COMPARATIVE ASSESSMENT OF RANDOM FOREST, XGBOOST AND MLP MODELS FOR PREDICTIVE GREENHOUSE CONTROL USING WIRELESS SENSOR NETWORK DATA

/

EVALUAREA COMPARATIVĂ A MODELELOR RANDOM FOREST, XGBOOST ȘI MLP PENTRU CONTROLUL PREDICTIV AL UNEI SERE, FOLOSIND DATE DINTR-O REȚEA DE SENZORI WIRELESS

Mihai Gabriel MATACHE[1], Carmen BĂLȚATU[*1], Dragoş SĂCĂLEANU[2], Adrian IOSIF[2], Tudor Adrian ENE[3], Cristina Mihaela DOBRE[4]

[1] INMA Bucharest/ Romania;

[2] National University for Science and Technology POLITEHNICA Bucharest, Romania;

[3] ICDP Brașov; [4] Beia Consult International SRL

E-mail: carmen.vasilachi@gmail.com

DOI: https://doi.org/10.35633/inmateh-79-51

Keywords: *greenhouse control, wireless sensor network, predictive control, machine learning, Random Forest, XGBoost, multilayer perceptron.*

ABSTRACT

Control of greenhouse microclimate requires the interpretation of several environmental and substrate-related parameters, including air temperature, relative humidity, light intensity, carbon dioxide concentration and soil moisture. This paper compares three machine learning models, Random Forest, Extreme Gradient Boosting (XGBoost) and a multilayer perceptron (MLP), for 30-minute-ahead prediction of irrigation, ventilation and shading states in a greenhouse. The models were trained using data collected from a wireless sensor network (WSN) with four sensor nodes installed in the protected cultivation area. After preprocessing, scaling, error removal, temporal synchronization and resampling at 10-minute intervals, a structured dataset of 7063 records was obtained. Individual sensor readings, spatially aggregated variables, time-related features, rolling averages and short-term differences were used to describe both the current and recent evolution of the greenhouse environment. The dataset was split chronologically into training, validation and test subsets, corresponding to 70%, 15% and 15% of the records, respectively, while preserving the temporal order of the measurements. The best average performance was obtained by XGBoost, with a mean F1-score of 0.9237, followed by Random Forest with 0.9045 and MLP with 0.5862. Random Forest achieved the best results for irrigation and ventilation control, with F1-scores of 0.9630 and 0.9611, respectively, while XGBoost achieved the best result for shading control, with an F1-score of 0.8571. XGBoost also achieved the highest mean accuracy, 0.9506, and mean balanced accuracy, 0.8904. The results indicate that tree-based ensemble models are more suitable than the tested MLP architecture for predictive greenhouse control based on multi-sensor WSN data. The proposed framework supports the selection of a final control model according to prediction performance, stability and interpretability.

REZUMAT

Controlul microclimatului din seră necesită interpretarea mai multor parametri de mediu și de substrat, inclusiv temperatura aerului, umiditatea relativă, intensitatea luminoasă, concentrația de dioxid de carbon și umiditatea solului. Această lucrare compară trei modele de învățare automată, Random Forest, Extreme Gradient Boosting (XGBoost) și perceptron multistrat (MLP), pentru predicția la 30 de minute a stărilor de irigare, ventilație și umbrire într-o seră. Modelele au fost antrenate folosind date colectate de la o rețea de senzori wireless (WSN) cu patru noduri instalate în spațiul protejat. După preprocesare, scalare, eliminarea erorilor, sincronizare temporală și reeșantionare la intervale de 10 minute, a fost obținut un set de date structurat, cu 7063 de înregistrări. Valorile individuale ale senzorilor, variabilele agregate spațial, variabilele temporale, mediile mobile și diferențele pe termen scurt au fost utilizate pentru a descrie atât starea curentă, cât și evoluția recentă a mediului din seră. Setul de date a fost împărțit cronologic în subseturi de antrenare, validare și testare, reprezentând 70%, 15% și, respectiv, 15% din înregistrări, cu păstrarea ordinii temporale a măsurătorilor. Cea mai bună performanță medie a fost obținută de XGBoost, cu un F1-score mediu de 0.9237, urmat de Random Forest, cu 0.9045, și MLP, cu 0.5862. Random Forest a obținut cele mai bune rezultate pentru controlul irigării și ventilației, cu F1-score-uri de 0.9630 și 0.9611, în timp ce XGBoost a obținut cel mai bun rezultat pentru controlul umbririi, cu un F1-score de 0.8571.

XGBoost a obținut, de asemenea, cea mai mare acuratețe medie, de 0.9506, și cea mai mare acuratețe echilibrată medie, de 0.8904. Rezultatele indică faptul că modelele ensemble bazate pe arbori sunt mai potrivite decât arhitectura MLP testată pentru controlul predictiv al serelor pe baza datelor multi-senzor WSN. Cadrul propus susține selectarea unui model final de control în funcție de performanța de predicție, stabilitate și interpretabilitate.

INTRODUCTION

Greenhouse cultivation allows a higher degree of control over plant growth conditions than open-field agriculture, but this advantage depends on the continuous regulation of several interacting environmental parameters. Air temperature, relative humidity, soil moisture, light intensity and carbon dioxide concentration influence plant development, water consumption, photosynthetic activity and thermal stress. If these parameters are not monitored and controlled properly, the greenhouse may consume excessive water and energy or may expose plants to unfavourable microclimate conditions.

Traditional greenhouse control systems are often based on fixed threshold rules. Such systems are simple and easy to implement, for example by activating irrigation when soil moisture drops below a predefined limit or by starting ventilation when air temperature or CO2 concentration exceeds a threshold. However, the greenhouse microclimate is dynamic and nonlinear. The need for irrigation, ventilation or shading may depend not only on one variable, but on the combined effect of temperature, humidity, radiation, CO2 concentration, soil moisture, time of day and recent trends in the measured data. Therefore, fixed rules may become insufficient when environmental variables interact or when borderline conditions occur. For this reason, modern greenhouse management increasingly relies on sensor-based monitoring, automatic data acquisition and decision support systems (*Dhanaraju et al., 2022; Săcăleanu et al., 2024*). Decision support systems have already shown practical value in horticultural irrigation management. For example, *Buono et al. (2022)* field-tested the Bluleaf DSS for kiwifruit irrigation management in Italy and reported water savings of approximately 20-25% compared with current farm management, without negative effects on yield and fruit quality. In this view, Wireless Sensor Networks (WSNs) are well suited for greenhouse monitoring and control applications because they allow the distributed measurement of environmental and substrate-related variables in different points of the protected cultivation area. Compared with single-point monitoring, multiple sensor nodes can capture spatial variability inside the greenhouse, which is important for irrigation, ventilation and shading decisions. Previous studies have shown that WSN-based systems can support real-time greenhouse monitoring, data aggregation and decision support, but also that the correct integration of sensing, communication, data processing and control remains a key challenge (*Kochhar and Kumar, 2019; Săcăleanu et al., 2024*).

IoT-based greenhouse control studies have emphasized the need for data-driven and optimization-based approaches able to improve climate regulation and resource efficiency (*Ullah et al., 2022; Chen et al., 2025*). *Ardiansah et al. (2022)* presented an IoT-enabled device for real-time monitoring and control of greenhouse irrigation water quality, using pH, TDS, EC, temperature and turbidity sensors. *Tan et al. (2021)* developed an intelligent irrigation system for a jujube orchard, consisting of a field control terminal and a remote client connected through a 4G network. The system monitored air temperature, relative humidity, light intensity, soil temperature and soil water content, and automatically controlled irrigation using lower and upper soil water content thresholds. Recent field studies have confirmed the importance of sensor-based and model-based irrigation strategies under increasing climatic variability. *Föll et al. (2026)* evaluated irrigation strategies in apple orchards over three years in Germany, comparing soil water tension-based irrigation, a web-based soil water balance model, interval irrigation and a non-irrigated control. Their results showed that sensor- and model-based irrigation strategies can improve water use efficiency compared with conventional interval irrigation.

Machine learning offers a practical alternative for transforming heterogeneous sensor data into control decisions. Tree-based ensemble models are particularly attractive for this type of application because they can handle nonlinear relationships, noisy measurements and mixed groups of input variables. A recent review focused on digital soil evaluation likewise identified Random Forest and boosting algorithms such as XGBoost as robust options for heterogeneous and multi-source agricultural datasets, while also emphasizing the associated trade-offs involving computational cost, hyperparameter sensitivity and model interpretability (*Nenciu et al., 2026*). Random Forest combines multiple decision trees in order to improve robustness and reduce overfitting, while also providing feature importance values that can be interpreted from an agronomic point of view (*Breiman, 2001*).

Recent studies in precision horticulture have also shown the value of integrating environmental monitoring, sensor fusion and machine learning for crop and fruit quality prediction. *Wang et al. (2026)* combined X-ray computed tomography, environmental data and machine learning models, including Random Forest and XGBoost, to analyse preharvest apple growth dynamics. Their results showed that environmental factors, especially temperature, can be used as relevant predictors in data-driven fruit quality modelling. Although their work focused on orchard fruit quality rather than greenhouse actuator control, it supports the broader use of multi-source sensor data and machine learning for predictive decision-making in horticultural system. XGBoost extends the tree-based approach through gradient boosting, where trees are added sequentially to correct previous errors, often leading to high predictive performance on structured tabular datasets (*Chen and Guestrin, 2016*).

Several studies have already demonstrated the potential of machine learning for greenhouse-related prediction tasks. *Cai et al. (2022)* developed a greenhouse temperature prediction model based on LightGBM using five years of environmental, control and temporal data collected from a Venlo-type greenhouse. Their comparison with BP neural network, RNN, XGBoost and stochastic gradient boosting showed that LightGBM achieved the best predictive performance, with an RMSE of 0.645 °C on the testing dataset and an average relative error below 2%. XGBoost also performed well, with an RMSE of 0.679 °C, confirming the suitability of gradient-boosted tree models for greenhouse microclimate prediction and real-time predictive control. *Tsai et al. (2020)* combined weather forecast data, numerical models and machine learning methods, including Random Forest, to predict soil temperature and volumetric water content in a greenhouse. Such results support the use of data-driven models in protected cultivation, especially when environmental variables are measured continuously and must be translated into management decisions. However, many studies focus on predicting individual environmental parameters, while fewer works address the direct prediction of control states for irrigation, ventilation and shading using multi-sensor greenhouse data (*Tsai et al., 2020*). *Wang et al. (2025)* compared several machine learning models for greenhouse cherry tomato irrigation prediction, including MLR, Ridge Regression, SVM, Random Forest, XGBoost, GRU, LSTM and Transformer. Their results showed that XGBoost provided the best performance when maximum and average values of temperature, relative humidity and light intensity were used as input variables. The model achieved R^2= 0.820 during the vegetative growth stage and R^2=0.840 during the reproductive growth stage. After deployment through a cloud platform and STM32-based irrigation control system, the model increased single fruit weight by 17.8% and improved water use efficiency by 20.7% compared with conventional irrigation management. *Mohammadi et al. (2024)* used IoT sensor data and machine learning for greenhouse water management in tomato production under water-limited conditions in Afghanistan. The authors collected real sensor data from a simulated greenhouse between June and August 2022 and compared six machine learning models. XGBoost achieved the best and most balanced classification performance among the tested models, indicating its suitability for predicting irrigation needs and avoiding excessive water use in greenhouse tomato production. The study also included an IoT-based visualization interface for farmers, supporting real-time supervision of greenhouse conditions and timely irrigation decisions. Another study compared several machine learning models, including Random Forest, XGBoost, LightGBM and MLP, for short-term greenhouse microclimate forecasting (*Cebolla-Alemany et al., 2026*). This study used temporal feature engineering, including lag variables and rolling statistics, and showed that tree-based ensemble models were highly suitable for predictive greenhouse control. In their case, LightGBM achieved the best temperature forecasting performance, with R^2 values above 0.99 for 5, 10 and 15-minute horizons, while XGBoost and Random Forest obtained comparable results. The integration of the forecasting model into a fuzzy predictive controller also reduced energy-related control costs by more than 60% compared with a conventional threshold-based strategy. Forecast-driven greenhouse control has also been investigated using time-series neural networks. *Elvanidi and Katsoulas (2023)* used microclimate and crop physiological data to detect tomato crop stress in greenhouses. Their Gradient Boosting classifier achieved 91% accuracy and 92% F1-score on the test set, while MLP reached 89% accuracy and 89% F1-score, supporting the use of ML models for greenhouse decision support. *Aborujilah et al. (2025)* developed an LSTM-based framework for smart greenhouse climate control, using multivariate environmental and actuator data to forecast temperature, relative humidity and CO_2 concentration. Their model achieved high prediction accuracy, with R^2=0.9835, and was used to generate actuator recommendations for heating, ventilation and lighting. *Ait Oussous et al. (2025)* compared several deep learning models for greenhouse internal temperature prediction and reported the best performance for PLSTM, with R^2=0.9999, RMSE = 0.0220 and MAE = 0.0160.

In this context, the present paper addresses the development of several machine learning-based decision models for greenhouse control using real data collected from a WSN installed in a protected cultivation environment. The monitored variables include air temperature, relative humidity, light intensity, CO2 concentration, atmospheric pressure, soil moisture and auxiliary diagnostic parameters collected from four distributed sensor nodes. The control problem is formulated as a supervised classification task, where the models estimate the future state of three greenhouse subsystems: irrigation, ventilation and shading.

The main objective of this study is to compare Random Forest, XGBoost and a multilayer perceptron neural network under the same experimental conditions and to identify the most suitable model for predictive greenhouse control. Unlike a same-time classification approach, where sensor values and actuator states correspond to the same timestamp, the final evaluation is formulated as a 30-minute-ahead prediction problem. In this formulation, the sensor-derived variables recorded at time t are used to estimate the control states at t+30 minutes. This approach is more relevant for practical greenhouse automation, because it evaluates the ability of the models to anticipate future control requirements rather than only reproduce the current state of the system.

MATERIALS AND METHODS

Experimental dataset and greenhouse monitoring system

The dataset used in this study was obtained from a wireless sensor network installed in a greenhouse environment. The monitoring system consisted of four distributed sensor nodes, identified as node_1, node_2, node_3 and node_4. Each node recorded several environmental and substrate-related parameters, allowing the characterization of the spatial variability of the greenhouse microclimate.

The raw dataset contained 1222305 individual sensor records stored in long format, where each row corresponded to one measurement from one sensor at a given timestamp. After cleaning, scaling, temporal synchronization and resampling at 10-minute intervals, the data were transformed into a wide-format dataset containing 7063 synchronized time steps. Each time step described the greenhouse state using 216 variables, including original sensor measurements, spatially aggregated microclimate indicators, temporal features, rolling statistics and the recorded control status variables.

Measured variables

The four WSN nodes included similar groups of sensors. The monitored variables covered the main physical quantities required for greenhouse control: air temperature, relative humidity, light intensity, carbon dioxide concentration, atmospheric pressure, soil moisture, local temperature measurements and battery voltage.

The following groups of variables were identified in the processed dataset:

Table 1

Types of process variables

Sensor group	Number of columns	Role in the control model
Air temperature	8	ventilation and shading control
Air relative humidity	8	ventilation control
Light intensity	4	shading control
CO2 concentration	4	ventilation control
Soil moisture	12	irrigation control

The air temperature group included measurements from HTS221 and SHT31 sensors installed on the four nodes. The air humidity group included the corresponding relative humidity measurements from the same sensor families. Light intensity was measured using TEMT6000 sensors, while CO2 concentration was measured using SCD41 sensors. Soil moisture was monitored using both tensiometer-based and capacitive sensors, providing several measurements per node. Besides these main variables, the dataset also included atmospheric pressure, UVA and UVB radiation, DS18B20 temperature measurements and battery voltage. These variables were retained in the processed dataset when their data quality was acceptable, but the main control decisions were based on the variables most directly related to greenhouse irrigation, ventilation and shading.

Raw data scaling

The raw sensor values were stored using an integer-based format. Therefore, a scaling operation was required before the data could be interpreted as physical measurements. The measured values were converted using the following relation:

$$x_{scaled} = \frac{x_{raw}}{10^6} \tag{1}$$

where x_{raw} is the original stored value and x_{scaled} is the corresponding physical or sensor-scale value used in the processed dataset.

This scaling step was applied uniformly to all measured values before the physical plausibility filters were applied.

Data cleaning and removal of erroneous values

The raw dataset contained abnormal values caused by sensor errors, communication issues or invalid readings. These values were removed or marked as missing before the temporal synchronization step. The cleaning process included:

- removal of missing timestamps, sensor identifiers and measurement values;
- conversion of timestamps to a standard datetime format;
- application of the scaling factor;
- identification of sensor-specific abnormal values;
- removal of physically impossible measurements;
- filtering of columns with excessive missing values.

Several physically based rules were applied according to the type of sensor. For example, air temperature values outside the plausible greenhouse range were removed, relative humidity values outside 0-100% were considered invalid, and CO2 values outside the expected measurement interval were discarded. Negative light intensity values were also removed. Similar checks were applied to atmospheric pressure, UV measurements and battery voltage.

Columns with a high proportion of missing values were removed from the final dataset. In the implemented preprocessing pipeline, a column was retained only if its missing data ratio was below the established threshold. This step reduced the risk of training models on unreliable variables.

The final report still indicated that some node-specific variables had relatively high missing ratios, especially for node 4 and for some UV-related measurements. For example, several node 4 variables had missing ratios around 0.54-0.58, while UVB and UVA measurements also showed notable missing data percentages. These variables were not used as primary control indicators.

Temporal synchronization

The raw WSN data were not directly synchronized, because the individual sensors and nodes did not necessarily transmit measurements at exactly the same timestamp. For machine learning, all input variables must refer to the same temporal state of the greenhouse. Therefore, the measurements were resampled using a fixed time interval.

A time step of 10 minutes was selected for the processed dataset. This interval provided a compromise between temporal resolution and data stability. It was short enough to capture relevant microclimate changes and long enough to reduce the effect of short-term noise or communication irregularities.

For each 10-minute interval, sensor measurements belonging to the same variable were aggregated using the mean value. The result was a wide-format dataset in which each row represented the greenhouse state at a specific time step, and each column represented one sensor-derived variable.

Interpolation of short data gaps

After temporal resampling, some missing values remained because not all sensors reported data in every 10-minute interval. Short gaps were filled using time-based interpolation. The interpolation was limited to a small number of consecutive missing intervals in order to avoid generating artificial data over long periods. In this study, interpolation was limited to three consecutive time steps. Since the sampling interval was 10 minutes, this corresponds to a maximum interpolated gap of approximately 30 minutes. Longer gaps were left as missing values and handled during the subsequent modelling stage.

This approach preserves the continuity of the dataset without overestimating the reliability of intervals where sensor information was not available.

Transformation from long format to machine learning format

The original dataset was stored in long format because of wireless communication optimization: **timestamp – sensor_id-value**

For machine learning, this format was transformed into a wide tabular structure:

timestamp	node_1_temp	node_1_rh	node_1_light	node_2_temp	...	control variables

This transformation allowed the models to receive, at each time step, a complete description of the greenhouse state. The final dataset retained the measurements from the individual nodes, so that the models could use spatial differences between measurement points. In figure 1 is presented a representative 3-day evolution of raw greenhouse variables recorded by node 1. The figure shows air and soil temperature, air humidity, soil moisture index and light intensity. The light intensity curve highlights the day-night alternation during the selected period The raw capacitive soil moisture signal was converted into a soil moisture index ranging from 0 to 100 using robust normalization based on the 5th and 95th percentiles calculated over the full monitoring period of the selected node. This index was used only for graphical representation and does not represent calibrated volumetric soil water content.

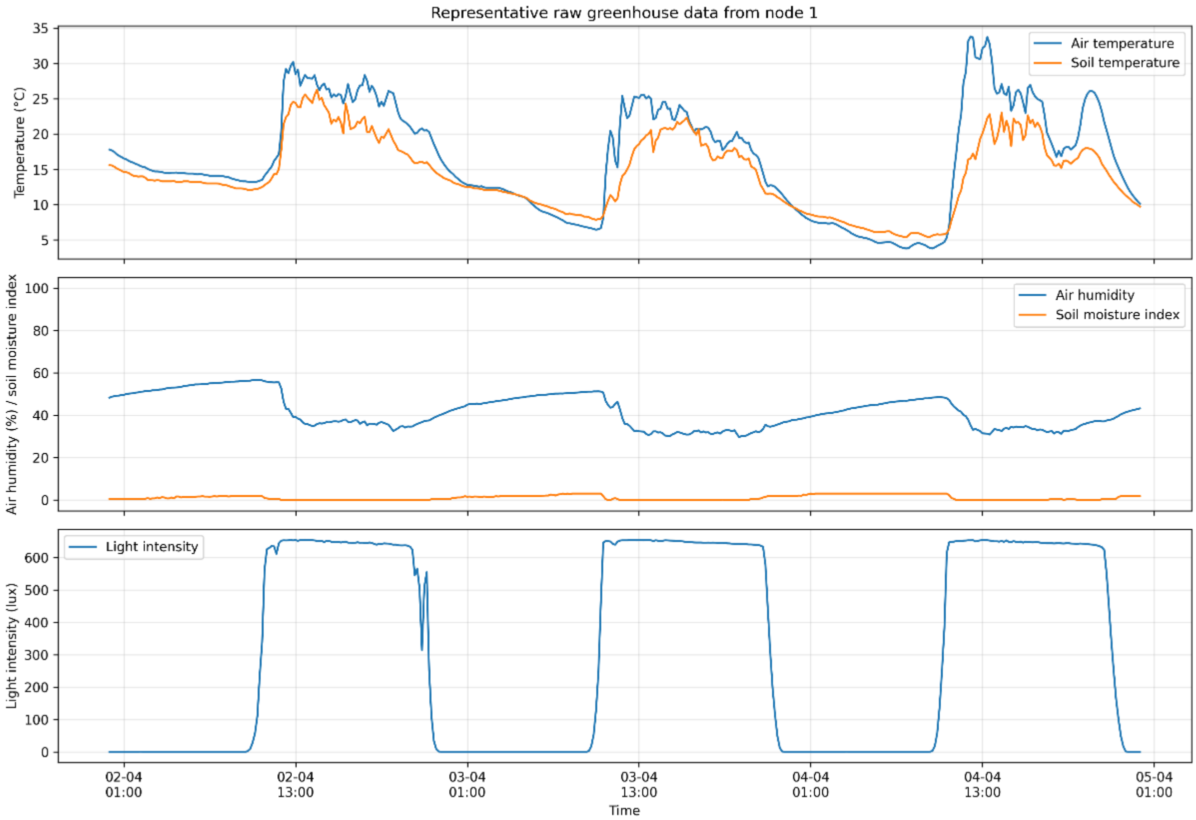

Fig.1 - Representative 3-day evolution of raw greenhouse variables recorded by node 1

Spatial aggregation of sensor data

Since the greenhouse was monitored using four WSN nodes, the dataset contains spatial information. The nodes were not treated as four independent control zones. Instead, they were used to describe the spatial variability of the greenhouse microclimate.

The control strategy considered in this study is a global control strategy:

4 WSN nodes→1 irrigation command, 1 ventilation command, 1 shading command

Therefore, in addition to the individual node measurements, aggregated variables were generated for each sensor group. For air temperature, air humidity, light intensity, CO2 concentration and soil moisture, the following statistics were calculated: x_{median}, x_{mean}, x_{min} , x_{max}, x_{range}, x_{std} , where:

$$x_{range} = x_{max} - x_{min} \tag{2}$$

Temporal feature generation

In addition to sensor measurements, time-related variables were generated. These variables help the machine learning models account for daily and seasonal patterns.

The following temporal variables were included:

- hour of the day;
- month;
- day of the year;
- sine and cosine encoding of hour;
- sine and cosine encoding of day of year.

Cyclic encoding was used for the hour of the day and the day of year in order to represent the daily and seasonal periodicity of greenhouse microclimate conditions. This transformation avoids the artificial discontinuity introduced by simple numerical encoding. For example, 23:00 and 00:00 are temporally close, although their numerical values are distant if the hour is represented as an integer. The cyclic representation allows the models to use time-related patterns associated with solar radiation, temperature evolution, ventilation demand and shading activation. The cyclic representation was calculated as:

$$\begin{cases} hour_{sin} = sin\left(2\pi \frac{hour}{24}\right) \\ hour_{cos} = cos\left(2\pi \frac{hour}{24}\right) \end{cases} \tag{3}$$

and similarly for the day of year.

Rolling and dynamic features

Greenhouse control decisions depend not only on the current measured values, but also on their recent evolution. Rolling and dynamic features were generated in order to describe the recent evolution of the greenhouse microclimate, not only its instantaneous state. For selected sensor variables, rolling means over 30 and 60 minutes were calculated, together with the difference from the previous time step. Since the dataset was resampled at 10-minute intervals, the 30-minute rolling mean used three consecutive samples, while the 60-minute rolling mean used six consecutive samples. These variables allow the models to capture short-term trends, such as increasing air temperature, decreasing soil moisture or rapid changes in light intensity, which are relevant for anticipating irrigation, ventilation and shading decisions at the 30-minute prediction horizon.

This feature engineering step was particularly important because the final modelling task was formulated as a 30-minute-ahead prediction problem. Therefore, the models had to use both the current state and the recent trend of the greenhouse environment to estimate future actuator states.

Control status variables

The final dataset included three binary control status variables: irrigation_control (OFF/ON), ventilation_control (OFF/ON) and shading_control (OFF/ON). In the methodological framework of this paper, these variables were treated as measured control status values associated with the greenhouse control system. They represent the observed operating state of the irrigation, ventilation and shading subsystems at each synchronized time step.

The distribution of the control status variables in the final dataset was:

Table 2

Control variables allocation

Control variable	OFF	ON
irrigation_control	5469	1594
ventilation_control	1213	5850
shading_control	5560	1503

This distribution shows that irrigation and shading were active in fewer time steps than they were inactive, while ventilation was active in most of the synchronized records. Because of this class imbalance, the training procedure used classification metrics that are suitable for imbalanced datasets. Precision, recall, F1-score and confusion matrices were used for each control variable.

Final dataset for machine learning

After preprocessing, the dataset was suitable for supervised machine learning. Each row represented one synchronized 10-minute greenhouse state, while the columns included:

1. individual sensor measurements from the four WSN nodes;
2. aggregated variables for each sensor group;
3. time-based features;
4. rolling and differential features;
5. reference control states for irrigation, ventilation and shading.

Agronomic operating limits were used only to interpret the control philosophy and to discuss model behaviour. They were not used as input variables and were not used to generate the target labels. The target variables were the recorded actuator states for irrigation, ventilation and shading.

Control philosophy

The control strategy adopted in this study was based on a greenhouse-wide decision model. The four WSN nodes captured the spatial variability of the greenhouse environment, whereas the control output represented the overall actuation state of each subsystem. The model did not assign separate irrigation, ventilation, or shading commands to individual sensor nodes. Instead, measurements from all nodes were combined to determine whether irrigation, ventilation, shading, or a combination of these operations was required.

The control philosophy was defined in relation to agronomic operating limits previously considered for greenhouse management: air temperature in the range of 18-35 °C, relative humidity between 45-85%, soil moisture between 20-60%, CO_2 concentration between 400-1200 ppm and light intensity between 500-20000 lux. In addition, reference control thresholds such as low soil moisture for irrigation, high air temperature or CO_2 concentration for ventilation and high light intensity for shading were used only as interpretative limits

Three algorithms were selected to cover complementary machine learning approaches suitable for tabular greenhouse sensor data. Random Forest was used as a robust and interpretable ensemble baseline, able to handle nonlinear relationships and noisy sensor measurements. XGBoost was included as a more advanced gradient boosting method, commonly effective on structured datasets and capable of improving predictive performance through sequential error correction. Multilayer perceptron was used to evaluate whether a neural network model could capture more complex interactions between microclimate variables and actuator states. This comparison allowed the final control model to be selected not only according to classification performance, but also according to stability and interpretability.

Random Forest was configured with 500 trees in order to reduce the variance of individual decision trees and improve prediction stability, while the maximum tree depth was limited to 15 to reduce excessive model complexity. The Gini criterion was used as a standard splitting rule, and balanced class weights were applied because the OFF and ON states were not equally represented for all control variables.

For XGBoost, a binary logistic objective was used because each control target was formulated as a binary classification problem. The number of boosting estimators was set to 500, with a learning rate of 0.05, in order to allow gradual error correction during training. The maximum tree depth was limited to 4 to control model complexity and reduce overfitting. Row and feature subsampling were both set to 0.8, providing additional regularization and improving generalization on the test subset.

The MLP architecture was kept relatively compact, with three hidden layers of 128, 64 and 32 neurons, in order to evaluate a neural network alternative without introducing an excessively large model for the available dataset size. The ReLU activation function and Adam optimizer were selected because they are widely used for tabular supervised learning problems. Input features were standardized before training, and early stopping was applied to reduce overfitting. The same chronological train-validation-test split and the same input feature set were used for all three models, allowing the comparison to reflect model behaviour rather than differences in data preparation.

The selected configurations should therefore be interpreted as controlled baseline settings suitable for comparative evaluation. Further hyperparameter optimization may improve individual model performance, but this was outside the scope of the present study.

Each algorithm was evaluated separately for irrigation, ventilation and shading control. The final model was selected according to prediction performance, stability and interpretability. Feature importance analysis was used for the tree-based models to identify the most relevant variables involved in each control decision.

To avoid direct reproduction of threshold-like actuator states, the machine learning models were trained to predict the control status at a future decision horizon of 30 minutes. The input vector at time t was therefore mapped to the actuator state at time t+30 min. In addition, the most direct aggregated control indicators, such as maximum air temperature, maximum CO_2 concentration, maximum light intensity and minimum soil moisture, were removed from the input feature set. This reduced the risk of data leakage and provided a more realistic evaluation of the ability of the models to anticipate irrigation, ventilation and shading states from multi-sensor data.

Random Forest, XGBoost and MLP were trained under identical experimental conditions. The same predictive dataset was used for all models, where sensor-derived variables recorded at time t were mapped to actuator states at t+30 minutes. The dataset was split chronologically into training, validation and test subsets, corresponding to 70%, 15% and 15% of the records, respectively. This approach preserved the temporal structure of the greenhouse data and prevented future information from being included in the training phase. The same evaluation metrics were then computed for each model and each control target, allowing a direct comparison of their predictive performance.

Software implementation

The experiments were conducted on a workstation running Windows 10 Pro 64-bit, equipped with an Intel Core i9-9900K processor at 3.60 GHz, 64 GB of RAM and an NVIDIA Quadro P4000 GPU with 8 GB of dedicated memory. All data preprocessing, feature engineering, model training and evaluation were performed using Python 3.9.13. Data handling and numerical operations were carried out using pandas 2.3.3 and NumPy 2.0.2. Data standardization, Random Forest and MLP implementation, and calculation of the evaluation metrics were performed using scikit-learn 1.6.1. The Random Forest and MLP models were executed on the CPU, whereas GPU acceleration was enabled for XGBoost by setting device="cuda" in XGBoost 2.1.4. Matplotlib 3.9.4 was used to generate the graphical representations of the results.

Evaluation metrics

The performance of the machine learning models was evaluated using accuracy, balanced accuracy, precision, recall, F1-score and ROC-AUC. These metrics were selected because the control states of the greenhouse subsystems were not equally distributed between the OFF and ON classes, especially for irrigation and shading. Therefore, the evaluation could not rely only on overall accuracy.

Accuracy represents the proportion of all correctly classified samples. It provides a general indication of model performance, but it can be misleading when one class is much more frequent than the other. For example, if the OFF state is dominant, a model may obtain high accuracy by mostly predicting OFF, while still failing to detect important ON events.

Balanced accuracy was used to provide a more reliable evaluation under class imbalance. Unlike standard accuracy, it considers the classification performance for both classes and gives equal importance to the OFF and ON states. This is important in greenhouse control, because missing an ON command for irrigation, ventilation or shading may have practical consequences even if the overall number of errors is small. Precision indicates how reliable the positive predictions are. In this study, it shows how many of the predicted ON states were actually correct. A high precision value means that the model does not frequently activate a control subsystem when it is not required. This is relevant for reducing unnecessary water use, energy consumption or shading activation.

Recall indicates how well the model detects the actual ON states. It shows whether the model correctly identifies the moments when irrigation, ventilation or shading should be active. A high recall value is important because failing to detect an ON state may lead to water stress, excessive temperature, high humidity, high CO_2 concentration or excessive solar radiation inside the greenhouse.

The F1-score combines precision and recall into a single performance indicator. It is especially useful when the dataset is imbalanced, because it reflects both the reliability of ON predictions and the ability to detect actual ON events. For this reason, F1-score was used as one of the main criteria for comparing the models.

ROC-AUC was used to evaluate the ability of each model to distinguish between the OFF and ON classes over different decision thresholds. A high ROC-AUC value indicates that the model assigns higher probabilities to true ON cases than to OFF cases, even before a final classification threshold is applied. This metric is useful for assessing the separability of the two control states.

In this study, the final model selection was based mainly on F1-score, balanced accuracy and recall, rather than accuracy alone. This approach was adopted because the practical objective of the control system is not only to maximize the total number of correct predictions, but also to reliably identify the moments when each actuator should be activated.

RESULTS

An initial same-time classification test was first performed, in which the sensor-derived variables and the actuator states corresponded to the same timestamp. This evaluation produced very high scores, with several F1-

score values close to or equal to 1.00. Although this confirmed that the models could reproduce the observed actuator states from the available sensor data, the results were considered overly optimistic for a practical control application. In such a configuration, the model may learn relationships that are too close to the current control logic, especially when aggregated variables directly related to control decisions are present in the input dataset. Therefore, the final evaluation was reformulated as a predictive task, where the sensor data recorded at time t were used to estimate the actuator states at t+30 minutes. This second configuration was considered more relevant for real greenhouse control, because it evaluates the ability of the models to anticipate future irrigation, ventilation and shading requirements. In addition, the most direct aggregated control indicators, including maximum air temperature, maximum relative humidity, maximum CO_2 concentration, maximum light intensity and minimum soil moisture, were removed from the input dataset to reduce the risk of direct threshold reproduction.

Random Forest classifier achieved strong predictive performance for irrigation and ventilation control. For irrigation, the model reached an F1-score of 0.9630, with a precision of 0.9811 and a recall of 0.9455. This indicates that the model was able to detect most future irrigation activation events while keeping false positive predictions low. Ventilation control also showed high performance, with an F1-score of 0.9611 and a recall of 0.9934, showing that most future ventilation ON states were correctly identified. Shading control was more difficult to predict, with an F1-score of 0.7894, mainly due to the lower recall of 0.6772 for the ON class. This suggests that Random Forest was conservative in predicting future shading activation, producing relatively reliable ON predictions but missing some actual shading events.

Table 3

Metrics for Random Forest classifier

Control	Accuracy	Balanced accuracy	Precision	Recall	F1-score	ROC-AUC
Irrigation	0.9962	0.9722	0.9811	0.9455	0.9630	0.9971
Ventilation	0.9311	0.7734	0.9309	0.9934	0.9611	0.9744
Shading	0.9027	0.8315	0.9461	0.6772	0.7894	0.9772

XGBoost classifier achieved high predictive performance on the 30-minute horizon dataset. For irrigation control, the model obtained an F1-score of 0.9533, with high precision and recall. For ventilation, the F1-score was 0.9606, very close to the Random Forest result. The best XGBoost result was obtained for shading control, where the model reached an F1-score of 0.8571, outperforming both Random Forest and MLP. This suggests that XGBoost captured more effectively the nonlinear relationships involved in shading decisions, especially those related to light intensity, temperature evolution and spatial variability within the greenhouse.

Table 4

Metrics for XGBoost classifier

Control	Accuracy	Balanced accuracy	Precision	Recall	F1-score	ROC-AUC
Irrigation	0.9953	0.9631	0.9808	0.9273	0.9533	0.9790
Ventilation	0.9311	0.8123	0.9437	0.9780	0.9606	0.9622
Shading	0.9254	0.8958	0.8843	0.8316	0.8571	0.9789

MLP model showed mixed performance. It achieved acceptable results for ventilation and shading, with F1-scores of 0.9455 and 0.8132, respectively. However, it failed to detect irrigation activation, obtaining an F1-score of 0.0000 for irrigation despite an accuracy of 0.9481. This indicates that the model mainly predicted the majority class for irrigation and did not correctly identify the ON state. The result confirms that accuracy alone is not sufficient for evaluating greenhouse control models when the class distribution is imbalanced. Balanced accuracy, recall and F1-score are more informative for this application.

Table 5

Metrics for MLP classifier

Control	Accuracy	Balanced accuracy	Precision	Recall	F1-score	ROC-AUC
Irrigation	0.9481	0.5000	0.0000	0.0000	0.0000	0.0444
Ventilation	0.9056	0.7836	0.9373	0.9538	0.9455	0.9187
Shading	0.8876	0.8943	0.7358	0.9088	0.8132	0.9480

The comparative evaluation was performed using the predictive dataset, where the sensor-derived variables recorded at time t were used to estimate the irrigation, ventilation and shading states at t+30 minutes.

This formulation provided a more realistic evaluation of the models for greenhouse control, since the objective was to anticipate future actuator states rather than reproduce the current control status. This result does not necessarily indicate that neural networks are unsuitable for greenhouse control in general, but that the tested MLP configuration was not robust enough for the irrigation target under the available dataset size and class distribution.

Table 6

Comparative results for the three investigated classifiers

Model	Accuracy mean	Balanced accuracy mean	Precision mean	Recall mean	F1-score mean	ROC-AUC mean
Random Forest	0.9433	0.8590	0.9527	0.8720	0.9045	0.9829
XGBoost	0.9506	0.8904	0.9363	0.9123	0.9237	0.9734
MLP	0.9138	0.7260	0.5577	0.6209	0.5862	0.6370

The predominance of tree-based ensemble models in the present study agrees with recent greenhouse research. *Mohammadi et al. (2024)* also reported XGBoost as the most balanced classifier for predicting greenhouse irrigation needs, while *Cebolla-Alemany et al. (2026)* found gradient-boosted tree models effective for short-term greenhouse microclimate forecasting. *Elvanidi and Katsoulas (2023)* obtained 91% accuracy and a 0.92 F1-score with Gradient Boosting, compared with 89% for both metrics with MLP, for greenhouse crop-stress detection. In the present study, XGBoost achieved a mean accuracy of 0.9506 and a mean F1-score of 0.9237 across the three actuator targets, whereas the tested MLP reached 0.9138 and 0.5862, respectively. These values should not be interpreted as a direct ranking across studies because the prediction targets, forecast horizons, datasets and validation protocols differ. The specific contribution of the present study is the direct 30-minute-ahead prediction of irrigation, ventilation and shading states from a distributed WSN, using a chronological data split and excluding the most direct aggregated control indicators to reduce data leakage.

To provide a more practical interpretation of the classification results, the predicted and actual actuator states were also compared over five days for a representative interval from the test dataset. The following figures show the 30-minute-ahead predicted states for the recommended model of each subsystem: Random Forest for ventilation, and XGBoost for irrigation and shading. The temporal representation confirms that the models generally followed the actual control transitions, especially for irrigation and ventilation. Shading control showed more deviations, which is consistent with the lower F1-score obtained for this subsystem and indicates that shading activation is more difficult to anticipate from the available sensor-derived variables.

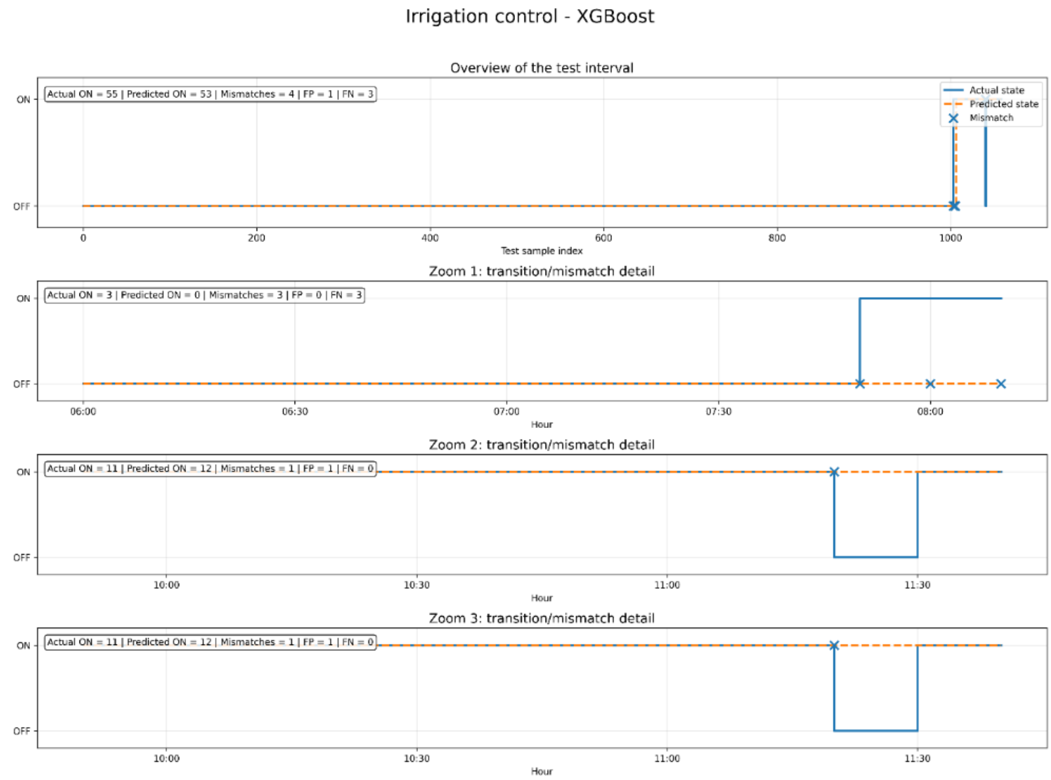

Fig. 2 - Temporal comparison between actual and predicted irrigation control states using the XGBoost model

Because irrigation control events were sparse and concentrated in a limited number of transition periods, a direct full-window visualization was not sufficiently informative. Therefore, the irrigation results were presented using an overview panel together with zoomed transition windows highlighting the agreement between actual and predicted states and the few mismatch cases. The upper panel provides an overview of the test interval, while the lower panels show zoomed details around transition and mismatch events.

For irrigation control, the temporal representation was adapted to the sparse occurrence of ON states. Therefore, an overview of the full test interval was combined with zoomed views around transition and mismatch events. The XGBoost model produced 53 predicted ON states compared with 55 actual ON states, with only four mismatches in the test interval. Most errors occurred near transition moments, indicating that the model captured the dominant irrigation state but remained sensitive to the exact timing of ON/OFF changes.

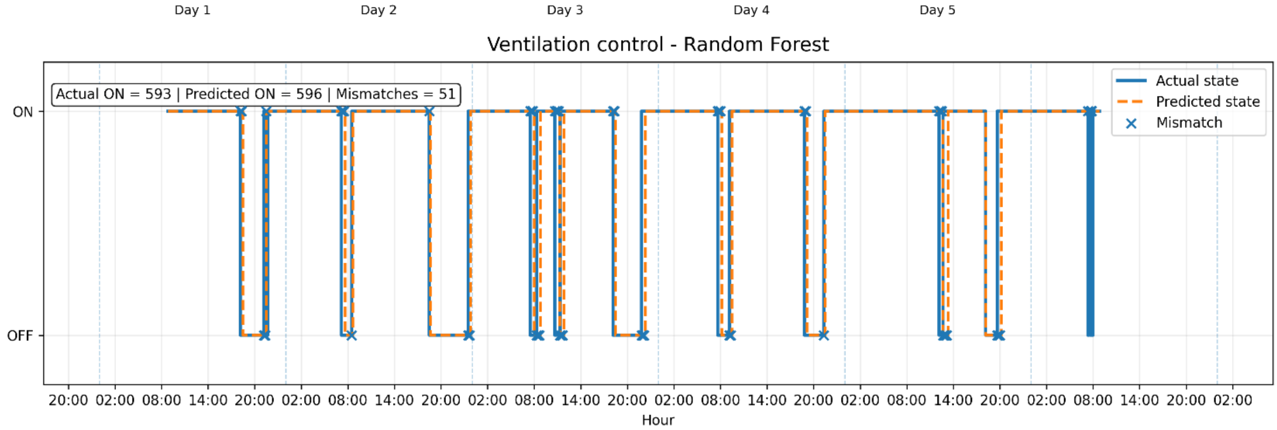

Fig. 3 - Actual and predicted ventilation control states obtained with the Random Forest model over a representative five-day test interval

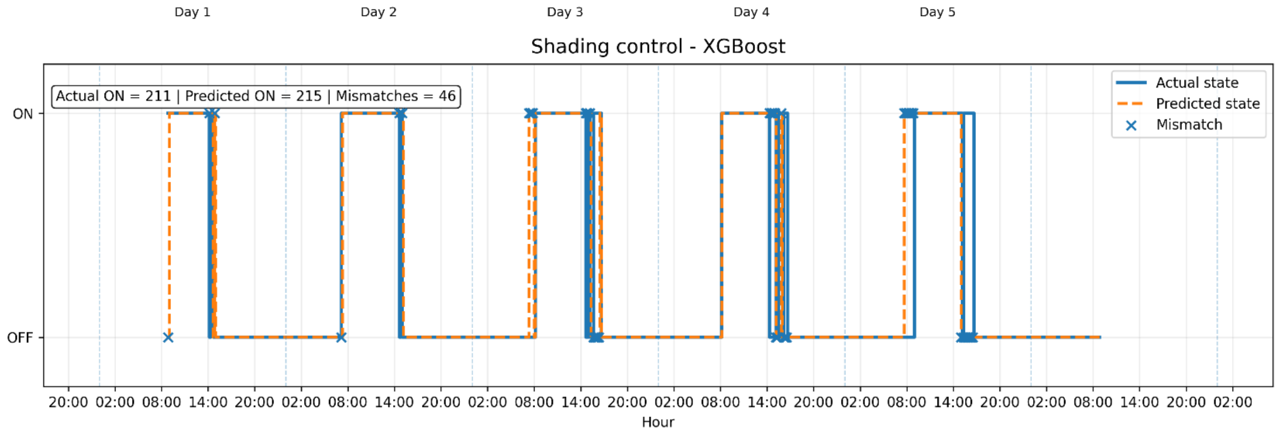

Fig. 4 - Actual and predicted shading control states obtained with the XGBoost model over a representative five-day test interval

The temporal comparison shows that the predicted states generally followed the actual actuator states. Most mismatches occurred close to OFF/ON or ON/OFF transitions, indicating that the main difficulty was not the identification of the dominant control state, but the exact timing of state changes in the 30-minute-ahead prediction task. This behaviour is particularly visible for shading control, where activation depends on rapidly changing light and temperature conditions.

Although Random Forest obtained a slightly higher F1-score for irrigation, XGBoost was retained as the recommended irrigation model because it provided a high and balanced performance and is consistent with the single-model implementation scenario.

Table 7

Recommended classifier for each investigated process

Control	Recommended model	F1-score	Balanced accuracy
Irrigation	XGBoost	0.9533	0.9631
Ventilation	Random Forest	0.9611	0.7734
Shading	XGBoost	0.8571	0.8958

Overall, XGBoost provided the best average predictive performance, while Random Forest offered the best performance for ventilation and remained highly interpretable. For a single-model implementation, XGBoost can be selected due to its highest average F1-score and balanced accuracy. For a subsystem-specific implementation, Random Forest is recommended for irrigation and ventilation control, while XGBoost is recommended for shading control.

CONCLUSIONS

This paper presented a machine learning-based decision model for greenhouse irrigation, ventilation and shading control using multi-sensor data collected from four WSN nodes. The raw dataset was cleaned, scaled, synchronized and transformed into a structured dataset suitable for supervised learning. To obtain a more realistic control formulation, the final evaluation was performed as a 30-minute-ahead prediction task, where sensor-derived variables recorded at time t were used to estimate the future control states at t+30 minutes.

The comparative evaluation showed that tree-based ensemble models were better suited to the processed greenhouse dataset than the tested MLP neural network. Random Forest remains a robust and interpretable alternative, with very strong results for irrigation and ventilation control. XGBoost achieved the highest average predictive performance and the best result for shading control, with an F1-score of 0.8571. The MLP model performed reasonably for ventilation and shading, but failed to detect irrigation activation, confirming that accuracy alone is not sufficient for evaluating imbalanced greenhouse control data.

Future work will first validate the selected model in shadow mode after integration with the actuator nodes, followed by supervised and closed-loop greenhouse trials. These experiments will assess control latency, false activations, missed activation events, robustness to sensor faults and seasonal distribution shifts, as well as the effects of predictive control on water and energy use. Model recalibration across crop cycles and greenhouse operating conditions will also be investigated.

ACKNOWLEDGEMENT

This work was supported by a grant of the Ministry of Education and Research, CCCDI - UEFISCDI, project number PN-IV-P7-7.1-PTE-2024-0359, within PNCDI IV.

REFERENCES

[1] Aborujilah, A., Al-Sarem, M., Abu-Zanona, M.A., 2025. Forecast-driven climate control for smart greenhouses: Energy optimization using LSTM model. *Energies*, 18(21), article 5821. DOI: https://doi.org/10.3390/en18215821

[2] Ait Oussous, S., Lail, D.M., El Bouayadi, R., Amine, A., 2025. Deep learning innovations for greenhouse climate prediction: insights from a Spanish case study. *IEEE Access*, 13, pp. 64810-64821. DOI: https://doi.org/10.1109/ACCESS.2025.3559232

[3] Ardiansah, I., Calibra, R.G., Bafdal, N., Bono, A., Suryadi, E., Nurhasanah, S., 2022. An IoT-enabled design for real-time water quality monitoring and control of greenhouse irrigation systems. *INMATEH – Agricultural Engineering*, 68(3), pp. 417-426. DOI: https://doi.org/10.35633/inmateh-69-39

[4] Buono, V., Mastroleo, M., Lucchi, C., D’Amato, G., Manfrini, L., Morandi, B., 2022. Field-testing of a decision support system (DSS) to optimize irrigation management of kiwifruit in Italy: a comparison with current farm management. *Acta Horticulturae*, 1335, pp. 355-362. DOI: https://doi.org/10.17660/ActaHortic.2022.1335.44

[5] Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), pp. 5-32. DOI: https://doi.org/10.1023/A:1010933404324

[6] Cai, W., Wei, R., Xu, L., Ding, X., 2022. A method for modelling greenhouse temperature using gradient boost decision tree. *Information Processing in Agriculture*, 9(3), pp. 343-354. DOI: https://doi.org/10.1016/j.inpa.2021.08.004

[7] Cebolla-Alemany, J., Cheng, Y., Pinto-Espín, L., Albano, M., Macarulla, M., Gasso-Domingo, S., 2026. Short-term forecasting and predictive control of rooftop greenhouse microclimate using multi-horizon machine learning models. *Smart Agricultural Technology*, 13, article 101733. DOI: https://doi.org/10.1016/j.atech.2025.101733

[8] Chen, S., Yang, D., Tang, Y., Chen, X., Li, Y., 2025. A review of environmental control strategies and models for greenhouses. *Sensors*, 25(5), article 1388. DOI: https://doi.org/10.3390/s25051388

[9] Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794. DOI: https://doi.org/10.1145/2939672.2939785
[10] Dhanaraju, M., Chenniappan, P., Ramalingam, K., Pazhanivelan, S., Kaliaperumal, R., 2022. Smart farming: Internet of Things (IoT)-based sustainable agriculture. *Agriculture*, 12(10), article 1745. DOI: https://doi.org/10.3390/agriculture12101745
[11] Elvanidi, A., Katsoulas, N., 2023. Machine learning-based crop stress detection in greenhouses. *Plants*, 12(1), article 52. DOI: https://doi.org/10.3390/plants12010052
[12] Föll, S., McCormick, R.J., Biegert, K., 2026. Three-year evaluation of irrigation strategies on apple yield and quality in Germany. *Scientia Horticulturae*, 359, article 114764. DOI: https://doi.org/10.1016/j.scienta.2026.114764
[13] Kochhar, A., Kumar, N., 2019. Wireless sensor networks for greenhouses: An end-to-end review. *Computers and Electronics in Agriculture*, 163, article 104877. DOI: https://doi.org/10.1016/j.compag.2019.104877
[14] Mohammadi, M.N., Takahara, T., Sokout, H., 2024. Optimizing water consumption and improving productivity in Afghanistan's greenhouses through IoT and machine learning. *Proceedings of the 2024 11th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*, Malmö, Sweden, pp. 185-191. DOI: https://doi.org/10.1109/IOTSMS62296.2024.10710304
[15] Nenciu, F., Matache, M.G., Gageanu, I., Persu, I.C., Marin, F.B., Voicea, I.F., 2026. Emerging technologies for soil evaluation using spectrometry sensing, Internet of Things and machine learning. *Applied Sciences*, 16(9), article 4559. DOI: https://doi.org/10.3390/app16094559
[16] Săcăleanu, D.-I., Matache, M.-G., Roșu, Ș.-G., Florea, B.-C., Manciu, I.-P., Perișoară, L.-A., 2024. IoT-enhanced decision support system for real-time greenhouse microclimate monitoring and control. *Technologies*, 12(11), article 230. DOI: https://doi.org/10.3390/technologies12110230
[17] Tan, K., Sun, S., Du, L., Zhou, S., 2021. Design of an intelligent irrigation system for a jujube orchard based on IoT. *INMATEH – Agricultural Engineering*, 63(1), pp. 189-198. DOI: https://doi.org/10.35633/inmateh-63-19
[18] Tsai, Y.-Z., Hsu, K.-S., Wu, H.-Y., Lin, S.-I., Yu, H.-L., Huang, K.-T., Hu, M.-C., Hsu, S.-Y., 2020. Application of Random Forest and ICON models combined with weather forecasts to predict soil temperature and water content in a greenhouse. *Water*, 12(4), article 1176. DOI: https://doi.org/10.3390/w12041176
[19] Ullah, I., Fayaz, M., Aman, M., Kim, D., 2022. An optimization scheme for IoT based smart greenhouse climate control with efficient energy consumption. *Computing*, 104(2), pp. 433-457. DOI: https://doi.org/10.1007/s00607-021-00963-5
[20] Wang, W., Mengoli, D., Sun, S., Manfrini, L., 2026. Integration of X-Ray CT, Sensor Fusion, and Machine Learning for Advanced Modeling of Preharvest Apple Growth Dynamics. *Sensors*, 26(2), article 623. DOI: https://doi.org/10.3390/s26020623
[21] Wang, Z., Liu, Z., Yuan, M., Yin, W., Zhang, C., Zhang, Z., Hu, X., 2025. A machine learning-based irrigation prediction model for cherry tomatoes in greenhouses: Leveraging optimal growth data for precision irrigation. *Computers and Electronics in Agriculture*, 237, article 110558. DOI: https://doi.org/10.1016/j.compag.2025.110558