A Robust and Scalable Machine Learning Approach for Crop Yield Prediction Using Advanced Ensemble Models

***Manvendra Singh**
****Dr. Sanjeev Kumar Sharma**
*****Dr. Subhash Mishra**
******Saurabh Karsoliya**

ABSTRACT

Accurate crop yield prediction is a critical component of precision agriculture, enabling efficient resource management, improved food security, and informed decision-making for farmers and policymakers. Increasing climate variability and heterogeneous agricultural conditions have reduced the effectiveness of traditional statistical yield estimation techniques. To address these challenges, this paper presents a robust and scalable machine learning framework for crop yield prediction using advanced ensemble learning methods. The proposed study evaluates the performance of Random Forest, XGBoost, and Light Gradient Boosting Machine (LightGBM) on a large-scale agricultural dataset comprising climatic, soil, crop, and management-related attributes. A unified preprocessing and modeling pipeline is employed to ensure fairness and reproducibility. Both regression-based yield estimation and classification-based yield categorization are investigated to provide a comprehensive performance assessment. Experimental results demonstrate that LightGBM consistently outperforms other models, achieving a high coefficient of determination and superior classification accuracy. The findings confirm the effectiveness of gradient boosting techniques for large-scale agricultural analytics and highlight their suitability for real-world precision agriculture applications.

Keywords:- Crop Yield Prediction, Machine Learning, Ensemble Learning, Gradient Boosting, LightGBM, Precision Agriculture**.**

* Manvendra Singh, Technocrats Institute of Technology CSE, iammanvendrasingh@gmail.com.
**Dr. Sanjeev Kumar Sharma, Technocrats Institute of Technology CSE, spd0@gmail.com.
***Dr Subhash Mishra, Technocrats Institute of Technology CSE, subhashmishra67@gmail.com.
****Saurabh Karsoliya, Technocrats Institute of Technology CSE, karsoliya.saurabh@gmail.com.

I. INTRODUCTION

Agriculture plays a crucial role in sustaining global food security and economic stability, particularly in developing countries where a significant portion of the population depends on farming for livelihood. Accurate crop yield prediction is a key factor in effective agricultural planning, efficient resource utilization, and risk management. Reliable yield forecasting enables farmers and policymakers to make informed decisions related to irrigation scheduling, fertilizer application, crop selection, and market supply planning. However, increasing climate variability, unpredictable weather patterns, and heterogeneous soil conditions have made crop yield estimation a complex and challenging task[1]. Traditional crop yield prediction methods are

primarily based on historical averages, empirical models, or simple statistical techniques. While these approaches are easy to implement, they often fail to capture the complex, nonlinear relationships among climatic factors, soil properties, crop characteristics, and management practices. As a result, their prediction accuracy and generalization capability remain limited, especially when applied to large and diverse agricultural datasets. With the growing availability of agricultural data from multiple sources, there is a strong need for advanced computational techniques that can efficiently process large-scale data and provide accurate yield predictions. In recent years, machine learning (ML) techniques have emerged as powerful tools for agricultural analytics due to their ability to model nonlinear patterns and interactions among multiple variables. Among various ML approaches, ensemble learning methods have gained significant attention for crop yield prediction. Ensemble models combine multiple base learners to improve prediction accuracy and robustness while reducing overfitting. Random Forest is one of the most widely used ensemble techniques and has shown promising results in agricultural prediction tasks because of its ability to handle highdimensional and heterogeneous data. Gradient boosting algorithms further enhance ensemble learning by iteratively minimizing prediction errors through optimized tree construction. XGBoost has demonstrated high predictive accuracy and computational efficiency by incorporating regularization and parallel processing. More recently, Light Gradient Boosting Machine (LightGBM) has been developed to address scalability challenges associated with large datasets [2]. By employing histogram-based learning and leaf-wise tree growth strategies, LightGBM significantly reduces training time and memory consumption while maintaining high prediction accuracy. This paper proposes a robust and scalable machine learning framework for crop yield prediction using ensemble and gradient boosting techniques. The study evaluates the performance of Random Forest, XGBoost, and LightGBM on a large-scale agricultural dataset using both regression-based yield estimation and classification-based yield categorization. By providing a comprehensive comparative analysis, the proposed work aims to demonstrate the effectiveness of advanced ensemble learning models for real-world precision agriculture applications. In recent years, deep learning techniques have emerged as powerful alternatives to traditional machine learning models for agricultural prediction tasks. Among these, Convolutional Neural Networks (CNNs) have demonstrated strong capability in learning complex feature representations from high-dimensional data. Originally developed for image analysis, CNN architectures have been successfully adapted for structured, spatial, and time-dependent agricultural datasets. The convolutional layers in CNNs enable automatic extraction of hierarchical feature patterns, while pooling operations enhance

generalization by reducing dimensional redundancy[4]. Unlike conventional machine learning models that rely heavily on manual feature engineering, CNNbased approaches learn discriminative features directly from raw input data, thereby improving modeling efficiency and predictive performance. In crop yield prediction, CNN models can capture intricate relationships among climatic variables, soil characteristics, and crop management factors, especially when spatial or multi-dimensional data representations are involved. Furthermore, CNN architectures can be integrated with other predictive frameworks to enhance robustness and scalability. By incorporating CNN into the proposed framework, this study enables a systematic comparison between advanced ensemble learning methods and deep learning-based approaches. Such integration not only broadens the analytical perspective but also strengthens the reliability and generalization capability of the crop yield prediction system for large-scale precision agriculture applications.

II. RELATED WORK

Machine learning-based crop yield prediction has gained increasing attention due to its ability to model complex agricultural systems. Several studies have applied traditional ML classifiers such as Decision Trees, K-Nearest Neighbors, and Random Forest to estimate crop productivity using soil and climatic data. While these approaches offer interpretability and moderate accuracy, their scalability and predictive performance remain limited for large datasets. Recent research has explored gradient boosting and deep learning techniques to overcome these limitations. Gradient boosting models have demonstrated improved accuracy and robustness by sequentially correcting prediction errors. However, many existing studies are restricted to moderate-sized datasets, singlecrop analysis, or classification-only evaluation. Deep learning approaches, including CNN and LSTM-based architectures, have also been proposed, particularly for time-series and remote sensing data. Despite achieving high accuracy, these models often require extensive computational resources and complex data acquisition processes. The limitations identified in existing literature highlight the need for scalable, efficient, and accurate machine learning frameworks capable of handling large-scale agricultural data while supporting both regression and classification tasks. This study addresses these gaps by leveraging advanced ensemble learning techniques within a unified experimental framework. El-Kenawy et al. (2025) explored the application of machine learning and deep learning techniques for potato crop yield prediction using climatic and soilrelated features. Their comparative analysis of Random Forest, Gradient Boosting, and LSTM models reported strong predictive performance. However, the study was confined to a single crop type, and issues related to scalability and computational efficiency on

large-scale datasets were not examined in depth [2]. Jabed et al. (2024) provided an extensive review of machine learning and deep learning methods for crop yield prediction, covering models such as Random Forest, CNN, LSTM, and ensemble-based approaches across diverse datasets. Although the review highlighted commonly used evaluation metrics, including accuracy and R^2, it lacked experimental validation and did not propose a unified framework for large-scale implementation [3]. Nagesh et al. (2024) proposed a boosting-based machine learning approach for crop yield classification using soil and climatic parameters. Their AdaBoost model achieved high accuracy and precision; however, the study was limited to classification tasks and did not consider regression-based yield estimation. Furthermore, advanced gradient boosting algorithms such as LightGBM were not included in the analysis [4]. Pukrongta et al. (2024) introduced an IoT-enabled machine learning framework for maize yield prediction by integrating sensor-derived environmental data with historical yield records. While the system achieved promising classification results, its heavy reliance on IoT infrastructure reduced its applicability to regions lacking sensor-based data, and scalability to conventional datasets was not addressed [5]. Kalmani et al. (2025) developed a hybrid CNN–LSTM deep learning model for crop yield prediction using time-series climatic data. Although the proposed model outperformed traditional machine learning techniques, it required substantial computational resources and extended training time. In addition, efficient ensemble learning methods such as XGBoost and LightGBM were not evaluated [6]. Vijayabaskaran (2025) reviewed recent advancements in machine learning and deep learning techniques for crop yield forecasting, emphasizing the performance improvements achieved through ensemble models. Despite these insights, the study did not provide experimental benchmarking using a common dataset or analyze scalability for large agricultural data [7]. Shawon et al. (2024) conducted a systematic review of machine learning-based crop yield prediction studies, summarizing reported performance metrics such as accuracy, RMSE, and R^2. While ensemble models were identified as consistently high-performing, the absence of empirical comparisons highlighted the need for large-scale experimental validation [8]. Bondre and Mahagaonkar (2022) applied Random Forest and traditional machine learning algorithms for crop yield prediction and fertilizer recommendation using regional datasets. Although the models demonstrated moderate accuracy and interpretability, the limited dataset size and lack of advanced boosting techniques constrained scalability and robustness [9]. Sun et al. (2022) employed deep CNN–LSTM architectures with remote sensing data to capture spatial and temporal patterns in crop yield prediction. Despite achieving strong predictive performance, the dependence on satellite imagery increased data complexity and cost, and structured tabular data-

based machine learning approaches were not investigated [10]. Shastry and Sanjay (2023) proposed a hybrid ensemblebased machine learning framework using climatic and soil parameters, which improved prediction stability. However, modern gradient boosting techniques such as LightGBM were not incorporated, and large-scale data efficiency was not evaluated [11]. Sengaliappan and Bharathkumar (2025) examined crop yield prediction using structured agricultural data with traditional machine learning algorithms, including KNN, Decision Tree, Random Forest, and Logistic Regression. While improved classification performance was reported, the study relied on moderate-sized datasets and did not explore advanced gradient boosting models or regressionbased yield prediction [1].

Table1.Comparative Analysis of Existing Crop Yield Prediction Studies

Ref.	Authors(Year)	Dataset / Crop Type	Methods Used	Limitations
[1]	Sengaliappan & Bharath kumar (2025)	Structured agricultural data	KNN, Decision Tree, Random Forest, Logistic Regression	Moderate dataset size; advanced boosting models (XGBoost, LightGBM) not explored; Regression not considered
[2]	El-Kenawyet al. (2025)	Potato crop, climatic & Soil data	Random Forest, Gradient Boosting, LSTM	Limited to a single crop; scalability and computational efficiency not analyzed
[3]	Jabedetal. (2024)	Multiple agricultural datasets	RF, CNN, LSTM, Ensemble models (Review)	Review-based study; no experimental validation or unified framework
[4]	Nageshet al. (2024)	Soil & climatic data	AdaBoost	Regression-based prediction not included; LightGBM not evaluated
[5]	Pukrongtaet al. (2024)	Maize yield with IoTsensor data	Ensemble ML models	Heavy reliance on IoT infrastructure; limited general applicability
[6]	Kalmaniet al. (2025)	Time-series climatic data	CNN–LSTM hybrid	High computational cost; ensemble boosting models not considered
[7]	Vijaya baskaran (2025)	Various agricultural datasets	RF, Gradient Boosting, CNN, RNN (Review)	No benchmarking on common dataset; scalability not discussed

[8]	Shawonet al. (2024)	Multiple crop yield studies	ML-based models (Review)	No empirical comparison; large-scale validation lacking
[9]	Bondre & Mahagaonkar (2022)	Regional agricultural data	Random Forest, traditional ML	Small dataset size; advanced Boosting techniques not explored
[10]	Sunetal. (2022)	Remote sensing data	CNN–LSTM	High data complexity and cost; tabular ML models not evaluated
[11]	Shastry & Sanjay(2023)	Climatic& soil parameters	Hybrid ensemble ML	LightGBM not included; large-scale efficiency not addressed

III. PROPOSED FRAMEWORK

The proposed framework aims to provide an efficient and scalable solution for crop yield prediction by integrating advanced ensemble learning algorithms with structured agricultural data. The framework consists of data preprocessing, model training and performance evaluation stages. Random Forest is employed as a baseline ensemble model, while XGBoost and LightGBM are utilized to capture complex nonlinear relationships and improve scalability. Both regression and classification perspectives are considered. For regression analysis, the models predict continuous yield values, while for classification analysis, yield values are categorized into high- and low-yield classes. This dual evaluation strategy enables a more comprehensive assessment of model performance and practical applicability in decision-support systems as shown blow The overall architecture of the proposed crop yield prediction framework is illustrated in Figure 1.

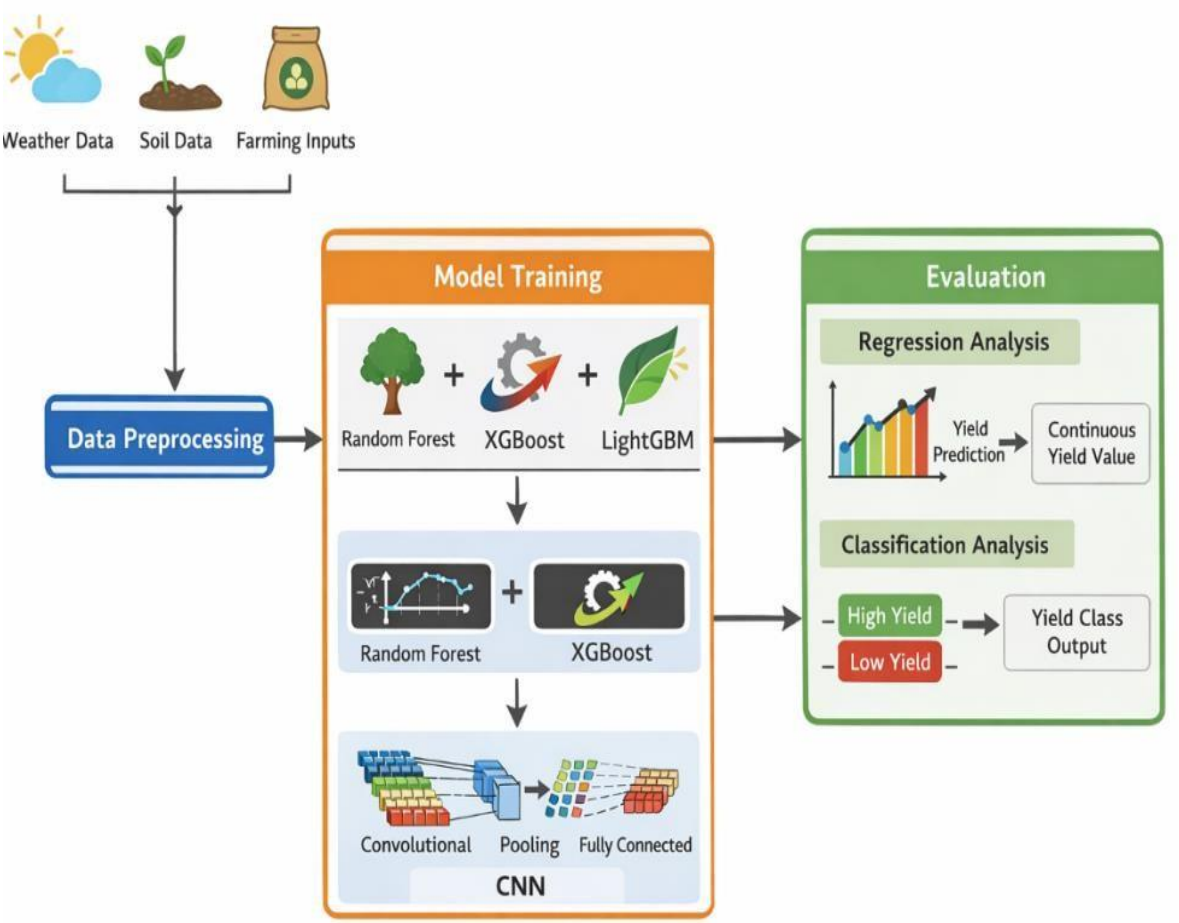

Figure 1. Architecture of the Proposed Crop Yield Prediction Framework

IV. METHODOLOGY

4.1. Dataset Description

The dataset used in this study is sourced from a publicly available agricultural repository and contains one million records with multiple attributes related to crop production. The features include region, soil type, crop type, rainfall, temperature, fertilizer usage, irrigation status, weather condition, and days to harvest. Crop yield measured in tons per hectare serves as the target variable.

4.2. Data Preprocessing

A structured preprocessing pipeline is employed to handle both categorical and numerical features. Categorical variables are encoded using one-hot encoding, while numerical attributes are processed directly due to the scale-invariant nature of tree-based models. A unified preprocessing pipeline ensures consistency across all experiments and prevents data leakage.

4.3.Model Training and Evaluation

The dataset is divided into training and testing subsets using an 80:20 split. Random Forest, XGBoost, and LightGBM models are trained within a common experimental pipeline. Regression performance is evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2). Classification performance is assessed using accuracy, precision, recall, and F1-score.

V. CONCLUSION

This paper presented a robust and scalable machine learning framework for crop yield prediction using ensemble and gradient boosting techniques. By evaluating Random Forest, XGBoost and LightGBM on a large-scale agricultural dataset, the study demonstrated the superior performance of LightGBM in both regression and classification tasks. The proposed framework addresses key limitations in existing studies by supporting large-scale data analysis and dual-task evaluation. The findings highlight the potential of advanced ensemble learning models for real-world precision agriculture applications and provide a strong foundation for future research in data-driven agricultural decision-making.

REFERENCES

1. 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.
2. E. M. El-Kenawy, A. A. Alhussan, N. Khodadadi, S. Mirjalili, and M. M. Eid, “Predicting Potato Crop Yield with Machine Learning and Deep Learning for Sustainable Agriculture,” Potato Research, vol. 68, pp. 759– 792, 2025, doi: 10.1007/s11540-024-09753-w.
3. M. A. Jabed, M. A. A. Murad, M. A. Rahman, and M. S. Islam, “Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches,” Heliyon, vol. 10, no. 24, p. e40836, 2024, doi: 10.1016/j.heliyon.2024.e40836.
4. O. S. Nagesh, P. K. Reddy, and S. Kumar, “Boosting enabled efficient machine learning technique for accurate prediction of crop yield towards precision agriculture,” Discover Sustainability, vol. 5, art. 78, 2024, doi: 10.1007/s43621-024-00254-x.
5. N. Pukrongta, T. Sattayasai, and P. Rakwatin, “Enhancing Crop Yield Predictions with PEnsemble 4: IoT and Machine Learning Integration,” Applied Sciences, vol. 14, no. 8, art. 3313, 2024, doi: 10.3390/app14083313.
6. V. H. Kalmani, N. V. Dharwadkar, and V. Thapa, “Crop Yield Prediction Using Deep Learning Algorithm Based on CNN-LSTM with Attention Layer and Skip Connection,” Indian Journal of Agricultural Research, vol. 59, no. 8, pp. 1303–1311, 2025, doi: 10.18805/IJARe.A-6300.
7. V. P. S. Vijayabaskaran, “Crop Yield Forecasting Using Machine Learning and Deep Learning Approaches: A Comprehensive Review,” International Journal of Communication and Computer Technologies, vol. 13, no. 2, pp. 83–91, 2025.
8. S. M. Shawon, M. R. Hasan, and M. T. Rahman, “Crop yield prediction using machine learning: An extensive and systematic literature review,” AI and Agriculture, vol. 9, pp. 100–118, 2024, doi: 10.1016/j.aiia.2024.100118.

9. D. Sun, X. Wang, and Y. Zhang, "Crop yield prediction using CNN-LSTM models with remote sensing data," Remote Sensing, vol. 14, no. 6, art. 1421, 2022, doi: 10.3390/rs14061421.
10. A. Bondre and A. Mahagaonkar, "Prediction of crop yield and fertilizer recommendation using machine learning techniques," International Journal of Engineering Research & Technology, vol. 11, no. 4, pp. 512–518, 2022.
11. Kaggle, "Crop Yield Prediction Dataset," 2024. [Online]. Available: https://www.kaggle.com/. Accessed: Jan. 2025. Yadav, V., & Mahajan, P. Optimizing Loan Approval Processes with XGBoost and Ensemble Models. Journal of Artificial Intelligence and Applications, Vol. 10, Issue 1, pp. 55–7
12. A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," Computers and Electronics in Agriculture, vol. 147, pp. 70–90, 2018, doi: 10.1016/j.compag.2018.02.016.
13. R. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," Sensors, vol. 18, no. 8, art. 2674, 2018, doi: 10.3390/s18082674.
14. J. Jeong, J. P. Resop, N. D. Mueller et al., "Random forests for global and regional crop yield predictions," PLOS ONE, vol. 11, no. 6, p. e0156571, 2016, doi: 10.1371/journal.pone.0156571.
15. P. Priya, U. M. Reddy, and K. S. Rao, "Crop yield prediction using machine learning algorithms," International Journal of Engineering & Technology, vol. 7, no. 2.8, pp. 170–174,
16. A. Cristea, D. Ioannou, and G. Georgopoulos, "Crop yield prediction using XGBoost and environmental data," Procedia Computer Science, vol. 192, pp. 1571–1580, 2021, doi: 10.1016/j.procs.2021.09.329.
17. M. Khaki, L. Wang, and S. Archontoulis, "A CNN–RNN framework for crop yield prediction," Frontiers in Plant Science, vol. 10, art. 1750, 2019, doi: 10.3389/fpls.2019.01750.
18. H. Shahhosseini, G. Hu, and S. Archontoulis, "Forecasting corn yield using ensemble machine learning," Agricultural and Forest Meteorology, vol. 252, pp. 9–19, 2018, doi: 10.1016/j.agrformet.2018.01.012.
19. S. M. Rahman, M. Hasanuzzaman, and M. A. Rahman, "An ensemble learning approach for crop yield prediction," International Journal of Advanced Computer Science and Applications, vol. 10, no. 6, pp. 1–8, 2019.
20. Y. Chen, L. Zhang, and Q. Yang, "Large-scale crop yield prediction using gradient boosting decision trees," Information Processing in Agriculture, vol. 8, no. 2, pp. 306–316, 2021, doi: 10.1016/j.inpa.2020.07.004. 2018.
21. S. You, J. Zhang, L. Xiong, and J. Sun, "Crop yield prediction based on machine learning and multisource data fusion," IEEE Access, vol. 8, pp. 100369–100381, 2020, doi: 10.1109/ACCESS.2020.2997707.