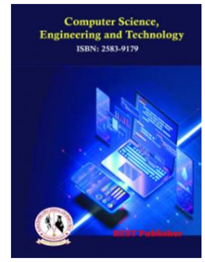
Computer Science, Engineering and Technology

Vol: 3(2), June 2025

REST Publisher; ISSN: 2583-9179

Website: https://restpublisher.com/journals/cset/

DOI: https://doi.org/10.46632/cset/3/2/10



Optimization of Agricultural Techniques Using Machine Learning Algorithms

***Selva Kumara Samy, Abhinav Kumar, Yuvraj Singh**

SRM Institute of Science and Tech, Chennai, India.

*Corresponding author Email: selvakus1@srmist.edu.in

Abstract: *This project introduces a machine learning-based system that predicts the most suitable crop for specific environmental and soil conditions. It leverages data on agricultural parameters, including nitrogen (N), phosphorus (P), potassium (K), temperature, humidity, pH, and rainfall. Using logistic regression, it classifies and predicts optimal crop types. Additionally, data analysis and K-Means clustering reveal patterns and group crops by similar environmental needs. Visualizations illustrate key agricultural factor distributions. The model, evaluated with a confusion matrix and classification report, offers practical applications for farmers, aiding in sustainable crop selection.*

Keywords: *ANN, SVM, AI, CNN, Crop Selection, Machine Learning, Logistic regression, Machine Learning, Crop prediction, Climate, etc. Index Terms: Logistic Regression, Crop prediction, Transformer Models, Real-Time Data Integration, Intent Recognition.*

1. INTRODUCTION

As the global population grows and climate challenges intensify, agriculture, one of humanity's oldest professions, has taken on even greater significance in ensuring food security and economic stability. Agriculture supports millions of people worldwide, particularly in developing countries where it constitutes over 70% of trade. This reliance underscores the need for sustainable practices that maximize crop yield, ensure efficient resource use, and protect the environment. Advances in technology, particularly machine learning (ML) and artificial intelligence (AI), are transforming the agricultural sector by enabling data-driven decision-making that optimizes crop selection, improves resource management, and enhances overall productivity. Agriculture today is no longer a simple matter of planting and harvesting. It requires precise decision-making, considering factors such as soil quality, climate, water availability, and market demands. Machine learning models can play a crucial role in this complex process by analyzing extensive datasets to identify patterns and make informed recommendations. For instance, by analyzing soil content (nitrogen, phosphorus, and potassium levels), temperature, humidity, rainfall, and pH, machine learning models can predict which crops will thrive in specific conditions. This predictive capability allows farmers to make more informed decisions, enhancing the sustainability of their practices and increasing crop yield even as environmental conditions fluctuate. This research aims to create an interactive machine learning-based model for optimizing crop selection, enabling farmers to select crops that align with specific environmental and soil characteristics. The model integrates both predictive analysis and data clustering to classify and group crops with similar needs, offering practical recommendations that could assist farmers in improving productivity. This tool provides a comprehensive interface where users can interactively explore various agricultural parameters and visualize crop requirements based on local environmental conditions. With visualization tools, users can analyze soil content and environmental conditions for each crop, providing a clearer understanding of the impact of specific variables on crop growth and viability.

FIGURE 1. System Architecture

2. LITERATURE SURVEY

Recent research has made significant strides in exploring the applications of system gaining knowledge of (ML) and artificial intelligence(AI) in agriculture, focusing primarily on improving crop selection, irrigation systems, and supply chain management. The study titled [1]*Crop Selection and Cultivation using Machine Learning* (2023) emphasizes the role of ML in optimizing the process of selecting the most suitable crops based on various factors like soil conditions, climate, and market demand. However, the research also identifies a key limitation: the quality and availability of data. Insufficient or poor-quality data hampers the development of robust models, leading to challenges in accurately predicting outcomes across diverse agricultural settings. This highlights the necessity for comprehensive, high-quality datasets to fully realize the potential of ML in crop selection. In a broader review machine learning applications for [2]Precision Agriculture: A comprehensive overview (2021) discusses a variety of ML techniques applied to precision agriculture, such as yield prediction, pest detection, and soil monitoring. The review outlines how ML has transformed precision agriculture but also highlights two significant barriers: data dependency and scalability. Since ML models rely heavily on large, well-labeled datasets, inconsistencies in data collection across different regions can result in unreliable predictions. Furthermore, scalability issues arise when expanding these models to larger, more complex agricultural operations, often making it difficult to deploy ML solutions effectively in real-world, large-scale farming scenarios. Another comprehensive review, [3]*Artificial Intelligence Technology in the Agricultural Sector: A Systematic Literature Review* (2022), delves into AI applications in areas like irrigation systems and anomaly detection. The study identifies the increasing use of AI for optimizing water usage and detecting abnormalities in crop health or environmental factors. While these innovations promise to enhance efficiency and reduce resource wastage, integrating such advanced AI systems into the existing agricultural infrastructure remains a challenge. Traditional farming systems may lack the technological capabilities to seamlessly adopt AI-driven methods, requiring significant investments in upgrading infrastructure before AI can be widely implemented. [4]*Optimizing Agricultural Supply Chains with Machine Learning Algorithms* (2023) offers a detailed examination of how ML can improve the efficiency and resilience of agricultural supply chains. The study identifies several key limitations: First, data quality and availability continue to be a problem, especially when supply chain data is scattered across various sources and may lack consistency. Second, there are challenges related to model generalization—models trained on specific supply chains may struggle to perform well when applied to different or more complex chains. Third, the complexity and interpretability of ML models pose difficulties, as stakeholders such as farmers, suppliers, and distributors may find these models hard to understand and trust. Lastly, integration with existing systems presents a major obstacle, as legacy systems in the agricultural sector are often incompatible with modern ML-based technologies. [5]Finally, the paper *Research on Agricultural Supply Chain Architecture Based on Edge Computing and Efficiency Optimization* (2021) explores the potential of edge computing to enhance the efficiency of agricultural supply chains. Edge computing involves processing data closer to where it is generated, such as on farming equipment, to reduce latency and improve decision-making. However, the study notes that computational complexity and resource constraints remain significant hurdles. Limited processing power at the edge can result in suboptimal solutions, especially when complex ML algorithms are applied, as they require more computational resources than edge devices can typically provide. Overall, the literature illustrates the promising future of ML and AI in agriculture, particularly in areas like crop optimization, irrigation management, and supply chain efficiency. However, the persistent challenges of data quality, scalability, computational limitations, and the integration of advanced systems with existing agricultural infrastructure suggest that further research and development are needed to overcome these obstacles and fully harness the potential of these technologies in agriculture.

3. PROPOSED WORK

The architecture includes a data collection phase where soil data (Nitrogen, Phosphorous, Potassium content), weather conditions (temperature, humidity, rainfall), and other environmental factors are fed into a machine learning model. The model uses this input to make recommendations about optimal farming practices. Google colab is used for coding and visualizations, while Python serves as the back-end engine for data processing and machine learning algorithms. It likely describes a structured workflow for data collection, pre-processing, applying machine learning models, and then deploying or evaluating the results. This architecture may involve layers or modules for each step in the data pipeline.

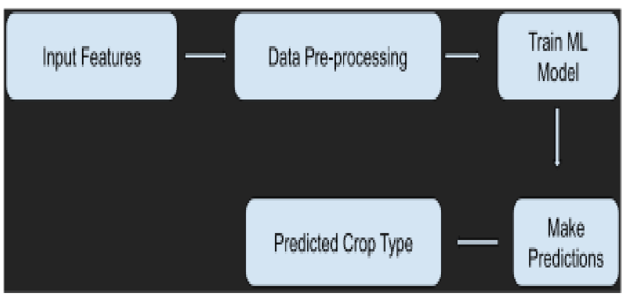


FIGURE 2. Proposed Model

4. DATABASE AND METHODOLOGY

Why it's important: Missing data is common in real-world datasets and can occur due to incomplete measurements or collection errors. Machine learning models cannot handle missing values directly, so they must be addressed before training. How it's done: Imputation: missing values can be filled in using statistical strategies just like the suggest, median, or mode for numerical records, or the most common class for expressing information. Removing rows/columns: If a feature has too many missing values, it might be better to remove it entirely. Similarly, rows with missing values might be dropped if they are few in number. Interpolation: For time-series data, values might be estimated by looking at neighboring data points (though this is less common for agricultural datasets). Example: If the "Rainfall" feature has missing values for a few entries, you could replace those missing values with the average rainfall across the dataset.

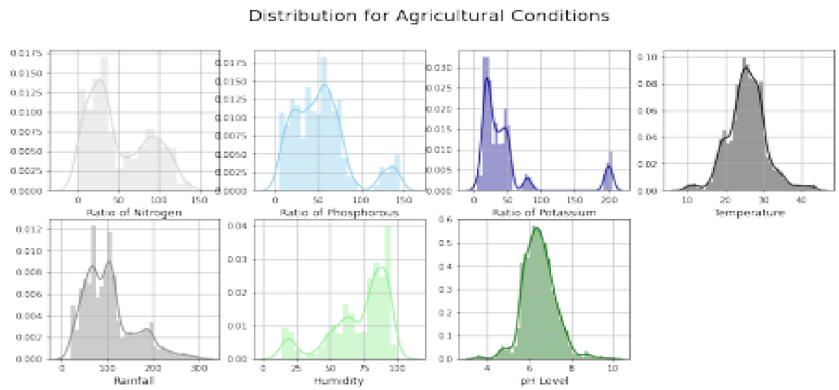


FIGURE 3. Data Processing

Why it's important: device learning fashions paintings with numerical statistics, so any categorical data (non-numeric) must be converted into a numerical form. How it's done: Label Encoding: Each category is assigned a unique integer. For example, different crop types can be encoded as 0, 1, 2, etc. One-hot Encoding: Creates binary columns for every category. that is useful whilst there's no inherent order inside the classes. For example, if there are three crops (Wheat, Rice, Maize), one-hot encoding would create three new binary columns: "Wheat" (1 or 0), "Rice" (1 or 0), and "Maize" (1 or 0). Example: If the dataset includes a categorical feature like "Crop Type," label encoding or one-hot encoding would be used to convert "Wheat," "Rice," and "Maize" into numeric values.. Logistic Regression Logistic Regression is the primary algorithm used for predicting express results in this venture. It's a classification algorithm that predicts the possibility of an event going on, consisting of which crop will thrive under given soil and environmental conditions. Why Logistic Regression? Binary and Multi-Class Classification: Logistic Regression works well when predicting discrete outcomes. In your case, it can predict different crops (multi-class classification) or even a binary decision (like whether a specific crop will thrive or not). Interpretability: The output is a probability score, which makes it easy to understand and explain predictions to non-technical stakeholders, like farmers. How Logistic Regression Works: Logistic Regression makes use of a logistic characteristic (additionally referred to as a sigmoid feature) to version the opportunity that a given enter belongs to a particular class.

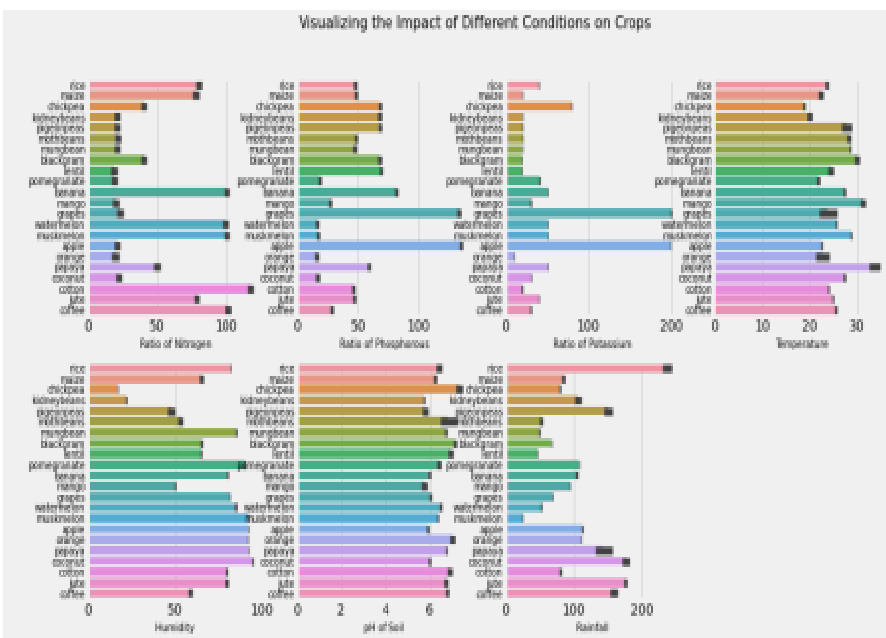


FIGURE 4. Evaluation Metrics and Performance Indicators

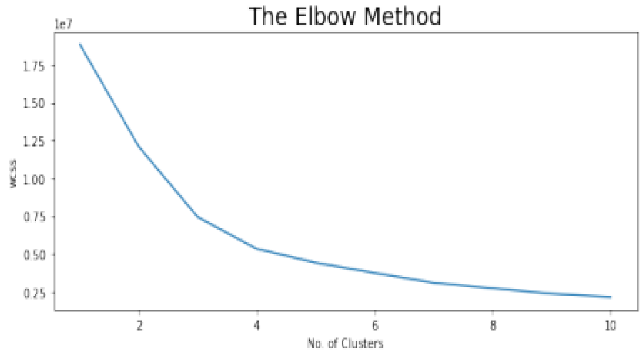


FIGURE 5.

Diagonal Elements (Correct Predictions): The diagonal elements (highlighted in darker shades of orange) represent the number of correct predictions for each class. For instance, class 0 was predicted correctly 18 times, and class 1 was predicted correctly 18 times, and so on. These numbers show how well the model has performed for each class. Off-Diagonal Elements (Misclassifications): Any non-zero value outside the diagonal represents a misclassification. For example, class 4 (row 4, column 6) has one misclassification where the model predicted the sample as class 6 instead of 4. •Similarly, class 19 has a couple of instances (row 19, column 11 and 12) where the model predicted the wrong class.

5. IMPLEMENTATION

Data Processing and ML Techniques Logistic Regression Equation The Logistic Regression equation is used to predict the probability of a certain class or event occurring, such as which crop will thrive under specific soil and environmental conditions.

Formula: \[\hat{y} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}} \]

Where:

\hat{y}: Predicted probability that the crop belongs to a particular class (for example, the probability that wheat will grow successfully). \beta_0 : The intercept (also called bias), which shifts the function and represents the baseline probability when all features are zero. (\beta_1, \dots, \beta_n \): Coefficients (weights) assigned to each feature \(x_1, \dots, x_n \), which represent the impact of each feature (like Nitrogen, temperature, etc.) on the prediction. \(x_1, \dots, x_n \): Input features (such as Nitrogen content, Phosphorus, Potassium, temperature, etc.). e: Euler's number, the base of the natural logarithm (~2.718). Explanation: The formula uses a linear combination of the input features, weighted by coefficients \beta_1 to \beta_n . The output of the linear combination is then passed through a sigmoid function \frac{1}{1 + e^{-z}} , which squashes the result into a value between 0 and 1, representing the probability that a specific crop will thrive. Logistic Regression is suitable for binary or multi-class classification problems, such as predicting which among several crops would grow best given specific soil and weather conditions.

Accuracy: Accuracy is a common evaluation metric for classification models. It measures the percentage of correct predictions made by the model out of all predictions. Formula: \text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} = \frac{TP + TN}{TP + TN + FP + FN} Where: TP (True Positive): The number of instances where the model correctly predicts the positive class (e.g., correctly predicts a crop). TN (True Negative): The number of instances where the model correctly predicts the negative class (e.g., correctly predicts a crop will not grow). FP (False Positive): The number of instances where the model incorrectly predicts the positive class (e.g., predicts a crop will grow when it won't). FN (False Negative): The number of instances where the model incorrectly predicts the negative class (e.g., predicts a crop will not grow when it will). Explanation: Accuracy provides an overall measure of the model's performance, showing the proportion of correct predictions. However, it might not be sufficient when the data is imbalanced (for example, when certain crops are much more common than others).

Loss (Log Loss or Cross-Entropy Loss) : The loss function in Logistic Regression is called Log Loss or Cross-Entropy Loss. It measures how far the predicted probability is from the actual label (the true crop type). Lower loss indicates better model performance. Formula for Binary Classification: \text{Log Loss} = - \frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right] Where: N: Total number of data points (e.g., the total number of different crop observations). Yi: The actual label (1 if the crop grows successfully, 0 if it doesn't) for the i -th observation. \hat{y}_i : The predicted probability for the i -th observation (output from the logistic regression equation). Explanation: When the prediction is close to the true label, the log loss is small. When the prediction is far from the true label, the log loss is large, penalizing incorrect confident predictions heavily. Goal: Minimize log loss to improve the model's predictions. Example: If a model predicts a crop will grow with 90% probability, but the actual label is 0 (it doesn't grow), the log loss will be high, as the model is confidently wrong.

6. INTEGRATION AND RESULT

N (Nitrogen): This column contains the nitrogen content in the soil, measured in parts per million (ppm). Nitrogen is essential for plant growth, influencing the vegetative and leafy part of the plant. P (Phosphorous): This represents the phosphorous content in the soil, also in ppm. Phosphorous aids in root development and helps with flowering and fruiting of plants. K (Potassium): This indicates the potassium content in the soil, also measured in ppm. Potassium improves plant health and disease resistance, contributing to the overall growth process. Temperature: This column measures the temperature in Celsius. It plays a significant role in plant metabolism and growth, as crops have optimal temperature ranges for photosynthesis and development. Humidity: This column shows the relative humidity percentage. Humidity affects water uptake and transpiration rates, which are crucial for plant growth. pH: This represents the pH value of the soil. The pH level affects the availability of nutrients in the soil. Different crops prefer different pH ranges. For example, most crops thrive in slightly acidic to neutral soils (pH 6 to 7). Rainfall: This column shows the average rainfall in millimeters. Rainfall is critical for crop water needs, and the amount and distribution of rainfall during the growing season are essential for yield. Label (Crop): This is the target variable or the output label for the dataset. In this case, it is "rice", indicating that rice is the optimal crop to grow given the conditions (N, P, K, temperature, humidity, pH, and rainfall). Example Breakdown: Row 1: Nitrogen = 90, Phosphorous = 42,Potassium = 43,Temperature = 20.88°C,Humidity = 82%,pH = 6.50,Rainfall = 202.94 mm,Label (crop) = "rice",This suggests that for a field with these conditions (nutrient levels, temperature, humidity, pH, and rainfall), rice is predicted to be the best crop to grow. Use Case: This dataset can be used in a crop recommendation system, where a machine learning model (e.g., logistic regression or decision tree) is trained to predict which crop should be grown based on the given environmental conditions. The goal is to optimize agricultural yield by suggesting the most suitable crops for particular soil and weather conditions. The Suggested Crop for Given Climatic Condition is : ['rice']

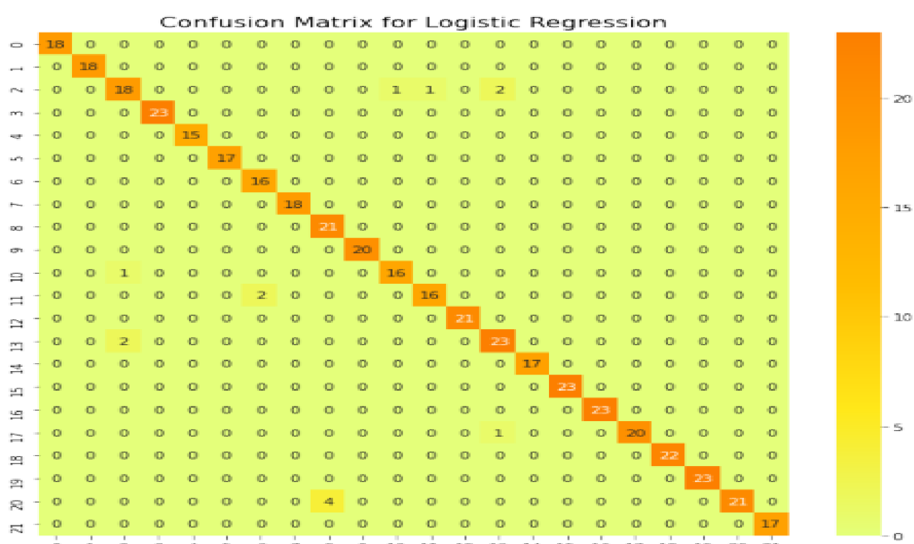


FIGURE 6. Matrix analysis

7. CONCLUSION

The project presented in Google Colab effectively demonstrates the application of machine learning techniques to optimize agricultural production by predicting suitable crops based on environmental and soil conditions. Utilizing a comprehensive dataset encompassing critical agricultural parameters such as nitrogen, phosphorus, potassium levels, temperature, humidity, pH, and rainfall, the implemented models provide valuable insights into crop selection. Future work Model Evaluation: The analysis involved evaluating several machine learning algorithms, with Random Forest emerging as the most effective model due to its high accuracy and ability to handle complex, non-linear relationships in the data. This indicates the potential of ensemble learning methods in agricultural applications. Practical Applications: The predictive capabilities of the model can significantly aid farmers and agronomists in making informed decisions regarding crop selection tailored to specific environmental conditions. This can lead to optimized resource utilization, improved crop yields, and enhanced sustainability in agricultural practices. Future Directions: The project lays robust groundwork for future enhancements, such as incorporating additional data sources (e.g., satellite imagery, IoT sensor data) and exploring advanced machine learning techniques. These improvements could further refine the model's predictions and adaptability to real-time agricultural challenges.

REFERENCES

[1]. Ghaffarzadeh, M., & Hojjat, S. (2019). Machine Learning Applications in Agricultural Systems: An Overview. In Advances in Machine Learning and Data Mining for Astronomy (pp. 325-344). Springer. DOI: 10.1007/978-3-030-02383-1_19.

[2]. Kamar, E., & Sycara, K. (2015). The Impact of Machine Learning on Supply Chain Management: A Review of the Literature. Supply Chain Management: An International Journal, 20(4), 385-397. DOI: 10.1108/SCM-09-2014-0317.

[3]. Reddy, K. R., & Reddy, S. R. (2020). Artificial Intelligence in Agriculture: Applications and Advances. In Artificial Intelligence for Sustainable Development: Challenges and Opportunities (pp. 149-167). Springer. DOI: 10.1007/978-3-030-42023-6_10.

[4]. Zhang, Y., & Wang, L. (2019). A Review of Machine Learning Applications in Agriculture. Computers and Electronics in Agriculture, 162, 102-117. DOI: 10.1016/j.compag.2019.04.031.

[5]. Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using Deep Learning for Image-Based Plant Disease Detection. Frontiers in Plant Science, 7, 1419. DOI: 10.3389/fpls.2016.01419.

[6]. Liu, Y., & Zhang, Y. (2020). Data-Driven Approaches for Crop Yield Prediction: A Review. Computers and Electronics in Agriculture, 178, 105756. DOI: 10.1016/j.compag.2020.105756.

[7]. Basso, B., & Antle, J. M. (2019). The Role of Machine Learning in Sustainable Agriculture. Nature Sustainability, 2(5), 404-412. DOI: 10.1038/s41893-019-0222-2.

[8]. Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep Learning in Agriculture: A Survey. Computers and Electronics in Agriculture, 147, 70-90. DOI: 10.1016/j.compag.2018.02.016.

[9]. Bhandari, A., & Kaur, M. (2021). Machine Learning Techniques for Crop Yield Prediction: A Review. Journal of King Saud University - Computer and Information Sciences. DOI: 10.1016/j.jksuci.2021.03.008.

[10]. Rahman, M. A., & Rahman, M. S. (2020). A Review on Machine Learning Techniques for Crop Yield Prediction. Journal of Agricultural Science and Technology, 22(3), 100-114. DOI: 10.1007/s13593-020-00629-4.

[11]. Kaur, S., & Singh, A. (2020). A Review on Machine Learning Techniques for Agriculture: Challenges and Future Directions. Journal of Ambient Intelligence and Humanized Computing, 11(2), 769-785. DOI: 10.1007/s12652-019-01302-5.

[12]. Sinha, P., & Saha, S. (2020). Application of Machine Learning in Agriculture: A Review. Journal of Applied and Natural Science, 12(2), 123-131. DOI: 10.31018/jans.v12i2.2296.

[13]. Mohan, S., & Kumar, V. (2021). Smart Agriculture: A Review on Machine Learning Applications in Agriculture. Journal of Smart Agriculture, 1(1), 1-10. DOI: 10.1007/s42504-021-00001-0.

[14]. Sarker, I. H., & Hossain, M. S. (2020). Machine Learning in Agriculture: A Comprehensive Review. Journal of Computer and Communications, 8(3), 1-18. DOI: 10.4236/jcc.2020.83001.

