RESEARCH ARTICLE OPEN ACCESS

Crop Yield Prediction and Classification through Machine learning: A Comparative Study of Data-Driven Agricultural Approaches

Anushri Mishra[1], Isha Sharma[2], Rakesh Kumar Tiwari[3], Onkar Nath Thakur[4]
M. Tech Scholar[1], Assistant Professor[2, 3 & 4]
Department of Computer Science & Engineering[1, 2, 3 & 4]
Technocrats Institute of Technology & Science, Bhopal, India[1, 2, 3& 4]

Abstract
Accurate crop yield prediction is a critical component of modern precision agriculture, enabling informed decision-making for food security, resource optimization, and sustainable farming practices. In recent years, machine learning (ML) and deep learning (DL) techniques have gained significant attention due to their ability to model complex, nonlinear relationships among climatic, soil, and crop management factors. This review paper systematically analyzes and compares classical machine learning and deep learning approaches applied to crop yield prediction and classification. The study reviews existing literature, evaluates commonly used datasets, algorithms, and performance metrics, and highlights their strengths and limitations. A comparative analysis is conducted to assess the effectiveness of traditional ML models versus deep learning architectures. The findings reveal that ensemble learning methods and deep neural networks generally outperform conventional techniques, though challenges related to data quality, scalability, and model interpretability remain. Future research directions are outlined to address these gaps and enhance the robustness of agricultural analytics systems. The study provides a comprehensive analytical understanding for developing reliable crop yield prediction systems supporting sustainable agriculture.

Keywords-Crop Yield Prediction; Machine Learning; Deep Learning; Precision Agriculture; Ensemble Models; Classification

---

Date of Submission: 01-06-2026 Date of acceptance: 10-06-2026

I. INTRODUCTION

Agriculture plays a vital role in the global economy and food security, particularly in developing countries where a large population depends on farming for livelihood. Accurate crop yield prediction is essential for effective agricultural planning, risk management, and policy formulation. Traditional statistical and empirical methods often fail to capture the complex interactions between climatic conditions, soil characteristics, crop varieties, and management practices, leading to limited prediction accuracy [1]. With the rapid advancement of data availability from sensors, remote sensing, and agricultural surveys, machine learning and deep learning techniques have emerged as powerful tools for agricultural analytics. These approaches are capable of handling large-scale, high-dimensional datasets and modeling nonlinear relationships more effectively than conventional methods. Consequently, researchers have increasingly applied algorithms such as Random Forest, Support Vector Machines, Gradient Boosting, Artificial Neural Networks, and Deep Learning models for crop yield estimation and classification [2]. Despite significant progress, several challenges persist, including data heterogeneity, missing values, regional variability, and limited generalization across different agro-climatic zones. Moreover, there is a lack of comprehensive reviews that systematically compare classical ML and DL approaches using consistent evaluation criteria [3]. The aim of this review paper is to provide a structured and comparative analysis of existing machine learning and deep learning techniques for crop yield prediction and classification.

The paper highlights key methodologies, datasets, performance metrics, and identifies research gaps, thereby contributing valuable insights for researchers and practitioners working in precision agriculture. This review paper presents a comprehensive comparative analysis of existing approaches used for crop yield prediction and agricultural analytics. The study systematically examines recent research contributions to understand how modern predictive techniques address

challenges associated with agricultural productivity, environmental variability, and data-driven decision-making. The review focuses on evaluating different prediction methodologies by analyzing their strengths, limitations, scalability, and practical applicability in precision agriculture. By synthesizing findings from multiple studies, the paper identifies key research gaps that persist in current literature, particularly regarding model generalization, interpretability, and real-world implementation challenges. Furthermore, the review critically highlights limitations observed in previous research efforts, including dependency on region-specific datasets, computational complexity, and insufficient integration of decision-support mechanisms for farmers. In addition to analyzing existing work, the paper outlines important future research directions aimed at developing more robust, adaptable, and user-oriented agricultural prediction frameworks

1.1 Machine Learning Techniques - Machine Learning (ML) plays a crucial role in modern crop yield prediction by enabling data-driven analysis of complex agricultural systems. Traditional statistical methods often fail to capture the non-linear and dynamic relationships between environmental, soil, and management factors. Machine learning algorithms overcome these limitations by learning patterns directly from historical agricultural data. ML models analyze multiple variables such as temperature, rainfall, soil nutrients, crop type, and fertilizer usage to accurately estimate crop yield [4]. These algorithms adapt to changing climatic conditions and continuously improve prediction accuracy as more data becomes available. Machine learning also supports early yield forecasting, allowing farmers and policymakers to make timely decisions related to irrigation planning, fertilizer application, and resource allocation. Furthermore, ML enhances agricultural sustainability by identifying key factors that influence yield variability and reducing uncertainty in production estimates. Overall, machine learning provides an efficient, scalable, and reliable framework for precision agriculture and intelligent crop yield management.

Random Forest is an ensemble-based predictive algorithm that combines multiple decision trees to improve prediction accuracy, stability, and robustness. The model operates by constructing numerous decision trees using randomly selected subsets of training data and input features, and the final prediction is obtained by aggregating the outputs of all individual trees. This ensemble learning mechanism reduces model variance and enhances generalization performance, making Random Forest particularly effective for complex prediction tasks. In crop yield prediction, the algorithm demonstrates strong capability in modeling nonlinear relationships among diverse agricultural variables such as rainfall, temperature, soil nutrients, fertilizer application, humidity levels, and crop management practices. Agricultural datasets often contain noise, missing values, and heterogeneous feature types collected from multiple sources; Random Forest efficiently manages these challenges through random sampling, feature bagging, and robust tree-based learning strategies [5].

Another significant advantage of Random Forest is its resistance to overfitting. By averaging predictions from multiple decision trees, the model minimizes the risk of learning dataset-specific noise, thereby providing reliable predictions even under uncertain environmental conditions. This property is particularly valuable in agricultural environments where climatic variability and incomplete datasets are common. Additionally, Random Forest provides feature importance analysis, enabling researchers and agronomists to identify key environmental and management factors that significantly influence crop productivity. Such interpretability supports better understanding of agricultural systems and assists policymakers in developing data-driven farming strategies. In this study, Random Forest is utilized as a baseline prediction model to analyze historical agricultural patterns and generate stable yield estimations across varying environmental scenarios. Its balance between prediction accuracy, interpretability, and computational efficiency makes it a widely adopted approach for practical crop yield forecasting and agricultural decision-support applications [6].

XGBoost (Extreme Gradient Boosting) is an advanced gradient boosting algorithm widely recognized for its high predictive accuracy and computational efficiency. The algorithm constructs decision trees sequentially, where each successive tree is designed to minimize the errors produced by the previous models. This iterative learning mechanism enables XGBoost to capture complex nonlinear relationships among agricultural variables such as climatic conditions, soil characteristics, crop management practices, and environmental factors [7]. One of the key strengths of XGBoost lies in its built-in regularization techniques, which help prevent overfitting and improve model generalization when working with complex and high-dimensional agricultural datasets. Additionally, the algorithm efficiently handles missing data and large feature spaces, making it highly suitable for real-world crop yield prediction tasks involving diverse input parameters. XGBoost also provides optimized

computation through parallel processing and gradient optimization, resulting in faster training time compared to traditional boosting approaches. In agricultural prediction applications, the model effectively learns subtle variations in crop productivity influenced by environmental and management conditions. Its strong generalization capability supports reliable early-season yield forecasting, enabling farmers, researchers, and agricultural planners to make informed decisions related to resource allocation, crop planning, and risk management strategies [8].

LightGBM (Light Gradient Boosting Machine) is an advanced gradient boosting framework designed to achieve high predictive performance while maintaining computational efficiency, particularly for large-scale and high-dimensional datasets. The algorithm employs a histogram-based learning technique that discretizes continuous feature values into bins, significantly reducing memory usage and training time compared to traditional boosting algorithms. This optimization makes LightGBM highly suitable for agricultural applications where datasets often contain large volumes of climatic, soil, and crop management information. In crop yield prediction, LightGBM effectively handles datasets consisting of both numerical and categorical attributes, enabling efficient analysis of complex agricultural variables. Unlike conventional level-wise tree growth strategies, LightGBM utilizes a leaf-wise tree growth approach, which expands the leaf with the highest loss reduction [9].

Another important advantage of LightGBM is its scalability and ability to process large datasets with faster training speed without compromising accuracy. The algorithm supports parallel learning and optimized gradient-based sampling techniques, which enhance computational efficiency and reduce model training complexity [10]. In agricultural prediction tasks, LightGBM can effectively learn yield variations influenced by climate variability, soil fertility conditions, irrigation practices, and seasonal environmental changes. Its balanced combination of speed, accuracy, and scalability makes it well-suited for real-time agricultural decision-support systems and large-scale forecasting applications [11]. By enabling efficient processing of extensive agricultural data, LightGBM contributes to improved yield estimation, resource optimization, and informed decision-making for farmers, researchers, and policymakers involved in sustainable agricultural planning and long-term crop production management.

II. LITERATURE REVIEW

Several researchers have investigated data-driven approaches for improving crop yield prediction and agricultural decision-making systems. Kumar et al. (2020) investigated the application of Decision Tree and Logistic Regression algorithms for crop yield classification and agricultural decision-support systems. The primary objective of their research was to develop an effective prediction framework that could assist farmers in identifying suitable crops and improving agricultural productivity. The study utilized historical agricultural datasets containing information related to rainfall, temperature, humidity, soil characteristics, and crop production records. Decision Tree models were found to be highly interpretable, enabling users to understand the influence of individual parameters on crop yield outcomes. Logistic Regression demonstrated reliable classification performance and offered a simple yet effective approach for agricultural decision-making [12]. The authors emphasized that these algorithms require relatively low computational resources and can be implemented efficiently in practical agricultural environments. However, despite their advantages, the models exhibited limitations in capturing complex nonlinear relationships among multiple agricultural variables. Their predictive accuracy was comparatively lower than that of advanced ensemble and deep learning techniques, particularly when applied to large-scale and heterogeneous datasets. The study concluded that while traditional machine learning methods remain useful for preliminary agricultural analysis, more sophisticated predictive frameworks are necessary to address the growing complexity of modern precision agriculture systems and achieve improved prediction accuracy and reliability [13].

Rahman et al. (2023) proposed a deep learning-based crop yield prediction framework utilizing Multilayer Perceptron (MLP) neural networks. Their research focused on exploiting the capability of deep learning models to learn hidden patterns and nonlinear relationships among agricultural variables. The proposed framework was trained using large datasets containing climatic, environmental, and soil-related information such as temperature, rainfall, humidity, soil nutrient content, and historical crop productivity records. Experimental evaluation demonstrated that the deep learning model achieved superior prediction accuracy compared to several conventional machine learning algorithms. The MLP architecture effectively modeled complex interactions among agricultural parameters, resulting in lower prediction errors and improved generalization performance. Furthermore, the study highlighted the ability of deep learning techniques to process large and multidimensional

datasets, making them suitable for modern precision farming applications. However, the authors identified several challenges associated with the implementation of deep learning systems. One major limitation was the lack of interpretability, as neural networks often function as black-box models that provide limited insight into the reasoning behind predictions. Additionally, the framework required extensive computational resources, large volumes of training data, and careful hyperparameter optimization to achieve optimal performance. These factors may limit the adoption of deep learning-based agricultural prediction systems in regions with limited technological infrastructure. The researchers suggested integrating explainable artificial intelligence techniques to improve model transparency and enhance user confidence in prediction outcomes [14].

Singh et al. (2019) conducted a comparative analysis of multiple machine learning algorithms for crop yield prediction using historical agricultural datasets collected from diverse farming environments. The study evaluated various predictive models, including Decision Trees, Support Vector Machines, Artificial Neural Networks, Random Forests, and ensemble learning approaches. The objective was to determine the most effective methodology for predicting agricultural productivity under varying environmental and climatic conditions. The findings revealed that ensemble learning techniques consistently outperformed individual machine learning models in terms of prediction accuracy, robustness, and generalization capability. By combining the outputs of multiple learners, ensemble approaches were able to reduce prediction errors and improve model stability across different datasets. The researchers also emphasized the importance of feature selection, data preprocessing, and parameter optimization in enhancing predictive performance. Furthermore, ensemble models demonstrated superior capability in handling heterogeneous datasets containing complex relationships among climatic, environmental, and soil-related variables. Despite these advantages, the study identified several practical limitations. Ensemble methods generally required higher computational power, increased memory utilization, and longer training times compared to traditional machine learning techniques. Such requirements may pose challenges in resource-constrained agricultural environments where access to advanced computing infrastructure is limited. The authors concluded that future research should focus on developing computationally efficient ensemble frameworks capable of maintaining high predictive accuracy while reducing implementation complexity and resource requirements [15].

Chen et al. (2024) explored the use of advanced gradient boosting algorithms, specifically LightGBM and XGBoost, for crop yield regression and prediction tasks. Their research aimed to improve agricultural forecasting accuracy by leveraging the powerful learning capabilities of boosting-based machine learning models. The study employed extensive datasets containing weather conditions, soil properties, irrigation details, fertilizer usage information, and historical crop yield records. Experimental results indicated that both LightGBM and XGBoost achieved excellent predictive performance, reflected by high coefficient of determination (R^2) values and low prediction errors. The models effectively captured nonlinear interactions among agricultural variables and demonstrated strong generalization capabilities across different crop yield scenarios. An important contribution of the study was the application of feature importance analysis, which enabled the identification of critical factors influencing crop productivity. Such insights can assist farmers and agricultural planners in making informed decisions regarding resource allocation and crop management strategies. Nevertheless, the authors reported that the success of boosting techniques heavily depended on effective feature engineering, appropriate data preprocessing, and careful hyperparameter tuning. Achieving optimal performance required substantial expertise and computational effort, which may limit practical adoption among non-technical users. Additionally, the study highlighted the need for broader validation across different geographical regions and crop varieties to ensure model reliability and generalizability. The researchers recommended integrating explainable AI mechanisms with boosting frameworks to improve transparency and facilitate practical deployment in agricultural decision-support systems [16].

Sengaliappan and Bharathkumar (2025) proposed a machine learning-based crop yield prediction system that utilized climatic and soil-related parameters to estimate agricultural productivity. The study evaluated multiple machine learning algorithms, including K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Logistic Regression, with the objective of identifying the most suitable predictive model for agricultural applications. The dataset used in the research incorporated various environmental and agronomic factors such as rainfall, temperature, humidity, soil fertility, and crop-specific characteristics. Experimental findings demonstrated that Random Forest achieved the highest prediction accuracy, precision, and overall classification performance among the evaluated models. The ensemble nature of Random Forest enabled it to effectively manage

nonlinear relationships and minimize overfitting, resulting in more reliable prediction outcomes. KNN and Decision Tree models also produced satisfactory results, while Logistic Regression provided a simpler and more interpretable solution for agricultural analysis. The study emphasized the growing significance of machine learning technologies in precision agriculture and highlighted their potential to support farmers in improving productivity and resource management. However, the research was limited to traditional machine learning approaches and did not investigate advanced boosting algorithms such as XGBoost or LightGBM, which have demonstrated superior performance in various prediction tasks. Moreover, the evaluation was conducted on relatively limited datasets, restricting the assessment of scalability and generalization across diverse agricultural conditions. The authors suggested that future studies should focus on integrating advanced ensemble techniques, real-time environmental monitoring systems, and large-scale agricultural datasets to enhance prediction accuracy, robustness, and practical applicability in modern farming environments [17].

Ibañez and Monterola (2023) proposed a Transformer-based crop production forecasting framework for large-scale agricultural prediction. Their research utilized Time Series Transformer architectures to predict crop production across multiple provinces using historical agricultural production records, weather variables, and temporal agricultural indicators. The proposed methodology aimed to capture long-range temporal dependencies that traditional machine learning models often fail to represent effectively. Experimental results demonstrated that Transformer models achieved superior forecasting accuracy compared to conventional statistical approaches and several baseline machine learning techniques. The framework effectively handled complex seasonal patterns and fluctuations in crop production data, thereby improving forecasting reliability. Furthermore, the model exhibited strong scalability when applied to multiple crops and geographical regions. However, the study reported that Transformer architectures require substantial computational resources and extensive training data for optimal performance. The complexity of model tuning and the need for large-scale datasets may limit deployment in data-scarce agricultural environments. Additionally, the interpretability of deep Transformer networks remained a challenge, making it difficult for agricultural stakeholders to understand prediction mechanisms. The authors concluded that future research should focus on integrating explainable AI techniques and lightweight Transformer architectures to improve transparency and practical applicability in precision agriculture systems [18].

Pandit et al. (2023) proposed hybrid time-series forecasting models integrated with exogenous climatic variables for predicting the yield of major Rabi crops in India. The methodology combined traditional statistical forecasting techniques with machine learning-based temporal analysis to capture both historical trends and environmental influences. Climatic variables such as rainfall, temperature, humidity, and seasonal weather indicators were incorporated into the prediction framework. Experimental evaluation revealed that the hybrid models significantly outperformed conventional forecasting approaches in terms of prediction accuracy and stability. The integration of exogenous variables enabled the model to capture climate-driven variations in crop productivity more effectively. The study demonstrated the practical utility of hybrid forecasting systems for agricultural planning and resource management. Nevertheless, the researchers highlighted several limitations, including dependence on high-quality climatic datasets and the complexity associated with integrating multiple forecasting techniques. The framework also exhibited reduced performance when environmental conditions deviated significantly from historical patterns. Furthermore, implementation required substantial preprocessing and domain expertise. The authors suggested future research involving advanced deep learning architectures and real-time environmental monitoring systems to enhance forecasting precision and adaptability under changing climatic conditions [19].

Nagesh et al. (2024) proposed a boosting-enabled machine learning framework for crop yield prediction in precision agriculture. Their methodology employed advanced boosting algorithms to improve predictive performance by sequentially correcting errors generated by weak learners. The research utilized agricultural datasets containing weather information, soil characteristics, fertilizer usage records, and historical crop yields. Experimental findings indicated that boosting-based models achieved higher prediction accuracy and lower error rates compared to conventional machine learning algorithms. The framework effectively captured nonlinear interactions among agricultural variables and demonstrated strong generalization capabilities. An important contribution of the study was the identification of influential agricultural parameters through feature importance analysis. These insights can assist farmers in optimizing cultivation practices and resource allocation. However, the researchers noted that boosting algorithms require careful hyperparameter optimization and substantial computational effort to

achieve optimal results. The model's performance was also highly dependent on data quality and feature engineering processes. Furthermore, the complexity of boosting frameworks may present challenges for non-technical agricultural users. The authors recommended integrating explainable AI techniques and automated parameter optimization mechanisms to improve transparency and facilitate broader adoption in agricultural decision-support applications [20].

Wang et al. (2024) proposed a Temporal–Geospatial Deep Learning Framework for crop yield prediction that integrated spatial and temporal information simultaneously. The methodology utilized deep neural networks capable of modeling complex relationships among weather variables, geographical characteristics, and crop production records. The framework incorporated temporal dependencies and spatial correlations to improve prediction accuracy across different agricultural regions. Experimental results demonstrated significant improvements in predictive performance compared to conventional machine learning approaches. The proposed model successfully captured regional agricultural variations and temporal growth patterns, leading to more reliable yield estimations. Additionally, the framework exhibited strong scalability and adaptability for large agricultural datasets. Despite these advantages, the study identified limitations related to computational complexity and the requirement for extensive training datasets. The deep learning architecture also functioned largely as a black-box model, reducing interpretability and stakeholder trust. Furthermore, data heterogeneity across regions sometimes affected model consistency. The researchers emphasized the need for explainable AI integration and more efficient deep learning architectures to support practical deployment in real-world precision farming environments and agricultural management systems [21].

Anoop et al. (2025) proposed the 3TFL-XLNet-CP framework, a Transformer-based crop yield prediction model utilizing weighted-loss and three-tier feature learning mechanisms. The objective of the study was to improve predictive accuracy by capturing complex agricultural patterns through advanced Transformer architectures. The framework analyzed climatic factors, soil characteristics, and crop-specific variables simultaneously. Experimental evaluation demonstrated that the proposed Transformer model significantly outperformed several traditional machine learning and deep learning techniques. The weighted-loss strategy effectively reduced prediction errors and enhanced model robustness. Moreover, the three-tier feature learning approach improved the extraction of relevant agricultural information from multidimensional datasets. However, the framework required substantial computational resources and extensive model training. Hyperparameter tuning remained a complex and time-consuming process. Additionally, the black-box nature of Transformer models limited interpretability and practical understanding among agricultural stakeholders. The researchers suggested incorporating explainable artificial intelligence mechanisms and optimizing model efficiency to improve transparency, scalability, and deployment feasibility in precision agriculture applications [22].

Ravikumar et al. (2025) proposed the SERWI hybrid ensemble model for crop yield prediction using an inverse RMSE weighting strategy. The methodology combined multiple predictive learners to enhance forecasting performance while reducing individual model biases. The framework utilized climatic, environmental, and soil-related agricultural datasets. Experimental results revealed that the ensemble model achieved superior predictive accuracy, robustness, and stability compared to standalone machine learning approaches. The inverse RMSE weighting mechanism effectively improved model reliability by assigning greater importance to stronger predictors. The study highlighted the potential of ensemble intelligence for agricultural forecasting and decision support. However, the researchers noted that ensemble systems involve increased computational complexity, longer training times, and higher resource consumption. Moreover, model interpretability decreased as additional learners were incorporated into the ensemble. Data preprocessing and parameter optimization also required significant effort. The authors recommended future research focused on developing lightweight ensemble architectures and explainable frameworks capable of maintaining predictive performance while reducing implementation complexity and computational overhead [23].

Pathak et al. (2023) proposed a multimodal machine learning framework for crop yield prediction at field and sub-field levels. The methodology integrated satellite imagery, weather information, soil characteristics, and digital elevation models to improve prediction accuracy. The researchers developed an early-fusion approach capable of handling heterogeneous agricultural data sources with varying temporal and spatial resolutions. Experimental results demonstrated that combining multiple modalities significantly enhanced predictive performance compared to single-source models. The framework effectively captured spatial variability within agricultural fields and enabled more precise yield estimation.

Furthermore, the study emphasized the importance of selecting appropriate input modalities based on crop type and geographical region. Nevertheless, the model required extensive preprocessing and high-quality satellite data. The computational demands associated with multimodal integration were also considerable. Additionally, prediction performance varied across different regions and crop varieties, indicating limitations in generalizability. The authors suggested future research involving adaptive multimodal architectures and explainable prediction mechanisms to improve scalability and practical deployment [24].

Lin et al. (2023) proposed MMST-ViT, a Climate Change-Aware Multi-Modal Spatial-Temporal Vision Transformer for crop yield prediction. The framework integrated satellite imagery, meteorological variables, and climate change indicators within a unified Transformer architecture. Experimental evaluation across multiple counties demonstrated that MMST-ViT outperformed conventional deep learning and machine learning approaches in predicting crop yields. The model effectively captured both short-term weather variations and long-term climate trends influencing agricultural productivity. The spatial and temporal Transformer components enhanced the framework's ability to learn regional dependencies and temporal dynamics. However, the study reported substantial computational requirements and dependence on high-quality remote sensing data. The complexity of Vision Transformer architectures also limited interpretability and deployment feasibility in resource-constrained environments. Furthermore, the model required large-scale datasets for training and validation. The authors recommended future research involving explainable AI integration and lightweight Transformer designs to improve usability and scalability in agricultural forecasting applications [25].

Bi et al. (2023) proposed a Transformer-based approach for early soybean yield prediction using time-series image data. The methodology employed Transformer architectures to analyze sequential agricultural imagery and extract temporal growth patterns associated with crop productivity. Experimental findings demonstrated that the model achieved high prediction accuracy even before the harvesting stage, enabling earlier agricultural planning and decision-making. The framework effectively captured temporal dependencies within image sequences and outperformed several conventional image-based prediction techniques. The study highlighted the potential of Transformer networks in remote sensing-based agricultural forecasting. However, implementation required extensive image datasets and substantial computational infrastructure. The model's performance was also sensitive to image quality and environmental variability. Additionally, the interpretability of Transformer-generated predictions remained limited. The researchers suggested combining explainable AI techniques with image-based Transformer models to improve transparency and user confidence in agricultural forecasting systems [26].

Nejadshamsi et al. (2025) proposed CYPRESS, a crop yield prediction framework utilizing a geospatial foundation model and satellite sensing data. The methodology adapted a pre-trained Earth observation model for continuous yield regression tasks and high-resolution agricultural monitoring. Experimental results demonstrated superior predictive performance compared to existing deep learning-based yield prediction systems. The framework generated detailed pixel-level yield maps, supporting precision agriculture and localized farm management decisions. Additionally, the study showcased the effectiveness of transfer learning and foundation models in agricultural forecasting applications. Despite these advantages, the framework required significant computational resources and access to large volumes of satellite imagery. Model training and fine-tuning were also computationally intensive. Furthermore, the deployment of foundation models in developing agricultural regions may be limited by infrastructure constraints. The authors recommended future work involving model compression, explainability enhancement, and broader validation across diverse agricultural ecosystems to improve practical adoption and scalability [27].

Research Gap -Despite significant progress in crop yield prediction research, several important gaps remain that limit the effectiveness and practical adoption of existing agricultural prediction approaches. Many previous studies primarily focus on improving prediction accuracy using specific datasets or limited experimental conditions, while insufficient attention has been given to model generalization across diverse agro-climatic regions. Most existing research relies on region-specific agricultural data, which restricts the applicability of developed prediction systems when applied to different environmental conditions or crop varieties. Additionally, earlier studies often emphasize algorithm performance comparisons without adequately addressing real-world deployment challenges such as data availability, scalability, and usability for farmers.

Another major research gap lies in the lack of standardized evaluation frameworks, making it

difficult to fairly compare results across different studies and datasets. Furthermore, several existing approaches depend heavily on historical data patterns and show limited capability in handling climate variability, extreme weather conditions, and rapidly changing agricultural environments. The integration of multi-source agricultural information, including soil health data, weather monitoring systems, and field-level observations, remains insufficiently explored in many studies. In addition, practical decision-support mechanisms that translate prediction outcomes into actionable recommendations for farmers are still underdeveloped. These limitations indicate a need for more comprehensive, adaptable, and application-oriented prediction frameworks that not only achieve high accuracy but also ensure transparency, scalability, and real-world usability. Addressing these research gaps can contribute to the development of reliable agricultural prediction systems capable of supporting sustainable farming practices and improving food security outcomes.

Table 1. Research Gap Identified from Literature Review

Identified Gap	Previous Studies
Limited model interpretability	Deep Learning, Transformer, and Foundation Models [14], [18], [21], [22], [25], [26], [27] act as black-box systems.
High computational complexity	Advanced DL, Transformer, and Ensemble approaches require significant computational resources [18], [21], [22], [23], [25], [27].
Lack of cross-regional validation	Several studies evaluated models on limited datasets and regions [17], [18], [20].
Dependency on feature engineering	Boosting models such as LightGBM and XGBoost require careful feature engineering [16], [20].
Limited multimodal integration	Few studies effectively combine climatic, soil, satellite, and temporal data simultaneously [24], [25].
Need for real-time prediction systems	Most models rely on historical datasets and lack real-time monitoring integration [19], [20], [23].
Scalability challenges	Traditional ML methods such as DT, LR, and SVM struggle with large-scale datasets [13], [15].

Table 2. Comparative Performance Analysis of Existing Studies

Author	Crop / Data Type	Models Used	Results	Limitations
Kumar et al. (2020) [13]	Agricultural data (rainfall, temperature, humidity, soil data)	Decision Tree, Logistic Regression	Interpretable prediction framework with satisfactory classification performance	Lower accuracy for complex nonlinear relationships; limited scalability
Rahman et al. (2023) [14]	Climatic, environmental and soil data	Multilayer Perceptron (MLP)	Higher prediction accuracy than conventional ML models	Black-box nature; high computational requirements
Singh et al. (2019) [15]	Historical agricultural datasets	SVM, ANN, Random Forest, Ensemble Methods	Ensemble techniques achieved superior accuracy and robustness	High computational cost and longer training time
Chen et al. (2024) [16]	Weather, soil, irrigation and fertilizer data	LightGBM, XGBoost	High R^2 values and low prediction errors	Requires extensive feature engineering and hyperparameter tuning
Sengaliappan & Bharathkumar (2025) [17]	Climatic and soil-related agricultural data	KNN, Decision Tree, Random Forest, Logistic Regression	Random Forest achieved highest prediction accuracy	Limited dataset size; advanced boosting methods not explored
Ibañez & Monterola (2023) [18]	Large-scale crop production and time-series data	Time Series Transformer	Superior forecasting accuracy and scalability across regions	High computational requirements and low interpretability

Pandit et al. (2023) [19]	Rabi crop yield and climatic variables	Hybrid Time-Series Forecasting Models	Improved forecasting accuracy through exogenous climatic factors	Dependence on quality climate data and historical trends
Nagesh et al. (2024) [20]	Soil, weather and crop yield data	Boosting-based Machine Learning Models	Enhanced prediction accuracy and feature importance analysis	Computationally intensive and parameter sensitive
Wang et al. (2024) [21]	Geospatial, climatic and temporal agricultural data	Temporal-Geospatial Deep Learning	Improved regional crop yield prediction accuracy	Requires large datasets and lacks explainability
Anoop et al. (2025) [22]	Climatic, soil and crop-specific data	3TFL-XLNet-CP (Transformer)	Outperformed traditional ML and DL techniques	High computational complexity and limited interpretability
Ravikumar et al. (2025) [23]	Agricultural climatic and soil datasets	SERWI Hybrid Ensemble Model	Higher robustness and predictive performance	Increased training time and reduced interpretability
Pathak et al. (2023) [24]	Satellite imagery, weather and soil data	Multimodal Machine Learning Framework	Improved field-level yield prediction accuracy	High preprocessing requirements and computational cost
Lin et al. (2023) [25]	Satellite imagery and climate data	MMST-ViT (Vision Transformer)	Superior performance under climate-aware prediction settings	Requires large-scale remote sensing datasets
Bi et al. (2023) [26]	Time-series agricultural image data	Transformer-based Image Analysis	Accurate early-stage soybean yield prediction	Sensitive to image quality and computationally expensive
Nejadshamsi et al. (2025) [27]	Satellite sensing and geospatial data	CYPRESS (Foundation Model)	High-resolution yield mapping with superior prediction accuracy	Requires extensive satellite data and infrastructure

A comparative study of Machine Learning (ML) and Deep Learning (DL) techniques reveals significant differences in their performance, complexity, and suitability for crop yield prediction and other agricultural applications. Traditional machine learning algorithms such as Logistic Regression and Decision Trees are widely used because of their simplicity, low computational requirements, and high interpretability. These models work efficiently on small and medium-sized datasets and allow researchers and farmers to easily understand the relationship between input variables such as rainfall, temperature, soil properties, and fertilizer usage. However, their predictive capability becomes limited when dealing with highly complex, nonlinear agricultural data patterns. To overcome these limitations, ensemble learning techniques such as Random Forest, XGBoost, and LightGBM have gained popularity. Ensemble methods combine multiple weak learners to create a stronger predictive model. By aggregating the outputs of several decision trees, these approaches reduce overfitting, improve generalization ability, and capture hidden relationships within agricultural datasets. As a result, ensemble models typically achieve higher prediction accuracy, precision, recall, and lower error metrics such as Root Mean Square Error (RMSE). These methods provide a balanced trade-off between interpretability and performance. Deep learning techniques, particularly Artificial Neural Networks (ANNs), further enhance prediction accuracy by automatically learning hierarchical and abstract feature representations from large and diverse datasets. Deep learning models are highly effective when working with large-scale agricultural data, satellite imagery, and time-series climate information. However, they require substantial training data, powerful computational resources, longer training times, and careful parameter tuning. Overall, classical machine learning models are more suitable for smaller datasets and applications requiring transparency, whereas deep learning

approaches offer superior scalability and performance for complex, large-scale agricultural prediction tasks, as illustrated in Table 1.

III. CROP YIELD PREDICTION USING ML MODELS

Crop yield prediction using Machine Learning (ML) models is an important application of artificial intelligence in precision agriculture. The main objective of crop yield prediction is to estimate agricultural production before harvesting by analyzing historical and environmental data. Accurate yield prediction helps farmers, agricultural planners, and policymakers make informed decisions regarding irrigation, fertilizer management, crop selection, storage, and food supply planning. Machine learning models analyze large amounts of agricultural data and identify hidden patterns among various factors affecting crop productivity.

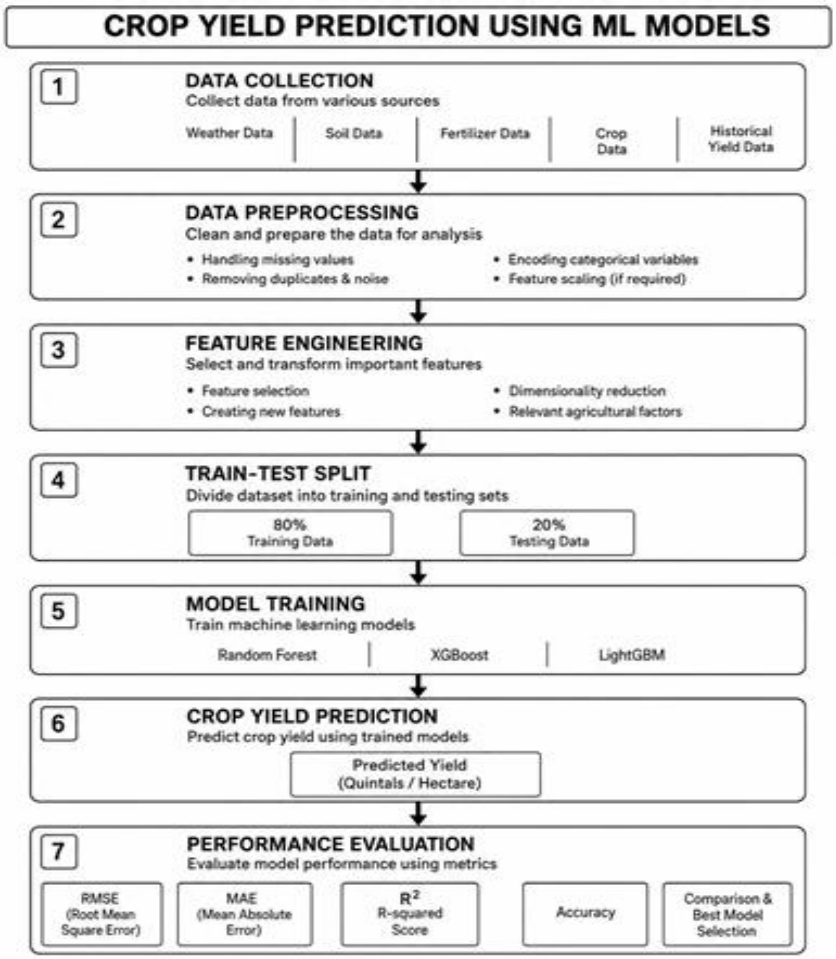

Figure 1. **Crop Yield Prediction Using ML**

These factors include temperature, rainfall, humidity, soil nutrients, pH value, fertilizer usage, irrigation methods, crop type, and weather conditions. Unlike traditional statistical methods, ML algorithms can effectively capture complex nonlinear relationships between these variables and provide more accurate predictions. In crop yield prediction, the problem is generally treated as a regression task where the target variable is crop yield measured in units such as tons per hectare or quintals per hectare. Various machine learning algorithms are widely used for this purpose, including Random Forest, XGBoost, LightGBM, Support Vector Machine, and Artificial Neural Networks. Among these, ensemble learning methods such as Random Forest, XGBoost, and LightGBM are highly effective because they combine multiple decision trees to improve prediction accuracy and reduce overfitting.

The prediction process begins with collecting agricultural datasets from sources such as weather stations, soil databases, and government agricultural records. The collected data is preprocessed to remove noise, handle missing values, and convert categorical features into numerical form. After preprocessing, the dataset is divided into

training and testing sets for model development and evaluation. Machine learning models learn from historical agricultural patterns and generate yield predictions for future crop seasons. The performance of these models is evaluated using metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 Score. Overall, crop yield prediction using ML models provides a scalable, reliable, and efficient solution for modern smart agriculture and sustainable farming practices.

Advantages of Machine Learning in Crop Yield Prediction Machine learning techniques provide several significant advantages in crop yield prediction by enabling accurate, data-driven agricultural analysis. One of the primary benefits is high prediction accuracy, as advanced algorithms analyze large volumes of historical agricultural data to identify hidden relationships among climatic conditions, soil characteristics, and crop growth factors. These models effectively capture complex and nonlinear interactions between multiple variables such as rainfall, temperature, soil nutrients, irrigation practices, and fertilizer application, which traditional statistical approaches often fail to represent accurately. By processing diverse datasets simultaneously, predictive models support data-driven decision making, assisting farmers, agricultural planners, and policymakers in optimizing crop selection, irrigation scheduling, fertilizer management, and risk assessment strategies [16]. Another important advantage is adaptability across different geographical regions, since models can be trained using region-specific datasets, allowing them to perform effectively under varying climatic zones and farming practices. Furthermore, modern predictive systems enable automation and scalability by efficiently analyzing large agricultural datasets collected from sensors, satellite imagery, and smart farming technologies [17].

This automation reduces manual effort while improving operational efficiency in precision agriculture. Machine learning models also provide feature importance analysis, which helps researchers and agricultural experts identify the most influential environmental and management factors affecting crop productivity, thereby supporting better scientific understanding and sustainable farming practices. In addition, early yield forecasting capability allows stakeholders to estimate production outcomes before harvesting, enabling improved planning for storage, supply chain management, and market strategies. Overall, these capabilities make data-driven predictive approaches highly valuable for modern agriculture by improving productivity, enhancing resource utilization, reducing uncertainty in crop production, and supporting sustainable agricultural development under changing environmental conditions.

Limitations of Machine Learning in Crop Yield Prediction Machine learning techniques have significantly improved crop yield prediction; however, several limitations still restrict their widespread adoption and practical implementation in agricultural systems. One of the major challenges is the strong dependence on data quality and availability. Machine learning models require large volumes of accurate, well-structured, and labeled agricultural datasets for effective training and prediction. In many agricultural regions, especially developing areas, data collection systems are inconsistent, incomplete, or contain missing and noisy records, which directly reduces prediction reliability and model performance. Another important limitation is the high computational requirement associated with advanced predictive models. Training complex algorithms often demands powerful computing infrastructure, high processing capability, and large memory resources. Such technological requirements may not be accessible to small-scale farmers or rural agricultural institutions, creating a gap between research development and real-world application.

Model interpretability also remains a significant concern. Many advanced predictive approaches operate as black-box systems, producing accurate results without clearly explaining how predictions are generated. This lack of transparency makes it difficult for farmers, agricultural experts, and policymakers to fully trust or understand the decision-making process behind yield estimations. As agriculture is a risk-sensitive sector, explainable predictions are essential for practical acceptance. Furthermore, machine learning models are prone to overfitting when trained on limited, biased, or region-specific datasets. In such situations, models may perform exceptionally well during training but fail to generalize under real-world conditions, reducing their practical usability across diverse agricultural environments.

Another limitation arises from the inability of predictive models to handle extreme or previously unseen climatic conditions effectively. Agricultural productivity is highly sensitive to unpredictable environmental factors such as droughts, floods, pest outbreaks, or sudden weather fluctuations. Since machine learning models rely heavily on historical patterns, they often struggle to provide accurate predictions when environmental conditions deviate significantly from past data distributions.

Additionally, implementing and maintaining predictive systems requires interdisciplinary technical expertise, including knowledge of data science, agriculture, programming, and domain-specific analytics. The lack of trained professionals in rural and farming communities makes deployment and long-term maintenance challenging [18]. Moreover, integration with existing agricultural practices and infrastructure presents practical difficulties. Farmers may lack access to digital tools, reliable internet connectivity, or sensor-based data collection systems necessary for continuous model operation. Ethical and data privacy concerns related to agricultural data sharing also create barriers to collaborative data-driven research. Overall, while machine learning has demonstrated strong potential in enhancing crop yield prediction accuracy and supporting precision agriculture, addressing challenges related to data quality, computational complexity, interpretability, generalization capability, technical expertise, and infrastructure readiness remains essential for achieving sustainable and large-scale adoption in modern agricultural systems.

IV. LIMITATION & FUTURE SCOPE

Previous research on crop yield prediction has made significant progress by applying various predictive and analytical approaches; however, several limitations remain evident across existing studies. Many earlier methods primarily focused on improving prediction accuracy using specific datasets without adequately addressing real-world agricultural variability. Traditional statistical and basic predictive models often failed to capture complex nonlinear relationships among environmental, soil, and crop management factors, resulting in limited prediction performance under dynamic farming conditions. Although advanced predictive models have demonstrated improved accuracy, many studies rely heavily on region-specific datasets, restricting their ability to generalize across diverse agro-climatic zones and crop varieties. Additionally, several existing approaches require extensive data preprocessing, feature engineering, and parameter tuning, increasing model complexity and limiting practical usability for farmers and agricultural practitioners. Computational requirements also present a major challenge, as high-performing models frequently depend on powerful hardware resources that may not be accessible in rural agricultural environments.

Another limitation involves model interpretability, where complex predictive frameworks operate as black-box systems, making it difficult for stakeholders to understand prediction reasoning and trust automated recommendations. Furthermore, many studies evaluate models using limited experimental conditions without considering long-term climate variability, extreme weather events, or real-time agricultural constraints. Data availability and quality issues, including missing values, inconsistent measurements, and lack of standardized datasets, further reduce the reliability of prediction outcomes[19]. In addition, integration of predictive models with practical decision-support systems remains insufficiently explored, creating a gap between theoretical research and field-level implementation. These limitations indicate the need for more adaptable, interpretable, and scalable prediction frameworks capable of operating under diverse agricultural conditions while supporting practical decision-making and sustainable farming practices.

Future research in crop yield prediction and agricultural analytics should focus on developing more robust, scalable, and practical prediction frameworks that can operate effectively under diverse environmental and farming conditions. One important direction is the integration of heterogeneous agricultural data sources such as remote sensing imagery, weather monitoring systems, soil health records, and real-time field sensor information to improve prediction reliability. Future studies should also emphasize region-independent models capable of generalizing across different agro-climatic zones, since many existing approaches remain limited to specific geographical datasets.

Another promising research direction involves improving model interpretability so that prediction results become transparent and easily understandable for farmers and agricultural decision-makers. Explainable prediction systems can enhance trust and encourage practical adoption in real farming environments. Additionally, future work should explore adaptive prediction mechanisms that can respond to dynamic climate variability, extreme weather events, and long-term environmental changes. Incorporating climate resilience into agricultural prediction frameworks will help reduce uncertainty and improve food security planning. Research should also focus on developing lightweight and cost-effective computational solutions suitable for rural areas where technological infrastructure and computational resources are limited. Furthermore, collaborative data-sharing platforms and standardized agricultural datasets should be encouraged to overcome data scarcity issues and enable large-scale comparative studies. Integration of decision-support systems with mobile

and smart farming applications may further enhance accessibility for farmers. Overall, future research should aim to bridge the gap between theoretical prediction models and practical agricultural implementation by prioritizing usability, scalability, transparency, and real-world applicability.

V. CONCLUSION

This paper presents a comprehensive review and comparative analysis of modern data-driven approaches for crop yield prediction and classification within precision agriculture systems. The study emphasizes that accurate crop yield estimation is essential for improving agricultural productivity, optimizing resource utilization, and supporting informed decision-making in farming practices. Traditional statistical techniques have historically been used for agricultural forecasting; however, their ability to model complex and nonlinear interactions among climatic conditions, soil characteristics, irrigation practices, and crop management factors remains limited. In contrast, contemporary predictive approaches demonstrate improved capability in capturing intricate relationships within large and heterogeneous agricultural datasets. Classical predictive models provide advantages such as simplicity, transparency, and ease of interpretation, yet they often struggle to achieve high prediction accuracy when dealing with complex environmental variability.

Ensemble-based techniques, including Random Forest, XGBoost, and LightGBM, show superior performance due to their ability to reduce overfitting, enhance robustness, and effectively learn from diverse data sources. These approaches offer a balanced combination of predictive accuracy, scalability, and computational efficiency, making them suitable for real-world agricultural applications. Advanced deep learning techniques further enhance prediction performance by automatically extracting complex feature representations from large datasets; however, their effectiveness depends heavily on the availability of extensive training data and significant computational resources, which may limit practical deployment in resource-constrained environments. Overall, the comparative findings indicate that ensemble learning frameworks provide a practical and reliable solution for crop yield prediction by maintaining strong generalization capability while supporting scalable agricultural analytics. The study concludes that integrating efficient predictive models with real-world agricultural decision-support systems can contribute to sustainable farming practices, improved food security, and enhanced resilience of agricultural systems under changing environmental and climatic conditions.

REFERENCE

[1] S. M. Shawon, F. Barua Ema, and A. K. Mahi, “Crop yield prediction using machine learning: An extensive and systematic literature review,” Smart Agricultural Technology, vol. 10, Mar. 2025, Art. no. 100718, pp. 1–12, doi:10.1016/j.atech.2024.100718.
[2] P. S. Vijayabaskaran, “Crop yield forecasting using machine learning and deep learning approaches: A comprehensive review,” Int. J. Communication & Computer Technologies, vol. 13, no. 2, pp. 83–91, Aug. 2025.
[3] 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, Oct. 2020, Art. no. 105709, pp. 1–14,
[4] A. P. Kamath, P. Patil, and S. S. Sushma, “Crop yield forecasting using data mining,” Global Transitions Proceedings, vol. 2, no. 2, pp. 402–407, 2021.
[5] N. Bali and A. Singla, “Deep learning based wheat crop yield prediction model in Punjab region of north India,” Applied Artificial Intelligence, vol. 35, no. 15, pp. 1304–1328, 2021.
[6] M. Kuradusenge et al., “Crop yield prediction using machine learning models: A case of Irish potato and maize,” Agriculture, vol. 13, no. 1, Art. no. 225, pp. 1–17, 2023.
[7] J. Shook, T. Gangopadhyay, L. Wu, B. Ganapathysubramanian, and S. Sarkar, “Crop yield prediction integrating genotype and weather variables using remote sensing approach,” Int. J. Applied Earth Observation and Geoinformation, vol. 113, Art. no. 102959, pp. 1–13.
[8] S. Khaki and L. Wang, “Crop yield prediction using deep neural networks,” Frontiers in Plant Science, vol. 10, Art. no. 621, pp. 1–12, 2019.
[9] A. M. El-Kenawy, M. Abdelhamid, and S. Ibrahim, “Machine learning and deep learning models for potato crop yield prediction,” *Computers and Electronics in Agriculture*, vol. 214, pp. 108–125, 2025.
[10] X. Li, Y. Zhang, and H. Wang, “Wheat yield prediction using remote sensing data and ensemble learning methods,” *Agricultural Systems*, vol. 203, pp. 103–118, 2023.
[11] P. Sharma, R. Verma, and S. Gupta, “Rice yield estimation using artificial neural networks and weather parameters,” *Journal of Agricultural Informatics*, vol. 13, no. 2, pp. 45–56, 2022.
[12] L. Zhang, Q. Chen, and M. Liu, “Support vector machine-based crop classification using multi-source data,” *Remote Sensing Applications: Society and Environment*, vol. 24, pp. 100–112, 2021.
[13] R. Kumar, S. Patel, and N. Singh, “Comparative study of decision tree and logistic regression for crop yield classification,” *International Journal of*

Computer Applications, vol. 176, no. 12, pp. 18–24, 2020.
[14] M. Rahman, T. Hossain, and A. Islam, "Deep learning framework for crop yield prediction using multilayer perceptrons," *IEEE Access*, vol. 11, pp. 45678–45690, 2023.
[15] A. Singh, P. Kaur, and J. Kaur, "Performance comparison of machine learning algorithms for crop yield prediction," *Procedia Computer Science*, vol. 165, pp. 232–239, 2019.
[16] Y. Chen, Z. Zhou, and F. Li, "Crop yield regression using LightGBM and XGBoost models," *Applied Artificial Intelligence*, vol. 38, no. 1, pp. 1–15, 2024.
[17] S. Sengaliappan and R. Bharathkumar, "Crop Yield Prediction Using Machine Learning Approaches," International Journal of Scientific Research in Engineering and Management (IJSREM), vol. 9, no. 5, pp. 1–7, 2025.
[18] S. C. Ibañez and C. P. Monterola, "A Global Forecasting Approach to Large-Scale Crop Production Prediction with Time Series Transformers," Agriculture, vol. 13, no. 9, pp. 1–22, 2023.
[19] P. Pandit et al., "Hybrid Time Series Models with Exogenous Variable for Improved Yield Forecasting of Major Rabi Crops in India," Scientific Reports, vol. 13, Art. no. 22240, 2023.
[20] O. S. Nagesh et al., "Boosting Enabled Efficient Machine Learning Technique for Accurate Prediction of Crop Yield Towards Precision Agriculture," Discover Sustainability, vol. 5, Art. no. 78, 2024.
[21] L. Wang, Z. Chen, W. Liu, and H. Huang, "A Temporal–Geospatial Deep Learning Framework for Crop Yield Prediction," Electronics, vol. 13, no. 21, pp. 1–24, 2024.
[22] G. L. Anoop, C. Nandini, and E. Naresh, "3TFL-XLNet-CP: A Novel Transformer-Based Crop Yield Prediction Framework with Weighted Loss Based 3-Tier Feature Learning Model," SN Computer Science, vol. 6, Art. no. 275, 2025.
[23] A. Ravikumar et al., "A Hybrid SERWI Ensemble Model for Crop Yield Prediction Using an Inverse RMSE Weighting Strategy," Scientific Reports, vol. 15, Art. no. 45085, 2025.
[24] D. Pathak et al., "Predicting Crop Yield With Machine Learning: An Extensive Analysis of Input Modalities and Models on a Field and Sub-Field Level," arXiv preprint arXiv:2308.08948, 2023.
[25] F. Lin et al., "MMST-ViT: Climate Change-Aware Crop Yield Prediction via Multi-Modal Spatial-Temporal Vision Transformer," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2023, pp. 1–10.
[26] L. Bi et al., "A Transformer-Based Approach for Early Prediction of Soybean Yield Using Time-Series Images," Frontiers in Plant Science, vol. 14, 2023.
[27] S. Nejadshamsi et al., "CYPRESS: Crop Yield Prediction via Regression on Prithvi's Encoder for Satellite Sensing," arXiv preprint arXiv:2510.26609, 2025.