iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83744

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Crop Yield Prediction Using the XG-Boost Algorithm in Indian Agroclimatic Conditions

Ranjana[1], Sandeep Ranjan[2]
[1]*Scholar*, [2]*Professor, Department of Computer Science and Engineering, CT Institute of Engineering, Management and Technology (CTIEMT), Jalandhar, India*

Abstract: ***Accurate assessment of crop yield is necessary for the use of resources in agroclimatic zones. This study emphasizes the need for an optimized approach for yield prediction utilizing the XGBoost machine learning algorithm using Indian agricultural datasets from the Kaggle library. The phases of data pre-processing, feature engineering, Bayesian hyperparameter optimization, and explanation analysis using SHAP make up the optimization approach. Performance measures like RMSE, MAE, and R2 have been used to assess a variety of machine learning methods, including linear regression, decision trees, random forests, and XGBoost. According to experimental data, XGBoost can beat other algorithms with low mistakes and an R2 value of 0.9666. Additionally, by determining the different weights of agricultural factors influencing crop production, the use of SHAP analysis was essential in increasing the model's accuracy. The accuracy of intelligent agricultural forecasts is anticipated to significantly increase with an optimized approach.***

Keywords: Bayesian Hyperparameter Optimisation, Crop Yield Prediction, Explainable Artificial Intelligence (XAI), Machine Learning, SHAP, XGBoost.

I. INTRODUCTION

Agriculture is really important for the Indian economy. Agriculture helps with food safety, rural development, and economic growth. In Indian agroclimatic conditions, crop yield forecasting is very important. Rainfall, irrigation, soil quality, temperature, and the season are all factors that need to be considered. These things can affect the crops. Using the XG-Boost Algorithm for Crop Yield Forecasting can be very useful. Indian agriculture is very dependent on these agroclimatic factors. We need to consider all these things to make sure we have a crop yield. The need to create a system for accurate crop production forecasting emerges as a result of growing concerns about climate change and other factors that contribute to the problem of water scarcity[1] .

Note that the classical models for building the forecast imply the use of statistics and expert experience. The problem is that these can provide for the development of relevant forecasts but lack certain capabilities regarding the discovery of interconnections between agricultural, climatic, and environmental factors. Hence, the need to concentrate on the use of intelligent techniques in the analysis of agricultural data[2] .

Machine learning and remote sensing technologies have come together to provide a very potent tool for yield prediction. A multitude of useful information regarding crop conditions, development, soil, and environmental elements may be obtained by remote sensing technology, which includes satellite imaging, UAV-based monitoring, and GIS data[4] . According to a recent study, machine learning techniques may be used to estimate agricultural output under changing weather circumstances by using meteorological and environmental parameters. It has been demonstrated that climate-based models are effective in determining how climate variables like temperature and precipitation affect agricultural productivity[5] .

Furthermore, because deep learning algorithms can extract intricate patterns from massive amounts of data, they have significantly improved agricultural forecasts. Deep learning algorithms achieved accurate predictions using data from satellites, soil, and the environment[6] .

The hybrid modeling technique, which combines process-based agricultural models with machine learning and remote sensing data, has already shown promise in predicting crop yields. The forecasts are much more accurate because hybrid models make it easier to incorporate both environmental and agronomic data. For agricultural surveillance and precision farming, hybrid models work well[7] .

Researchers have been inspired to employ more creative techniques for estimating crop production due to the growth in agricultural data. Machine learning models are capable of efficiently analyzing a variety of data pertaining to soil, climate, and agricultural conditions and identifying certain trends that affect plant output. Deep learning techniques have been shown to greatly enhance machine learning estimations in agriculture[8].

The idea of Explainable Artificial Intelligence (XAI), which aims to increase the interpretability and transparency of machine learning algorithms, is another area of agricultural analytics progress. By using methods like SHAP, scientists and analysts can comprehend how aspects connected to agriculture influence forecasts[9] .

Although there have been many developments in crop yield prediction modeling, problems including overfitting, prediction uncertainty, models that are difficult to understand, and generalization because of agroclimatic variability still exist. This study suggests the best crop production forecast modeling strategy based on XGBoost algorithms, which includes feature engineering, data pretreatment, Bayesian optimization for hyperparameter tweaking, and SHAP for explanation. Regression analysis has been used to assess such a model using traditional metrics[10] .

II. LITERATURE REVIEW

A machine learning technique for estimating agricultural output was suggested by authors Chawla and Singh[11] , taking into account ecological and agrarian factors. In comparison to traditional statistical methods, the experiment demonstrated the effectiveness of employing machine learning algorithms to identify the intricate interconnections of the factors involved in crop production estimation.

The capabilities of XGBoost and Random Forest machine learning algorithms used for agricultural yield estimation were examined in Bisht and Nahar's investigations [3] . Because of its effective gradient boosting and nonlinearity modeling, experiments demonstrated the superiority of the XGBoost technique.

To estimate crop production using Indian agricultural data, Bhavika et al.[12] devised a method based on the XGBoost machine learning technique. In calculating agricultural yield, it emphasized the benefits of ensemble learning.

Furthermore, to enhance the machine learning agricultural yield estimate, Dubey et al[13] suggested applying interpretable methods based on SHAP. According to the study, explainable AI can improve model transparency without lowering prediction quality.

To forecast rice yield, Sah et al.[5] employed machine learning techniques in conjunction with optical and SAR data for remote sensing. Their findings demonstrated the importance of remote sensing data in enhancing crop forecasting and monitoring.

Machine learning techniques for crop yield prediction based on agricultural factors, including cultivation area, production, yield, and irrigation area, have been proposed by Adlin Jebakumari and Jayanthiladevi[14] . In order to demonstrate how machine learning techniques might benefit agriculture, they employed many supervised learning algorithms in their research. Using satellite, meteorological, and soil data, Ashfaq et al.[6] developed a deep learning method for predicting wheat production.

For instance, Menon et al.[15] predicted agricultural output at the field level using machine learning approaches in conjunction with multi-temporal satellite data. Temporal data was shown to be crucial for determining agricultural development patterns and, consequently, for precisely forecasting crop yields.

In addition, Yenkikar et al.[9] proposed using Explainable AI in their hybrid machine learning model, which is based on SHAP and LIME approaches. Significantly, these two machine learning approaches not only achieved great prediction accuracy but also explained how the models function.

Likewise, Ashfaq et al.[6] created a deep learning model that included weather, soil, and remote sensing data (satellite photos) to estimate wheat yields. This study showed that the deep learning system is capable of handling intricate agroclimatic data and making precise agricultural production predictions.

Even though machine learning, deep learning, remote sensing, and Explainable AI helped achieve the encouraging results mentioned above, a number of difficulties were encountered during the model-building process, including model overfitting, poor model generalization ability, and low interpretability. Thus, combining the Bayesian hyperparameter tuning approach and explainability using SHAP, the current work suggests creating an optimized XGBoost model.

III. PROPOSED METHODOLOGY

Our goal in this study is to use XGBoost modeling methods to create the best machine learning system for accurate crop production prediction under Indian agroclimatic conditions. Data pre-processing techniques, feature development, hyperparameter tuning using Bayesian optimization, and XAI technology are all used in this strategy.

This agricultural experiment uses 19,689 samples of data that were obtained from Kaggle[16] . Crop factors and agroclimatic variables, such as crop kinds, year, season, state, crop cultivation area, production, fertilizer and pesticide application rates, and crop yields, are included in the information. Numerous studies emphasize how crucial it is to use big data to estimate agricultural productivity[5] .

Before training the model, the data was split into an 80-20 distribution for the train and test sets. The data was pre-processed in a number of ways to enhance its quality and facilitate effective model training. These pre-processing methods included data normalization, categorical data encoding, and duplicate data elimination [1] .

Feature engineering has been used to identify the key agricultural characteristics that contribute to crop yield and to create new features that show how current characteristics are interdependent. The accuracy of yield prediction has been shown to significantly increase when suitable characteristics are chosen and transformed[15] .

Linear regression, decision trees, random forests, and XGBoost are the machine learning methods that are employed and evaluated for agricultural yield prediction. The XGBoost technique is the best model for prediction since it can handle nonlinearity in the high-dimensional data space extremely well. The model's hyperparameters were optimized using Bayesian approaches in order to improve the model's accuracy and avoid overfitting. It will aid in determining the optimal setup that will provide the most accurate model[18] .

RMSE, MAE, and R2 scores are the metrics utilized to complete the review procedure. To further define the function of each characteristic in the decision-making process, the SHAP (explainable artificial intelligence) technique was employed [19] .

Preprocessing and feature selection are important factors in improving the accuracy of crop yield prediction. Appropriate pretreatment of environmental and agricultural characteristics maximizes the algorithm's learning ability and guarantees that there is no repetition of data[20] . Crop yield forecasting has advanced significantly thanks to data-driven learning algorithms that facilitate the examination of several agricultural variables. This ensures wise decision-making and the creation of sustainable farming systems[14] .

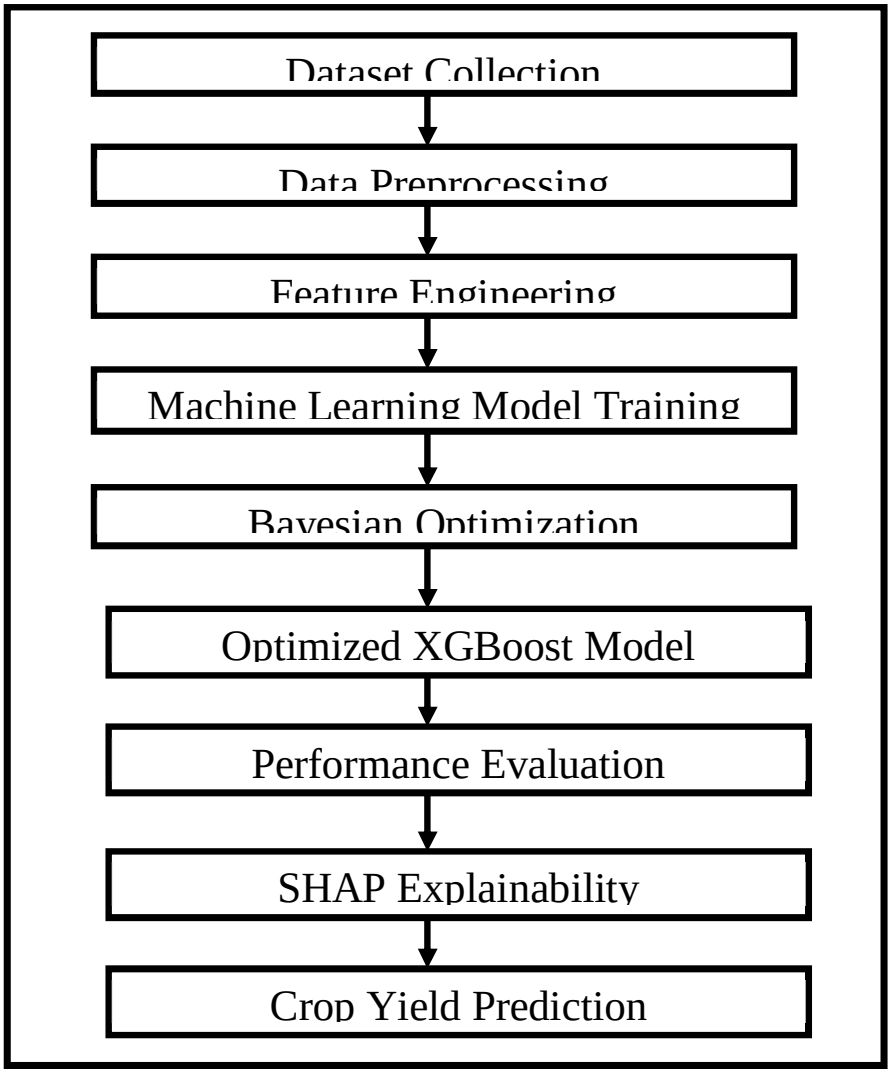


Figure 1. Proposed Workflow of Optimised XGBoost-Based Crop Yield Prediction Framework

IV. RESULTS AND DISCUSSION

In this chapter, we show some experimental results attained through the use of various machine learning algorithms for the prediction of crop yields under the agro-climatic conditions in India using Python on the Google Colab environment. The different types of machine learning algorithms used include linear regression, decision trees, random forests, and optimised XGBoost models.

A. Experimental Results from Data Preprocessing

Different agro-climatic features of crop production, land allocation of the crop, fertiliser application, pesticide application, and yield were identified in the agricultural data set. The process of transforming, normalising, removing duplicates, and encoding the data is carried out in the pre-processing stage.

B. Comparison among the Machine Learning Algorithms

Machine learning models are employed in predicting crop yield. These models include Linear Regression, Decision Tree, Random Forest, and optimised XGBoost models. Comparison of all these models is presented in the table.

Table 1. Comparative Machine Learning Model Performance

Model	RMSE	MAE	R^2 Score
Linear Regression	692.9976	139.4293	0.4006
Decision Tree	322.9946	11.7066	0.8698
Random Forest	270.9966	10.3657	0.9083
Optimized XGBoost	163.6337	19.3717	0.9666

As per the results from the experiment, it can be stated that the Optimised XGBoost model has performed better compared to other models, as this has produced the lowest value for RMSE, which is 163.6337, with an R2 score of 0.9666, indicating a highly accurate and generalised model. The linear regression algorithm has turned out to be the worst algorithm, while the Decision Tree and Random Forest models have shown a very good result due to their capability to establish non-linear agricultural relationships.

C. Bayesian Hyperparameter Tuning Result

Bayesian hyperparameter tuning was used to improve the predictive capability of the XGBoost algorithm as well as prevent overfitting. When we were working on the XGBoost algorithm, there were some settings that we needed to adjust. The XGBoost algorithm had things like learning rate, max_depth, n_estimators, subsample, and columnsample that we had to get just right. The XGBoost model did a lot better after we made these adjustments. The result of the XGBoost model was pretty good with an RMSE of 163.6337, an MAE of 19.3717 and an R2 Score of 0.9666, for the XGBoost model.

Table 2. Optimized XGBoost Hyperparameters

Hyperparameter	Optimized Value
n_estimators	174
max_depth	3
learning_rate	0.1005
subsample	0.8046
colsample_bytree	0.5418

D. SHAP Analysis and Feature Importance

The parameters with the strongest influence in predicting crop yield in agriculture have been identified using the feature importance and SHAP analysis approach. The parameters included Production, Type of Crop, Land Area, and State.

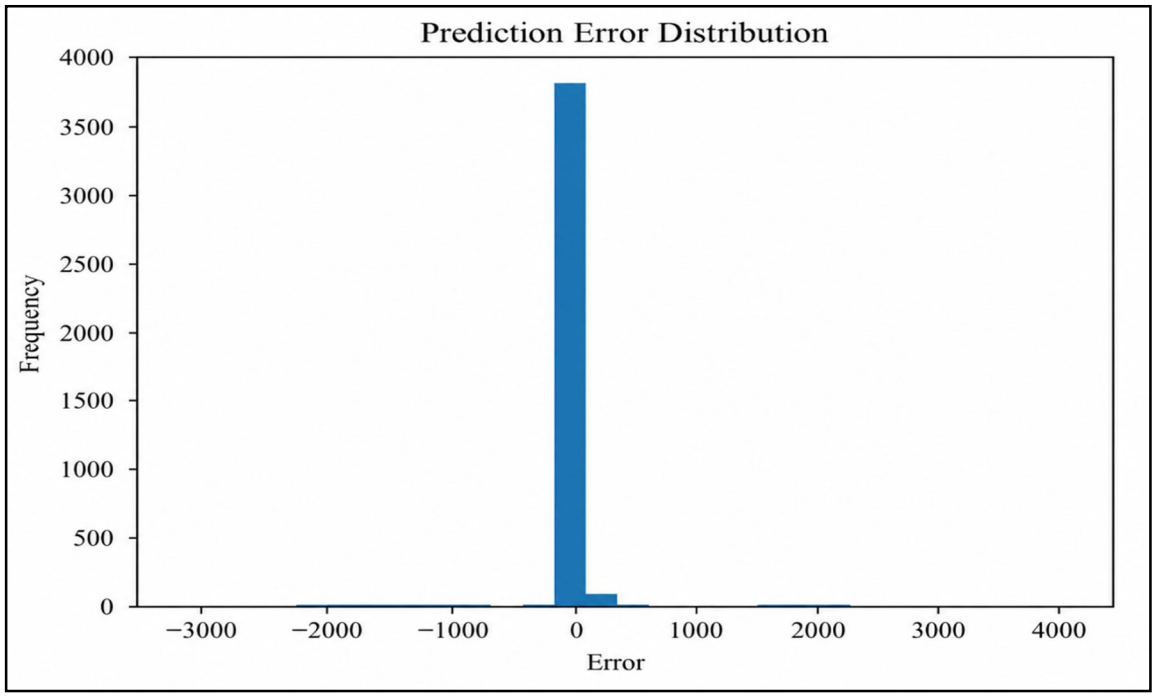


Figure 1. Analysis of Feature Importance

The production variable had the largest effect on yield prediction, followed by fertiliser and pesticide variables, which had a medium effect.

Explainability was enhanced by using SHAP analysis to understand how each attribute contributed to prediction results and to verify that XAI methods were effective for crop yield prediction.

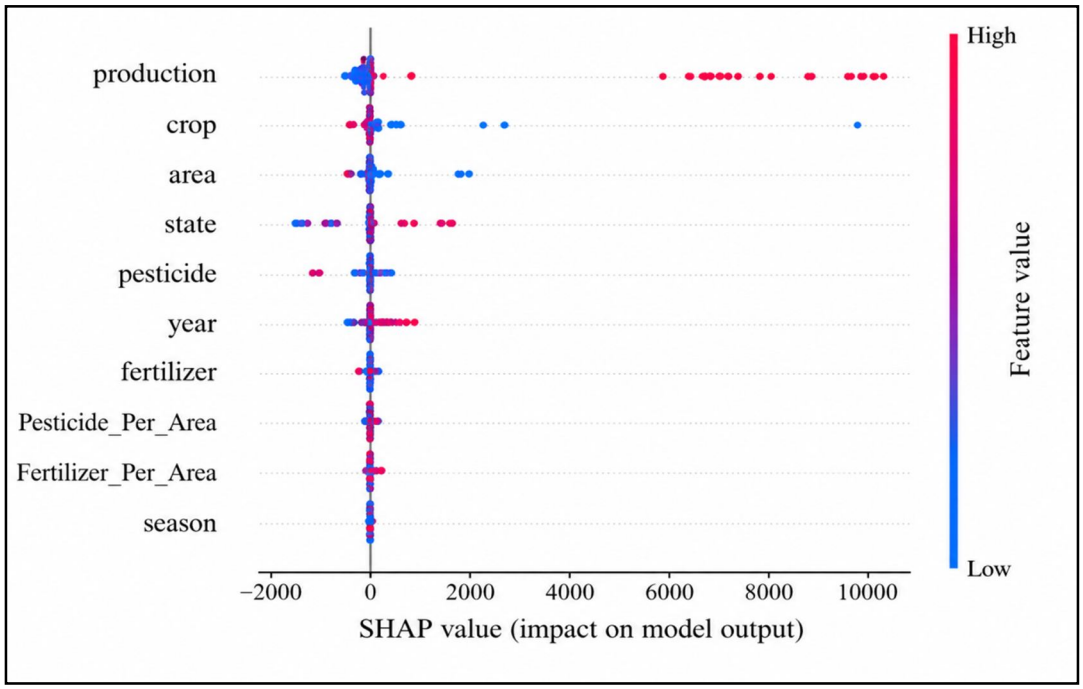


Figure 2. Plotting SHAP Summary for Feature Contribution

E. Analysis of Actual vs. Predicted Yield

The plot for actual and predicted yields showed that there was a strong connection between the actual yield values and the predicted yield values generated by optimising the XGBoost model. Predictions are clearly quite accurate because they typically fall within the trend zone.

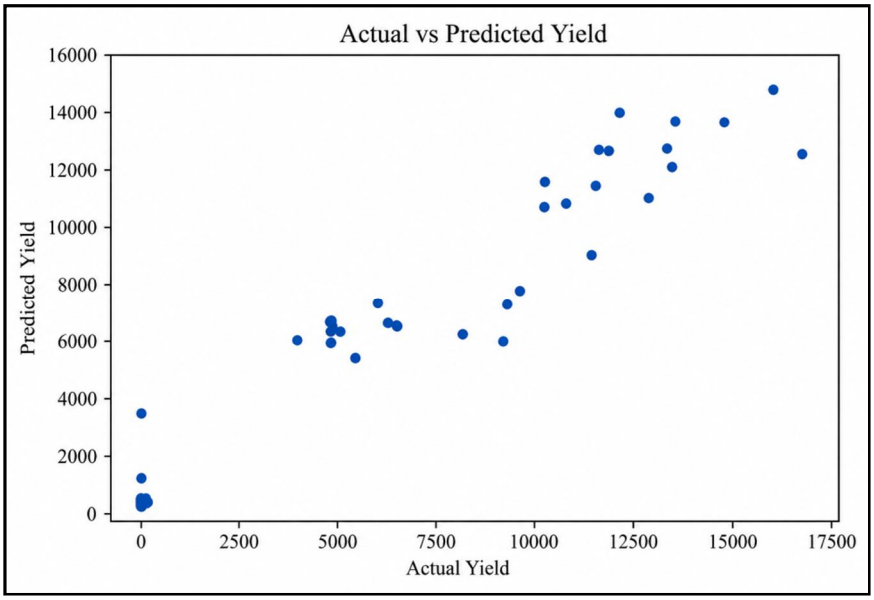


Figure 3. Actual and Predicted Data Scatter Plot

From the results of visual validation, it is evident that the model successfully learned the nonlinear relations existing between agriculture and crop yields.

F. Analysis of Prediction Errors

In order to establish the stability of forecasts, as well as to calculate the residuals of the model, an analysis of the distribution of prediction errors was conducted. The majority of the errors are around zero when examining the chart that displays the errors made with forecasts. This shows that our predictions were reasonably accurate and failed to differ significantly from reality. The determining errors converged around zero, meaning that our predictions were largely accurate.

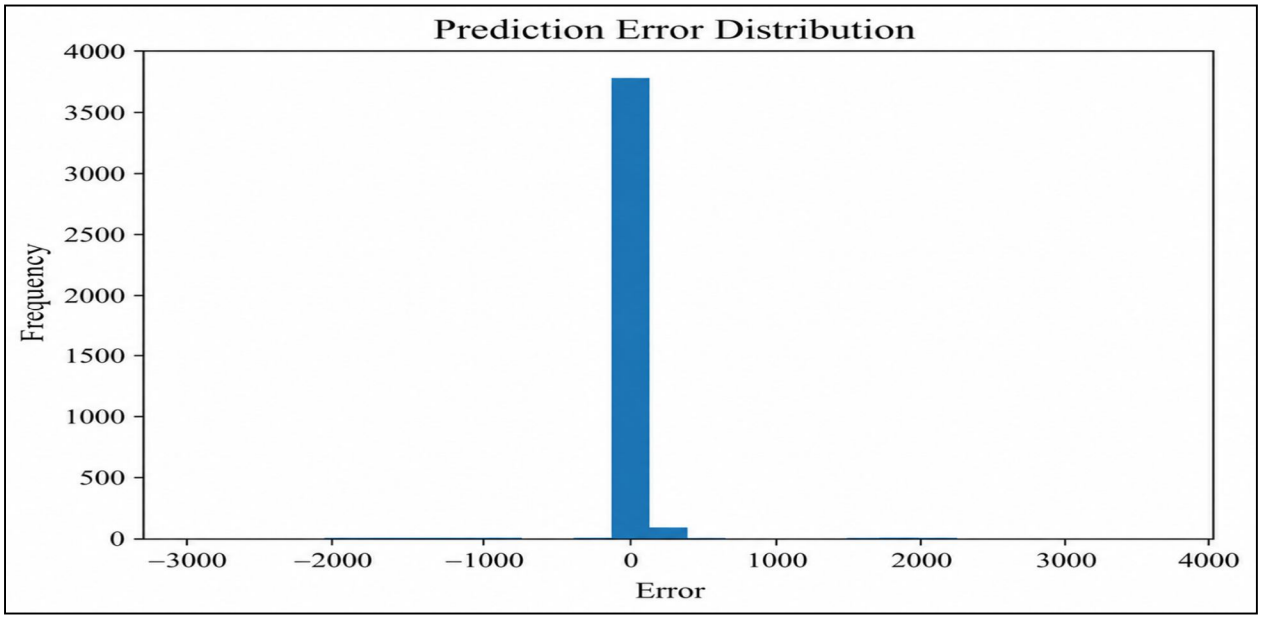


Figure 4. Distribution and Prediction Errors

Though there were outliers regarding the prediction of the error value, most of the residuals were within a very limited range, indicating stability of the XGBoost algorithm.

G. Cross-Validation Evaluation

To evaluate the ability of the model to generalise and avoid overfitting, cross-validation is carried out, and from the results obtained from the analysis, it can be seen that the optimised XGBoost model was consistent with regard to prediction.

Table 3. Optimised XGBoost Model's Cross-Validation Performance

Fold	RMSE	MAE	R^2 Score
Fold 1	312.25	46.41	0.9371
Fold 2	37.72	12.74	0.9849
Fold 3	165.31	21.57	0.8532
Fold 4	355.09	20.69	0.7216
Fold 5	88.79	13.11	0.9792

The model performed quite well, with an average RMSE of 191.8339, MAE of 22.9078, and R2 of 0.8952.

H. Discussion

According to the experimental findings, machine learning methods can accurately forecast crop production under the agroclimatic conditions of India. With an R2 score of 0.9666 and the lowest prediction error values, the improved XGBoost algorithm outperformed all other assessed models in terms of prediction accuracy.

The major reasons for XGBoost's better performance are its inherent regularization process, which lessens overfitting, and its capacity to grasp intricate nonlinear correlations among agricultural variables. By determining the optimal parameter settings, Bayesian hyperparameter optimization enhanced the model's predictive power.

Production, crop type, farmed area, and state were shown to be the most significant factors influencing crop yield forecast based on feature significance and SHAP analyses. The suggested framework is also good because it is stable and it works well with various kinds of data, which is something that the cross-validation results showed.

The study got some results, but it has some limitations. The study used data, but it did not have the latest weather and remote sensing information. The framework can be made better in the future by adding these things to make the weather forecast more accurate and more useful for the framework and for the people who use it.

V. CONCLUSION AND FUTURE WORK

A. Conclusion

In this study, a more advanced and optimised version of the XGBoost algorithm for machine learning to predict crop yield intelligently from the Indian agricultural climate system is proposed. It is found from the experiments that the optimised XGBoost algorithm works excellently for predicting the crop yield intelligently, where the maximum R^2 score is 0.9666 with a minimal error rate. The Bayesian hyperparameter optimisation of the algorithm has helped to improve the performance of the algorithm and minimise the chances of overfitting. By implementing the SHAP values, the model's interpretability has been enhanced using the features of agriculture to affect crop productivity.

B. Future Work

A few possible directions for the future may include the integration of real-time climatic data, satellite image analysis, and remote sensing data. Also, other machine learning techniques can be implemented along with deep learning methods to analyse agricultural data.

REFERENCES

[1] A. Jabed, M. Azrifah, and A. Murad, "Heliyon Crop yield prediction in agriculture : A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability," Heliyon, vol. 10, no. 24, p. e40836, 2024, doi: 10.1016/j.heliyon.2024.e40836.

[2] T. Van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning : A systematic literature review," Comput. Electron. Agric., vol. 177, no. July, p. 105709, 2020, doi: 10.1016/j.compag.2020.105709.

[3] S. Bisht and S. Nahar, "Crop Yield Prediction Accuracy Using XGBoost and Random Forest," Int. J. Sci. Res. Eng. Trends, vol. 11, no. 3, pp. 1–6, 2025.

[4] And R. P. A. Kumar, I. Singh, M. Kashyap, A. Kumar, N. B. Devi, S. Singh, S. Sharma, "Integration of machine learning and remote sensing in crop yield prediction: A review," Int. J. Res. Agron., vol. 8, no. 1S, pp. 549–562, Jan. 2025, doi: 10.33545/2618060x.2025.v8.i1sh.2496.

[5] S. Sah, D. Haldar, R. N. Singh, B. Das, and A. S. Nain, "Rice yield prediction through integration of biophysical parameters with SAR and optical remote sensing data using machine learning models," Sci. Rep., vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-72624-4.

[6] M. Ashfaq, I. Khan, D. Shah, S. Ali, and M. Tahir, "Predicting wheat yield using deep learning and multi-source environmental data," Sci. Rep., vol. 15, no. 1, pp. 1–20, 2025, doi: 10.1038/s41598-025-11780-7.

[7] A. M. S. Kheir et al., "Hybridisation of process-based models, remote sensing, and machine learning for enhanced spatial predictions of wheat yield and quality," Comput. Electron. Agric., vol. 234, Jul. 2025, doi: 10.1016/j.compag.2025.110317.

[8] K. Jhajharia, P. Mathur, S. Jain, and S. Nijhawan, "ScienceDirect Procedia ScienceDirect Crop Yield Prediction using Machine Learning and Deep Learning Crop Yield Prediction using Techniques Machine Learning and Deep Learning Techniques," Procedia Comput. Sci., vol. 218, pp. 406–417, 2023, doi: 10.1016/j.procs.2023.01.023.

[9] A. Yenkikar, V. Prakash, M. Bali, and T. Ara, "MethodsX An explainable AI-based hybrid machine learning model for interpretability and enhanced crop yield prediction ☆," MethodsX, vol. 15, no. June, p. 103442, 2025, doi: 10.1016/j.mex.2025.103442.

[10] N. Iqbal et al., "Analysis of Wheat-Yield Prediction Using Machine Learning Models under Climate Change Scenarios," Sustain., vol. 16, no. 16, pp. 1–26, 2024, doi: 10.3390/su16166976.

[11] H. S. Chawla and D. Singh, "Development of a Machine Learning Model for Crop Yield Prediction in Agriculture," The Bioscan, vol. 20, no. Supplement 2, pp. 827–832, 2025, doi: 10.63001/tbs. 2025. v20.i02.s2.pp827-832.

[12] B. A. Bhavika, G. Samaira, D. Kumari, and R. Kusum, "Predicting Annual Crop Yields in India ' s States : Leveraging XGBoost Techniques for a Web-Based Machine Learning Model," " Int. J. Res. Trends Innov., vol. 10, no. 3, pp. 140–146, 2025.

[13] Y. Dubey, A. Sakhare, A. Tasare, S. Kakad, and R. Umate, "Explainable Model for Agricultural Crop Yield Prediction in Indian Conditions with SHAP Analysis," SSRG Int. J. Electron. Commun. Eng., vol. 12, no. 1, pp. 236–244, 2025 doi: 10.14445/23488549/IJECE-V12I1P118.

[14] S. Adlin Jebakumari and A. Jayanthiladevi, "AgriYield-ML: Enhancing Agricultural Productivity through Machine Learning: A Model for Accurate Crop Yield Prediction," J. Eng. Sci., vol. 53, no. 5, pp. 155–169, Sep. 2025, doi: 10.21608/jesaun.2025.367985.1450.

[15] A. S. Menon, J. Aravinth, R. Sankaran, and P. Kiran, "Smart Agricultural Technology Deep learning-based farm-level crop yield prediction using multi-temporal satellite data for complex engineering application," Smart Agric. Technol., vol. 12, no. August, p. 101562, 2025, doi: 10.1016/j.atech.2025.101562.

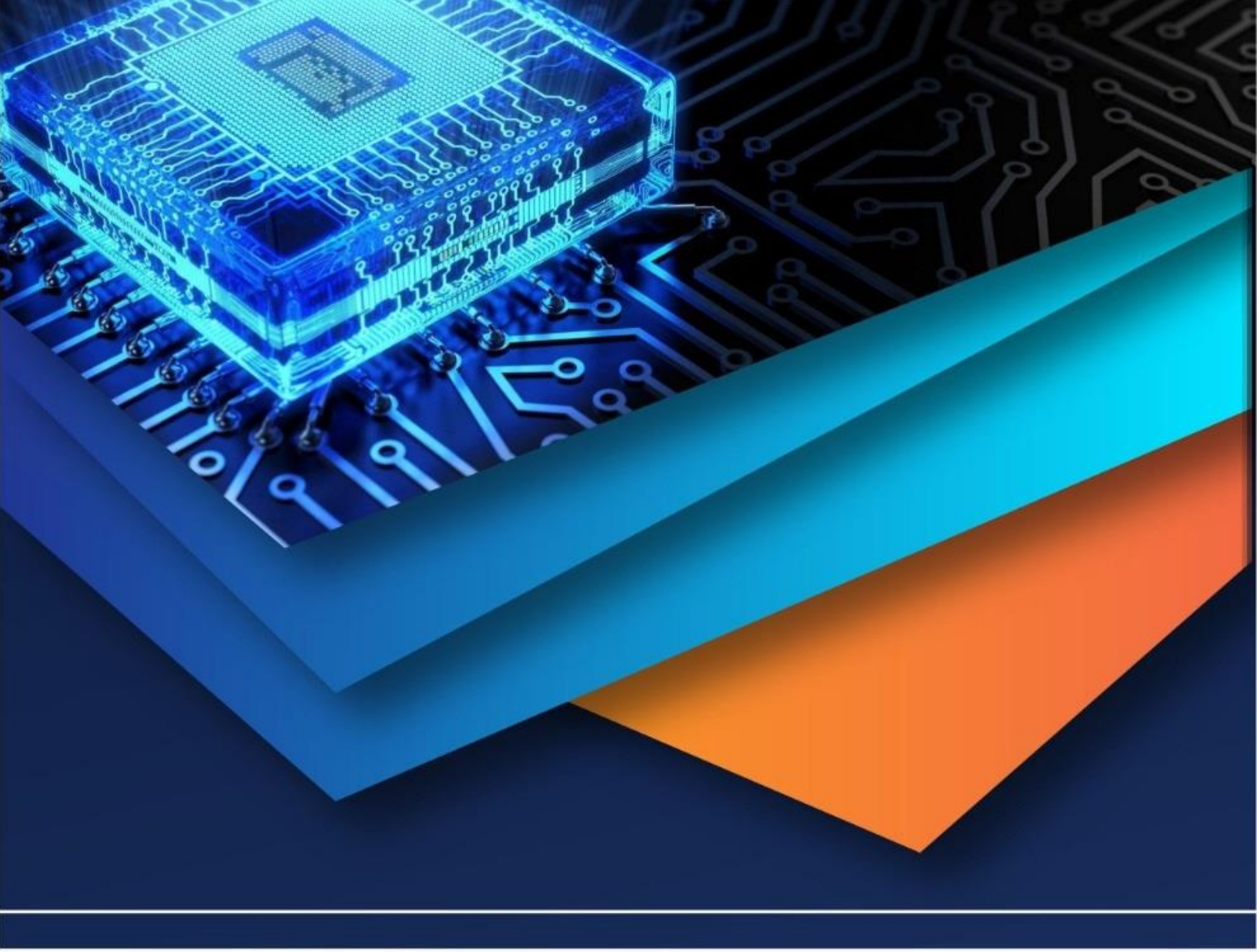
[16] Anshumish, "Crop Yield Data with Soil and Weather Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/anshumish/crop-yield-data-with-soil-and-weather-dataset.

[17] D. De Clercq and A. Mahdi, "Feasibility of machine learning-based rice yield prediction in India at the district level using climate reanalysis and remote sensing data," Agric. Syst., vol. 220, 2024, doi: 10.1016/j.agsy.2024.104099.

[18] K. P. S. Attwal, "Integrated machine learning model for wheat yield prediction using agronomic and meteorological factors: A case study from Punjab, India," Int. J. Agric. Food Sci., vol. 7, no. 8, pp. 385–400, 2025, doi: 10.33545/2664844x.2025.v7.i8f.636.

[19] J. Lu et al., "Estimation of rice yield using multi-source remote sensing data combined with crop growth model and deep learning algorithm," Agric. For Meteorol., vol. 370, no. January, p. 110600, 2025, doi: 10.1016/j.agrformet.2025.110600.

[20] S. Sarode, P. Sharma, P. Tidke, and N. Panghate, "Crop Yield Prediction Using Machine Learning," Lect. Notes Electr. Eng., vol. 1270, no. 4, pp. 335–343, 2025, doi: 10.1007/978-981-97-7876-8_30.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY