International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 28-02-2026 Accepted: 02-04-2026 Published: 10-04-2026

Smart Crop Prediction Using Soil Moisture by Machine Learning

Subrhamanyam Kolusu
Assistant Professor
Department of CSE
Seshadri Rao Gudlavalleru
Engineering College
Gudlavalleru, Andhra Pradesh, India
subramanyamkolusu53@gmail.com

Attili Nithish
Department of CSE
Seshadri Rao Gudlavalleru
Engineering College
Gudlavalleru, Andhra Pradesh, India
attilinithish128@gmail.com

Chandika Vishnu Vardhan
Department of CSE
Seshadri Rao Gudlavalleru
Engineering College
Gudlavalleru, Andhra Pradesh, India
vishnuvarmach2344@gmail.com

Chilakala Prasanna Lakshmi
Department of CSE
Seshadri Rao Gudlavalleru
Engineering College
Gudlavalleru, Andhra Pradesh, India
prasannalakshmichilakala2005@gmail.com

Anaparthy Anurag
Department of CSE
Seshadri Rao Gudlavalleru
Engineering College
Gudlavalleru, Andhra Pradesh, India
anuraganaparthy@gmail.com

Abstract — Accurate crop selection is essential for improving agricultural productivity and ensuring sustainable resource utilization. Conventional crop planning methods often rely on experiential judgment rather than systematic analysis of soil and environmental conditions, which can lead to inefficient input usage and reduced yields. Our approach treats crop selection as a classification challenge where soil nutrient profiles and climatic variables serve as the primary predictors. We scrutinized the performance of several widely used models, from basic Decision Trees to high-performance gradient-boosted machines, to see how well they could map these inputs to specific crop suitability. To guard against overfitting and ensure the model generalizes well to new regions, we applied a combination of standard hold-out validation and iterative cross-validation techniques. Experimental results demonstrate that XGBoost achieves superior predictive performance and generalization capability, attaining an accuracy of approximately 99% on the test dataset. To enhance interpretability, SHAP (SHapley Additive exPlanations) was applied to analyze feature contributions. The explainability analysis indicates that climatic factors, particularly humidity and rainfall, exert the highest influence on crop classification decisions. The proposed framework provides a reliable and interpretable decision-support tool for data-driven crop planning. By combining high predictive accuracy with model transparency, the system contributes toward intelligent agricultural management and sustainable farming practices.

Keywords — *Soil–Climate Feature Modeling, Crop Classification System, Gradient Boosted Trees, Model Interpretability Analysis, Agricultural Decision Support, Data-Driven Cultivation Planning.*

I. Introduction

Crop selection is a critical decision-making process that directly influences agricultural productivity and economic stability. Variations in soil composition and dynamic climatic patterns have made crop planning increasingly complex. Selecting an unsuitable crop may result in reduced yield, inefficient use of fertilizers and water resources, and financial losses for farmers. Therefore, aligning crop choice with local agro-climatic and soil conditions is essential for sustainable agricultural development.

In many regions, crop planning still relies on traditional knowledge and prior cultivation experience. Although such methods may be effective under stable environmental conditions, they often lack adaptability to dynamic weather patterns and soil variability. This limitation highlights the need for systematic and data-driven approaches to improve crop suitability assessment.

Agronomic development is driven by a synergy of variables, ranging from macronutrient profiles (N, P, K) and pH to ambient thermal and hydrometeorological conditions. The nonlinear

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 28-02-2026 Accepted: 02-04-2026 Published: 10-04-2026

relationships among these parameters make conventional analytical techniques insufficient for accurate prediction. Recent advancements in supervised learning have demonstrated strong performance in classification-based decision systems across multiple domains [1]–[3], motivating their application to agricultural prediction tasks.

Machine learning models have increasingly been employed in agricultural analytics to improve crop yield forecasting and resource optimization [4]–[9]. More advanced learning frameworks, including ensemble techniques and deep learning models, have further enhanced predictive capability [10]–[17]. Among these methods, Extreme Gradient Boosting (XGBoost) has shown superior performance in structured datasets due to its regularization mechanism and ability to model complex feature interactions [18].

Although predictive accuracy is essential, interpretability remains equally critical in agricultural applications. Decision-support systems must provide transparency regarding how environmental variables influence crop recommendations. Diagnostic frameworks like SHAP facilitate a granular understanding of model behavior by quantifying the specific impact of each input variable [19]. Current research increasingly prioritizes the deployment of transparent algorithmic architectures within the domain of precision agronomy [20].

This research assesses various supervised learning paradigms—ranging from linear and proximity-based estimators to tree-based ensembles like XGBoost—for mapping environmental and pedological variables to optimal cultivar selections. Model performance is assessed using standard accuracy metrics and cross-validation techniques. The final framework integrates XGBoost for classification and SHAP for interpretability, thereby delivering a high-accuracy and interpretable decision-support framework for sustainable crop planning.

II. Literature Review

Machine learning has been widely adopted for solving classification problems across multiple domains. In their comparative analysis, Latha et al. [1] showed that differences in predictive accuracy are often linked to how learning algorithms process dataset characteristics. Markapudi et al. [2] formulated a hybrid predictive structure that strengthened prediction results by integrating complementary modeling approaches. In a separate domain, Indira et al. [3] applied deep neural architectures to complex medical datasets, illustrating the adaptability of advanced learning models to real-world scenarios. Together, these contributions highlight the importance of comparative experimentation and hybrid modeling strategies in supervised learning research.

In the agricultural domain, early research emphasized the use of climatic variables for predictive modeling. Veenadhari et al. [4] showed that machine learning-based forecasting methods outperform traditional estimation techniques in crop yield prediction. Kumar [5] proposed a crop selection framework based on soil and environmental features, highlighting the importance of algorithm-driven recommendations in precision agriculture. Shakoor [6] further demonstrated that incorporating multiple input parameters improves agricultural output prediction.

Subsequent studies validated the criticality of data normalization and dimensional transformation. Nigam [7] reported that model performance is strongly influenced by dataset characteristics and feature selection strategies. Kalimuthu et al. [8] emphasized structured preprocessing for improved classification robustness, while Nishant et al. [9] highlighted the adaptability of data-driven models to diverse agro-environmental contexts.

Recent developments in machine learning have expanded modeling flexibility for complex prediction tasks. Elavarasan and Vincent [10] explored reinforcement learning for agricultural yield estimation, addressing variability in environmental conditions. Reddy and Kumar [11], along with Gajula et al. [12], demonstrated that combining soil nutrient attributes with climatic variables enhances predictive precision by capturing multidimensional interactions.

Comprehensive reviews have further examined the role of machine learning in modern agriculture. Dharani et al. [13] analyzed deep learning applications and identified operational constraints affecting model deployment. Sharma et al. [14] presented a broad overview of precision agriculture

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 28-02-2026 Accepted: 02-04-2026 Published: 10-04-2026

systems, emphasizing the transition toward data-driven management. Sharma et al. [15] combined regression and deep learning methods for yield estimation, while Kulyal and Saxena [16] identified ensemble learning as an effective approach for managing agricultural data variability. Kumar [17] reinforced the effectiveness of supervised learning models for agricultural prediction tasks.

Despite these advancements, most existing studies emphasize yield forecasting rather than comprehensive crop suitability classification. Additionally, limited research integrates high-performing ensemble models with interpretability frameworks in agricultural decision systems. This study addresses these gaps by evaluating multiple supervised models and incorporating XGBoost with SHAP-based interpretation to deliver both predictive accuracy and model transparency.

III. Proposed Methodology

The proposed crop prediction framework leverages supervised learning techniques to support crop selection decisions using soil and environmental attributes as inputs. The overall methodology follows a sequential process that begins with data acquisition and preparation, proceeds through model training and evaluation, and concludes with crop recommendation generation.

A. System Workflow

The overall system architecture is illustrated in Fig. 1. The workflow begins with soil and climatic data collection. The dataset undergoes preprocessing, including cleaning and feature scaling. Multiple supervised classification algorithms are then trained and comparatively evaluated. Based on the observed validation fidelity, the most resilient model was prioritized for deployment to ensure the reliability of the automated crop selection system. This modular architecture ensures scalability and facilitates future integration of real-time environmental data.

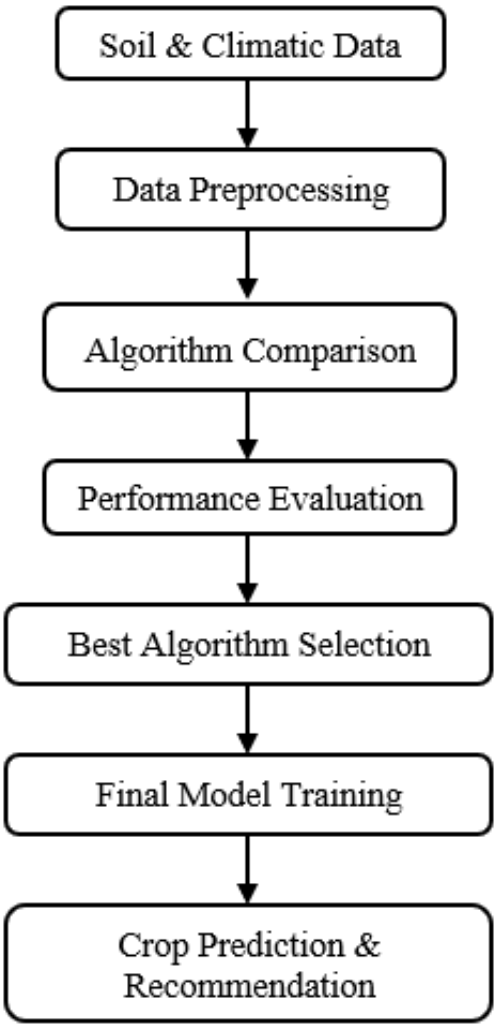

Fig. 1. Computational workflow of the introduced crop selection architecture.

B. Dataset Description

The study utilizes a publicly accessible agricultural dataset containing soil nutrient measurements and environmental variables frequently employed in crop recommendation research. Each record contains the attributes such as nitrogen, phosphorus, potassium, and pH, along with environmental variables including temperature, humidity, and rainfall. Each data instance represents a unique combination of soil and climatic conditions and is mapped to a crop label that indicates the most suitable crop for those conditions. Collectively, these attributes provide sufficient information to learn meaningful patterns for crop suitability prediction. The dataset comprises 2200 observations characterized by seven numerical input attributes related to soil nutrients and climatic conditions. The target variable represents 22 distinct crop categories

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 28-02-2026 Accepted: 02-04-2026 Published: 10-04-2026

suitable for varying environmental profiles. The class distribution is reasonably balanced, enabling unbiased model training and reliable performance evaluation across different crop types.

TABLE I. DESCRIPTION OF DATASET ATTRIBUTES

Attribute	*Description*
Nitrogen	Amount of Nitrogen present in soil
Phosphorus	Level of Phosphorus in soil
Potassium	Potassium concentration in soil
pH	Indicator of soil acid–base balance
Temperature	Average temperature of the region
Humidity	Average atmospheric humidity
Rainfall	Average annual rainfall
Crop	Target crop label

C. Data Preprocessing

The dataset underwent an initial quality inspection to identify incomplete records and structural inconsistencies. Continuous attributes were normalized to prevent scale imbalance during learning. The crop labels were transformed into numerical form to enable supervised classification. A standard 80% to 20% hold-out split was applied to the finalized dataset to facilitate a rigorous evaluation of the system's predictive stability.

D. Feature Selection

All soil nutrient and climatic attributes are retained due to their direct relevance to crop growth. These features are known to significantly influence agricultural productivity and are therefore included in the final modeling process without dimensionality reduction.

E. Model Training

To examine predictive strength across different algorithmic approaches, several supervised classifiers were fitted to the prepared dataset. The investigated frameworks encompass linear estimators, proximity-based learners, decision trees, support vector architectures, and gradient-boosted ensembles. Each classifier was trained using identical feature inputs to ensure a consistent experimental framework.

During training, algorithm-specific optimization procedures were applied to minimize misclassification error on the training partition. Logistic Regression establishes a probabilistic linear boundary, KNN assigns classes based on neighborhood proximity, and Decision Trees construct rule-based partitions of the feature space. The SVM model determines optimal separating hyperplanes, while XGBoost iteratively builds boosted decision trees under a regularized objective to improve predictive precision and reduce overfitting effects.

Through this structured training approach, each algorithm learns the complex interactions between soil attributes, environmental conditions, and crop suitability labels. The comparative training framework enables an objective assessment of model generalization capability and robustness prior to final model selection.

The final model employs XGBoost within a standardized preprocessing pipeline to ensure consistent feature scaling prior to boosting. The boosting configuration includes 300 trees with a shrinkage factor of 0.05 and a depth constraint of six levels to control model expressiveness. Row and feature sampling proportions were both fixed

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 28-02-2026 Accepted: 02-04-2026 Published: 10-04-2026

at 0.8 to enhance generalization stability. A fixed initialization seed was applied for experimental consistency, and multi-class logarithmic loss served as the optimization objective throughout training.

F. Model Evaluation and Selection

Model effectiveness was primarily assessed using testing accuracy and cross-validation consistency. Comparative results reveal that XGBoost achieved the strongest predictive outcome among the examined classifiers. Its boosting strategy enables efficient modeling of nonlinear interactions between soil composition and environmental conditions. Based on quantitative performance comparison, XGBoost was selected as the final predictive model. SHAP analysis was subsequently incorporated to quantify feature influence and enhance interpretability within the recommendation framework.

IV. Experimental Results And Discussion

The following discussion provides a quantitative scrutiny of the developed crop recommendation engine. Multiple supervised classifiers were trained and systematically compared to determine the most dependable classifier for crop prediction. The evaluation focuses on computed performance measures, including training accuracy, testing accuracy, and cross-validation results.

A. Experimental Setup

Numerical modeling was performed using a standardized agricultural dataset that pairs substrate nutrient concentrations (N-P-K and pH) with atmospheric drivers, ensuring a comprehensive analysis of the environmental factors governing crop suitability. The output variable represents recommended crop categories under defined environmental conditions. Continuous variables were scaled prior to training, and the dataset was partitioned using an 80:20 split for evaluation. Cross-validation was additionally applied to verify model stability. All experiments were conducted in Python under uniform preprocessing conditions to ensure methodological fairness.

B. Comparative Model Performance

The numerical comparison of classifier performance is presented in Table II. The analysis includes training accuracy, testing accuracy, and cross-validation results to evaluate both learning capacity and generalization strength.

TABLE II. Comparative Results Of Supervised Models

Model Name	***Train Accuracy***	***Test Accuracy***	***CV Accuracy***
Logistic Regression (LR)	0.9738	0.9727	0.9713
K-Nearest Neighbor Classifier (KNN)	0.9778	0.9705	0.9705
Decision Tree Classifier (DT)	1.0000	0.9795	0.9868
Support Vector Classifier (SVC)	0.9858	0.9841	0.9827
XGBoost	1.0000	0.9932	0.9941

The results indicate that Logistic Regression and K-Nearest Neighbors provide strong baseline performance but show slightly lower generalization compared to more advanced models. The Decision Tree classifier achieves perfect training accuracy, indicating strong fitting capability; however, the difference between training and testing accuracy suggests minor overfitting. The Support Vector Machine demonstrates stable and competitive performance across training and testing phases.

Among all evaluated models, XGBoost achieves the highest testing accuracy of 99.32% along with the strongest cross-validation performance. The consistency between its training and validation metrics confirms its superior generalization capability. The gradient boosting framework effectively captures nonlinear relationships between soil nutrients and climatic conditions, leading to improved predictive performance.

Fig. 2 highlights the performance disparities between the models, offering a graphic

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 28-02-2026 Accepted: 02-04-2026 Published: 10-04-2026

representation of the accuracy benchmarks discussed in Table II.

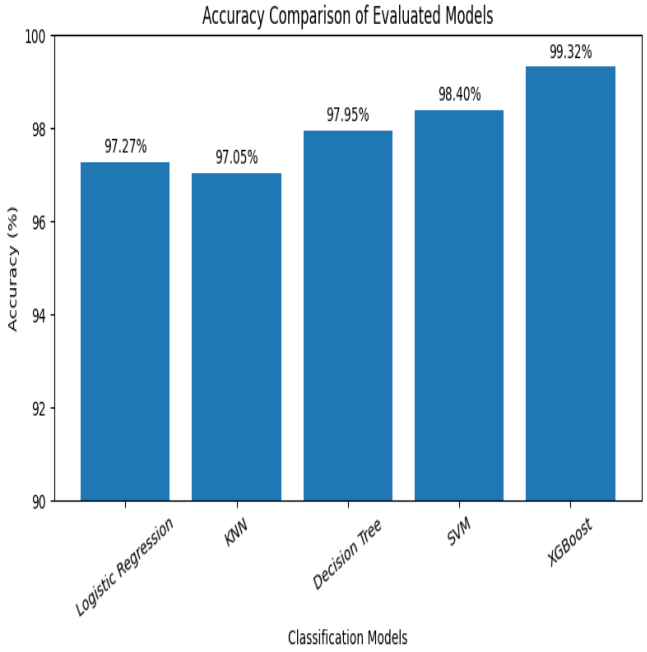

Fig. 2. Comparative analysis of predictive resolution across various computational architectures.

C. Confusion Matrix Analysis

To evaluate prediction consistency across individual crop classes, the confusion matrix corresponding to the selected XGBoost model is depicted in Fig. 3.

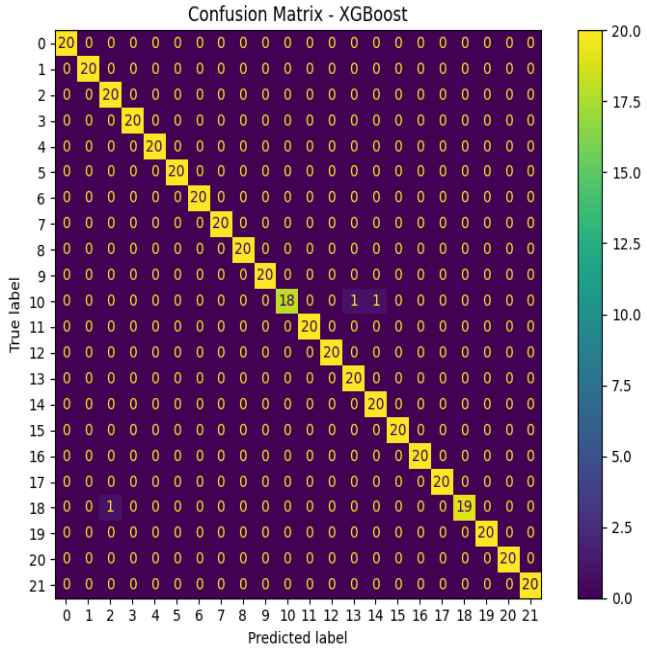

Fig. 3. Classification performance of the XGBoost model illustrated via a multi-class confusion matrix.

The confusion matrix exhibits strong diagonal dominance, indicating that most crop categories are correctly classified. Only a minimal number of off-diagonal elements are observed, representing minor misclassifications between crop classes with similar environmental requirements. The overall distribution confirms the robustness and high discriminative capability of the XGBoost model across multiple crop categories.

D. Feature Importance Analysis

Understanding the contribution of individual features is critical for practical agricultural decision-making. The relative feature contribution values obtained from the XGBoost classifier are illustrated in Fig. 4.

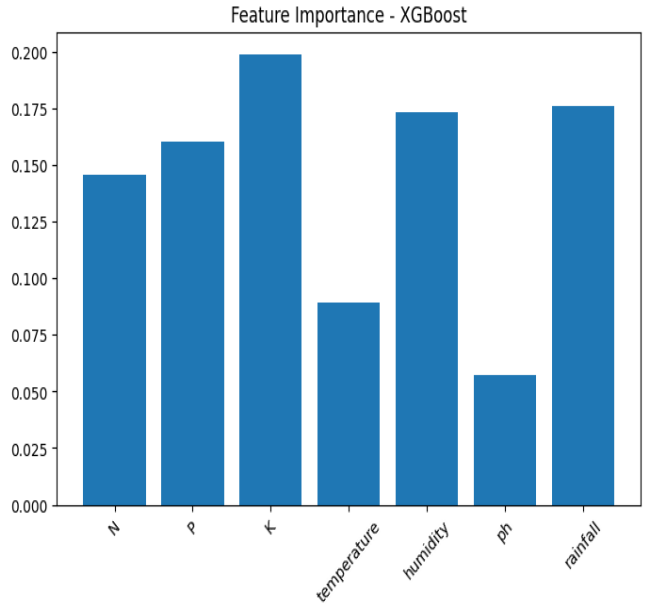

Fig. 4. Feature importance obtained from the XGBoost Model.

The results indicate that Potassium (K) has the highest contribution to the classification process, followed by Rainfall and Humidity. Nitrogen and Phosphorus demonstrate moderate influence, while Temperature and soil pH exhibit comparatively lower importance. These findings suggest that nutrient availability and moisture-related variables play a dominant role in determining crop suitability within the evaluated dataset.

E. SHAP-Based Interpretation

While traditional feature importance provides a global ranking of attributes, it does not quantify their overall impact on prediction magnitude.

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 28-02-2026 Accepted: 02-04-2026 Published: 10-04-2026

Therefore, SHAP analysis was performed to evaluate feature influence more comprehensively. A SHAP-based summary of feature influence is provided in Fig. 5.

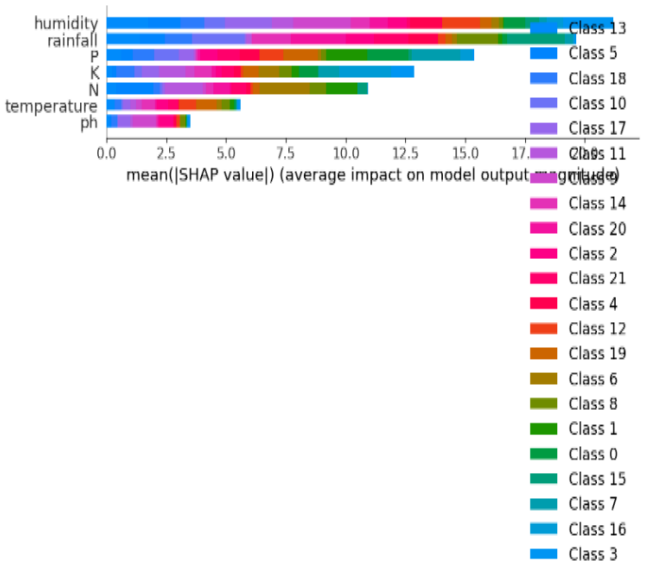

Fig. 5. Aggregate SHAP value distribution illustrating feature contribution magnitude.

The SHAP analysis confirms that Humidity and Rainfall exert the highest average influence on prediction outcomes. Potassium and Phosphorus also contribute significantly to crop classification decisions. In contrast, pH and Temperature demonstrate relatively lower impact. The SHAP framework enhances transparency by providing a quantitative understanding of how soil and climatic variables influence the model's output.

F. Discussion

The experimental results demonstrate that ensemble-based gradient boosting significantly outperforms traditional classification models in crop suitability prediction tasks. XGBoost achieves superior accuracy and maintains stable performance across validation measures. The confusion matrix confirms strong class-level reliability, while feature importance and SHAP analyses provide interpretability and insight into environmental factor contributions.

Overall, the integration of boosting-based classification with interpretability analysis establishes a reliable and transparent framework suitable for intelligent crop recommendation systems.

V. Conclusion And Future Work

The present study engineers a supervised classification architecture tailored to map pedological profiles and climatic variables to optimal agricultural outputs. By combining macronutrient measurements, soil acidity levels, and atmospheric variables such as temperature, humidity, and precipitation, the system facilitates structured, evidence-driven crop selection tailored to specific growing conditions. Multiple supervised classification algorithms were evaluated to determine the most reliable predictive model.

Comparative experimental analysis demonstrates that ensemble learning approaches provide superior predictive performance compared to traditional classifiers. Among the evaluated models, XGBoost achieved the highest testing and cross-validation accuracy, confirming its strong generalization capability and robustness in modeling nonlinear relationships within agricultural datasets.

The confusion matrix analysis further validates the stability of the selected model, showing minimal misclassification across crop categories. In addition, feature importance evaluation highlights the significant influence of nutrient concentration and moisture-related variables in determining crop suitability.

To enhance interpretability, SHAP-based explainability was incorporated to quantify feature contributions to prediction outcomes. This synthesis guarantees high-resolution forecasting while offering traceable insights into the specific environmental parameters that dictate the system's output logic.

Future research will aim to integrate live environmental data streams, broaden the range of supported crop types, and implement the system within web-based or mobile applications to enhance accessibility and practical usability. These enhancements will improve scalability, adaptability, and accessibility, enabling broader application in smart and precision agriculture systems. Additionally, integrating advanced ensemble tuning and region-specific calibration mechanisms may further enhance predictive robustness across heterogeneous ecological landscapes.

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 28-02-2026 Accepted: 02-04-2026 Published: 10-04-2026

REFERENCES

[1] K. Latha, M. Baburao, and C. Kavitha, "A comparative study on Logit Leaf Model (LLM) and Support Leaf Model (SLM) for predicting customer churn," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 5, pp. 1628–1632, May 2019.

[2] B. Markapudi, K. J. Latha, and K. Chaduvula, "A new hybrid classification algorithm for predicting customer churn," in *Proc. Int. Conf. Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, 2021, pp. 1–4.

[3] D. N. Indira, V. S. Lakshmi, B. R. Markapudi, A. Yannam, M. B. Prasad, C. S. Babu, and K. K. Rao, "Detection of cardiac arrhythmia using multi-perspective convolutional neural network for ECG heartbeat classification," *Revue d'Intelligence Artificielle*, vol. 36, no. 4, pp. 629–634, Aug. 2022.

[4] S. Veenadhari, B. Misra, and C. D. Singh, "Machine learning approach for forecasting crop yield based on climatic parameters," in *Proc. Int. Conf. Computer Communication and Informatics (ICCCI)*, 2014, pp. 1–5.

[5] R. Kumar, "Crop selection method to maximize crop yield rate using machine learning technique," in *Proc. Int. Conf. Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, 2015.

[6] M. T. Shakoor, "Agricultural production output prediction using supervised machine learning techniques," in *Proc. 1st Int. Conf. Next Generation Computing Applications (NextComp)*, 2017.

[7] A. Nigam, "Crop yield prediction using machine learning algorithms," in *Proc. Fifth Int. Conf. Image Information Processing (ICIIP)*, 2019.

[8] M. Kalimuthu, P. Vaishnavi, and M. Kishore, "Crop prediction using machine learning," in *Proc. Third Int. Conf. Smart Systems and Inventive Technology (ICSSIT)*, 2020.

[9] P. S. Nishant, P. S. Venkat, B. L. Avinash, and B. Jabber, "Crop yield prediction based on Indian agriculture using machine learning," in *Proc. Int. Conf. for Emerging Technology (INCET)*, 2020, pp. 1–4.

[10] D. Elavarasan and P. D. Vincent, "Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications," *IEEE Access*, vol. 8, pp. 86886–86901, 2020.

[11] D. J. Reddy and M. Rudra Kumar, "Crop yield prediction using machine learning algorithm," in *Proc. 5th Int. Conf. Intelligent Computing and Control Systems (ICICCS)*, 2021.

[12] A. K. Gajula et al., "Prediction of crop and yield in agriculture using machine learning technique," in *Proc. 12th Int. Conf. Computing Communication and Networking Technologies (ICCCNT)*, 2021, pp. 1–5.

[13] M. K. Dharani, R. Thamilselvan, P. Natesan, P. C. Kalaivaani, and S. Santhoshkumar, "Review on crop prediction using deep learning techniques," *J. Phys.: Conf. Ser.*, vol. 1767, no. 1, Art. no. 012026, 2021.

[14] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, "Machine learning applications for precision agriculture: A comprehensive review," *IEEE Access*, vol. 9, pp. 4843–4873, 2020.

[15] P. Sharma, P. Dadheech, N. Aneja, and S. Aneja, "Predicting agriculture yields based on machine learning using regression and deep learning," *IEEE Access*, vol. 11, pp. 111255–111264, 2023.

[16] M. Kulyal and P. Saxena, "Machine learning approaches for crop yield prediction: A review," in *Proc. 7th Int. Conf. Computing, Communication and Security (ICCCS)*, 2022, pp. 1–7.

[17] Y. J. N. Kumar, "Supervised machine learning approach for crop yield prediction in agriculture sector," in *Proc. 2020 5th Int. Conf. Communication and Electronics Systems (ICCES)*, 2020.

[18] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785–794.

[19] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[20] I. Attri, L. K. Awasthi, and T. P. Sharma, "Machine learning in agriculture: A review of crop management applications," *Multimedia*

Tools and Applications, vol. 83, no. 5, pp. 12875–12915, 2024.