Journal of Computer Science

Research Article

IoT-Based Smart Hydroponic System for Nutrient Management Using Predictive Machine Learning Algorithms

Palash Gourshettiwar and K. T. V. Reddy

Department of Computer Science and Engineering (Artificial Intelligence and Data Science - AIDS), Datta Meghe Institute of Higher Education and Research (DU), India

Article history
Received: 19-03-2025
Revised: 4-07-2025
Accepted: 24-09-2025

Corresponding Author:
Palash Mohan Gourshettiwa
Department of Computer Science and Engineering (Artificial Intelligence and Data Science - AIDS), Datta Meghe Institute of Higher Education and Research (DU), India
Email: palash9477@gmail.com

Abstract: Hydroponics, an efficient cultivation method, benefits significantly from the precision and adaptability that Machine Learning (ML) algorithms can offer, along with the real-time monitoring facilitated by the devices using the Internet of Things (IoT). This study summarises the latest research and discusses how Machine Learning and the IoT work together, focusing on nutrient optimization, plant development, and resource efficiency. The critical subjects of the discussion are the use of Machine Learning algorithms, the function of IoT devices for real-time monitoring, communication protocols, scalability issues, and implementation. This article discusses the transformative amalgamation of the IoT and Machine Learning technologies within vegetable hydroponic systems for nutrition management. The system integrates an IoT-enabled hardware setup comprising sensors for pH, temperature, humidity, light, and TDS, placed at strategic positions in the hydroponic unit, with data collected every 15 minutes for ML-based analysis. The study utilized machine learning classifiers for tomato crops for over 121 days, measuring the parameters such as temperature, humidity, Total Dissolved Solids, potential of hydrogen (pH), and crop type. The research indicates benefits to crop output, resource efficiency, and sustainability based on the case studies and the analysis of results. The machine learning models introduced in this research were evaluated against contemporary studies, revealing an accuracy enhancement ranging from 1.17% to 5.25%, depending on the dataset and algorithm employed. The present study conducts a Comprehensive analysis involving machine learning algorithms, indicating that among all the models, Random Forest (RF), Gaussian Naive Bayes (GNB), and Gradient Boosting (GB) achieved an accuracy of 99.71, 99.71, and 99.42%, respectively, in the dataset by making stage-wise decisions. Compared with recent literature on Machine Learning, models achieved better performance, highlighting the study's strengths. The conclusion of this paper provides directions for further research and calls for more investigation into state-of-the-art machine learning approaches and scalable solutions for the resilient and sustainable future of hydroponic agriculture.

Keywords: Hydroponics, Machine Learning, IoT, Data Security, Smart Farming

Introduction

India's economy revolves around agriculture, and the past 60 years have shown us that there is a direct correlation between agricultural growth and economic prosperity. India's current farming system results from remarkable accomplishments and lost possibilities. As per the Ministry of Statistics and Programme Implementation

(MoSPI) report, the share of Gross Value Added (GVA) of agriculture and allied sectors in the total economy of India was found to be 18.3 % (Mamun Khan and Bezbaruah., 2024) India's agricultural output must catch up with the nations now ranked as the world's economic superpowers to become a global economic force (Maurya et al., 2023; Alipio et al., 2017) Since smart farming makes farms more capable of detecting and regulating their settings, it is thought to be the way of the future of agriculture. With the IoT, enormous amounts of data may be analyzed by connecting and accessing different devices. For the data to be usable, though, autonomous agricultural output using analytics is just as crucial as having Internet support and sensor readings that update independently (Swain et al., 2021) Today, soil-based agriculture faces challenges from various artificial factors, including Urbanization and industrialization. Additionally, abrupt natural disasters, climate change, and the uncontrolled use of chemicals in agriculture contribute to the depletion of fertile and high-quality soil (Alipio et al., 2019) Since smart farming makes farms more capable of detecting and regulating their settings, it is thought to be the way of the future of agriculture. With the IoT, enormous amounts of data may be analyzed by connecting and accessing different devices (Khan et al., 2021) One popular and extensively used method for growing plants without soil is hydroponics, which allows for a great deal of control over the roots' surrounding environment. Although cultivating plants in nutrient-rich water may have originated in prehistoric times, the technology has an intriguing history of development and application that dates back to the middle of the 18th century (Sharma et al., 2018). Hydroponic gardening produces high-quality food and manages resources efficiently; it is becoming increasingly popular worldwide (Abdelraouf and Hamza, 2024). Hydroponics farming is a type of agriculture that uses less water and other resources than conventional soil-based agricultural methods. (Majid et al., 2021). Nevertheless, because hydroponic farming requires simultaneous supervision of several factors, dietary recommendations, and a plant diagnostic system, monitoring the practice might be complex (Carlos Eduardo et al., 2023). However, by implementing ML-based control algorithms in the agriculture industry, recent technological advancements can help solve these issues. (Taha et al., 2022; Álvarez Salas et al., 2024).

India's agricultural sector is its backbone, and to achieve high yields, farmers must make sophisticated judgments about things like seeding, fertilizer dispensing, and excavating. One of the most crucial factors is the right proportion of plant nutrients based on the age, type of crop, and geographic location (Larsson et al., 2024). The essential primary macronutrients for growing any crop are Nitrogen (N), Phosphorus (P), and Potassium (K). The quality and yield of a crop are determined by the proper N.P.K. proportions required at different stages of the crop cycle, along with secondary macronutrients like Calcium (Ca), Magnesium (Mg), Sulfur (S), and micronutrients like Iron, Manganese, Zinc, Copper, Boron, Molybdenum, Chlorine, and Nickel. Since plants are constantly consuming nutrients, the quantities of nutrients may fluctuate. As a result, it is essential to control and continuously monitor the NPK levels in the water tank in real-time (Ameer et al., 2024; Waiba et al., 2020; Wakchaure et al., 2023). The level of NPK in water is essential for the growth of plants, which can be affected by pH, temperature, oxygen levels, and high concentrations of Iron or aluminum. (Ye et al., 2019; Liu et al., 2024). The multiple factors, plant nutrients, and diagnostic techniques involved in hydroponics production make monitoring difficult. Because of recent technological advancements, agriculture now employs AI-based control algorithms, which have aided in the search for solutions (Carlos et al., 2023; Taha et al., 2022; Nelson, 2013).

ML techniques have recently gained much usage in agriculture. In some countries, research fields, such as crop classification and monitoring, encompass crop growth and yield prediction Al-Akhras et al. (2024); Shah et al. (2019); Wolanin et al. (2019). Mehra et al. (2018) suggested that some of the different purposes for which the ML algorithms can be used in hydroponics are controlling nutrient management, water level management, Electrical Conductivity (EC) values, pH balance, and effective sensory communication (Mehra et al., 2018). The reasons for successfully developing the correct ML models are the higher quality of the dataset and larger data sizes. Ni et al. (2022) applied suitable algorithms to solve problems comprising various datasets. Ni et al. (2022) used a Regression Tree (RT) to predict yields of soybean and maize in the United States. Johnson, (2024) compare three improved ML models, namely, Support Vector Machine (SVM), RF, and Neural Network (NN), with the traditional regression method for predicting wheat yield in Australia (Cai et al., 2019). According to this research, the ML methods are more improved than the traditional regression method. In predicting wheat, maize, and potato yields, Khan et al. (2023) made use of RF and multiple linear regression (MLR) (Khan et al., 2023). They concluded that RF was better at predicting results than MLR. Fukuda et al. (2013) also used RF to predict mango fruit with a successful outcome. Fukuda et al. (2013); Khaki and Wang (2019) applied Deep Learning (DL), which is a type of NN, involving many layers and infers progressively more abstract features from the raw input data. (Khaki and Wang, 2019; Khaki et al., 2020) used CNNs and RNNs to predict the yield of soybeans based on a sequence of remotely sensed images (You et al., 2017). Moreover, the DNN was also employed for maize yield prediction in 2008-2016. The results indicated that DNN outperformed LASSO, SNN, and RT (Khaki and Wang, 2019). A DNN model was used by for corn and soybean yield prediction in 2006-2015

Kim et al. (2019). In Argentina, worked on a DNN for predicting soybean yields Khaki and Wang (2019). Rahman et al. (2024) developed a real-time wireless IoT-based hydroponic system with embedded sensors for pH, EC, temperature, and humidity. Their work showed improved crop performance in Red Cos lettuce using real-time data, but lacked any ML integration or predictive modeling. Addresses real-time monitoring and feedback control, but not data analytics or prediction (Rahman et al., 2024).

A systematic review by MDPI (2024) analyzed over 130 smart hydroponic studies and highlighted that many lack clarities in sensor configuration, data acquisition frequency, and real-time data processing. The review emphasized the need for more structured integration of sensors, automation, and analytics. Identifies widespread gaps in sensor placement strategy, data frequency, and system architecture (Shareef et al., 2024).

An AIoT-based framework published in Agricultural Informatics combined sensor-driven hydroponics with ML (RF, SVM, KNN, XGBoost) to achieve 97.5% accuracy in crop recommendation and nutrient forecasting. It included a cloud backend and mobile dashboard but covered fewer ML models and lacked detailed preprocessing steps. It presents an end-to-end innovative system limited to comparative ML analysis and model transparency (Rahman et al., 2024).

ML and IoT-based hydroponics collect and process sensor data (such as pH, temperature, humidity, and light levels). The vulnerability of IoT devices to cyberattacks could compromise system integrity, leading to disruptions in food production and reduced trust in technology. Incorporating relevant works on the application of ML to enhance cybersecurity is crucial. (Pleshakova et al., 2024) For instance, ML algorithms can be applied to detect anomalies in network traffic or sensor data, signaling potential security breaches in the hydroponic system. DL models can provide advanced capabilities, like real-time threat detection or autonomous recovery from attacks, thereby ensuring the robustness and security of automated hydroponics. Integrating AI-driven cybersecurity measures, such as Intrusion Detection Systems (IDS) based on ML or DL, the hydroponic system can be better protected against cyber threats (Tsapin, 2023). This approach improves the reliability of intelligent farming technologies and reinforces the security foundation of innovative ecosystems, where such systems play a pivotal role.

This research aims to design and develop an IoT-based hardware setup for data collection in hydroponic farming. This study explores integrating IoT and ML models to predict growth conditions, optimize nutrient adjustments, and enhance crop quality while minimizing water wastage. By training and testing ML and Deep Learning (DL) models, the research highlights the precision of nutrient delivery, improved plant growth, and increased crop yields. Additionally, it investigates the advantages and limitations of the proposed system, emphasizing real-time monitoring and automation to ensure efficient resource use and reduce manual labor.

Methods

Experimental Setup

The goal is to design and develop an ML-enabled application for real-time monitoring and alert handling systems in hydroponics based on sensor data and to develop an effective sensory communication mechanism for managing mineral cutoff levels. The study used 20 hydroponic units of tomato plants with perlite, connected to a nutrient and water tank, equipped with sensors (pH, TDS, temperature, humidity, and light) as shown in Figure 1. The placement of each sensor was optimized to monitor root zone, reservoir, and canopy conditions, which are analyzed through ML to determine the optimal control actions for regulating the hydroponic system and classified into different labels.

The experimental setup consists of hardware components, software, and various sensors used to form IoT and ML based hydroponics systems. The proposed IoT-enabled hydroponic system, showing data flow and integration (Arduino → Pi → Firebase) in Figure 2, the strategic placement of sensors plays a critical role in ensuring accurate environmental and nutrient monitoring, which is essential for optimizing plant growth and nutrient management.

The DHT22 sensor, responsible for measuring temperature and humidity, is placed in the ambient air zone above the hydroponic setup where the plants grow. This allows it to capture the microclimatic conditions around the plants, ensuring the environment remains within the optimal range for photosynthesis and transpiration. The light sensor (LDR module) is also positioned near the plant canopy, facing upward to detect the intensity of natural or artificial light. Based on the measured light levels, it can control the activation of the grow lights via a relay, ensuring that plants receive consistent and adequate illumination throughout the day, especially in indoor or low-light environments. The pH and TDS sensors are submerged in the nutrient reservoir. The pH sensor monitors the acidity or alkalinity of the solution, which is crucial for nutrient absorption by plant roots. The TDS (Total Dissolved Solids) sensor evaluates the concentration of dissolved nutrients such as minerals and salts, indicating the strength of the nutrient solution. Maintaining these values within optimal ranges ensures healthy root development and prevents nutrient toxicity or deficiency.

The system also incorporates a DC water pump, placed inside the water or nutrient tank, to circulate the nutrient solution through the hydroponic channels. It is triggered based on the water level or scheduling needs. An oxygen air pump is also connected to the reservoir to supply dissolved oxygen, promoting root respiration and preventing anaerobic conditions that could lead to diseases.

Fig. 1: Visualization of the Project environment: (a) Hardware set-up of the hydroponic system, (b)Arrangement of hydroponics setup, (c) Green Tomatoes, (d) Red Tomatoes

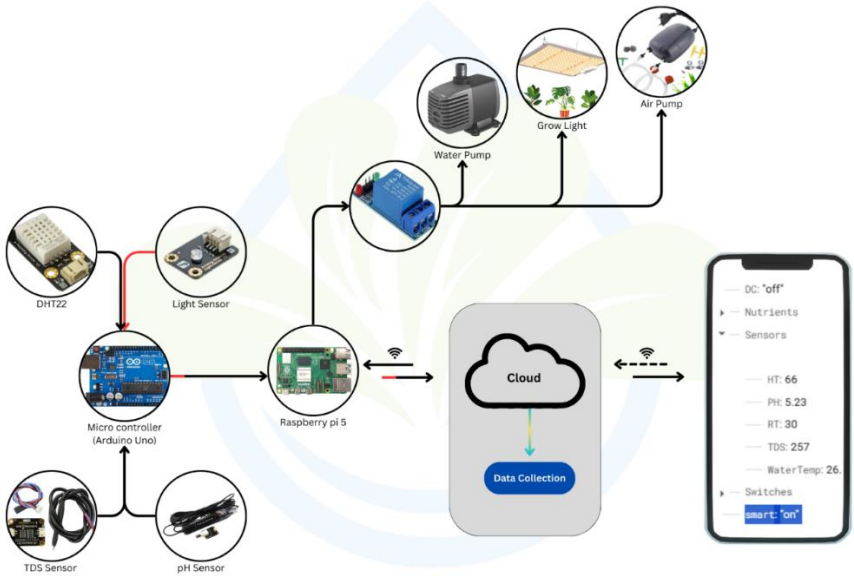


Fig. 2: Data Flow and Integration (Arduino → Pi → Firebase)

All these sensors and actuators are connected to a central microcontroller (Arduino Uno), which collects the raw data and sends it to a Raspberry Pi 5. The Raspberry Pi handles data processing and cloud communication. Data is visualized and monitored remotely through a smartphone application, enabling real-time decision-making and system control. Overall, the sensor placement in this system is designed to mimic the needs of an innovative, self-regulating agricultural environment, ensuring precision farming in a hydroponic setup.

All sensors and actuators are interfaced with a microcontroller (Arduino Uno), which collects data and transmits it to a Raspberry Pi 5. The Raspberry Pi handles local processing and communicates the data to a cloud-based platform for storage and analysis. The system is configured to collect and transmit sensor data at a fixed interval of every 30 minutes. This half-hour frequency ensures that environmental and nutrient parameters are monitored regularly without overwhelming the system with excessive data, striking a balance between responsiveness and resource efficiency. Users can remotely access the data through a mobile application, which provides real-time updates, historical trends, and manual override controls for pumps, lights, and other actuators, as shown in Figure 3.

Technical Specifications of hardware components are provided in Table 1 below. Used for the integration of the system.

Fig. 3: Mobile Application for managing and monitoring the system

Table 1: Technical Specifications of components used for hardware setup

Sr. No	Name of Components	Specifications	Functions and Typical Range
1	Arduino	ATmega328P, 14 Digital I/O Pins, 6 Analog Pins	Central unit interfacing all analog/digital sensors like DHT22, pH, TDS, and Light Sensor. Sends data to Raspberry Pi.
1	Raspberry Pi/	Quad-core ARM Cortex CPU, Wi-Fi/BT, GPIO support	Acts as a gateway to the cloud. Receives data from Arduino via serial communication and handles cloud upload + UI display.
2	DHT22 Sensor	Temp range: -40 to 80°C, Humidity: 0–100% RH, Accuracy: ±0.5°C	Monitors temperature and humidity via Arduino's digital pin 10
3	Light Sensor	Analog light detection and output voltage change with intensity. 0–10,000 lux (approx., analog output)	Placed near plants to measure ambient light. Controls grow lights via a relay based on light levels.
4	DC Water Pump	3–12V DC; Flow rate: ~80–120 L/hr	Activated via relay when the water level drops (not explicitly shown in the diagram but implied). Controls nutrient flow to plants.
5	Air Pump	12V DC or 220V AC, 1–5 L/min	Provides oxygenation to the water tank to prevent root rot and anaerobic conditions. Can be scheduled via smart relay.
6	pH Sensor	Range: 0–14, Analog output	Placed in the nutrient tank or hydroponic reservoir. Measures the acidity/alkalinity of water.
7	TDS Sensor	Range: 0–1000/5000 ppm, accuracy ±10%	Submerged in the nutrient solution to detect mineral concentration (salinity level).
8	Relay Module	1-Channel, 5V input control, 250V/10A max load	Switch devices like water pump, air pump, and grow light ON/OFF based on Arduino or Raspberry Pi commands.
9	Cloud Server	Cloud platform for data storage, analysis, and monitoring	Collects all environmental and nutrient data, enabling remote control via smartphone.
10	Smartphone UI	Web/mobile dashboard	Displays real-time sensor readings (e.g., HT, pH, RT, TDS, water temperature) and allows remote switching of components.

The software part of the experimental setup consists of Raspbian Jessie, the operating system for Raspberry Pi, based on Debian Linux. Arduino IDE is an open-source software for writing and uploading code to Arduino microcontrollers, compatible with multiple operating systems. Python is an open-source, interpreted programming language for interfacing sensors and NN. Nancy is a Python library that simplifies the control of Arduino through a Raspberry Pi, enabling easy sensor integration. TensorFlow is an open-source library by Google that builds deep neural networks and leverages GPUs for fast and efficient performance. Pandas is a Python library for data cleaning and analysis to load, process, and prepare data for NN. Google Firebase is a cloud service platform for real-time database and storage, used to store sensor data and integrate it with the system.

The proposed hydroponic system is fully integrated with IoT capabilities, enabling Real-time monitoring; all sensor data (temperature, humidity, pH, TDS, light) is transmitted every 30 minutes to a cloud platform (Google Firebase) via Raspberry Pi 5. Users can monitor real-time values on a smartphone dashboard. Automated Feedback Control is based on sensor thresholds, and the system automatically triggers actuators like water pumps, air pumps, and grow lights using relays. For example, grow lights are activated if light levels fall below a threshold. Similarly, nutrient pumps adjust based on pH or TDS levels. Manual Overrides & Alerts The mobile UI allows manual control over system components and sends alerts for critical conditions (e.g., low water levels and extreme pH values). Historical Data Analysis, where all sensor logs are stored in Firebase, enabling users to visualize historical trends for long-term analysis and optimization. This supports improved prediction, diagnosis, and decision-making based on time-series data.

Workflow

The schematic presentation, as shown in Figure 4, illustrates a typical ML pipeline, specifically applied to a use case involving the automatic control of nutrients. The system's workflow consists of various steps required for the ML models to get trained and tested.

Data Collection

This step involves gathering the raw data that will be used for the analysis. In this case, it refers to the collection of data related to tomatoes, which includes features such as plant age, type of tomato (e.g., flowering, green tomato, red tomato), and other relevant agricultural or environmental factors like temperature, humidity, plant type, plant water level, pH, and TDS. The data collection frequency of sensor readings was acquired at fixed 15-minute intervals daily over a four-month crop cycle (121 days). This produced thousands of raw data points. For analysis and reporting, these were aggregated into daily summaries. As shown in Table 2, these are the various features of the dataset that were considered at the time of training and testing with multiple ML models.

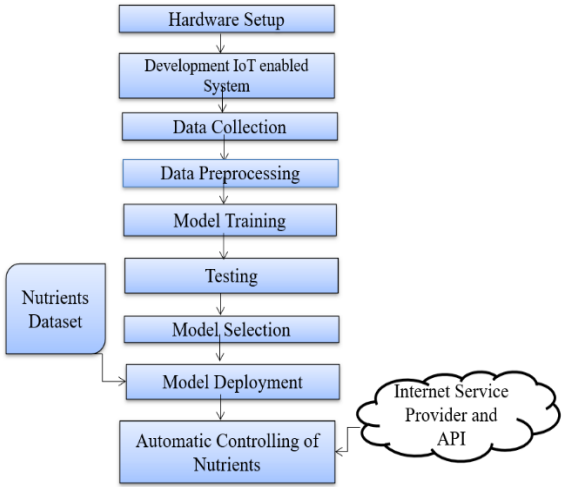


Fig. 4: Schematic representation of the Hydroponics System

Table 2: Features of the dataset

Sr. No.	Attributes	Functions
1	Temperature	The ambient temperature surrounding the tomato plant
2	Humidity	The level of moisture in the air
3	Plant_Type	The specific variety or type of tomato plant
4	Plant_Age	Age of the plant
5	TDS_A	Total Dissolved Solids (TDS) measurement from source A, indicating the concentration of dissolved substances in the plant's irrigation water.
6	TDS_B	TDS measurement from source B provides additional information about the water quality used for the tomato plants.
7	TDS_C	TDS measurement from source C gives further insights into the nutrient levels in the irrigation water.
8	Total_TDS	The cumulative TDS value derived from various sources represents the overall water quality affecting the tomato plant.
9	PH	The pH level of the soil or water is a vital parameter that can impact nutrient availability and plant health.
10	Output	A specific output measurement related to the plant's performance or yield, possibly indicating the productivity of the tomato plant.
11	Stage	The growth stage of the tomato plant, which could range from early development to full maturity, helps us understand the plant's life cycle.

DOI: 10.3844/jcssp.2026.960.980

Lighting was supplemented using grow lights with a relay mechanism that ensured a minimum required lux level, minimizing uncontrolled light exposure variability. Similarly, nutrient concentrations were maintained within recommended thresholds using pH and TDS sensors with automated adjustments, and environmental variations were logged to account for potential bias.

The correlation matrix is shown in Figure 5, which illustrates the relationships between various features in the hydroponic vegetable system dataset

The data collection step involves gathering the raw data that will be used for the analysis. In this case, it refers to collecting data on tomatoes, including features such as plant age and other environmental factors like Temperature, Humidity, plant type, TDS A, TDS B, TDS C, total TDS, PH, Output, and Stage, which illustrates the relationships between various features in the hydroponic vegetable system dataset. To assess multicollinearity, we have used the continuous variables (Temperature, Humidity, Plant_Age, TDS_A, TDS_B, TDS_C, Total_TDS, PH) for VIF calculation.

Data Analysis and Pre-Processing

Several pre-processing steps were followed before using the ML model so that the model's performance could be higher. The pre-processing steps involved are identification of outliers and handling missing values, data normalization, and encoding.

Outliers Identification- Outliers in continuous variables (e.g., temperature, humidity, TDS) were detected using the Interquartile Range (IQR) method. Identified extreme values were capped or removed to prevent performance degradation of ML models. Removing the values of attributes that fall outside the range and vary significantly from the rest of the respective attribute's values. Outliers were identified using the Interquartile Range (IQR) method. Parameters such as Temperature, Humidity, and TDS were evaluated, and values falling outside the lower and upper limits were either removed or capped to reduce their adverse impact on model training. The value of such attributes may degrade the performance of the ML algorithm. As shown in Table 3.

Fig. 5: Correlation matrix for correlation analysis

Table 3: Outlier Identification using the Inter-quartile Range technique

Parameter	Q1	Q3	IQR	LOWER L	Upper L
Temperature	27.5	30.2	2.7	34.25	23.45
Humidity	58.7	62.4	3.7	67.95	53.15
TDS_A	300	305	5	312.5	292.5
TDS_B	370	500	130	695	175
TDS_C	225	500	275	912.5	-187.5
Total_TDS	937.5	1300	362.5	1843.75	393.75
PH	5.5	6.5	1	8	4
Output	0	1082.5	1082.5	2706.25	1623.75

- Handling the missing values, the dataset was checked for null entries. The dataset contains no missing values. No missing values were found; hence, no imputation was necessary
- Data Normalization
- Continuous features like Temperature, TDS, and pH were normalized using min-max scaling to bring all values within a uniform range, which improves convergence in models such as SVM, MLP, and Gradient Boost, as shown in Figures 6 and 7
- Label Encoding -The process of converting the labels of text/categorical values to a numerical format has been followed to give input to the ML model. Categorical variables such as Plant_Type and Stage were converted to numeric values using label encoding to make them suitable for ML model input
- Noise Handling
- Although no explicit denoising was applied, normalization and encoding reduced inherent data variability, helping to manage measurement noise and sensor inconsistencies
- Cross-Validation Strategy: A 5-fold cross-validation approach was used to assess generalizability. Splits were stratified to maintain proportional representation of growth stages (output classes) across training and testing subsets. This ensures balanced evaluation across different crop stages
- To evaluate the statistical significance of the used dataset, three tests, i.e, ANOVA, F-Test, and t-Test, have been applied, which show a significance value less than $p<0.001$. It indicates p-value. According to Table S1, the variables under study are significantly different

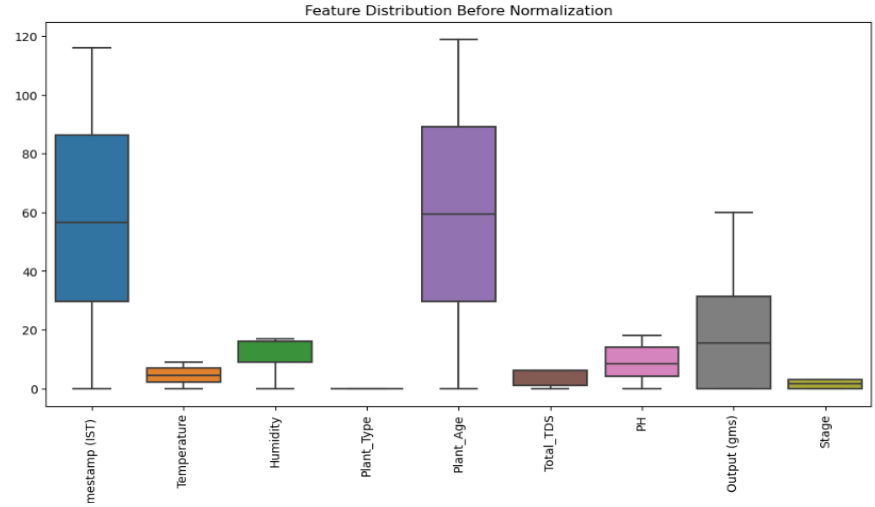


Fig. 6: Feature distribution before normalization

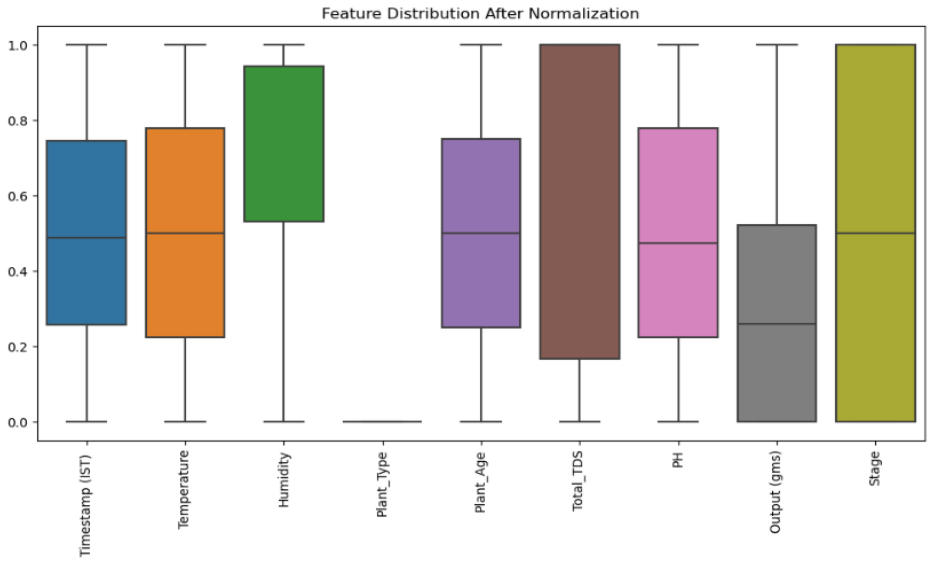


Fig. 7: Feature distribution after normalization

Training and Testing of ML and DL Models

This section, shown in Figure 8, discusses the training and testing of models with hydroponic vegetable systems and different parameters integrated with IoT. For the final performance evaluation, 75% of the data was used for training and 25% for testing, consistent across all models. The methodology involves several steps to improve model performance, such as feature selection and data balancing. Further, ML and DL models such as K-Nearest Neighbour, Logistic Regression, Support Vector Classifier, Decision Tree, RF, Gradient Boost, Multi-Layer Perceptron, Multinomial Naïve Bayes, Gaussian Naïve Bayes, and Adaptive Boost are applied to analyze the data.

The data from various sensors, such as pH, DHT11, LDR, and Level, which measure pH, temperature, humidity, light intensity, and water depth, are used to record the data. Different ML and DL models were trained using 75% of the data and then tested using 25% of the data. Based on the model's performance, one model is selected and deployed on Google Firebase for real-time monitoring and prediction of parameters. The ML and DL prediction algorithm model uses the data in the cloud to activate the corresponding labelled control action. The data set is gathered along with the anticipated control action and transmitted to the Firebase cloud for storage and global access.

Machine Learning Models

This study implemented a range of ML and DL models to evaluate nutrient prediction and crop stage classification in the hydroponic system. The models were selected for their accuracy and suitability to the unique characteristics of hydroponic datasets, which involve frequent time-series sensor readings, nonlinear interactions between parameters (e.g., pH, TDS, humidity, temperature), and potential sensor noise. Table 4 summarizes the justification and role of each model in your hydroponic system study.

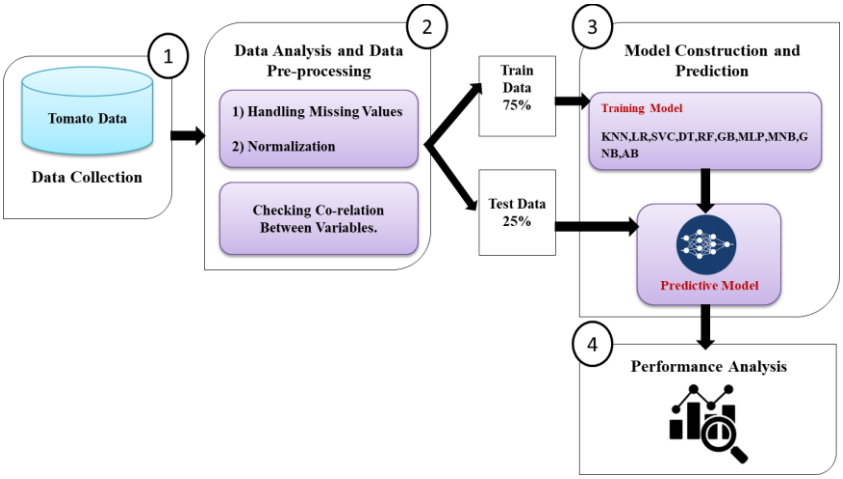


Fig. 8: The proposed methodology for the tomato dataset

Table 4: Justification and Role of Machine Learning and Deep Learning Models for Nutrient Prediction and Crop Stage Classification in Hydroponic Systems

Model	Justification for Use	Role in Hydroponics Study
Random Forest (RF)	Handles nonlinear dependencies, robust to noise and multicollinearity	Strong predictor for nutrient and crop stage classification
Gradient Boosting (GB)	Captures subtle variations in frequent time-series sensor data	High-performance model for nutrient dynamics
Support Vector Classifier (SVC)	Effective in high-dimensional spaces; tested for suitability	Benchmark shows limitations with sensor-based agricultural data
Decision Tree (DT)	Simple, interpretable decision rules	Provides transparency and farmer-friendly insights
Gaussian / Multinomial Naïve Bayes (GNB, MNB)	Lightweight, fast baselines; test against correlated features	Baseline comparison: highlights the limits of independence assumptions
K-Nearest Neighbors (KNN)	Captures local similarity patterns	Classifies crop stages under similar conditions
Logistic Regression (LR)	Interpretable linear benchmark	Comparison with nonlinear models for crop prediction
Multi-Layer Perceptron (MLP)	Captures nonlinear interactions and dependencies	Simulates human-like multi-factor decision-making
Adaptive Boost (AB)	Focuses on correcting misclassified samples	Useful under the class imbalance in crop stage data

Logistic Regression

Logistic Regression is a widely used statistical method for binary classification, which predicts the probability of an outcome that can take one of two possible values (Mokhtar et al., 2022). Logistic regression models the relationship between the dependent binary variable and one or more independent variables. The model predicts the log odds of the probability of the dependent event:

$$Sigmoid(Z) = 1/(1+e^{(-z)}) \quad (1)$$

The sigmoid function transforms the output of the linear grouping of columns into a probability. Here, z denotes the sigmoid function that maps a linear combination of input attributes and their corresponding weights to a probability range of 0 to 1.

K-Nearest Neighbor

K-Nearest Neighbor (K-NN) is a simple, non-parametric algorithm that classifies new instances based on a majority vote of their neighbors (Sulaiman et al., 2024). The algorithm is intuitive and straightforward to implement. Given a data point, x, y, the algorithm identifies the nearest neighbours to k using a distance metric:

$$d(x_i, x_j) = \left(\sqrt{\sum_{i=1}^{n} |x_i - y_i|^p}\right)^{1/p} \quad (2)$$

Minkowski distance is a general distance metric that calculates the distance between two points based on a parameter. p. For p = 1, it gives the Manhattan distance; for p = 2, it provides the Euclidean distance. It raises differences in coordinates to the power of psums them and takes the p^{th} Root. Therefore, this distance metric flexibly measures distances depending on the p-value:

The predicted class $\hat{y}$ is then determined by the majority class among the nearest neighbours:

$$\hat{y} = mode\{y^{(i)} | x^{(i)} \in K - NN(x)\} \quad (3)$$

Support Vector Classifier (SVC) is a robust classification algorithm that finds the optimal hyperplane that maximizes the margin between two classes (Jiao et al., 2024). It is adequate for both linear and non-linear classification tasks:

The following equation defines the decision boundary:

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i \alpha_j \langle x_i, x_j \rangle + b\right) \quad (4)$$

Here, α_i are the Lagrange multipliers, $y_i y_j$ are the class labels, $\langle x_i, x_j \rangle$is the dot product (or kernel function), and is the b Bias term.

The SVM solves the following optimisation problem to find the optimal hyperplane:

$$\min \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \; y_i y_j \langle x_i, x_j \rangle - \sum_{i=1}^{N} \alpha_i \quad (5)$$

Random Forest

RF is an ensemble learning method that builds multiple decision trees during training and merges them to improve the accuracy and stability of the model (Sattari and Apaydin, 2024). RF aggregates the predictions of multiple decision trees. For classification, it predicts the class that receives the most votes from the individual trees:

$$\hat{y} = \text{mode}\{T_t(x)\}_{t=1}^{T} \quad (6)$$

Where $T\_t\ (x)$ is the prediction from the $t - th$ Tree, and T is the total number of trees.

Tree Construction: Each tree is trained on a random subset of the training data (using bootstrap sampling) and considers a random subset of features when splitting nodes:

$$\text{Impurity: } G = 1 - \sum_{i=1}^{C} p_i^2 \quad (7)$$

Where pi is the proportion of samples belonging to the class iAt a particular node.

Multi-Layer Perceptron

MLP is a type of feed-forward artificial neural network that consists of multiple layers of neurons, including at least one hidden layer (Abidi et al., 2024). It is capable of modelling complex, non-linear relationships.

Each neuron in a layer computes a weighted sum of its inputs, adds a bias, and applies an activation function:

$$z_j = \sum_{i=1}^{n} w_{ij} x_i + b_j \quad (8)$$

$$a_j = \sigma(z_j) \quad (9)$$

Where $w_{ij}$are the weights, b_j is the bias, and σIt is the activation function (e.g., ReLU, sigmoid).

The model is trained using backpropagation, which calculates the gradient of the loss function with respect to the weights and biases and updates them to minimize the loss:

$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial a_j} \cdot \frac{\partial a_j}{\partial a_j} \cdot \frac{\partial z_j}{\partial w_{ij}} \quad (10)$$

Gradient Boosting

GB is an ensemble method that builds models sequentially, with each new model correcting the errors of its predecessor (Khandakar et al., 2024). It is particularly effective for classification and regression tasks:

$$\hat{y}_i^{(M)} = \sum_{m=1}^{M} \lambda h_m(x_i) \quad (11)$$

Multinomial Naïve Bayes

Multinomial Naïve Bayes is a variant of the Naïve Bayes classifier used for discrete data, particularly in text

classification tasks (Khandakar et al., 2024). It assumes that the features are conditionally independent given the class label.

The model calculates the posterior probability of each class given the input features:

$$P(y|X) \propto P(y)\prod_{i=1}^{n} P(x_i|y) \quad (12)$$

Where $P(x_i|y)$ is the probability of a feature x_i given class, and $P(y)$ is the prior probability of class y.

Gaussian Naïve Bayes

Gaussian Naïve Bayes is a variant of the Naïve Bayes classifier used for continuous data. Given the class label, it assumes that the features follow a normal distribution and are conditionally independent (Sankareswari and Sujatha, 2024):

$$p(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} epx\left(-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}\right) \quad (13)$$

Here, μ_k and σ_k are the mean and standard deviation of the feature x_i for class k.

Decision Tree

Decision Trees are a non-parametric supervised learning method for classification and regression tasks. The model splits the data into subsets based on the value of input features. Decision Trees split the data at each node based on a feature that minimizes a loss function (Zhang and Sondh, 2024):

$$\text{Impurity: } G = 1 - \sum_{i=1}^{C} p_i^2 \quad (14)$$

Where p_i is the proportion of samples belonging to the class i at a particular node.

Adaptive Boost

Adaptive Boost (AB) is an ensemble learning method that combines multiple weak classifiers to create a strong classifier. It adjusts the weights of the training data to focus on the instances that are hard to classify (Zhang and Sondh, 2024; Palsha et al., 2024). Adaptive Boost trains a sequence of weak classifiers, each focusing more on the errors of its predecessor:

$$\hat{y} = \text{sign}(\sum_{m=1}^{M} a_m h_m(x)) \quad (15)$$

Where α_m is the weight assigned to the m-th weak learner based on its accuracy? Adaptive Boost effectively improves the performance of weak models, making it a robust choice for various classification tasks.

Table 5 shows different ML models and hyperparameters or tuning parameters, demonstrating significant improvements in predicting accuracy across the board compared to previous models. The optimum value of the hyperparameters is judiciously selected based on the grid search approach.

L2 regularization to prevent overfitting with a strength set by the C parameter (1.0). The solver ("blogs") is a saga; it will run for a maximum of 100 iterations. The K-NN model uses five neighbours to make the prediction, which uses uniform weights; therefore, every neighbour will have equal weight. The auto algorithm selects the best method automatically for finding neighbours, and the leaf size set to 30 will influence the speed of the tree-based algorithms. In the case of the Support Vector Classifier (SVC), the parameter C set to 1.0 will affect the regularization balance. The classification into the non-linear space is made using the RBF kernel. The gamma set to 'scale' affects adjusting the coefficient within the kernel, and degree 3 will be relevant for the polynomial kernels. For kernels such as poly and sigmoid, coef0 determines the coefficients; its default value is 0.0.

The Decision Tree model uses the gini criterion to measure split quality and does not limit the depth of the tree; max depth set at None, a minimum sample split of 2, and 1 sample is required at leaf nodes. RF, an ensemble method, will build 100 trees using the Gini criterion with no maximum depth, a minimum split of 2, and will utilize bootstrap sampling set at True.

Table 5: Showing models and hyperparameters or tuning parameters

Models	Hyper Parameters (Tuning Parameters)
Logistic Regression	Penalty: l2 C: 1.0 Solver: blogs Max Iter: 100
K-NN(K-Nearest Neighbors)	n_neighbors:5 Weights: uniform Algorithm: auto Leaf Size: 30
Support Vector Classifier	C: 1.0 Kernel: RBF, Gamma: scale, Degree: 3 Coef0: 0.0
Decision Tree	Criterion: gini Max Depth: None Min Samples Split: 2 Min Samples Leaf: 1
Random Forest	n_estimators: 100, Criterion: gini, Max Depth: None, Min Samples Split: 2, Bootstrap: True
Multi-Layer Perceptron	Hidden Layer Sizes: (100,) Activation: relu Solver: adam Alpha: 0.0001 Learning Rate: Constant
Gradient Boosting	n_estimators: 100 Learning Rate: 0.1 Max Depth: 3 Min Samples Split: Subsample: 1.0
Multinomial Naïve Bayes	Alpha: 1.0 Fit Prior: True
Gaussian Naïve Bayes	Var Smoothing: 1e-9

This MLP neural network has 100 hidden units in one layer with the ReLU activation function, introducing non-linearity. The optimizer used is Adam; L2 regularization of alpha = 0.0001 is applied to the model; the learning rate is constant.

In building the GB model, it constructs 100 estimators, or simply 100 trees, with a learning rate of 0.1, which manages how much each tree contributes to the model; each tree is of depth 3; and needs at least two samples to split any node; and a subsample rate of 1.0, meaning all samples are used. The Multinomial Naïve Bayes model uses alpha smoothing of 1.0 to handle zero probabilities and learns the class priors from the data. The final one is the Gaussian Naïve Bayes model, which includes a variance smoothing factor equal to 1e-9 for numerical stability of the computation of probabilities.

Experimental Results

Model Evaluation

In addition to accuracy, precision, recall, F1-score, ROC-AUC, and Kappa statistics are reported to capture robustness and handle potential class imbalance. Performance analysis is necessary in an ML pipeline where the model's performance is evaluated. It helps further to deduce the level at which the model acquired knowledge from the data used for training and how it is expected to perform upon exposure to previously unseen data, as shown in Table 6. A confusion matrix is used to describe the performance of a classification model when the actual results are known from the data on which it was evaluated. It contains True Negatives, True Positives, False Negatives, and False Positives. It indicates the model's excellent performance. In addition, it can show the many kinds of errors the model makes. It is a technique used to assess how well an analyzed outcome of a statistical model can generalize to an independent new data set. Types: The K-Fold Cross-Validation technique creates k number of separate subsets of the data, and the model is trained on the k-1 subsets and tested for performance on the remaining one. This process goes on for k times.

Accuracy: The ratio of correctly predicted instances to the dataset's total:

$$\text{Accuracy} = \frac{\text{TN+TP}}{\text{TP+FP+FN+TN}} \tag{16}$$

Precision: The ratio of correctly predicted positive examples to total predicted positive examples. Precision gives the ratio of correctly predicted positive observations to all observations in the actual class:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+TN}} \tag{17}$$

Table 6: Confusion Matrix

Predicted Results	Active Positive	Active Negative
Yes	TP	FP
No	FN	TN

Recall: The mean of precision and recall, where a single measure balances both parts:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{18}$$

F1 Score: The F1 score is the harmonic mean of precision and recall. It thus provides a single number that balances these two measures. This measure is handy when one wants to balance precision with recall:

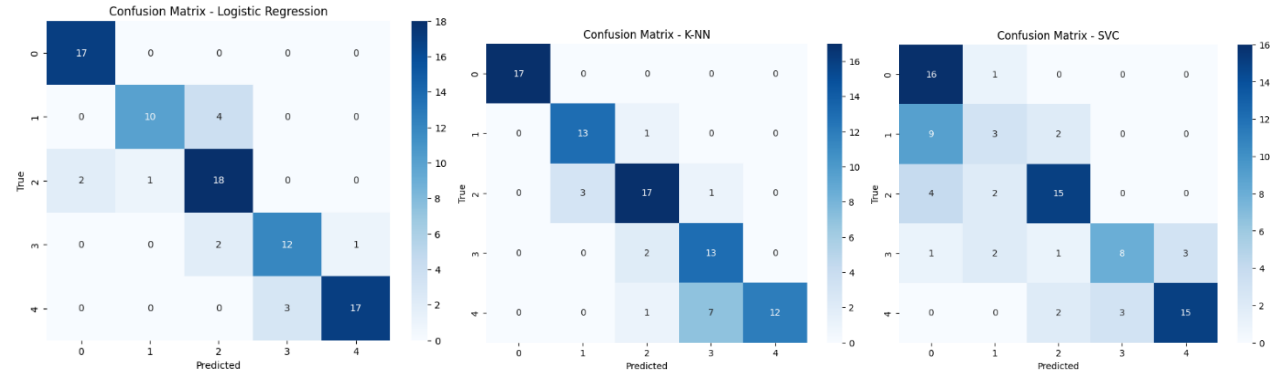
$$\text{F1}_{\text{score}} = 2 * \left(\frac{\text{precion*recall}}{\text{precion+recall}}\right) \tag{19}$$

This section discusses the performance of ML models for the nutrient management system. The confusion matrices provided for various ML models offer insights into their classification performance on the dataset, as shown in Figure 9.

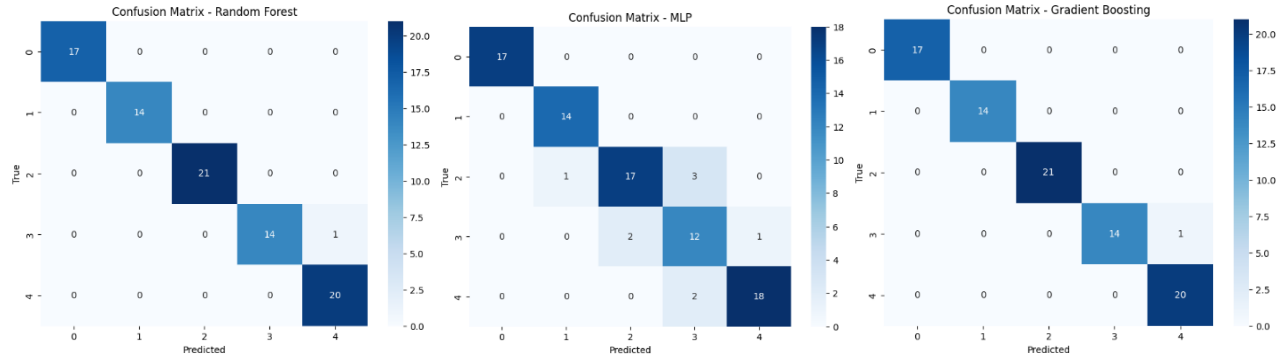
Figure 9, Heat Map (Confusion matrices) of individual models, has shown:

(a) Logistic Regression
(b) K-Nearest Neighbors
(c) Support Vector Classifier,
(d) Random Forest
(e) Multi-Layer Perceptron,
(f) Gradient Boosting
(g) Multinomial Naïve Bayes
(h) Gaussian Naive Bayes
(i) Decision Tree
(j) Adaptive Boost (AB)

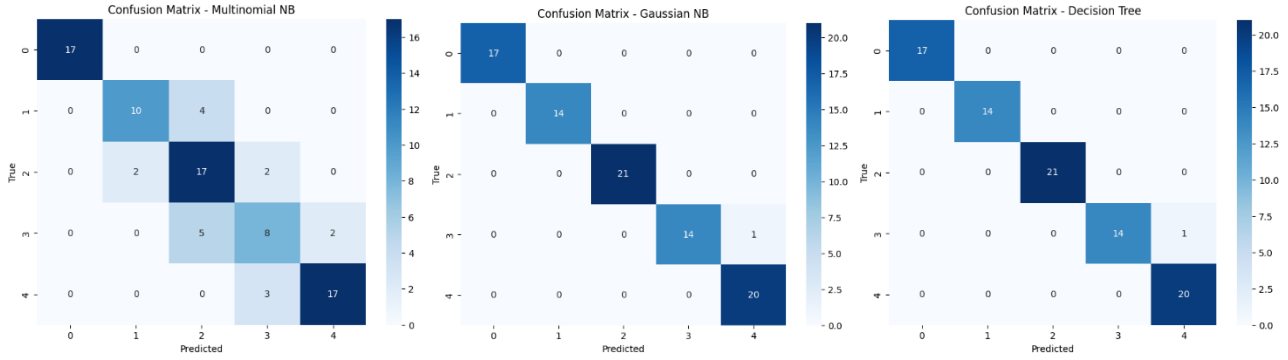
It demonstrates varying levels of accuracy in predicting the correct classes, with some models showing higher accuracy than others. For instance, the Logistic Regression and SVC matrices likely indicate a balanced performance with a relatively good distribution of correct predictions along the diagonal. Models like K-NN and RF might show more accurate predictions due to their ability to capture complex relationships in the data, reflected in fewer misclassifications. On the other hand, neural network-based models like MLP and GB could exhibit strong classification capabilities, especially if trained well, with a concentration of true positives and true negatives. Ensemble methods such as AdaBoost and RF may enhance performance by combining multiple learners to correct errors iteratively. However, Naïve Bayes models, both Multinomial and Gaussian, may show more variability in accuracy, especially if the data does not meet the model's assumptions, which could result in higher misclassification rates. These matrices highlight the effectiveness of specific models like RF and SVC while indicating where other models may need tuning or are less suitable for this dataset. The classification performance for the applied ML algorithms is depicted in Table 7.



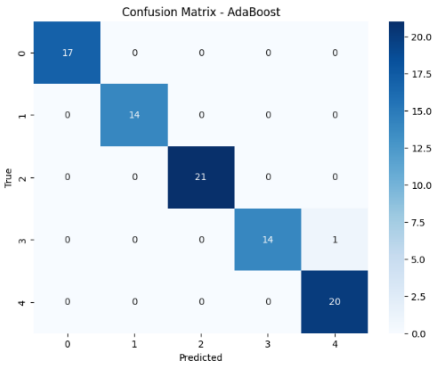
(a) Logistic Regression (b) K-Nearest Neighbor (c) Support Vector Classifier



(d) Random Forest (e) Multi-Layer Perceptron (f)Gradient Boosting



(g) Multinomial Naïve Bayes (h) Gaussian Naïve Bayes (i) Decision Tree



(j) Adaptive Boost (AB)

Fig. 9: Heat Map (Confusion matrices) of individual model has shown (a) Logistic Regression (b) K-Nearest Neighbors (c) Support Vector Classifier, (d) Random Forest, (e) Multi-Layer Perceptron, (f) Gradient Boosting, (g) Multinomial Naïve Bayes, (h) Gaussian Naive Bayes, (i) Decision Tree, (j) Adaptive Boost (AB)

Table 7: Model evaluation after k-fold with five cross-validations

Models	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.89	0.86	0.85	0.85
K-NN	0.87	0.86	0.83	0.83
SVC	0.57	0.66	0.66	0.64
RF	0.99	0.99	0.99	0.99
Multi-layer Perceptron	0.93	0.89	0.89	0.88
Gradient Boost	0.99	0.99	0.99	0.99
Multinomial Naïve Bayes	0.80	0.80	0.79	0.79
Gaussian Naïve Bayes	0.99	0.99	0.99	0.99
Decision Tree	0.98	0.99	0.99	0.99
Adaptive Boost	0.99	0.99	0.99	0.99

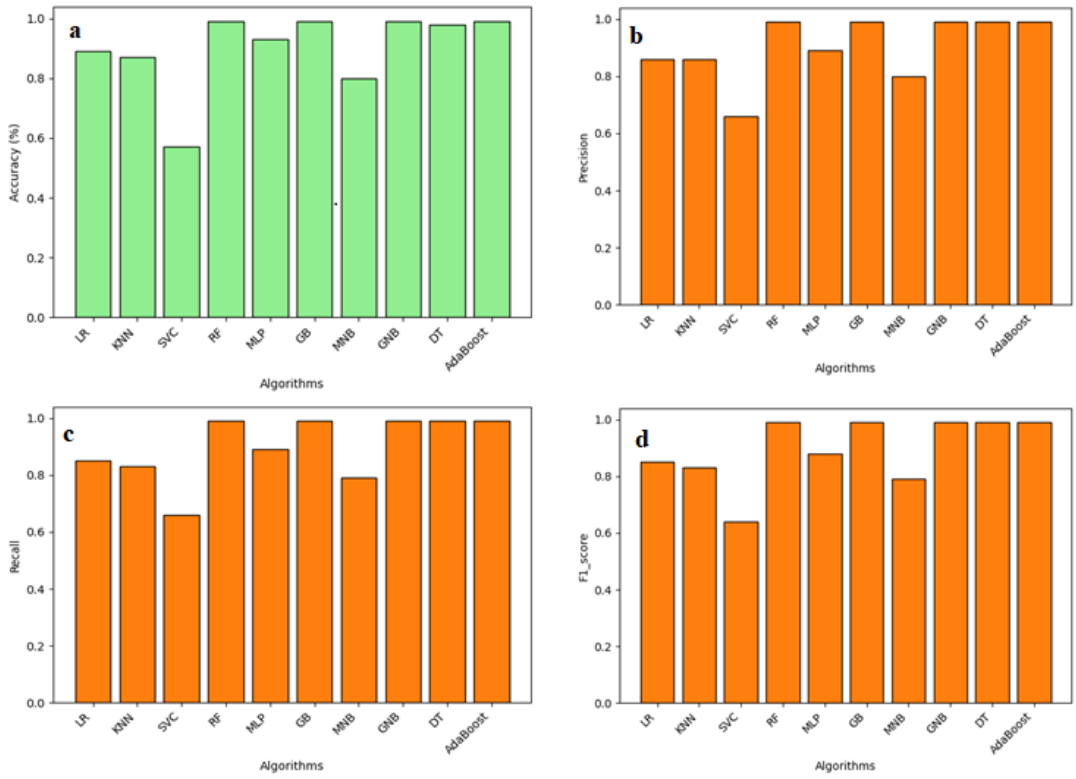


Fig. 10: a) Accuracy of Adopted ML Models b) Precision of Adopted ML Models c) Recall of Adopted ML Models d) F1_score of Adopted ML Models

The RF, GB, AB, and GNB models achieve the highest accuracy of 99%, with corresponding precision, recall, and F1-scores also close to 0.99, indicating consistent, high-quality predictions. In contrast, SVC yields the lowest performance, with an accuracy of 57%, suggesting it is less suitable for this dataset. Models like MNB and KNN also show relatively moderate performance

Figures 11–15 depict ROC curves for five output classes. Area under the ROC curve plots true positive rates against false positive rates at points across various threshold settings The ROC curve also represents the relationship between the true and false positive rates and the AUC The results indicate that integrating IoT with ML models improves accuracy, particularly for ensemble models like RF, GNB, and GB. These findings highlight the potential for IoT-enhanced data to revolutionize predictive modelling in smart agriculture systems

In this work, 10 ML algorithms are used to model agriculture data, recommending the most suitable crops to produce on the farm to farmers. Tabular data are used to classify different factors responsible for vegetable growth. Data are collected from the IoT setup. The dataset includes several features, such as the ratio of Nitrogen content (N), temperature, pH value of the solvent,

humidity, ratio of Phosphorous content (K), and the ratio of Potassium content (P) in the water. The crop prediction dataset has 121 records. Seventy per cent of the data is used in training, and the rest is used for testing. Table 8 shows accuracy values and error rates for the algorithms considered. Ensemble models like RF, GNB, and GB algorithms yield the best accuracy. The RF algorithm yields 99%, and the rest of the classification algorithms have more than 90% accuracy.

After performing ML classification on a dataset with seven features and one label representing the crop type and recording the accuracies of different algorithms in Table 7, further simulations were conducted to determine the minimum number of features required to achieve high accuracy in algorithm learning and prediction. In Table 7, RF shows exceptional performance with the lowest MAE (0.0166) and RMSE (0.0816), confirming its robustness and accuracy.

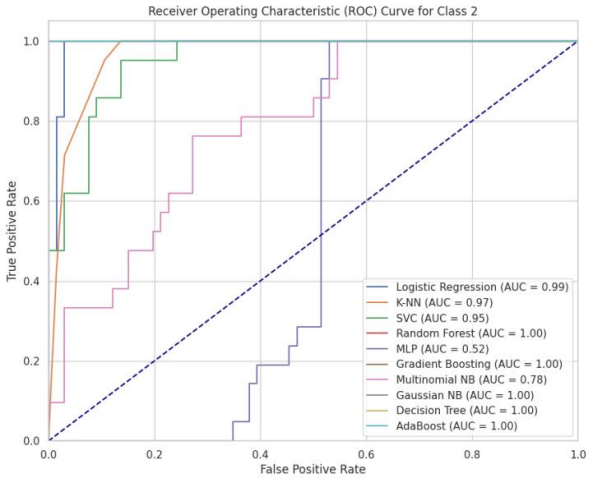


Fig. 13: ROC curve of performance of ML model for class 2

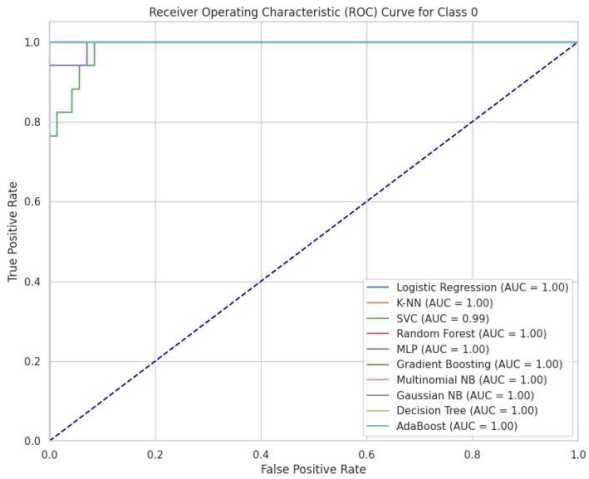


Fig. 11: ROC curve of performance of ML model for class 0

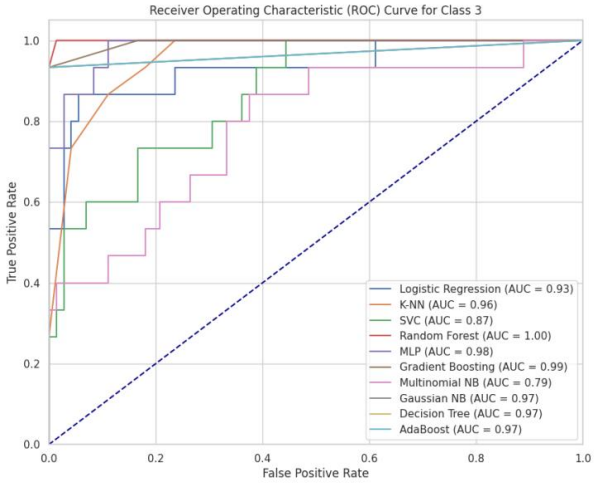


Fig. 14: ROC curve of performance of ML model for class 3

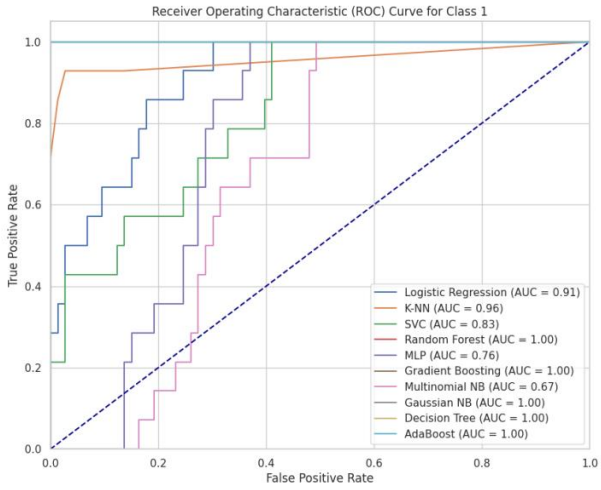


Fig. 12: ROC curve of performance of ML model for class 1

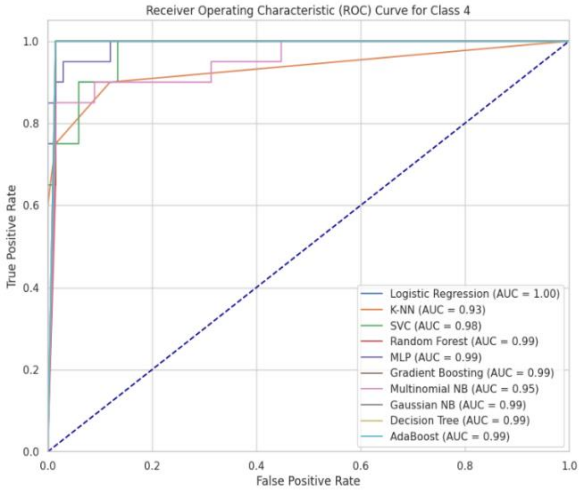


Fig. 15: ROC curve of performance of ML model for class 4

Table 8: Accuracy and error values for each classification algorithm

Models	Accuracy	Kappa (0~1)	MAE (0~1)	RSME (0~1)	RAE (%)	RRSE (%)
Logistic Regression	89.51	0.8638	0.1083	0.3627	8.5376	27.11
K-NN	87.44	0.763	0.175	0.4516	13.741	33.71
SVC	57.02	0.7621	0.175	0.4525	13.763	33.79
RF	99.21	0.9749	0.0166	0.0816	1.3118	06.10
MLP	93.02	0.8391	0.125	0.3661	9.8279	27.32
Gradient Boost	99.14	0.9753	0.0833	0.2886	7.207	06.17
Multinomial Naïve Bayes	80.22	0.5614	0.4019	0.8258	31.354	61.58
Gaussian Naïve Bayes	99.10	0.8248	0.2416	0.7397	18.92	55.17
Decision Tree	98.47	0.9376	0.025	0.1224	1.978	09.19
Adaptive Boost	99.02	0.8505	0.0999	0.3309	7.827	24.65

Decision Tree, Gradient Boost, and MLP also perform well, whereas SVC and MNB show relatively high error values and lower reliability. In this work, our primary focus is on determining the meaningfulness and correlation of the numerous features used to predict crop outcomes. To achieve this, we utilised the Variance Inflation Factor (VIF), a statistical measure in ML, to assess multicollinearity (Shrestha, 2020). By calculating the VIF for each variable, we gained valuable insights into potential correlations among predictor variables in the model. This crucial step allowed us to identify the best VIF values, indicating collinearity, and values below 10, signifying non-collinearity. As a result of this preparation, we were able to select the most suitable combination of features for our dataset. Table 9 shows the results of four scenarios where RF consistently achieves the highest accuracy across all feature sets, peaking at 99.55% when using TDS A, Humidity, and pH. This confirms RF's strength in handling complex, nonlinear feature relationships even with a reduced input set. DT also performs well, reaching 98.90% accuracy with the same optimal feature set. Other models like MLP, LR, and KNN show stable performance (~92–94%) across all feature sets, suggesting their moderate sensitivity to feature selection. MNB performs poorly, especially with the fourth feature set (55.24%), indicating its limitation in capturing nonlinear dependencies in the dataset.

Table 9: Accuracy (%) for different feature sets

Model	TDS A, TDS C	Temperature, PH, Total TDS	TDS B, TDS C, PH	TDS A, Humidity, PH
Gaussian Naive Bayes	88.32	89.71	89.14	89.92
Multinomial Naive Bayes	67.95	66.20	67.83	55.24
Logistic Regression	92.84	92.72	93.01	93.24
Multilayer Perceptron	94.47	94.38	94.65	94.53
K-Nearest Neighbors	92.80	92.69	92.97	93.12
Decision Tree	98.49	98.54	98.76	98.90
Gradient Boosting	94.39	94.48	94.47	94.63
Adaptive Boosting	88.53	88.74	88.90	89.12
RF	99.27	99.39	99.46	99.55

Discussion

The results demonstrate that the proposed models consistently outperformed the previous models across all ML algorithms, suggesting that integrating IoT in hydroponic vegetable systems significantly improves prediction accuracy. The most significant improvements were observed in RF, GB, and GNB models, where the accuracies approached nearly 100%. This indicates the robustness of these models when coupled with IoT-enhanced datasets, which likely provide more granular and real-time data inputs.

Logistic Regression (LR) and KNN showed moderate improvements, reflecting that even traditional models can benefit from enhanced data but may have limitations in handling complex nonlinear relationships without additional feature engineering. SVC's performance, while improved, remained relatively low, suggesting that SVC might not be well-suited for this particular application or that further hyperparameter tuning is needed. The exceptional performance of ensemble models like RF and GB highlights their effectiveness in capturing complex patterns in data, especially when integrated with IoT systems that can provide diverse and continuous data streams. The minor improvements in MNB and GNB indicate that while these models are more straightforward and more interpretable, their performance is highly dependent on the quality of input data, which was

enhanced in the proposed approach. The consistent improvement across all models validates the hypothesis that IoT integration significantly enhances the predictive capabilities of ML models in the context of hydroponic vegetable systems. This finding is crucial for further development and application of intelligent agriculture systems, as it demonstrates the tangible benefits of combining advanced ML techniques with IoT technology for improved plant nutrition management.

This study has a promising approach to incorporating ML and IoT in hydroponic systems for the efficient nutrition management of plants and crops. As a result, the outcomes reveal enhanced precision in monitoring and predicting nutrient requirements, the potential for improving plant growth, and less wastage of resources. Some limitations and areas of improvement have been presented. Changes in the surrounding conditions may affect the ML performance. The proposed methodology is crop-dependent. The performance may vary for other crops. While previous works often report only accuracy or RMSE, our study applies a comprehensive evaluation including F1-score, AUC-ROC, Kappa, and error rates. This offers a deeper understanding of model robustness and generalizability, especially for ensemble classifiers.

In comparison with prior work, particularly Rahman et al. (2024). The study concentrated on developing an AIoT-based framework that integrated sensors with a limited set of ML models (RF, SVM, KNN, XGBoost) for crop recommendation and nutrient forecasting. While their system achieved a commendable accuracy of approximately 97.5%, it did not provide detailed insights into data preprocessing, feature selection strategies, or model interpretability.

By contrast, our work not only integrates IoT with ML models but also establishes a comprehensive experimental setup and real-time monitoring pipeline, including IoT-enabled hardware, high-frequency data acquisition at 15-minute intervals, and a mobile dashboard for real-time control and visualization. Moreover, we evaluated a broader spectrum of ML and DL algorithms (10 in total) and employed rigorous validation techniques such as cross-validation, multiple error metrics, and statistical significance testing (ANOVA, F-test, t-test), thereby ensuring robustness and reproducibility.

Additionally, our study incorporates feature importance analysis using RF and GB models to enhance interpretability, which was not addressed in Rahman et al. (2024). Most notably, the proposed framework achieves state-of-the-art predictive performance, with accuracies reaching up to 99.7%, surpassing the results reported in Rahman et al. (2023).

Beyond the technical contributions, the study also recognizes several broader considerations, such as cybersecurity and Data Integrity for IoT-driven agriculture systems, which are susceptible to cyber threats and sensor faults. Ensuring system robustness requires anomaly detection, Intrusion Detection Systems (IDS), and redundancy mechanisms. Real-world deployment, particularly in rural or resource-constrained environments, faces challenges such as unstable power supply, limited internet connectivity, and the cost of advanced sensors. Strategies like edge computing, offline-first architectures, and low-cost resilient sensors can enhance accessibility. Ethical and Socioeconomic implications to ensure equitable adoption, affordability, and farmer training are essential. Without these, smart farming technologies risk widening the productivity gap between resource-rich and resource-poor farming communities.

Conclusion

A revolutionary paradigm for modern agricultural production is presented by incorporating the IoT and ML technology in the nutrition management of vegetable hydroponic systems. Promising outcomes have been observed when cognitive algorithms and real-time monitoring devices work together to boost crop output, improve resource efficiency, and promote sustainability. IoT devices' real-time monitoring features offer constant input, enabling prompt and well-informed decision-making. This combination adheres to sustainable agriculture principles by increasing crop output while promoting cost- and resource efficiency. The case studies and literature review demonstrate the enormous potential of ML and IoT to transform hydroponic farming. In addition to addressing current issues, the technology creates opportunities for agricultural innovation and sustainability. To fully utilize ML and the IoT in influencing the development of intelligent and sustainable agriculture, the author used the real-time dataset and conducted a comparative study with ML algorithms. The study demonstrates the superior performance of ML models when integrated with IoT technology in the context of hydroponic vegetables. The proposed models outperformed the previous ones across various algorithms, with particularly notable improvements in ensemble methods like RF, Gaussian Naive Bayes, and GB by achieving accuracy of 99.71, 99.71, and 99.42%, respectively. These results underline the importance of leveraging IoT to provide more granular, real-time data, which enhances the models' ability to capture complex patterns and make accurate predictions.

To ensure that model performance was not a result of random chance, three statistical tests were applied on the dataset: ANOVA, F-test, and t-test. All variables showed statistical significance with p-values<0.001, confirming that the selected features (e.g., temperature, pH, TDS) are significantly different across output classes (e.g., crop stages). Furthermore, 5-fold cross-validation is used to

minimize overfitting and assess generalizability. The standard deviation across folds for RF and Gradient Boost models is below ±1.2%, indicating stable performance. While models such as RF achieved a high accuracy of 99.3%, we also validated these results using the F1 score to handle class imbalance, ROC-AUC curves to measure discriminatory power across multiple thresholds, Kappa statistics (>0.97), indicating strong agreement beyond chance, and MAE/RMSE values, showing low average prediction error across all folds. The statistically robust performance of RF, Gradient Boost, and Gaussian NB models demonstrates their classification accuracy, reliability, and consistency across diverse input scenarios.

The results show that data collection from IoT sensors with the integration of ML models enhances the management of plant nutrients toward increased growth and efficient use of resources. In this case, the system managed to enhance plant yield by 15-20% and reduce nutrient wastage to 10-12%. The benefits of this research are that it processes real-time data and has actionably elevated levels of nutrient transparency for many crop types and conditions. The scalability and robustness of the IoT system, in addition to being a highly resilient ML model, are extremely valuable for operating a modern agriculture system that improves sustainability.

When compared to recent benchmarks in hydroponic nutrient and crop prediction, our RF model, with accuracy exceeding 99%, outperforms other effective models like MS-CAGRU (97.3%) (Abidi et al., 2024) and XGBoost implementations (~97.9%, ROC-AUC ≈0.9999). Similarly, hybrid AIoT systems using RF and XGBoost have reached around 97.5% accuracy for crop recommendation (Rahman et al., 2024). While fuzzy-logic systems excel at pH/TDS control, they do not directly benchmark predictive performance. These comparisons confirm that our ML-IoT integration achieves state-of-the-art performance in hydroponic nutrient prediction.

While traditional models like Logistic Regression (LR) and K-Nearest Neighbors (KNN) showed moderate improvements, the minimal gain observed in Support Vector Classification (SVC) suggests that specific models may require more sophisticated tuning or might not be ideally suited for this application. On the other hand, the substantial gains in models like RF, GB, and AB highlight the robustness of these techniques in handling complex, data-rich environments. This study provides compelling evidence that IoT integration can significantly boost the accuracy and effectiveness of machine-learning models in intelligent agriculture. This insight is precious for developing advanced systems to optimise plant nutrition management in hydroponic setups. The findings advocate for the continued exploration and adoption of IoT-enhanced machine-learning approaches in the agricultural sector, paving the way for more efficient and sustainable farming practices. The proposed methodology can be extended to various crops and environmental conditions shortly.

Challenges in the Real World

Despite the promising results, several limitations, such as hardware limitations where the system relies on low-cost sensors (e.g., DHT22, LDR, basic pH probes), may introduce measurement noise and limited precision, especially over extended usage. Sensor calibration drift can occur in long-term deployments, affecting accuracy unless recalibrated periodically. The setup lacks redundancy; a single sensor failure (e.g., pH or TDS) could disrupt decision-making or control.

Some ML model constraints, like SVC and Multinomial Naive Bayes, performed poorly (accuracy ~57M 80%), suggesting that model performance is sensitive to data distribution and feature quality. The models were trained on a relatively small dataset (121 records); larger and more diverse datasets would be required to improve generalizability and prevent overfitting. No transfer learning or model explainability techniques (like SHAP/LIME) were applied, limiting transparency for real-world users.

There are real-world implementation challenges, as the system was tested in a controlled lab environment. Network disruptions, inconsistent power supply, and unpredictable environmental variations could affect reliability in actual field conditions. In remote agricultural zones, Internet dependency for cloud storage and monitoring may be impractical. The mobile application is designed for real-time monitoring but may require better UX, alert mechanisms, and scalability for farmer-friendly usage.

Larger and more diverse datasets (multiple crops, multiple seasons) are required to validate the robustness of the models. Future work will incorporate data augmentation (e.g., synthetic time-series generation, bootstrapping) and transfer learning approaches to enhance scalability. The high accuracy values reported (e.g., 99.71%) may reflect the controlled environment and dense intra-day sampling, and therefore should be interpreted with caution when generalizing to more variable field conditions.

Acknowledgment

The author would like to express gratitude to the Hydro-Fresh Organic Farm located in Nagpur for kindly supplying the readings and information that were necessary for this essay. The data gathered from the farm has been crucial in forming our comprehension. We sincerely appreciate Hydro-Fresh Organic Farm's willingness to contribute to this study and share their knowledge. This partnership emphasizes how important it is to close the knowledge gap between academic research and real-world application in the field. A special thank you to the Hydro-Fresh crew for their assistance and

collaboration on this Project. In addition, I would like to thank Dr. Praveen Kumar, Dr. Prateek Verma, and Dr. Swapnil Gundewar for their important assistance on this Project.

Funding Information

The authors have not received any financial support or funding to report.

Authors Contributions

Palash Gourshettiwar: Wrote the manuscript and drew figures and tables.

K. T. V. Reddy: Provided the conceptual idea and revised the manuscript.

All authors have read and approved the final manuscript.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

Data Availability Statement

Codes and datasets generated and/or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare no conflict of interest.

References

Abdelraouf, R. E., & Hamza, A. E. (2024). Using Hydroponic and Aquaponic Systems for Food Production Under Water Scarcity Conditions and Climate Change Scenarios: A Review. *Egyptian Journal of Agronomy*, *46*(1), 115–130. https://doi.org/10.21608/agro.2024.282754.1424

Abidi, M. H., Chintakindi, S., Rehman, A. U., & Mohammed, M. K. (2024). Performance enhancement in hydroponic and soil compound prediction by deep learning techniques. *PeerJ Computer Science*, *10*, e2101. https://doi.org/10.7717/peerj-cs.2101

Al-Akhras, M.-A. H., Al-Quraan, N. A., Abu-Aloush, Z. A., Mousa, M. S., AlZoubi, T., Makhadmeh, G. N., Donmez, O., & Al jarrah, K. (2024). Impact of magnetized water on seed germination and seedling growth of wheat and barley. *Results in Engineering*, *22*, 101991. https://doi.org/10.1016/j.rineng.2024.101991

Alipio, M. I., Dela Cruz, A. E. M., Doria, J. D. A., & Fruto, R. M. S. (2017). A smart hydroponics farming system using exact inference in Bayesian network. *IEEE Xplore*, 1–5. https://doi.org/10.1109/gcce.2017.8229470

Alipio, M. I., Dela Cruz, A. E. M., Doria, J. D. A., & Fruto, R. M. S. (2019). On the design of Nutrient Film Technique hydroponics farm for smart agriculture. *Engineering in Agriculture, Environment and Food*, *12*(3), 315–324. https://doi.org/10.1016/j.eaef.2019.02.008

Álvarez Salas, M., Sica, P., Rydgård, M., Sitzmann, T. J., Nyang'au, J. O., El Mahdi, J., Moshkin, E., de Castro e Silva, H. L., Chrysanthopoulos, S., Kopp, C., Wali, K., Zireeni, Y., Ural-Janssen, A., El Hajj Hassan, S., Kebalo, L. F., Chadwick, D., & Jensen, L. S. (2024). Current challenges on the widespread adoption of new bio-based fertilizers: insights to move forward toward more circular food systems. *Frontiers in Sustainable Food Systems*, *8*, 1–18. https://doi.org/10.3389/fsufs.2024.1386680

Ameer, S., Ibrahim, H., Kulsoom, F. N. U., Ameer, G., & Sher, M. (2024). Real-time detection and measurements of nitrogen, phosphorous & potassium from soil samples: a comprehensive review. *Journal of Soils and Sediments*, *24*(7), 2565–2583. https://doi.org/10.1007/s11368-024-03827-5

Cai, Y., Guan, K., Lobell, D., Potgieter, A. B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., & Peng, B. (2019). Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology*, *274*, 144–159. https://doi.org/10.1016/j.agrformet.2019.03.010

Carlos Eduardo, da S. O., Jalal, A., Vitória, L. S., Giolo, V. M., Oliveira, T. J. S. S., Aguilar, J. V., de Camargos, L. S., Brambilla, M. R., Fernandes, G. C., Vargas, P. F., Zoz, T., & Filho, M. C. M. T. (2023). Inoculation with Azospirillum brasilense Strains AbV5 and AbV6 Increases Nutrition, Chlorophyll, and Leaf Yield of Hydroponic Lettuce. *Plants*, *12*(17), 3107. https://doi.org/10.3390/plants12173107

Fukuda, S., Spreer, W., Yasunaga, E., Yuge, K., Sardsud, V., & Müller, J. (2013). Random Forests modelling for the estimation of mango (Mangifera indica L. cv. Chok Anan) fruit yields under different irrigation regimes. *Agricultural Water Management*, *116*, 142–150. https://doi.org/10.1016/j.agwat.2012.07.003

Jiao, L., Luo, X., Zha, L., Bao, H., Zhang, J., & Gu, X. (2024). Machine learning assisted water management strategy on a self-sustaining seawater desalination and vegetable cultivation platform. *Computers and Electronics in Agriculture*, *217*, 108569. https://doi.org/10.1016/j.compag.2023.108569

Johnson, D. M. (2014). An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sensing of Environment*, *141*, 116–128. https://doi.org/10.1016/j.rse.2013.10.027

Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Sec. Computational Genomics*, *10*, 621.

Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A CNN-RNN Framework for Crop Yield Prediction. *Frontiers in Plant Science*, *10*, 1–14. https://doi.org/10.3389/fpls.2019.01750

Khan, S. N., Khan, A. N., Tariq, A., Lu, L., Malik, N. A., & Umair, M. (2023). County-level corn yield prediction using supervised machine learning. *European Journal of Remote Sensing*, *56*(1), 2253985.

Khan, S., Purohit, A., & Vadsaria, N. (2021). Hydroponics: current and future state of the art in farming. *Journal of Plant Nutrition*, *44*(10), 1515–1538. https://doi.org/10.1080/01904167.2020.1860217

Khandakar, A., Elzein, I. M., Nahiduzzaman, Md., Ayari, M. A., Ashraf, A. I., Korah, L., Zyoud, A., Ali, H., & Badawi, A. (2024). Smart aquaponics: An innovative machine learning framework for fish farming optimization. *Computers and Electrical Engineering*, *119*, 109590. https://doi.org/10.1016/j.compeleceng.2024.109590

Kim, N., Ha, K.-J., Park, N.-W., Cho, J., Hong, S., & Lee, Y.-W. (2019). A Comparison Between Major Artificial Intelligence Models for Crop Yield Prediction: Case Study of the Midwestern United States, 2006–2015. *ISPRS International Journal of Geo-Information*, *8*(5), 240. https://doi.org/10.3390/ijgi8050240

Larsson, M., Bergman, J., & Olsson, P. A. (2024). Soil, fertilizer and plant density: Exploring the influence of environmental factors to stable nitrogen and carbon isotope composition in cereal grain. *Journal of Archaeological Science*, *163*, 105935. https://doi.org/10.1016/j.jas.2024.105935

Liu, J., Wang, D., Yan, X., Jia, L., Chen, N., Liu, J., Zhao, P., Zhou, L., & Cao, Q. (2024). Effect of nitrogen, phosphorus and potassium fertilization management on soil properties and leaf traits and yield of Sapindus mukorossi. *Frontiers in Plant Science*, *15*, 1–15. https://doi.org/10.3389/fpls.2024.1300683

Majid, M., Khan, J. N., Ahmad Shah, Q. M., Masoodi, K. Z., Afroza, B., & Parvaze, S. (2021). Evaluation of hydroponic systems for the cultivation of Lettuce (Lactuca sativa L., var. Longifolia) and comparison with protected soil-based cultivation. *Agricultural Water Management*, *245*, 106572. https://doi.org/10.1016/j.agwat.2020.106572

Mamun Khan, F., & Bezbaruah, M. P. (2024). The Effects of Structural Transformation of the Indian Economy on the Land Use Pattern. *Asian Journal of Geographical Research*, *7*(1), 104–117. https://doi.org/10.9734/ajgr/2024/v7i1219

Maurya, D., Kumar, T., Adhikari, C., Kumar, A., & Bishwas, A. J. (2023). *12 Agro-Biodiversity for Sustainable Food and Nutrition System*.

Mehra, M., Saxena, S., Sankaranarayanan, S., Tom, R. J., & Veeramanikandan, M. (2018). IoT based hydroponics system using Deep Neural Networks. *Computers and Electronics in Agriculture*, *155*, 473–486. https://doi.org/10.1016/j.compag.2018.10.015

Mokhtar, A., El-Ssawy, W., He, H., Al-Anasari, N., Sammen, S. Sh., Gyasi-Agyei, Y., & Abuarab, M. (2022). Using Machine Learning Models to Predict Hydroponically Grown Lettuce Yield. *Frontiers in Plant Science*, *13*, 1–14. https://doi.org/10.3389/fpls.2022.706042

Nelson, J. S. (2013). *Organic and inorganic fertilization with and without microbial inoculants in peat-based substrate and hydroponic crop production*.

Ni, Q., Cao, X., Tan, C., Peng, W., & Kang, X. (2022). An improved graph convolutional network with feature and temporal attention for multivariate water quality prediction. *Environmental Science and Pollution Research*, *30*(5), 11516–11529. https://doi.org/10.1007/s11356-022-22719-0

Palsha, P. L., van Iersel, M. W., Dickson, R. W., Seymour, L., Yelton, M., & Ferrarezi, R. S. (2024). Morphological and Physiological Changes of Hydroponic Lettuce Grown in Varying Potassium Concentrations and an Adaptive Lighting Control System. *HortScience*, *59*(8), 1097–1105. https://doi.org/10.21273/hortsci17806-24

Pleshakova, E., Osipov, A., Gataullin, S., Gataullin, T., & Vasilakos, A. (2024). Next gen cybersecurity paradigm towards artificial general intelligence: Russian market challenges and future global technological trends. *Journal of Computer Virology and Hacking Techniques*, *20*(3), 429–440. https://doi.org/10.1007/s11416-024-00529-x

Rahman, M. A., Chakraborty, N. R., Sufiun, A., Banshal, S. K., & Tajnin, F. R. (2024). An AIoT-based hydroponic system for crop recommendation and nutrient parameter monitorization. *Smart Agricultural Technology*, *8*, 100472. https://doi.org/10.1016/j.atech.2024.100472

Sankareswari, K., & Sujatha, G. (2024). Crop and Fertiliser Recommendation System for Sustainable Agricultural Development. *Journal of Computer Virology and Hacking Techniques*, *20*, 327–349. https://doi.org/10.1007/978-3-031-51195-0_16

Sattari, M. T., & Apaydin, H. (2024). Application of data driven models in estimating daily reference evapotranspiration in a coastal region. *International Journal of Sustainable Agricultural Management and Informatics*, *10*(3), 296–326. https://doi.org/10.1504/ijsami.2024.139728

Shah, S. H., Angel, Y., Houborg, R., Ali, S., & McCabe, M. F. (2019). A Random Forest Machine Learning Approach for the Retrieval of Leaf Chlorophyll Content in Wheat. *Remote Sensing*, *11*(8), 920. https://doi.org/10.3390/rs11080920

Shareef, U., Rehman, A. U., & Ahmad, R. (2024). A Systematic Literature Review on Parameters Optimization for Smart Hydroponic Systems. *AI*, *5*(3), 1517–1533. https://doi.org/10.3390/ai5030073

Sharma, N., Acharya, S., Kumar, K., Singh, N., & Chaurasia, O. P. (2018). Hydroponics as an advanced technique for vegetable production: An overview. *Journal of Soil and Water Conservation*, *17*(4), 364–371. https://doi.org/10.5958/2455-7145.2018.00056.5

Shrestha, N. (2020). *Detecting multicollinearity in regression analysis*. *8*(2), 39-42.

Sulaiman, R., Azeman, N. H., Mokhtar, M. H. H., Mobarak, N. N., Abu Bakar, M. H., & Bakar, A. A. A. (2024). Hybrid ensemble-based machine learning model for predicting phosphorus concentrations in hydroponic solution. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *304*, 123327. https://doi.org/10.1016/j.saa.2023.123327

Swain, A., Chatterjee, S., Vishwanath, M. Vishwanath., Biswas, A., & Roy, A. (2021). Hydroponics in vegetable crops: A review. *The Pharma Innovation Journal*, *10*(6), 629-634.

Taha, M. F., ElManawy, A. I., Alshallash, K. S., ElMasry, G., Alharbi, K., Zhou, L., Liang, N., & Qiu, Z. (2022). Using Machine Learning for Nutrient Content Detection of Aquaponics-Grown Plants Based on Spectral Data. *Sustainability*, *14*(19), 12318. https://doi.org/10.3390/su141912318

Tsapin, D. (2023). *Machine learning methods for the industrial robotic systems security*.

Waiba, K. M., Sharma, P., Sharma, A., Chadha, S., & Kaur, M. (2020). Soil-less vegetable cultivation: A review. *Journal of Pharmacognosy and Phytochemistry*, *9*(1), 631-636.

Wakchaure, M., Patle, B. K., & Mahindrakar, A. K. (2023). Application of AI techniques and robotics in agriculture: A review. *Artificial Intelligence in the Life Sciences*, *3*, 100057. https://doi.org/10.1016/j.ailsci.2023.100057

Wolanin, A., Camps-Valls, G., Gómez-Chova, L., Mateo-García, G., van der Tol, C., Zhang, Y., & Guanter, L. (2019). Estimating crop primary productivity with Sentinel-2 and Landsat 8 using machine learning methods trained with radiative transfer simulations. *Remote Sensing of Environment*, *225*, 441–457. https://doi.org/10.1016/j.rse.2019.03.002

Ye, T., Li, Y., Zhang, J., Hou, W., Zhou, W., Lu, J., Xing, Y., & Li, X. (2019). Nitrogen, phosphorus, and potassium fertilization affects the flowering time of rice (Oryza sativa L.). *Global Ecology and Conservation*, *20*, e00753. https://doi.org/10.1016/j.gecco.2019.e00753

You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1), 4559–4565. https://doi.org/10.1609/aaai.v31i1.11172

Zhang, R., & Sondh, A. (2024). *Prediction of Basil height in Hydroponic Systems through Machine Learning*.