

Article

AI-Based Potato Crop Abiotic Stress Detection via Instance Segmentation

Emmanouil Savvakis ¹, Dimitrios Kapetas ¹, María del Carmen Martínez-Ballesta ², Nikolaos Katsoulas ³ and Eleftheria Maria Pechlivani ^{1,*}

¹ Centre for Research and Technology Hellas, Information Technologies Institute, 57001 Thessaloniki, Greece; msavvakis@iti.gr (E.S.); dimikape@iti.gr (D.K.)

² Department of Agricultural Engineering, Technical University of (UPCT), Paseo Alfonso XIII 48, 30203 Cartagena, Spain; mcarmen.ballesta@upct.es

³ Laboratory of Agricultural Constructions and Environmental Control, Department of Agriculture Crop Production and Rural Environment, University of Thessaly, 38446 Volos, Greece; nkatsoul@uth.gr

* Correspondence: riapechl@iti.gr; Tel.: +30-231-125-7751

Abstract

Background: Automated monitoring of crop health and the precise detection of abiotic stress, such as herbicide damage, are demanding challenges for modern agriculture. Abiotic stresses are a demanding challenge for modern agriculture, responsible for up to 82% of yield losses in major food crops. To address this, researchers are increasingly leveraging artificial intelligence (AI) to automate the detection and management of these stressors.

Methods: In particular, this paper presents an instance segmentation framework to precisely detect interveinal chlorosis and leaf curling on potato leaves, two common symptoms of herbicide damage and soft wind. Within the context of precision agriculture and the need to address the inherent ambiguity in manual leaf assessment, this study employs a partial label learning approach to refine the dataset. This method utilizes an EfficientNet-b1 model to classify ambiguous samples, generating high-confidence pseudo-labels for instances that are difficult to categorize visually. The core of the proposed framework is a Mask2Former model, which is first fine-tuned on general potato leaf dataset to enhance its segmentation capabilities and then transferred on the refined, pseudo-labeled dataset.

Results & Conclusions: This two-stage approach yields a highly accurate segmentation tool, achieving 89% mAP₅₀ and a pseudo-label classification accuracy of 95%, designed for integration into smart agriculture systems like ground level robotics or unmanned aerial vehicles for real-time, automated crop monitoring.

Keywords: Mask2Former; deep learning; partial label learning; precision agriculture; interveinal chlorosis; leaf curling



Academic Editor: Arslan Munir

Received: 21 January 2026

Revised: 3 March 2026

Accepted: 12 March 2026

Published: 16 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Potato crops represent a cornerstone of the global food chain, ranking as the third most important food supply worldwide. It is essential to ensure and prevent the impact of stressors that can affect and impair their production. Abiotic stresses are a rigorous test for modern agriculture, resulting in yield losses in major food crops that can reach up to 82% and millions of USD [1]. In this study, two major symptoms of herbicide damage and soft wind are examined—interveinal chlorosis and leaf curling. Interveinal chlorosis is recognized by the yellowing of leaf tissue, triggered by the reduction or destruction of chlorophyll—the pigment essential for photosynthesis [2]. Leaf curling is a plant's physical

response and can be caused by “wind stress”, generally characterized by the leaf margins curling upwards, potentially browning on the edges due to water loss [3,4]. To reduce the costs resulting from these factors, precise symptom detection using technological tools is vital.

Automated detection of abiotic stressors, using modern solutions like artificial intelligence, can play a decisive role in combating and reducing the number of crops affected by them. Traditional methods that include irrigation practices and fertilizer application have massive limitations, leading to salinity buildup in soil and pollution. Furthermore, manual scouting is labor-intensive, subjective and often fails to detect early-stage symptoms that are not yet visually distinct. To overcome these limitations, Park et al. (2025) [5] utilized hyperspectral imaging to identify drought and heat stress with high precision by analyzing multi-spectral bands and Jia et al. (2024) [6] demonstrated the efficacy of UAV-based spectral sensors for detecting viral diseases on potato plants at a commercial grower’s field. While powerful, these spectral approaches often require expensive, specialized hardware that is possibly unavailable to most farmers. Moreover, Lapajne et al. (2025) [7] noted that they can struggle with diagnostic ambiguity in potato crops since spectral signatures of abiotic stresses can often overshadow those of biotic infections, making precise diagnosis complicated. In contrast, deep learning models applied to computer vision can recognize complex patterns and make accurate predictions based on the hidden features learned from RGB leaf imagery [8,9]. These methods can effectively transform modern agriculture practices into more sustainable and economically viable pesticide management and water efficiency.

In particular, Convolutional Neural Network-based models have shown great results in abiotic plant disease detection classification [10]. The limitation of these models is their tendency to focus on local feature extraction, like texture and edges, which may lead to suboptimal performance in some cases [11]. Certain symptoms are not identified just in a region, but may show global patterns and long-range relationships in an image. To address these challenges, Transformer-based models have been proven to excel in general object detection and have proven effective in a variety of applications [12].

Transformers have played a pivotal role in the development of Neural Networks, since attention mechanism solves limitations in data’s long-distance dependencies [13]. Even though originally designed for Natural Language Processing, self-attention was adapted for image classification by the Vision Transformer (ViT) [14], which divided images into patches, treating them as a sequence of tokens and proving that Transformers could deliver powerful results in Computer Vision. A significant leap occurred when Carion et al. (2020) [15] introduced Detection Transformer (DETR), a Transformer model with learnable object queries, which are trained to “ask” about specific objects in an image, that are passed to a Transformer decoder. This innovation led to the elimination of anchor boxes [16] and non-maximum suppression techniques that were used excessively to detect varying-sized objects and filter candidate bounding boxes. Cheng et al. (2021) [17] extended this query-based approach to segmentation by MaskFormer, reframing segmentation as a “mask classification” problem, where queries predict pairs of masks and class labels, unifying semantic and instance segmentation. To deal with MaskFormer’s cross-attention intensive computational cost, Cheng et al. (2022) [18] proposed “masked attention” within its decoder, defining the Masked-attention Mask Transformer (Mask2Former) model, which constraints each query to attend only to localized features within its predicted mask region, achieving faster convergence, higher performance and efficiency. Mask2Former architecture has been a breakthrough to the image segmentation task, since it combines the Masked Attention mechanism to a single framework to solve instance, semantic and panoptic problems, detecting multi-scale object with high accuracy.

Recently, Zhang and Liu (2025) [19] proposed a Swin-Transformer based model, utilizing TokenEmbedder and Axial Transformer modules to significantly reduce the computational cost and number of parameters by 46% compared to the base architecture and exploited it to detect a variety of plant diseases, such as leaf blight on corn and septoria leaf spots on tomato leaves. In order to segment plant and leaf instances Darbyshire, Sklar and Parsons (2023) [20] followed an adapted architecture that uses separate Transformer decoders while also using a single layer for each feature level, instead of three as used in the original Mask2Former implementation, scoring the Panoptic Quality [21] of crops and leaves equal to 70.18% and 66.91% respectively in the panoptic segmentation task. Wei et al. (2024) [22] introduced a large-scale growth monitoring technique for lettuce seedlings, integrating multidimensional collaborative attention into Mask2Former for precise instance segmentation and localization, reaching mean Average Precision (mAP) at an Intersection over Union of 0.5 to 0.95 ($\text{mAP}_{0.50:0.95}$) equal to 78.23%.

In related research, Christakakis et al. (2024) [23] optimized the early detection of *Botrytis cinerea* in cucumber plants using ViTs and “Cut-and-Paste” data augmentation to address dataset imbalance, achieving an overall accuracy of 92%. Building on these methodologies, Kapetas et al. (2024) [24] established a framework for early *B. cinerea* detection in tomato crops using YOLOv8 segmentation and Transformer ensembles, achieving 79.41% accuracy. This work was later advanced by Kapetas et al. (2025) [25], who transitioned to a YOLOv11-based architecture and incorporated five derived vegetation indices (CVI, GNDVI, NDVI, NPCI, and PSRI) to detect symptoms on pepper plants. This refined approach accomplished a significantly improved overall accuracy of 87.42% and successfully validated the model’s sensitivity against molecular RT-qPCR assays for fungal biomass estimation.

This study presents a tool for instance segmentation of potato leaves and identification of interveinal chlorosis and leaf curling, using deep learning methods. The training pipeline consists of a Mask2Former model, selected specifically for its ability to model the long-range dependencies inherent in abiotic stress symptoms, fine-tuned on a binary dataset to develop a strong segmentation foundation on potato leaves. Current methods in precision agriculture rely on definitive, discrete labels, which forces a critical trade-off when facing ambiguity: ambiguous samples are either discarded or incorrectly labeled, introducing label noise. To overcome this limitation, this study’s dataset was enabled to have ambiguous samples and a Partial Labeling Learning (PLL) [26] approach is introduced, allowing for a candidate set of labels with a single true, latent label. This latent label is extracted using an EfficientNet-b1 [27] classifier, trained with instance masks cropped from the dataset while exploring different disambiguation and mask cropping methodologies. Finally, the initial Mask2Former model is transferred and fine-tuned on this refined, pseudo-labeled dataset, targeting the two abiotic stress symptoms, yielding precise instance masks while maximizing data utilization. Experimental evaluations demonstrate the efficacy of this framework, achieving an instance segmentation mAP of 89% and a label disambiguation accuracy of 95% on the validation set.

The structure of this paper is organized as follows. Section 2, Materials and Methods, details the dataset acquisition and the proposed two-stage training pipeline, including the initial fine-tuning of the Mask2Former model and the implementation of the partial label learning (PLL) framework using an EfficientNet-b1 classifier to refine ambiguous annotations. Section 3, Results, presents the experimental evaluation of the segmentation model across different backbone configurations and image resolutions, analyzing the impact of the proposed methodology on detection accuracy. Finally, Section 4, Discussion, concludes the study by summarizing the key findings and discussing potential future applications for integration into smart agriculture systems.

2. Materials and Methods

Figure 1 presents a high-level flowchart that outlines key stages of the proposed methodology, which are discussed in detail in the following sections.

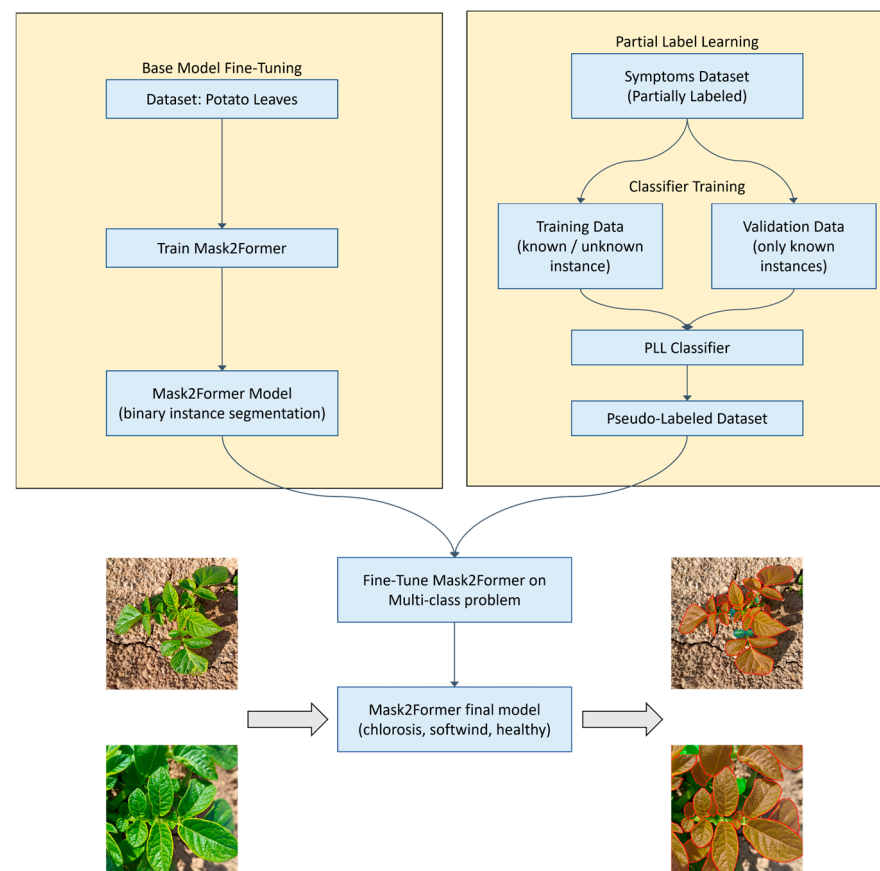


Figure 1. High-level flowchart of the procedure followed. At first, Mask2Former model is fine-tuned on a potato leaves dataset in a binary class setting. Then, a partial label learning approach is followed to classify unknown labeled instance. Finally, Mask2Former model is transferred to the multi-class instance segmentation problem. The masks in red color are example segmentations made by the model.

2.1. Dataset Collection

The collection of data was performed utilizing a standard commercial smartphone camera (manufacturer: Xiaomi, Beijing, China) with approximately 12 MP effective resolution to introduce device-agnostic variability into the dataset. This approach ensures the model's generalization capabilities across different hardware typically available to farmers. All images were captured in standard 'Photo' mode with automatic exposure and white balance settings to mimic real-world scouting conditions. High Dynamic Range (HDR) and AI scene enhancement features were strictly disabled to preserve raw texture fidelity. Images were saved in RGB color mode, JPEG format, without applying any external post-processing or color grading during the data collection phase. This approach ensures the system's ease of use and eliminates the need for specialized equipment or specific photographic know-how.

The capture target classes were healthy potato leaves, or potato leaves with symptoms of interveinal chlorosis or leaf curling. The data capture was performed by expert phytopathologists in fields located in Volos, Magnesia, Greece for healthy potato plants on fields, and in La Palma, Cartagena, Murcia, Spain for symptomatic potato plants and further healthy leaf captures. Figure 2 exhibits one example capture for each of the target

classes for this work. The data collection comprises a total of 222 images and at a resolution of 3456×3456 .

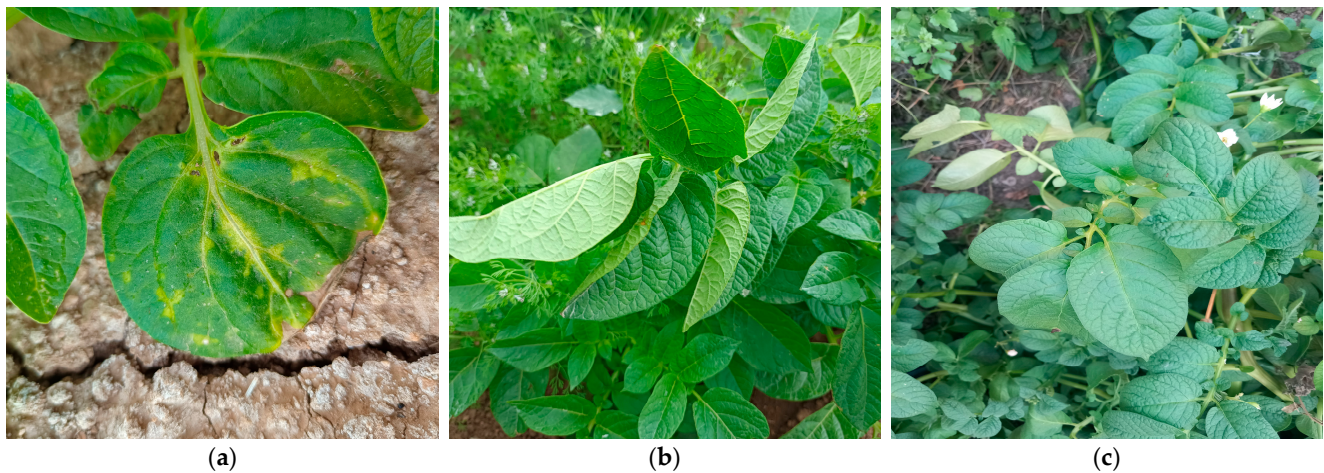


Figure 2. Example dataset images: (a) leaf with interveinal chlorosis symptoms; (b) curled leaves suffering from constant soft wind; (c) healthy potato leaves.

For the data capturing process the aim of each capture was to depict one main leaf where a specific known class would be assigned to it, while the rest of the leaves potentially depicted on the images were assigned one of the other known classes if the characterization was possible or they were assigned as an unknown class in cases where explicit class assignment through the image was not feasible.

2.2. Dataset Annotation and Split

To enhance the model's segmentation capabilities, an initial fine-tuning phase was conducted on the images collected. The objective was to refine the model's generalization of potato leaf morphology, ensuring precise instance masking and effective separation between foreground and background elements.

To achieve a robust representation of abiotic stress features, it was vital to annotate and expand images collected based on four labels, Interveinal Chlorosis, Leaf Curling (Softwind), Healthy and Unknown. The annotation procedure was performed using Roboflow (<https://roboflow.com>) [28], resulting in 8203 leaf annotations. An annotated image sample can be viewed in Figure 3.



Figure 3. Image sample from the dataset: (a) raw image captured in the open field; (b) the annotation result—the red colored masks indicate the annotation overlay to the RGB image.

Approximately 15% of these annotations were classified as “unknown” due to visual ambiguity, necessitating a partial label learning (PLL) classifier to disambiguate this class and generate high-confidence pseudo-labels. The dataset used for the classifier’s training was constructed by extracting instance-level masks from the multi-class symptoms dataset used for Mask2Former training. To explore the classifier’s sensitivity to background noise, annotation masks were cropped from the original images and processed using two different configurations: one with background pixels replaced by black padding and another with the original background pixels intact. The number of annotations in the training and validation splits, for each class, is presented in Table 1.

Table 1. Partially labeled dataset number of annotations for each class and split.

Class	Training	Validation	Total
chlorosis	641	161	802
softwind	656	164	820
healthy	3752	1590	5342
unknown	1239	-	1239

The relatively small number of annotations made the application of standardized dataset augmentation techniques necessary. The augmentations applied consisted of standardized augmentation techniques including horizontal, vertical flips with probability of 25% and rotation limited to 10 degrees and probability of 50% to replicate the different point of views an instance can be captured and teach the model orientation and rotational invariance, adjustments in brightness, contrast, saturation and hue to account for the fluctuations in lighting throughout the day and teach photometrics invariance. A final augmentation addition to bridge the gap between the different image domains introduced by the multiple capturing locations for the images (i.e., Spain and Greece) was the Fourier Domain Adaption (FDA) [29]. The augmentations applied and their corresponding probabilities are presented in Table 2.

Table 2. Augmentations applied with their corresponding arguments.

Augmentation	Value	Probability
Horizontal flip Vertical flip	-	25%
Rotation	$[-10^\circ, 10^\circ]$	50%
Random Scale	$[-10\%, 10\%]$	30%
Perspective	0.0001	20%
Elastic transform	$(\alpha, \sigma) = (50, 5)$	20%
Gaussian Noise or Blur	blur kernel = (3, 3)	20%
Hue shift Saturation shift brightness shift	$[-20, 20]$ $[-30, 30]$ $[-20, 20]$	50%
Fourier Domain Adaption	beta limit $\in [0.02, 0.09]$	80%

After completing 20% of the training, a coarse dropout augmentation was introduced, where black pixel “holes” are created on the instance mask to prevent overfitting and mimic real-world imperfections. This forces the model to learn the entire pattern of the disease rather than relying on its single most obvious symptom. The augmentations mentioned are solely applied to the training set, achieving a great amount of different leaf capturing scenarios and conditions.

Once the PLL classifier assigned pseudo-labels to the unknown instances (all of which were restricted to the training set during the disambiguation phase), the dataset was consolidated into three final classes and divided into training/validation subsets following an image-level grouping approach. To prevent data leakage during the two subsets split for model training, an iterative stratification strategy was employed, testing multiple dataset splits to select a split where each definitive class (Healthy, Chlorosis, and Leaf Curling) in the outcome distribution remained proportional between the training and validation sets with an approximate 75%/25% proportion. This helps ensure that the distribution of instances remained balanced, without breaking the image-level grouping. The resulting number of annotations for each class is shown in Table 3.

Table 3. Pseudo-labeled dataset statistics.

Class	Training Split	Validation Split
chlorosis	857	192
softwind	982	367
healthy	4122	1660

2.3. Evaluation Metrics

The performance of the Partial Labeling Learning classifier was determined using the model's predictive accuracy and F1-Score, which are standard classification benchmarks. Accuracy was calculated as the ratio of instances where the predicted label matched the single true ground-truth label. To account for potential class imbalances within the dataset, the F1-Score was employed as the harmonic mean of precision and recall, providing detailed insight into the classifier's disambiguation capabilities.

To assess the performance of the Mask2Former models employed for the leaf segmentation task, a comprehensive set of evaluation metrics was utilized. The key metrics selected were precision, recall, F1-Score and the mean Average Precision (mAP) at an Intersection over Union (IoU) for different threshold values. Precision indicates the fraction of positive predictions that are true leaf instances, recall measures the fraction of actual leaf instances correctly identified, while F1-Score balances them to a single metric. The mAP is a common benchmark for object detection tasks as it evaluates the accuracy of the predicted masks by counting a prediction as correct only if its overlap with the ground truth mask exceeds the threshold value applied. By analyzing these metrics alongside the parameter count for each model configuration, this study evaluates both the segmentation accuracy and the computational efficiency of the models.

2.4. Partial Label Learning

Partial label learning (PLL) [25] is a weakly supervised multi-class classification framework where each instance is associated with a set of candidate labels, only one of which is the ground truth. This framework is particularly advantageous for real-world problems characterized by inter-class ambiguity. In this context, distinguishing between mild symptoms and healthy leaves can be difficult to categorize definitively as either “diseased” or “healthy” during manual annotation. This method allows the model to learn features from unambiguous instances and subsequently disambiguate the partially labeled data, mitigating the risk of annotation noise. In this work, class labels were turned to multi-hot encoding, following the format shown in Table 4.

Within the PLL framework, unknown labeled instances are assigned to a complete or subset of the candidate set (Equation (1)).

$$S = \{\text{Chlorosis, Softwind, Healthy}\} \quad (1)$$

Table 4. Multi-hot encoding of dataset's labels.

Class	Multi-Hot Encoding
chlorosis	[1, 0, 0]
softwind	[0, 1, 0]
healthy	[0, 0, 1]
unknown	[1, 1, 0]
	[1, 0, 1]
	[0, 1, 1]
	[1, 1, 1]

This formulation states that the true label exists within this subset but remains latent. The core challenge is to train the model to disambiguate the candidate set by identifying the single, true latent label for each 'unknown' sample.

Classifier and Loss for Partial Labeling

The reliability of the label disambiguation process is pivotal for the downstream segmentation task. In order to balance between parameter efficiency and feature extraction capabilities, the EfficientNet-b1 [26] architecture pre-trained on ImageNet [30] is selected. This allows the model to identify subtle symptomatic differences in leaf texture while minimizing computational cost.

The custom training loss had to take into account that the ground truth is latent within the candidate set. Two different loss objectives were implemented: a Progressive Identification (PRODEN) [31] and a Naive Strategy [26]. The fundamental difference between these methods lies in their target distribution: while the Naive strategy computes a fixed uniform probability over the candidate set (treating all candidates equally), PRODEN dynamically weights the candidate labels based on the model's predicted confidence to identify the true latent label.

The goal here is to create a refined target distribution that is restricted to the candidate set and weighted by the model's current belief. To achieve this, the logits derived from the classifier are filtered, setting non-candidate logits to negative infinity ($-\infty$) and the rest to their raw value. The soft pseudo-target probability $q_{\text{soft},i}(j)$ is calculated by applying softmax function to these filtered logits.

The model's final prediction was made by applying a simple Test-Time Augmentation (TTA) [32], which means it receives three different [33] 224×224 patches of the image: the original leaf image and two that were flipped horizontally and vertically. The model generates three different predictions for each patch that are then averaged to create the final prediction. Test-Time Augmentation improves predictor's robustness and reduces variance, since the model might be slightly more sensitive to features in one part of the image. The results of the training process of the EfficientNet-b1 classifier and the different settings applied are presented in Table 5.

Table 5. EfficientNet-b1 average accuracy and F1-scores across all classes for different training settings.

Loss Strategy	Padded with Black Pixels	Accuracy	F1-Score
Naive	Yes	0.9319	0.9318
Naive	No	0.9369	0.9369
PRODEN	Yes	0.9498	0.9253
PRODEN	No	0.9352	0.9361

Different loss strategies and black padding of the masks extracted do not seem to affect the classification metrics dramatically. As shown in Figure 4, PRODEN strategy with black padded masks results in the most balanced performance across all classes while also breaking the 0.94 accuracy barrier for the healthy class.

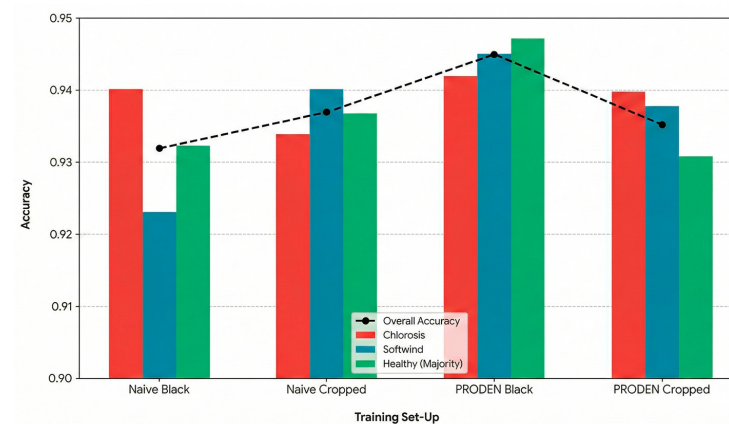


Figure 4. Accuracy scores (Y axis) for each class and each training setting.

Naive strategy with black padded masks delivers great score for the chlorosis class, which is the minority as shown in Table 5, in contrast with the softwind class that scores slightly inefficiently compared to the other settings. Directly cropping the instance masks and feeding them to the model, regardless of loss strategy, performs similarly on average with slight improvement in the chlorosis label score if PRODEN is selected, while the rest of the classes are raised faintly for the Naive strategy.

Based on the showcased analysis, the model used for disambiguation was the EfficientNet-b1 model that was trained with PRODEN loss strategy and was fed with black padded images. The resulting labels were updated in the initial dataset, creating a new multi-class dataset that contained only instances of known classes. An example of this procedure is shown in Figure 5.

A small subset of the disambiguation results was also examined from expert phytopathologist and concluded that possibly up to 20% of the samples could be incorrectly classified comparatively to the 95% accuracy metric that the model returned.

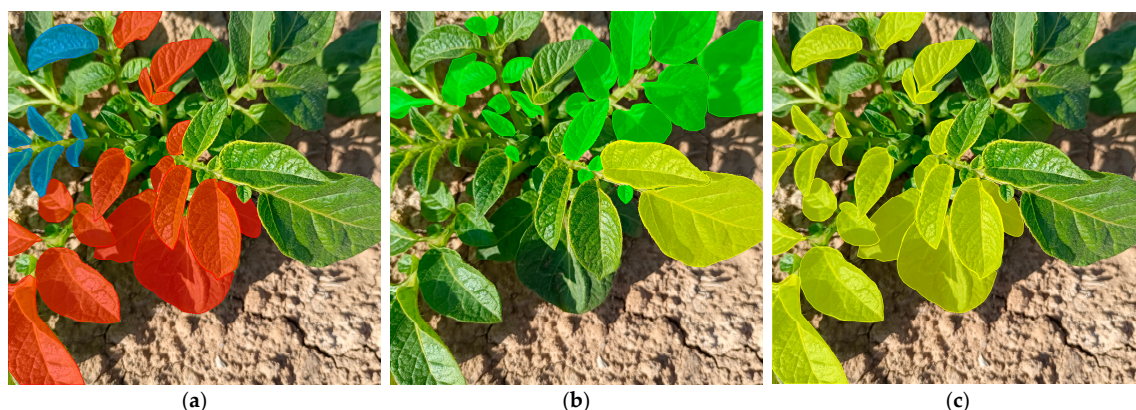


Figure 5. Disambiguation example: (a) Red colored masks are annotated as [0, 1, 1] (either showing chlorosis symptoms or healthy) and blue masks are annotated as [1, 1, 1] (the whole candidate set); (b) instances with unambiguous labels, using green color for [0, 0, 1] (healthy) and yellow for [1, 0, 0] (chlorosis); (c) partially labeled instances are classified by the model, resulting in chlorosis labeled masks. These predictions are valid, since this plant is showing interveinal chlorosis symptoms on the leaves shown in image (b) that are explicitly labeled as “chlorosis”.

2.5. Model Fine-Tuning in the Multi-Class Scenario

The Mask2Former model fine-tuned on the leaf segmentation setting is transferred to solve the instance segmentation task that includes three classes. All instances are labeled explicitly, either manually or pseudo-labeled by the EfficientNet-b1 predictor.

To improve the model's robustness in this scenario, images were augmented by using mosaic method or standard augmentation techniques. Mosaic combines four images into a single, 2×2 grid with probability of 25% and standard augmentations include spatial, geometric, color, noise/blur transforms, with application probabilities as mentioned in Table 2. This ensures the model is trained with a variety of different views of the images, enhancing generalization and accuracy in the feature extraction. Instance segmentation Mask2Former model predicts an instance mask and class for the object and since it was already fine-tuned to detect potato leaves sufficiently, it is essential to apply different learning rates in different parts of the network. To mitigate the model's ability to distinguish between classes without degrading the high-quality features already learned, parameters related to the class predictor layer are using a higher learning rate value than those responsible for the mask prediction. The key hyperparameters used for this setting are presented in Table 6.

Table 6. Key hyperparameters used for the multi-class setting experiment.

Hyperparameter	Value
Train/Evaluation batch size	1
Gradient Accumulation steps	32
Learning Rate (LR) for predictor layer	10^{-4}
LE for none predictor layer	10^{-5}
LR scheduler type	Polynomial
LR scheduler warmup steps	150
Optimizer type	AdamW

3. Results

3.1. Base Segmentation Model Performance Evaluation

The initial stage of the training pipeline focused on developing a robust segmentation foundation by fine-tuning Mask2Former on the potato leaf dataset. Two distinct configurations were evaluated, both utilizing the Swin Transformer [34] backbone: Swin-Tiny and Swin-Small. As shown in Table 7, the Swin-Small architecture without random crop augmentation yielded the highest performance, achieving $mAP_{50} = 0.839$ and $F1-Score_{50} = 0.879$.

Table 7. Mask2Former models evaluated in the initial segmentation task.

Model	Parameters	Random Crop	mAP_{50}	mAP_{75}	$F1-Score_{50}$	$mAR_{50:95}$
mask2former-swin-tiny-coco-instance	47.5M	Yes	0.766	0.660	0.827	0.739
mask2former-swin-tiny-coco-instance	47.5M	No	0.826	0.721	0.877	0.769
mask2former-swin-small-coco-instance	68.8M	Yes	0.778	0.671	0.845	0.754
mask2former-swin-small-coco-instance	68.8M	No	0.839	0.743	0.879	0.773

Random crop augmentation was proved detrimental to the model's performance. For the Swin-Small backbone, mAP_{50} dropped from 0.839 to 0.778 when random cropping was applied. This suggests that maintaining global spatial context is vital for the model to effectively distinguish potato leaf boundaries from the soil background. Consequently, the

base segmentation model which will be transferred to the multi-class instance segmentation problem is the Swin-Small architecture without the use of random crop augmentation.

3.2. Final Multi-Class Segmentation Model Evaluation

Table 8 presents the performance results of Mask2Former model in the abiotic stress detection and classification task on the pseudo-labeled dataset. The evaluation was conducted using three different input image sizes: 600×600 pixels, 800×800 pixels, and 1024×1024 pixels, to determine the effect of image resolution on model performance.

Table 8. Evaluation metrics based on input's size.

Image Size	mAP ₅₀	mAP ₇₅	mAP _{50:95}	F1-Score ₅₀	mAR _{50:95}
(600, 600)	0.813	0.761	0.696	0.830	0.749
(800, 800)	0.889	0.840	0.753	0.89	0.821
(1024, 1024)	0.815	0.761	0.703	0.833	0.763

By comparing the different image sizes, it is clear that the model's performance is optimized using the (800, 800) input size, scoring 88.9% mAP₅₀. Mean average recall scored across all IoU threshold is 82.1% which means the model efficiently detects potato leaves and creates reliable instance masks across different IoU thresholds applied.

A detailed breakdown of performance metrics across target classes at the optimal resolution is provided in Table 9. The Chlorosis class achieved the highest scores (AP₅₀ = 0.939), followed by Softwind (AP₅₀ = 0.901). The Healthy class exhibited the lowest metrics, particularly in AR_{50:95} (0.746), indicating that while the model identifies healthy leaves effectively, it faces higher ambiguity at stricter IoU thresholds.

Table 9. Evaluation metrics for each class.

Class	AP ₅₀	AP ₇₅	AP _{50:95}	F1-Score ₅₀	AR _{50:95}
chlorosis	0.939	0.891	0.820	0.951	0.873
softwind	0.901	0.864	0.779	0.910	0.844
healthy	0.827	0.765	0.661	0.836	0.746
total	0.889	0.840	0.753	0.899	0.821

To complement the results of Table 9, the confusion matrix at an IoU of 0.5 is presented in Figure 6. Ground truth instances are shown row-wise while the predictions made for each class are presented column-wise. The last row indicates that 76 masks were generated where no ground truth instance was present (False Positives) and the last column shows 108 instances were missed during segmentation (False Negatives).

An initial fine-tuning stage on potato leaves' features was conducted to assure the model develops a deep understanding of the morphology and shape of the leaves. Table 10 compares two models: one which followed the two-staged training pipeline proposed against a model that skipped the initial fine-tuning phase. The training settings, as well as the input image size, remained identical for both training procedures. The initial fine-tuning stage resulted in a 9.7% increase in mAP₅₀ and a 9.8% in mAR_{50:95}.

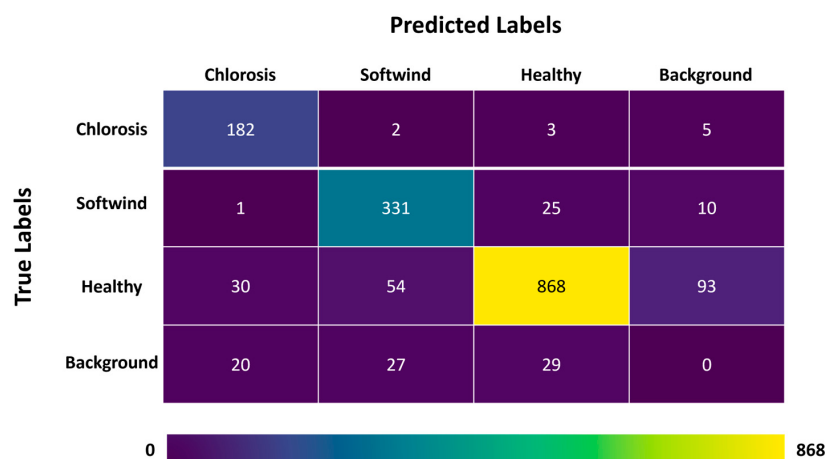


Figure 6. Confusion Matrix at IoU of 0.5. Each row displays the ground truth instances count and each column shows the number of predictions for each class.

Table 10. Comparative results of models utilizing or skipping the initial fine-tuning stage on potato leaf features.

Initial Fine-Tuning Stage	mAP ₅₀	mAP ₇₅	mAP _{50:95}	F1-Score ₅₀	mAR _{50:95}
Yes	0.889	0.840	0.753	0.899	0.821
No	0.792	0.726	0.663	0.815	0.723

Pseudo-labeled instances introduce uncertainty into the system which may downgrade the model's segmentation performance. To investigate this, an additional evaluation stage was performed where all the pseudo-labeled instances of the validation set were not included in the evaluation metrics calculation of the Mask2Former. The outcome of this evaluation procedure, compared to the original evaluation of the model is depicted in Table 11, indicating that the uncertainty derived from the inclusion of pseudo-labeled instances does not negatively affect the segmentation performance of the model, given that the average gap between the two approaches was below 1%.

Table 11. Comparative results of evaluation stages that included pseudo-labeled instances in the validation split of Mask2Former.

Pseudo-Label Instances Included	mAP ₅₀	mAP ₇₅	mAP _{50:95}	F1-Score ₅₀	mAR _{50:95}
Yes	0.889	0.840	0.753	0.894	0.821
No	0.873	0.835	0.768	0.888	0.812
Absolute Difference	0.016	0.005	0.015	0.006	0.009

To further investigate the model's ability to classify and generate instance masks on the potato leaves, a visual inspection of the outcome was also performed. Figure 7 presents three sample images alongside their annotations, predictions and confidence scores. The first row displays raw images from the validation dataset, while the second and third row show the ground truth and predicted masks, respectively. Chlorosis class is represented in the color red, softwind in purple, and the healthy class in blue. In addition, images overlaid with the predictions are marked in green for false positives and pink for false negatives. Predicted masks are filled with the color of the ground truth label and the edge color follows the color of the predicted class. Last row provides a confidence heatmap for each predicted mask, starting from the blue color—for completely uncertain mask—and reaching to yellow—for a fully confident prediction.



Figure 7. Visual assessment of the final model's performance on representative input images from the validation set: (a,d,g,j) healthy potato leaves; (b,e,h,k) plant with chlorosis symptoms; (c,f,i,l) leaves curled due to soft wind stress; (a–c) raw images; (d–f) ground truth masks, colored with distinct colors; (g–i) predictions made, using the color of the predicted class as the edge color; (j–l) confidence scores for each prediction. Color coding: for items (a–i) each color represents a segment overlay of the mask over the RGB image. Blue color indicates healthy class, red color indicates chlorosis class and softwind class is represented in purple. Green and pink colors show false positive and false negative predictions respectively; for items (j–l) the overlay colors depict a heatmap of the confidence scored for each prediction.

As shown in Figure 7, the model successfully segments the potato leaves and classifies them correctly across different abiotic stress scenarios: healthy plants (column 1), chlorosis symptoms (column 2), and wind-induced curling (column 3). Notably, many “false positives” in the healthy leaves image were actually correctly segmented leaves that were excluded from the manual annotation, as they were out of focus or out of the main leaves scene.

4. Discussion

The experimental results indicate that input resolution plays a critical role in the detection of subtle abiotic stress symptoms. While the 600×600 resolution was insufficient for capturing fine textural details, the 800×800 resolution proved to be the optimal trade-off, achieving a balance between feature extraction and computational efficiency. Interestingly, increasing the resolution to 1024×1024 did not yield further improvements and instead introduced higher computational demands, likely due to the model becoming more sensitive to background noise in the open-field setting.

Enhancing performance by initially fine-tuning the model on a binary segmentation task showed important boosting in the evaluation metrics. Segmenting potato leaves in open-field environments is inherently difficult due to complex backgrounds and overlapping foliage. Simplifying the task into simpler individual processes allowed the model to optimize its mask-prediction parameters without the added complexity of class-label noise, which highlights the importance of establishing a strong segmentation foundation before addressing multi-class symptom categorization.

A notable disparity in performance was observed across the three target classes. The Chlorosis class performed exceptionally well ($AP_{50} = 0.939$), most probably due to the sharp contrast of yellowing leaf tissue against the green leaf color that provides a clear signature for the Masked Attention mechanism. The Softwind class also showed strong results ($AP_{50} = 0.901$), as the structural changes, specifically leaf curling, create unique geometric boundaries that the Transformer-based queries can effectively isolate. However, the Healthy class recorded lower metrics ($AP_{50} = 0.827$). This can be attributed to two main factors: inter-class ambiguity and the annotation logic. Early-stage stress symptoms may mimic healthy leaf morphology, since the distinct characteristics are not fully visible, leading to inherent classification difficulty. In addition, as observed in the visual assessment and the confusion matrix of the experiment, many “false positives” were correctly segmented healthy leaves that were simply not part of the manual annotation split because they were in the background or out of focus. This suggests the model’s true precision for healthy leaves is higher than the numerical evaluation indicates.

The high-level performance of the proposed framework can be attributed to the architectural advantages of the Transformer models. Unlike CNN-based methods that focus on local feature extraction, the attention mechanism enables the capturing of global context and long-range dependencies, which proved to be vital for distinguishing abiotic stress symptoms across the leaf surface rather than sharp edges. In addition, the query-based, anchor-free methodology of Mask2Former enables a more flexible segmentation of highly irregular shapes that standard anchor-based models fail to enclose sufficiently. However, these advantages come with significantly higher computational cost and memory footprint, compared to CNN-based models. While the exclusive use of RGB imagery ensures the system’s ease of use, accessibility and economic viability for farmers, it inherently limits the spectral information which could potentially lead to an even better performance.

5. Conclusions

This study examined the detection and prediction of healthy potato leaves and two abiotic stress factors, leaf curling and interveinal chlorosis, using Deep Learning and Machine Learning techniques in the open field setting. The challenge of the instance segmentation task was successfully accomplished by employing a Masked Attention Mask Transformer model of different input image sizes from which the 800×800 input size resulted in the best metrics outcome, achieving a mAP₅₀ of 89%. To achieve this performance, a base instance segmentation model was trained to learn the potato leaves features, which was then transferred to solve the multi-class scenario of the final model presented. In addition, it was vital to disambiguate leaf instances that could not be explicitly labeled to a class and for that reason, an EfficientNet-b1 classifier was trained, scoring an accuracy of 95%.

This 95% accuracy of the PLL model, although very significant, is not clearly reflected on the on-image validation of the expert phytopathologist in a small subset of the dataset, who noted a possible false positive rate of up to 20% for the PLL classified samples. This creates a future examination point where further data collection is applied and further validation of the PLL model results is assessed. Furthermore, future work could involve the detection of more abiotic stress factors to further extend the Mask2Former model's capabilities. Another point for future work to be addressed during real-world operation of this system would be the inclusion of uncertainty estimation techniques and the integration of corrective mechanisms for high-uncertainty samples, such as expert phytopathologist validation, ensemble-base uncertainty modeling or incorporation of additional modalities (e.g., multispectral imagery) to improve robustness.

The use case of this framework is the real-time, automated monitoring of potato crops for the early and precise detection of abiotic stress symptoms related to herbicide and soft wind damage. The proposed system is designed for seamless integration into smart agriculture solutions, overcoming the limitations of traditional monitoring methods that include manual labor and subjective assessment. Deployment settings may include ground level robotics, for high-resolution, close-up detection and targeted treatment, or unmanned aerial vehicles, for rapid, large-scale field surveying and monitoring across vast areas. By providing accurate, early symptom detection, this implementation supports more sustainable and economically viable practices for potato crop management.

Author Contributions: Conceptualization: E.M.P. and M.d.C.M.-B.; methodology: E.S., D.K. and E.M.P.; software: E.S.; validation: D.K.; formal analysis: E.S.; investigation: E.S. and D.K.; resources: E.M.P., N.K. and M.d.C.M.-B.; data curation: E.S. and D.K.; writing—original draft preparation: E.S.; writing—review and editing: D.K. and E.M.P.; visualization: E.S.; supervision: E.M.P.; project administration: E.M.P., N.K. and M.d.C.M.-B.; funding acquisition: E.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: Dataset is available for download at <https://doi.org/10.5281/zenodo.18323649> (accessed on 21 January 2026).

Acknowledgments: This work supported by the E-SPFdigit project, funded by European Union's Horizon Europe research and innovation programme under the grant agreement No. 101157922. We thank Ioannis Naounoulis for his assistance in data collection.

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

References

1. Nehra, A.; Kalwan, G.; Gill, R.; Nehra, K.; Agarwala, N.; Jain, P.K.; Naeem, M.; Tuteja, N.; Pudake, R.N.; Gill, S.S. Chapter 1—Status of impact of abiotic stresses on global agriculture. In *Nanotechnology for Abiotic Stress Tolerance and Management in Crop Plants*; Pudake, R.N., Tripathi, R.M., Gill, S.S., Eds.; Academic Press: Cambridge, MA, USA, 2024; pp. 1–21. [\[CrossRef\]](#)
2. Barbaš, P.; Pietraszko, M.; Pszczółkowski, P.; Skiba, D.; Sawicka, B. Assessing phytotoxic effects of herbicides and their impact on potato cultivars in agricultural and environmental contexts. *Agronomy* **2023**, *14*, 85. [\[CrossRef\]](#)
3. Precheur, R.; Greig, J.; Armbrust, D. The effects of wind and wind-plus-sand on tomato plants. *J. Am. Soc. Hortic. Sci.* **1978**, *103*, 351–355. [\[CrossRef\]](#)
4. ECHO Community. The Many Causes of Leaf Curling. Available online: <http://edn.link/m4frzk> (accessed on 8 December 2025).
5. Brandon Bioscience. Overcoming Abiotic Stress Challenges: Innovative Solutions for Resilient Crops. Available online: <https://brandonbioscience.com/overcoming-abiotic-stress-challenges-innovative-solutions-for-resilient-crops/> (accessed on 26 February 2026).
6. Jia, T.; Smigaj, M.; Kootstra, G.; Kooistra, L. Detection of Diseased Potato Plants with UAV Hyperspectral Imagery. In *Proceedings of the 2024 14th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*; IEEE: Helsinki, Finland, 2024; pp. 1–5. [\[CrossRef\]](#)
7. Lapajne, J.; Susič, N.; Vončina, A.; Gerič Stare, B.; Viaene, N.; Van Beek, J.; Nuytens, D.; Širca, S.; Žibrat, U. Detecting nematodes in potato plants an explainable machine learning approach for detection of potato cyst nematode infections using hyperspectral imaging. *Plant Phenomics* **2025**, *7*, 100127. [\[CrossRef\]](#)
8. Upadhyay, A.; Chandel, N.S.; Singh, K.P.; Chakraborty, S.K.; Nandede, B.M.; Kumar, M.; Subeesh, A.; Upendar, K.; Salem, A.; Elbeltagi, A. Deep learning and computer vision in plant disease detection: A comprehensive review of techniques, models, and trends in precision agriculture. *Artif. Intell. Rev.* **2025**, *58*, 92. [\[CrossRef\]](#)
9. Grinblat, G.L.; Uzal, L.C.; Larese, M.G.; Granitto, P.M. Deep learning for plant identification using vein morphological patterns. *Comput. Electron. Agric.* **2016**, *127*, 418–424. [\[CrossRef\]](#)
10. Aggarwal, R.; Aggarwal, E.; Jain, A.; Aluvalu, R.; Maheswari, U.; Choudhury, T. Deep Learning Image Classification Abiotic Plant Disease Detection Classification Using MobileNet. In *Proceedings of the 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS)*; IEEE: Tashkent, Uzbekistan, 2024; pp. 526–530.
11. Papastratis, I. Comparison of Convolutional Neural Networks and Vision Transformers (ViTs). Medium, Sep. 2023. Available online: <https://medium.com/@iliaspapastratis/comparison-of-convolutional-neural-networks-and-vision-transformers-vits-a8fc5486c5be> (accessed on 11 December 2025).
12. Pereira, G.A.; Hussain, M. A review of transformer-based models for computer vision tasks: Capturing global context and spatial relationships. *arXiv* **2024**, arXiv:2408.15178. [\[CrossRef\]](#)
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [\[CrossRef\]](#)
14. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
15. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the 29th International Conference on Neural Information Processing Systems—Volume 1 (NIPS’15)*; MIT Press: Cambridge, MA, USA, 2015; pp. 91–99.
17. Cheng, B.; Schwing, A.; Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17864–17875.
18. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299.
19. Zhang, A.; Liu, W. Plant Disease Detection Using an Innovative Swin-Axial Transformer. *IEEE Access* **2025**, *13*, 111938–111952. [\[CrossRef\]](#)
20. Darbyshire, M.; Sklar, E.; Parsons, S. Hierarchical mask2former: Panoptic segmentation of crops, weeds and leaves. *arXiv* **2023**, arXiv:2310.06582. [\[CrossRef\]](#)
21. Kirillov, A.; He, K.; Girshick, R.; Rother, C.; Dollár, P. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 16–20 June 2019; pp. 9404–9413.
22. Wei, X.; Zhao, Y.; Lu, X.; Zhang, M.; Du, J.; Guo, X.; Zhao, C. A high-throughput method for monitoring growth of lettuce seedlings in greenhouses based on enhanced Mask2Former. *Comput. Electron. Agric.* **2024**, *227*, 109681. [\[CrossRef\]](#)
23. Christakakis, P.; Giakoumoglou, N.; Kapetas, D.; Tzovaras, D.; Pechlivani, E.-M. Vision Transformers in Optimization of AI-Based Early Detection of *Botrytis cinerea*. *AI* **2024**, *5*, 1301–1323. [\[CrossRef\]](#)

24. Kapetas, D.; Kalogeropoulou, E.; Christakakis, P.; Klaridopoulos, C.; Pechlivani, E.M. Multi-spectral image transformer descriptor classification combined with molecular tools for early detection of tomato grey mould. *Smart Agric. Technol.* **2024**, *9*, 100580. [\[CrossRef\]](#)
25. Kapetas, D.; Kalogeropoulou, E.; Christakakis, P.; Klaridopoulos, C.; Pechlivani, E.M. Comparative Evaluation of AI-Based Multi-Spectral Imaging and PCR-Based Assays for Early Detection of *Botrytis cinerea* Infection on Pepper Plants. *Agriculture* **2025**, *15*, 164. [\[CrossRef\]](#)
26. Cour, T.; Sapp, B.; Taskar, B. Learning from partial labels. *J. Mach. Learn. Res.* **2011**, *12*, 1501–1536.
27. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*; PMLR: Brookline, MA, USA, 2019; pp. 6105–6114.
28. Roboflow (Version 1.0). 2025. Available online: <https://roboflow.com> (accessed on 10 September 2025).
29. Yang, Y.; Soatto, S. FDA: Fourier Domain Adaptation for Semantic Segmentation. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Seattle, WA, USA, 2020; pp. 4084–4094.
30. 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*; IEEE: Miami, FL, USA, 2009; pp. 248–255.
31. Lv, J.; Xu, M.; Feng, L.; Niu, G.; Geng, X.; Sugiyama, M. Progressive identification of true labels for partial-label learning. In *Proceedings of the International Conference on Machine Learning*; PMLR: Brookline, MA, USA, 2020; pp. 6500–6510.
32. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [\[CrossRef\]](#)
33. Shanmugam, D.; Blalock, D.; Balakrishnan, G.; Gutttag, J. Better aggregation in test-time augmentation. In *Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, 10–17 October 2021; pp. 1214–1223.
34. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Montreal, QC, Canada, 2021; pp. 9992–10002. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.