Precision Farming Through AI: A Machine Learning-Based IoT Framework for Real-Time Agricultural Monitoring

1st Saud Alhajaj Aldossari
Department of Electrical Engineering
Prince Sattam bin Abdulaziz University
Wadi Addwasir, Saudi Arabia
Email: s.alhajaj@psau.edu.sa

2nd Salman Bader Hazza
Department of Electrical Engineering
Prince Sattam bin Abdulaziz University
Wadi Addwasir, Saudi Arabia
Email: salmanhuzaim@gmail.com

3rd Ibrahim Alhajouj
Department of Electrical Engineering
Prince Sattam bin Abdulaziz University
Wadi Addwasir, Saudi Arabia
Email: Ibrahim88hajouj@gmail.com

4th Abdullah Nader Aldossary
Department of Electrical Engineering
Prince Sattam bin Abdulaziz University
Wadi Addwasir, Saudi Arabia
Email: Eng.anm@hotmail.com

5th Moteb Abdullah Aldossary
Department of Electrical Engineering
Prince Sattam bin Abdulaziz University
Wadi Addwasir, Saudi Arabia
Email: Eng.moteb11@gmail.com

***Abstract*—With the upcoming farming revolution, this paper presents the development of an AI-powered agricultural monitoring system that integrates IoT devices with machine learning algorithms for real-time soil data analysis and nutrient prediction. The proposed system integrates a custom-built sensor-based device was designed to collect environmental data, including temperature, humidity, and essential soil nutrients (Nitrogen, Phosphorus, and Potassium).**

The methodology involved preprocessing the collected data to remove noise and inconsistencies, followed by training three types of models including Neural Networks, Random Forests, and CatBoost. These models were evaluated using key regression metrics such as MSE, MAE, and R^2 to determine their predictive accuracy. The results demonstrate that AI techniques can significantly enhance nutrient estimation and decision support in precision agriculture. By offering insights into soil health and nutrient availability, the solution can help reduce fertilizer use and improve crop yields. This study contributes to the growing field of smart farming by offering a low-cost, sensor-integrated solution for sustainable agricultural monitoring.

***Index Terms*—IoT, AI, catboost, random forest, neural network, farming technologies.**

I. Introduction

AI-powered agricultural monitoring systems are revolutionizing traditional farming practices by enhancing sustainability, productivity, and data-driven decision-making. These systems leverage Internet of Things (IoT) technology to enhance crop management and promote sustainable farming practices [1]. Given the increasing scarcity of natural resources and the impact of unpredictable climate conditions, intelligent monitoring systems are essential to ensuring global food security. Such systems support real-time monitoring of crop health, enable efficient resource usage, and improve overall productivity. IoT and machine learning technologies enable real-time data analysis to assist farmers in making informed decisions. This paper presents the development of a sensor-based IoT monitoring system that integrates AI algorithms to enhance sustainable farming. The IoT device collects real-time environmental data, including temperature, humidity, soil moisture, and concentrations of Nitrogen (N), Phosphorus (P), and Potassium (K).

Agriculture is changing as an outcome of machine learning (AI), which improves crop output, profitability, and production. It makes it possible to monitor supply chains, irrigation, weather, and pest management in real-time [2]. [3] selected drones for spraying and sensing, mapping, and harvesting to improve the water use efficiency and quality of crop yield. Moreover, they chose Smart Decision Support Systems (SDSS). While [4] explored IoT with the usage of big data and its analysis of massive data. [5] proposes a smart farming system in a limited, enclosed area wherein different sensors are strategically positioned to measure parameters such as moisture content, temperature, pressure, light intensity, and pH of the soil.

[6] has proposed a novel methodology for smart farming by linking a smart sensing system and a smart irrigator system through wireless communication technology. Jagannathan's system focuses on the measurement of physical parameters such as soil moisture content, nutrient content, and pH of the soil while our team focused on WSN other than the smart irrigator system. Futhermore, [7] developed a smart sensor-based monitoring system for an agricultural environment using a field programmable gate array (FPGA), which comprised of a wireless protocol, different types of sensors, a microcontroller, a serial protocol and the field programmable gate array with display element. Different types of sensors, such as tempera-

This project was funded by the Deanship of Scientific Research at Prince Sattam bin Abdulaziz University award number 2023/SDG/01.

ture, soil moisture, and relative humidity, sense the data in an agricultural environment and provide it to a microcontroller interfaced with the wireless Bluetooth module. While, [8] worked on an agricultural environment monitoring system using wireless sensor networks and IoT with low amount of data. However this project produced a device that has all-in-one generated data that will be used with AI algorithms for several purposes.

The primary contributions of this paper are as follows:

- Introducing the state of modern agriculture technologies.
- Presenting IoTs devices toward the agriculture world.
- Generating new agriculture data.
- Showing the integration of AI and agriculture data.
- Applying new AI algorithms with optimizations.

The remainder of this paper is organized as follows. Section 2 provides an overview of artificial intelligence in agriculture, with a focus on supervised learning and regression techniques relevant to our study. Section 3 describes how the agricultural dataset was generated using custom-built IoT sensors, followed by a detailed discussion of data preprocessing and cleansing methods. In Section 4, we present the experimental setup, model training procedures, and results obtained from Neural Networks, Random Forests, and CatBoost algorithms. Section 5 discusses the limitations of the current system, while Section 6 highlights its practical applications in real-world farming environments, including an analysis of cost and expected return on investment. Finally, Section 7 outlines future research directions, and Section 8 concludes the paper.

II. Artificial Intelligence in Agriculture

Artificial Intelligence (AI) is a branch of computer science focused on developing systems capable of performing tasks traditionally requiring human intelligence [9]. These tasks include learning from data, reasoning, problem-solving, perception, and language understanding. AI systems mimic human cognitive processes and are widely used to automate tasks, support decision-making, and generate predictions across multiple domains. AI research aims to develop machines with general intelligence that can understand and adapt to complex environments and tasks, potentially matching or exceeding human capabilities. AI has broad applications across fields such as engineering, healthcare, finance, transportation, and entertainment. As AI technologies continue to advance, they have the potential to revolutionize how we live, work, and interact with the world around us [10].

The convergence of AI and agriculture is expected to bring transformative changes to modern farming. Innovative solutions are essential to addressing global challenges such as resource scarcity, population growth, and climate change, all of which threaten food security. AI enhances traditional farming practices by improving efficiency and insight throughout the agricultural process [11] [12]. The following subsection focuses specifically on supervised learning techniques, while other AI approaches such as probabilistic modeling are discussed in detail in external references, including Taeho's book [13].

A. Supervised Learning

Supervised learning is a fundamental concept in machine learning, defined by the use of labeled datasets to train predictive models. Each training example in supervised learning contains an input feature vector paired with a corresponding target output.The goal is for the model to learn a mapping function that can generalize to accurately predict outputs for unseen inputs. As discussed in [14], supervised learning includes a range of algorithms such as linear regression, logistic regression, decision trees, support vector machines, and neural networks. Additionally, author of [15] applied supervised machine learning techniques in the context of agricultural applications. Classification is one category of supervised learning, where data is categorized based on predefined labels. Regression [16], the second type of supervised learning, is discussed in detail in the following subsection .

B. Regression

Regression focuses on modeling the relationship between input features and a continuous output variable. Regression is used for predicting continuous-valued output based on input features. The objective of regression analysis is to establish a quantitative relationship between independent variables (features) and a continuous dependent variable (target) [17]. The goal is to predict the target variable using the input features provided.

In machine learning, different regression techniques apply varying mathematical formulations depending on the problem type. Common supervised regression methods include linear regression, polynomial regression, and ensemble models each offering unique strengths and limitations. Linear regression, one of the most fundamental techniques, assumes a linear relationship between features and the target. It fits a line that minimizes the error between predicted and actual values using a least squares approach.

$$y_i = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

Several metrics are used to measure the error and evaluate the performance of a model. Each metric highlights a different aspect of the model's predictive performance. Following are the most common matrices:

The Root Mean Square Error (RMSE) is used to evaluate the average magnitude of prediction errors, and is calculated using the following equation: 2 is chosen as performance measure:

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

where $RMSE_v$ is Root-mean-square error (RMSE) of link residual time. y and $\hat{y}$ are the actual and predicted values respectively.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

where:

- y_i: Actual value of the i-th data point.
- $\hat{y}_i$: Predicted value of the i-th data point.
- n: Total number of data points.

The R^2 score, or coefficient of determination, measures how well the predictions approximate the actual data distribution:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

Moreover, Residual Sum of Squares (RSS) is measuring method in regression which is a metric used to measure the discrepancy between the observed data and the values predicted by a regression model. It quantifies the total squared difference between observed values (y_i) and the predicted values $(\hat{y}_i$ from a model.

The goal of RSS minimization is to find the parameters of a regression model (e.g., coefficients and intercepts) that minimize the RSS. This ensures that the model best fits the data by reducing the error in predictions. The Residual Sum of Squares (RSS) is a key regression metric that measures the total squared difference between observed and predicted values:

$$\text{RSS} = \min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2 \tag{5}$$

where:

- y_i: Observed value for the i-th instance
- $\hat{y}_i$: Predicted value for the i-th instance
- β_0: Intercept term
- β_j: Coefficient of the j-th independent variable (scalar), $j = 1, \ldots, p$
- x_{ij}: Value of the j-th feature for the i-th observation
- n: Number of observations
- p: Number of features

III. Data

The integration of Artificial Intelligence (AI) into agriculture has significantly improved productivity, resource optimization, and data-driven decision-making.AI-driven technologies such as precision farming, predictive analytics, and real-time crop monitoring allow farmers to optimize resource usage including water, fertilizers, and pesticides, while minimizing waste and environmental impact. Machine learning models have been applied to predict nutrient levels (NPK) and irrigation needs, thereby improving yield forecasting and operational efficiency. Figure 1 illustrates the process of integrated block diagram of the AI-based agricultural monitoring system. From the data that was generate, Figure 2 shows the relation between phosphorus levels and two environmental factors: temperature and humidity. Where phosphorus levels peak around 20°C and then gradually decline as the temperature increases and phosphorus levels remain stable at low humidity values. This suggests that phosphorus concentration is higher at moderate temperatures and decreases significantly at higher temperatures. After about 90% humidity, phosphorus levels start increasing, reaching a peak between 120–130%, then slightly decline.

In Figure 3 shows the relationship between nitrogen levels and two environmental variables: temperature and humidity. Nitrogen levels rise sharply with humidity up to 100% due to enhanced ammonification and reduced volatilization. The sudden drop at very high humidity levels likely reflects nutrient leaching and reduced aeration, which hinder nitrogen retention—consistent with known soil-water interactions.

A sharp increase starts after 60% humidity, reaching a peak around 100–110%. Then, nitrogen values decline gradually with further increases in humidity. This suggests that nitrogen is positively correlated with humidity up to a point, after which excessive humidity may lead to nutrient leaching or reduced retention.

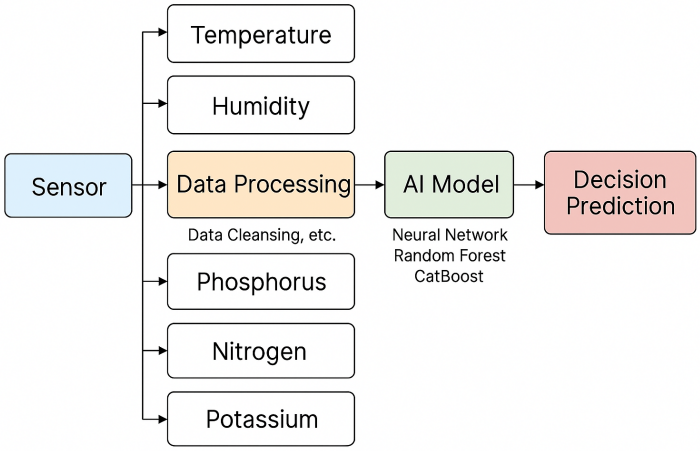

Fig. 1. Integrated Block Diagram of the AI Agricultural Monitoring System

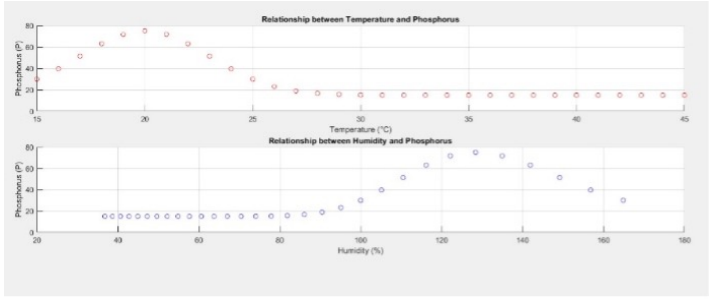

Fig. 2. Relationship Between Phosphorus, Temperature and Humidity.

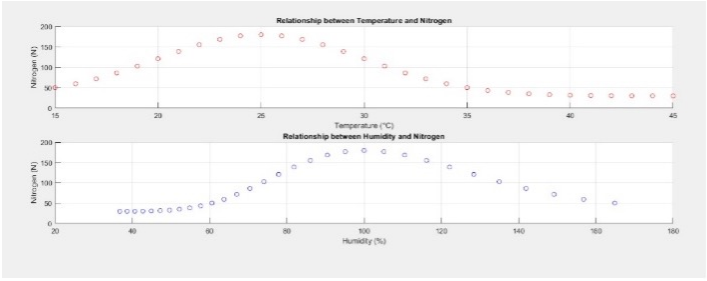

Fig. 3. Relationship Between Nitrogen, Temperature and Humidity

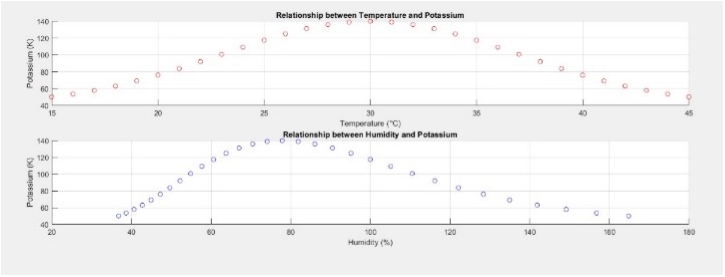

Fig. 4. Relationship Between Potassium, Temperature and Humidity

Moreover, Figure 4 illustrates relationship between potassium, temperature and humidity. Potassium increases with temperature until 30°C as root uptake and soil exchange activity intensify. The decline beyond 30°C is associated with reduced nutrient fixation and soil dryness. The humidity-based rise toward 80–90% further supports the role of moisture in enabling soil–K exchange, with oversaturation causing reduced mobility.

A. Data Cleansing

Data cleaning is a crucial preprocessing step in any AI-driven agricultural project, particularly when working with sensor-generated readings such as temperature, humidity, and nutrient levels (Nitrogen, Phosphorus, and Potassium). It ensures the collected data is accurate, consistent, and free from noise, errors, or outliers that could impair model performance. In this section, we'll cover the importance of data cleaning, common issues in raw data, and key techniques for cleaning data in our project.

In agricultural projects, the accuracy of sensor data is crucial for making informed decisions. For instance, incorrect NPK levels could lead to poor recommendations for soil treatment, and inaccurate temperature and humidity readings might result in incorrect assumptions about climate conditions. Data cleaning ensures that the dataset is reliable and suitable for analysis. Without proper data cleaning, even advanced machine learning models can generate misleading outcomes due to incomplete or noisy input.

1) Common Data Issues:

- Missing Values: Sensor failures, environmental interference, or transmission errors can result in missing values, creating gaps that compromise analysis integrity.
- Outliers: Outliers, which deviate significantly from the normal data distribution, may arise due to sensor malfunctions or extreme environmental conditions. Identifying whether these are legitimate or erroneous is essential for robust analysis.
- Inconsistent Data: Inconsistent formats may occur when data is collected from multiple sensors or across different time intervals. For instance, temperature data might mix Celsius and Fahrenheit units, necessitating standardization.
- Duplicate Entries: Duplicate entries are common in continuously streaming systems, where identical values may be recorded multiple times due to network lag or sensor delay.

Data cleaning is an integral part of any data usage. By ensuring that your data is free from errors, outliers, and inconsistencies, you increase the reliability of your analysis and improve the quality of your insights. Whether you're analyzing or optimizing, data cleansing is the foundation upon which accurate models and predictions are built.

IV. Results

A. Neural Network

An Artificial Neural Network (ANN) is a machine learning model inspired by the structure and function of the human brain. ANNs are core components in both AI and ML, capable of recognizing patterns, processing input data, and making predictions or decisions. They are particularly effective for tasks such as image classification, language modeling, and predictive analytics.

Modern ANN architectures, including deep learning models and transformers, have significantly expanded the applicability of neural networks to complex, real-world problems.

The mathematical foundation of an ANN is described by its forward propagation mechanism.. A common form for a single-layer neural network is:

$$\hat{y} = f\left(W \cdot x + b\right) \tag{6}$$

For a multi-layer ANN with L layers, the forward pass is defined as follows:

$$a^{[l]} = f^{[l]}\left(W^{[l]} \cdot a^{[l-1]} + b^{[l]}\right), \quad \text{for } l = 1, 2, \ldots, L \tag{7}$$

where $W^{[l]}$ = weight matrix at layer and $a^{[l-1]}$ activations (or inputs) from the previous layer. While $b^{[l]}$ is bias vector at layer and $f^{[l]}$ is activation function at layer.

B. Random Forests

Random Forest is an ensemble machine learning algorithm widely used for both classification and regression tasks. It constructs multiple decision trees during training and aggregates their outputs to improve accuracy and reduce overfitting. Random forests are particularly valued for their ability to handle complex datasets with a mix of categorical and numerical features, as well as for their robustness to overfitting. Each

decision tree t outputs a prediction y_t, calculated as the mean of all target values in the leaf node where input X is classified.

$$y_t = \frac{1}{n_t} \sum_{i \in \text{Leaf}_t(X)} y_i \tag{8}$$

Where:

y_t : Prediction from the t-th tree.
n_t : Number of samples in the leaf node of the t-th tree corresponding to input X.
$\text{Leaf}_t(X)$: Set of samples in the leaf node of the t-th tree where the input X falls.
y_i : Actual target value of the i-th sample in the leaf node.

The final regression output is the average of predictior from all T decision trees. While the final prediction in Rando Forest is made by aggregating predictions from all the trees

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} y_t \tag{9}$$

Where:

$\hat{y}$: Final predicted value of the Random Forest model.
T : Total number of decision trees in the ensemble.
y_t : Prediction from the t-th tree.

Random forests are widely used in various industries because of their reliability, simplicity, and versatility, making them an essential tool in the machine learning toolbox.[20]

C. CatBoost

CatBoost (Categorical Boosting) is a gradient boosting algorithm specifically designed to handle categorical features, developed by Yandex. It natively supports categorical variables, eliminating the need for preprocessing methods such as one-hot or label encoding. CatBoost is highly efficient, works well on structured/tabular data, and achieves state-of-the-art performance on classification and regression tasks. CatBoost is designed to handle both classification and regression problems by minimizing appropriate loss functions and leveraging its unique features like categorical data handling and gradient boosting.

CatBoost is part of the ensemble learning family, like other gradient boosting frameworks like XGBoost and LightGBM. However, it stands out because of its simplicity, ease of use, and automatic handling of categorical variables.

CatBoost minimizes a loss function that quantifies the difference between the predicted and actual values. The most common loss function used for regression is the Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{10}$$

Where:

MSE : MSE function minimized by CatBoost.
N : Total number of samples in the dataset.
y_i : Actual target value of the i-thsample.
$\hat{y}_i$: Predicted value of the i-thsample by the model.

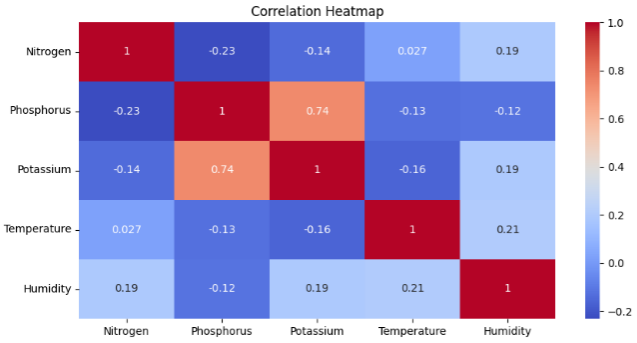

Fig. 5. Correlation Between the Agriculture Data

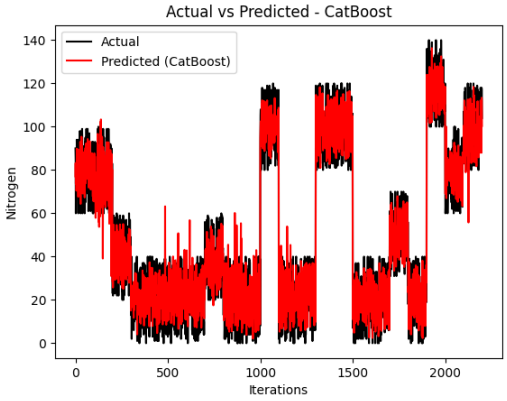

Fig. 6. CatBoost Modle

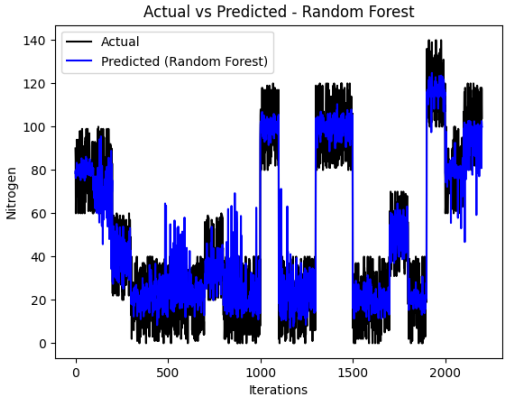

Fig. 7. Random Forests Model

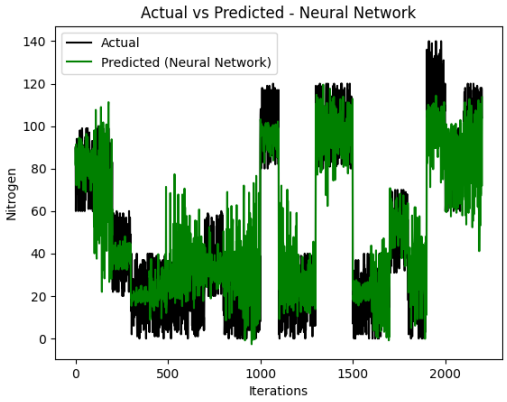

Fig. 8. Neural Network Model

Metric	NN Score	Random Forests Score	CatBoost Score
MSE	328	124	63
MAE	15	9	6
$\hat{R}^2$	0.76	0.91	0.95

TABLE I
COMPARISON OF PERFORMANCE METRICS FOR NEURAL NETWORK (NN), RANDOM FORESTS, AND CATBOOST ALGORITHMS.

V. FUTURE WORKS

Future iterations of the system will investigate the use of advanced communication technologies such as LPWAN, LoRa, and 5G to support high-speed, low-latency data exchange, particularly in remote agricultural environments. Future work will also involve expanding the dataset with samples from diverse climates, soil types, and crop varieties to improve model generalization and robustness. The incorporation of edge computing techniques will enable on-device processing, reducing latency and minimizing the need for continuous cloud connectivity. Designing lightweight AI models optimized for deployment on embedded systems will be prioritized to ensure efficiency and portability. Long-term deployments across varying seasonal and climatic conditions will be conducted to evaluate system resilience, reliability, and field usability. Energy sustainability will also be explored through the integration of solar panels and low-power design strategies to support autonomous operation in remote areas.

VI. CONCLUSIONS

The integration of Artificial Intelligence (AI) into agricultural monitoring systems marks a significant advancement in modern farming practices. By leveraging AI technologies such as machine learning, these systems enable farmers to monitor crop health, detect anomalies early, and optimize resource usage. AI facilitates real-time, data-driven decision-making, allowing farmers to reduce risk, respond proactively to field conditions, and enhance productivity. EThis proactive approach supports early detection of issues such as pest infestations, enabling timely interventions and minimizing crop damage. This paper presents a step toward revolutionizing agriculture through the application of AI and smart sensing technologies.By integrating machine learning and IoT technologies, the proposed system enhances monitoring of crop health, irrigation efficiency, and nutrient prediction. The system involved the generation and cleaning of real-world sensor data, followed by the implementation and optimization of various AI models for prediction tasks.

ACKNOWLEDGMENT

The authors appreciate the support of this project, which was funded by Prince Sattam bin Abdulaziz University.

REFERENCES

[1] King, A. Technology: The future of agriculture. *Nature*. **544**, S21-S23 (2017)
[2] Friha, O., Ferrag, M., Shu, L., Maglaras, L. & Wang, X. Internet of things for the future of smart agriculture: A comprehensive survey of emerging technologies. *IEEE/CAA Journal Of Automatica Sinica*. **8**, 718-752 (2021)
[3] Rawat, P., Gupta, P. & Soni, P. Smart IoT System for Agricultural Production Improvement and Machine Learning-Based Prediction. *Soft Computing Principles And Integration For Real-Time Service-Oriented Computing*. pp. 174-186 (2024)
[4] Li, C. & Niu, B. Design of smart agriculture based on big data and Internet of things. *International Journal Of Distributed Sensor Networks*. **16**, 1550147720917065 (2020)
[5] Gorli, R. & Yamini, G. Future of smart farming with Internet of things. *Journal Of Information Technology And Its Applications*. **2** (2017)
[6] Jagannathan, S., Priyatharshini, R. & Others Smart farming system using sensors for agricultural task automation. *2015 IEEE Technological Innovation In ICT For Agriculture And Rural Development (TIAR)*. pp. 49-53 (2015)
[7] Mathurkar, S., Patel, N., Lanjewar, R. & Somkuwar, R. Smart sensors based monitoring system for agriculture using field programmable gate array. *2014 International Conference On Circuits, Power And Computing Technologies [ICCPCT-2014]*. pp. 339-344 (2014)
[8] Marques, G. & Pitarma, R. Agricultural environment monitoring system using wireless sensor networks and IoT. *2018 13th Iberian Conference On Information Systems And Technologies (CISTI)*. pp. 1-6 (2018)
[9] Ertel, W. Introduction to artificial intelligence. (Springer Nature,2024)
[10] Russell, S. & Norvig, P. Artificial intelligence: a modern approach. (Pearson,2016)
[11] Shaikh, T., Rasool, T. & Lone, F. Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. *Computers And Electronics In Agriculture*. **198** pp. 107119 (2022)
[12] Liakos, K., Busato, P., Moshou, D., Pearson, S. & Bochtis, D. Machine learning in agriculture: A review. *Sensors*. **18**, 2674 (2018)
[13] Jo, T. Machine learning foundations. *Machine Learning Foundations. Springer Nature Switzerland AG. Https://doi. Org/10.1007/978-3-030-65900-4*. (2021)
[14] Choudhary, R. & Gianey, H. Comprehensive review on supervised machine learning algorithms. *2017 International Conference On Machine Learning And Data Science (MLDS)*. pp. 37-43 (2017)
[15] Kumar, Y., Spandana, V., Vaishnavi, V., Neha, K. & Devi, V. Supervised machine learning approach for crop yield prediction in agriculture sector. *2020 5th International Conference On Communication And Electronics Systems (ICCES)*. pp. 736-741 (2020)
[16] Nasteski, V. An overview of the supervised machine learning methods. *Horizons. B*. **4**, 56 (2017)
[17] Kumar, S. & Bhatnagar, V. A review of regression models in machine learning. *JOURNAL OF INTELLIGENT SYSTEMS AND COMPUTING*. **3**, 40-47 (2022)
[18] Saiz-Rubio, V. & Rovira-Más, F. From smart farming towards agriculture 5.0: A review on crop data management. *Agronomy*. **10**, 207 (2020)