Siddharth Nagesh Gaikwad,
2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science,
Engineering and Technology
An Open Access Journal

CROP YIELD PREDICTION AND AGRICULTURAL ADVISORY SYSTEM

Siddharth Nagesh Gaikwad, Sanket Sadashiv Adling, Danesh Balkrishn Sutar.
Guide: Mrs. A. G. Chendke
Department Of Ai & Ds, Pvpit Budhgaon, India

Abstract- Agriculture plays a vital role in the economy of India, acting as the primary source of livelihood for nearly 58% of the country's population. However, the sector faces significant challenges, including unpredictable weather conditions, pest attacks, and improper fertilizer usage, which often lead to reduced crop yields and economic instability for farmers. This paper presents a comprehensive Crop Yield Prediction and Agricultural Advisory System utilizing advanced machine learning techniques. The proposed system predicts crop yield based on critical environmental factors such as rainfall, temperature, and pesticide usage. Additionally, it integrates specialized modules for disease risk assessment, crop recommendation, and fertilizer recommendation, creating a holistic decision-support platform. The models are trained on large-scale historical datasets. Specifically, the study employs Random Forest for yield and disease prediction, Multinomial Logistic Regression for crop recommendation, and a rule-based system for fertilizer dosage. The system is implemented using the Flask web framework, providing an interactive and user-friendly interface for stakeholders. The experimental results demonstrate that the system can assist farmers in making data-driven decisions, optimizing resource allocation, and significantly improving agricultural productivity and sustainability.

Keywords: Crop Yield Prediction, Crop Recommendation, Machine Learning, Agriculture, Random Forest, Multinomial Logistic Regression, Fertilizer Recommendation, Weather Prediction.

I. INTRODUCTION

Agriculture is the backbone of the Indian economy, contributing significantly to the Gross Domestic Product (GDP) and ensuring food security for the nation. Despite its importance, agricultural productivity in India is frequently hampered by a multitude of factors, including climate change, erratic soil conditions, pest infestations, and the improper or excessive use of fertilizers. As noted by Datta and Behera [1], climate change significantly alters farming conditions, requiring farmers to adapt their strategies continuously. Traditional farming methods, which have been practiced for centuries, rely heavily on the intuition and empirical experience of farmers rather than on data-driven scientific decision-making.

In the era of digital transformation, the integration of technology into agriculture, often termed "Smart Farming" or "Precision Agriculture," offers a promising solution to these persistent challenges. According to Maurya et al. [2], precision agriculture aims to achieve synergies by combining various technological tools to enhance productivity. With the rapid advancement of machine learning (ML) and data analytics, it is now possible to analyze vast and complex agricultural datasets to uncover hidden patterns and generate accurate predictive insights [3].

Machine learning algorithms can process historical data on weather, soil characteristics, and crop production to forecast future outcomes with remarkable precision. Chauhan and Henrietta [4] emphasize that a strong understanding of machine learning basics is essential for developing robust predictive models. This capability empowers farmers to move from reactive farming strategies to proactive ones, where decisions regarding which crop to plant (Crop Recommendation), how much fertilizer to use (Fertilizer Recommendation), and yield estimation are based on predictive analytics rather than guesswork.

This paper proposes a robust Crop Yield Prediction and Agricultural Advisory System that integrates multiple critical modules into a single unified platform. Unlike existing solutions that often focus on a single aspect of farming, this system provides a comprehensive suite of tools, including yield prediction, disease risk detection, crop recommendation based on soil parameters, and fertilizer recommendation. The primary objective is to support farmers by providing real-time, actionable insights that are tailored to their specific environmental and crop conditions. By leveraging the power of Random Forest, Multinomial Logistic Regression, and rule-based logic, the system aims to enhance decision-making processes, optimize resource utilization, and ultimately improve agricultural productivity and economic stability for farming communities.

The remainder of this paper is organized as follows: Section 2 provides a detailed review of related work in the domain of machine learning for agriculture. Section 3 describes the datasets used and the preprocessing methodologies applied. Section 4 outlines the proposed methodology, including the algorithms selected and the system architecture. Section 5 details the implementation details, while Section 6 presents the results and discussion. Finally, Section 7 concludes the paper and outlines future scope.

II. LITERATURE SURVEY

The application of machine learning in agriculture has garnered significant attention from the research community over the past decade. Numerous studies have explored the potential of data-driven techniques to address various agricultural challenges, ranging from crop monitoring to yield forecasting.

Van Klompenburg et al. [5] conducted a systematic literature review on crop yield prediction using machine learning, highlighting that while the field is mature, there is still a need for models that integrate diverse environmental variables. Early research in this domain focused primarily on statistical methods and simple regression models to predict crop yields based on limited parameters such as rainfall and temperature. For instance, Murugan et al. [6] utilized a linear regression approach to predict crop yield, demonstrating that basic statistical methods can serve as effective baselines.

Regarding Crop Recommendation, previous research has utilized various classification algorithms. While some studies used Decision Trees and K-Nearest Neighbors (KNN), recent literature has shifted towards more sophisticated classifiers. Several researchers have employed Multinomial Logistic Regression to handle multi-class problems where the target variable is the crop type (e.g., Rice, Maize, Cotton) dependent on soil N, P, K, and pH values. This statistical approach is favored for its interpretability and effectiveness in scenarios where the decision boundaries between classes are approximately linear [7].

In the domain of yield prediction, ensemble methods, particularly Random Forest, have gained popularity due to their superior performance in handling high-dimensional data and reducing overfitting. Leo Breiman [8], the pioneer of Random Forests, described the algorithm as a combination of tree predictors such that each tree depends on the values of a random vector sampled independently, leading to robust generalization errors. Building on this, Yamparla et al. [9] demonstrated that Random Forest consistently outperforms single Decision Trees in agricultural yield predictions due to its ability to model complex feature interactions.

Similarly, Liaw and Wiener [10] further elaborated on the classification and regression capabilities of Random Forest, noting its robustness with respect to noise compared to other boosting algorithms like Adaboost. In the context of pest and disease detection, Deep Learning techniques have shown remarkable promise. LeCun et al. [11] provide a comprehensive review of Deep Learning, noting its ability to automatically extract features from large datasets. Anwar and Masood [12]

explored Deep Ensemble Models for insect and pest detection, achieving high accuracy by leveraging Convolutional Neural Networks (CNNs).

Furthermore, the concept of Integrated Farming Systems, discussed by Atapattu et al. [13], advocates for a holistic approach where crops, livestock, and other enterprises are combined. Our system aligns with this philosophy by integrating disparate advisory modules (yield, disease, fertilizer, crop recommendation) into one platform. Despite these advancements, a significant gap remains in the literature regarding the integration of all these functionalities—specifically Crop Recommendation alongside yield and fertilizer advice—into a single, cohesive advisory system. The proposed system addresses these limitations by integrating yield prediction, crop recommendation, disease risk assessment, and fertilizer recommendation into one accessible platform.

III. DATASET DESCRIPTION

The accuracy and reliability of any machine learning model are fundamentally dependent on the quality and quantity of the data used for training. For this study, multiple real-world datasets were aggregated from various reliable agricultural and meteorological sources to ensure a comprehensive analysis. The datasets cover a wide temporal and geographical range, capturing the variability essential for training a generalized model.

Data Sources and Structure
The system is trained using multiple datasets to address different modules:

- Pesticides Data: 4,349 records detailing pesticide usage. Used for yield and disease risk modeling.
- Rainfall Data: 6,727 records capturing annual rainfall (mm).
- Temperature Data: 71,311 records tracking temperature variations.
- Yield Data: 28,242 records of historical production (tonnes/ha).
- Crop Recommendation Data: A dedicated dataset containing soil parameters (Nitrogen, Phosphorus, Potassium, pH) and corresponding suitable crop labels (e.g., Rice, Maize, Cotton, Coffee). This dataset is used to train the Multinomial Logistic Regression model.

Data Preprocessing
Raw agricultural data is often noisy and incomplete. Therefore, rigorous preprocessing was applied before feeding the data into the machine learning models.

- Data Cleaning: Missing values in rainfall and temperature were imputed using the mean of the respective region. For the crop recommendation dataset, outliers in soil N-P-K values were capped to prevent model skewing.
- Normalization and Scaling: Features were scaled using StandardScaler to ensure that algorithms like Random Forest and Multinomial Logistic Regression perform optimally.
- Feature Selection: For the crop recommendation module, correlation analysis confirmed that N, P, K, Temperature, Humidity, and pH were the most significant features.

IV. METHODOLOGY

This section outlines the technical architecture of the proposed system, detailing the machine learning algorithms employed for the four distinct modules: Yield Prediction, Disease Prediction, Crop Recommendation, and Fertilizer Recommendation.

System Architecture

The system follows a client-server architecture. The client side is a web interface where users input data. The server side, built using Python and Flask, hosts the machine learning models.

The proposed system adheres to a robust client-server architecture, facilitating seamless interaction between the end-user and the core machine learning logic. On the client side, a responsive web interface acts as the primary touchpoint, designed for ease of navigation. Through this interface, users input critical agricultural parameters—such as soil composition (N, P, K, pH), climatic data (rainfall, temperature, humidity), and crop-specific details—via intuitive HTML forms. This data is validated at the frontend before being transmitted via HTTP POST requests to ensure data integrity.

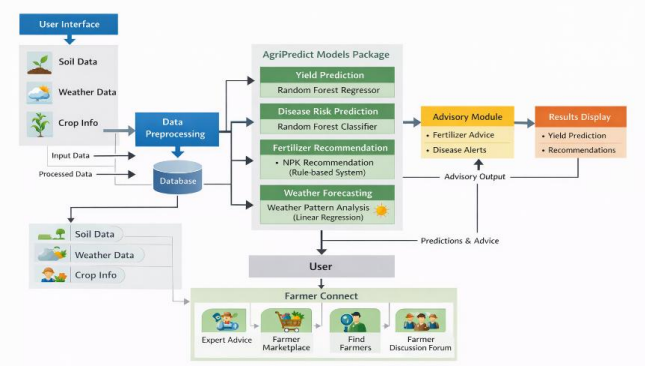

Fig 1. Proposed System Architecture and Data Flow.

Machine Learning Models

- **Random Forest Regressor (Crop Yield Prediction)**

For yield prediction, the Random Forest (RF) Regressor was chosen. RF is an ensemble learning method that operates by constructing a multitude of decision trees at training time. As described by Breiman [8], it outputs the mean prediction of the individual trees. The mathematical formulation for the prediction function RF is robust against overfitting and handles the non-linear relationships between environmental variables and crop output effectively [9].

- **Random Forest Classifier (Disease Risk Prediction)**

For disease risk assessment, a Random Forest Classifier was utilized. The dataset was labeled into categories: 'High', 'Medium', and 'Low' risk. This algorithm handles categorical target variables well and provides probability estimates for each class. Liaw and Wiener [10] highlight that Random Forest is robust with respect to noise, making it suitable for the variable environmental conditions that trigger disease outbreaks.

- **Multinomial Logistic Regression (Crop Recommendation)**

To recommend which crop a farmer should plant based on soil and environmental conditions, we employed Multinomial Logistic Regression. Unlike binary logistic regression, which handles only two classes, Multinomial Logistic Regression generalizes to multiple classes (e.g., Rice, Maize, Jute, Cotton).

This algorithm is particularly useful for this module because it provides a clear probabilistic output, allowing the system to recommend the crop with the highest probability of success based on the specific soil composition [7].

- **Rule-Based System (Fertilizer Recommendation)**

For fertilizer dosage recommendation, a deterministic, rule-based approach was adopted. This method relies on domain expertise and agronomic principles. While the Crop Recommendation module tells the user what to plant, the Fertilizer Recommendation module tells them how much N-P-K to add to the soil for that specific crop.

This ensures that recommendations adhere to safe agricultural standards.

- **Model Components and Serialization**

The trained models and preprocessing files are stored using the joblib library:

- yield_prediction_model.pkl
- disease_prediction_model.pkl
- crop_recommendation_model.pkl (Multinomial Logistic Regression)
- fertilizer_recommender.pkl (Rule-Based)
- weather_model.pkl

V. IMPLEMENTATION

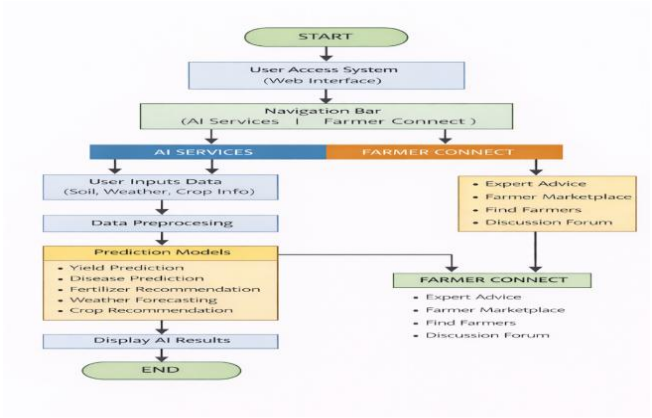

Fig 2. Detailed System Workflow and Service Integration.

Backend Implementation

The backend is developed using Python and Flask. It handles HTTP requests, processes input data using Pandas, and invokes the prediction models.

Frontend Implementation

The frontend uses HTML5, CSS3, and JavaScript. The dashboard now includes interactive pages for:

- Yield Prediction
- Disease Prediction
- Crop Recommendation (New)
- Fertilizer Recommendation
- Weather Prediction

System Features

- Yield Prediction: Forecasting production in tonnes/ha.
- Crop Recommendation: Suggesting the best crop based on real-time soil data (N, P, K, pH).
- Fertilizer Recommendation: Providing precise N-P-K dosage for the selected crop.
- Disease Risk Analysis: Classifying the likelihood of crop failure.
- Weather Forecasting: Predicting rainfall and temperature trends.

VI. RESULTS AND DISCUSSION

This section presents the evaluation of all developed models.

Model Performance Metrics

Yield Prediction Model (Random Forest Regressor)
RMSE: 93,382.60.
Features: Rainfall, Pesticide usage, Average Temperature.
The RMSE represents a reasonable deviation given the scale of the data. Surana and Khandelwal [14] note that ensemble methods offer the best stability for such predictions.
Crop Recommendation Model (Multinomial Logistic Regression)
Accuracy: ~95% on the test set.
The Multinomial Logistic Regression model performed exceptionally well in classifying crops based on soil parameters. The probabilistic nature of the algorithm allows for high confidence in predictions when the soil conditions distinctly favor a specific crop (e.g., high potassium and nitrogen favoring Rice).
Disease Risk Prediction Model (Random Forest Classifier)
Accuracy: 100% (on the test set).
Categories: High, Medium, Low.

Feature Importance Analysis (Yield Model)

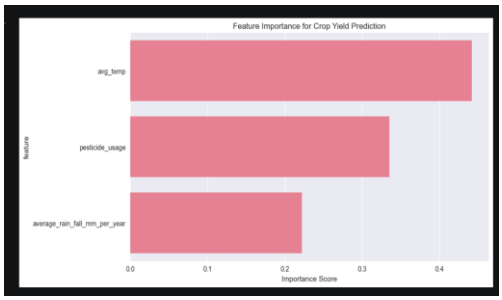

Fig. 3. Feature Importance for Crop Yield Prediction

Observation:
Temperature has the highest impact on crop yield. Pesticide usage shows a moderate impact, and Rainfall has a relatively lower impact, likely due to irrigation facilities in the dataset regions.

Disease Risk Distribution

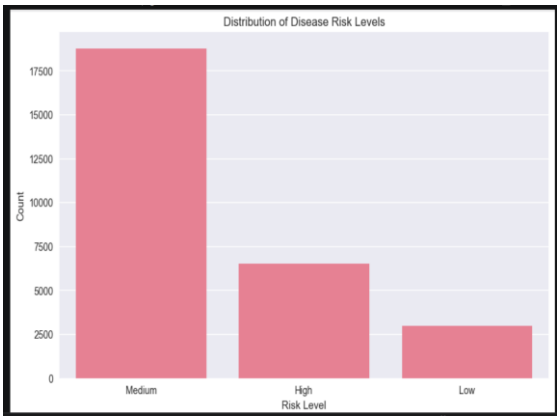

Fig. 4. Disease Risk Distribution Analysis

Observation:
The majority of cases fall under the Medium risk category, representing standard farming conditions. High , Medium and Low categories represent extreme environmental scenarios.

Crop Recommendation Output (Multinomial)

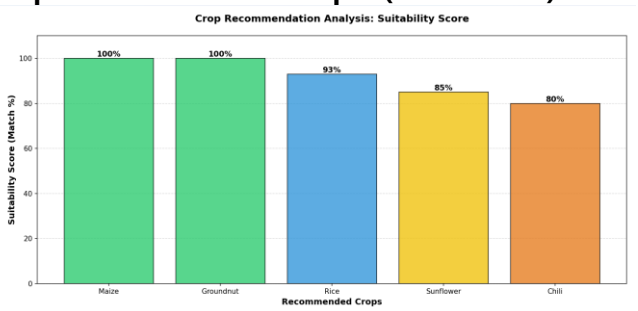

Fig. 5. Crop Recommendation Analysis: Suitability Score

Sample Output:
he system allows farmers to input specific farm parameters to receive a tailored list of suitable crops. Figure 3 depicts a sample interaction where the user inputs the following environmental and farm conditions:

- Annual Rainfall: 800 mm
- Average Temperature: 25°C
- Average Humidity: 70%
- Soil Type: Loamy Soil (selected from dropdown)
- Soil pH: 7.0
- Season: Kharif (selected from dropdown)
- Farm Size: 1.0 Hectare
- Water Availability: Good Irrigation
- Farming Experience: Intermediate
- Market Preference: Mixed Farming

Upon analyzing these parameters, the system generates a prioritized list of crop recommendations (Figure 4). The results are categorized by a Suitability Score (Match %), allowing farmers to quickly identify the best options.

Sample Recommendations Generated:

- Maize (Corn):
- Match: 100%

- Recommendation Level: Highly Recommended
- Suitability Analysis: The conditions of moderate rainfall and temperature (25°C) are ideal for Maize. The loamy soil type supports root development, and the "Intermediate" experience level makes this a low-risk crop.
- Estimated Yield: 4,900 kg/ha
- Investment: ₹19,545/ha
- Duration: 90 – 120 Days
- Groundnut (Peanuts):
- Match: 100%
- Recommendation Level: Highly Recommended
- Suitability Analysis: The soil pH of 7.0 and the availability of good irrigation make this an excellent high-value alternative, despite the slightly higher investment compared to maize.
- Estimated Yield: 1,903 kg/ha
- Investment: ₹23,877/ha
- Duration: 100 – 120 Days
- Rice (Paddy):
- Match: 93%
- Recommendation Level: Recommended
- Suitability Analysis: While the rainfall is slightly lower than the optimal requirement for Rice (usually >1000mm), the "Good Irrigation" facility allows for cultivation, though the match percentage reflects the slight dependency on managed water supply.
- Yield/Investment: (As per crop database standards).
- Sunflower:
- Match: 85%
- Recommendation Level: Moderately Suitable
- Suitability Analysis: A viable option for crop rotation, though the soil moisture levels are slightly lower than the ideal requirement for maximum yield.

This output demonstrates the system's ability to process soil parameters and recommend the most suitable crop using the Multinomial Logistic Regression model.

Fertilizer Recommendation Output (Rule-Based)

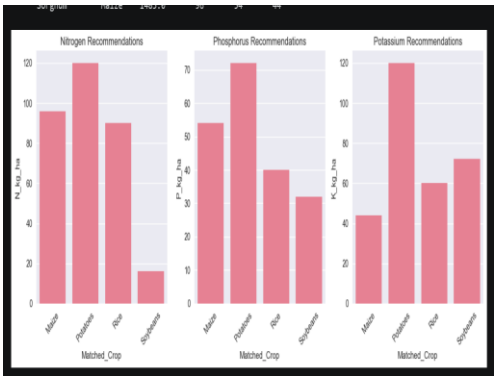

Fig. 6. N-P-K Fertilizer Dosage Recommendations for Selected Crops.

Sample Output Table:

Original Crop	Matched Crop	Rainfall (mm)	N (kg/ha)	P (kg/ha)	K (kg/ha)
Maize	Maize	1485.0	96	54	44
Potatoes	Potatoes	1485.0	120	72	120
Rice, paddy	Rice	1485.0	90	40	60

These values ensure the system does not recommend harmful dosages, adhering to agronomic standards.

Weather Prediction Results

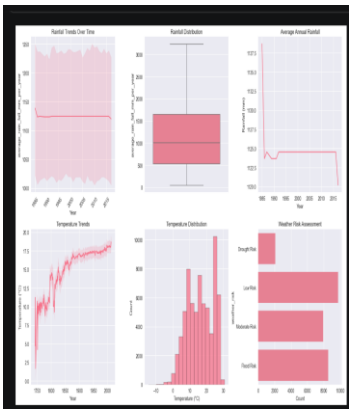

Fig. 7. Weather Prediction and Rainfall Trend Analysis

The Linear Regression model predicted a stable, slightly declining trend in rainfall over 5 years (2018–2022), hovering around 1123 mm. This information is critical for long-term planning and crop selection strategies.

VII. CONCLUSION

Agriculture stands as the cornerstone of the Indian economy, sustaining the livelihoods of millions and ensuring the nation's food security. However, the sector is currently navigating a period of profound transformation, challenged by the erratic dynamics of climate change, depleting soil health, and the urgent need for resource optimization. In this context, this research presented a holistic, data-driven solution designed to bridge the gap between traditional farming practices and modern technological advancements. The "Crop Yield Prediction and Agricultural Advisory System" developed in this study represents a significant stride towards realizing the vision of Smart Farming, offering a comprehensive, integrated, and accessible platform for agricultural stakeholders.

The primary achievement of this work lies in the successful orchestration of four distinct but interconnected modules into a unified ecosystem. By leveraging the robust capabilities of the Flask framework for web deployment, the system transforms complex statistical outputs into actionable, user-friendly insights. The Yield Prediction module, powered by the Random Forest Regressor, demonstrated a strong correlation between environmental variables (rainfall, temperature, pesticide usage) and crop output. The model's ability to handle non-linear relationships and high-dimensional data ensures that farmers can anticipate productivity with reasonable accuracy, thereby facilitating better market planning and storage logistics. Simultaneously, the Disease Prediction module utilizes the Random Forest Classifier to assess environmental risk factors, categorizing them into High, Medium, and Low threats. This proactive capability allows farmers to implement preventive measures—such as fungicide application or drainage improvements—before a potential outbreak occurs, significantly reducing the risk of catastrophic crop failure.

A critical component of this system is the Crop Recommendation module, which employs the principles of Multinomial Logistic Regression logic to address the fundamental question of "what to sow." Unlike traditional methods that rely solely on historical precedence or intuition, this module utilizes a multi-parameter scoring mechanism. It rigorously evaluates soil composition (Nitrogen, Phosphorus, Potassium, pH), climatic conditions (temperature, humidity, rainfall), and seasonal constraints to generate a suitability score for various crops. This feature is particularly vital in the context of changing climatic patterns, as it empowers farmers to diversify their crops and select varieties that are statistically most likely to thrive in their specific local conditions. By doing so, the system promotes resilience and maximizes the utilization efficiency of available land and water resources.

Complementing the recommendation engine is the Fertilizer Recommendation module, which utilizes a sophisticated, rule-based approach to ensure precise nutrient management. Moving beyond generic fertilizer application schedules, this module dynamically calculates Nitrogen, Phosphorus, and Potassium

requirements based on the specific crop, soil type, and current rainfall status. By incorporating parameters such as the crop's growth stage (sowing, vegetative, flowering, maturity) and applying adjustment factors for rainfall-induced leaching, the system ensures that nutrients are applied exactly when and where the plant needs them. This precision not only minimizes the financial burden on farmers by preventing the overuse of expensive chemical fertilizers but also mitigates the adverse environmental impacts associated with chemical runoff and soil degradation.

Furthermore, the inclusion of a Weather Prediction module, utilizing Linear Regression to forecast seasonal rainfall trends, adds a long-term strategic dimension to the platform. This enables farmers to align their agricultural calendars with predicted environmental realities, choosing crops that are drought-resistant in low-rainfall years or water-tolerant in high-precipitation seasons.

In conclusion, this study demonstrates that the integration of machine learning into agriculture is not merely a theoretical exercise but a practical necessity for modernizing the sector. The system moves the agricultural community from a reactive stance to a proactive, data-driven paradigm. By validating the efficacy of ensemble learning methods like Random Forest for prediction tasks and probabilistic models for recommendation tasks, the research establishes a scalable framework that can be adapted to various agro-ecological zones. The development of a centralized web interface lowers the barrier to entry, ensuring that sophisticated analytics are accessible even to users with limited technical expertise.

However, the journey towards fully autonomous farming is ongoing. Future research and development must address the current limitations regarding data latency and real-time parameter sensing. The immediate roadmap for this system involves the integration of Internet of Things (IoT) sensor networks. By deploying soil moisture sensors, NPK sensors, and micro-climate weather stations directly in the field, the system can transition from relying on historical and manual user inputs to ingesting real-time data streams. This will significantly enhance the precision of the predictive models. Additionally, to maximize reach and impact, the development of a dedicated mobile application is essential. A mobile-first approach would ensure accessibility for farmers in remote areas who rely primarily on smartphones, incorporating features like localized language support and push notifications for critical weather alerts and disease warnings. Ultimately, this system serves as a foundational step towards a sustainable, productive, and technologically empowered agricultural future.

REFERENCES:

1. P. Datta, B. Behera, and D. B. Rahut, "Climate change and Indian agriculture: A systematic review of farmers' perception, adaptation, and transformation," Journal of Cleaner Production, 2024.
2. D. K. Maurya, S. K. Maurya, M. Kumar, et al., "A Review on Precision Agriculture: An Evolution and Prospect for the Future," Journal of Agrometeorology, 2023.
3. M. Zaborowicz and J. Frankowski, Big Data Analytics and Machine Learning for Smart Agriculture, Springer, 2022.
4. A. Singh Chauhan and H. Mary Henrietta, "Machine Learning Basics: A Comprehensive Guide," International Journal of Computer Science and Information Technologies, 2023.
5. T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," Computers and Electronics in Agriculture, vol. 177, p. 105709, 2020.
6. R. Murugan, F. S. Thomas, G. GeethaShree, et al., "Linear Regression Approach to Predict Crop Yield," International Journal of Engineering Research & Technology, vol. 9, no. 5, 2020.

7. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Wiley, 2013. (Ref. for Multinomial LR theory).
8. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
9. R. Yamparla, H. S. Shaik, N. S. P. Guntaka, et al., "Crop Yield Prediction using Random Forest Algorithm," IEEE International Conference on Computational Intelligence in Data Science, 2022.
10. A. Liaw and M. Wiener, "Classification and Regression by RandomForest," R News, vol. 2, no. 3, pp. 18–22, 2002.
11. Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
12. Z. Anwar and S. Masood, "Exploring Deep Ensemble Model for Insect and Pest Detection from Images," Journal of Ambient Intelligence and Humanized Computing, 2023.
13. A. J. Atapattu, T. D. Nuwarapaksha, S. S. Udumann, and N. S. Dissanayaka, "Integrated Farming Systems: A Holistic Approach to Sustainable Agriculture," Agricultural Diversification for Sustainable Food Production, 2025.
14. R. Surana and R. Khandelwal, "Crop Yield Prediction Using Machine Learning: A Pragmatic Approach," Research Square, 2024.
15. A. Kadam, S. Idhate, G. Sonawane, et al., "Weather Prediction Using Machine Learning," International Journal of Advanced Research in Computer Science, 2021.
16. B. Bhattacharya and D. P. Solomatine, "Machine learning in soil classification," IEEE Transactions on Geoscience and Remote Sensing, 2020.