

Smart Crop Recommendation System Using Machine Learning

Raveena Kumari¹, Shristi², Shreya KV³, Dr. Sreenivasa B C⁴

^{1,2,3}*Department of Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India*

⁴*Associate Professor of Department of Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India*

Abstract—The Smart Crop Recommendation System is a data-driven agricultural decision-support framework that leverages machine learning to identify the most suitable and economically beneficial crops for a given region. The model evaluates key soil attributes such as nitrogen, phosphorus, potassium, pH, along with meteorological variables including temperature, humidity, and rainfall. Using a Random Forest-based prediction engine, the system processes these multi-dimensional inputs to generate the top three crop options best aligned with local conditions. Real-time weather information is incorporated through the OpenWeatherMap API, enabling adaptive and location-specific recommendations. The system also provides additional insights through profit estimation, irrigation guidance, and farmer feedback analysis, ensuring a comprehensive decision-making workflow. By integrating environmental intelligence with predictive analytics, the proposed system supports sustainable farming practices, improves resource utilization, and empowers farmers to make informed and profitable crop choices.

Index Terms—Machine Learning, Random Forest Classifier, Real-Time Weather Integration, Smart Crop Recommendation, Soil Nutrient Analysis.

I. INTRODUCTION

Agriculture remains a cornerstone of the economic and social framework of developing countries like India, where a significant share of the population depends on farming as its primary occupation. However, crop selection in many regions is still driven by traditional practices, individual judgment, and inherited experience, rather than by a scientific assessment of soil and climatic conditions. Such an approach often leads to inappropriate crop choices, reduced

productivity, inefficient use of resources, and avoidable financial setbacks. These challenges are further exacerbated by increasing climate variability, soil degradation, and unpredictable weather patterns, highlighting the urgent need for intelligent systems that can support farmers with evidence-based decision-making.

With the rapid progress of Artificial Intelligence (AI) and Machine Learning (ML), agriculture has witnessed transformative changes similar to those observed in other domains. ML techniques are capable of processing complex, multi-variable datasets and identifying intricate relationships that cannot be easily recognized through manual analysis. These capabilities enable the development of systems that provide accurate, data-centric recommendations by evaluating soil nutrients, environmental conditions, and climatic influences, thereby assisting farmers in making informed cropping decisions.

In this context, the present research introduces a Smart Crop Recommendation System that utilizes machine learning to recommend the most suitable and profitable crops for a specific region. The system analyzes crucial environmental factors such as nitrogen, phosphorus, potassium, pH level, temperature, humidity, and rainfall to generate the top three crop suggestions. A Random Forest classifier serves as the core predictive model due to its strong performance, high accuracy, and reliability across diverse agricultural datasets.

To ensure real-time relevance, the system integrates the OpenWeatherMap API to gather up-to-date temperature, humidity, and rainfall information, thereby enabling dynamic and location-specific predictions. The system further incorporates modules

for profit estimation and irrigation guidance, helping farmers make comprehensive decisions that combine environmental suitability with economic considerations. A simple and accessible user interface ensures that individuals with minimal technical experience can also benefit from the system.

By promoting scientific, data-driven crop planning, the proposed system has the potential to enhance crop yield, improve resource efficiency, reduce uncertainties, and support sustainable farming practices. Ultimately, this work advances the scope of smart agriculture by offering farmers reliable, technologically enhanced insights for better and more profitable crop selection.

II. LITERATURE SURVEY

The growing application of machine learning in agriculture has accelerated the development of intelligent systems aimed at improving crop selection, enhancing soil management, and increasing agricultural productivity. Researchers have explored various approaches for crop recommendation, soil fertility assessment, and yield prediction, each employing different machine learning or hybrid techniques. The following review summarises significant contributions in the field, along with their methodologies, outcomes, and limitations.

One of the earlier contributions was made by Pudumalar et al., who implemented a precision agriculture framework using multiple algorithms such as KNN, CHAID, Naïve Bayes, and Random Tree to recommend crops based on soil attributes. Although their model performed well, its reliance on large datasets and the absence of real-time environmental inputs restricted its practical deployment. Building on this concept, Saranya and Mythili proposed a soil classification approach centred on macronutrient levels and experimented with several machine learning models. Their results showed that SVM achieved a maximum accuracy of 96%, yet the dependence on separate soil and crop datasets reduced the system's suitability for dynamic field applications. Further research by Venkat Narayana Rao et al. involved the use of Random Forest and Decision Tree models to predict soil quality by analysing physical, chemical, and biological indicators. Despite demonstrating good classification capability, the performance was limited due to a narrower feature set.

Sharavani et al. also explored soil categorisation and crop suggestion using SVM and ensemble methods, obtaining promising results, though their study did not extend to fertilizer recommendations.

A more extensive multi-parameter analysis was conducted by Muneshwara M.S. et al., who combined SVM, KNN, and Random Forest models to evaluate soil fertility using NPK levels, pH, temperature, and moisture. Their findings emphasised the advantage of ensemble learning techniques but highlighted scalability constraints and dataset limitations. Rahman et al. also reported strong performance, achieving 94.5% accuracy with SVM for soil and land-type classification; however, their model lacked economic considerations such as profitability and market demand.

Several advanced studies introduced hybrid and multi-input models. Ashok Kumar et al. developed a crop and bio-fertilizer recommendation system that integrated KNN with image processing to analyse nutrient data and remote sensing images. While innovative, the approach required significant computational resources, limiting its accessibility for small-scale farmers. Archana and Saranya implemented an ensemble-based crop yield and fertilizer prediction model, attaining about 92% accuracy. Nonetheless, the applicability of their system was limited by dataset size and predefined assumptions.

Review-based studies, including the comprehensive survey by Shabari Shedthi B. et al., highlighted the diverse strengths of models such as KNN, SVM, Decision Trees, Random Forest, Naïve Bayes, ANN, and XGBoost across different agricultural datasets. XGBoost, in particular, demonstrated high predictive capability, achieving accuracies up to 99.31% in multi-feature crop recommendation tasks. The review consistently noted that hybrid and ensemble approaches offer superior reliability across varied agricultural conditions.

Other recent works have explored real-time and large-scale deployment challenges. Lakshmi et al. integrated weather patterns, land characteristics, and water usage into a big-data framework, which improved predictive performance but restricted applicability to a specific crop set. Viviliya and Vaidhehi proposed a hybrid crop recommendation model incorporating NPK levels, pH, and organic carbon, achieving 96% accuracy while remaining confined to a limited range of crops.

Systems such as AgroConsultant and mobile-based soil pH prediction solutions demonstrate the rapid adoption of user-friendly, AI-enabled agricultural platforms.

From the reviewed literature, several consistent research gaps emerge:

1. Limited incorporation of real-time environmental parameters such as humidity, rainfall, and temperature.
2. Minimal emphasis on economic considerations like profitability and market fluctuations.
3. Many existing systems provide a single recommendation rather than multiple ranked alternatives.
4. Insufficient integration between soil testing results and automated decision-support systems.
5. Lack of irrigation guidance or water-use recommendations in most models.

The Smart Crop Recommendation System proposed in this research addresses these gaps by integrating real-time weather data using the OpenWeatherMap API, applying a robust Random Forest model for multi-feature prediction, and offering the top three suitable crops along with profitability estimation and irrigation suggestions. This comprehensive approach enhances regional adaptability and provides farmers with actionable, data-driven support that surpasses the capabilities of existing systems.

III. METHODOLOGY

3.1 Dataset description

We trained our model on a curated dataset containing agronomic and climatic descriptors for multiple crops (`final_crop_dataset.csv`). Each sample records the crop label, soil type, season, categorical nutrient levels (N/P/K), and climate/soil measurements (temperature, rainfall, humidity, pH). Additionally, a `profit_data.csv` mapping is used to estimate the profitability of predicted crops.

3.2 Data Preprocessing

Raw data were cleaned and normalized using deterministic preprocessing: numeric columns were median-imputed, categorical columns mode-imputed, nutrient categories ordinally encoded (low=0, medium=1, high=2), and nominal categorical variables (soil, season, and derived interaction features) one-hot encoded. Continuous inputs were

standardized using a fitted `StandardScaler` to match model training. Derived features capturing soil–season and soil–NPK interactions were included to improve representational capacity.

3.3 Model and training

We used a Random Forest classifier (`scikit-learn`) as the primary predictive model because of its robustness and interpretability for structured agronomic data. Hyperparameter tuning was performed using `GridSearchCV` with a 5-fold `StratifiedKFold` cross-validation strategy to identify the optimal model parameters. The final model configuration selected by the tuning process used 800 trees (`n_estimators=800`), maximum depth of 18, `min_samples_split=3`, `min_samples_leaf=2`, `class_weight='balanced_subsample'`, and `oob_score=True`. Model performance was evaluated using both 5-fold cross-validation accuracy and the Random Forest Out-Of-Bag (OOB) error estimate. This combination ensured a well-regularized and generalizable model suitable for multi-feature agricultural prediction.

3.4 Top-3 Crop Prediction Module

After preprocessing, the Random Forest model generates class probabilities for all crops. The system ranks these probabilities and selects the top three most suitable crops for the given soil and environmental conditions.

3.5 Profitability Estimation Module

Using the external `profit_data.csv` file, the system computes the estimated profit for each of the top three predicted crops. The crop with the highest profit value is identified as the most economically beneficial option for the farmer.

3.6 Irrigation Recommendation Module

A rule-based irrigation engine analyses the predicted crop, soil type, temperature, humidity, and rainfall to recommend an appropriate irrigation technique such as drip, sprinkler, or furrow/flood irrigation.

3.7 System Architecture

The system architecture of the Smart Crop Recommendation System is structured as a modular and sequential pipeline that integrates data processing, predictive modelling, and decision-support

mechanisms. The architecture begins with the acquisition of soil nutrient data, environmental variables, and supporting agricultural datasets. These inputs undergo an extensive preprocessing phase that includes noise removal, normalization, label encoding, and feature refinement to ensure that only clean and meaningful data are used for model training. A Stratified K-Fold cross-validation strategy is employed to maintain class distribution and prevent overfitting, allowing the model to generalize effectively across diverse agricultural conditions.

Once preprocessing is completed, the Random Forest classifier selected for its robustness, interpretability, and high predictive stability is trained using optimized hyperparameters. The trained model is then stored and integrated into the system's decision-making module. During real-time execution, the user provides soil parameters alongside the geographical location. The system uses this location to fetch real-time weather parameters such as temperature, humidity, and rainfall from the OpenWeatherMap API. These dynamically retrieved environmental inputs are combined with soil characteristics to form a comprehensive feature vector, which is passed to the trained model for inference.

The architecture includes a multi-stage prediction pipeline that generates the top three most suitable crops for the given conditions. In parallel, a profitability estimation module evaluates the economic feasibility of each predicted crop by analyzing market price and yield expectations from the economic dataset. This module computes the expected profit margins and identifies the most economically viable crop for the farmer. To support effective agricultural practices, an irrigation recommendation module further examines crop type, soil category, and prevailing weather conditions to suggest the most appropriate irrigation method, ensuring efficient water utilization.

Furthermore, the architecture follows a rigorous and standardized data flow that ensures reliability at each stage of processing. All user inputs pass through a parameter validation layer that checks for consistency, realistic range adherence, and completeness. Any anomalies are automatically corrected or flagged using predefined validation rules. The validated inputs are then transformed into the same feature representation used during the training phase, ensuring compatibility and eliminating discrepancies between training and inference pipelines. This unified transformation

workflow guarantees stable, unbiased, and reproducible predictions that align closely with agronomic standards.

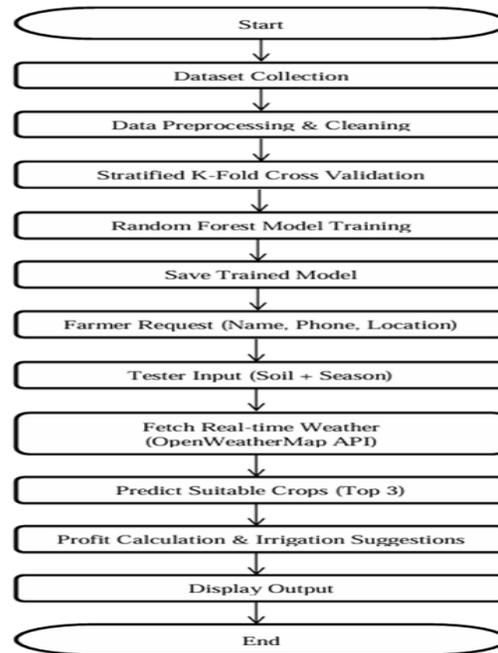


Fig. 1. System Design Flowchart

IV. IMPLEMENTATION

4.1 Workflow

When a farmer requests a crop suggestion, they enter their name, location, and phone number through the frontend interface. This request is stored in `farmer_requests.csv` for tracking. The tester dashboard displays all farmer requests categorized as pending or completed. The tester selects a pending request, visits the farmer's field, collects the soil sample, and determines the soil type, N-P-K nutrient categories (low, medium, or high), and pH value through laboratory analysis. The tester also records the current season to complete the soil assessment report. All verified readings, along with the farmer's details, are submitted as a structured report and stored in `linked-tester-data.csv`, and the status of the original request is updated.

The Flask backend coordinates the entire workflow by acting as the bridge between the user interface, datasets, and the Smart Crop Recommendation model. Once a tester report is submitted, Flask retrieves real-time weather data for the farmer's location using the

OpenWeatherMap API, combines it with the verified soil features, and applies the preprocessing pipeline before invoking the trained Random Forest classifier. The system generates the Top-3 crop recommendations, identifies the most profitable option using the profit_data.csv dataset, and computes suitable irrigation methods through irrigation_utils.py. These results are displayed to the farmer when they revisit the portal by entering their name, location and phone number. Farmers may additionally provide feedback, which is stored in feedback.csv and used to refine future system updates. All interactions between frontend and backend occur through structured Flask REST APIs, ensuring a smooth, reliable, and fully traceable end-to-end workflow. To maintain transparency and traceability, every stage of the workflow is logged and time-stamped, allowing administrators to audit request progress and verify data integrity. The system’s modular workflow ensures that each component user request handling, soil assessment, weather integration, model prediction, and feedback collection operates in a coordinated manner. This structured approach not only enhances reliability but also ensures that each recommendation is backed by verifiable data, making the entire process scientifically consistent and operationally efficient.

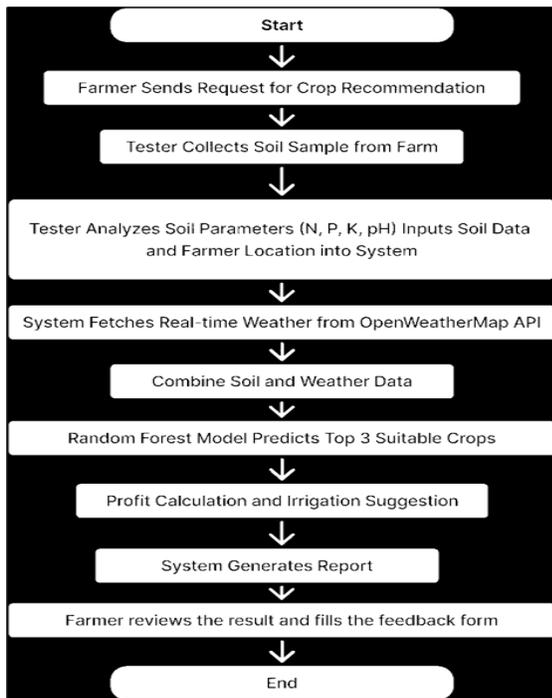


Fig. 2. Overall Workflow (Farmer and Tester interaction)

4.2 Technologies used

1. Python – Core language for backend development, preprocessing, and model execution.
2. NumPy & Pandas – For numerical computation, dataset handling, and CSV operations.
3. Scikit-learn (Random Forest Classifier) – Used for building, training, and deploying the crop prediction model.
4. Joblib – Supports model and preprocessing pipeline serialization and loading.
5. Flask – Backend framework providing REST APIs connecting the UI, model, datasets, and weather API.
6. Requests Library – Used to fetch real-time data from the OpenWeatherMap API.
7. OpenWeatherMap API – Provides temperature, rainfall, and humidity data for location-based predictions.
8. HTML, CSS, JavaScript – Used for building the farmer and tester user interfaces.
9. CSV-based Runtime Storage (farmerrequests.csv, linked-tester-data.csv, predictions.csv, feedback.csv) – Maintains logs for requests, tester inputs, predictions, and feedback.

V. RESULT AND DISSCUSSIONS

A. Performance Matrix

To analyze the effectiveness of the proposed Smart Crop Recommendation System using Machine Learning, five widely-used machine learning algorithms were evaluated: Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbour (KNN). The performance comparison was based on accuracy obtained from the trained model on the integrated agricultural dataset.

Table 1. Comparison of Classification Accuracies

Algorithm	Accuracy (%)
Random Forest	93
Naive Bayes	89
SVM	88
Decision Tree	85
KNN	83

From the performance matrix, it is evident that the Random Forest classifier achieves the highest accuracy among all evaluated models. Random Forest benefits from:

- Ensemble bagging strategy
- Robustness to noise and missing data
- Superior ability to handle multi-dimensional agricultural features
- Reduced risk of overfitting compared to standalone Decision Trees

Naive Bayes demonstrated comparatively strong performance, attributed to its efficiency in handling probabilistic feature dependencies. However, its assumption of feature independence limits its effectiveness for complex agricultural interactions.

Support Vector Machine (SVM) performing well on linearly separable patterns but facing limitations when dealing with the non-linear relationships present in soil nutrient and weather data.

Decision Tree and KNN performed lower than the ensemble-based Random Forest. Decision Tree is prone to overfitting, while KNN is sensitive to feature scaling and suffers in high-dimensional datasets such as agricultural multi-parameter inputs.

Thus, Random Forest is clearly the most reliable and accurate model for the crop recommendation task, making it the preferred choice for this system.

B. Evaluation Metrics Used

1. Cross-Validation Accuracy (5-Fold)

The model was trained and tested on five stratified folds to ensure class balance across splits. Accuracy for each fold is averaged to obtain the final cross-validation performance.

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

Where:

- TP – True Positive
- TN – True Negative
- FP – False Positive
- FN – False Negative

This ensures reliable estimation of model generalization capability.

2. Precision (Macro-Averaged)

Precision measures how many of the crops predicted as “correct” are actually correct.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Macro-average is used because the dataset contains multiple crop classes with varying frequencies.

3. Recall (Macro-Averaged)

Recall indicates how well the model identifies all relevant crop classes.

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

This ensures the model does not ignore minority crops.

4. F1-Score (Macro-Averaged)

F1 Score is a harmonic mean of Precision and Recall and is useful for imbalanced datasets.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

A high macro F1 confirms strong performance across all crop types.

5. Out-Of-Bag (OOB) Score

The Out-of-Bag score is used in this project to check both overfitting and the model’s ability to generalize to unseen data. Since each tree in the Random Forest is trained on a bootstrap sample, the samples excluded from that tree act as internal validation data. The model’s predictions on these OOB samples provide an unbiased estimate of performance. In this work, the OOB score is compared with the test accuracy, and the close agreement between the two indicates that the model is not overfitting and that the predictions are stable, consistent, and reliable for real-world soil and weather conditions.

6. Training-Set Classification Report

The code generates a complete classification report showing:

- Class-wise Precision
- Class-wise Recall
- Class-wise F1 Score
- Support (number of samples per class)

This gives insight into which crops are predicted strongly and which require improvement.

7. Confusion Matrix

The training set confusion matrix illustrates the count of:

- Correctly classified crops
- Misclassified crops
- Confusion between crops with similar profiles (e.g., cereals or pulses)

This allows deeper analysis of prediction patterns.

C. Graphical Representation of Model Accuracy

A graphical comparison of the accuracies of all five models is shown in Fig. 3 (bar graph). The visualization clearly highlights the superior performance of the Random Forest model.

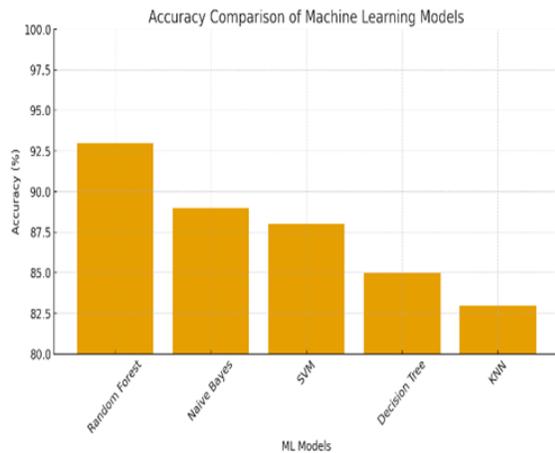


Fig. 3. Accuracy Comparison of ML Models

Graph Interpretation

- Random Forest stands at the top with 93% accuracy, indicating strong generalization and robustness.
- Naive Bayes and SVM show moderate but reliable performance.
- Decision Tree shows performance degradation due to overfitting.
- KNN performs the lowest, as expected, due to the high-dimensional nature of the dataset.

This graphical analysis supports the decision to adopt Random Forest as the core model for the Smart Crop Recommendation System.

VI. CONCLUSION

The experimental findings clearly demonstrate that the Random Forest classifier is the most robust and reliable model for crop recommendation among all the evaluated machine learning techniques. Achieving an accuracy of 93%, Random Forest consistently outperforms Naive Bayes, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbours (KNN), not only in predictive accuracy but also in model stability, handling of nonlinear relationships, and resilience to noisy input variables.

By integrating soil nutrient levels (N, P, K), environmental conditions, and weather parameters, the system provides a highly comprehensive analysis of factors influencing crop suitability. The incorporation of real-time weather API data further enhances the model's effectiveness, enabling dynamic, location-specific, and seasonally adaptive recommendations that adjust according to current climatic variations.

An important strength of the system is its ability to recommend the three most suitable crops and the most profitable crop among them, providing farmers with multiple choices based on expected yield, market profitability, and environmental fit. Each of these recommended crops is accompanied by accurate irrigation suggestions, ensuring that farmers receive actionable guidance not only on crop selection but also on optimal water management practices.

Furthermore, the system includes a farmer feedback feature, allowing users to share their experiences and results. This feedback loop enables continuous improvement of the model, making the system more adaptive, farmer-friendly, and aligned with real-world agricultural conditions.

Overall, the proposed Smart Crop Recommendation System is accurate, scalable, and practical for deployment in real agricultural environments. Its data-driven insights empower farmers to make informed decisions, reduce uncertainty in crop planning, and optimize agricultural productivity. With its strong performance, adaptability, and user-centric features, the system holds significant potential to support sustainable farming practices and contribute meaningfully to modern precision agriculture initiatives.

VII. FUTURE ENHANCEMENT

Although the proposed Smart Crop Recommendation System performs efficiently using machine learning and real-time weather data, several enhancements can significantly broaden its applicability, accuracy, and usability. The following future improvements are recommended:

1. Deployment of IoT-Enabled Real-Time Data Collection

The system can be upgraded by integrating IoT sensors for continuous monitoring of soil moisture, pH, nutrient levels, and environmental conditions. Real-time data acquisition will minimize manual data

entry, enhance prediction precision, and support automated decision-making for farmers.

2. Incorporation of Advanced Deep Learning Models
Beyond traditional machine learning algorithms, future versions may utilize deep learning techniques such as LSTM, GRU, or hybrid CNN-LSTM architectures. These models are capable of capturing complex temporal and spatial patterns, enabling more accurate seasonal forecasting and long-term crop planning.

3. Utilization of Satellite and Remote Sensing Technologies

Integrating satellite-based indices like NDVI, EVI, and thermal imaging can help assess large-scale vegetation health, soil moisture dynamics, and disease-affected regions. Remote sensing data will strengthen prediction capabilities and enable region-wide agricultural monitoring.

4. Development of a Pest and Disease Early-Warning System

A dedicated module can be added to predict potential pest infestations and crop diseases using climate patterns, image-based detection, and historical outbreak data. This feature will help farmers adopt preventive measures and reduce crop losses.

5. Multi-Language, Voice-Assisted User Interface

To improve accessibility for farmers across different regions, the system can include a voice-enabled interface supporting multiple Indian languages. This enhancement will make the platform user-friendly, especially for farmers with limited literacy or technical exposure.

6. Drone-Assisted Field Diagnostics

Integrating drone-based monitoring equipped with multispectral and thermal cameras can automate large-area field assessment. Drones can identify nutrient deficiencies, water stress, and disease outbreaks faster than manual inspections, enabling timely interventions.

REFERENCES

[1] Ayesha Siddiq, Shreya G., Shambhavi S. V., Umme Hani, and Varshaa S. K., "Agri vision: AI-Enhanced Yield Prediction and Smart Crop Recommendation," International Conference on

Advancements in Smart, Secure and Intelligent Computing (ASSIC), 2024, IJRASET.

[2] Sangeetha Allam, Bollimuntha Manjusha, Asha Vuyyuru, P. Veeranjanyulu, K. Pushpa Rani, and Sushruta Mishra, "Agriculture Crop Recommendation System using Machine Learning," International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC), 2024, IEEE.

[3] Shwetha A. N. and Prabodh C. P., "Machine Learning-based Smart Crop Recommendation System," 4th International Conference on Intelligent Technologies (CONIT), 2024, IEEE.

[4] Ramachandra A. C. and G. V. Ankitha, Idupulapati Divya, Parimi Vandana, and H. S. Jagadeesh, "Crop Recommendation using Machine Learning," International Conference on Data Science and Network Security (ICDSNS), 2023, IEEE.

[5] J. Ashok Kumar, N. Parimala, and R. Pitchai, "Crop Selection and Yield Prediction using Machine Learning Algorithms," Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), 2023, IEEE.

[6] Shabari Shedthi B, Anusha, Anisha Shetty, Rakshitha RShetty, B. A Divyashree Alva, and Aishwarya D Shetty, "Machine Learning Techniques in Crop Recommendation based on soil and Crop Yield Prediction System – Review," International Conference on Artificial Intelligence and Data Engineering (AIDE), 2022, IEEE.