Exploring Machine Learning and Ensemble Methods for Crop Yield Prediction: A Review

*Bhumika Tiwari[1], Navneet Kaur[2], and Paurav Goel[3]

[1,2,3]Department of CSE, Chandigarh University, Punjab, India

[1]tiwaribhumika2002@gmail.com, [2]navneetsehgal5@gmail.com,
[3]pauravg5986@gmail.com

Abstract. The population growth rate all over the world, especially in countries like India, has increased the pressure on food production. Hence, there has been an increasing need for advancements in agriculture. Current technologies like data mining, machine learning (ML), remote sensing, and image analysis offer accurate ways of measuring the cultivated land and the expected agricultural harvest. By using factors like temperature, precipitation, soil type, and planted area, it is possible to develop machine learning and advanced ensemble learning models that reliably predict crop yield. This study takes a comprehensive look at the application of supervised, unsupervised, and ensemble machine learning models in agronomic yield forecasting. Studying this deeply proved that classification methods, especially those based on ensemble models, perform better than regression and unsupervised models in agricultural production forecasting. Agricultural management can benefit from these results as well, as they demonstrate how ensemble learning can be used to enhance prediction accuracy in the context of ever-increasing food requirements.

Keywords: Crop Yield Prediction, Machine Learning, Ensemble Learning, Classification methods, Supervised Learning, Unsupervised learning, Regression Methods, Bagging, Boosting

1 Introduction

Agriculture is a major contributor to economic growth and a powerful solution to global food security challenges. Crop forecasting is crucial to optimize agricultural output and food availability, as well as inform prudent policy decisions. Accurate estimations can reduce problems associated with unstable food supply, wasteful resource allocation, and excess caused by extreme climate or plant disease [1]. Many factors, including rainfall, soil type, fertilization habits, and pest control, greatly influence a farm's yield. However maximizing production must also consider economic limitations, rural community social realities, and long-term environmental responsibility. Although technology can help model intricate cropping systems, local knowledge is still necessary to test theories and apply locally relevant solutions. Contemporary farmers, especially those in developing countries, are employing new methods to transform agriculture from an old-fashioned profession into a more analytical, efficient enterprise [2].

N. Vyas et al. (eds.), *Proceedings of the International Conference on Advancements in Computing Technologies and Artificial Intelligence (COMPUTATIA 2025)*, Advances in Intelligent Systems Research 189,
https://doi.org/10.2991/978-94-6463-700-7_6

Ensemble methods have been crucial to developing the application of statistical learning in identifying forecasting insights from huge agricultural data sets. Contemporary study on crop production estimates has become feasible due to the availability of big data in the agriculture industry. Random Forest, Decision Trees, and Neural Networks are some of the machine learning methods that have been extensively used to predict intricate relationships between climate, soil, and agricultural data [3]. Ensemble learning methods have greatly enhanced prediction accuracy and stability by aggregating predictions from numerous models. With their capacity to process numerous datasets and uncover latent patterns, these approaches outperform conventional statistical models [4].

In addition, the integration of machine learning with IoT and remote sensing technologies has enhanced real-time data collection and analysis, strengthening agricultural production forecasting systems. Such advancements not only enable farmers to make more informed resource allocation decisions, but also assist in tackling challenges like climate change, insect infestations, and soil degradation [5]. This research discusses recent advances in agricultural production forecasting with a focus on the relative merits of machine learning and ensemble approaches. The research outlines a typical architecture for agricultural output forecasting systems, emphasizes algorithmic advancements, and presents a comparative analysis to determine the most effective techniques for future study and application.

2 Crop Yield Prediction System

2.1 Input Data

The typical input data consists of meteorological parameters (rainfall, humidity, and temperature), agronomic factors, soil characteristics, and past yield data. Satellite imagery, weather stations, and farm records are some of the numerous potential sources of such data. Use various factors to retrieve information from the meteorological, agricultural, and historical departments.

2.2 Data Preprocessing

This is the critical procedure by which the data becomes clean, consistent, and usable. Preprocessing involves handling missing values, removing outliers, normalizing or scaling numerical attributes, and categorical data encoding.

2.3 Data Splitting

To create and assess a machine learning model, the dataset is divided into sub-sets for testing and training. The model is built using the training set, and its performance is evaluated using the testing set. The 70-30 and 80-20 splitting ratios are frequently used. In this stage, overfitting is prevented and the model is ensured to generalize effectively to new, and like unseen data.

2.4 Modeling

This phase involves selecting an appropriate agricultural Yield Estimation through machine learning. The selected model is trained on the training dataset, helping it to discover patterns and correlations in the data.

2.5 Pattern Visualization

This phase involves visualizing patterns to predict results. The predictive rules help to visualize these patterns, aiding in accurate predictions.

2.6 Model Evaluation

The performance is evaluated against the expected outcomes and the samples that were supplied in the agricultural knowledge base. To determine the efficiency of the anticipated outcomes, the farmer or expert may need to interact. Fig. 1 displays the overall architecture of agricultural yield prediction as a block diagram.

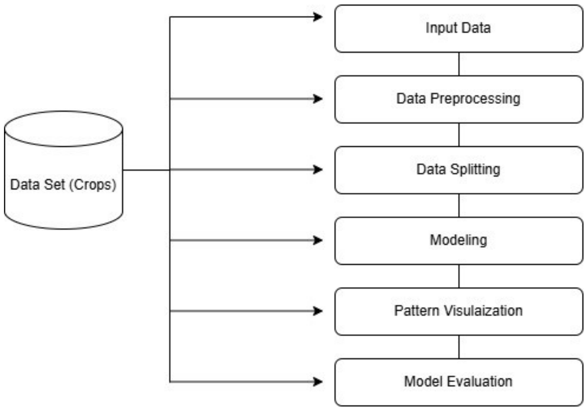

Fig. 1. Architecture of agricultural yield prediction

3 Literature Survey

To identify machine learning and ensemble learning models for agricultural output prediction, a comprehensive literature study has been conducted. We looked at several supervised, unsupervised, and advanced machine-learning algorithms created for agricultural yield prediction. A summary of the review that was done on the gathered publications is provided in Table 1. The next subsections discuss several supervised learning models for regression and classification, unsupervised learning models for association and clustering, and ensemble learning models for yield prediction using bagging and boosting.

3.1 Supervised Algorithms

Supervised machine learning techniques rely on labeled data for training algorithms that categorize inputs or forecast results based on established associations. Using examples given during training, it maps inputs to outputs. Classification and regression are the two Key supervised machine-learning techniques employed in crop yield prediction.

Classification: Yield prediction for a new crop can be accomplished using classification models that categorize crop yields based on historical data or information about similar crops. In Table 1, various classification techniques, including K-Nearest Neighbors (K-NN), Logistic Regression (LR), and Neural Networks (ANN), have been explored for this purpose [6]. Experimental findings indicate that ANN offers superior prediction accuracy for crops like rice, potato, and wheat, while linear regression demonstrates better performance for crops such as Boro and Amon. For rice crops, a multilayer perceptron neural network [7] provides 97.5% accuracy, 96.3% sensitivity, and 98.1% specificity [8]. Two techniques were developed for rice crop yield: Bayes Net and Naïve Bayes. Bayes Net performed better than Naïve Bayes. The SVM approach [9] was used to estimate rice crop production, and the experimental results exceeded those of Naïve Bayes, Bayes Net, and MLP [10]. The suggested technique outperforms PCA in predicting the yield of wheat crop remote sensing data [11]. The researcher used ANN and MLP to study the prediction of rapeseed crops, and there was an absolute percentage of inaccuracy of less than 9.43% on average.

For sugar cane [12], models based on Random Forest and Gradient Boosting Trees were developed, and the Random Forest model outperformed the Gradient Boosting Tree model in the testing [13]. Using soil factors, rainfall, and yield, the scientists employed a recurrent neural network to forecast rice crop production. In every year of the research, the Root Mean Square Value is less than 45%. Fully connected neural network algorithms RNN-CNN Framework [14], Random Forest, and Deep were designed to predict soybean and corn crop yield, and the RNN-CNN framework produced better results with an RMSE of 4.15. The accuracy of the Deep Recurrent Q-Network [15] model for paddy crop prediction was 93.7%.

Regression: The impact of environmental influences on agricultural production is examined using regression models. Numerous studies use both regression and classification techniques to predict agricultural productivity [16]. A multi-linear regression analysis was performed to predict the tea leaf output in Assam during the past 30 years using climatic parameters. The accuracy of the model was 82% [17]. To predict groundnut yield, multilinear regression, a KNN classifier, and an ANN model were built. The K-nearest classifier outperformed the others with an RMSE value. Several regression algorithms [18] are used to calculate maize crop production using operational field images from remotely detected Landsat.

The Boosted Random Tree method outperformed Random Forest Regression in each of the research years, and a trustworthy technique for forecasting maize yield is random forest regression [19]. The production of rice crops was predicted using MLR, ANN, and the recently suggested hybrid model MLR-ANN. The hybrid MLR-ANN produced better results in terms of RMSE and MAE values. Wheat crop models include fuzzy logic, MLR, and ANFIS [20]. The AN- FIS approach performs better than fuzzy logic and MLR approaches based on the RMSE value. Researchers [9] suggested forecasting and comparing several harvests using density-based clustering and multilinear regression. Multi Linear Regression approaches perform better than Density Based Clustering. Uncertain time series [21] was used to predict rice crop yields using regression, which proved to be accurate. [22] In contrast to earlier techniques, stacked regression achieved an RMSE value below 1 since the complex regression algorithms Kernel Ridge, Lasso, and Enet were developed for several crops.

3.2 Unsupervised Algorithms

Unsupervised machine learning techniques use unlabeled training data to teach computers to uncover insights and hidden patterns. Clustering and association rules are the two main unsupervised learning techniques utilized in agricultural production prediction. To ascertain the relationship among soil, crop range, and rainfall [23], the wheat crop was treated to an association rule technique. It was discovered that the association rule may yield valuable agricultural production data. A neural network [24] was employed to forecast the wheat crop, and the fuzzy C-mean clustering algorithm's RMSE value was 0.17. Clustering techniques [25] such as K-Means, DBSCAN, CLARA, and PAM have been applied to classify Bangladeshi soils, with PAM delivering superior outcomes. For yield prediction, methods including Stepwise, Simple Linear, and Multiple Linear Regression are utilized.

3.3 Ensemble Learning Algorithms

Supervised methods of learning such as Naïve Bayes, Random Forest, and Logistic Regression are frequently used due to their high prediction capability. A comparison of the two approaches revealed that both were effective in forecasting yields, particularly RF's ability to handle high-dimensional data [26]. For yield prediction, hybrid frameworks have also been investigated. By combining several machine learning methods, these models improve performance by using more effective feature extraction and combination techniques. A new hybrid learning approach, for example, showed exceptional accuracy by using supervised models to identify intricate correlations in agricultural production data [27]. Additionally, yield prediction has greatly improved due to techniques such as Deep Learning. Recurrent neural networks (RNN) and convolutional neural networks (CNN) have shown to be accurate and durable, especially for time-series and spatial data, exceeding conventional models [28]. These developments highlight how important machine learning is for handling issues like resource optimization in agriculture and climate unpredictability.

Bagging: It is a method that involves training several models on various data subsets and averaging their results to lower the variance of predictions. The process entails training a base learner, usually a decision tree, on each subset of the original dataset using bootstrap samples, which are randomly selected with replacement. Bagging minimizes overfitting and maximizes predictive accuracy by aggregating the predictions of numerous models [29]. Bagging is effective in handling large volumes of highly variable agricultural data. In agricultural use, Random Forest (RF), a widely used bagging-based algorithm, has been extremely effective. Literature indicates that Random Forest (RF) is capable of working on noisy and multi-modal agricultural data and making accurate predictions for many types of crop types. For instance, one study showcased its capabilities to forecast the yield of crops like rice and sugarcane in regions with varied environmental conditions [30]. In addition, in comparison to standalone models, bagging in multi-modal datasets has demonstrated improvements in prediction accuracy. Improved accuracy of agricultural production forecasts even when data is missing or inaccurate has been achieved through research using Bagging's ability to combine estimates [1].

Boosting: It increases the prediction accuracy of weak learners by iteratively sequentially training the model. Each new model corrects the mistakes of the old model by giving more weight to examples that were misclassified. To create a single, potent prediction model, several weak learners (typically decision trees) are coupled using well-known boosting algorithms like Extreme Gradient Boosting (XGBoost) and Adaptive Boosting (AdaBoost) [31]. Some studies show how AdaBoost [32] may be used to estimate crop yields with excellent accuracy for crops like maize and rice. The model demonstrated the effectiveness of boosting in managing noisy datasets by utilizing input parameters including temperature, rainfall, and soil conditions. The comparison of Machine Learning techniques and ensemble learning will be summarized and presented in Section IV in Table 1.

Table 1. Comparing Machine Learning with Ensemble Learning Techniques.

Technique	Dataset used	Culture	Features	Results
CLASSIFICATION TECHNIQUES				
LR, K-NN (K-Nearest Neighbors), ANN	Historical crop yield data	Rice, Potato, Wheat, Boro, Amon	Environmental and historical yield data	For rice, potato, and wheat, ANN demonstrated higher accuracy; for Boro and Amon, LR demonstrated greater accuracy [6].
Multilayer Perceptron (MLP) Neural Network	Experimental rice datasets	Rice	Sensitivity, Specificity, Accuracy	Achieved 97.5% accuracy, 96.3% sensitivity, 98.1% specificity [7].
Bayes Net, The Naïve Bayes	Historical rice crop yield	Rice	Yield parameters	Bayes Net outperformed Naïve Bayes [8].
(SVM) Support Vector Machine	Experimental datasets	Rice	Environmental parameters	Outperformed Naïve Bayes, Bayes Net, MLP [9].
(PCA) Principal Component Analysis	Remote sensing data	Wheat	Soil and climatic factors	Suggested approach outperformed PCA in wheat crop yield prediction [10].
ANN, MLP	Experimental datasets	Rapeseed	Yield and soil parameters	Achieved Mean Absolute Percentage Error (MAPE) < 9.43% [11].
Gradient Boosting Trees (GBT) and Random Forest (RF)	Soil and environmental data	Sugarcane	Soil factors, rainfall	RF outperformed GBT [12].
Recurrent Neural Network (RNN)	Soil, rainfall, yield data	Rice	Soil factors, rainfall, yield	RMSE < 45% annually [13].
RNN-CNN Framework, RF, Deep Neural Network	Experimental datasets	Soybean, Corn	Environmental parameters	RNN-CNN framework performed best with RMSE = 4.15 [14].
Deep Recurrent Q-Network (DRQN)	Paddy datasets	Paddy	Environmental and soil factors	Achieved 93.7% accuracy [15].
REGRESSION TECHNIQUES				
Multilinear Regression (MLR)	30 years climatic data	Tea Leaf (Assam)	Climatic parameters	Accuracy of 82% [16].
MLR, KNN Classifier, ANN	Groundnut yield dataset	Groundnut	Yield parameters	KNN outperformed others with lowest RMSE [17].
Boosted Random Tree, RF	Remotely sensed Landsat images	Maize	Operational land pictures	Boosted Random Tree outperformed RF [18].

MLR, ANN, Hybrid MLR-ANN	Rice crop data	Rice	Yield parameters	Hybrid MLR-ANN gave superior RMSE and MAE [19].
Fuzzy Logic, MLR, ANFIS	Wheat crop data	Wheat	Environmental and climatic parameters	ANFIS outperformed Fuzzy Logic and MLR [20].
MLR, Density Based Clustering	Multiple crop dataset	Multiple Crops	Environmental parameters	MLR was more successful than clustering [21].
Stacked Regression (Kernel Ridge, Lasso, Enet)	Multi-crop dataset	Multiple Crops	Environmental and complex parameters	RMSE value below 1 (Stacked Regression superior) [22].
UNSUPERVISED LEARNING				
Association Rule	Soil, crop variety, rainfall	Wheat Crop	Relationship between soil, variety, rainfall	Yielded valuable production data [23].
Fuzzy C-Means Clustering	Wheat yield data	Wheat Crop	Yield prediction parameters	RMSE value of 0.17 [24].
K-Means, DBSCAN, CLARA, PAM	Soil classification data	Bangladeshi Soil	Soil classification parameters	PAM performed best among methods [25].
ENSEMBLE LEARNING				
Random Forest (RF)	Multimodal agricultural datasets	Rice, Sugarcane	Random Forest (RF)	Effective in handling high-dimensional data and noisy datasets, improving prediction accuracy [26].
Hybrid Frameworks	Combined agricultural datasets	Agricultural production	Hybrid Frameworks	Improved accuracy through combination of machine learning models [27].
Bagging (Random Forest)	Bootstrap sampled datasets	Rice, Sugarcane	Bagging (Random Forest)	Reduces overfitting and variance; effectively predicts crop yield in variable datasets [29].
Boosting (AdaBoost, XGBoost)	Noisy agricultural datasets	Maize, Rice	Boosting (AdaBoost, XGBoost)	Improved accuracy by correcting errors in predictions; highly effective for noisy datasets [32].

Historically, neural networks, hybrid techniques based on Deep Neural Networks (DNNs) and Recurrent Neural Networks (RNNs) have been investigated to build various enhancements. They outperform classic ML algorithms on some occasions, making advantage of their ability to optimize for time-series sequential data.

4 Conclusion

Many studies have shown the strength of supervised learning techniques, especially classification and regression methods, for predicting crop yields. Despite the parameter complexity being higher in classification methods, as they can handle several crop types, sometimes it would be. Methods of ensemble learning, which include but are not limited to: bagging, boosting, and stacking, address the weaknesses of individual models by combining multiple base learners to directly reduce bias, and variance, and improve prediction accuracy. It is demonstrated that ensemble strategies especially outperform in the case of multi-modal agrarian datasets which improve reliability for inter-corresponding prediction while also making use of heterogeneous data. By contrast, unsupervised learning methods such as clustering and association rule mining are relatively untouched for complex crop yield prediction problems. Such methods are fine for solving multi-dimensional problems as multi-core bi-cluster identification can reveal patterns in crop data that may go unnoticed in traditional methods. Future work will focus on constructing adaptive hybrid ensemble models that might provide significant improvement in prediction efficiency and performance. Ensemble learning Combined with advanced machine learning strategies can improve prediction performance, promote Sustainable practices, and empower data-driven decision-making. A comprehensive focus in this direction will have a vast impact on food security, economic growth, and the sustainable development of world agriculture.

References

1. M. Keerthana, K. Meghana, S. Pravallika, K. Modepalli, An ensemble algorithm for crop yield prediction, 2021, pp. 963–970. doi:10.1109/ICICV50876.2021.9388479.
2. T. Klompenburg, A. Kassahun, C. Catal, Crop yield prediction using machine learning: A systematic literature review, Computers and Electronics in Agriculture 177 (2020) 105709. doi:10.1016/j.compag.2020.105709.
3. D. S. Rao, S. S. Chaganti, S. S. Chelikani, Y. V. Nandamuri, P. V. Nippun, Crop yield prediction using stacking ensemble model, in: R. Tiwari, M. Saraswat, M. Pavone (Eds.), Proceedings of International Conference on Computational In- telligence (ICCI 2023), Algorithms for Intelligent Systems, Springer, Singapore, 2024.
4. D. Tripathi, S. K. Biswas, An expert system using ensemble learning for crop yield prediction: Eescyp-i, in: 2022 International Conference on Advances in Computing, Communication and Materials (ICACCM), 2022.
5. M. Keerthana, K. Meghana, S. Pravallika, K. Modepalli, An ensemble algorithm for crop yield prediction, 2021, pp. 963–970. doi:10.1109/ICICV50876.2021.9388479.
6. A. S. Ahamed, Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in bangladesh, in: SNPD, 2015.
7. N. Gandhi, Rice crop yield prediction using artificial neural networks, in: IEEE International Conference on Technological Innovations.
8. N. Gandhi, Predicting rice crop yield using bayesian networks, in: 2016 Interna- tional Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016.
9. S. H. A. Aditya Shastry, M. H., A parameter based anfis model for crop yield prediction, in: IEEE International Advance Computing Conference, 2015.
10. Y. Zhang, Winter wheat yield estimation with ground based spectral information, in: 2018 IEEE International Geoscience and Remote Sensing Symposium, Milan, Italy, 2018.
11. G. Niedbała, Simple model based on artificial neural network for early prediction and simulation winter rape seed yield, Journal of Integrative Agriculture 18 (1) (2019) 54–61.
12. P. Charoen-Ung, Sugarcane yield grade prediction using random forest and gra- dient

boosting tree techniques, in: 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE), Nakhonpathom, Thailand, 2018.
13. S. Kulkarni, S. N. Mandal, G. S. Sharma, M. R. Mundada, Meeradevi, Predictive analysis to improve crop yield using a neural network model, in: 2018 Interna- tional Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018.
14. S. Khaki, A cnn-rnn framework for yield prediction, Frontiers in Plant Science 10 (2020).
15. M. D. VINCENT, Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications, IEEE Access 8 (2020).
16. R. D. Baruah, Use of data mining technique for prediction of tea yield in the face of climate change of assam, india, in: International Conference on Information Technology, Bhubaneswar, 2016.
17. V. Shah, Groundnut crop yield prediction using machine learning techniques, In- ternational Journal of Scientific Research in Computer Science, Engineering and Information Technology 3 (5) (2018).
18. R. S. Hossein Aghighi, Machine learning regression techniques for the silage maize yield prediction using time-series images of landsat 8 oli, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11 (12) (2018).
19. R. Bhargavi, A novel approach for efficient crop yield prediction, Computers and Electronics in Agriculture (2019).
20. A. Shastry, A parameter based anfis model for crop yield prediction, in: IEEE International Advance Computing Conference, Bangalore, 2015.
21. A. Garg, A robust and novel regression based fuzzy time series algorithm for pre- diction of rice yield, in: International Conference on Intelligent Communication and Computational Techniques, Jaipur, 2017.
22. P. S. Nishant, Crop yield prediction based on indian agriculture using machine learning techniques, in: International Conference for Emerging Technology, Bel- gaum, 2020.
23. L. J. Armstrong, Agricultural decision support framework for visualization and prediction of western australia crop production, in: 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016.
24. A. Verma, Crop yield prediction of wheat using fuzzy c means clustering and neural network, International Journal of Applied Engineering Research 13 (11) (2018).
25. S. Afrin, Analysis of soil properties and climatic data to predict crop yields and cluster different agricultural regions of bangladesh, in: 17th International Confer- ence on Computer and Information Science (ICIS), Singapore, 2018.
26. J. Babber, P. Malik, V. Mittal, K. C. Purohit, Analyzing supervised learning algo- rithms for crop prediction and soil quality, in: 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), 2022, pp. 969–973. doi:10.1109/ICCMC53470.2022.9753798.
27. S. M. M. Abdel-salam, N. Kumar, A proposed framework for crop yield predic- tion using hybrid feature selection approach and optimized machine learning, Neu- ral Computing and Applications 36 (2024) 20723–20750. doi:10.1007/s00521-024- 10226-x.
28. S. Z. Addu, Assessing environmental impact: Machine learning for crop yield pre- diction, E3S Web of Conferences 529 (2024). doi:10.1051/e3sconf/202452903008.
29. B. P. D. Tripathi, S. K. Biswas, A comparative analysis of ensemble learning techniques for crop yield prediction: Cypela, in: 2023 4th International Conference on Computing and Communication Systems (I3CS), 2023, pp. 1–6. doi:10.1109/I3CS58314.2023.10127268.
30. P. Kishore, Enhancing crop yield prediction using ensemble learning techniques on multi-modal agricultural data, International Journal of Intelligent Systems and Applications in Engineering 12 (20s) (2024) 262–268.
31. S. K. O.S. Nagesh, R.R. Budaraju, Boosting enabled efficient machine learning technique for accurate prediction of crop yield towards precision agriculture, Dis- cover Sustainability 5 (2024). doi:10.1007/s43621-024-00254-x.
32. D. B. Rohit Ravi, Crop yield prediction using xg boost algorithm, International Journal of Recent Technology and Engineering (IJRTE) 8 (2020) 3516–3520. doi:10.35940/ijrte.D9547.018520