technologies

MDPI

Article

Application of a Hybrid Model for Data Analysis in Hydroponic Systems

Kuanysh Bakirov [1], Jamalbek Tussupov [1,*], Akhmet Tussupov [2,*], Ibraheem Shayea [3,4] and Aruzhan Shoman [2]

[1] Faculty of Information Technology, L. N. Gumilyov Eurasian National University, Astana 010000, Kazakhstan; bakir.kuanysh@gmail.com
[2] Research and Innovation Center "AgroTech", Astana IT University, Astana 010000, Kazakhstan; a.shoman@astanait.edu.kz
[3] Electronics and Communication Engineering Department, Faculty of Electrical and Electronics Engineering, Istanbul Technical University, Sarıyer 34467, Turkey; ibr.shayea@gmail.com
[4] Department of Intelligent Systems and Cybersecurity, Astana IT University, Astana 010000, Kazakhstan
* Correspondence: tussupov@mail.ru (J.T.); akhmet.tussupov@astanait.edu.kz (A.T.)

Abstract: This study presents a hybrid data analysis approach to optimize the growing conditions for beetroot and tarragon microgreens cultivated in hydroponic systems. Maintaining precise microclimate control is essential, as even minor deviations can significantly affect the yield and product quality, but traditional monitoring methods fail to adapt promptly to changing conditions. To overcome this limitation, an automated monitoring system integrating machine learning methods XGBoost 3.0.0, principal component analysis (PCA), and fuzzy logic was developed. The model continuously identifies the deviations in environmental parameters and recommends corrective actions to stabilize the growth conditions. Experimental evaluation demonstrated superior predictive performance by using XGBoost, achieving an accuracy and F1-score of 97.88%, ROC-AUC of 99.99%, and computational efficiency (training completed in 2.3 s), outperforming RandomForest and GradientBoosting algorithms. Real-time data collection was facilitated through IoT sensors transmitting readings via Wi-Fi every 5 s to a local server, accumulating approximately 17,280 records per day. The analysis highlighted air humidity, solution humidity, and temperature as critical influencing factors. This research confirms the developed system's effectiveness in intelligent hydroponic monitoring, with future work aimed at integrating IoT and IIoT technologies for scalable management across diverse crops.

Keywords: hybrid model; hydroponic systems; microgreens growth; machine learning; fuzzy logic; environmental parameters; automated monitoring; crop yield prediction; XGBoost; data analysis

Academic Editor: Piero Cosseddu

Received: 13 March 2025
Revised: 11 April 2025
Accepted: 14 April 2025
Published: 22 April 2025

Citation: Bakirov, K.; Tussupov, J.; Tussupov, A.; Shayea, I.; Shoman, A. Application of a Hybrid Model for Data Analysis in Hydroponic Systems. *Technologies* **2025**, *13*, 166. https://doi.org/10.3390/technologies13050166

1. Introduction

The growing global demand for environmentally sustainable agriculture has significantly increased the interest in hydroponic systems [1–3]. Hydroponics allows efficient crop production in controlled environments, addressing challenges like limited arable land and rapid urbanization by enabling cultivation in minimal space without traditional soil [4–6]. Among the various crops suitable for hydroponics, microgreens, particularly beetroot and tarragon, have gained popularity due to their high nutritional value and short growth cycle. A couple of weeks are needed for the microgreens to grow and be fully harvested. This is such a short span that any changes in the growing conditions,

even minor ones, can make a big difference in the result, taste, and nutritional value of the product [6–8]. Therefore, cultivating microgreens requires the meticulous management of several critical environmental parameters, such as temperature, air and solution humidity, lighting intensity, nutrient solution pH, and electrical conductivity. Even slight deviations from the optimal conditions can adversely affect the yield quality and productivity [7–11].

Internet of Things (IoT)-based monitoring solutions able to record environmental data in real time have lately brought advances in modern technologies [9,10]. For commercial-scale hydroponics, these systems provide scalability benefits; but, they also create massive, continuous data streams, which makes efficient analysis a major difficulty. Growers must thus routinely calibrate and control sensors, nutrient solutions, and illumination schedules—tasks that become more difficult in dense vertical farming or greenhouse activities. The traditional environmental monitoring methods often lack the precision and adaptability required for real-time adjustments, resulting in a reduced crop quality and higher operational costs [12]. Recent advancements [13] propose integrating machine learning and intelligent systems to overcome these limitations, enhancing the system responsiveness and efficiency. Previous studies [14,15] have explored hybrid machine learning models, including RandomForest, Support Vector Machines, K-nearest neighbors, and ensemble approaches, demonstrating significant improvements in prediction accuracy and system performance [13–16].

In another direction, machine learning approaches have shown growing importance for agricultural uses by using neural networks and gradient boosting to maximize inputs and anticipate the yield [17]. Notwithstanding these encouraging advances, the current methods usually fail to adequately capture the multivariate relationships among the environmental factors [17]. For instance, a temporary rise in humidity can reduce the negative effects of an elevated temperature or nutrient imbalance; yet, many single-model solutions ignore these complex interactions [18]. To close this discrepancy, this research aims to develop a hybrid data analysis model that combines fuzzy logic and machine learning, specifically XGBoost and principal component analysis (PCA), to automate the monitoring of the critical environmental conditions in hydroponic systems. The proposed system identifies parameter deviations in real time and suggests corrective actions, thus optimizing the growing environment for microgreens. The experimental findings confirm that the developed hybrid model achieves high prediction accuracy (Accuracy and F1-score of 97.88%, ROC-AUC of 99.99%) and offers rapid computational performance. Consequently, the approach promises enhanced productivity and sustainability in hydroponic agriculture, providing scalability for commercial applications.

By methodically combining approaches instead of depending on a single predictive model, the suggested approach seeks to provide better precision, computational efficiency, and useful insights, therefore broadening the application area for scalable, profitable hydroponics microgreen cultivation.

2. Materials and Methods

Starting with a review of pertinent work and the literature, this part highlights the current machine learning methods in hydroponics systems and points out areas of future research needs. It then describes how the important environmental parameters were acquired. Emphasizing the synergy among gradient boosting, principal component analysis, and fuzzy logic for real-time predictions, it then outlines the justification for the method choice. A quick overview of every method including fuzzy inference and decision-tree ensembles is given to help streamline their use in a hydroponics setting. The section ends with feature importance and selection details, which show how data-driven metrics guided the choice of the parameters most affecting microgreen development and environmental stability.

2.1. Related Work and Literature Review

Driven by the worldwide demand for resource-efficient, high-yielding agricultural systems, soilless cultivation techniques, especially hydroponics, have seen notable developments in recent years [18]. Hydroponics provides greater autonomy over environmental parameters including nutrient delivery, acidity, and temperature conditions by removing the dependability on conventional soil-based farming, so increasing both production and sustainability [19]. With an emphasis on studies [20–22] that combine modern data analysis approaches, this work investigates the major themes developing from up-to-date research and points out the main gaps. It also emphasizes the increasing relevance of real-time IoT solutions allowing for ongoing monitoring at brief reporting intervals.

Investigated for their efficiency in maximizing the yield and minimizing water use are a variety of hydroponics systems: Nutrient Film Technique (NFT), Deep Water Culture (DWC), Ebb and Flow, and drip-based approaches [23–27]. From leafy greens to root vegetables, these methods have shown that diverse crops can be grown in conditions free from soil-borne pests and erratic weather patterns. Despite the achievements, researchers [28] point out that data-driven optimization is still underused when trying to capture the interactions among several parameters (temperature, humidity, nutrient concentrations, pH, electrical conductivity). Although a study [29] shows the need for balancing elements like water quality and nutrient composition, not all projects include advanced machine learning methods. Conventional methods might just look at correlations or simple regression to see how pH or temperature affects yields [30]. Actually, these parameters are quite interdependent; changes in one factor like humidity may either reduce or magnify the effects of another, such as temperature stress [31]. Many times, approaches lacking in holistic analytics ignore these multivariate dependencies.

Analyzing tabular data effectively and strongly dealing with outliers and missing values [32] have shown especially promise for Extreme Gradient Boosting (XGBoost). While research [33] concentrates on artificial neural networks for real-time hydroponics management, they do not fully utilize XGBoost's advanced regularization to address overfitting, a crucial factor in agricultural datasets including natural noise and variability. Moreover, although XGBoost shines in handling vast amounts of data, few studies [30–33] methodically utilize it in hydroponics to control fast changing environmental conditions in short reporting periods. XGBoost, for example, can use PCA-derived feature sets to improve predictions, but this synergy is understudied in hydroponics even if real-world data include sensor readings from several modalities including water chemistry, lighting, and air temperature [33]. Table 1 lists the main conclusions and approaches applied in most recent research.

Table 1. Summary of methods and identified gaps from main related studies.

Ref.	Study Focus	Methods	Key Findings	Identified Gaps
[29]	Benchmarks several ML techniques (including XGBoost) on four different hydroponics setups	XGBoost, Linear Regression, DNN, Federated Split Learning	XGBoost produces promising yield predictions by efficiently handling tabular data in several hydroponic settings	Does not incorporate fuzzy logic interpretability or real-time short interval data
[30]	Reviews IoT-based hydroponics automation solutions and possible ML synergy	IoT architectures, Naive Bayes approach	Emphasizes controlling pH and nutrient levels, and the benefits of sensor-driven data	Does not use advanced ensemble techniques or PCA to handle high-dimensional sensor data

Table 1. *Cont.*

Ref.	Study Focus	Methods	Key Findings	Identified Gaps
[31]	Predicts nutrient intake in hydroponics using ML. Emphasizes K-nearest neighbors and ensemble techniques.	KNN, ensemble models	Shows that using data-driven approaches results in the improvement in the absolute crop growth rate	Mostly uses daily or weekly aggregates; does not include short-interval data or fuzzy logic
[32]	IoT-based method focusing on a low heavy-metal content for monitoring aquaponic (and partially hydroponics) nutrient solutions.	Linear SVM, feature selection approaches	Accomplishes real-time sensors' integration with an eye toward nutrient optimization	Restricted cooperation with advanced ML (such as XGBoost and PCA). Does not apply fuzzy logic; short intervals under explored
[33]	General conversation on IoT sensor data use for hydroponics plant monitoring. Machine learning applied in pH and temperature control	DNN-based classification/control logic	Highlights sensor-based decisions and real-time control for a nutrient solution.	Does not mention advanced ensemble or fuzzy layering for interpretability
[34]	Investigates a greenhouse hydroponics system including an ML control factor. Emphasizes short-interval monitoring to guarantee stable conditions	KNN algorithm, limited IoT-based data pipeline	Shows strong yield gains in constant surroundings	Fails to include short-interval integrated PCA or fuzzy logic interpretability
[35]	Suggests an IoT-based sensor–actuator feedback loop for pH, water level, and heavy metal monitoring	Basic threshold-based control	Achieves a stable aquaponic/hydroponics environment with low manual intervention required	There are no advanced ML component past threshold triggers. Though not extensively tested, short intervals are theoretically possible

High-frequency environmental variable measurement by IoT-based sensor networks provides an unparalleled dataset to optimize hydroponics [34]. Though some prototypes save sensor outputs in the cloud and allow remote user dashboards [34], many struggle with real-time information connection at second or sub-second intervals. Coarse-grained data logging (e.g., every 30 min) may miss important inflection points affecting the crop yield and quality in a dynamic system—where humidity, temperature, and nutrient flows can change rapidly [35]. Studies usually apply solely straightforward threshold-based control—adjusting the flow of nutrients or pH as indicators deviate from nominal ranges—even when IoT sensors provide data frequently [35]. Still mostly unexplored is the relationship among short-interval IoT data and sophisticated machine learning—including online or gradual learning approaches. Moreover, methods that do combine ML often focus on a single crop or a limited set of environmental conditions, so limiting their relevance to more general multi-crop or large-scale environments [35].

Offering effective resource use and high-quality yields, hydroponics is fast rising as the main answer to agricultural environmental constraints. Regardless of this development, one of the main challenges still is the lack of a consistent, end-to-end strategy. The combination of XGBoost, PCA, and fuzzy logic provides, specifically, an effective but underused paradigm:

- XGBoost generates quick, accurate predictions on tabular sensor data;

- PCA lowers dimensionality and noise;
- Fuzzy logic reads partial truths for interventions that would be friendly for growers.

Moreover, short-interval IoT data streams open the door to real-time control, but advanced online learning algorithms are rarely used. Dealing with this complexity calls for an integrated system that balances interpretability, real-time responsiveness, and robust performance. Such a system would drive hydroponics to new frontiers in sustainability, adaptability, and yield.

2.2. Dataset Collection

The data [27] were collected using IoT sensors that transmit measurements to a local server via Wi-Fi at a frequency of 5 s. As Figure 1 illustrates, the hydroponics system integrates temperature, humidity, light, pH, and electrical conductivity (EC) sensors connected to an ESP Arduino microcontroller. The microcontroller was programmed to transmit data via HTTP. The data received by the server were automatically saved in the CSV format, broken down by days, with a separate file containing 17,280 lines (one record every 5 s) collected for each day during a week for the experiment. This architecture allows for the real-time monitoring and detailed analysis and forecasting of the critical parameters. The structured storage of data in the CSV format provides the ease of subsequent processing, including statistical analysis, visualization, and machine learning to identify patterns and predict potential changes in the monitored environment.

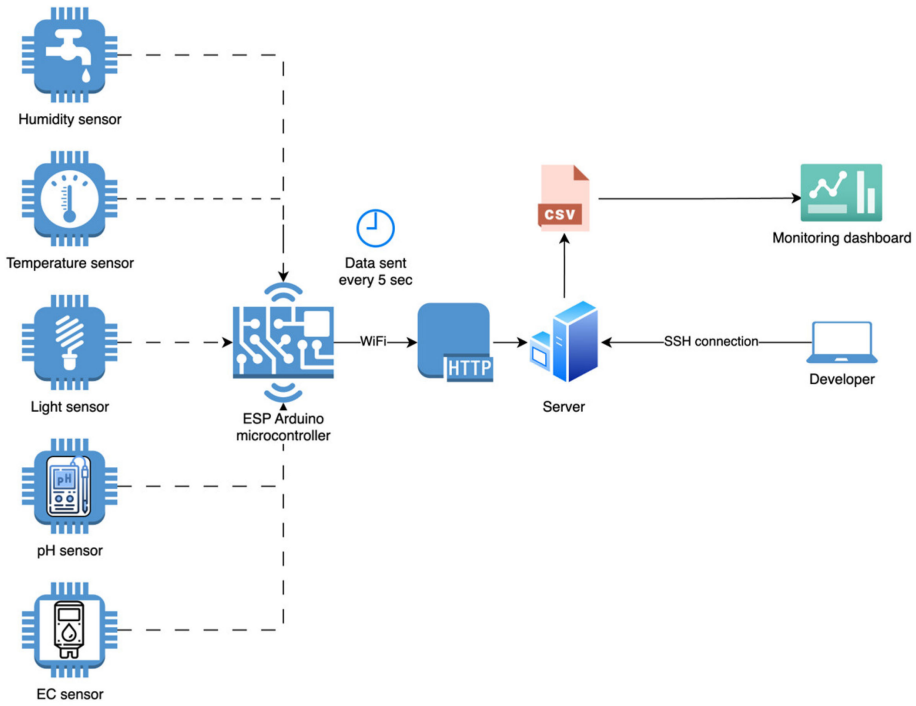

Figure 1. Overall architecture of data collection and storage in a hydroponics system.

The original data presented in the study are openly available in FigShare at https://doi.org/10.6084/m9.figshare.28667981.v1 (accessed on 26 March 2025).

2.3. Rationale for a Method Selection

A hybrid data analysis model developed by combining several machine learning [17] and data mining [18,19] methods was used for the study. Hybrid analysis combines several algorithms, each performing a specific role in data analysis. This approach combines four methods: XGBoost [20], StandardScaler [21], principal component analysis (PCA) [22–25],

and fuzzy logic [26]. The main goal of hybrid analysis is to identify the deviations of factors from the norm and assess the degree of their influence on microgreens (beetroot or tarragon). Each selected for purposes, the method combines XGBoost, principal component analysis (PCA), and fuzzy logic:

- XGBoost's great computational speed, efficient handling of tabular data, and robust performance in real-time or near-real-time scenarios helped it to be chosen in preference over other ensemble methods including RandomForest and GradientBoosting. Its built-in regularizing systems also reduce overfitting, making XGBoost fit for agricultural settings where environmental noise can be common;
- The main components are included mostly to preserve the necessary information while capturing the most important variance in the dataset. Although PCA is often used for noise reduction, in this framework, it also acts as a feature-ranking mechanism. XGBoost feature-importance measurements are combined with the resultant component loadings to produce a composite metric that retains the elements most pertinent to microgreen development;
- A layer of fuzzy logic helps to further fit the natural gradations in environmental parameters. Temperature deviations, for example, could be seen as "slightly above the norm" or "significantly above the norm", and fuzzy sets let one more finely classify these intermediate states. In hydroponics, where practitioners need clear, practical advice, this extra interpretability has great value;
- The algorithm presented in Figure 2 describes the process of a hybrid data analysis for monitoring and managing the microgreen growth parameters in hydroponic systems. At the beginning of the algorithm, data are prepared that include vital parameters such as the type of microgreens, temperature, air and solution humidity, illumination, solution pH, and solution conductivity. These parameters are necessary for subsequent analysis and presentation in numerical format, which is essential for operating machine learning algorithms. The integrated framework intended to evaluate new data points using XGBoost classification, feature-importance extraction, data scaling, PCA, and a hybrid scoring method based on fuzzy logic. There are several different phases to the process, each of which helps to provide a thorough evaluation of "negative" or undesirable conditions in a microgreen's dataset. Each phase of the pipeline is explained in the next paragraphs, which also help to clarify their interconnection:

    1. The pipeline starts with a dataset including several characteristics related to farming micro-greens. Usually, including temperature, air humidity, solution humidity, light intensity, pH, and electrical conductivity (EC) of the nutrient solution, these features also include an encoded category (type) denoting the microgreen variety. Furthermore, included in the dataset is a target label, referred to as negative_class, which indicates whether an observation reflects an unwelcome (negative) outcome. Every row represents one observation and captures the target label together with the variable measurements;
    2. Using temp, humidity_air, humidity_solution, light, ph_solution, ec_solution, and type_encoded as predictors and negative_class as the target variable, an XGBoost model is trained once the data are assembled. With strong predictive performance and easily understandable feature-importance measures, XGBoost is a gradient-boosting method fit for tabular datasets. Following the training process, a vector of importance coefficients is generated to show the relative contribution of every feature in target prediction;
    3. The raw data go through standardized scaling concurrently with training. Subtracting the mean and dividing by the standard deviation seen in the training set turns each feature to have almost zero mean and unit variance. This change

guarantees that, especially PCA, features with different units or scales do not dominate in the next analytical steps. The scaled features maintain the structure of the original data but change each dimension for more balanced contributions in later stages;

4. Designed to fit the microgreen type, the framework uses a "Norma Interval" block to save known or empirically based normal (acceptable) ranges for every feature. For example, kale microgreens might have a pH range that supports the best growth, or a temperature interval recommended for use. Later, these intervals—which represent lower (L) and upper (H) bounds for every feature—are used in a fuzzy-logic membership function to evaluate how well new observations fit these normative thresholds;
5. PCA is used for the data following standard scaling to lower the dimensionality and find the main components explaining the biggest variance. This stage produces PCA components together with a matching weight or loadings. These weights expose the degree of contribution that each original feature makes to a principal component. PCA aids in the capture of the necessary variance in a reduced-dimensional space by simplifying complicated, maybe correlated features into a set of orthogonal components. Complementing the XGBoost importance scores, the resulting PCA-based weights act as a second indicator of feature relevance;
6. The "Using hybrid analysis for new data" block, in which a hybrid score is computed for every feature, forms the essence of the framework. The technique specifically aggregates XGBoost importances, PCA weights, and fuzzy-logic membership values;
7. Using a Min–Max scaler, the resultant negative or risk scores are then scaled between 0 and 1. Since all scores now lie in a consistent range and allow comparisons across many features or observations, this last change provides a practical means to interpret the results. In the end, these accepted risk values can be shown in a bar chart or another kind of visual aid to offer an at-a-glance overview of the possible problems. This can be particularly helpful for rapidly spotting traits that significantly deviate from the expected conditions and so call for closer examination or corrective action.

Taken together, the figure presents a strong, multifarious framework for evaluating whether microgreen growing conditions (or related characteristics in other fields) point to a negative outcome. The pipeline provides a single score that balances the predictive power, statistical variance, and domain-specific acceptability thresholds by using the complementarity of XGBoost classification, feature importance analysis, PCA-based dimension weighting, and fuzzy logic-derived interval checks. This method can direct decisions by pointing out early hazards or anomalies before they become major concerns.

2.4. Feature Importance and the Selection Process

Hybrid analysis combines principal component analysis (PCA), gradient boosting (XGBoost), and fuzzy logic. This combination allows for considering various aspects of the data, including its variance, nonlinear dependencies, and impact on the target parameters, providing more complete and accurate modeling:

1. PCA finds the data's proportion of explained variance for individual features. PCA components mirror the primary directions of variability; feature weights are computed as the total of the weighted component coefficients. These weights show the general contribution of every feature to the structure of the data;

2. XGBoost ranks features in order of importance depending on lowering the model error. The obtained importance values show the degree of influence that every feature has on the correctness of model predictions;
3. Their product aggregates PCA feature weights and XGBoost feature importance. This enables one to take the local influence on the target variable (XGBoost) and global influence of features on the data variance (PCA) into consideration. The degree of negative impact of every feature is then computed via fuzzy logic. In fuzzy logic, membership functions define the degree of influence that the value of a feature deviates from the norm has.

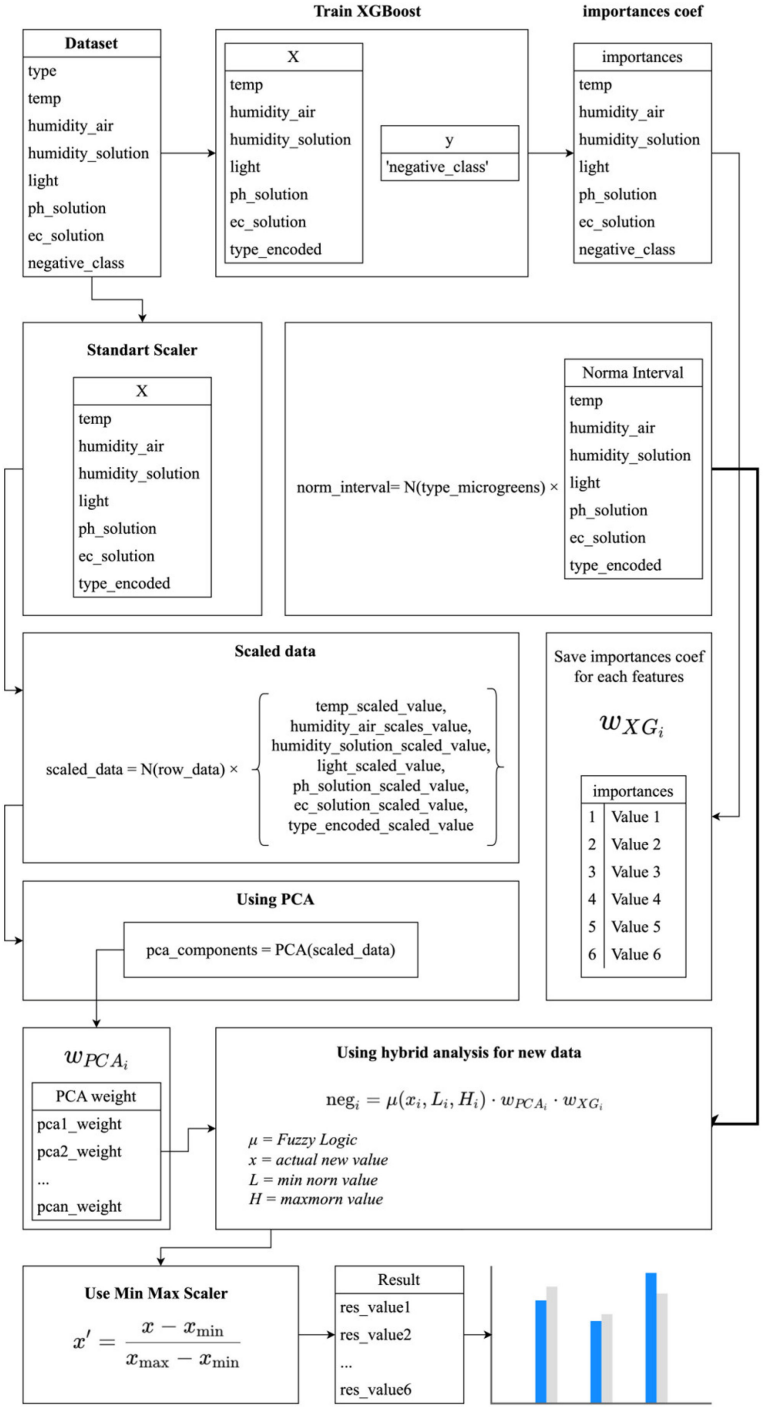

Figure 2. Algorithm for performing work.

To identify the "most significant features", a combined approach is used:

- The XGBoost model training produces an importance score for every feature, that is, an indicator mostly expressed by the reduction of the regression or classification error, made by the feature. These measures, for instance, offer insightful data on the parameters (e.g., temperature, humidity), which most affect the predictive performance of the model;
- Viewed after PCA application are all the loadings of the main components. According to it, all the main characteristics are covered with all their variation even until six are relined. Then, the weights of every feature are split according to their rank, which reflects the data's fluctuation;
- PCA brings XGBoost and together generates a "global importance index" via the complex ways of obtaining from PCA-direction loaded weights (eigenvalues) and the feature-importance scores of the XGBoost method. In the next stage, variance-based dimensional analysis is merged with predictive relevancy producing a one-dimensional measure;
- The pH levels of the nutrient solution, which aside from the innovation of the use of the modern working model in the development of plants, are still relevant to the sustaining of nutrients, and hence, they must still be included in the final model. This mixed approach seeks to cover the mostly important elements with extra potential contributions through the control of the other features resulting from their natural influence on green development.

In the first stage, the XGBoost algorithm [26] selects the most significant features. Its task is to build a classification model based on decision trees, where the target variable y (negative_class) is predicted. The algorithm is trained with the prediction error minimization, and the result is a matrix of feature importances, $w = [w_1, w_2, \ldots, w_n]$, where each w_i corresponds to the importance of feature x_i. XGBoost is used to select the most significant features among the set X, which includes the following variables (1):

$$X = [x_{temp}, x_{humidity\_air}, x_{humidity\_solution}, x_{light}, x_{ph\_solution}, x_{ec\_solution}] \quad (1)$$

The target variable in this analysis is $y = negative\_class$, where each factor is assigned a class reflecting the degree of deviation from the norm. The XGBoost algorithm is trained on the dataset and calculates the importance of each feature. The result of the training is a matrix of feature importances (2):

$$w = [w_{temp}, w_{humidity\_air}, w_{humidity\_solution}, w_{light}, w_{ph\_solution}, w_{ec\_solution}] \quad (2)$$

where w_i is the importance of feature x_i. These values are used to select the most significant features, sorted by importance. Finally, the essential features from the set X are chosen based on the calculated weights w_i and are used in the following stages of the analysis. After selecting the significant features, they are standardized using the StandardScaler method. Standardization transforms each feature so that its mean becomes zero, and its standard deviation becomes one, using the following Formula (3):

$$X_{scaled} = \frac{X - \mu_X}{\sigma_X}, \quad (3)$$

where μ_X is the mean value of a feature, and σX is the standard deviation of a feature. This process helps to balance the influence of each feature in the subsequent stages of analysis, especially in scale-sensitive methods such as PCA. It ensures that no single feature dominates due to differences in units or ranges of values. PCA is a dimensionality reduction technique that transforms data into a more compact representation while preserving the essential information about the features. In this analysis, PCA was used to create new

features (principal components) without reducing the original dimensionality of the data. Six components were used, which allowed the preservation of the complete information about each feature and consider all the factors that affect the growth of microgreens. Each principal component, such as PCA1, PCA2, and others, reflects the variations in the data, and these components are used to calculate weights that are applied in the hybrid analysis. The weights help estimate how much each feature contributes to the deviations of the factors from normality based on the importance obtained using XGBoost. The application of PCA (4) in this case is not aimed at reducing the dimensionality but at redistributing the weights between features for a more accurate analysis of the microgreen growth conditions:

$$Z = X_{scaled} \times W_{PCA_i}, \tag{4}$$

where X_{scaled}—standardized features, W_{PCA_i}—matrix of eigenvectors for transformation, $Z = [z_1, z_2, \ldots, z_n]$—a set of new features corresponding to the original ones. When applying PCA, each z_i component reflects a portion of the total variation in the data. In this case, six components are used, each corresponding to one of the original features, which means that each z_i preserves and explains the portion of the data variance that was associated with the corresponding feature (5):

$$W_{PCA_i} = \frac{|z_i|}{|z_1| + |z_2| + |z_3| + |z_4| + |z_5| + |z_6|} \tag{5}$$

In the classical approach, the PCA1 and PCA2 components usually explain the most significant part of the feature variance together, which allows for a substantial dimensionality reduction. However, since the number of components is equal to the number of features in this analysis, the goal is not to reduce the dimensionality. Instead, the feature variations are redistributed across each element, allowing complete information to be preserved. This makes the data suitable for further processing, including fuzzy analysis, where each element is weighted based on its contribution to the overall data variance. A study uses fuzzy logic based on the principal components' weights and the features' significance. For each feature, the deviation from the norm is calculated using the membership function (6):

$$\mu(x_i) = \begin{cases} \min\left(1, \frac{L_i - x_i}{0.1\, L_i}\right), \; if \; x_i < L_i \\ 0, \; if \; L_i \leq x_i \leq H_i \\ \min(1, \; \frac{x_i - H_i}{0.1\, H_i}), \; if \; x_i > H_i \end{cases}, \tag{6}$$

where $\mu(x_i)$—degree of membership for parameter x_i, L_i and H_i—lower and upper limits of the normal range for parameter i, x_i—the parameter value for a given plant. The deviation of each feature is adjusted, considering the importance obtained in the previous stages, according to the following Formula (7):

$$neg_i = \mu(x_i) * W_{PCA_k} \times w_i, \tag{7}$$

where neg—the final deviation for feature i, W_{PCA_k}—the weight of the main component, w_i—the significance of the feature obtained in the analysis step using XGBoost. In the final step, the deviation results are normalized to 0 and 1 using MinMaxScaler 1.6.1 version. This transforms all the results to the same scale, where zero means that the parameter is within the normal range and 1 is the maximum deviation (8):

$$neg_i^{norm} = \frac{neg_i - \min(neg_i)}{\max(neg_i) - \min(neg_i)} \tag{8}$$

This normalization allows for the correct interpretation of deviations, where 0 means full compliance with the norm, and 1 is the maximum deviation. In hybrid analysis, each method performs its specialized task, and their interaction forms a multi-layered approach to data assessment, providing a more accurate and flexible understanding of the deviations in factors affecting microgreen growth. First, XGBoost is used to identify the significant features by analyzing all the data and creating a model based on gradient-boosted decision trees that classify the data by reference to the target class negative_class. This step allows us to reduce the number of features, highlighting those with the most significant impact on deviations from the norm. Next, the selected features are standardized using StandardScaler to equalize their mean values and variance, which is essential for the correct application of the principal component analysis (PCA), which is sensitive to the scale of the data. After standardization, PCA is used to redistribute the weights between the significant features, where each component reflects its contribution to the overall variability of the data, and these weights are then used for further assessment using fuzzy logic. Fuzzy logic helps estimate parameter deviations from regular intervals, where deviations are expressed in values from 0 to 1. Finally, MinMaxScaler is applied to normalize the fuzzy logic results, bringing all values into the range from 0 to 1, simplifying their interpretation and comparison. As a result, the hybrid analysis completes its process by providing a summary dataset, where each row contains information about the degree of deviation of each parameter from the norm in percentage values.

3. Results

When choosing a model to predict the target variables such as changes in temperature, humidity, and other factors, it is essential to consider the prediction accuracy and the model performance. The experiment tested three methods: RandomForest, GradientBoosting, and XGBoost. Each method was evaluated based on the mean squared error (MSE) and training time. The main goal was to identify a model that balances accuracy, speed, and computational efficiency.

3.1. Comparison Results of the Main Classification Methods

RandomForest demonstrated the lowest MSE values, for example, for Delta_temp (1.57×10^{-8}) and Delta_ph (1.78×10^{-10}), making it the leader in accuracy among the tested methods. However, the training time was 3 min, which makes this method less suitable for processing large amounts of data or tasks that require timeliness. Long training times can be a critical limitation in real-world settings, especially when retraining the model on new data. GradientBoosting showed acceptable accuracy with MSE, for example, for Delta_temp (0.0044) and Delta_ec (1.77×10^{-5}). Its training time was 1 min 16 s, which makes it more efficient than RandomForest. However, this method still requires a significant processing time and does not show clear performance advantages when working with large datasets. This limits its application in scalable and high-speed systems. XGBoost showed competitive MSE values within the acceptable accuracy range for Delta_temp (0.0107) and Delta_ph (0.0001288). However, its main advantage is the training time of only 2.3 s. Such high computational efficiency makes the method most suitable for problems that require fast data processing, scalability, and a low computational load. The speed of training and prediction allows XGBoost to be used for real-time problems where responsiveness is essential.

XGBoost was chosen due to its excellent training speed of 2.3 s, which is significantly faster than that of GradientBoosting and RandomForest. The difference in accuracy, measured via MSE, between XGBoost and other methods is minimal and does not significantly affect the quality of predictions. However, its high computational efficiency

makes XGBoost an optimal choice for scalable problems and real-world applications where processing speed, flexibility, and minimal computational costs are key. The importance of features extracted by using XGBoost showed that the most significant feature is humidity_air (0.125534), indicating a key role. Features such as humidity_solution (0.122299) and temp (0.110875) also demonstrated high importance, confirming their influence on the classification process. According to the model training results, XGBoost achieved a high accuracy (Accuracy: 0.9788, F1-score: 0.9788) and a near perfect ROC-AUC: 0.9999, indicating its ability to distinguish between classes accurately. Moreover, the code execution time on the final dataset was only 1 min 22 s, demonstrating its high computational efficiency compared to the other methods. This makes XGBoost an optimal choice for tasks involving large amounts of data, providing a balance between the processing speed and the importance of the feature determined by GradientBoosting, which suggests that the most significant feature was temp (0.180000), which emphasizes the critical role of temperature in the analyzed problem. This is followed by humidity_solution (0.150000) and humidity_air (0.140000), significantly impacting the model predictions. Parameters such as ph_solution (0.130000) and ec_solution (0.100000) also play an important role, showing the contribution of solution characteristics to the model's overall performance. Figure 3 illustrates the important features identified by the three models.

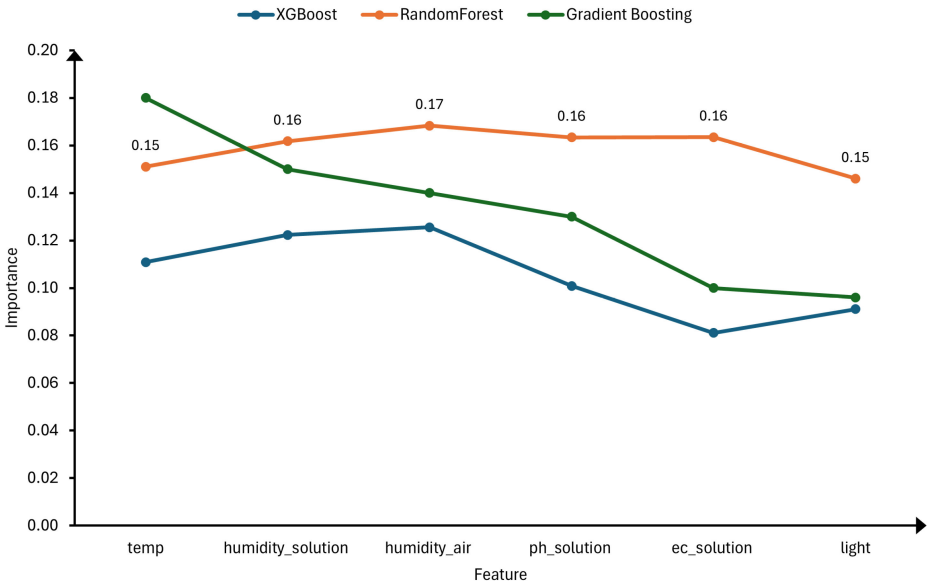

Figure 3. Result of feature importance calculation by using XGBoost, RandomForest, and GradientBoosting.

The GradientBoosting model showed promising results in terms of accuracy (Accuracy: 0.9343, F1-score: 0.9343, ROC-AUC: 0.989) and an acceptable Log Loss of 0.1006, indicating its ability to distinguish between classes. However, the execution time was 20 min 45 s, making it the least efficient among the tested methods regarding the computational cost. This makes GradientBoosting less suitable for tasks that require high-speed processing of big data. Table 2 compares the three classification algorithms, RandomForest, GradientBoosting, and XGBoost, using the key metrics—Accuracy, F1-score, ROC-AUC score, Log Loss, and execution time. The RandomForestClassifier showed the best accuracy (Accuracy: 0.9993) and minimal Log Loss (0.0906), but it was slower (1 min 35 s). GradientBoosting showed acceptable accuracy (Accuracy: 0.9343) but was the least time-efficient (20 min 45 s). XGBoost's balanced accuracy (Accuracy: 0.9788, ROC-AUC: 0.9999) and execution speed (1 min 22 s) make it the most efficient choice.

Table 2. Comparison of efficiency and performance of three methods.

Metric	RandomForest	GradientBoosting	XGBoost
Top Feature	humidity_air (0.16)	temp (0.18)	humidity_air (0.12)
Second Feature	ec_solution (0.16)	humidity_solution (0.15)	humidity_air (0.12)
Third Feature	ph_solution (0.16)	humidity_air (0.14)	humidity_solution (0.12)
Accuracy	0.9993	0.9343	0.9788
F1-score	0.9993	0.9343	0.9788
ROC-AUC score	0.9993	0.989	0.9999
Log Loss	0.0906	0.1006	0.1846

The model's adaptability was tested on an initial dataset that included only two microgreen species, beetroot and tarragon. The data for other species, such as radish, basil, mustard, watercress, spinach, parsley, and cilantro, were added to increase the versatility and validate the model. Using this approach, the tested models, RandomForest, GradientBoosting, and XGBoost, demonstrated how well they cope with changing environmental conditions, including new species and variations in temperature, humidity, light, pH, and conductivity. If necessary, the model easily adapts to removing or adding data for specific microgreen species, confirming its flexibility and versatility. The performance metric shows that each model demonstrated its adaptability. XGBoost achieved an accuracy of 0.9752, F1-score of 0.9746, ROC-AUC score of 0.9994, and minimum Log Loss of 0.1863 with a training time of 30 s, making it the best performer in regard to speed and accuracy. RandomForest achieved the best accuracy (Accuracy: 0.9984, F1-score: 0.9984, ROC-AUC score: 0.9996) with the minimum Log Loss (0.0912), but it took 1 min 25 s to train, which slightly reduces its applicability to large datasets. GradientBoosting achieved decent results with an accuracy of 0.9412, F1-score of 0.9407, and ROC-AUC score of 0.9901, but the training time was 4 min 33 s, making it less suitable for scalable problems. These results confirm that all the models effectively adapt to changing data and different environments, maintaining a high accuracy when considering the diversity of the input features. However, XGBoost stands out due to its speed and ability to work with large amounts of data. It is the most suitable choice for problems with dynamic changes and increasing data complexity.

3.2. Comparative Analysis with Other Modern Techniques

A comparison was made outside RandomForest, GradientBoosting, and XGBoost to include a basic feedforward neural network (as a representative deep-learning-based model) and LightGBM (an alternative boosting framework), so validating our hybrid approach. Every model was trained with identical data splits, so guaranteeing a consistent basis for evaluation. Table 3 highlights the trade-offs among speed, accuracy, and interpretability by aggregating the performance measures and training times for all five techniques.

Table 3. Performance results of the models.

Model	Accuracy	F1-Score	ROC-AUC	Log Loss	Training Time
RandomForest	0.9984	0.9984	0.9996	0.0912	1 min 25 s
GradientBoosting	0.9412	0.9407	0.9901	0.2451	4 min 33 s
XGBoost (Proposed Core)	0.9752	0.9746	0.9994	0.1863	30 s
Feedforward Neural Network	0.9500	0.9510	0.9950	0.2785	3 min 10 s
LightGBM	0.9690	0.9685	0.9980	0.2102	35 s

RandomForest comes with a rather longer training time (1 min 25 s), yet as Table 3 shows, it achieves the lowest Log Loss (0.092) and the highest accuracy (0.9984). Although gradient boosting offers a good accuracy (0.9412), it is not very suitable for real-time or highly dynamic situations since it takes over four minutes to train. Although both LightGBM and the feedforward neural network show competitive performances—accuracy of 0.9500 and 0.9690, respectively—their training times and tuning complexity can rise dramatically with increasing datasets. Concurrent with the strong interpretability and robustness across diverse hydroponics conditions, the proposed XGBoost-based hybrid approach offers an outstanding trade-off: it maintains a high accuracy (0.9752) and one of the fastest training times (30 s). For real-time hydroponics and adaptive control, the hybrid approach is especially appealing even if deep learning or alternative ensembles can match or surpass some accuracy measurements through speed, clarity, and computational efficiency.

3.3. Impact and Relationships of the Environmental Parameters

During the research on the cultivation of microgreens, the key factors and optimal intervals that ensure their health and development were identified. Specific standard parameters were established for two types of microgreens, beetroot and tarragon, which include temperature, air and solution humidity, illumination, pH level, and solution electrical conductivity (EC). The following parameters are considered optimal for beetroot. The temperature should be 18–24 °C, supporting active photosynthesis and cell division processes. An air humidity of 50–60% helps to avoid excessive transpiration and maintain a balanced water regime. Solution humidity in the range of 0.8–1.2 maintains the necessary ratio of water and nutrients for the root system. Illumination should be 12–16 h a day to ensure effective photosynthesis and the accumulation of biomass. A pH level of 6.0–6.5 provides the availability of nutrients, and an electrical conductivity of 1.0–1.4 EC helps plants receive enough nutrition without the risk of salt stress. For tarragon, the optimal conditions are slightly different. The temperature should be 20–25 °C, which promotes the accumulation of biologically active substances, such as essential oils. An air humidity of 60–70% maintains healthy leaves, reducing the risk of dehydration. The moisture of the solution, as for beets, is 0.8–1.2, providing optimal conditions for the root system. Tarragon requires 12–16 h of light for active growth and development. The pH level is preferably within 5.5–6.0, which creates favorable acidity for the absorption of nutrients. The electrical conductivity of the solution for tarragon is slightly higher at −1.2–1.6 EC, which is due to its increased need for nutrients.

Each of these factors has a significant impact on plant growth and health. The temperature determines the rate of photosynthesis and cell growth, and its deviations can lead to heat stress. The air humidity affects plants' transpiration and water balance: too low humidity causes dehydration, and too high humidity promotes the development of fungal infections. Solution humidity regulates the water supply to the root system, and its balance is necessary to prevent root hypoxia. Light provides plants with the energy for photosynthesis, and its deficiency or excess can lead to slow growth or photostress. The pH controls the availability of nutrients, and electrical conductivity (EC) reflects the concentration of these substances in the solution—its imbalance can lead to nutrient deficiencies or salt stress. Thus, compliance with these parameters provides optimal conditions for growing microgreens and promotes healthy development. Table 4 demonstrates the optimal growing conditions of two types of microgreens: beetroot and tarragon. The data were obtained via sensors measuring the key parameters such as temperature, air and solution humidity, light, pH level, and solution conductivity. For each type of microgreen, the table presents the corresponding values for these parameters. For example, for beets, the temperature is 20.6 °C, and the air humidity is 54.5%, and for tarragon, it is 23.5 °C and

64.9%, respectively. These values help to assess the conditions maintained in the growing system (grow box) for each type of plant.

Table 4. Optimal growing conditions for the beetroot and tarragon.

Type	Beetroot	Tarragon
temp	20.60482	23.52371
humidity_air	54.54698	64.97496
humidity_solution	0.941568	1.052053
light	14.36484	13.99901
ph_solution	6.206704	5.675794
ec_solution	0.506409	2.248434
type_encoded	0	1
negative_class	1	65

The key element of the dataset is the negative_class column, which reflects the number of parameter deviations from the norm. This column indicates how many factors are outside the optimal values for each plant type. If all the parameters are within the norm, the deviation class will be 0. For example, if only temperature deviation is recorded for beetroot, the deviation class will be (1). If deviations are recorded for two parameters, such as temperature and air humidity, the class will be (2). For tarragon, given six parameters, the deviation class can take values of up to 65 if all the parameters are simultaneously outside the norm. This approach to classifying deviations allows for the creation of different combinations of factors that can be used for automated monitoring and forecasting in microgreen management systems. For example, class 1 for beetroot indicates a deviation of one factor, such as temperature, while class 65 for tarragon indicates multiple deviations in several parameters simultaneously. This classification mechanism enables the more accurate and flexible management of the growing conditions, which is especially important for automated monitoring systems. Figure 4 shows a correlation matrix showing the relationship between the critical parameters affecting microgreen growth, such as temperature (temp), air humidity (humidity_air), solution humidity (humidity_solution), light (light), solution pH (ph_solution), and solution electrical conductivity (ec_solution). The color scale on the right shows the degree of correlation: warm shades of red indicate a positive correlation, and blue shades indicate a negative correlation. The correlation values range from -1 (strong negative correlation) to 1 (strong positive correlation), with 0 indicating no relationship between the variables.

Temperature (temp) has a weak positive correlation with air humidity (0.11) and solution electrical conductivity (0.056), and its influence on the other parameters, such as solution substrate, light, and pH, is minimal. Air humidity (humidity_air) shows a negative correlation with solution pH (-0.35), indicating that increasing humidity can decrease the pH and has a moderate positive correlation with electrical conductivity (0.2). Solution humidity (humidity_solution) has virtually no significant correlation with the other parameters, except for a weak relationship with light and electrical conductivity, indicating a low interdependence. Light has a weak positive correlation with solution humidity, possibly due to chance. Solution pH (ph_solution) shows a noticeable negative correlation with air humidity and a weak negative correlation with temperature and electrical conductivity. The electrical conductivity of the solution (ec_solution) is positively correlated with air humidity, which may indicate the need to adjust the composition of the solution at high humidity to maintain the optimal conditions. Figure 5 shows a mutual information analysis diagram that reflects the relationship between factors such as air humidity, solution pH, temperature, and the target attribute, negative_class, representing the likely negative impact of parameter deviations on the health of microgreens. Mutual

information determines how much information one feature contains about another, in this case, the target variable. This indicator helps to understand how strongly each factor, such as air humidity or pH, relates to the target class and its significance for predicting the health of microgreens.

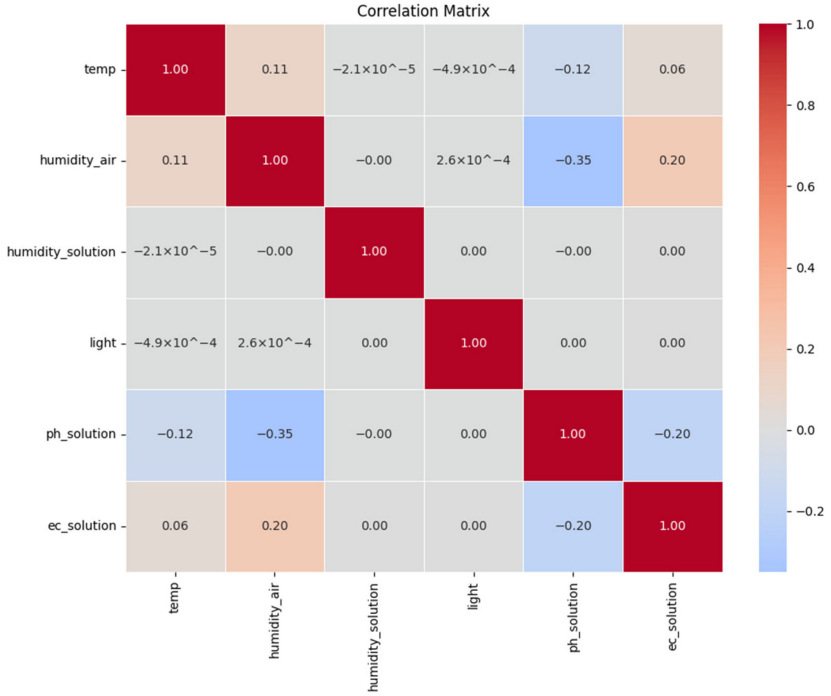

Figure 4. Relationship between critical parameters affecting the growth of microgreens.

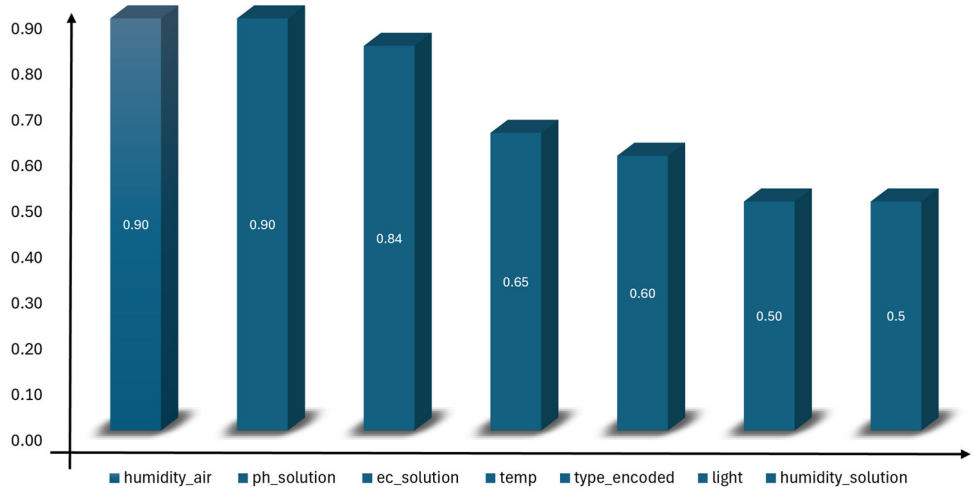

Figure 5. How each factor relates to the target class.

Based on Figure 5, it can be concluded that the humidity_air and pH of the solution (ph_solution) have the most significant impact on the target class, indicating that these parameters are strongly associated with negative deviations in the condition of microgreens, affecting their health. The electrical conductivity of the solution (ec_solution) and temperature (temp) also play an essential role, but their influence is slightly smaller com-pared to the first two parameters. Plant type encoding (type_encoded) also shows a significant value, which may indicate differences in the sensitivity of different plant types, such as beetroot and tarragon, to parameter deviations. Light and humidity_solution show the lowest mutual information, indicating their relatively low contribution to predicting

negative deviations. This analysis helps to highlight the key factors to consider when building predictive models for microgreen health analysis, which can further im-prove the accuracy of machine learning models. Figure 6 shows the feature importance diagram, indicating the factors significantly impacting the prediction of the target class negative_class. As a result of the analysis, each feature was rated according to its importance for accurately predicting the negative impact on microgreens.

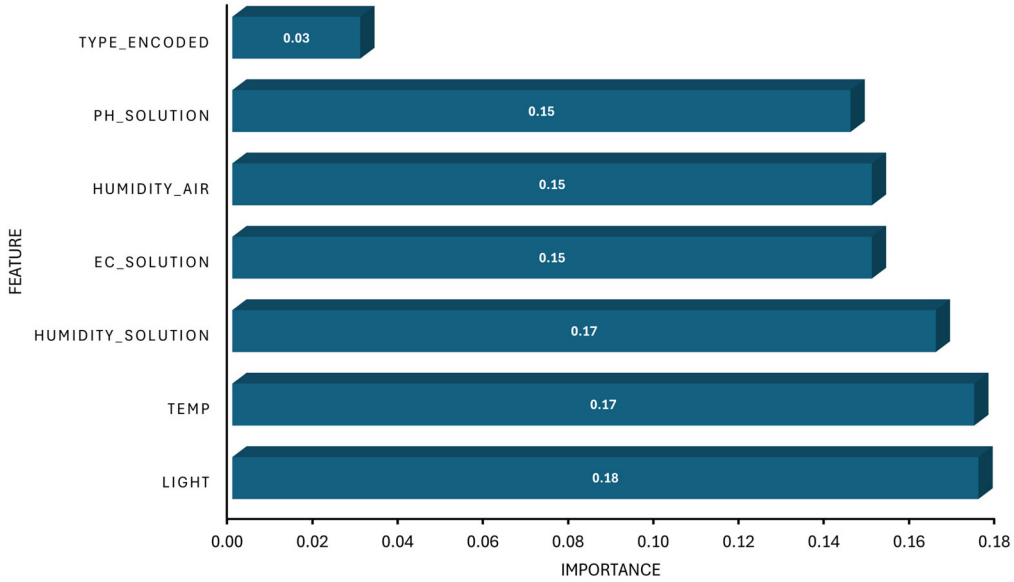

Figure 6. Feature importance diagram.

After training, the model rated the importance of each feature based on how often and how strongly it influences the decisions in each tree. The more a feature helps reduce the prediction error, the higher its importance. The diagram shows that light and temperature were the most important for predicting the negative class, as they significantly influenced the model's decision on the deviations that affect microgreen growth. Next were humidity_solution and ec_solution, which had a significant but slightly smaller influence. The least important feature was the encoded plant type feature, which may indicate a minor impact on the overall parameter deviations compared to other factors. This feature importance analysis provides valuable information about which parameters are critical to control to maintain the optimal microgreen growth conditions and prevent negative deviations. Table 5 shows the initial data for beets, where the key parameters are listed: temperature, air humidity, solution moisture, light, solution pH, and solution electrical conductivity (EC). These data are compared with the optimal ranges for each parameter. For example, the temperature for beets is 26 °C, which is 2 °C above the norm (18–24 °C). This deviation is then evaluated in the second table, which shows the predicted deltas (deviations from the norm) and their negative impact on the plant. In this case, temperature has a 100% negative impact on beets, while the other parameters, such as air humidity and solution pH, are within the norm and do not have a negative impact.

Table 5. Initial data for beets.

Input Data: Beet (Normal Conditions)			
Parameter	Value	Unit	Norm
Temperature	26	°C	18–24
Air Humidity	60	%	50–60
Solution Humidity	1.0	Ratio	0.8–1.2
Light	13	Hours	12–16
pH of Solution	6.1	pH	6.0–6.5
EC of Solution	1.3	EC	1.0–1.4
Deltas and Negative Impact			
Parameter	Delta	Unit	Negative Impact
Temperature	2.000	°C	100.00%
Air Humidity	0.008	%	0.00%
Solution Humidity	0.000	Ratio	0.00%
Light	0.000	Hours	0.00%
pH of Solution	0.000	pH	0.00%
EC of Solution	0.000	EC	0.00%

Figure 7 includes boxplots showing the distribution of values for parameters such as temperature, air humidity, solution moisture, illumination, solution pH, and solution electrical conductivity (EC). Red dots indicate abnormal values; for example, 15 °C and an air humidity of 76% are outside the permissible values. Solution moisture, illumination, pH, and electrical conductivity are within the normal range, as evidenced by green dots.

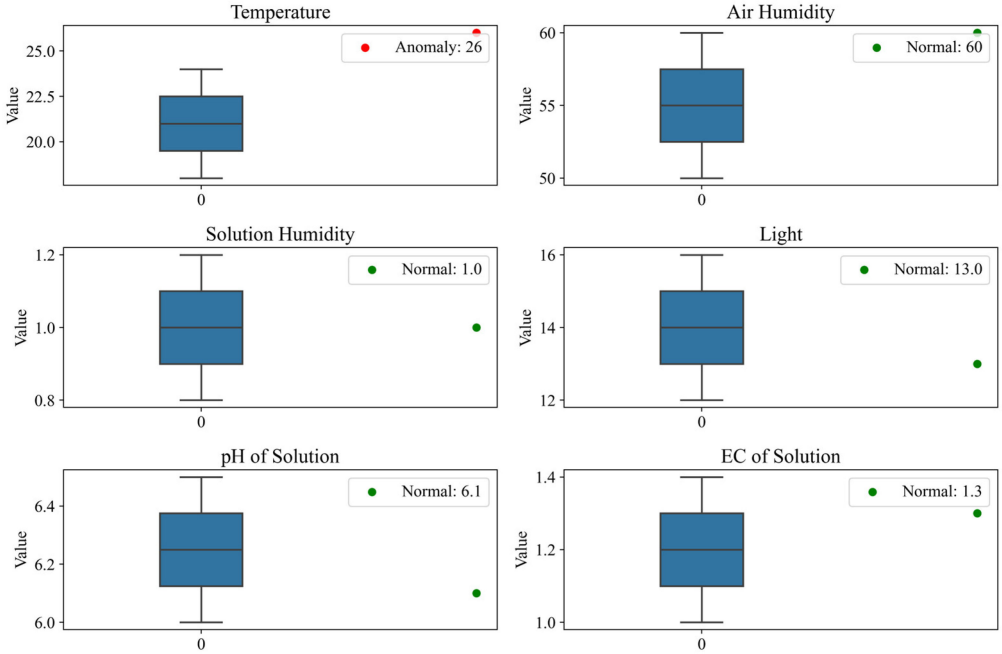

Figure 7. Visualization of microclimate parameter deviations for growing beet microgreens.

Table 6 shows a similar analysis process but with a changed value of air humidity (65%), which is 5% above the upper limit of the norm. This change in substrate resulted in a 100% negative impact on the plant. However, even though the temperature remained above the norm, its negative impact decreased to 23.04%. This demonstrates the interaction between the parameters: an increase in air humidity mitigated the adverse effects of temperature, emphasizing the importance of analyzing the relationships between factors.

Table 6. Updated data for beets with 5% above the upper limit.

Input Data: Beet (Normal Conditions)			
Parameter	Value	Unit	Norm
Temperature	26	°C	18–24
Air Humidity	65	%	50–60
Solution Humidity	1.0	Ratio	0.8–1.2
Light	13	Hours	12–16
pH of Solution	6.1	pH	6.0–6.5
EC of Solution	1.3	EC	1.0–1.4
Deltas and Negative Impact			
Parameter	Delta	Unit	Negative Impact
Temperature	2.000	°C	23.04%
Air Humidity	5.000	%	100.00%
Solution Humidity	0.000	Ratio	0.00%
Light	0.000	Hours	0.00%
pH of Solution	0.000	pH	0.00%
EC of Solution	0.000	EC	0.00%

Figure 8 shows boxplots showing the distribution of values for the six key parameters. Red dots indicate abnormal values that are outside the normal ranges. For example, the temperature is 26 °C, above normal, and humidity is 65%, also above the optimal range, which are shown as anomalies. The parameters of solution moisture, illumination, pH, and solution conductivity are within the normal limits, as evidenced by the green dots. These results indicate that temperature and humidity must be adjusted to maintain the optimal plant growth conditions, while the other parameters do not require changes.

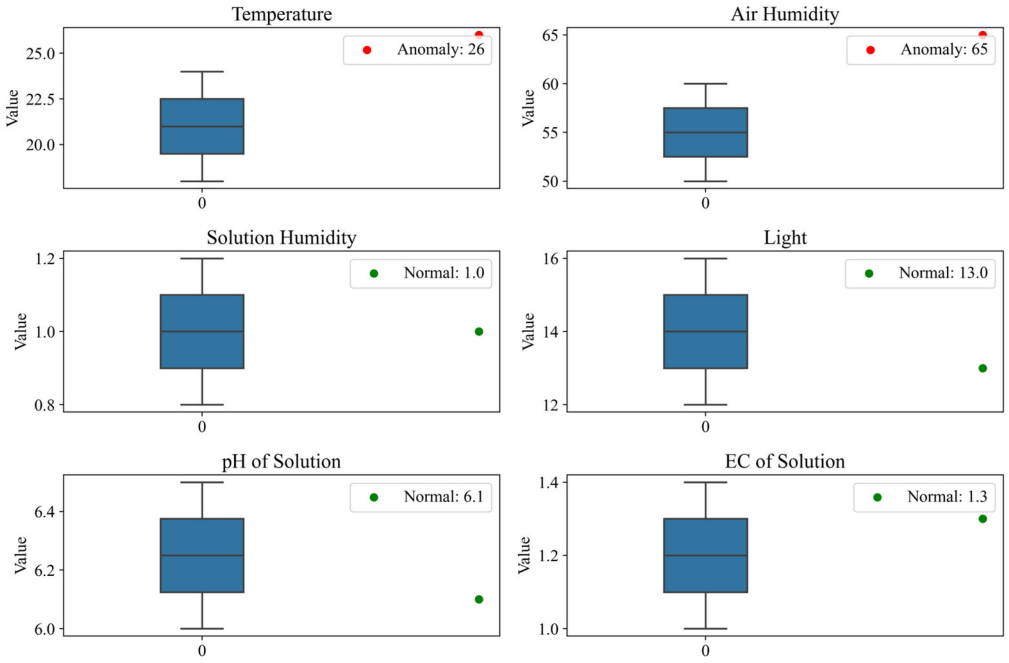

Figure 8. Distribution of microclimate parameters with the identification of anomalies.

It is recommended to reduce the temperature to 18–24 °C and air humidity to 50–60% to avoid heat stress and the development of fungal infections, which can negatively affect plant growth. The other parameters should continue to be monitored to maintain stable conditions. These measures will help stabilize the growing conditions, improving the quality and yield of microgreens. Let us consider, for example, the temperature for beets

x_{temp} = 20.96, where the norm is set as L_{temp} = 18, and H_{temp} = 24. In this case, x_{temp} is within the norm, so that the membership degree will be equal to 0 (9):

$$\mu_{temp}(x_{temp}) = 0 \tag{9}$$

Now let us consider the illumination x_{light} = 17.24, where the normal range is [12,16]. Then, we calculate the degree of deviation (10):

$$\mu(x_{light}) = \min\left(1, \frac{17.24 - 16}{0.1 \times 16}\right) = 0.774, \tag{10}$$

this value indicates that the illumination exceeds the norm by 0.774 units from 0 to 1. Thus, hybrid analysis uses a combination of methods to evaluate the factors that affect microgreens.

3.4. Evaluation and Accuracy Metrics

Each stage plays its role: XGBoost selects significant features, PCA transforms the features into a new space for better data representation, and fuzzy logic allows the evaluation of each factor's deviation from the norm. Accurate data are collected in a controlled hydroponic grow box to test the hybrid model. The collected data are used to validate the model, refine its algorithms, and test its ability to suggest corrective measures in the face of dynamic and changing parameters. Figure 9 shows a bar chart plotting the model quality metrics, including MAE, MSE, RMSE, R^2, AIC, and BIC. The MSE and MAE metrics have minimal values (almost zero), and the R^2 determination coefficient shows the high accuracy of the model with a result of 0.99. The AIC and BIC metrics have negative values (−25.76 and −27.01), indicating a good model adaptation to the data. This plot illustrates how effectively the model copes with the predictions and interpretation of data.

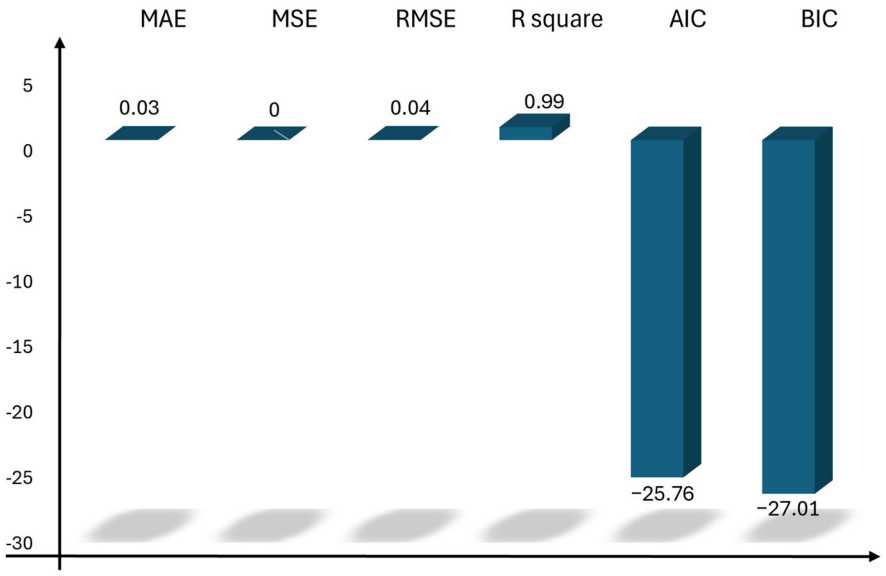

Figure 9. Hybrid analysis performance evaluation result.

A bar chart in Figure 10 illustrates the model quality metrics, including MAE, MSE, RMSE, R^2, AIC, and BIC. The MAE and MSE metrics have low values (0.76 and 1.19, respectively), indicating minor average errors in the model. The R^2 coefficient of determination is 1.00, indicating the high accuracy of the model. The AIC and BIC metrics showed values of 13.03 and 11.78, reflecting the level of adaptation of the model to the data. The graph

highlights the overall quality of the model and its ability to make accurate predictions for the given parameters.

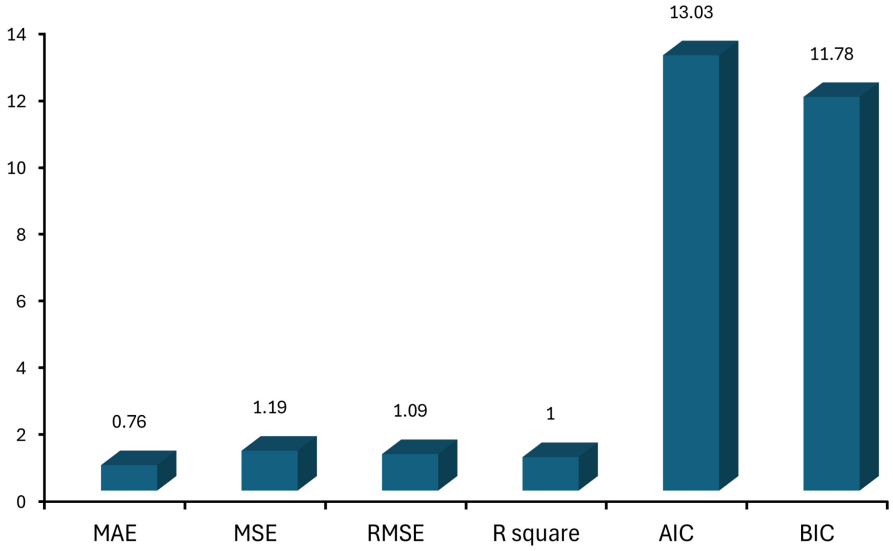

Figure 10. Basic analysis performance evaluation result.

The hybrid method of analysis using weights and fuzzy logic demonstrated a high prediction accuracy (MAE: 0.03, MSE: 0.00, R^2: 0.99) and profound data interpretation. It highlights the critical parameters (air temperature and humidity), making it suitable for complex analysis systems. The AIC (−25.76) and BIC (−27.01) metrics confirm the effectiveness and adaptation of the model. The traditional method based on regression models showed higher errors (MAE: 0.76, MSE: 1.19) and limited interpretation. Although R^2 is 1.00, the lack of consideration of parameter significance makes it less effective. This method is suitable for simple analysis but is inferior to the hybrid approach in problems requiring deep analysis and considering the influence of factors.

4. Discussion

Integrating XGBoost, PCA, and fuzzy logic in a hybrid model has its own unique advantages for hydroponic growers. First and foremost is that by performing resource usage optimization, the technique allows one to identify and adjust the most important parameters before the other ones; thus, it becomes feasible that the inputs, such as nutrients, temperature, or lighting, are used only as required, thereby saving both energy and water. The same is applicable to the use of control strategies for the timely detection of the deviations from the standard environmental setup; thus, no crop loss is experienced, and the risk that fewer crops will have a stable odor flavor, and nutrient density does not occur. Such an upscale of automated and interpretable operations besides less labor is an appreciated attribute. Automatic warnings and fuzzy logic results assist in making the decision-making process much simpler. Also, they enable smaller teams to manage larger cultivation areas.

Despite all these pro factors, some problems remain. In the hydroponic systems that are undergoing constant change, the sensor recalibration is a matter to be resolved periodically for the sake of data accuracy. Moreover, executing such systems on a large scale becomes more complex, thus requiring the resolution of data processing and model retraining limitations. A smooth conclusion of the work involved may comprise the search for more advanced sensor fault detection, the utilization of incremental or online learning approaches, which adjust the model without fully retraining, and the adoption of

distributed computing frameworks (e.g., edge computing) for the efficient processing of the increasingly large data volumes.

The proposed hybrid strategy not only presents a lot of potential for being implemented in a wider range of crops but also for finding more use cases. The system could act as a "one-fits-all" configuration for many different types of green plants, herbs, and fruits by means of changing the main features, i.e., pH thresholds, temperature ranges, and nutrient schedules so that they fit well with different plant species. This issue can be solved easily by using a mini computer; in other words, if low-energy-consuming systems are in place, ranging from single sensor networks to thousands of stations, the coverage of data aggregation and primary processing can be simple short paths, which as a direct consequence would cut off the latent response and offer more efficient real-time progress in large-scale installations.

One key benefit of using this methodology is its high performance, which comes from the capability of the XGBoost part to provide fast and accurate predictions for the case of tabulated data, as well as from the use of more sophisticated regularization techniques. The fuzzy logic unit does the job of making the input sensor readings, which are continuously updated, clear, and easy to understand for human beings, for instance, in the form of "slightly acidic" or "significantly above norm", thus enabling the farmers to take rapid actions. Also, apart from the fact that the generated system can work in real time, the feature is designed to be in a potential state, and it is the automated feedback which can help in adjusting the environmental parameters quite quickly, which is further augmented using the short reporting periods and the efficiency of XGBoost.

A major shortcoming of the suggested approach, though quite attractive in its aspects, is that complexity grows because of three algorithms being utilized simultaneously—XGBoost, PCA, and fuzzy logic. The method will hence need people who are specialized in both the field and subdomain for the model parameterization purpose and model maintenance support. The best hardware had better be used if we are to keep the possibility of working with sensor data streams in real time. The hardware can, in this case, be the edge (or cloud) and reliable sensor networks as well as robust sensor data streams. The above also means that the effectiveness of the data processing system totally depends on the quality of the data collected; a breakdown in the data stream or instability thereof caused by sensors will result in poor model performance unless sensor fault detection or data-imputation strategies are not capable of identifying and handling the unique characteristics of the system.

5. Conclusions

This study developed a hybrid model for data analysis in hydroponic systems, enabling the efficient monitoring and control of microgreens' growth conditions, specifically beets and tarragon. The application of machine learning methods improved the accuracy of predicting the deviations in environmental parameters and facilitated timely corrective measures. The high efficiency in analyzing the key parameters such as temperature, air and solution humidity, illumination, pH, and electrical conductivity ensured the optimal plant growth conditions.

A dataset comprising daily records from IoT sensors was used to train and test the model, allowing for a comprehensive consideration of the various factors affecting microgreen cultivation. The experimental results demonstrated that the hybrid model effectively reduced deviations from the norm, enhancing the stability and quality of the produced microgreens. The model integrated XGBoost, principal component analysis (PCA) with six components, and fuzzy logic, providing more accurate and detailed data analysis. Unlike simple correlation methods that only identify linear dependencies, the proposed

approach detected complex multivariate nonlinear relationships between the parameters while preserving all the critical information through principal components.

The ensemble of XGBoost, principal component analysis (PCA), and fuzzy logic was seen to exhibit strong predictive capability and thus achieved an accuracy of around 97.88%, an F1-score of nearly 0.98, and an unbelievably high ROC-AUC score of approximately 0.9999. This collection of key figures not only has a contingent monetary value; it also delves deeper into the issue of environmental sustainability. Improved accuracy, by way of an example, can be achieved by the farmer who manages to keep close to the most suitable environmental conditions; thus, the farmer cuts down on the wastage of crops and becomes more reliable in achieving the target quantity of yield from time to time. A high F1-score is an assurance that the errors will be rare and recognized with both recall and precision. Such a situation helps to rid the process of false alerts and at the same time ensures that there is no interruption in the detection process at any point. Finally, the almost-perfect ROC-AUC score verifies the model's ability to segregate faultlessly a variety of irregularities, which is the premise of more intricate, data-influenced resolutions.

However, several challenges related to system scalability were identified during the study. These include the high computational resource requirements for real-time data processing in large-scale installations, data variability when adapting the model to different crops and environmental conditions (necessitating significant retraining efforts), and hardware limitations, such as the need to deploy and maintain numerous sensors. Further refinement is required to enhance the model's adaptability to heterogeneous datasets and dynamic environmental conditions, ensuring its scalability for larger installations. As sensor arrays are being upscaled, data challenges—ranging among missing or corrupted data, sensor drift, and hardware malfunctions—arise as a significant issue that needs to be resolved. Thus, the ongoing actions should concentrate on implementing the advanced methods for sensor fault detection, outlier handling, and data imputation in a streaming sensor environment. The application of anomaly-detection algorithms, probably in combination with PCA or fuzzy logic strategies, would add to the system's reliability and consequently assure that the process flow will not be disrupted, even if there is a partial sensor failure, or noisy readings are present.

The hybrid method, so far, has been put to the test for tarragon and beets. Many profitable microgreens and horticultural plants have similar vulnerabilities to the surrounding conditions as those observed in beets and tarragon. In the future, the method may also find applications with other kinds of fast-growing greens (e.g., spinach, kale, basil) and later with fruiting vegetables like tomatoes and peppers. By tuning the parameter thresholds and the membership functions in the fuzzy logic module, it is possible to accommodate these different crops, thereby making the system more versatile and ready for the market.

Author Contributions: Conceptualization, K.B., J.T., A.T., I.S. and A.S.; data curation, A.S.; formal analysis, J.T., A.T., I.S. and A.S.; funding acquisition, A.S.; investigation, K.B., J.T., A.T., I.S. and A.S.; methodology, K.B., J.T., A.T., I.S. and A.S.; project administration, A.S.; resources, K.B., J.T., A.T., I.S. and A.S.; software, A.T. and I.S.; supervision, J.T.; validation, J.T., A.T., I.S. and A.S.; visualization, A.T. and I.S.; writing—original draft, K.B., J.T. and A.T.; writing—review and editing, K.B., J.T., A.T. and I.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No.BR24992852 "Intelligent models and methods of Smart City digital ecosystem for sustainable development and the citizens' quality of life improvement").

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in FigShare at https://doi.org/10.6084/m9.figshare.28667981.v1 (accessed on 26 March 2025).

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

Abbreviations

The following abbreviations are used in this manuscript:

Ref	Reference to the papers listed in the References section
PCA	Principal component analysis
IoT	Internet of Things
IIoT	Industrial Internet of Things
ROC-AUC	Receiver Operating Characteristic—Area Under the Curve
XGBoost	Extreme Gradient Boosting
F1-score	F1 Measure (Harmonic Mean of Precision and Recall)
ESP	Arduino Microcontroller
HTTP	Hypertext Transfer Protocol (HTTP) is an application-layer protocol for transmitting hypermedia documents.
EC	Electrical Conductivity Sensor
pH	A pH meter is an instrument used to measure the hydrogen ion activity in solutions.
AIC	The Akaike information criterion (AIC) is an estimator of prediction error.
BIC	The Bayesian Information Criterion, or BIC for short, is a method for scoring and selecting a model.
MSE	Mean Squared Error represents the average of the squared difference between the original and predicted values in the dataset.
MAE	The mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset.
R square	R-squared represents the proportion of the variance in the dependent variable
RMSE	Root Mean Squared Error is the square root of the Mean Squared Error.
temp	Temperature in the hydroponics grow box
ph_solution	Level of pH activity in the hydroponics grow box
ec_solution	Level of electrical conductivity in the hydroponics grow box
humidity_air	General relative humidity of the grow box
humidity_solution	Humidity level of the substrate in the hydroponics
type_encoded	Two types of microgreen (beetroot and tarragon)
negative_class	Indication whether an observation represents an undesirable outcome

References

1. Singh, A.; Singh, J.; Kaur, S.; Gunjal, M.; Kaur, J.; Nanda, V.; Ullah, R.; Ercisli, S.; Rasane, P. Emergence of microgreens as a valuable food, current understanding of their market and consumer perception: A review. *Food Chem. X* **2024**, *23*, 101527. [CrossRef] [PubMed]
2. Abaajeh, A.R.; Kingston, C.E.; Harty, M. Environmental factors influencing the growth and pathogenicity of microgreens bound for the market: A review. *Renew. Agric. Food Syst.* **2023**, *38*, e12. [CrossRef]
3. Nair, B.R. Microgreens: A Future Super Food. In *Conservation and Sustainable Utilization of Bioresources*; Springer: Singapore, 2023; pp. 103–122. [CrossRef]
4. Atherton, H.R.; Li, P. Hydroponic cultivation of medicinal plants—Plant organs and hydroponic systems: Techniques and trends. *Horticulture* **2023**, *9*, 349. [CrossRef]
5. Fernandes, A.S.; Bragança, I.; Homem, V. Personal care products in soil-plant and hydroponic systems: Uptake, translocation, and accumulation. *Sci. Total Environ.* **2023**, *912*, 168894. [CrossRef]
6. Kumara, V.; Mohanaprakash, T.A.; Fairooz, S.; Jamal, K.; Babu, T.; Sampath, B. Experimental study on a reliable smart hydroponics system. In *Human Agro-Energy Optimization for Business and Industry*; IGI Global: Pennsylvania, PA, USA, 2023; pp. 27–45. [CrossRef]

7. Yessenova, M.; Abdikerimova, G.; Ayazbaev, T.; Murzabekova, G.; Ismailova, A.; Beldeubayeva, Z.; Ainagulova, A.; Mukhanova, A. The effectiveness of methods and algorithms for detecting and isolating factors that negatively affect the growth of crops. *Int. J. Electr. Comput. Eng.* **2023**, *13*, 1669–1679. [CrossRef]
8. Zhidekulova, G.; Mustafayev, Z.; Kozykeyeva, A. Regulation of irrigation: Modelling of bioclimatic coefficients of agricultural cultures. *Res. Crops* **2018**, *19*, 132–143. [CrossRef]
9. Yessenova, M.; Abdikerimova, G.; Sadirmekova, Z.B.; Glazyrina, N.; Adikanova, S.; Tanirbergenov, A.; Karipola, M.; Mukhamedrakhimova, G. Features of growth of agricultural crops and factors negatively affecting their growth. *Indones. J. Electr. Eng. Comput. Sci.* **2023**, *30*, 625–632. [CrossRef]
10. Sulaiman, R.; Azeman, N.H.; Mokhtar, M.H.H.; Mobarak, N.N.; Bakar, M.H.A.; Bakar, A.A.A. Hybrid ensemble-based machine learning model for predicting phosphorus concentrations in hydroponic solution. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2024**, *304*, 123327. [CrossRef]
11. Shrivastava, A.; Nayak, C.K.; Dilip, R.; Samal, S.R.; Rout, S.; Ashfaque, S.M. Automatic robotic system design and development for vertical hydroponic farming using IoT and big data analysis. *Mater. Today Proc.* **2023**, *80*, 3546–3553. [CrossRef]
12. Chikkasiddaiah, C.; Govindaswamy, P.; Srikantaswamy, M. An efficient hydro-crop growth prediction system for nutrient analysis using machine learning algorithm. *Int. J. Electr. Comput. Eng.* **2023**, *13*, 6681–6690. [CrossRef]
13. Rathor, R.; Choudhury, A.S.; Sharma, S.; Nautiyal, P.; Shah, G. Empowering vertical farming through IoT and AI-Driven technologies: A comprehensive review. *Heliyon* **2024**, *10*, e34998. [CrossRef] [PubMed]
14. Balik, S.; Dasgan, H.Y.; Ikiz, B.; Gruda, N.S. The performance of growing-media-shaped microgreens: The growth, yield, and nutrient profiles of broccoli, red beet, and black radish. *Horticulturae* **2024**, *10*, 1289. [CrossRef]
15. Dhaka, A.S.; Dikshit, H.K.; Mishra, G.P.; Tontang, M.T.; Meena, N.L.; Kumar, R.R.; Ramesh, S.V.; Narwal, S.; Aski, M.; Thimmegowda, V.; et al. Evaluation of growth conditions, antioxidant potential, and sensory attributes of six diverse microgreens species. *Agriculture* **2023**, *13*, 676. [CrossRef]
16. Tussupov, J.; Abdikerimova, G.; Ismailova, A.; Kassymova, A.; Beldeubayeva, Z.; Aitimov, M.; Makulov, K. Analyzing disease and pest dynamics in steppe crop using structured data. *IEEE Access* **2024**, *12*, 71323–71330. [CrossRef]
17. Mamatha, V.; Kavitha, J.C. Machine learning based crop growth management in greenhouse environment using hydroponics farming techniques. *Meas. Sens.* **2023**, *25*, 100665. [CrossRef]
18. Phukan, A. Hydroponics using IOT and Machine Learning. *Int. Res. J. Eng. Technol.* **2022**, *9*, 207–211.
19. Verma, M.S.; Gawade, S.D. A machine learning approach for prediction system and analysis of nutrients uptake for better crop growth in the Hydroponics system. In Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 25–27 March 2021; IEEE: New York, NY, USA, 2021; pp. 150–156. [CrossRef]
20. Tussupov, J.; Yessenova, M.; Abdikerimova, G.; Aimbetov, A.; Baktybekov, K.; Murzabekova, G.; Aitimova, U. Analysis of formal concepts for verification of pests and diseases of crops using machine learning methods. *IEEE Access* **2024**, *12*, 19902–19910. [CrossRef]
21. Venkatasaichandrakanth, P.; Iyapparaja, M. Pest detection and classification in peanut crops using CNN, MFO, and EViTA algorithms. *IEEE Access* **2023**, *11*, 54045–54057. [CrossRef]
22. Bakirov, K. Datasets from IoT Devices. Available online: https://figshare.com/articles/dataset/Datasets_from_IoT_devices/28667981 (accessed on 26 March 2025). [CrossRef]
23. Idoje, G.; Mouroutoglou, C.; Dagiuklas, T.; Kotsiras, A.; Muddesar, I.; Alefragkis, P. Comparative analysis of data using machine learning algorithms: A hydroponics system use case. *Smart Agric. Technol.* **2023**, *4*, 100207. [CrossRef]
24. Susanto, F.; Suryani, N.K.; Darmawan, P.; Prasiani, K.; Satria, I.M. A Comprehensive review on automation in hydroponic agriculture using machine learning and IoT. *RSF Conf. Ser. Eng. Technol.* **2021**, *1*, 86–95. [CrossRef]
25. Dhal, S.B.; Mahanta, S.; Gumero, J.; O'Sullivan, N.; Soetan, M.; Louis, J.; Gadepally, K.C.; Mahanta, S.; Lusher, J.; Kalafatis, S. An IoT-Based Data-Driven Real-Time Monitoring System for Control of Heavy Metals to Ensure Optimal Lettuce Growth in Hydroponic Set-Ups. *Sensors* **2023**, *23*, 451. [CrossRef] [PubMed]
26. Dhal, S.B.; Jungbluth, K.; Lin, R.; Sabahi, S.P.; Bagavathiannan, M.; Braga-Neto, U.; Kalafatis, S. A Machine-Learning-Based IoT System for Optimizing Nutrient Supply in Commercial Aquaponic Operations. *Sensors* **2022**, *22*, 3510. [CrossRef] [PubMed]
27. Tatas, K.; Al-Zoubi, A.; Christofides, N.; Zannettis, C.; Chrysostomou, M.; Panteli, S.; Antoniou, A. Reliable IoT-Based Monitoring and Control of Hydroponic Systems. *Technologies* **2022**, *10*, 26. [CrossRef]
28. Săcăleanu, D.-I.; Matache, M.-G.; Roșu, Ș.-G.; Florea, B.-C.; Manciu, I.-P.; Perișoară, L.-A. IoT-Enhanced Decision Support System for Real-Time Greenhouse Microclimate Monitoring and Control. *Technologies* **2024**, *12*, 230. [CrossRef]
29. Subahi, A.F. Advancing Sustainable Cyber-Physical System Development with a Digital Twins and Language Engineering Approach: Smart Greenhouse Applications. *Technologies* **2024**, *12*, 147. [CrossRef]
30. Zito, F.; Giannoccaro, N.I.; Serio, R.; Strazzella, S. Analysis and Development of an IoT System for an Agrivoltaics Plant. *Technologies* **2024**, *12*, 106. [CrossRef]

31. Theodorakopoulos, L.; Karras, A.; Theodoropoulou, A.; Kampiotis, G. Benchmarking Big Data Systems: Performance and Decision-Making Implications in Emerging Technologies. *Technologies* **2024**, *12*, 217. [CrossRef]
32. Christakakis, P.; Papadopoulou, G.; Mikos, G.; Kalogiannidis, N.; Ioannidis, D.; Tzovaras, D.; Pechlivani, E.M. Smartphone-Based Citizen Science Tool for Plant Disease and Insect Pest Detection Using Artificial Intelligence. *Technologies* **2024**, *12*, 101. [CrossRef]
33. Bakthavatchalam, K.; Karthik, B.; Thiruvengadam, V.; Muthal, S.; Jose, D.; Kotecha, K.; Varadarajan, V. IoT Framework for Measurement and Precision Agriculture: Predicting the Crop Using Machine Learning Algorithms. *Technologies* **2022**, *10*, 13. [CrossRef]
34. Kang, K.-D. A Review of Efficient Real-Time Decision Making in the Internet of Things. *Technologies* **2022**, *10*, 12. [CrossRef]
35. Govindarajan, U.H.; Zhang, C.; Raut, R.D.; Narang, G.; Galdelli, A. A Review of Academic and Patent Progress on Internet of Things (IoT) Technologies for Enhanced Environmental Solutions. *Technologies* **2025**, *13*, 64. [CrossRef]