STM JOURNALS

Research & Reviews: A Journal of
Crop Science and Technology

ISSN: 2319-3395
Volume 15, Issue 1, 2026
January–April
DOI (Journal): 10.37591/RRJoCST

https://journals.stmjournals.com/rrjocst

Review RRJoCST

A Data-Driven Analysis of Machine Learning Classification Models for Reliable Crop Yield Prediction

Raghunath Maji[1], Swarup Ghosh[2,*], Dipankar Barui[3], Sourav Chowdhury[4], Shreya Patra[5], Biswajit Gayen[6]

Abstract

The adoption of ML technologies in agriculture is reshaping farming practices, empowering producers to make informed, data-oriented decisions that improve yields, sustainability, and long-term resilience. In mango cultivation, ML analyzes data from weather, soil, and pests to optimize irrigation, fertilization, and pest control. Predictive analytics help forecast ideal farming practices, minimizing resource wastage and improving yield. Real-time monitoring and image-based disease detection allow timely interventions to maintain plant health and fruit quality. After harvesting, machine learning improves supply chain operations by forecasting market needs and limiting product deterioration. Combined with satellite imagery and drones, ML supports precision and eco-friendly farming. Additionally, ML-based fertilization and pest detection reduce chemical use and promote sustainability. Integration with blockchain ensures transparency and food safety. Overall, ML empowers mango farmers with precision tools to improve crop resilience, efficiency, and profitability amid changing climatic conditions. The adoption of ML-based decision support systems encourages data-backed planning rather than traditional intuition-driven farming, assisting farmers in selecting suitable mango varieties, optimizing planting density, and scheduling harvest operations to maximize market value. ML-powered mobile and cloud platforms enhance accessibility for small and marginal farmers by providing real-time insights, alerts, and recommendations at a low cost. By integrating historical trends with real-time sensor data, ML helps reduce uncertainty in farming operations and improves risk management. As climate variability intensifies, such intelligent systems play a critical role in ensuring stable production and long-term agricultural sustainability. In addition, continuous model learning enables adaptive responses to evolving field conditions, ensuring scalable deployment across diverse agro-climatic zones and production systems, ultimately strengthening food security while supporting farmer livelihoods and environmental conservation.

***Author for Correspondence**
Swarup Ghosh
E-mail: swarupghosh0207@gmail.com

[1]Assistant Professor, Department of Computer Science and Engineering, Greater Kolkata College of Engineering and Management, Baruipur, Kolkata, West Bengal, India.
[2]Student, Department of Computer Science and Engineering, Greater Kolkata College of Engineering and Management, Baruipur, Kolkata- West Bengal, India.
[3]Assistant Professor, Department of AI & ML, St. Thomas College of Engineering and Technology, Diamond Harbour Rd, Kidderpore, Kolkata, West Bengal, India.
[4]Student, Department of Computer Science and Engineering, Greater Kolkata College of Engineering and Management, Baruipur, Kolkata, West Bengal, India.
[5]Student, Department of Computer Science and Engineering, Dr. Sudhir Chandra Sur Institute of Technology and Sports Complex, Dum Dum, Kolkata, West Bengal, India
[6]Assistant Professor, Department of Basic Science, Greater Kolkata College of Engineering and Management, Baruipur, Kolkata, West Bengal, India

Received: December 8, 2025
Accepted: December 17, 2025
Published: January 10, 2026

Citation: Raghunath Maji, Swarup Ghosh, Dipankar Barui, Sourav Chowdhury, Shreya Patra, Biswajit Gayen. A Data-Driven Analysis of Machine Learning Classification Models for Reliable Crop Yield Prediction. A Journal of Crop Science and Technology. 2026; 15(1):12–17p.

Keywords: Machine learning, precision farming, mango production, crop yield, predictive analytics, sustainability, real-time monitoring, disease detection, post-harvest optimization, satellite imagery, drone technology, soil fertilization, blockchain, agricultural resilience, climate change

INTRODUCTION

Machine Learning (ML) offers innovative solutions to optimize mango crop production

through data analysis, yield prediction, disease detection, and efficient resource management, thereby enhancing productivity and sustainability [1]. By employing algorithms such as regression, classification, and clustering, farmers can make data-driven decisions to improve mango quality, minimize losses, and optimize irrigation and fertilization schedules. The integration of Internet of Things (IoT) devices with ML models enables real-time monitoring and automated responses, advancing precision and scalability in sustainable mango farming. Recent research highlights ML's effectiveness in disease detection, yield forecasting, and precision agriculture, showing its potential to transform mango farming for greater profitability and environmental balance. ML facilitates early detection of pests, nutrient deficiencies, and diseases, ensuring timely interventions that preserve crop health and maximize output. By leveraging predictive analytics, agricultural inputs such as water, fertilizers, and pesticides can be applied with greater precision, helping reduce unnecessary expenses while also protecting the environment from overuse.

By integrating remote sensing and IoT data, ML assists in identifying optimal harvesting periods, predicting pest outbreaks, and improving supply chain logistics. These predictive capabilities enable efficient harvest planning, reduced post-harvest losses, and improved market forecasting. Historical crop yield data can train ML models to anticipate future yields, enhancing decision-making and risk management [2].

Despite challenges such as incomplete or inconsistent datasets, ML's ability to identify patterns and develop predictive models enables continuous learning and adaptation without explicit programming. Both supervised and unsupervised learning approaches are applied depending on data type and objectives.

LITERATURE REVIEW

This paper presents a study on corn yield prediction using UAV-based multispectral and RGB imagery combined with machine learning models, focusing on the impact of spatial structure in data evaluation. By comparing traditional 10-fold cross-validation with spatial cross-validation methods, the authors demonstrate that ignoring spatial dependencies can lead to overoptimistic performance estimates. The study finds that mid-season multispectral imagery yields the best results, with Random Forest performing well on high-resolution data. Simple vegetation indices like NDVI RedEdge and NG emerged as strong predictors. The research highlights the need for spatially aware validation for accurate yield prediction in precision agriculture [4].

This study explores the use of a novel optimization technique, Randomized Search cross-validation (RScv), to enhance the accuracy of machine learning models in predicting annual yields of four major crops – Barley, Oats, Rye, and Wheat – across 20 European countries using satellite-based climate and soil data from NASA missions. By applying RScv to ensemble ML algorithms (AdaBoost, Gradient Boost, Random Forest, and Extra-Tree), the research demonstrates significant improvements in prediction accuracy, with the RScv-optimized AdaBoost model achieving over 90% accuracy. The study underscores the potential of combining remote sensing data with optimized ML techniques for scalable, accurate, and transferable yield prediction tools to support global food security and agricultural planning amid climate change challenges [5].

This paper presents a systematic literature review on the industrial maturity of machine learning (ML) applications in the food industry, highlighting a significant gap between research and practical deployment. It introduces a six-dimensional maturity assessment framework covering both technical (robustness, scalability, integrability) and human/process-related (explainability, usability) factors. The study finds that most ML solutions remain at low to medium maturity levels, particularly in integrability, usability, and packaging/logistics phases. The paper emphasizes the need for holistic approaches, real-world data usage, regulatory alignment, and industry collaboration to achieve higher levels of operational readiness and successful ML deployment across the food value chain [6].

This review paper explores advanced methods for automatic detection and classification of plant leaf diseases using machine learning (ML), deep learning (DL), and few-shot learning (FSL) techniques. It emphasizes the significance of early and accurate disease detection to minimize agricultural losses and ensure food security. The paper discusses various stages in the detection pipeline, including image acquisition, preprocessing, segmentation, feature extraction, and classification, highlighting popular algorithms like SVM, CNN, ANN, and decision trees. Despite promising results, the study also outlines several challenges such as overfitting, feature selection complexity, real-time image variability, and the shortage of annotated datasets. It concludes by recommending a more robust, adaptable, and hybrid approach for future plant disease detection systems, integrating molecular diagnostics and mobile-based technologies for better scalability and real-world application [7].

This study addresses the critical issue of crop loss due to plant diseases, which threatens global food security, by developing real-time datasets for rice, wheat, and maize – three essential food grains. These datasets include various common fungal and bacterial diseases captured in real-life scenarios with different severity levels. Due to limited initial image availability, data augmentation techniques were employed to expand the dataset, which was then used to train and evaluate eight fine-tuned deep learning models and a newly proposed CNN model (MRW-CNN). The best-performing models achieved high testing accuracy across all three crops, with Xception and MobileNet excelling in maize disease detection, MobileNetV2 and MobileNet in wheat, and Xception and InceptionV3 in rice. The proposed MRW-CNN model also demonstrated robust performance, surpassing some pre-trained models. A practical application of the system includes a plant disease protection app that cannot guide farmers – especially small-scale growers – by identifying diseases, suggesting treatment measures, and providing environmental risk alerts, thereby supporting timely and accurate decision-making in the field [8].

This study presents a robust methodology that integrates satellite-based climate data and machine learning techniques to accurately predict crop yields, which is vital for global food security and sustainable agriculture. Utilizing remote sensing inputs from NASA's GPCP and GLDAS missions, the researchers applied a novel optimization method – Randomized Search Cross-Validation (RScv) – to four ensemble ML models (AdaBoost, Gradient Boosting, Random Forest, and Extra-Trees) across four crops (Barley, Oats, Rye, and Wheat) in 20 European countries over two decades. Among these, the RScv-optimized AdaBoost (RScv-AB) achieved the highest prediction accuracy ($R^2 \approx 0.90$), outperforming traditional AI models like ANN, KNN, and SVR. Spatial error analysis showed a slight tendency towards underestimation, with Barley showing more prediction uncertainty than Wheat. Compared to existing literature, this RScv-ML approach demonstrated superior performance and transferability, proving useful for policymakers and farmers alike in managing water allocation, mitigating climate change effects, and enhancing crop productivity. Despite its regional focus and reliance on climate parameters, the model's adaptability to broader geographies and inclusion of socio-economic variables in future work makes it a promising tool for scalable agricultural forecasting [9].

This paper presents a comprehensive review of the integration of deep learning with smart agricultural equipment for effective weed and crop recognition. It highlights the limitations of traditional weed control methods, such as excessive herbicide use, and showcases how technologies like drones, satellites, and agricultural robots offer eco-friendly, precise, and cost-effective alternatives. The study emphasizes the importance of accurate image acquisition, feature extraction, and optimized algorithms for high recognition accuracy, with CNNs and models like DeepLabv3+ and Swin Transformer showing promising results. While smart equipment like UAVs and robots enable real-time monitoring and targeted weed removal, challenges remain, including dataset annotation, environmental variability, and visual similarity between weeds and crops. The paper concludes by suggesting improvements in model robustness, generalization, and sensor integration as vital steps toward sustainable, intelligent farming [10].

This paper presents a comprehensive review of various machine learning and statistical techniques applied in the domain of agricultural crop production forecasting. It emphasizes the necessity for accurate and timely crop yield predictions, which are vital for effective policymaking and resource planning. The authors discuss multiple methodologies such as regression analysis, clustering (notably k-means), principal component analysis (PCA), Bayesian belief networks, time series forecasting, and Markov chain models. Each technique is evaluated through relevant case studies and empirical data, highlighting their effectiveness in capturing the complex, multivariate factors influencing crop yield. The paper notably concludes that machine learning approaches, when combined with domain knowledge and diverse data sources (e.g., weather, soil, and remote sensing), can significantly enhance the precision and reliability of crop production forecasts [11].

This study emphasizes the effectiveness of machine learning and deep learning techniques in accurately predicting potato crop yields, which is essential for sustainable agriculture and food security. Various models, including K-nearest neighbors (KNN), gradient boosting, XGBoost, and multilayer perceptron (MLP), were evaluated alongside advanced deep learning models such as graph neural networks (GNNs), gated recurrent units (GRUs), and long short-term memory networks (LSTMs). The models were assessed using comprehensive performance metrics like mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). Results showed that while gradient boosting and XGBoost performed well among machine learning models, GNNs outperformed all, achieving the lowest MSE (0.02363) and the highest R^2 (0.51719), indicating superior precision and the ability to capture complex spatial-temporal patterns. LSTMs and GRUs also demonstrated strong performance, highlighting the growing potential of deep learning in enhancing agricultural forecasting and decision-making [12].

This study highlights the importance of leveraging machine learning to enhance agricultural productivity amidst limited natural resources and growing population demands. The research introduces an automated crop yield prediction model that uses the Grey Level Co-occurrence Matrix (GLCM) for effective feature selection and applies AdaBoost Decision Tree, Artificial Neural Network (ANN), and K-Nearest Neighbor (KNN) for classification. Using a comprehensive dataset from the Food and Agriculture Organization (FAO) and World Data Bank, comprising 33 features related to crops like potato, maize, rice, wheat, and soybean, the AdaBoost classifier with GLCM feature selection achieved the highest accuracy (98%), precision (99%), and recall (98%), outperforming ANN and KNN. The results confirm that GLCM feature selection significantly boosts the performance of machine learning models, with AdaBoost emerging as the most reliable technique for crop yield prediction under varying conditions [13].

This study focuses on predicting crop yields using machine learning and deep learning techniques in the context of smart farming, which integrates technologies like Artificial Intelligence (AI) and the Internet of Things (IoT) to improve agricultural sustainability. Given the challenges of food security, climate change, and resource scarcity, the study employs models like Random Forest, Decision Tree, XGBoost, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) to predict agricultural productivity in India. The results demonstrate that Random Forest provides the highest accuracy (98.96%) for crop yield predictions, while CNN outperforms LSTM in terms of test loss. The study underscores the potential of machine learning and deep learning in enhancing crop yield forecasting, offering a solution for small-scale farmers to optimize production planning amidst growing global food demands and climate uncertainties [14].

The ID3 approach utilizes both entropy and information gain measurements to build a decision tree. At each stage of the analysis on the dataset, the entropy of the features is computed. The features with the highest information gain and the lowest entropy are chosen as split attributes. This process is repeated for each new subset of data until a perfect fit is achieved for all the data in the predefined categories. In this method, terminal nodes of the tree are labeled with class labels, while non-terminal

nodes are defined by the split attributes. By using a decision tree model built with the ID3 algorithm, heart disease can be identified and predicted with high accuracy. Decision tree modeling techniques, such as Classification and Regression Tree (CART) and ID3, are commonly used in building prediction models for large health datasets. A tenfold cross-validation method is often applied to ensure the reliability of the results. Based on the data, decision tree classification techniques, such as ID3, provide an accurate and cost-effective way to create prediction models. Additionally, the ID3 approach can be particularly useful for analyzing smaller datasets quickly, with results that may be more accurate. In contexts like electronic health records, where data is continuously updated, the ID3 approach can be efficient despite the large storage space requirements for handling dynamic data [3, 15].

METHODOLOGY

Data Processing

Developing a machine learning model for crop production classification begins not with algorithms, but with data – its quality, structure, and relevance. The effectiveness of any AI model heavily depends on how well the data is prepared before training. In this project, we placed significant emphasis on building a strong data foundation by collecting, cleaning, integrating, and analyzing agricultural datasets from multiple trusted sources. This detailed data processing phase ensures that the subsequent machine learning steps are built on a reliable and meaningful dataset as shown in Figure 1.

CONCLUSION

The integration of Machine Learning (ML) in agriculture has led to remarkable advancements in farming practices, particularly in the cultivation of high-value crops like mangoes. By leveraging data-driven technologies, ML enhances productivity and sustainability by enabling precise crop management. Through predictive analytics, farmers can make informed decisions regarding irrigation, fertilization, and pest control, which significantly reduces resource wastage and maximizes efficiency. The ability to monitor mango orchards in real-time through IoT sensors ensures timely interventions, preserving plant health and improving fruit quality throughout the growth cycle.

Additionally, ML plays a pivotal role in early disease detection, utilizing sensor and image data to mitigate crop loss and ensure the consistent quality of mangoes. Post-harvest, the technology optimizes the supply chain by predicting demand, minimizing spoilage, and streamlining distribution, which is critical for perishable fruits like mangoes. ML also facilitates large-scale environmental monitoring through satellite imagery and drone technology, enabling farmers to monitor stressors and allocate resources more effectively.

The ability to precisely fertilize crops based on soil nutrient analysis promotes eco-friendly practices and resource conservation. Accurate harvest prediction, enabled by ML, improves labor planning and reduces waste, while weather-forecasting models enhance proactive crop protection. ML also contributes to reducing pesticide usage by identifying potential pest outbreaks early through environmental and crop data evaluation, fostering healthier ecosystems.

Moreover, the integration of blockchain technology ensures transparency and traceability across the mango supply chain, enhancing consumer confidence and compliance with food safety standards. The customization of farming strategies based on local conditions, such as microclimates and soil types, further strengthens mango crop resilience against climate change. With the support of drone-assisted surveillance, real-time insights into soil moisture and nutrient levels enable precise irrigation and fertilization, optimizing plant growth and resource usage. The synergy between machine learning, smart technologies, and data analytics offers mango farmers innovative solutions to tackle agricultural challenges effectively. This integration not only ensures consistent harvests but also promotes sustainable practices, contributing to a profitable and resilient future in food production. As

collaborations with research institutions continue to foster innovation, mango producers will remain equipped with the latest technologies to drive agricultural development in a rapidly changing world.

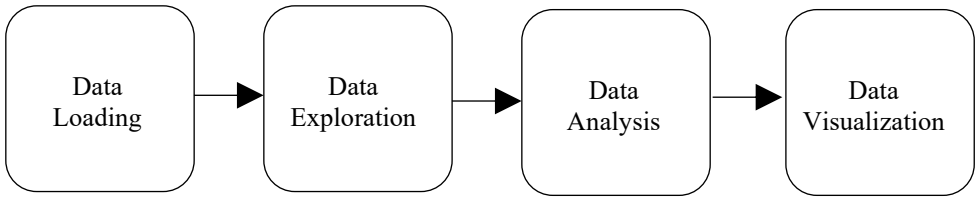

Figure 1. Steps of Data Processing to build on a reliable and meaningful dataset.

REFERENCES

1. Bhanu KN, Jasmine HJ, Mahadevaswamy HS. Machine learning implementation in IoT-based intelligent system for agriculture. In: Proc Int Conf Emerg Technol (INCET); 2020. p. 1–5. doi: 10.1109/INCET49848.2020.9153978.
2. Sharma A, Jain A, Gupta P, Chowdary V. Machine learning applications for precision agriculture: a comprehensive review. IEEE Access. 2021;9:4843–4873. doi: 10.1109/ACCESS.2020.3048415.
3. Raghuvanshi A, Singh U, Sajja G, Pallathadka H, Asenso E, Kamal M, et al. Intrusion detection using machine learning for risk mitigation in IoT-enabled smart irrigation in smart farming. J Food Qual. 2022;2022:1–8. doi: 10.1155/2022/3955514.
4. Killeen P, Kiringa I, Yeap T, Branco P. Corn grain yield prediction using UAV-based high spatiotemporal resolution imagery, machine learning, and spatial cross-validation. Remote Sens. 2024;16(4):683. doi: 10.3390/rs16040683.
5. Haji Seyed Asadollah SB, Jodar-Abellan A, Pardo MA. Optimizing machine learning for agricultural productivity: a novel approach with RScv and remote sensing data over Europe. Agric Syst. 2024;218:103955. doi: 10.1016/j.agsy.2024.103955.
6. Gradl L, Reis L, Buettner R. Industrial maturity of machine learning solutions within the food industry. IEEE Access. 2025;13:25534–25544. doi: 10.1109/ACCESS.2025.3558091.
7. Bhargava A, Shukla A, Goswami OP, Alsharif MH, Uthansakul P, Uthansakul M, et al. Plant leaf disease detection, classification, and diagnosis using computer vision and artificial intelligence: A review. IEEE Access. 2024;12:12345–12357. doi: 10.1109/ACCESS.2024.3373001.
8. Joseph DS, Pawar PM, Chakradeo K. Real-time plant disease dataset development and detection of plant disease using deep learning. IEEE Access. 2024;12:12345–12358. doi: 10.1109/ACCESS.2024.3358333.
9. Haji Seyed Asadollah SB, Jodar-Abellan A, Pardo MA. Optimizing machine learning for agricultural productivity: A novel approach with RScv and remote sensing data over Europe. Agric Syst. 2024;218:103955. doi: 10.1016/j.agsy.2024.103955.
10. Qu HR, Su WH. Deep learning-based weed–crop recognition for smart agricultural equipment: A review. Agronomy. 2024;14(2):363. doi: 10.3390/agronomy14020363.
11. Mishra S, Mishra D, Santra GH. Applications of machine learning techniques in agricultural crop production: A review paper. Indian J Sci Technol. 2016;9(38):1–8. doi: 10.17485/ijst/2016/v9i38/95032.
12. El-Kenawy EM, Alhussan AA, Khodadadi N, Mirjalili S, Eid MM. Predicting potato crop yield with machine learning and deep learning for sustainable agriculture. Potato Res. 2025;68:759–792. doi: 10.1007/s11540-024-09753-w.
13. Nagesh OS, Budaraju RR, Kulkarni SS, Vinay M, Ajibade SSM, Chopra M, et al. Boosting-enabled efficient machine learning technique for accurate prediction of crop yield towards precision agriculture. Discover Sustain. 2024;5:78. doi: 10.1007/s43621-024-00254-x.
14. Harinath D, Patil A, Bandi M, Raju AVS, Ramana Murthy MV, Spandana D. Smart farming system—an efficient technique for predicting agriculture yields using machine learning. Technische Sicherheit. 2024 Dec. Available from: https://www.researchgate.net/publication/387306143

15. Zamani LS, Anand KPR, Prabhu P, Buttar AM, Pallathadka ARH, Dugbakie BN. Performance of machine learning and image processing in plant leaf disease detection. J Food Qual. 2022;2022:1–7. doi: 10.1155/2022/1598796.

