Smart Agriculture: Leveraging IoT and Machine Learning for Sustainable Farming

Vidhi Gupta, Department of Computer Applications, Invertis University, Bareilly, India, gvidhi957@gmail.com
Ridhima Singh, Department of Computer Applications, Invertis University, Bareilly, India, ridhima415singh@gmail.com
Divas Mishra, Department of Computer Applications, Invertis University, Bareilly, India, mishradivas18@gmail.com
Pratha Sexena, Department of Computer Applications, Invertis University, Bareilly, India, prathasaxena222@gmail.com
Navnika Kapoor, Department of Computer Applications, Invertis University, Bareilly, India, navnikakapoor344@gmail.com

Abstract-

The increasing global demand for food, along with the challenges posed by climate change and limited natural resources, calls for a shift from conventional farming to more intelligent, data-centric methods. This study investigates the use of Internet of Things (IoT) devices, cloud computing, and Machine Learning (ML) algorithms to support sustainable agricultural practices. A dataset containing 10,001 entries—including variables such as environmental conditions, soil nutrients, and crop data—was analysed to forecast crop yield. Multiple regression models were tested, with the Random Forest Regressor delivering the highest accuracy at 98.48%, significantly outperforming baseline models like Linear Regression, which scored 76.42%. The integration of cloud services facilitates scalable, real-time data handling and allows efficient processing of sensor data alongside predictive modelling. This research highlights the effectiveness of ensemble learning methods and connected infrastructure in delivering actionable insights for precision agriculture. In order to increase productivity and ensure sustainable resource use, the suggested framework encourages more intelligent choices in areas such as crop planning, soil management, and yield enhancement.

Keywords- Sustainable farming, IoT in agriculture, machine learning, regression models, random forests, smart farming, and crop prediction.

1. Introduction

Ensure food security, economic advancement, and the welfare of rural populations, the agricultural sector remains a vital component of global stability. Climate change, depleting natural resources, and rising food demand due to population increase, however, are putting growing strain on it. To fulfil the demands of a predicted 9.7 billion people by 2050, the Food and Agriculture Organization (FAO) estimates that world food production must increase by almost 70% [1].

Traditional farming methods, which mostly rely on human labor and traditional expertise, are not meeting these changing difficulties, a major technical change is taking place in agriculture. New developments like cloud computing, machine learning, and the internet of things (IoT) are increasingly guiding the sector toward "Smart Agriculture," a datacentric approach that makes decisions more sustainable and well-informed [2].

Soil moisture sensors, weather tracking devices, and GPS enabled equipment are examples of contemporary products that make it easier to monitor crops and the environment in real time. When this constant stream of data is analysed by machine learning algorithms, it allows for predictive forecasting, pattern detection, and farming process automation [3]. Particularly for critical agricultural decisions like production prediction, insect identification, irrigation optimization, and soil health analysis machine learning (ML) provides useful insights [4].

The Random Forest Regressor is one of the most successful machine learning approaches. It works well as an ensemble model for handling intricate, nonlinear connections and is resistant to overfitting, which makes it a good option for agricultural prediction applications [5]. However, the availability of high-quality data is crucial for these models to perform well, highlighting the necessity of dependable sensor networks and effective data pipelines. Since cloud computing offers scalable platforms for real-time data processing and storage, it also plays a crucial role. Small-scale farmers especially benefit from these cloud-based infrastructures since they provide centralized access to computing resources without requiring expensive onsite systems [6].

This study offers a comprehensive framework for smart agriculture that combines cloud-based storage, ML based analytics, and IoT-driven data collecting to promote sustainable farming. With a 98.48% accuracy rate, the Random Forest Regressor was the best performer when various regression models were applied to a dataset with 10,001 samples, which included data on crop types, weather, and soil nutrient

This demonstrates how it can enhance yield predictions and encourage resource management based on data. In conclusion, by utilizing real-time data and cognitive analytics, this study suggests a solid, scalable strategy for contemporary farming. In addition to increasing production, the combination of IoT, cloud computing, and machine learning promotes long-term environmental sustainability in the agricultural sector

2. Literature Review

By enabling real-time data collecting, smarter automation, and improved resource management, the integration of Internet of Things (IoT) technology in agriculture has greatly advanced precision farming. Using wireless sensor networks (WSNs) and cloud platforms, Zhang et al. [7] presented a smart farming system that tracks important variables including temperature and soil moisture.

Using real-time environmental data, Rajeswari et al. [8] deployed an Internet of Things (IoT)based irrigation system that maximized water use. A review of IoT applications in agriculture was conducted by Kamilaris et al. [9], who also looked at frequent use cases, architectural patterns, and important difficulties. Khodve et al. [10] used inexpensive microcontrollers like Arduino to automate farm tasks, and Zhang et al. [11] used cloud computing and IoT to remotely monitor crops and irrigation. For agricultural prediction tasks, Random Forest (RF) has become a potent algorithm in the field of machine learning (ML).

Anticipate agricultural yield, Jeong et al. [12] used RF, showcasing its ability to identify intricate data pattern. In their summary of the application of machine learning techniques in precision agriculture, Liakos et al. [13] pointed out that RF continuously performed well on a variety of tasks. To offer practical insights for precision farming, Chlingaryan et al. [14] suggested a sensor-data fusion framework improved by machine learning. The potential of image-based models in agriculture, demonstrated by Gandhi et al. [15], who used deep learning for plant disease recognition. According to Kamilaris and Prenafeta Boldú [16], RF's accuracy and resilience make it one of the best machine learning models for agricultural analytics.

Adaptive smart agricultural systems can be developed with the help of IoT and ML integration. A greenhouse monitoring system developed by Jabbar et al. [17] uses sensor data and machine learning to control variables like temperature and humidity. Automate watering schedules, Zhou et al. [18] developed an intelligent irrigation model driven by regression techniques. A cloud based framework for visualizing and making decisions from farm data was introduced by Fukatsu and Hirafuji [19]. Standardization, sharing, and interoperability are crucial issues in agricultural big data systems that must be, addressed by Wolfert et al. [20] to scale smart farming solutions.

In smart agriculture, sustainability is still a key component. A cloud-based Internet of Things system for tracking soil pH and nutrients was created by Jayaraman et al. [21] with the goal of promoting long-term soil health. An irrigation model based on decision trees that effectively controlled water use was presented by Panday et al. [22].

Tzounis [24] and Boursianis et al. [23] highlighted the role that IoT-ML integration plays in environmentally conscious agriculture. Ramesh and Vani [25] also highlighted how smart farming supports sustainability goals, including climate adaptation and reduced chemical usage.

Emerging challenges include data security, system scalability, and real-time processing. Kumar et al. [26] explored blockchain as a solution for secure data sharing in agriculture. Manogaran et al. [27] recommended fog computing to reduce latency in real-time IoT systems. These developments underline the transformative potential of combining IoT and ML, particularly Random Forest, in creating scalable, efficient, and sustainable agricultural systems.

3. System Architecture

The architecture of the proposed IoT-enabled smart farming system combines environmental sensing, wireless data transmission, cloud-based data management, and machine learning-based analytics. It is built to enable real-time tracking and intelligent regulation of key agricultural factors such as soil moisture, temperature, humidity, and light

intensity—elements essential for promoting sustainability and precision in modern farming operations. IOT Sensor system for model evaluation is shown in Figure 1.

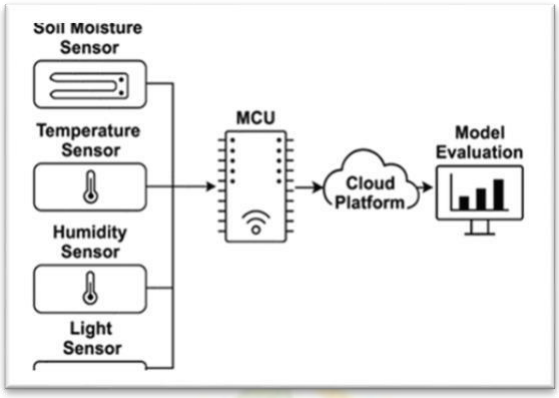


Figure 1. IoT sensor system for model evaluation.

3.1. Sensor Layer

The sensor layer forms the foundation of the system and is responsible for data acquisition from the farm environment. The following sensors are deployed:

Soil Moisture Sensor: A capacitive soil moisture sensor shown in Figure 2 is utilized to measure the volumetric water content in the soil. This sensor is chosen for its durability and resistance to corrosion, making it suitable for continuous field deployment.

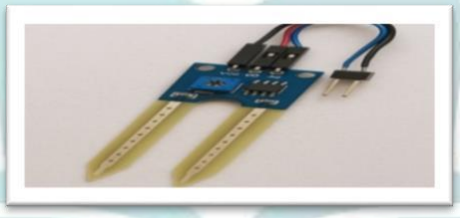


Figure 2. Soil moisture sensor module

DHT11/DHT22 Sensor: These sensors shown in Figure 3 measures ambient temperature and humidity. DHT22 is preferred for higher accuracy and wider operating ranges, though DHT11 is also a low-cost alternative for small-scale applications.

Figure 3. Temperature and humidity sensor module

Light Sensor (LDR/BH1750): A light-dependent resistor (LDR) or BH1750 is used to monitor light intensity, which influences photosynthesis and crop growth. The BH1750 shown in Figure 4 is selected when higher precision and digital output are needed.

Figure 4. BH1750 light intensity sensor module

These sensors are interfaced with a microcontroller for data acquisition at predefined intervals.

3.2. Microcontroller Unit

The system employs an ESP32 microcontroller, which serves as the central processing unit. ESP32 is selected due to its built-in Wi-Fi and Bluetooth capabilities, low power consumption, and sufficient GPIOs for multiple sensor inputs. It also supports edge-level processing and real-time data transmission to the cloud.

3.3. Communication and Networking Layer

Sensor data is transmitted via Wi-Fi to a cloud-based IoT platform using the microcontroller's wireless capabilities. This architecture eliminates the need for complex gateway devices and supports direct communication between the device and the cloud server.

3.4. Cloud Platform

The system uses Thing Speak, a free, open-source IoT analytics platform powered by MATLAB. ThingSpeak facilitates:
Real-time data storage
Visualization through dashboards
Data export in CSV format for offline analytics
This platform is ideal for prototyping and small-scale deployments due to its ease of integration with ESP32 and support for HTTP protocols.

3.5. Data Analysis & Machine Learning

After collecting the data on Thing Speak, the dataset was exported for offline preprocessing and analysis. In the preprocessing phase, the dataset shown in Figure 5 underwent rigorous cleaning to ensure data quality and model reliability. Missing values were identified and imputed using appropriate statistical methods based on the distribution of each feature.

	Unnamed: 0	Crop_type	Crop	N	P	K	pH	rainfall	temperature	Area_in_hectares	Production_in_tons	target
0	0	kharif	cotton	120	40	20	5.46	654.34	29.266667	7300	9400	1.287671
1	1	kharif	horsegram	20	60	20	6.18	654.34	29.266667	3300	1000	0.303030
2	2	kharif	jowar	80	40	40	5.42	654.34	29.266667	10100	10200	1.009901
3	3	kharif	maize	80	40	20	5.62	654.34	29.266667	2800	4900	1.750000
4	4	kharif	moong	20	40	20	5.68	654.34	29.266667	1300	500	0.384615
...	...	...	...	...	...	...	...	...	...	...	...	...
9996	9996	summer	maize	80	40	20	5.40	34.81	34.666667	152	154	1.013158
9997	9997	summer	moong	20	40	20	5.60	34.81	34.666667	488	211	0.432377
9998	9998	whole year	onion	120	60	65	5.94	689.88	29.037273	752	9080	12.074468
9999	9999	whole year	potato	180	60	90	5.02	689.88	29.037273	7595	167455	22.048058
10000	10000	kharif	maize	80	40	20	5.48	579.75	34.010000	11247	3385	0.300969

Figure 5. Dataset with environmental and yield features.

Univariate analysis depicted in Figure 6,7,8 was performed on key numerical variables such as nitrogen (N), phosphorus (P), potassium (K), pH, temperature, rainfall, and so on, enabling the visualization of their distributions and understanding of underlying patterns.

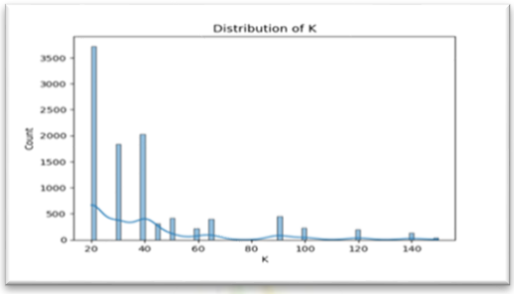


Figure 6. Distribution plot of potassium (K) levels

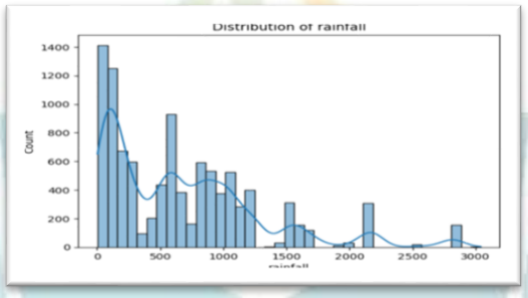


Figure 7. Distribution plot of rainfall levels.

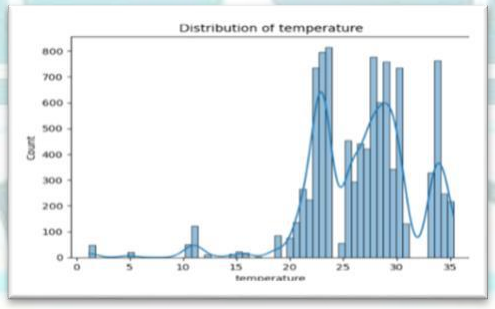


Figure 8. Distribution plot of temperature levels.

Box plots and distribution plots in Figure 9,10,11 were used to detect and interpret outliers, which were subsequently treated using interquartile range (IQR)-based filtering. The data was then re-visualized to confirm successful outlier removal.

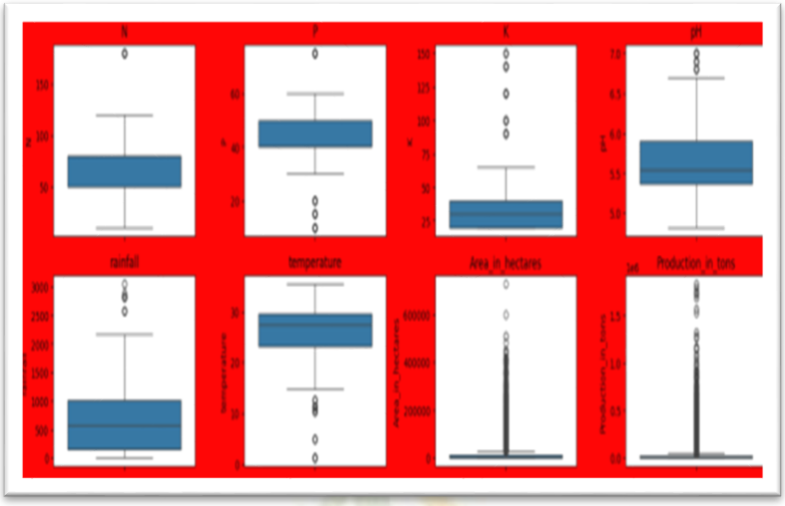


Figure 9. Boxplots of feature distributions with outliers.

```
or col in num_col:
    Q1 = data[col].quantile(0.25)
    Q3 = data[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    data[col] = np.where(data[col] < lower_bound, lower_bound,
                         np.where(data[col] > upper_bound, upper_bound, data[col]))
```

Figure 10. IQR-based outlier capping applied to numerical features.

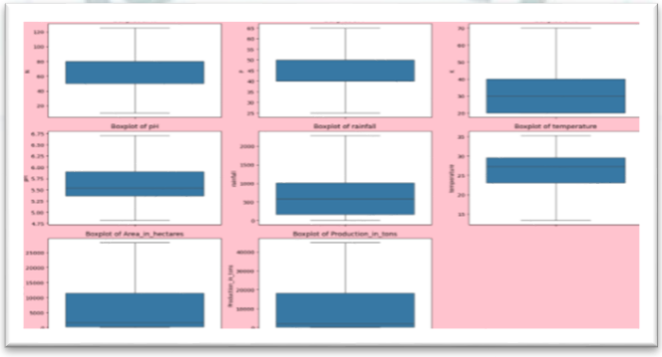


Figure 11. Boxplots after outlier handling.

Understand multivariate relationships, a Pearson correlation heatmap in Figure 12 was generated, highlighting both strong and weak associations among the variables.

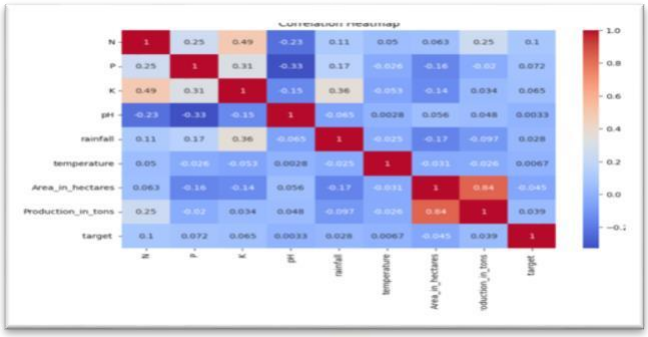


Figure 12. Correlation heatmap of features and target.

Additionally, statistical metrics like skewness and kurtosis shown in Figure 13 were computed to assess the symmetry and peak of each distribution, aiding in decisions around feature scaling and transformation.

```
for i in data.select_dtypes(include='number'):
    kurtosis=data[i].kurt()
    print("The kurtosis of",i,"is",kurtosis)
```

```
The kurtosis of N is -0.7301032553683915
The kurtosis of P is -0.6683414127428859
The kurtosis of K is -0.46160178977684563
The kurtosis of pH is -0.23630498985886517
The kurtosis of rainfall is 0.531391366126706
The kurtosis of temperature is 0.28378409803257787
The kurtosis of Area_in_hectares is -0.2792433054587127
The kurtosis of Production_in_tons is -0.29750692222787656
The kurtosis of target is 1560.9117230489421
```

Figure 13. Kurtosis values indicating data distribution sharpness.

Since the target variable in the dataset was categorical in nature, it required numerical transformation before model training. This transformation was achieved using Label Encoding, wherein each distinct category was assigned a unique integer value. This encoding was applied specifically to the target column, enabling supervised learning algorithms to interpret categorical outputs in a numerically meaningful form. Label Encoding is particularly suitable for target variables as it maintains compact integer values without inflating feature space. Following data cleaning and feature scaling, the dataset was divided into training and testing subsets using an 80:20 split ratio, whereby 80% of the data was set aside for training the machine learning model and the remaining 20% was set aside for assessing its performance on unseen data. This ensured that the target variable could be used in regression models effectively without exceeding data type constraints, and numerical features were standardized using the Standard Scaler to ensure uniform scaling and model convergence. This resulted in 7000 samples in the training set (X train: 7000*10, Y train: 7000) and 3001 samples in the testing set (X test: 3001*10, Y test: 3001). Splitting the data in this manner ensures that the model learns from a substantial portion of the dataset while preserving a sufficient portion for validation. This

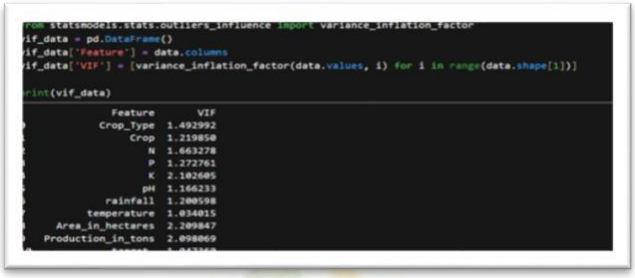
```
rom sklearn.model_selection import train_test_split

_train, X_test, Y_train, Y_test = train_test_split(x, y, test_size=0.30, random_state=42)
rint(f"X_train shape: {X_train.shape}")
rint(f"X_test shape: {X_test.shape}")
rint(f"Y_train shape: {Y_train.shape}")
rint(f"Y_test shape: {Y_test.shape}")
```

```
_train shape: (7000, 10)
_test shape: (3001, 10)
_train shape: (7000,)
_test shape: (3001,)
```

approach helps in assessing the model's generalization capability—i.e., its ability to make asccurate predictions on new, unseen data—thus reducing the risk of overfitting (in Figure 14).

Figure 14. Dataset split into 70% training and 30% testing sets.

Variance Inflation Factor (VIF) analysis depicted in Figure 15 was also performed to check for multicollinearity, helping to refine the feature set and prevent redundancy.



ss

Figure 15. VIF analysis confirms low multicollinearity.

Forecast crop yield based on agro-climatic and soil parameters, several regression models were tested, including Linear Regression, KNN, Decision Tree, and Random Forest. Linear Regression, implemented via sklearn. linear model, served as the baseline, assuming a direct linear relationship between input features and yield. After training on preprocessed data, the model produced an MSE of 69,468,470.20, RMSE of 8334.77, and MAE of 5285.77. The R^2 score was 0.7642, suggesting it explained about 76.42% of the variance in yield. Although the model captured broad trends, its performance was limited by an inability to represent complex, non-linear interactions, making it less effective than more adaptive methods like Decision Trees or Random Forests.

The K-Nearest Neighbours (KNN) regression model, implemented with K Neighbors Regressor from sklearn. neighbours, used k=3k = 3k=3 to predict crop yield. This non-parametric method estimates output by averaging the target values of the three nearest data points based on Euclidean distance. Trained on scaled data, the model achieved an MSE of 31,650,724.99, an RMSE of 5625.90, and an MAE of 2422.78. Its R^2 score was 0.8925, meaning it explained roughly 89.25% of the variation in crop yield. While KNN effectively modeled non-linear patterns, its performance lagged ensemble models like Random Forest, which offered better accuracy and generalization.

The Decision Tree Regressor, implemented via DecisionTreeRegressor from sklearn. tree, was used to predict crop yields by recursively splitting the feature space based on input variables. The model builds a hierarchical tree where internal nodes represent feature-based decisions and leaves indicate output predictions. Trained on scaled data, it yielded an MSE of 31,650,724.99, an RMSE of 5625.90, and an MAE of 680.92. Its R^2 score reached 0.9754, explaining over 97% of the yield variance. These metrics indicate strong performance and the model's ability to capture complex, non-linear interactions among agro-climatic and soil factors. However, due to the model's tendency to overfit, especially with deep trees, regularization techniques like pruning may be needed to ensure better generalization on unseen data.

The Random Forest Regressor implemented using Random Forest Regressor from sklearn. ensemble, leverages an ensemble of decision trees trained on random data subsets with feature sampling to enhance prediction accuracy With an RMSE of 5625, MAE of 517.03, and an R2 score of 0.984 following training on the processed dataset, the model demonstrated impressive performance, explaining almost 98.49% of the yield variation. This performance demonstrates how well Random Forest models intricate, nonlinear interactions between soil and agroclimatic factors, which makes it particularly useful and generalizable for precise crop yield estimation.

4. Acknowledgement

Although agriculture is essential to India's economic development, conventional farming methods frequently encounter difficulties like low crop yields and unfavourable weather. By combining IoT and machine learning, farmers may get useful information that will help them make better decisions, cut expenses, and increase output. With the potential to benefit millions of people nationwide, these data driven suggestions provide a reliable and scalable solution. Ensemble methods like Random Forest were the most successful among the models studied in managing intricate agricultural data, identifying non-linear patterns, and reducing overfitting. These models improve the

accuracy of yield prediction and facilitate data-driven decision-making. Going forward, the model can be further enhanced by adding additional data sources including weather forecasts pest and disease reports, satellite imaging, and irrigation patterns. Additionally, sophisticated deep learning models like CNNs or LSTMs may improve temporal and geographical analysis. Furthermore, the system's deployment as a real time decision support tool for farmers and policymakers can help close the gap between research and practical application. Additional methods to enhance performance and transparency include Explainable AI (XAI), feature selection, and hyperparameter tuning.

References

[1] Food and Agriculture Organization (FAO), "The State of Food and Agriculture 2017: Leveraging Food Systems for Inclusive Rural Transformation".
[2] W. Wolfert, L. Ge, C. Verdouw, and M. Bogaardt, "Big Data in Smart Farming – A review".
[3] X. Zhang, X. Xu, L. Zhang, and Z. Lu, "Internet of Things for Smart Agriculture".
[4] M. Kamilaris and F. Prenafeta-Boldú, "A Survey of the Applications of Machine Learning in Agriculture".
[5] Patel, S. Jain, and A. Agarwal, "Application of Random Forest for Crop Yield Prediction in Precision Agriculture".
[6] R. Kumar, A. Sharma, and A. Saini, "Cloud Computing in Agriculture: A Systematic Review".
[7] X. Zhang, X. Wang, and Y. Yang, "A wireless solution for greenhouse monitoring and control system based on ZigBee," Procedia Environmental Sciences.
[8] R. Rajeswari and E. Anand, "Smart irrigation system using IoT," International Journal of Engineering Science and Computing.
[9] Kamilaris, A. Kartakoullis, and F. X. Prenafeta-Boldú, "A review on the practice of big data analysis in agriculture," Computers and Electronics in Agriculture.
[10] S. Khodve, R. Shelke, and R. Deshmukh, "Smart agriculture monitoring system using IoT," International Journal of Recent Technology and Engineering.
[11] J. Zhang, J. Wang, and H. Wang, "Cloud-based intelligent agriculture monitoring system with IOT technology," Agricultural Engineering International: CIGR Journal.
[12] J. Jeong, J. Resop, and N. Mueller, "Random Forest model for forecasting crop yields using remote sensing data," Remote Sensing.
[13] B. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," Sensors.
[14] Chlingaryan, S. Sukkarieh, and B. Whelan, "Machine learning approaches for crop yield prediction and
[15] agricultural input optimization: A review," Computers and Electronics in Agriculture.
[16] N. Gandhi, B. Patel, and A. Parmar, "Plant disease detection using CNN and deep learning," International Journal of Computer Applications.
[17] Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," Computers and Electronics in Agriculture.
[18] S. Jabbar, M. A. Khan, and R. Ullah, "Smart environment monitoring system using wireless sensor networks in agriculture," Sustainable Computing: Informatics and Systems.
[19] Y. Zhou, Y. Zheng, and H. Liu, "Intelligent irrigation system using regression models based on real-time data," Journal of Ambient Intelligence and Humanized Computing.
[20] T. Fukatsu and M. Hirafuji, "Field monitoring using sensor-nodes with a web server," Journal of Robotics and Mechatronics.
[21] S. Wolfert, L. Ge, C. Verdouw, and M. J. Bogaardt, "Big data in smart farming–A review," Agricultural Systems.
[22] Jayaraman, M. Yavari, D. Georgakopoulos, A. Morshed, and A. Zaslavsky, "Internet of things platform for smart farming: Experiences and lessons learnt".
[23] R. Panday and R. R. Gautam, "Decision tree-based intelligent irrigation system," International Journal of Scientific & Technology Research.
[24] Boursianis, M. Papadopoulou, and P. G. Sarigiannidis, "Internet of Things (IoT) and agricultural drones for smart farming: A review," Internet of Things.
[25] Tzounis, N. Katsoulas, T. Bartzanas, and C. Kittas, "Internet of Things in agriculture, recent advances and future challenges," Biosystems Engineering.
[26] P. Ramesh and K. Vani, "Climate smart agriculture using predictive analytics," International Journal of Computer Applications.
[27] R. Kumar, D. Tripathi, and M. S. Khan, "Blockchain-based secure framework for agriculture data sharing using cloud," Sustainable Computing: Informatics and Systems.
[28] G. Manogaran, D. Lopez, and C. Thota, "Fog computing-based smart health care system for heart disease monitoring," Future Generation Computer Systems.