28], and rice pest [7,29–31] are a few other insects and pests that have been studied for classification using CNN. Ref. [32] used a CNN combined with a Euclidean distance map (EDM) to automatically recognise the brown planthopper captured on a sticky pad. Their method demonstrated successful results with 95% accuracy in identifying the BPH. However, the dataset used in the research was relatively small, comprising only 1374 samples. As imperfect planthoppers exhibit more variations, the addition of new samples may lead to misclassification. Moreover, the research only focused on differentiating between BPH and benign insects.

Therefore, to address the research gap, this study proposes a novel method for classifying four types of planthoppers using five different deep convolutional neural networks and a large dataset. Planthopper images were cropped from the full image captured by the light trap, resulting in a total of 7328 planthopper images. These images were then divided into training, validation, and test sets with a ratio of 80:10:10. Augmentation techniques were applied to the training dataset to create a larger dataset, resulting in a total of 187,456 samples for the training dataset. Subsequently, the dataset was trained using five different CNN architectures, namely ResNet-50, ResNet-101, ResNet-152, VGG-16, and VGG-19. The results of this experiment demonstrated the feasibility and validity of the model architecture for accurately classifying the four types of planthoppers. Quick and accurate classification can significantly reduce the time required to identify pests captured by the light trap. Implementing this approach can help reduce the reliance on manual labour while minimising the risk of human error during the classification process.

2. Materials and Methods

This study included three major stages: image acquisition, pre-processing, and classification. The flowchart for this study is shown in Figure 1.

2.1. Data Collection

The study area was located in Felcra Seberang Perak, Malaysia (4.072710082450001, 100.86747760853657). The on-field data collection was conducted by an officer from MARDI Seberang Perai, Pulau Pinang, during the paddy planting season in 2020. To collect the data, a light trap device was utilised, which consisted of a light bulb, a stand pole, and a sticky trap. The sticky trap was created using a clear plastic sheet with the dimensions of a sheet of A3 paper, onto which sticky glue was sprayed on one side. The sticky light trap was then wrapped around the light bulb to capture any pests attracted to it. The light bulb was turned on from 7:30 p.m. to 8:30 p.m., when the insects were most active. Insects were drawn to the light source, flew toward it, and became trapped on the sticky trap in various positions. Some of them sustained damage in their attempt to escape from the light trap. The following day, the sticky trap was collected and taken to the lab for the image acquisition process.

2.2. Image Acquisition

Figure 2 illustrates a machine vision system used to capture light trap images. The system consisted of an industrial camera, a fixed focal length lens, a diffused LED white light (DLW2-60-070-1-W-24V, TMS Lite, Pulau Pinang, Malaysia), a flat platform for the light trap, and a 3-axis jig. To eliminate external light interference, all components were housed inside a black box. The LED light was powered by a 24 VDC light controller (SD-1000-D1, TMS Lite, Malaysia), with the controller output set to its maximum value of 2 A. A 6 Megapixel (MP) camera MV-CA060-10GC (HIK Vision, Hangzhou, China) with a sensor resolution of 2.4 μm per pixel was used to capture the images. The camera was paired with a 35 mm focal length lens (MVL-HF3528M-6MP, HIK Vision, Hangzhou, China) and

mounted on top of the platform. The distance between the platform and the lens was set to 127 mm, as depicted in Figure 3. The combined camera and lens configuration provided a field of view (FOV) of 24 mm in width and 15 mm in length. The camera was set to capture images in the red, green, and blue (RGB) colour format, with a size of 3072 × 2048 pixels. An example of a captured image is presented in Figure 4. The pixel density (ppcm) of the captured image was determined by dividing the number of pixels in the FOV by the actual measurement in centimetres. In this case, the pixel density of the captured image was calculated as 1229 pixels per centimetre (ppcm), indicating that each pixel on the image represented 0.0081 mm of the actual measurement.

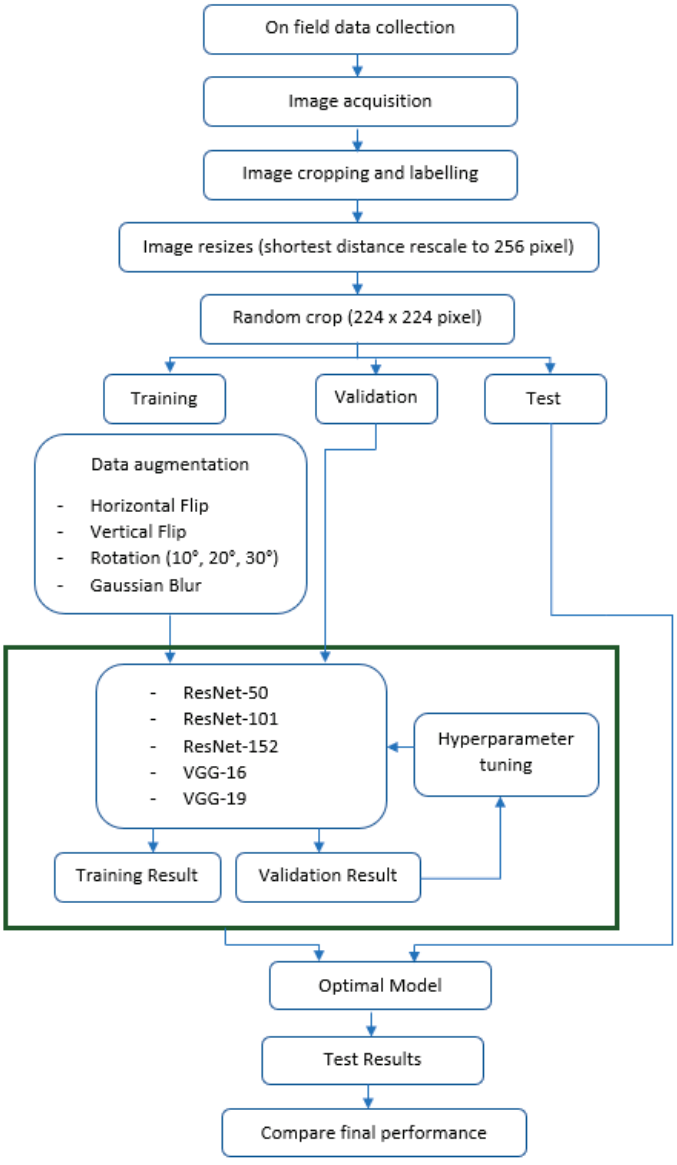


Figure 1. Flow chart of the study.

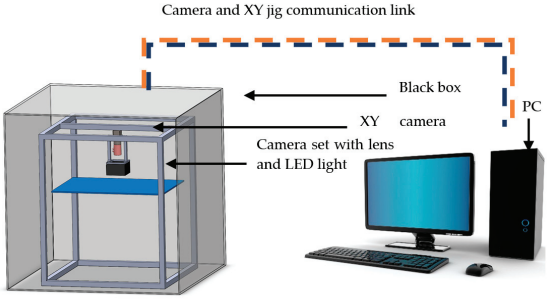


Figure 2. Machine vision system used for image acquisition.

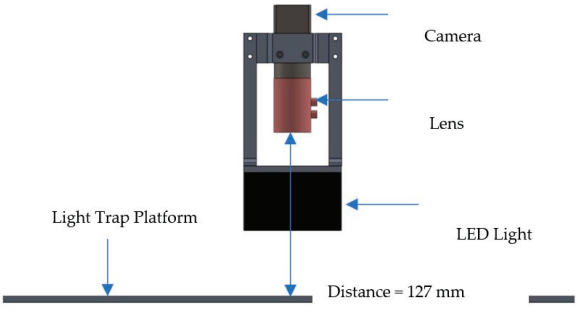


Figure 3. Camera and lens setup for the system.

Figure 4. Sample of an image captured using a machine vision system with actual dimensions.

The light trap was integrated into the machine vision system to facilitate the image acquisition process. Within the machine vision system, an xy-jig was employed to move the camera. The dimensions of the light trap were 420 mm in width and 294 mm in length. However, the field of view of the camera (FOV) could only cover an area of 24 mm × 15 mm. As depicted in Figure 5, the region occupied by the insects was measured to be 336 mm × 245 mm. Therefore, the camera only needed to be moved across 19 × 17 grids to capture the entire populated area. The xy-jig utilised 2 stepper motors to enable camera movement in the x and y directions. As a result, the stepper motor shifted the camera from one grid to another, covering a total of 323 grids. The operation of the enclosed black box, including stepper motor movement and image acquisition, was controlled using LabVIEW software (National Instruments, Austin, TX, USA) running on a Windows-based computer system equipped with a Ryzen 5-2600X CPU@3.6GHz).

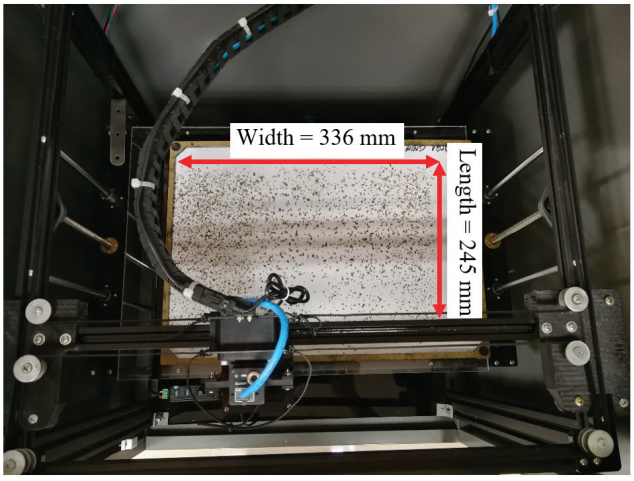


Figure 5. Sample of light trap inside the black box.

2.3. Dataset

Figure 6 illustrates the four types of planthoppers utilised in this study: BPH, GLH, WBPH, and ZIGZAG. The captured images from the sticky trap were manually cropped to extract individual planthopper images. In total, 7328 planthopper images were cropped from the light trap image and labelled according to their respective types. The labelling process was carried out manually with the assistance of experts from MARDI, who relied on the visual features and morphology of the planthoppers. The dataset was then divided into training, validation, and test sets in an 80:10:10 ratio. This resulted in 5858 samples for the training set, 730 samples for the validation set, and 736 samples for the test set. To enhance the variety of the training sample, augmentation techniques were applied. Firstly, the images were horizontally flipped. Then, they were rotated at three different angles: 10°, 20°, and 30°. Finally, a Gaussian blur was applied as the last step of the augmentation process. After augmentation, the training set comprised a total of 187,456 samples. Figure 7 provides examples of damaged and multi-orientated planthopper images from the dataset.

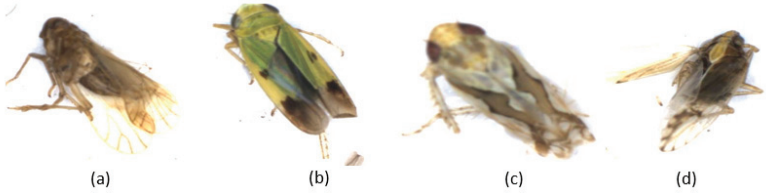


Figure 6. Four types of planthoppers gathered for the data collection. (**a**) BPH, (**b**) GLH, (**c**) ZIGZAG, and (**d**) WBPH.

2.4. Model Architecture

In this study, two types of model architecture were used, namely Residual Network (ResNet) and Visual Geometry Group (VGG) network.

2.4.1. ResNet Model Architecture

Instead of learning unreferenced functions, Residual Networks (ResNets) train residual functions with reference to the layer inputs. They have an extremely deep architecture and are high-performance networks that enable the process of propagation of information to take place more directly through the network [33]. Residual nets allow each layer to match a residual mapping rather than requiring each few stacked layers to exactly match a desired

underlying mapping. This method, sometimes known as a "skip connection", connects the activation of one layer to subsequent levels by bypassing some layers in between. It creates a network by stacking residual blocks on top of one another; e.g., ResNet-50 uses 50 layers of these blocks. This skip connection was designed to tackle the problem of CNN accuracy degradation. The inutile layer will be skipped during training. Table 1 shows the network structure of ResNet-50, ResNet-101, and ResNet-152; these three models were used in this study.

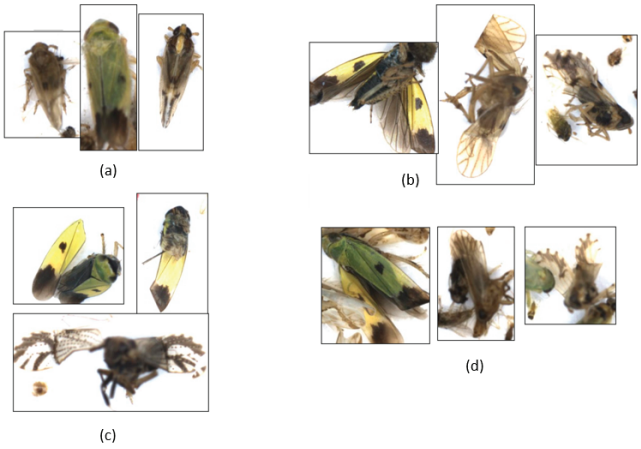


Figure 7. Samples of planthopper images in different conditions (**a**) good, (**b**) misoriented, (**c**) damaged, and (**d**) overlapped.

Table 1. Model architectures for ResNet. The number of stacked building blocks is indicated in brackets. Downsampling is performed by conv3(1), conv4(1), and conv5(1) with a stride of 2.

Layer Name	Output Size	50-Layer	101-Layer	152-Layer
Conv1	112 × 112	7 × 7, 64, stride 2		
Conv2(x)	56 × 56	3 × 3 max pool, stride 2		
		$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$
Conv3(x)	28 × 28	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 8$
Conv4(x)	14 × 14	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 36$
Conv5(x)	7 × 7	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$
	1 × 1	Average pol, 1000-d fc, softmax		
FLOPs		3.8×10^9	7.6×10^9	11.3×10^9

2.4.2. VGG Model Architecture

Visual geometry group (VGG) is a simple and effective CNN architecture proposed by [34]. A convolutional layer, an activation layer, a pooling layer, and a dense layer make up the VGG hierarchical structure of the CNN. Among these, the convolutional layer is one that is essential. By implementing "local perception" and "parameter sharing" in

two different methods, the objectives of feature extraction and dimensionality reduction processing are achieved. The convolution kernel is the main element of the convolution layer. The convolution kernel enables the retrieval of the shape of an object from several spots within an image, which minimises the dimensionality and the number of parameters that must be learned [35]. Smaller filters (3×3) were used in this network, which minimised its computational complexity by lowering the number of parameters.

The VGG architecture starts by passing the image dataset through a stack of convolutional layers. VGG-16 has 13 convolutional layers and 3 fully connected layers, whereas VGG-19 has 16 convolutional layers and 3 fully connected layers. Both VGG models require an image size of $224 \times 224 \times 3$, which is an RGB image with pixel size of 224×224. A filter with size of 3×3 captures the concept of the left, right, top, bottom, and centre of the image. The convolution process was achieved with a 1 pixel stride. Spatial padding was utilised to maintain the spatial resolution of an image following convolution. Two 64 convolution kernels were processed in the first stage of the convolution process, and two 128 convolution kernels were processed in the second stage. During the third to fifth stage, VGG-16 and VGG-19 used different numbers of convolution. For VGG-16, 256, 512 and 512 convolution kernels were used from third to fifth stage, where in each stage the convolution was repeated 3 times. VGG-19 had the same convolution kernels with VGG-16, with the difference being on the number of convolution processes, which is 4. After each stage, max-pooling was performed on the output. Maximum pooling was carried out with a 2×2 pixel window and a stride size of 2. The next step was performing fully connected layers 3 times, followed by a soft-max layer. Table 2 shows the two VGG architectures used in this study, i.e., VGG-16 and VGG-19.

2.5. Experimental Setup

All of the training, validation, and testing processes were conducted using Jupyter Notebook (version 6.4.12) on a 64-bit Windows 11 operating system. The system was equipped with an AMD Ryzen 5 2600X processor running at 3.6 GHz and 16 GB RAM. The model training utilised the processing power of an NVIDIA GeForce RTX 3060 GPU with 12 GB VRAM, using CUDA API version 11.2. The algorithm was implemented using Keras, which was a deep learning API that operated on the TensorFlow library for machine learning platform. The experimental setups were carefully tuned to fully utilise the memory capacity of the GPU.

2.6. Pre-Processing

A total of 7328 original sample images were used in the experiment. These samples were randomly divided into the training set, validation set, and test set in the proportions of 80:10:10. The training samples were augmented, resulting in a total of 187,456 images after the augmentation process. The distribution of classes in the training dataset was as follows: BPH (35,264), GLH (40,992), ZIGZAG (29,568), and WBPH (81,632). Table 3 shows the detail of the dataset that been used for this experiment.

The training data exhibited class imbalance. Nevertheless, the extensive number of samples utilised in this study for training purposes could sufficiently mitigate concerns related to overfitting and bias. Additionally, to preserve the integrity of the images and prevent distortion caused by resizing, which could affect the geometry and shape of the samples, each image was scaled to 256 pixels in its smallest dimension while maintaining its original aspect ratio. Subsequently, the images were randomly cropped to a size of 224×224 pixels. These measures were implemented to expedite the model training process.

The model weights were randomly initialised. The group parameters of the last dense layer were modified for each model to accommodate four classes, which represented the total number of classes in this study. All models utilised a softmax activation function in the final layer. The SGD optimiser was employed, with a categorical cross-entropy loss function. The optimiser was configured with a learning rate of 0.0005, momentum of 0.9, and without Nesterov. A batch size of 32 was set for all models. The models were

trained for 20 epochs, with early stopping implemented when there was no improvement in validation loss after 3 epochs. An epoch refers to a complete iteration of the training data through the algorithm.

Table 2. Architecture for VGG-16 and VGG-19.

VGG-16	**VGG-19**
16 weight layers	19 weight layers
Input (224 × 224 RGB image)	
conv3-64 conv3-64	conv3-64 conv3-64
maxpool	
conv3-128 conv3-128	conv3-128 conv3-128
maxpool	
conv3-256 conv3-256 conv3-256	conv3-256 conv3-256 conv3-256 conv3-256
maxpool	
conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool	
conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool	
FC-4096	
FC-4096	
FC-1000	
soft-max	

Table 3. Details of planthopper dataset.

Planthopper Name	**Number of Image Sample**				
	Original Image	**Training Image**	**Training Image (Augmented)**	**Validation Image**	**Testing Image**
Brown Planthopper	1379	1102	35,264	139	137
Green Leafhopper	1603	1281	40,992	161	160
Zigzag Leafhopper	1156	924	29,568	116	115
White-Backed Planthopper	3190	2551	81,632	320	318
Image Total	7328	5858	187,456	736	730

2.7. Performance Metrics

The primary objective of this research was to develop a classification model for distinguishing four different planthopper types. The accuracy of classification was deemed

the most crucial aspect of this multi-class task. Therefore, performance metrics based on the confusion matrix, including accuracy, precision, recall, and F1-score, were utilised to compare the performance of various models. Accuracy, which represents the proportion of correctly predicted observations to all observations, is the simplest performance metric to understand. However, when dealing with imbalanced datasets, accuracy may not provide a clear picture of the performance of the model, as imbalanced datasets often exhibit a bias toward the dominant class [36]. Hence, the F1-score, which combines precision and recall, is a more suitable metric for imbalanced datasets. Precision measures the proportion of correctly predicted positive observations out of all projected positive observations, while recall assesses the proportion of correctly predicted positive observations out of all instances in the true positive class. Accuracy, precision, recall, and F1-score are computed using indices such as true positive (tp_i), false positive (fp_i), true negative (tn_i), and false negative (fn_i). True positive and true negative refer to the model correctly predicting the positive class or negative class, respectively. A false positive occurs when the model incorrectly predicts the positive class, whereas a false negative occurs when the model inaccurately predicts the negative class. In this study, the Scikit-learn library was utilised to plot the performance metrics for the training and test results. The equations used to calculate accuracy, precision, recall, and F1-score for each individual class (i) are provided in Table 4. Table 5 displays the performance metrics used to calculate the average performance across all classes (n), which include average accuracy, macro-average precision, macro-average recall, and macro-average F1-score.

Table 4. Metrics for assessing the performance of individual classification classes.

Measure	Formula
Accuracy_i	$\frac{tn_i + tp_i}{tp_i + fp_i + tn_i + fn_i}$
$\text{Precision}_\text{i}$	$\frac{tp_i}{tp_i + fp_i}$
Recall_i	$\frac{tp_i}{tp_i + fn_i}$
F1-Score_i	$2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i}$

Table 5. Metrics for assessing the average performance of all classification classes.

Measure	Formula
Average Accuracy	$\frac{\sum_{i=1}^{n} Accuracy_i}{n}$
Macro Average Precision	$\frac{\sum_{i=1}^{n} Precision_i}{n}$
Macro Average Recall	$\frac{\sum_{i=1}^{n} Recall_i}{n}$
Macro Average F1-Score	$\frac{\sum_{i=1}^{n} F1{-}Score_i}{n}$

3. Results and Discussion

This section provides an in-depth analysis of the training, validation, and test outcomes achieved using the sample. The results are presented by comparing the performance metrics of all of the models.

3.1. Model Comparisons

The stopping criterion for each model was set to 20 epochs. However, we also introduced an early stopping criterion where the model would stop if there was no improvement in the validation loss after three epochs. Figure 8 displays the validation loss results, while Figure 9 illustrates the validation accuracy results for each of the five CNN models. ResNet-152 and ResNet-50 stopped at the fifth epoch, ResNet-101 stopped at the eighth epoch, VGG-16 stopped at the ninth epoch, and VGG-19 stopped at the 15th epoch. Initially, ResNet-152 and ResNet-50 exhibited high loss values, surpassing 1. On the other hand, the

loss values for VGG-16 and VGG-19 models showed gradual and minimal fluctuations. ResNet-101 exhibited fluctuations in the loss rate, with a final loss rate of 0.5647 before stopping, compared to its lowest value of 0.181. ResNet-152 had the lowest loss rate of 0.1348, followed by ResNet-101 with 0.1819. The highest loss value was observed in VGG-16, with a value of 0.2515.

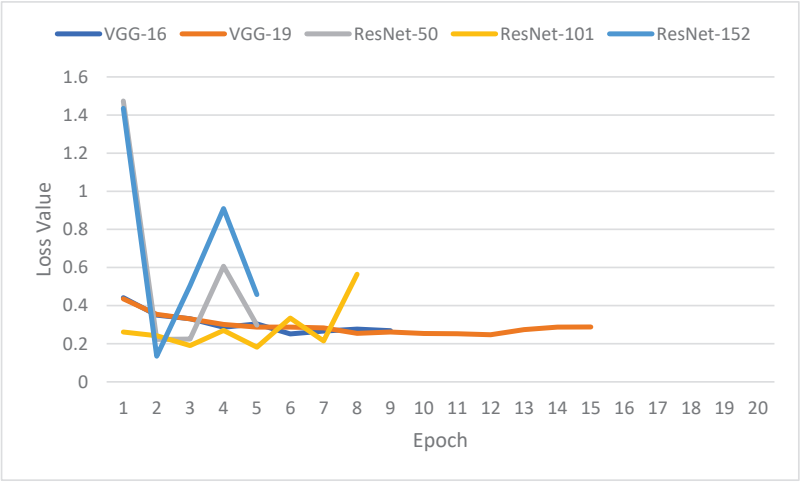


Figure 8. Value of validation loss for all models.

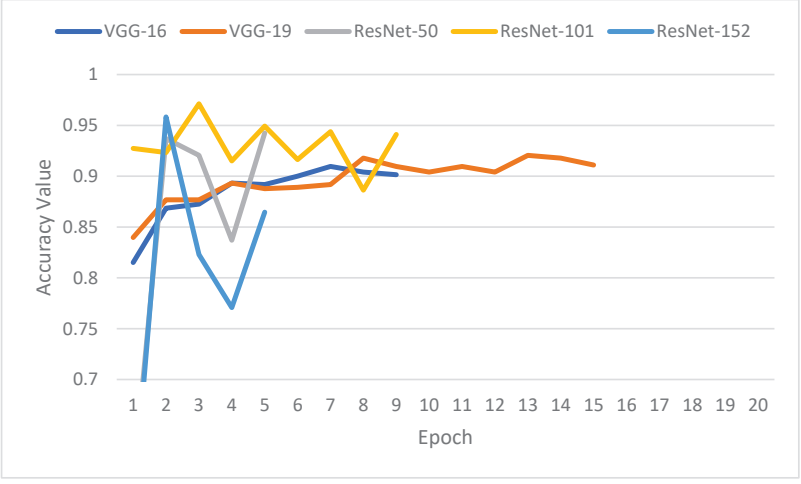


Figure 9. Value of validation accuracy for all models.

In terms of validation accuracy, ResNet-152 achieved the highest accuracy with a value of 0.9583 at its lowest loss, followed by ResNet-50 with a value of 0.937. VGG-16 had the lowest accuracy value of 0.9 at its lowest loss.

Figure 10 displays the average time taken by each model to complete one epoch of training and validation. The plot reveals that ResNet-152 required the longest time, taking 50.17 min, while VGG-16 had the shortest time, at 12.53 min. This pattern indicated that models with more layers required more time to complete the training and validation process. Table 6 presents the average prediction time for a single sample. According to the table, VGG-16 exhibited the fastest prediction time, taking 0.022 s, followed by VGG-16

with 0.023 s. On the other hand, ResNet-152 had the longest prediction time, of 0.051 s. These results indicated that the training time directly influenced the prediction time for all models.

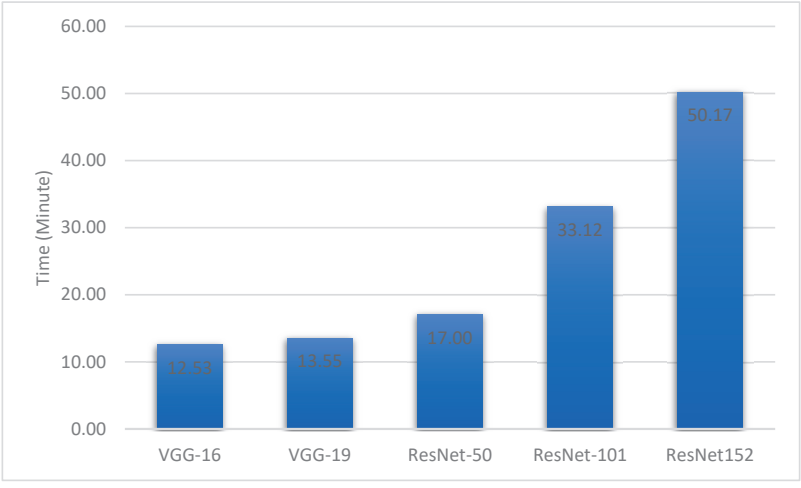


Figure 10. Model comparison based on time taken to complete training and validation.

Table 6. Average time for predicting an image sample.

Model	Prediction Time (s)
ResNet-50	0.026
ResNet-101	0.037
ResNet-152	0.051
VGG-16	0.022
VGG-19	0.023

Figure 11 presents a comparison of the results for each model based on accuracy, precision, recall, and F1-score for the test dataset. Each model achieved an average accuracy greater than 93.68%, which is impressive and meets the expected performance due to the utilisation of a large dataset. Among the models, ResNet-50 exhibited slightly superior performance with an accuracy value of 97.28%, while the model with the lowest performance was ResNet-152 with an accuracy of 93.68%. ResNet-101 demonstrated the second-highest performance with an accuracy of 97.15%, followed by VGG-16 (96.81%) and VGG-19 (95.92%). In terms of all performance metrics, ResNet-50 outperformed other models, with the lowest score in precision at 92.05%. Despite the imbalanced datasets, the performance of the models based on F1-scores exhibited a pattern similar to accuracy with slightly lower scores. ResNet-50 achieved the highest F1-score value at 93.07%, followed by ResNet-101 (91.91%), VGG-16 (91.34%), VGG-19 (89.36%), and ResNet-152 (86.59%). These results indicated that the utilisation of a large number of samples for training and testing in this study provided sufficient data to mitigate overfitting and bias concerns. Despite not having the fastest training time, ResNet-50 was considered the best choice for the classification model, as it only required 0.026 s to classify one sample during testing. Training was only conducted if there were new varieties of planthopper samples.

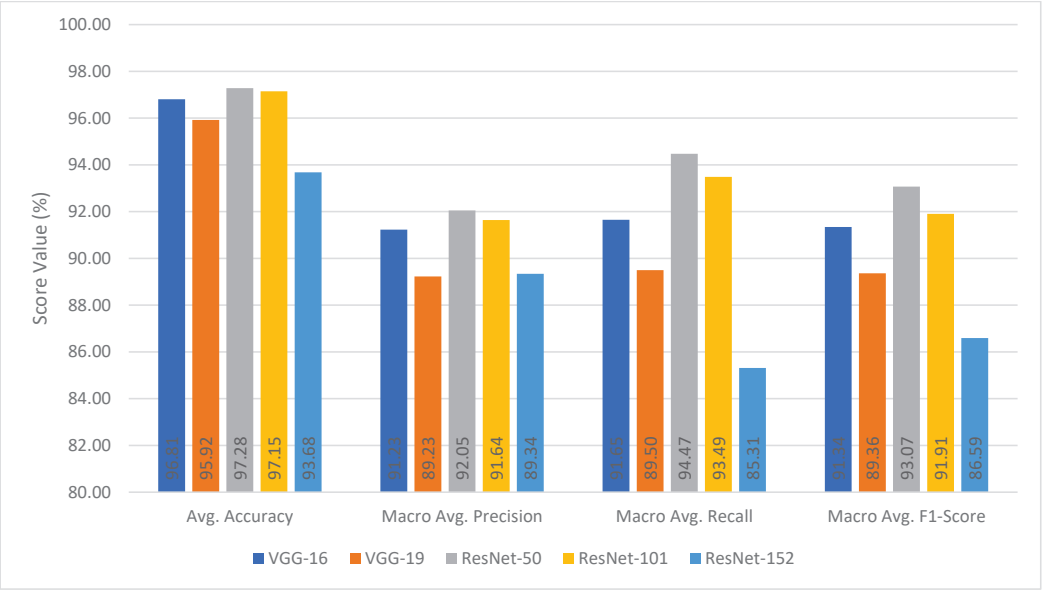


Figure 11. Model comparison based on average accuracy, macro average precision, macro average recall, and macro average F1-score in percentage.

3.2. Error Analysis

This section discusses the analysis of the error on the prediction performed by the best performing model, which was ResNet-50. Figure 12 shows the confusion matrix for the ResNet-50 model. Recall was measured as the ratio of samples that were correctly assigned to a class to the total samples in that class. As for precision, it was determined by dividing the number of samples that were correctly classified by their classification by the number of anticipated classes.

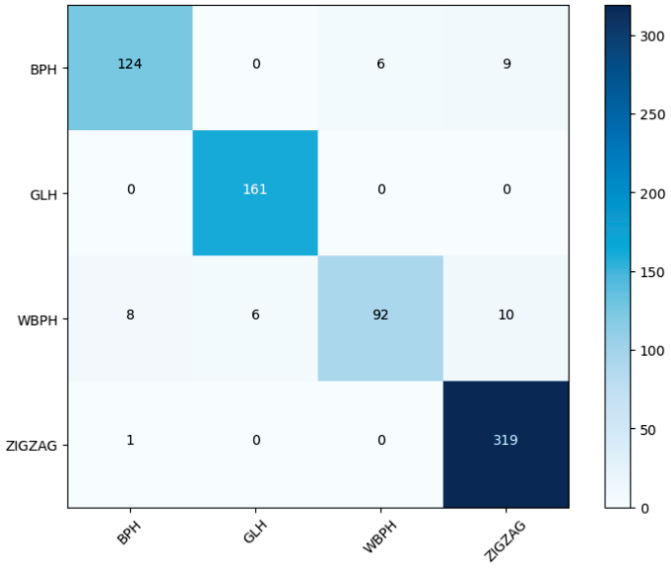


Figure 12. Confusion matrix for ResNet-50 model.

The confusion matrix revealed that the model accurately categorised 161 GLH samples, demonstrating almost perfect classification. The majority of misclassifications occurred between BPH and WBPH, with 8 out of 116 WBPH samples misclassified as BPH, and 6 out of 139 BPH samples misclassified as WBPH. This was because WBPH and BPH were nearly identical in size and shape, with identical head and body proportions. When the backs of the planthoppers were clearly visible, these two groups could be easily distinguished. WBPH had a white line on their head and wings, while BPH had a distinct shape on their wings. However, if the backs of the planthoppers were not visible, it became extremely difficult to identify them. Alternative methods included reviewing the body colour of a BPH, which should be entirely brown, or searching for a white stripe on the face or side of a WBPH. WBPH generally had a darker body colour than BPH, but it could still be challenging to distinguish them if the BPH had a body colour similar to that of the WBPH. For GLH, it could be distinguished from other classes by its green body with a black stripe on its back. However, if viewed from the side, other samples could also be misclassified as GLH. As indicated by the confusion matrix, six WBPH samples were misclassified as GLH, while nine ZIGZAG samples were misclassified as GLH. This was because GLH also had a black body beneath the wings, sharing a characteristic with the other classes.

Figure 13 presents a comparison of accuracy, recall, precision, and F1-score for each planthopper class classified using ResNet-50. From the plot, it can be observed that GLH had the highest values for accuracy (95.52%), precision (100%), recall (96.41%), and F1-score (98.17%). On the other hand, WBPH had the lowest values for accuracy (95.52%), precision (79.31%), and F1-score (85.58%). Comparing these results with those obtained by [32], our model outperformed their proposed method with 2.28% higher accuracy for four types of planthoppers instead of only two types of planthoppers.

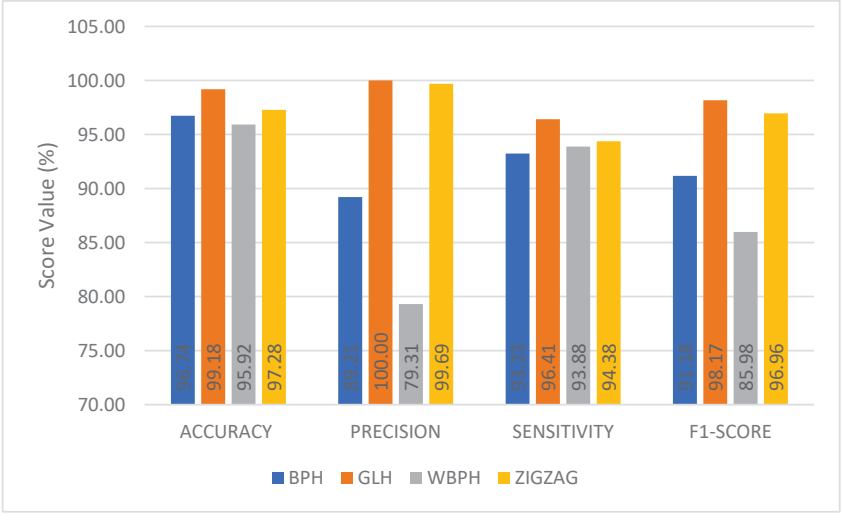


Figure 13. Comparison of accuracy, precision, recall and F1-score for each planthopper class using ResNet-50.

4. Conclusions

Detecting insect pests is crucial in agriculture, particularly in paddy fields, as it facilitates the assessment of their population dynamics and density. Accurate detection allows for precise and targeted application of pesticides. However, automatically detecting insects using image processing presents challenges due to the unpredictable nature of their environment. The presence of imperfect samples of trapped insects and inconsistent categorisation by humans further complicates the task. To overcome these challenges, this study proposed an automated detection method for planthoppers using deep CNN.

Five models, namely ResNet-50, ResNet-101, ResNet-152, VGG-16, and VGG-19, were employed to train on randomly initialised weights and identify the characteristics of four planthopper classes: BPH, GLH, WBPH, and ZIGZAG. The planthopper images were converted to RGB format and augmented to increase the sample size. A total of 7328 images were used, with 80% allocated for training, 10% for validation, and 10% for testing. The performance of these five approaches was evaluated using accuracy, precision, recall, and F1-score. The results demonstrated that ResNet-50 achieved the highest performance, with an average classification accuracy of 97.28% and individual class accuracies of 96.74% for BPH, 99.18% for GLH, 95.52% for WBPH, and 97.28% for ZIGZAG. It is important to note that the classification was performed on image samples that were previously manually cropped by an expert. Although the proposed method demonstrated promising results, it had a limitation in the case of borderline cases. In this study, we observed that these borderline cases were frequently misclassified. Furthermore, overlapping samples posed a substantial issue in the classification process. Overlapping insects were more prevalent when data collection was conducted over a longer duration. We limited the sample collection duration to one hour per sample. Consequently, the collected samples contained fewer overlapping insects. The overlapping issue can be addressed in the future work to enhance the robustness of the classification process. Developing methods to handle overlapping samples could significantly improve the performance of the classification process. Furthermore, an additional effort can be undertaken to integrate the processes of object detection and classification to integrate them into a single step, aiming to fully automate the counting of planthoppers. Additionally, the capability of other deep learning architectures for planthopper classification can also be studied in the future.

Author Contributions: Conceptualisation, M.F.I. and S.K.-B.; methodology, M.F.I. and S.K.-B.; software, M.F.I.; formal analysis, M.F.I. and S.K.-B.; validation, M.F.I. and S.K.-B.; investigation, M.F.I. and M.A.M.B.; resources, M.F.I. and M.A.M.B.; data curation, M.F.I. and M.A.M.B.; writing—original draft preparation, M.F.I.; writing—review and editing, S.K.-B.; visualisation, M.F.I. and S.K.-B.; supervision, S.K.-B., M.H., M.J. and F.S.A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to restrictions.

Acknowledgments: The authors would like to give thanks to MARDI for providing a machine vision system for data collection, assisting in field sampling and helping in identifying the pest dataset.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bray, F. Rice in Malaya: A Study in Historical Geography. *AAG Rev. Books* **2014**, *2*, 127–129. [CrossRef]
2. Che Omar, S.; Shaharudin, A.; Tumin, S.A. The Status of the Paddy and Rice Industry in Malaysia. 2019. Available online: http://www.krinstitute.org/assets/contentMS/img/template/editor/20190409_RiceReport_FullReport_Final.pdf (accessed on 14 January 2023).
3. Zakaria, M.B.; Nik Abdul Ghani, N.A.R. An Analysis of Rice Supply in Malaysia Post COVID-19—From an Agriculture-Related Fiqh Perspective. *Int. J. Acad. Res.Account. Financ. Manag. Sci.* **2022**, *12*, 150–160. [CrossRef] [PubMed]
4. Norliza, A.B.; Pritchard, J. Identification of candidate genes involved in brown planthopper resistance in rice using microarray analysis. *J. Trop. Agric. Food Sci.* **2016**, *44*, 49–62.
5. Hong-Xing, X.; Ya-Jun, Y.; Yan-Hui, L.; Xu-Song, Z.; Jun-Ce, T.; Feng-Xiang, L.; Qiang, F.; Zhong-Xian, L. Sustainable Management of Rice Insect Pests by Non-Chemical-Insecticide Technologies in China. *Rice Sci.* **2017**, *24*, 61–72. [CrossRef]
6. Witzgall, P.; Kirsch, P.; Cork, A. Sex Pheromones and Their Impact on Pest Management. *J. Chem. Ecol.* **2010**, *36*, 80–100. [CrossRef]
7. Bookeri, M.A.M.; Masaruddin, M.F.; Shah, N.A.A.; Noh, A.M.; Samsuri, N.S.; Abu Bakar, B.H.; Khadzir, M.K. Evaluation of Light Trap System in Monitoring of Rice Pests, Brown Planthopper (*Nilaparvata lugens*). *Adv. Agric. Food Res. J.* **2021**, *3*, 1–7. [CrossRef]

8. Kumar, R.; Martin, V.; Moisan, S.; Sophia, I.; Méditerrannée, A. Robust Insect Classification Applied to Real Time Greenhouse Infestation Monitoring. In Proceedings of the 20th International Conference on Pattern Recognition on Visual Observation and Analysis of Animal and Insect Behavior Workshop, Istanbul, Turkey, 22 August 2010; pp. 1–4.
9. Mundada, R.G.M.R.G. Detection and Classification of Pests in Greenhouse Using Image Processing. *IOSR J. Electron. Commun. Eng.* **2013**, *5*, 57–63. [CrossRef]
10. Kiratiratanapruk, K.; Temniranrat, P.; Sinthupinyo, W.; Prempree, P.; Chaitavon, K.; Porntheeraphat, S.; Prasertsak, A. Development of Paddy Rice Seed Classification Process using Machine Learning Techniques for Automatic Grading Machine. *J. Sens.* **2020**, *2020*, 7041310. [CrossRef]
11. Bhupendra; Moses, K.; Miglani, A.; Kumar Kankar, P. Deep CNN-based damage classification of milled rice grains using a high-magnification image dataset. *Comput. Electron. Agric.* **2022**, *195*, 106811. [CrossRef]
12. Hassanzadeh, T.; Essam, D.; Sarker, R. EvoDCNN: An evolutionary deep convolutional neural network for image classification. *Neurocomputing* **2022**, *488*, 271–283. [CrossRef]
13. Weng, S.; Tang, P.; Yuan, H.; Guo, B.; Yu, S.; Huang, L.; Xu, C. Hyperspectral imaging for accurate determination of rice variety using a deep learning network with multi-feature fusion. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2020**, *234*, 118237. [CrossRef]
14. Setiawan, A.; Yudistira, N.; Wihandika, R.C. Large scale pest classification using efficient Convolutional Neural Network with augmentation and regularizers. *Comput. Electron. Agric.* **2022**, *200*, 107204. [CrossRef]
15. Zheng, T.; Yang, X.; Lv, J.; Li, M.; Wang, S.; Li, W. An efficient mobile model for insect image classification in the field pest management. *Eng. Sci.Technol. Int. J.* **2023**, *39*, 101335. [CrossRef]
16. Peng, Y.; Wang, Y. CNN and transformer framework for insect pest classification. *Ecol. Inform.* **2022**, *72*, 101846. [CrossRef]
17. Wei, D.; Chen, J.; Luo, T.; Long, T.; Wang, H. Classification of crop pests based on multi-scale feature fusion. *Comput. Electron. Agric.* **2022**, *194*, 106736. [CrossRef]
18. Huang, M.-L.; Chuang, T.-C.; Liao, Y.-C. Application of transfer learning and image augmentation technology for tomato pest identification. *Sustain.Comput. Inform. Syst.* **2022**, *33*, 100646. [CrossRef]
19. Li, J.; Tang, Y.; Zou, X.; Lin, G.; Wang, H. Detection of Fruit-Bearing Branches and Localization of Litchi Clusters for Vision-Based Harvesting Robots. *IEEE Access* **2020**, *8*, 117746–117758. [CrossRef]
20. Ding, W.; Taylor, G. Automatic moth detection from trap images for pest management. *Comput. Electron. Agric.* **2016**, *123*, 17–28. [CrossRef]
21. Sun, Y.; Liu, X.; Yuan, M.; Ren, L.; Wang, J.; Chen, Z. Automatic in-trap pest detection using deep learning for pheromone-based Dendroctonus valens monitoring. *Biosyst. Eng.* **2018**, *176*, 140–150. [CrossRef]
22. Jiao, L.; Xie, C.; Chen, P.; Du, J.; Li, R.; Zhang, J. Adaptive feature fusion pyramid network for multi-classes agricultural pest detection. *Comput. Electron. Agric.* **2022**, *195*, 106827. [CrossRef]
23. Hadipour-Rokni, R.; Asli-Ardeh, E.A.; Jahanbakhshi, A.; Paeen-Afrakoti, I.E.; Sabzi, S. Intelligent detection of citrus fruit pests using machine vision system and convolutional neural network through transfer learning technique. *Comput. Biol. Med.* **2023**, *155*, 106611. [CrossRef]
24. Shi, Z.; Dang, H.; Liu, Z.; Zhou, X. Detection and Identification of Stored-Grain Insects Using Deep Learning: A More Effective Neural Network. *IEEE Access* **2020**, *8*, 163703–163714. [CrossRef]
25. Li, R.; Jia, X.; Hu, M.; Zhou, M.; Li, D.; Liu, W.; Wang, R.; Zhang, J.; Xie, C.; Liu, L.; et al. An Effective Data Augmentation Strategy for CNN-Based Pest Localization and Recognition in the Field. *IEEE Access* **2019**, *7*, 160274–160283. [CrossRef]
26. Xia, D.; Chen, P.; Wang, B.; Zhang, J.; Xie, C. Insect detection and classification based on an improved convolutional neural network. *Sensors* **2018**, *18*, 4169. [CrossRef]
27. Liu, Y.; Liu, S.; Xu, J.; Kong, X.; Xie, L.; Chen, K.; Liao, Y.; Fan, B.; Wang, K. Forest pest identification based on a new dataset and convolutional neural network model with enhancement strategy. *Comput. Electron. Agric.* **2022**, *192*, 106625. [CrossRef]
28. Xing, S.; Lee, M.; Lee, K.-K. Citrus pests and diseases recognition model using weakly dense connected convolution network. *Sensors* **2019**, *19*, 3195. [CrossRef]
29. Rahman, C.R.; Arko, P.S.; Ali, M.E.; Khan, M.A.I.; Apon, S.H.; Nowrin, F.; Wasif, A. Identification and recognition of rice diseases and pests using convolutional neural networks. *Biosyst. Eng.* **2020**, *194*, 112–120. [CrossRef]
30. Yao, Q.; Lv, J.; Liu, Q.-J.; Diao, G.-Q.; Yang, B.-J.; Chen, H.-M.; Tang, J. An Insect Imaging System to Automate Rice Light-Trap Pest Identification. *J. Integr. Agric.* **2012**, *11*, 978–985. [CrossRef]
31. Yao, Q.; Feng, J.; Tang, J.; Xu, W.-G.; Zhu, X.-H.; Yang, B.-J.; Lü, J.; Xie, Y.-Z.; Yao, B.; Wu, S.-Z.; et al. Development of an automatic monitoring system for rice light-trap pests based on machine vision. *J. Integr. Agric.* **2020**, *19*, 2500–2513. [CrossRef]
32. Nazri, A.; Mazlan, N.; Muharam, F. PENYEK: Automated brown planthopper detection from imperfect sticky pad images using deep convolutional neural network. *PLoS ONE* **2018**, *13*, e0208501. [CrossRef]
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
34. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2019; pp. 1–14.

35. Shin, H.-C.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; Summers, R.M. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging* **2016**, *35*, 1285–1298. [CrossRef] [PubMed]
36. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
www.mdpi.com

Agriculture Editorial Office
E-mail: agriculture@mdpi.com
www.mdpi.com/journal/agriculture

Academic Open
Access Publishing

mdpi.com

ISBN 978-3-0365-8851-3