IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15530

A Review on Machine Learning Techniques for Crop Yield Prediction

A Renukamma[1], Arathi C G[2], Aruna B[3], K Ananya[4], Dr. Muhibur Rahman T.R[5]

6th Sem B.E.(CS&E), Ballari Institute of Technology Management (BITM), Ballari, Karnataka - 583104, India[1-4]

Associate Professor, Department of Computer Science and Engineering,

Ballari Institute of Technology and Management (BITM), Ballari, Karnataka - 583104, India[5]

Abstract: Crop yield prediction plays a crucial role in ensuring food security, efficient resource management, and sustainable agricultural planning. With the rapid advancement of artificial intelligence, machine learning (ML) techniques have emerged as powerful tools for predicting crop productivity using diverse datasets such as weather conditions, soil characteristics, and remote sensing data. This paper presents a comprehensive review of machine learning techniques applied to crop yield prediction. It analyzes commonly used algorithms, including linear regression, decision trees, random forests, support vector machines, and deep learning models such as artificial neural networks and convolutional neural networks. Studies show that environmental factors like temperature, rainfall, and soil type are the most significant features influencing prediction accuracy.

Keywords: Crop Yield Prediction, Machine Learning, Precision Agriculture, Remote Sensing, Predictive Modeling

I. INTRODUCTION

Agriculture is a fundamental sector that supports global food security and economic stability. Accurate crop yield prediction is essential for effective decision-making by farmers, policymakers, and agricultural stakeholders. It helps in planning resource allocation, managing supply chains, and minimizing risks associated with climate variability and market fluctuations. Traditional methods of yield estimation rely heavily on historical data and manual observations, which are often time consuming and less accurate.

In recent years, the advancement of machine learning (ML) techniques has significantly transformed agricultural practices. Machine learning enables the analysis of large and complex datasets, including weather conditions, soil properties, crop characteristics, and remote sensing data, to generate more precise yield predictions. Various algorithms such as linear regression, decision trees, support vector machines, random forests, and deep learning models have been widely applied in this domain, offering improved accuracy and efficiency.

This paper presents a comprehensive review of machine learning techniques used for crop yield prediction. It aims to analyze different models, compare their performance, and identify the most influential factors affecting prediction accuracy. Additionally, the paper discusses the challenges associated with data quality, model generalization, and scalability. By examining recent developments and trends, this review provides valuable insights into the potential of machine learning in enhancing agricultural productivity and supporting sustainable farming practices.

II. BACKGROUND AND MOTIVATION

The increasing global demand for food, combined with climate change and limited agricultural resources, has made accurate crop yield prediction more important than ever. Variability in environmental conditions such as rainfall, temperature, and soil quality significantly affects crop productivity. Machine learning techniques provide data-driven solutions that can handle these complexities. The motivation of this study is to explore and evaluate different machine learning approaches to improve agricultural decision-making and productivity.

III. DATA SOURCES AND FEATURES USED

Crop yield prediction using machine learning relies on multi-source heterogeneous data, as crop growth is influenced by environmental, biological, and management factors. Recent journal studies emphasize that integrating multiple data sources significantly improves prediction accuracy compared to using a single dataset.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15530

A. Weather Data

Weather data plays a fundamental role in crop yield prediction, as climatic conditions directly influence plant growth and development. Variables such as temperature, rainfall, humidity, solar radiation, and wind speed are commonly used in predictive models. Temperature affects physiological processes like photosynthesis and respiration, while rainfall determines water availability for crops. Irregular weather patterns, such as droughts or excessive rainfall, can significantly reduce yield. Machine learning models are particularly effective in capturing the nonlinear relationships between weather variables and crop productivity.

B. Soil Data

Soil characteristics are another critical component in determining crop yield, as they directly affect nutrient availability and water retention capacity. Important soil features include soil type, pH level, organic matter content, and essential nutrients such as nitrogen, phosphorus, and potassium. Soil moisture also plays a vital role in supporting plant growth, especially in regions dependent on rainfall. Machine learning models use these parameters to assess soil fertility and its impact on crop productivity. Research indicates that variations in soil properties across regions can lead to significant differences in yield, making it essential to include localized soil data in prediction models. C. Management Practices Agricultural management strategies, such as irrigation scheduling, fertilization regimes, and planting density, serve as key modifiers of the final harvest.

C. Remote Sensing Data

Remote sensing data has become increasingly important in modern agricultural analysis due to its ability to provide large-scale and real-time information about crop conditions. Satellite imagery and vegetation indices such as the Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) are widely used to monitor crop health, biomass, and growth stages. These indices reflect the greenness and photosynthetic activity of plants, which are closely related to yield potential. Remote sensing allows continuous monitoring of agricultural fields without the need for physical inspection, making it highly efficient for large-scale farming. Studies have demonstrated that combining remote sensing data with machine learning techniques significantly improves prediction accuracy.

D. Crop Management Data

Crop management practices represent the human factors influencing agricultural productivity. These include fertilizer application, irrigation methods, crop variety, planting schedules, and pest control measures. Even under similar environmental conditions, differences in management practices can lead to variations in yield. Machine learning models incorporate these variables to capture the impact of human decision-making on crop performance. For instance, optimized fertilizer usage can enhance soil fertility, while efficient irrigation techniques can improve water utilization. Research highlights that integrating management data with environmental and remote sensing data provides a more comprehensive understanding of crop growth dynamics.

E. Integration of Multi-Source Data

The integration of multiple data sources is a key factor in improving the performance of crop yield prediction models. Weather data provides information about climatic conditions, soil data reflects nutrient availability, remote sensing captures real-time crop health, and management data accounts for human interventions. Combining these datasets allows machine learning models to learn complex interactions among various factors affecting crop growth. Recent studies have shown that multi-source data integration leads to more robust and accurate predictions compared to models that rely on a single data type. This approach also enables scalable solutions that can be applied across different regions and crop types.

IV. MACHINE LEARNING TECHNIQUES FOR CROP YIELD PREDICTION

A. Regression Models

Regression models are among the earliest and most widely used techniques for crop yield prediction due to their simplicity and interpretability. Linear regression assumes a direct relationship between input variables such as temperature, rainfall, and soil properties and the crop yield. It is useful when the relationship between variables is approximately linear and the dataset is relatively small. However, real-world agricultural data often exhibits nonlinear behavior, which limits the performance of simple linear models. To address this, polynomial regression extends linear regression by introducing higher-degree terms, allowing the model to capture curved relationships between variables.

B. Decision Tree-Based Models

Decision tree-based models are widely used in crop yield prediction due to their ability to handle nonlinear relationships and complex datasets. A decision tree works by splitting the dataset into smaller subsets based on feature values, forming

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15530

a tree-like structure of decisions. These models are highly interpretable, as they clearly show how input variables influence predictions. Random Forest, an extension of decision trees, improves prediction accuracy by combining multiple decision trees and averaging their outputs. This ensemble approach reduces overfitting and increases robustness. Decision tree-based models can effectively handle both numerical and categorical data, making them suitable for agricultural datasets that include diverse features such as soil type and crop variety.

C. Support Vector Machines (SVM)

Support Vector Machines (SVM) are powerful supervised learning algorithms used for both classification and regression tasks. In crop yield prediction, Support Vector Regression (SVR) is commonly applied. SVM works by finding an optimal hyperplane that best fits the data while minimizing prediction error. One of the key advantages of SVM is its effectiveness in handling high-dimensional data and small sample sizes. It uses kernel functions, such as linear, polynomial, and radial basis function (RBF), to transform data into higher-dimensional spaces where complex relationships can be captured. This makes SVM particularly suitable for agricultural datasets that contain nonlinear patterns. However, SVM models can be computationally expensive and require careful tuning of parameters, which may limit their scalability for large datasets.

D. Neural Networks and Deep Learning

Neural networks and deep learning models have gained significant attention in recent years due to their ability to capture highly complex and nonlinear relationships in data. Artificial Neural Networks (ANN) consist of interconnected layers of neurons that process input features and learn patterns through training. These models are capable of modeling intricate relationships between environmental factors and crop yield. Convolutional Neural Networks (CNN), a type of deep learning model, are particularly effective when working with image-based data such as satellite imagery. CNNs can automatically extract features from remote sensing data, making them valuable for large-scale agricultural monitoring. Deep learning models generally provide higher accuracy compared to traditional methods, especially when large datasets are available.

V. COMPARATIVE ANALYSIS OF MODELS

The performance of machine learning models for crop yield prediction varies significantly depending on the nature of the dataset, the quality of features, and the complexity of relationships among variables. Simpler models such as linear and polynomial regression are easy to implement and interpret, but they often struggle to capture nonlinear interactions present in agricultural data. As a result, their prediction accuracy is generally lower when compared to more advanced techniques, especially in scenarios involving multiple influencing factors such as weather variability, soil diversity, and crop management practices.

Decision tree-based models, particularly Random Forest, have demonstrated superior performance in many studies due to their ability to model nonlinear relationships and handle heterogeneous data. These models are less sensitive to noise and can effectively manage missing values, making them suitable for real-world agricultural datasets.

VI. PERFORMANCE EVALUATION METRICS

The evaluation of machine learning models for crop yield prediction is essential to determine their accuracy, reliability, and generalization capability. Various statistical metrics are used to measure how well a model's predicted values match the actual observed values. These metrics provide quantitative insight into model performance and enable comparison among different algorithms.

A. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) measures the average magnitude of errors between predicted and actual values without considering their direction. It provides a straightforward interpretation of prediction accuracy by calculating the average absolute difference. In crop yield prediction, MAE indicates how close the predicted yield values are to the actual yields. A lower MAE value represents better model performance. One advantage of MAE is that it treats all errors equally, making it less sensitive to outliers compared to other metrics.

B. Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) measures the square root of the average squared differences between predicted and actual values. Unlike MAE, RMSE gives higher weight to larger errors due to the squaring process. This makes RMSE particularly useful when large prediction errors are undesirable. In agricultural applications, RMSE helps identify models

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15530

that produce consistent and reliable predictions. A lower RMSE value indicates better model accuracy, but it is more sensitive to outliers than MAE.

C. R-squared (R^2)

R-squared (R^2), also known as the coefficient of determination, measures the proportion of variance in the dependent variable that is explained by the model. It ranges from 0 to 1, where a value closer to 1 indicates a better fit. In crop yield prediction, a high R^2 value means that the model successfully captures the relationship between input features and yield outcomes. However, R^2 alone may not fully reflect model performance, especially in cases of overfitting.

D. Accuracy

Accuracy is commonly used as an evaluation metric in classification problems, where predictions are categorized into discrete classes. In the context of crop yield prediction, accuracy may be used when yield levels are classified into categories such as low, medium, or high. It represents the proportion of correct predictions made by the model out of the total predictions. While accuracy is simple to understand, it may not be sufficient for evaluating regression-based yield prediction models, where continuous output values are more common.

VII. APPLICATIONS IN PRECISION AGRICULTURE

Machine learning-based crop yield prediction plays a significant role in advancing precision agriculture by enabling data-driven decision-making and efficient farm management. Precision agriculture focuses on optimizing agricultural practices by using technology to monitor and manage field variability. The integration of machine learning models with diverse data sources allows farmers and stakeholders to improve productivity, reduce costs, and ensure sustainable farming practices.

A. Efficient Resource Management

One of the primary applications of machine learning in precision agriculture is efficient resource management. By analyzing data related to soil conditions, weather patterns, and crop requirements, machine learning models can recommend the optimal use of resources such as water, fertilizers, and pesticides. This helps in minimizing waste and reducing environmental impact. For example, predictive models can determine the exact amount of irrigation needed based on soil moisture and weather forecasts, thereby conserving water. Similarly, fertilizer recommendations can be tailored to specific soil nutrient levels, improving crop health and yield while reducing excessive chemical usage.

B. Early Yield Forecasting

Machine learning enables early and accurate prediction of crop yields before the harvesting season. By using historical data along with current environmental conditions, models can forecast expected yield levels at different growth stages. Early yield forecasting helps farmers and policymakers make informed decisions regarding storage, transportation, and market planning. It also allows for timely interventions in case of expected low yield due to unfavorable conditions such as drought or pest infestations. This proactive approach reduces risks and enhances agricultural productivity.

C. Crop Monitoring Using Satellite Data

Remote sensing and satellite-based monitoring have become essential components of precision agriculture. Machine learning models analyze satellite imagery and vegetation indices to monitor crop health, detect stress conditions, and track growth patterns over time. This enables large-scale monitoring of agricultural fields without the need for manual inspection. For instance, changes in vegetation indices can indicate issues such as nutrient deficiency, water stress, or disease outbreaks. Early detection allows farmers to take corrective actions promptly, thereby preventing yield losses and improving overall crop performance.

D. Decision Support for Farmers and Policymakers

Machine learning-based systems provide valuable decision support tools for farmers, agricultural experts, and policymakers. These systems integrate data from multiple sources to generate actionable insights and recommendations. Farmers can use these insights to choose suitable crops, determine optimal planting times, and adopt effective management practices. Policymakers can utilize predictive models to plan food distribution, manage supply chains, and develop agricultural policies. Additionally, decision support systems help in addressing challenges such as climate change, population growth, and food security by enabling informed and strategic planning.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15530

VIII. CHALLENGES AND LIMITATIONS

Despite significant advancements in machine learning techniques for crop yield prediction, several challenges and limitations continue to affect the performance, reliability, and real-world adoption of these models. These challenges arise mainly due to the complexity of agricultural systems, variability in data sources, and practical constraints in implementation.

A. Poor Data Quality and Missing Values

One of the major challenges in crop yield prediction is the availability of high-quality data. Agricultural datasets often contain missing, incomplete, or inconsistent values due to errors in data collection, sensor failures, or lack of proper monitoring systems. For example, weather stations may not record data continuously, and soil measurements may be taken irregularly. Such gaps in data can negatively impact model training and reduce prediction accuracy. Additionally, noisy data and outliers can mislead machine learning algorithms, leading to unreliable results. Although techniques such as data imputation and preprocessing can help address these issues, they may not fully eliminate the impact of poor data quality.

B. Lack of Standardized Datasets

Another significant limitation is the absence of standardized and universally accepted datasets for crop yield prediction. Agricultural data is often region-specific, varying in format, scale, and quality across different countries and institutions. This lack of standardization makes it difficult to compare results across studies and limits the generalizability of machine learning models. Models trained on one dataset may not perform well when applied to another region due to differences in climate, soil conditions, and farming practices. As a result, there is a need for standardized data collection frameworks and publicly available datasets to improve consistency and reproducibility in research.

C. High Computational Requirements

Advanced machine learning models, particularly deep learning techniques, require significant computational resources for training and deployment. These models often involve large datasets and complex architectures, which demand high processing power, memory, and specialized hardware such as GPUs. This can be a major barrier, especially in developing regions where access to such resources is limited. Additionally, the time required for training and tuning these models can be substantial, making it challenging to implement real-time prediction systems. While cloud computing and distributed systems offer potential solutions, they also introduce additional costs and infrastructure requirements.

D. Limited Model Interpretability

Many advanced machine learning models, especially deep learning and ensemble methods, are often considered “black-box” systems. This means that while they may provide highly accurate predictions, it is difficult to understand how they arrive at those results. In agriculture, interpretability is important because farmers and stakeholders need to trust and understand the recommendations provided by the system. Lack of transparency can reduce user confidence and hinder adoption. Efforts are being made to develop explainable AI techniques that provide insights into model decisions, but achieving a balance between accuracy and interpretability remains a challenge.

IX. FUTURE RESEARCH DIRECTIONS

Future research in crop yield prediction using machine learning is expected to focus on improving model accuracy, scalability, and real-world applicability by leveraging emerging technologies and advanced methodologies. As agriculture continues to face challenges such as climate change, resource scarcity, and population growth, there is a growing need for more intelligent and adaptive prediction systems.

A. Integration of IoT and Real-Time Data

One of the most promising directions is the integration of Internet of Things (IoT) technologies with machine learning models. IoT devices such as soil sensors, weather stations, and smart irrigation systems can continuously collect real-time data related to soil moisture, temperature, humidity, and crop conditions. Incorporating this real-time data into predictive models enables dynamic and up-to-date yield predictions. This approach allows farmers to respond quickly to changing environmental conditions, improving decision-making and reducing risks. Future systems are expected to combine IoT data with cloud computing and edge computing to enable real-time analytics and automation in agriculture.

B. Development of Interpretable Models

While advanced machine learning models provide high accuracy, their lack of interpretability remains a major concern. Future research should focus on developing models that are not only accurate but also transparent and explainable.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.15530

Interpretable models help farmers and stakeholders understand how different factors influence crop yield predictions, increasing trust and adoption. Techniques such as explainable AI (XAI), feature importance analysis, and visualization tools are expected to play a key role in making machine learning models more understandable. Balancing accuracy with interpretability will be essential for practical deployment in agricultural systems.

C. Use of Hybrid and Ensemble Techniques

Hybrid and ensemble approaches are gaining attention as they combine the strengths of multiple models to achieve better performance. Future research is likely to explore more advanced combinations of machine learning and deep learning techniques, such as integrating statistical models with neural networks or combining different ensemble strategies. These approaches can improve prediction accuracy, robustness, and generalization across different datasets. Additionally, hybrid models can effectively handle diverse data types, including numerical, categorical, and image-based data, making them highly suitable for agricultural applications.

X. CONCLUSION

Machine learning techniques have significantly improved crop yield prediction by providing accurate and scalable solutions. This review highlights the strengths and limitations of various models and emphasizes the importance of data quality and model selection. Future advancements in technology and data integration will further enhance prediction accuracy and support sustainable agriculture.

REFERENCES

[1]. P. Pankaj, P. K. Bharti, and B. Kumar, “Crop Yield Prediction using Machine Learning: A Review of Recent Approaches,” International Journal of Computer Applications, vol. 185, no. 24, ISSN: 0975-8887, Jul. 2023, pp. 27–32.

[2]. A. Maheswary, S. Nagendram, K. U. Kiran, S. H. Ahammad, P. P. Priya, and M. A. Hossain, “Intelligent Crop Recommender System for Yield Prediction Using Machine Learning Strategy,” Journal of The Institution of Engineers (India): Series B, vol. 105, ISSN: 2250-2106, Aug. 2024, pp. 979–987.

[3]. S. Sah, D. Haldar, R. N. Singh, B. Das, and A. S. Nain, “Rice Yield Prediction through Integration of Biophysical Parameters with SAR and Optical Remote Sensing Data Using Machine Learning Models,” Scientific Reports, vol. 14, ISSN: 2045-2322, Sep. 2024, Article no. 21674.

[4]. S. Sharma et al., “Comparative Analysis on Crop Yield Forecasting using Machine Learning Techniques,” Rural Sustainability Research, vol. 52, ISSN: 2256-0939, Dec. 2024, pp. 63–77.

[5]. U. V. Nikhil, A. M. Pandiyan, S. P. Raja, and Z. Stamenkovic, “Machine Learning-Based Crop Yield Prediction in South India: Performance Analysis of Various Models,” Computers, vol. 13, no. 6, ISSN: 2073-431X, May 2024, Article no. 137.

[6]. M. P. Mahesh and R. Soundrapandiyan, “Yield Prediction for Crops by Gradient-Based Algorithms,” PLOS ONE, vol. 19, no. 8, ISSN: 1932-6203, Aug. 2024, e0291928.

[7]. 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, ISSN: 0168-1699, Oct. 2020, pp. 105709–105709.

[8]. M. Rashid, B. S. Bari, N. Khan, et al.,“A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction,” IEEE Access, vol. 9, ISSN: 2169-3536, Jan. 2021, pp. 63406–63439