

Article

Reconstructing High-End Soil Sensor Measurements from a Low-Cost 7-in-1 Device in Hass Avocado Orchards Using Random Forest

Andrés Felipe Parra Barragán * , Danny Alexandro Múnera Ramírez  and Natalia Gaviria Gómez

Applied Telecommunications Research Group (GITA), Faculty of Engineering, University of Antioquia, Medellín 050010, Colombia; danny.munera@udea.edu.co (D.A.M.R.); natalia.gaviria@udea.edu.co (N.G.G.)

* Correspondence: andres.parrab@udea.edu.co

Abstract

Soil monitoring is a key component of precision agriculture and environmental sensing systems, where reliable measurements support irrigation management and crop monitoring. Although high-end sensing platforms provide accurate measurements, their cost limits widespread adoption, particularly in resource-constrained agricultural environments. Low-cost soil sensors, such as widely available 7-in-1 probes capable of measuring soil moisture, temperature, electrical conductivity, and pH, offer a scalable alternative for distributed monitoring; however, their limited accuracy raises concerns regarding their reliability for decision-support systems. This study investigates whether measurements from a single low-cost 7-in-1 soil sensor contain sufficient information to reconstruct the outputs of a commercial high-end sensing platform (CropX), specifically volumetric water content (VWC) and pore-water electrical conductivity (EC_{pw}). Field data were collected in a tropical Hass avocado orchard in Colombia, and four machine learning models were evaluated to reconstruct CropX measurements from low-cost sensor signals at three soil depths (20, 41, and 66 cm). Random Forest achieved the highest reconstruction performance, with coefficient of determination R^2 values between 0.9965 and 0.9986 and consistently low root mean square error (RMSE) and mean absolute error (MAE) across depths. Out-of-bag validation and multi-seed stability analyses confirmed the robustness of the models despite the limited dataset size. A chronological validation (80–20%) showed substantially reduced performance, indicating that the proposed approach is more suitable for reconstructing high-end sensor signals under concurrent measurement conditions than for strict temporal extrapolation. Therefore, the framework should be interpreted as a virtual sensing strategy for reconstructing simultaneous CropX measurements from low-cost sensor observations rather than as a standalone model for predicting future soil conditions without periodic recalibration. These results demonstrate that low-cost multi-parameter sensors can support high-fidelity virtual reconstruction of high-end soil measurements, contributing to the development of scalable and cost-effective soil monitoring systems for precision agriculture.



Academic Editor: José Miguel Molina Martínez

Received: 16 March 2026

Revised: 24 April 2026

Accepted: 8 May 2026

Published: 11 July 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.

Keywords: low-cost soil sensors; 7-in-1 soil sensor; Random Forest; virtual sensing; soil moisture; pore-water electrical conductivity (EC_{pw}); precision agriculture; Hass avocado

1. Introduction

Smart agriculture, understood as the application of digital technologies for the efficient management of agricultural resources, has become a central pillar of agro-industrial transformation. This approach integrates artificial intelligence (AI), the Internet of Things (IoT),

Big Data, machine learning (ML), robotics, and cloud computing to automate, monitor, and connect productive system processes. These tools optimize the efficiency of the use of water, energy and fertilizer while increasing crop yield and stability through data-driven management and real-time decision making [1–3]. Smart farming methods could help mitigate the effects of climate change by improving water use efficiency, reducing chemical inputs, and reducing carbon emissions. Moreover, their adoption promotes more sustainable and traceable agricultural operations, strengthening the adaptation capacity to climatic variability [4,5]. In addition to environmental challenges, modern agriculture faces water scarcity, soil degradation, rising food demand, and technological gaps between rural and urban regions. In this context, agricultural digitalization emerges as a comprehensive strategy to produce more with less, ensuring the sustainability and competitiveness of the agro-food sector [6,7].

IoT represents one of the key pillars of smart agriculture by enabling interconnection between sensors, autonomous devices, and integrated management platforms. Several studies highlight their contributions in different dimensions: improvement of crop monitoring through AI and IoT networks [1], integration of renewable energy to improve operational efficiency [2], and implementation of cloud-based platforms to strengthen agricultural management [3]. This infrastructure enables the collection and transmission of real-time data in soil, atmospheric, and crop physiological conditions, allowing automatic adjustments to irrigation, fertilization or ventilation systems [1–3]. Applications such as automated irrigation, greenhouse environmental control, soil nutrient monitoring, and early detection of stress or disease demonstrate the versatility of IoT to improve operational efficiency and reduce management variability [4–6]. This connectivity transforms production processes into more sustainable evidence-based systems, consolidating the role of IoT as the cornerstone of modern agriculture [7].

Despite advances in digitalization, the adoption of sensor-based technologies remains limited among small- and medium-scale farms. The high installation and maintenance costs of commercial systems, together with the requirements for technological infrastructure, restrict their implementation in rural areas with limited resources. As a result, this situation widens the digital divide between large-scale farmers, who can access advanced technological solutions, and those with fewer resources who face economic and operational barriers to adopting smart agriculture tools [7]. In response, recent research has promoted the development of low-cost and energy-efficient sensors designed to balance cost with functionality and reduce dependence on complex infrastructures without compromising environmental sustainability [7,8]. These innovations represent a significant step toward technological democratization, enabling producers of different scales to adopt data-driven monitoring and management tools [8]. However, this accessibility introduces an important challenge related to the accuracy and reliability of measurements. Low-cost sensors represent an accessible alternative for digital agriculture, although their precision and stability are lower than those of more robust sensors [9]. For this reason, several studies have explored strategies to improve the quality of the data obtained from these devices. Various studies have shown that, after calibration processes and predictive modeling, these sensors can approximate the measurements of reference equipment with high reliability [9,10]. This is key to the development of systems capable of adjusting inaccurate measurements through artificial intelligence algorithms, particularly those based on machine learning (ML). In this context, the integration of affordable sensors such as the 7-in-1 sensor, which measures basic edaphic variables, together with predictive tools, can provide results comparable to more robust platforms such as CropX, which measures soil moisture, temperature, and electrical conductivity, and further integrates infiltration, water retention, and evapotranspiration analysis through cloud processing [11,12]. The use of low-cost sensors, complemented by

AI models, therefore represents a promising strategy to reduce technological gaps between high-cost agricultural systems and accessible solutions, which could improve access to smart agriculture [9,10].

Predictive models are essential tools in modern agriculture because they transform sensor and IoT data into actionable information for evidence-based decision-making. These models employ AI algorithms, including ML, that process large volumes of environmental, edaphic, and physiological data to anticipate critical conditions and optimize resource use. The integration of IoT and AI has also enabled agricultural systems capable of monitoring soil and climate variables, predicting irrigation needs, and detecting pest or disease risks, thereby improving productivity and sustainability [13].

Although affordable sensors have expanded access to smart agriculture technologies, their performance still faces important limitations in terms of accuracy and reliability, which restricts their application in critical decision-support contexts [6]. Devices such as 7-in-1 multiparameter sensors simultaneously record edaphic variables including soil moisture, temperature, pH, electrical conductivity, and nutrients. However, despite measuring similar environmental variables, these sensors and robust commercial platforms operate under different measurement principles and data interpretations.

For example, the 7-in-1 sensor reports relative soil moisture and bulk electrical conductivity, whereas platforms such as CropX estimate volumetric water content (VWC) and pore-water electrical conductivity (EC_{pw}), parameters that are more directly associated with water availability and nutrient transport in the soil profile. According to manufacturers' technical specifications, CropX systems additionally incorporate multilayer sensing and cloud-based processing to derive hydrological parameters such as infiltration dynamics, water retention characteristics, and evapotranspiration indicators [11,12]. These differences illustrate the technological gap between low-cost sensing solutions and robust commercial platforms designed for advanced soil monitoring.

Despite these limitations, several studies have demonstrated that the performance of low-cost sensors can be substantially improved through calibration strategies and computational correction techniques. In particular, predictive models based on artificial intelligence (AI) and machine learning (ML) can adjust sensor measurements, reduce systematic errors, and improve their consistency. For example, as reported in [14], integrating sensor observations into AI algorithms significantly improved the estimation of evapotranspiration and the crop water stress index, achieving coefficient of determination values close to 0.99 together with low prediction errors (RMSE $\approx$ 0.15–0.25; MAE $\approx$ 0.10–0.19). These findings suggest that approaches combining affordable sensing devices with AI-based predictive models could represent an effective strategy to reduce the technological gap between low-cost monitoring systems and high-end sensing platforms. By leveraging data-driven correction models, it may be possible to reconstruct or approximate advanced soil measurements using signals obtained from simpler sensors. Nevertheless, limited research has investigated the extent to which AI models can reproduce measurements obtained from advanced soil monitoring systems using only low-cost sensor data. From this gap arises the following research question: How can AI techniques be applied to reconstruct or approximate measurements from advanced sensors, such as CropX, using data obtained from low-cost sensors like the 7-in-1 under concurrent measurement conditions? Therefore, this study evaluates the capacity of machine learning models to approximate measurements from high-end soil sensors using affordable sensing devices, proposing a methodological framework that integrates both technologies within smart agricultural monitoring systems. The main contributions of this study are threefold: (i) the evaluation of machine learning models for reconstructing advanced soil measurements from low-cost sensors, (ii) the com-

parison of multiple predictive algorithms under field conditions, and (iii) the validation of the approach across multiple soil depths in a tropical Hass avocado orchard.

2. Related Work

The integration of smart sensors with Internet of Things (IoT) technologies has enabled continuous monitoring of soil and environmental conditions in precision agriculture. Several studies have focused on the development of distributed sensing architectures based on low-cost sensors to support irrigation and crop management. For example, Placidi et al. [15] designed a LoRaWAN-based wireless sensor network using low-cost sensors to measure soil water content together with environmental monitoring devices. Their system demonstrated the feasibility of modular IoT architectures built with inexpensive components and validated the measurements against a commercial reference sensor. However, the study reported non-constant sensitivity in the low-cost volumetric water content sensor, highlighting the challenges associated with measurement reliability when using inexpensive sensing devices.

Other studies have implemented IoT-based monitoring systems using commercial-grade sensors to ensure higher measurement accuracy. Kumari et al. [16] developed an IoT-enabled monitoring platform integrating TEROs-12 sensors to measure soil moisture, temperature, and electrical conductivity under both laboratory and field fertigation conditions. Their system achieved strong agreement with laboratory reference measurements ($R^2 = 0.999$), demonstrating the reliability of high-end sensing platforms for monitoring soil electrical conductivity and moisture dynamics. Nevertheless, the use of such sensors increases the overall system cost, limiting large-scale deployment in agricultural environments with constrained resources.

To mitigate the accuracy limitations of low-cost sensors, several studies have explored machine learning approaches for sensor calibration and measurement correction. Setiawan et al. [17] compared multiple machine learning models to improve the prediction of volumetric water content obtained from a low-cost capacitive soil moisture sensor. Using gravimetric measurements as ground truth, the authors reported that the K-Nearest Neighbours (KNN) algorithm achieved the best performance with an RMSE of 0.029, outperforming linear regression and decision tree models. These findings demonstrate that machine learning techniques can significantly improve the predictive performance of inexpensive sensing devices.

Recent studies have also investigated optimized ensemble learning strategies for soil variable prediction. Bouhacina et al. [18] proposed a hybrid framework combining tree-based models with a Tri-phase Opposition Adaptive Random Search (TOARS) optimization algorithm to predict soil temperature at shallow depths. Their results showed that metaheuristic optimization can improve the predictive accuracy of Decision Tree, Random Forest, and XGBoost models, achieving R^2 values above 0.999. These findings reinforce the suitability of nonlinear ensemble approaches for modeling complex soil–environment relationships.

Another relevant research direction focuses on cross-sensor calibration and transfer learning to harmonize measurements between heterogeneous sensors. Lin et al. [19] proposed SensorTalk3, an edge-based transfer learning framework designed to recalibrate soil electrical conductivity sensors affected by long-term drift. The approach combines ensemble machine learning models, including Random Forest and Extreme Gradient Boosting, and performs calibration directly on edge devices. Experimental results showed calibration errors as low as 1.7%, with overall accuracies exceeding 98%. These results demonstrate the potential of AI-driven calibration strategies to improve the reliability of low-cost sensors in IoT-based agricultural monitoring systems.

Related approaches have explored the use of machine learning models to estimate difficult-to-measure environmental variables using indirectly related sensor signals. Ouadi et al. [20] proposed a hybrid Gradient Boosting framework optimized using the Salp Swarm Algorithm (GBSSA) to predict the Silt Density Index (SDI) in desalination systems from pressure-based measurements. Their results showed that indirect physical signals can be effectively transformed into accurate estimates of complex operational indicators, highlighting the potential of data-driven models to replace or complement costly measurement procedures.

Deep learning and transfer learning techniques have also been applied to soil monitoring and irrigation management. Burri et al. [21] developed a soil moisture prediction framework using transfer learning based on the ResNet50 architecture combined with IoT sensor data. The model achieved an area under the receiver operating characteristic curve (AUC) of 0.95, demonstrating the ability of deep learning approaches to capture complex relationships between environmental variables and soil moisture dynamics. However, these models focus on predictive irrigation scheduling rather than on reproducing measurements from high-end sensing platforms.

Taken together, previous studies demonstrate significant advances in IoT-based soil monitoring, machine learning-driven sensor calibration, and predictive modeling for irrigation management. However, most existing research focuses either on improving the performance of low-cost sensors through calibration or on developing predictive irrigation models based on environmental data. Limited attention has been given to reconstructing measurements from high-end commercial sensing platforms using only signals from low-cost multiparameter sensors. Furthermore, studies evaluating such approaches under real field conditions in tropical crops remain scarce.

Addressing this gap, the present work investigates whether measurements obtained from a single low-cost 7-in-1 soil sensor can be used to reconstruct key variables measured by a commercial sensing platform (CropX), including volumetric water content and pore-water electrical conductivity, under field conditions in a tropical Hass avocado orchard.

3. Materials and Methods

This section describes the experimental design and methodological framework used to evaluate whether measurements obtained from a low-cost soil sensor can reconstruct the variables provided by a commercial sensing platform. First, the study site and the sensor instrumentation deployed in the field are described. Next, the structure of the dataset and the preprocessing steps applied to synchronize and clean the measurements are presented. Subsequently, the selection of predictor and target variables is justified based on their physical relevance and statistical behavior. Finally, the machine learning models evaluated and the training and validation procedures used to assess predictive performance are detailed.

3.1. Study Site

The study was conducted at the Finca Agrícola Sol de Oriente S.A.S., a commercial avocado (*Persea americana* cv. Hass) production farm located in the rural area “Compañía Abajo” of San Vicente Ferrer, Antioquia, Colombia (6°14′57.9″ N, 75°25′36.6″ W). The farm lies at an altitude ranging from 2165 to 2227 m above sea level, with an average annual temperature of 17 °C. It covers a total area of 20.47 ha, of which 11.41 ha are currently under production, with a planting density of approximately 305 trees per hectare.

The study was conducted in a field covering approximately 0.82 ha, planted in a 5 × 5 m spacing design. This field corresponds to a four-year-old avocado crop, established in 2021.

3.2. Data

The dataset used in this study was obtained from two soil sensors installed on a randomly selected avocado tree. The sensors were buried at three depths (20 cm, 41 cm, and 66 cm) near the tree's root zone, corresponding to the most active layer for root activity. The sensors continuously recorded edaphic parameters throughout the experimental period.

3.2.1. 7-in-1 Soil Sensor (JXBS-3001-NPK-RS)

The 7-in-1 soil sensor (JXBS-3001-NPK-RS, JXCT Electronics Co., Ltd., Weihai, China) is a low-cost multiparameter probe capable of simultaneously measuring seven key soil variables: nitrogen (N, mg/kg), phosphorus (P, mg/kg), potassium (K, mg/kg), electrical conductivity (EC, $\mu\text{S}/\text{cm}$), soil moisture (%), pH, and soil temperature ($^{\circ}\text{C}$). The device operates via an RS485 Modbus communication protocol and is designed for long-term underground installation with IP68 protection. Its measurement ranges and accuracies are summarized in Table 1. The approximate market price of the JXBS-3001-NPK-RS sensor ranges between USD 80 and USD 150 depending on the distributor and configuration, which classifies it as a low-cost soil sensing device compared to commercial precision agriculture systems.

Table 1. Technical specifications of the 7-in-1 Soil Sensor (JXBS-3001-NPK-RS).

Parameter	Range	Accuracy
Nitrogen (mg/kg)	0–1999	$\pm 2\%$ FS
Phosphorus (mg/kg)	0–1999	$\pm 2\%$ FS
Potassium (mg/kg)	0–1999	$\pm 2\%$ FS
Electrical conductivity ($\mu\text{S}/\text{cm}$)	0–10,000	$\pm 10 \mu\text{S}/\text{cm}$
Soil moisture (%)	0–100	$\pm 3\text{--}5\%$
pH	3–9	± 0.3
Temperature ($^{\circ}\text{C}$)	−40 to 80	$\pm 0.5 \text{ }^{\circ}\text{C}$

Note: “FS” denotes full scale. Accuracy given as “FS” refers to the percentage of the sensor maximum range.

3.2.2. CropX Soil Sensor V04

The CropX V04 sensor (CropX Technologies Ltd., Wellington, New Zealand) is a commercial-grade probe designed for high-resolution soil monitoring. It measures volumetric water content (VWC, %), soil temperature ($^{\circ}\text{C}$), and pore water electrical conductivity (EC_{pw}, dS/m).

Each unit includes three sensing points positioned at depths of 20 cm, 41 cm, and 66 cm, implementing proprietary CropX algorithms for data calibration and interpretation. The device transmits data to the CropX cloud platform via a cellular (2G/3G/4G) network and a low-energy Bluetooth interface. The principal technical specifications are summarized in Table 2. In addition to the hardware cost, the CropX system requires access to its proprietary cloud platform for data visualization and analytics. The approximate cost of the CropX sensor unit ranges between USD 900 and USD 1200, with additional subscription fees associated with the platform services. This cost difference highlights the economic gap between commercial sensing platforms and low-cost alternatives.

Table 2. Technical specifications of the CropX Soil Sensor V04.

Parameter	Range	Accuracy
Volumetric water content (VWC, %)	0–60	± 1
Pore water EC (dS/m)	0.05–18	± 0.1
Temperature ($^{\circ}\text{C}$)	−15 to 60	± 1

3.2.3. Sensor Installation and Spatial Configuration

Both sensing systems were installed within the same soil pit to ensure that measurements corresponded to the same pedological environment and root zone conditions. The 7-in-1 sensor array and the CropX integrated probe were positioned at a horizontal separation distance of approximately 30–40 cm between the central axis of the 7-in-1 sensor column and the CropX probe. This spacing minimized physical interference between devices while maintaining both sensors within the same representative soil volume (Figure 1).



Figure 1. Installation of the 7-in-1 sensor array and the CropX multi-depth probe within the same soil pit.

The 7-in-1 configuration consisted of three independent sensors installed at depths of 20, 40, and 60 cm. These sensors were vertically aligned within the same soil column and separated by less than 5 cm, allowing measurement of soil properties along a consistent vertical profile. The CropX probe is an integrated multi-depth sensing device installed vertically in the soil, with sensing elements corresponding to similar depths (20, 41, and 66 cm).

To minimize soil disturbance and ensure adequate soil–sensor contact, installation procedures followed a layer-by-layer backfilling protocol. After excavation to 60 cm depth, the soil pit was saturated with water to reduce air gaps and facilitate soil compaction. The deepest 7-in-1 sensor was installed first, followed by gradual backfilling and manual compaction of the soil to approximate the original bulk density. The same procedure was repeated sequentially for the 40 cm and 20 cm sensors.

The CropX probe was installed according to the manufacturer’s guidelines using a specialized auger to create a narrow borehole. The hole was saturated twice with water prior to insertion to improve soil contact, after which the probe was screwed vertically into the soil profile.

These procedures ensured consistent soil contact conditions for both sensing systems and reduced potential measurement bias associated with installation disturbance.

3.2.4. Comparison of Measured Variables

Although both devices quantify soil electrical and moisture properties, they measure different aspects of these parameters. The 7-in-1 sensor records the bulk electrical conductivity (EC) of the soil, representing the total salinity from both the solid and liquid phases. In contrast, the CropX sensor measures the electrical conductivity of the soil pore water (EC_{pw}), which reflects the dissolved ion concentration in the water available for plant uptake. Thus, EC_{pw} is more directly related to the salinity experienced by roots, while EC provides a broader indication of overall soil salinity. For consistency in the modelling

workflow, CropX ECpw values were converted from dS/m to $\mu\text{S}/\text{cm}$ during preprocessing before model training and evaluation.

This conceptual mismatch between EC and ECpw implies that the model must learn a nonlinear transformation rather than a direct scaling relationship [22].

Regarding water content, the 7-in-1 sensor reports relative humidity (%RH) within the soil matrix, which is an indirect measure influenced by temperature and vapor equilibrium conditions. Conversely, the CropX sensor estimates the volumetric water content (VWC, %), a direct measurement of the volume of water relative to the total soil volume. Therefore, VWC is more representative of the water available to plants, whereas %RH provides a relative indicator of soil wetness. These conceptual differences are essential for interpreting and aligning data between the two devices.

3.2.5. Predictor and Target Variable Selection

Following an initial exploration of the data collected by both sensors, the following variables were selected as inputs for the predictive modeling:

- Soil moisture, soil temperature, electrical conductivity, and pH measured by the 7-in-1 sensor at three depths (20 cm, 40 cm, and 60 cm).

The target variables to be predicted were:

- Volumetric water content (VWC) and pore water electrical conductivity (ECpw) measured by the CropX sensor at depths of 20 cm, 41 cm, and 66 cm.

Soil moisture, pH, temperature, and electrical conductivity measured by the 7-in-1 sensor at depths of 20, 40, and 60 cm were selected as predictor variables due to their direct physical relationship with soil water dynamics and salinity processes. Soil moisture reflects water availability within the soil matrix, temperature influences evaporation and water redistribution, electrical conductivity is related to the concentration of dissolved ions and solutes, and pH represents a complementary chemical property associated with soil solution conditions. Together, these variables represent key drivers of the hydrological and physicochemical processes governing volumetric water content (VWC) and pore water electrical conductivity (ECpw).

An exploratory statistical analysis confirmed that these variables also exhibited the strongest correlations and temporal coherence with the target variables measured by the CropX sensor. Including measurements from multiple depths further allowed the models to capture vertical gradients in soil water movement and solute transport within the root zone.

Exploratory correlation analysis revealed that the N, P, and K measurements were strongly correlated with soil electrical conductivity, particularly at the deeper soil layers. For example, correlations above 0.9 were observed between bottom-layer nutrient measurements and electrical conductivity. This indicates that these variables contain largely redundant information, since soil electrical conductivity reflects the concentration of dissolved ions present in the soil solution. Including NPK variables would therefore introduce multicollinearity without substantially improving predictive performance. Consequently, these chemical variables were excluded from the final predictor set to improve model interpretability and reduce redundancy.

Unlike N, P, and K, soil pH was retained in the predictor set because it represents a distinct chemical property of the soil rather than a direct proxy of ionic concentration. Exploratory correlation analysis showed that pH maintained relevant associations with several target variables, suggesting that it provides complementary information for the reconstruction task.

3.2.6. Data Exploration, Cleaning, and Preprocessing

The dataset was generated from two independent sensing platforms operating with different native sampling frequencies and communication protocols. The CropX sensor records measurements at fixed intervals determined by its cloud-based monitoring system, while the 7-in-1 sensor produces readings according to the acquisition frequency configured in the local data logger. As a result, measurements from both sensors were not perfectly synchronized in time. To address this issue and create a consistent dataset for modeling, a temporal alignment and aggregation procedure was implemented.

First, invalid records, duplicated entries, and non-numeric textual outputs (e.g., sensor failure flags) were identified and removed. For the 7-in-1 sensor, rows were discarded only when missing or invalid values affected the core predictor variables used for modeling, while auxiliary variables not used in the final model were retained whenever possible for exploratory analysis. This step ensured that only complete and reliable observations were retained for subsequent analyses.

Next, a temporal rounding procedure was applied at 30 min intervals, whereby each timestamp was adjusted to the nearest 0- or 30 min mark. This standardization reduced inherent temporal noise caused by irregular sampling, homogenized the sampling resolution between sensors, and allowed for precise synchronization of measurements from different sources. The choice of a 30 min interval reflects the native sampling frequency of the CropX sensor and the need to capture meaningful variations in soil moisture, temperature, and electrical conductivity without compromising the natural temporal variability of the system.

During the preprocessing stage, an additional inspection of extreme values was performed to detect potential sensor anomalies. Two observations presenting abnormally high electrical conductivity values at the deepest measurement depth (66 cm) were identified. These values exceeded the typical operating range observed during the monitoring period and were interpreted as sensor artifacts. Consequently, these two records were removed prior to model training to prevent distortion of the learning process.

Additionally, CropX EC_{pw} values, originally recorded in dS/m, were converted to $\mu\text{S}/\text{cm}$ during preprocessing by multiplying by 1000. This transformation was applied before model fitting; therefore, all EC_{pw} ranges, predictions, RMSE values, and MAE values reported in Section 4 are expressed in $\mu\text{S}/\text{cm}$.

Finally, after temporal rounding, multiple measurements falling within the same interval were aggregated using the arithmetic mean, producing a representative value for each time point and preparing the dataset for comparative analyses and subsequent modeling stages. After this aggregation step, the two sensor datasets were merged using their synchronized timestamps, retaining only the observations available in both platforms.

After applying the cleaning, synchronization, and merging procedures, the final dataset contained 10,490 paired observations and 35 variables. The monitoring period extended from 21 February 2025 to 17 October 2025, covering approximately eight months of continuous soil measurements during the avocado production cycle. Each record in the final dataset therefore corresponds to temporally aligned observations from both the 7-in-1 and CropX sensing platforms.

3.2.7. Correlation Analysis Procedure

Before training the machine learning models, an exploratory correlation analysis was conducted to examine the statistical relationships between the variables measured by both sensing platforms. Pearson correlation coefficients were computed between all predictor variables from the 7-in-1 sensor and the target variables measured by the CropX sensor.

This baseline analysis served two purposes. First, it provided an initial estimate of the linear relationships between the sensors and helped identify the most relevant predictors. Second, it established a reference benchmark for evaluating the improvement achieved by the machine learning models over simple statistical associations.

Additionally, correlations among the predictor variables themselves were analyzed to identify potential multicollinearity effects. This analysis confirmed strong correlations between soil nutrient measurements (NPK) and soil electrical conductivity, supporting the decision to exclude the nutrient variables from the final modeling dataset.

These correlation analyses were used as a preliminary statistical filter to support the selection of predictors and to identify potentially redundant variables prior to machine learning model training.

3.3. Model

3.3.1. Model Inputs and Outputs

The predictive models were designed to reconstruct measurements obtained from the CropX sensing platform using signals from the 7-in-1 soil sensor.

Model inputs (predictors):

- Soil moisture (%)
- Soil temperature (°C)
- Electrical conductivity ($\mu\text{S}/\text{cm}$)
- pH (pH).

These variables were measured by the 7-in-1 sensor at depths of 20 cm, 40 cm, and 60 cm.

Model outputs (targets):

- Volumetric water content (VWC, %)
- Pore water electrical conductivity (EC_{pw}, $\mu\text{S}/\text{cm}$)

These variables were measured by the CropX sensor at depths of 20 cm, 41 cm, and 66 cm.

3.3.2. Predictive Models Evaluated

To estimate the target variables obtained from the CropX sensor volumetric water content (VWC) and pore water electrical conductivity (EC_{pw}) at different depths, a set of supervised learning models from distinct methodological families was implemented. These algorithms were selected due to their documented use in agronomic and environmental prediction tasks and their ability to capture nonlinear relationships, handle edaphic variability, and tolerate the instrumental noise characteristic of soil sensing systems. Several studies have demonstrated that these models can effectively process multivariate sensor data and represent complex hydrological and physicochemical interactions in agricultural environments [13,23,24].

The first model considered was Multiple Linear Regression, a classical statistical approach widely used for modeling relationships between predictor and response variables [25]. Although its ability to represent nonlinear interactions among soil moisture, temperature, and electrical conductivity is limited, its interpretability facilitates the identification of general trends and physical relationships between the variables measured by the 7-in-1 sensor and those obtained from CropX.

The second model evaluated was Random Forest, an ensemble of decision trees built through bootstrap aggregation. Its capacity to model complex interactions, robustness to noise, and stable performance with heterogeneous datasets make it particularly suitable for agronomic applications, where soils often exhibit substantial vertical and temporal

variability. Previous studies highlight its strong performance in agricultural and IoT-sensor contexts due to its tolerance to imperfect or noisy data [24]. In this study, the Random Forest model was implemented with 100 decision trees using bootstrap aggregation and out-of-bag (OOB) estimation to internally assess model generalization. The algorithm was trained using parallel processing to improve computational efficiency. Because bootstrap sampling was enabled, out-of-bag (OOB) estimation was also used as an internal validation mechanism to further assess the generalization capacity of the model. Additionally, intrinsic feature importance values from the Random Forest model were extracted to provide preliminary insight into the relative contribution of predictor variables to the reconstruction task.

The third model included was XGBoost, a gradient boosting algorithm recognized for its high predictive accuracy in tasks involving complex and nonlinear relationships. Its iterative optimization of residual errors and incorporated regularization allow it to capture high-order interactions among moisture, temperature, and soil electrical conductivity. XGBoost has been widely applied in agronomic prediction and remote sensing analysis, consistently demonstrating strong performance with multivariate sensor datasets [23].

Finally, a Multilayer Perceptron (MLP) Regressor was implemented. This feed-forward neural network is capable of approximating highly nonlinear functions. Standardizing the inputs prior to training improves model stability, particularly when combining variables with different scales such as temperature, moisture, and electrical conductivity. MLPs have been successfully applied in digital agriculture to model multivariate sensor time series, performing well in scenarios characterized by nonlinear interactions and complex temporal dynamics [13].

To ensure reproducibility and a fair comparison across model families, all algorithms were implemented using consistent baseline hyperparameter configurations (Table 3). The objective of this comparison was not exhaustive hyperparameter optimization, but rather to evaluate the relative predictive capacity of representative model classes under homogeneous conditions. All models employed `random_state = 42` to ensure reproducibility. In particular, XGBoost was not subjected to extensive hyperparameter tuning, as the focus of this study is on virtual sensing feasibility rather than maximizing model-specific performance.

Table 3. Main hyperparameters used for the evaluated machine learning models.

Model	Main Hyperparameters
Linear Regression	default scikit-learn implementation
Random Forest	<code>n_estimators = 100</code> , <code>oob_score = True</code>
XGBoost	<code>n_estimators = 100</code> , <code>learning_rate = 0.1</code> , <code>max_depth = 3</code>
MLPRegressor	<code>hidden_layer_sizes = (50,30)</code> , <code>max_iter = 2000</code> , <code>early_stopping = True</code>
Preprocessing	StandardScaler applied before MLP

3.3.3. Training and Evaluation Procedure

To ensure a fair and replicable comparison among models, a standardized training and evaluation workflow was employed. The final dataset was randomly divided into training (80%) and test (20%) subsets using a fixed random seed to ensure reproducibility. The training subset was used for model fitting and internal validation, whereas the test subset was reserved exclusively for evaluating final generalization performance.

Within the training set, a 5-fold cross-validation scheme was applied to obtain robust performance estimates and assess model stability across different partitions of the data. After cross-validation, each model was refitted on the full training set and evaluated on the held-out test set.

The evaluation relied on three complementary performance metrics: the coefficient of determination (R^2), the root mean squared error (RMSE), and the mean absolute error (MAE). RMSE and MAE quantify the magnitude of prediction errors, while R^2 reflects the proportion of variance in the target variables explained by the model.

Model comparison was based on the balance between predictive accuracy, stability across folds, and agronomic coherence in the relationships captured between the 7-in-1 predictors and the CropX target variables. This strategy ensured that the selected model not only performed well statistically but also remained consistent with the physical and hydrological processes governing soil behavior.

For temporal generalization assessment, the dataset was sorted chronologically, using the earliest 80% of observations for training and the most recent 20% for testing, enabling evaluation under future unseen temporal conditions. This approach avoids information leakage caused by temporal autocorrelation, ensuring that training and testing data correspond to independent temporal segments.

3.3.4. Performance Evaluation Metrics

The performance of the predictive models was assessed using three widely adopted metrics in agricultural and environmental machine learning: the coefficient of determination (R^2), the root mean squared error (RMSE), and the mean absolute error (MAE). These metrics are standard in studies involving soil moisture prediction, crop yield modeling, and digital soil assessment due to their robustness and their ability to quantify different aspects of the discrepancy between observed and predicted values [26–30].

Coefficient of Determination (R^2). Expresses the proportion of the total variance in the target variable explained by the model. It is defined as:

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

where y_i represents the observed value, $\hat{y}_i$ the predicted value, and $\bar{y}$ the mean of the observed values. In agronomic modeling, R^2 is commonly used because it quantifies how well the model captures the underlying soil and crop-related dynamics [26,27].

Root Mean Squared Error (RMSE). It measures the average magnitude of the prediction error and penalizes larger deviations more strongly due to the squared term. It is defined as:

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

RMSE captures absolute predictive accuracy in the same units as the target variable and is particularly relevant for soil sensors, where abrupt changes in moisture or EC may indicate meaningful edaphic events. It is one of the most frequently used metrics in agricultural machine learning and environmental time-series modeling [26,27].

Mean Absolute Error (MAE). Quantifies the average absolute difference between predictions and observations. It is defined as:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|.$$

MAE is more robust than RMSE to outliers and provides a clear interpretation of the typical deviation between predicted and true values. This makes it particularly useful in field sensor studies with noisy measurements [30].

The combined use of R^2 , RMSE, and MAE is standard in evaluating machine learning models applied to agricultural and environmental prediction, capturing model fit, error magnitude, and robustness across algorithms [26–29].

All data preprocessing, correlation analysis, model training, and evaluation steps were implemented in Python 3.11.9 using standard scientific computing libraries including pandas 3.0.1, NumPy 2.4.3, and scikit-learn 1.8.0. The complete codebase used in this study, including scripts for dataset construction, exploratory analysis, correlation computation, model training, and figure generation, is maintained in a GitHub repository. Due to institutional data restrictions, the repository is currently private but can be made available for academic and research purposes upon reasonable request. The repository is hosted on GitHub: https://github.com/AndresParra11/7en1_cropx (accessed on 20 April 2026).

4. Results

This section presents the main findings obtained from the statistical exploration and machine learning reconstruction of CropX measurements using signals from the 7-in-1 soil sensor. First, the linear relationships between predictors and target variables are examined through correlation analysis. Next, the comparative performance of the evaluated regression models is reported. Finally, the predictive behavior, residual structure, and robustness of the best-performing model are analyzed.

4.1. Exploratory Correlation Analysis

Pearson correlation analysis revealed clear depth-dependent relationships between the variables measured by the 7-in-1 sensor and the CropX targets (Figure 2). Among the selected predictors, the strongest positive association was observed between bottom-layer soil moisture and CropX moisture at 66 cm ($r = 0.92$), confirming a strong linear correspondence at the deepest soil layer.

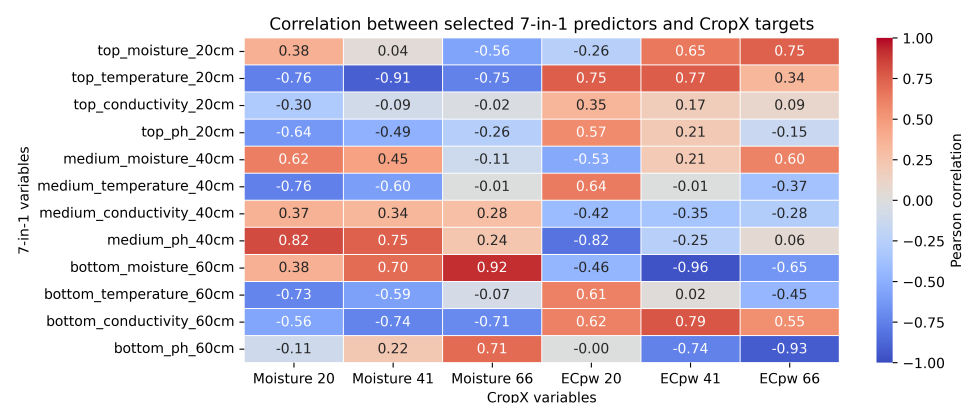


Figure 2. Pearson correlation coefficients between the selected 7-in-1 predictors and the CropX target variables.

Relevant correlations were also observed for soil pH. In particular, medium-layer pH (40 cm) showed strong positive correlations with Moisture at 20 cm ($r = 0.82$) and moisture at 41 cm ($r = 0.75$). In addition, bottom-layer pH (60 cm) exhibited a strong association with ECpw at 66 cm ($r = 0.93$). These relationships suggest that soil pH provides complementary information related to the chemical conditions of the soil solution, which can influence both water dynamics and ion mobility within the soil profile.

The analysis of excluded nutrient variables further revealed that N, P, and K measurements were strongly correlated with soil electrical conductivity, especially at matching depths (Figure 3). Bottom-layer nutrient measurements presented correlations above 0.91 with bottom electrical conductivity, indicating substantial redundancy between these vari-

ables. These results support the methodological decision to exclude NPK variables from the final predictor set while retaining pH as an informative and non-redundant chemical variable.

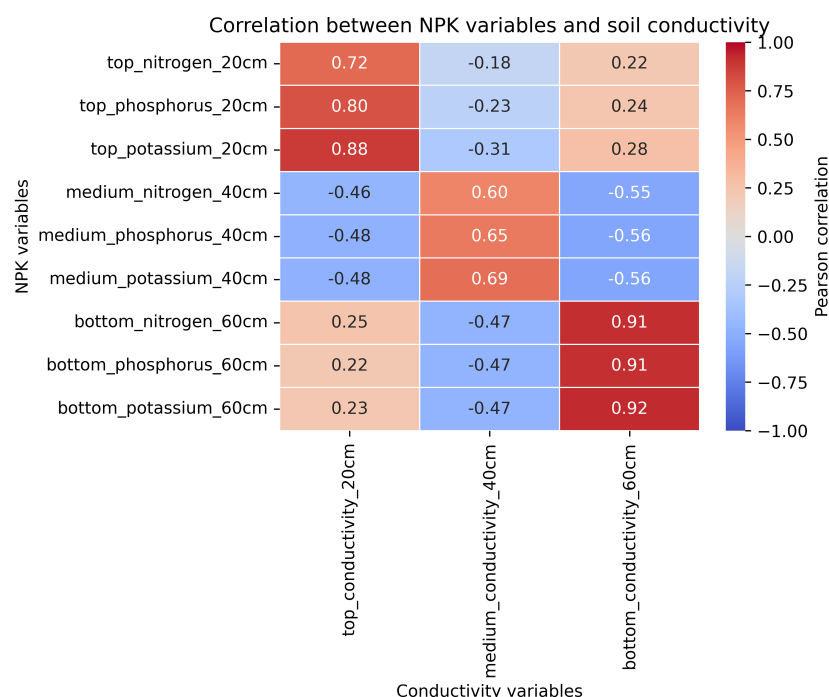


Figure 3. Pearson correlation coefficients between NPK variables and soil electrical conductivity at the three evaluated depths.

4.2. Comparative Performance of Predictive Models

Four regression algorithms were evaluated for reconstructing CropX measurements: Multiple Linear Regression, Random Forest, XGBoost, and a Multilayer Perceptron (MLP) regressor. Model performance was assessed using the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). The overall comparison of model performance across all variables and depths is summarized in Table 4.

Across all evaluated targets, Random Forest consistently achieved the highest predictive accuracy, combining the highest R^2 values with the lowest RMSE and MAE. For ECpw at 20 cm, Random Forest reached an R^2 of 0.9974, with RMSE = 59.7292 $\mu\text{S}/\text{cm}$ and MAE = 11.6511 $\mu\text{S}/\text{cm}$, outperforming XGBoost ($R^2 = 0.9931$), the MLP regressor ($R^2 = 0.9810$), and linear regression ($R^2 = 0.8511$). Similar performance differences were observed at 41 and 66 cm.

For moisture prediction, Random Forest again achieved the best performance. At 20 cm, the model obtained $R^2 = 0.9990$, with RMSE = 0.0870 and MAE = 0.0267. The remaining models showed lower predictive accuracy, particularly linear regression, whose R^2 values ranged between 0.8865 and 0.9281 depending on depth.

These results indicate that the relationships between the 7-in-1 sensor variables (soil moisture, temperature, electrical conductivity, and pH measured at multiple depths) and the CropX variables (VWC and ECpw) are strongly nonlinear. Ensemble tree-based models, particularly Random Forest, are better able to capture these interactions than linear models or single nonlinear learners.

Table 4. Comparative performance of the evaluated machine learning models for moisture and ECpw at different depths.

Target	Model	R^2	RMSE	MAE
ECpw 20 cm	Random Forest	0.9974	59.7292	11.6511
	XGBoost	0.9931	98.1041	49.5195
	MLPRegressor	0.9810	163.0899	97.6081
	Linear Regression	0.8511	456.3335	330.3436
ECpw 41 cm	Random Forest	0.9950	77.9966	19.3922
	XGBoost	0.9922	97.4155	48.2055
	MLPRegressor	0.9875	123.4153	76.0642
	Linear Regression	0.9638	210.2702	165.8347
ECpw 66 cm	Random Forest	0.9986	13.9292	3.7861
	XGBoost	0.9960	23.1591	9.6820
	MLPRegressor	0.9891	38.2817	18.0405
	Linear Regression	0.9006	115.4577	81.8562
Moisture 20 cm	Random Forest	0.9990	0.0870	0.0267
	XGBoost	0.9948	0.1983	0.1182
	MLPRegressor	0.9869	0.3155	0.1700
	Linear Regression	0.8865	0.9287	0.6882
Moisture 41 cm	Random Forest	0.9976	0.0755	0.0220
	XGBoost	0.9939	0.1190	0.0749
	MLPRegressor	0.9845	0.1903	0.1151
	Linear Regression	0.9281	0.4095	0.3246
Moisture 66 cm	Random Forest	0.9971	0.2396	0.0319
	XGBoost	0.9962	0.2730	0.0992
	MLPRegressor	0.9970	0.2429	0.1371
	Linear Regression	0.9000	1.4032	1.0172

4.3. Performance of the Random Forest Model by Variable and Depth

Given its superior performance across all targets, Random Forest was selected as the primary model for detailed analysis. Table 5 summarizes the performance of Random Forest for each variable and depth.

Table 5. Performance of the Random Forest model for moisture and ECpw at different depths.

Target	MAE	RMSE	R^2
Moisture 20 cm	0.0267	0.0870	0.9990
Moisture 41 cm	0.0220	0.0755	0.9976
Moisture 66 cm	0.0319	0.2396	0.9971
ECpw 20 cm	11.6511	59.7292	0.9974
ECpw 41 cm	19.3922	77.9966	0.9950
ECpw 66 cm	3.7861	13.9292	0.9986

For moisture prediction, R^2 values ranged between 0.9971 and 0.9990, with RMSE values between 0.0755 and 0.2396. The lowest prediction errors were observed at 41 cm, where RMSE reached 0.0755 and MAE was 0.0220.

For ECpw, Random Forest achieved R^2 values between 0.9950 and 0.9986. The highest predictive accuracy was obtained at 66 cm ($R^2 = 0.9986$), where RMSE and MAE reached 13.9292 and 3.7861 $\mu\text{S}/\text{cm}$, respectively.

These results indicate that Random Forest can accurately reconstruct CropX sensor measurements across the full soil profile using only signals from the low-cost 7-in-1 sensor. However, these results should be interpreted with caution, as temporal validation indicates

that such performance is not maintained under chronological data separation, highlighting the influence of temporal autocorrelation in the dataset.

4.4. Feature Importance Analysis

To provide additional insight into model behavior and improve model interpretability, intrinsic feature importance values were computed from the Random Forest models. Figure 4 resents the mean importance values averaged across all predicted targets (soil moisture and ECpw at 20, 41, and 66 cm depths).

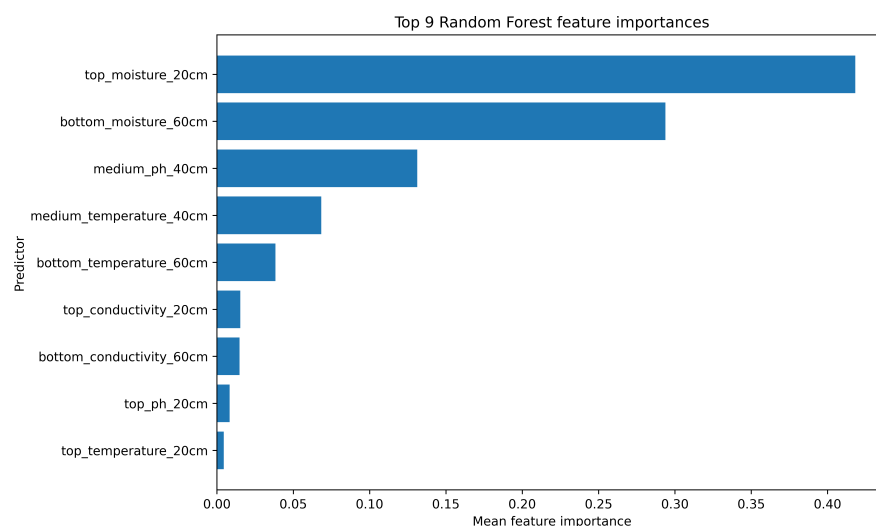


Figure 4. Mean intrinsic feature importance analysis.

The results indicate that moisture-related predictors dominate the importance ranking, particularly top_moisture_20cm and bottom_moisture_60cm. Intermediate soil pH and temperature variables also showed moderate contributions, while conductivity predictors exhibited smaller but consistent influence.

Overall, the importance distribution indicates that the model primarily relies on variables directly related to soil water status, with secondary contributions from physico-chemical soil properties.

4.5. Performance Under Time-Based Validation

Under the temporal split (80% training, 20% testing), predictive performance decreased compared to the random split, particularly for ECpw variables (Table 6). ECpw 41 retained moderate predictive capability ($R^2 = 0.246$), while the remaining variables exhibited negative R^2 values, indicating limited extrapolation ability under future temporal conditions.

Table 6. Comparison of R^2 between random and temporal validation strategies.

Variable	R^2_{random}	$R^2_{temporal}$
Moisture 20	0.9990	−4.2631
Moisture 41	0.9976	−3.3748
Moisture 66	0.9971	−0.4612
ECpw 20	0.9974	−4.6944
ECpw 41	0.9950	0.2463
ECpw 66	0.9986	−2.3081

4.6. Graphical Assessment of Model Predictions

The predictive performance of the Random Forest model was further examined through graphical diagnostics (Figures 5 and 6). Scatter plots of observed versus predicted values show a strong concentration of points along the 1:1 line, indicating close agreement between measured and reconstructed values under random partitioning. However, this apparent agreement should be interpreted in light of the temporal dependence structure of the dataset, as adjacent observations are not independent and may contribute to optimistic performance estimates under random splits.

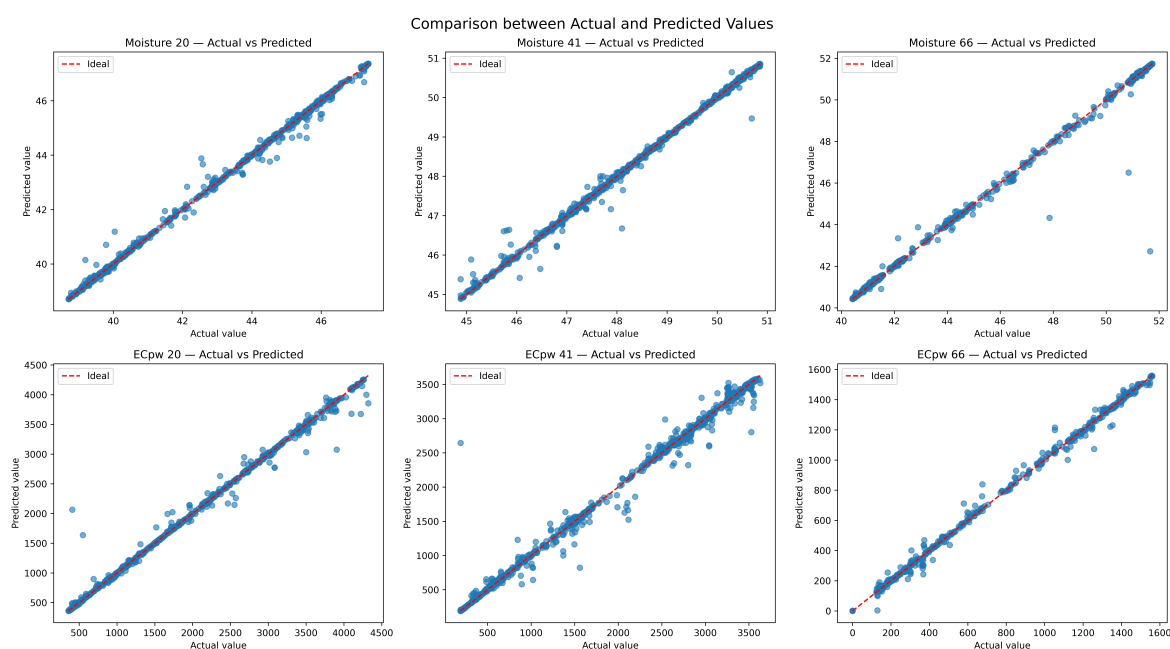


Figure 5. Observed versus predicted values for moisture and ECpw across the three evaluated depths.

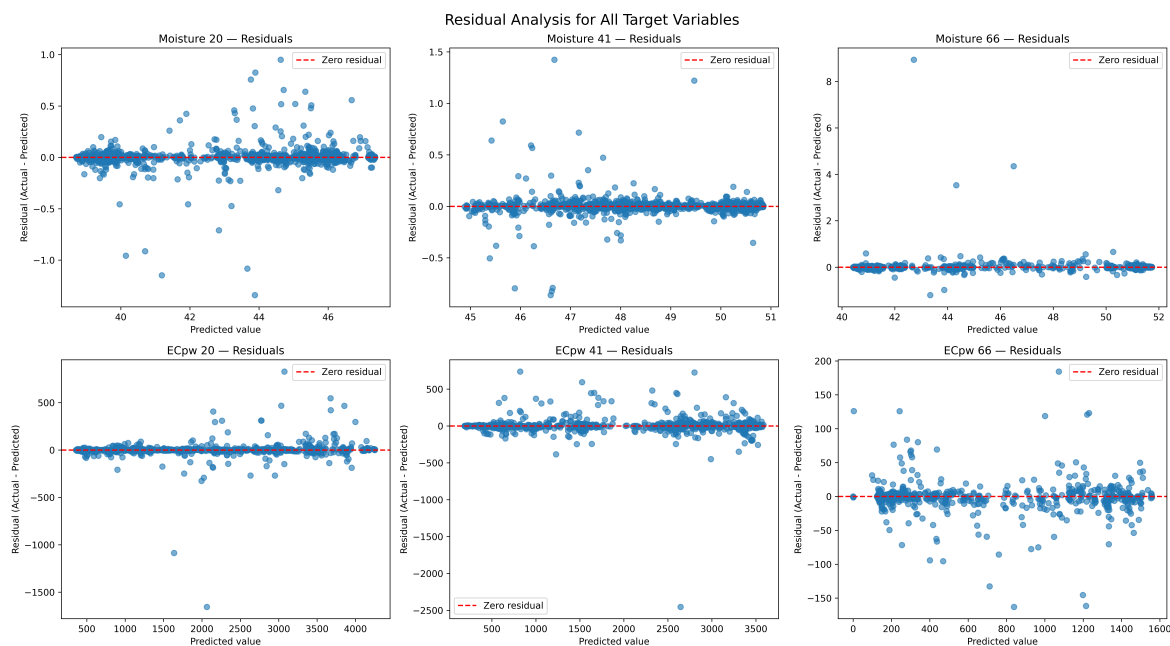


Figure 6. Residual plots for Random Forest predictions across all variables and depths.

Residual plots revealed distributions centered around zero, with no noticeable systematic patterns or heteroscedasticity. These results indicate that the remaining prediction errors are small and randomly distributed. Therefore, although graphical diagnostics

suggest strong predictive performance, these results primarily reflect the model's ability to reconstruct concurrent measurements rather than its capacity to generalize to future unseen temporal conditions.

4.7. Internal Validation and Robustness Analysis

The out-of-bag (OOB) performance estimates obtained during Random Forest training were consistently high for all targets (Table 7). OOB scores ranged from 0.9980 to 0.9991, confirming that the model maintains excellent predictive accuracy when evaluated using internal bootstrap samples.

Table 7. Out-of-bag (OOB) scores of the Random Forest model for moisture and ECpw at different depths.

Target	OOB Score
Moisture 20 cm	0.9989
Moisture 41 cm	0.9988
Moisture 66 cm	0.9991
ECpw 20 cm	0.9988
ECpw 41 cm	0.9980
ECpw 66 cm	0.9984

Durbin–Watson statistics were close to 2.0 for all targets (ranging from 1.95 to 2.05), indicating an absence of significant autocorrelation in the residuals.

To further evaluate the stability of the model, six independent Random Forest models were trained using different random seeds. The resulting performance metrics exhibited extremely low variability (Table 8), with standard deviations of R^2 on the order of 10^{-4} .

Table 8. Sensitivity analysis of the Random Forest model to random seed initialization.

Variable	R^2_{mean}	R^2_{std}	RMSE _{mean}	RMSE _{std}	MAE _{mean}	MAE _{std}
Moisture 20 cm	0.9990	0.00003	0.0867	0.0014	0.0262	0.0003
Moisture 41 cm	0.9978	0.00010	0.0725	0.0015	0.0215	0.0003
Moisture 66 cm	0.9971	0.00009	0.2399	0.0035	0.0319	0.0002
ECpw 20 cm	0.9974	0.00008	60.5772	0.9582	11.6121	0.0843
ECpw 41 cm	0.9950	0.00007	78.3128	0.5256	19.4480	0.0739
ECpw 66 cm	0.9986	0.00004	13.8038	0.1896	3.7447	0.0410

These results demonstrate that the proposed modeling approach is highly stable and robust with respect to stochastic initialization and internal sampling procedures.

5. Discussion

The results of this study demonstrate that low-cost soil sensors, when combined with appropriate machine learning techniques, can reproduce the behavior of high-end probes such as CropX with high accuracy. Across all depths, the Random Forest model achieved R^2 values above 0.995 with low RMSE and MAE, placing its performance at the upper range of values reported in the soil monitoring and virtual sensing literature. This effect explains, in part, the very high performance observed under random partitioning, as the model benefits from temporal proximity between training and testing samples.

However, temporal validation revealed negative R^2 values for most variables, indicating that predictive performance decreases substantially when models are evaluated under chronological data separation. This behavior suggests that part of the predictive relationships learned under random splits is influenced by temporal autocorrelation structures

present in the dataset. In particular, adjacent time observations are not independent due to the gradual evolution of soil moisture and ionic dynamics, which can lead to overly optimistic performance estimates when random partitioning is applied.

The exploratory correlation analysis provides important insights into the physical relationships underlying the predictive models. Strong correlations were observed between moisture variables measured by the 7-in-1 sensor and CropX volumetric water content at matching depths, particularly in the deeper soil layers. In addition, soil pH showed relevant associations with both moisture and EC_{pw}, suggesting that chemical soil conditions contribute complementary information that helps explain variations in water and ion dynamics within the soil profile. By contrast, NPK variables exhibited extremely high correlations with electrical conductivity at the same depths, indicating strong redundancy between these measurements. This supports the methodological decision to exclude NPK variables from the predictive models while retaining pH as an informative predictor.

It is important to note that correlations involving soil pH should not be interpreted strictly as direct causal relationships with moisture or EC_{pw} dynamics. Low-cost multi-parameter sensors may exhibit cross-sensitivity effects, where measured variables respond jointly to environmental changes such as rainfall events, temperature fluctuations, or variations in soil ionic concentration. Consequently, pH should be understood primarily as an informative predictive feature capturing shared variability in soil physicochemical conditions, rather than as an isolated mechanistic driver. This interpretation aligns with the objective of virtual sensing, where predictive usefulness does not necessarily imply direct physical causality.

A second line of comparison relates to research focused on improving the reliability of low-cost soil sensors. Setiawan et al. [17] demonstrated that KNN-based calibration can significantly improve the accuracy of capacitive soil moisture sensors ($R^2 \approx 0.93$), while Kumari et al. [16] showed that affordable sensors such as YL-69 and EC-5 can reach performance comparable to commercial probes after appropriate calibration. Other studies have explored virtual sensing and cross-device harmonization approaches to reconcile measurements from heterogeneous sensors [19,21]. The results presented here extend this line of work by demonstrating that a single multi-parameter 7-in-1 sensor can reconstruct multiple CropX measurements across three depths with very high accuracy ($R^2 = 0.9950\text{--}0.9990$). This suggests that low-cost multisensor platforms may contain sufficient information to emulate more expensive monitoring systems when combined with appropriate data-driven models. Recent studies have also explored optimized ensemble learning strategies for soil variable prediction. Bouhacina et al. [18] proposed a hybrid metaheuristic optimization framework for tree-based models, reporting improvements in predictive accuracy for soil temperature across multiple depths. These findings further support the suitability of ensemble-based approaches for capturing nonlinear relationships in soil environmental variables. Comparable evidence has been reported in other environmental monitoring applications. Ouadi et al. [20] showed that pressure-based variables can be used to accurately estimate the Silt Density Index (SDI) in desalination systems using Gradient Boosting optimized with the Salp Swarm Algorithm. Although the physical system differs, both studies support the broader principle that easily measurable variables may contain sufficient information to approximate complex physicochemical indicators through machine learning models, reinforcing the feasibility of virtual sensing approaches in resource-constrained monitoring scenarios.

An additional contribution of this study lies in its agroecological and operational context. Many studies on virtual sensing or sensor calibration are conducted under controlled experimental conditions or in highly instrumented farms where calibration resources are readily available [15,31,32]. In contrast, this research was conducted in a commercial

tropical Hass avocado orchard in Colombia, where economic and infrastructural constraints often limit the adoption of high-end monitoring technologies. Previous research in Latin America has mainly focused on irrigation management, climatic variability, and yield responses in avocado production [33,34], or on AI-based prediction of edaphic and postharvest attributes [35–38]. However, studies specifically evaluating the capacity of low-cost sensors to reproduce high-end soil measurements in avocado orchards remain scarce. Therefore, the present work contributes both methodological and regional novelty to the literature.

Model robustness was further supported by out-of-bag validation and multi-seed sensitivity analyses, which revealed negligible variability in R^2 , RMSE, and MAE across different stochastic initializations. This stability is particularly relevant in agricultural applications, where data collection is often limited by logistical constraints and environmental variability. The consistency observed across different random seeds indicates that the predictive performance of the model is not strongly dependent on stochastic training conditions. Despite the very high predictive performance obtained, several limitations should be acknowledged. Additionally, both sensing systems were installed within the same soil pit and at a short horizontal separation distance, ensuring that measurements correspond to the same representative soil volume. This spatial configuration reduces potential bias associated with soil heterogeneity and supports the validity of the proposed virtual sensing framework under concurrent measurement conditions.

The feature importance analysis supports this interpretation, as moisture-related predictors were consistently ranked as the most influential variables across depths. This behavior is consistent with the known dependence of pore water electrical conductivity on soil water content and ion mobility. The contribution of pH and temperature variables further reflects expected physicochemical interactions affecting conductivity measurements in porous media.

Temporal validation using a chronological 80–20 partition produced negative R^2 values for most variables, confirming limited extrapolation capability when predicting future observations. This result indicates that part of the performance observed under random splits is associated with temporal dependence structures in the dataset, rather than purely independent predictive relationships between sensors. Such behavior is consistent with the physical nature of soil processes, where moisture and ion dynamics evolve gradually over time and exhibit strong temporal continuity influenced by irrigation patterns, evapotranspiration cycles, and environmental variability.

Therefore, the proposed approach should be interpreted primarily as a virtual sensing framework for reconstructing concurrent measurements rather than as a fully standalone temporal forecasting model without periodic recalibration or updating. In practical deployments, recalibration may be required at regular intervals (e.g., seasonal) or following significant environmental changes such as rainfall events or irrigation cycles, to maintain predictive accuracy under evolving soil conditions.

Although the objective of this study was to reconstruct high-end sensor measurements rather than directly optimize agronomic management decisions, accurate signal reconstruction constitutes an essential step toward cost-effective digital agriculture systems. Demonstrating that low-cost sensor signals can be transformed into high-fidelity estimates of moisture and EC_{pw} enables applications such as variable-rate irrigation management, soil salinity monitoring, crop stress detection, and yield prediction. Future research should evaluate adaptive recalibration strategies, multi-site datasets, and hybrid modeling approaches that integrate data-driven methods with agronomic knowledge to improve temporal robustness under changing environmental conditions.

6. Conclusions and Future Work

This study demonstrates that a low-cost 7-in-1 soil sensor contains sufficient information to approximate key soil variables measured by a high-end CropX probe under concurrent measurement conditions. Using machine learning models, measurements from the low-cost device were translated into accurate estimates of volumetric water content (VWC) and pore water electrical conductivity (EC_{pw}) across three soil depths. These results indicate that strong nonlinear relationships exist between both sensing systems and that data-driven models can effectively support virtual sensing strategies for soil monitoring.

From an operational perspective, the findings suggest that accessible soil measurements can be leveraged to approximate high-end sensor outputs under similar environmental conditions in Hass avocado orchards, potentially reducing economic barriers associated with advanced monitoring technologies. This capability is particularly relevant in regions where cost constraints limit the adoption of high-resolution soil sensing infrastructure. The proposed framework contributes to ongoing efforts to develop scalable and cost-effective digital agriculture solutions.

However, temporal validation revealed that predictive performance decreases substantially under chronological data separation, with negative R^2 values observed for most variables. This result suggests that the models capture relationships that are strongest when low-cost and high-end sensor measurements are observed concurrently, while extrapolation to future unseen temporal conditions remains more challenging. Therefore, the proposed approach should be interpreted primarily as a virtual sensing framework for reconstructing concurrent measurements rather than as a fully standalone temporal forecasting model without periodic recalibration or updating.

Future work should evaluate the proposed methodology across multiple growing seasons, sensor units, and geographical locations to better characterize its robustness under varying environmental conditions. Adaptive calibration strategies, including online learning, drift-aware models, and domain adaptation techniques, may improve temporal generalization under changing soil and climate dynamics. Additionally, integrating data-driven approaches with soil process knowledge or agronomic expertise may support the development of hybrid virtual sensing frameworks capable of complementing high-end monitoring systems in resource-constrained agricultural environments.

Author Contributions: Conceptualization, A.F.P.B.; methodology, A.F.P.B., D.A.M.R. and N.G.G.; software, A.F.P.B.; validation, A.F.P.B.; formal analysis, A.F.P.B.; investigation, A.F.P.B.; data curation, A.F.P.B.; writing—original draft preparation, A.F.P.B.; writing—review and editing, A.F.P.B., D.A.M.R. and N.G.G.; visualization, A.F.P.B.; supervision, D.A.M.R. and N.G.G.; project administration, D.A.M.R. and N.G.G.; funding acquisition, D.A.M.R. and N.G.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Colombian General System of Royalties (Sistema General de Regalías, SGR) through the project “Development of an energy-autonomous smart agriculture platform for continuous monitoring of relevant variables aimed at improving productivity and mitigating environmental impacts in horticultural crops of Antioquia and Quindío”, executed by Universidad de Antioquia, grant number BPIN 2022000100012.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request within the framework of the research project BPIN 2022000100012 led by Universidad de Antioquia. The complete codebase used in this study, including scripts for dataset construction, exploratory analysis, correlation computation, model training, evaluation, and figure generation, is maintained in a private GitHub repository. Due

to institutional data restrictions, the repository is currently private but can be made available for academic and research purposes upon reasonable request: https://github.com/AndresParra11/7en1_cropx (accessed on 20 April 2026).

Acknowledgments: The authors acknowledge Universidad de Antioquia and the research project BPIN 2022000100012 for providing the technological infrastructure and sensor network used to collect the soil monitoring data analyzed in this study.

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

AI	Artificial intelligence
CWSI	Crop water stress index
EC	Electrical conductivity
ECpw	Pore-water electrical conductivity
ETc	Crop evapotranspiration
IoT	Internet of Things
MAE	Mean absolute error
ML	Machine learning
MLP	Multilayer perceptron
OOB	Out-of-bag
RF	Random Forest
XGB	Extreme Gradient Boosting
RMSE	Root mean square error
VWC	Volumetric water content

References

- Sharma, K.; Shivandu, S.K. Integrating artificial intelligence and Internet of Things (IoT) for enhanced crop monitoring and management in precision agriculture. *Sens. Int.* **2024**, *5*, 100292. [\[CrossRef\]](#)
- Bouali, E.T.; Abid, M.R.; Boufounas, E.M.; Hamed, T.A.; Benhaddou, D. Renewable Energy Integration into Cloud IoT-Based Smart Agriculture. *IEEE Access* **2022**, *10*, 1175–1191. [\[CrossRef\]](#)
- Aijaz, N.; Lan, H.; Raza, T.; Yaqub, M.; Iqbal, R.; Pathan, M.S. Artificial intelligence in agriculture: Advancing crop productivity and sustainability. *J. Agric. Food Res.* **2025**, *20*, 101762. [\[CrossRef\]](#)
- Karunathilake, E.M.; Le, A.T.; Heo, S.; Chung, Y.S.; Mansoor, S. The Path to Smart Farming: Innovations and Opportunities in Precision Agriculture. *Agriculture* **2023**, *13*, 1593. [\[CrossRef\]](#)
- Khan, S.; Sachan, H.K.; Krishna, D. The Role of Smart Farming Technologies in Mitigating Climate Change and Enhancing Agricultural Sustainability. *Int. J. Environ. Clim. Change* **2025**, *15*, 138–159. [\[CrossRef\]](#)
- Mizik, T. How can precision farming work on a small scale? A systematic literature review. *Precis. Agric.* **2023**, *24*, 384–406. [\[CrossRef\]](#)
- Sajib, M.M.H.; Sayem, A.S.M. Innovations in Sensor-Based Systems and Sustainable Energy Solutions for Smart Agriculture: A Review. *Encyclopedia* **2025**, *5*, 67. [\[CrossRef\]](#)
- Schwamback, D.; Persson, M.; Berndtsson, R.; Bertotto, L.E.; Kobayashi, A.N.A.; Wendland, E.C. Automated Low-Cost Soil Moisture Sensors: Trade-Off between Cost and Accuracy. *Sensors* **2023**, *23*, 2451. [\[CrossRef\]](#)
- Abdelmoneim, A.A.; Kalaany, C.M.A.; Khadra, R.; Derardja, B.; Dragonetti, G. Calibration of Low-Cost Capacitive Soil Moisture Sensors for Irrigation Management Applications. *Sensors* **2025**, *25*, 343. [\[CrossRef\]](#)
- González-Teruel, J.D.; Torres-Sánchez, R.; Blaya-Ros, P.J.; Toledo-Moreo, A.B.; Jiménez-Buendía, M.; Soto-Valles, F. Design and calibration of a low-cost SDI-12 soil moisture sensor. *Sensors* **2019**, *19*, 491. [\[CrossRef\]](#)
- Medicio. *Soil Sensor V04: Product Specifications and Technical Datasheet*; CropX Technologies Ltd.: Wellington, New Zealand, 2023. Available online: <https://cropx.com/wp-content/uploads/2023/10/Datasheet-V04-ES.pdf> (accessed on 20 April 2026).

12. Jiangsu Jingxun Changtong Electronic Technology Co., Ltd. (JXCT). *RS485 Modbus JXBS-3001-NPK-RS Soil NPK Sensor User Manual*; Jiangsu Jingxun Changtong Electronic Technology Co., Ltd. (JXCT): Weihai, China, 2020. Available online: <https://manuals.plus/m/d86fb8b39ce379db332b450ae70742cbb5e411cef3a7bc8cdb61397f01e2015e> (accessed on 20 April 2026).
13. Qazi, S.; Khawaja, B.A.; Farooq, Q.U. IoT-Equipped and AI-Enabled Next Generation Smart Agriculture: A Critical Review, Current Challenges and Future Trends. *IEEE Access* **2022**, *10*, 21219–21235. [[CrossRef](#)]
14. Katimbo, A.; Rudnick, D.R.; Zhang, J.; Ge, Y.; DeJonge, K.C.; Franz, T.E.; Shi, Y.; Liang, W.-Z.; Qiao, X.; Heeren, D.M.; et al. Evaluation of artificial intelligence algorithms with sensor data assimilation in estimating crop evapotranspiration and crop water stress index for irrigation water management. *Smart Agric. Technol.* **2023**, *4*, 100176. [[CrossRef](#)]
15. Placidi, P.; Morbidelli, R.; Fortunati, D.; Papini, N.; Gobbi, F.; Scorzoni, A. Monitoring soil and ambient parameters in the IoT precision agriculture scenario: An original modeling approach dedicated to low-cost soil water content sensors. *Sensors* **2021**, *21*, 5110. [[CrossRef](#)]
16. Kumari, S.; Ali, N.; Dagati, M.; Dong, Y. IoT-Enabled Soil Moisture and Conductivity Monitoring Under Controlled and Field Fertigation Systems. *AgriEngineering* **2025**, *7*, 207. [[CrossRef](#)]
17. Setiawan, I.; Musa, M.D.T.; Afriza, D.A.; Hafidah, S.N. Comparing Machine Learning Algorithms to Enhance Volumetric Water Content Prediction in Low-Cost Soil Moisture Sensor. *J. Appl. Inform. Comput. (JAIC)* **2025**, *9*, 140–145. [[CrossRef](#)]
18. Bouhacina, O.; Khatir, A.; Matoug, S.A.; Tamine, T. A Novel Hybrid TOARS-Optimized Ensemble of Tree-Based Models for Predicting Soil Temperature at Shallow Depths. *Sustainability* **2025**, *18*, 294. [[CrossRef](#)]
19. Lin, Y.W.; Lin, Y.B.; Chang, T.C.; Lu, B.X. An Edge Transfer Learning Approach for Calibrating Soil Electrical Conductivity Sensors. *Sensors* **2023**, *23*, 8710. [[CrossRef](#)] [[PubMed](#)]
20. Ouadi, B.; Khatir, A.; Magagnini, E.; Mokadem, M.; Abualigah, L.; Smerat, A. Optimizing silt density index prediction in water treatment systems using pressure-based gradient boosting hybridized with Salp Swarm Algorithm. *J. Water Process Eng.* **2024**, *68*, 106479. [[CrossRef](#)]
21. Burri, S.R.; Agarwal, D.K.; Vyas, N.; Duggar, R. Optimizing Irrigation Efficiency with IoT and Machine Learning: A Transfer Learning Approach for Accurate Soil Moisture Prediction. In *2023 World Conference on Communication & Computing (WCONF)*; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2023. [[CrossRef](#)]
22. Hilhorst, M.A. A pore water conductivity sensor. *Soil Sci. Soc. Am. J.* **2000**, *64*, 1922–1925. [[CrossRef](#)]
23. Garofalo, S.P.; Modugno, A.F.; Carolis, G.D.; Sanitate, N.; Tesemma, M.N.; Scarascia-Mugnozza, G.; Tegegne, Y.T.; Campi, P. Explainable Artificial Intelligence to Predict the Water Status of Cotton (*Gossypium hirsutum* L., 1763) from Sentinel-2 Images in the Mediterranean Area. *Plants* **2024**, *13*, 3325. [[CrossRef](#)]
24. van Klompenburg, T.; Kassahun, A.; Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **2020**, *177*, 105709. [[CrossRef](#)]
25. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*; Wiley: Hoboken, NJ, USA, 2012.
26. Chlingaryan, A.; Sukkarieh, S.; Whelan, B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Comput. Electron. Agric.* **2018**, *151*, 61–69. [[CrossRef](#)]
27. Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine learning in agriculture: A review. *Sensors* **2018**, *18*, 2674. [[CrossRef](#)]
28. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
29. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
30. Genuer, R.; Poggi, J.M.; Tuleau, C. Random Forests: Some methodological insights. *arXiv* **2008**, arXiv:0811.3619. [[CrossRef](#)]
31. Kim, H.N.; Park, J.H. Monitoring of soil EC for the prediction of soil nutrient regime under different soil water and organic matter contents. *Appl. Biol. Chem.* **2024**, *67*, 1. [[CrossRef](#)]
32. Amado, T.M.; Alvarez, A.E.D.; Ocampo, A.; Paz, V.A.F.; Punongbayan, A.J.N.; Yumena, L.M.; Padilla, M.V.C.; Madrigal, G.A.M.; Tolentino, L.K.S.; Monilar, E.G. Development of an IoT-Based Soil Macronutrient Analysis System Utilizing Electrochemical Sensors and Machine Learning Algorithms. In *2023 International Conference on Network, Multimedia and Information Technology (NMITCON)*; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2023. [[CrossRef](#)]
33. Erazo-Mesa, E.; Ramírez-Gil, J.G.; Sánchez, A.E. Avocado cv. Hass needs water irrigation in tropical precipitation regime: Evidence from Colombia. *Water* **2021**, *13*, 1942. [[CrossRef](#)]
34. Rodríguez, C.M.C.; Alvarez, G.E.G.; Porras, J.C.C.; Acevedo, C.A.J.; Vélez, J.P.A.; Martínez, A.A.P. Formulación de un modelo para determinar las zonas actuales y potenciales de cultivo de aguacate Hass (*Persea americana* Mill) en el departamento de Risaralda a partir de variables edafoclimáticas y de calidad del fruto. *Entre Cienc. Ing.* **2024**, *18*, 41–45. [[CrossRef](#)]
35. Zapata, C.; Castro, A.; Espino, R.; Yauri, R. Monitoring and Prediction of Hass Avocado Conditions Using Support Vector Machines on an Embedded Device. *SSRG Int. J. Electr. Electron. Eng.* **2025**, *12*, 1–10. [[CrossRef](#)]
36. Yauri, R.; Cuyubamba, L.; Nuñez, S. Crop Monitoring System Using IoT, Solar Energy and Decision Tree Algorithm. *Emerg. Sci. J.* **2025**, *9*, 603–614. [[CrossRef](#)]

37. Ochoa-Ornelas, R.; Gudiño-Ochoa, A. Implementation of models to predict avocado exports. *J. Math. Quant. Methods* **2022**, *6*, 6–17. [[CrossRef](#)]
38. Moreno-Bernal, P.; Arizmendi-Peralta, P.; Hernández-Aguilar, J.A.; del Carmen Peralta-Abarca, J.; Velásquez-Aguilar, J.G. IoT Platform for Monitoring Nutritional and Weather Conditions of Avocado Production. In *Proceedings of the Smart Cities; Communications in Computer and Information Science (CCIS)*; Springer Science and Business Media Deutschland GmbH: Cham, Switzerland, 2023; Volume 1706, pp. 95–109. [[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.