Journal of Information Systems Engineering and Management
2025, 10(18s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Performance Testing of Advanced Data Mining for Crop Yield Prediction Using SMGBR

M.Parthiban[1], Subhash Bhagavan Kommina[2], M.V.V.S. Nagendranath[3], Sudheerkumar Pulapa[4], E.V. Sandeep[5], T.Vinay[6]

[1345]*Department of Computer Science and Engineering, Sasi Institute of Technology and Engineering,*
[26]*Department of Information Technology, Sasi Institute of Technology and Engineering,*
West Godavari, Andhra Pradesh, India
parthimitit@sasi.ac.in

ARTICLE INFO

Received: 16 Dec 2024
Revised: 25 Jan 2025
Accepted: 10 Feb 2025

ABSTRACT

Crop plants are essential to economic viability, especially in agricultural areas of INDIA. Planning, resource planning, and successful practice in agriculture must possess a successful model of forecasting productive crop yield. Strategic Modified Gradient Boosting Regression (SMGBR) is proposed in this research with the objective of improving the accuracy of crop yield prediction. Unlike the typical Gradient Boosting Regression (GBR) models, the SMGBR model possesses a learning rate dynamic factor that it is better equipped to handle complex, non-linear patterns in farm-level data. This paper uses an extremely large dataset that was collected over a period of ten years (2010-2021) and contains pertinent factors like crop yield histories, advanced weather observations, and soil quality observations. Conservative performance testing was also done to check the predictability of the model for performance, efficiency, and reliability compared to conventional methods. The outcome of the work strongly indicates that SMGBR performs better than conventional methods in terms of performance but is also less computationally expensive with more dependable yield predictions. Increased reliability of the SMGBR model advantages farmers, policymakers, and farm planners the most. The model will be applied in improving better decision-making for agriculture long-term planning, resource utilization, and crop management towards economic and environmental sustainability. By the creation of sound data-driven predictions of yields, this research will improve agricultural production by limiting wastage and attaining equilibrium in ecosystems for crop development. The results are evidence of the importance of state-of-the-art machine learning methods and sound performance assessment in creating resilience and sustainability in farming communities.

Keywords: Crop Yield Prediction, Machine Learning, Gradient Boosting Regression, Performance testing, SMGBR

INTRODUCTION

Agriculture continues to be the cornerstones of the global economy, driving economic growth and food supply. Crop yield prediction maximizes the utilization of resources, planning, and decision-making. For farming that is vital in Banaskantha District, North Gujarat, precise prediction of yield enables sustainable agriculture through reduced risk due to uncertainty of yields. It also enables policymakers to plan markets and resource planning and, hence, is a priority for agriculture. Jeong et al. [1] examined Random Forest (RF) and Gradient Boosting Machines (GBMs) for yield prediction in 2016. Big data was handled effectively by RF, while GBMs provided stronger generalization capacity with issues of overfitting. In 2017, ensemble learning and neural networks were utilized by Li and Zhang [8]. Neural networks consumed patterns cost-effectively but were easier to interpret using GBMs. Their study fostered adaptive learning towards improved performance.

Li et al. [2] integrated dynamic learning rates into GBM models in 2018 to improve prediction performance. Singh et al. [5] developed a unique hybrid RF-Neural Network model in 2019 that was computationally intensive which will improve accuracy. Zhao et al. [6] proposed a XGBoost and LightGBM for large datasets, introducing dynamic learning rate updates in recent years. Sarkar et al. [6] compared RF, GBM, and SVMs and reported RF more stable and GBM more generalizing. They emphasized the usage of adaptive learning rates in GBMs to maximize accuracy. Ahmed et al. [7] in 2021 utilized XGBoost with a fixed learning rate in different conditions and had average success but proposed dynamic parameter optimization keeping in mind the climate changes. Awan et al. [4] utilized feature importance methods to RF models and found weather and soil characteristics as key factors influencing yield. Chen et al. [10] showed up to 2022 that adaptive learning rates to the boosting algorithm profoundly enhanced precision and served as the basis of the Strategic Modified Gradient Boosting Regression (SMGBR) model.

Sun et al. [11] proposed SIDEST technique, a super-resolution image reconstruction-artificial intelligence-based segmentation platform for finding accurate agricultural field boundary delineation, improving dataset quality and yield estimation. Dembani et al. [12] considered Federated Learning (FL) method with is mainly for privacy-preservation methods such as Homomorphic Encryption (HE) and Secure Multi-Party Computation (SMPC) to facilitate safe data sharing. Vahidi et al. [13] integrated two algorithms called Artificial Neural Networks (ANN) and 1D Convolutional Neural Networks (CNN) for predicting multi-depth soil moisture, which is very crucial for yield prediction. Raj et al. [14] proposed Support Vector Machines (SVM) performed better than K-Nearest Neighbors (KNN) in examining irrigation suitability with 97% accuracy for yield prediction. Mengmeng et al. [15] described vectorized parcel extraction via PLR-Net, a machine learning architecture with attraction field maps and multi-task feature interactions for precise satellite-based parcel definition. They highlighted the value of high-resolution geospatial data in reducing yield prediction model complexity. Yield prediction has grown from conventional statistics to sophisticated machine learning. Coupling dynamic learning rates, feature selection, and hybrid models improved prediction accuracy much more. Emerging research will call for adaptive algorithms to be sharpened, utilize high-resolution remote sensing, and adopt secure data-sharing practices such as Federated Learning. SMGBR is a milestone success, improving accuracy and adaptability for real-world use in agriculture.

OBJECTIVES

The The main objective of this research is to develop a strong and dynamic model to predict the yield of crop in the Banaskantha District of North Gujarat. farming is economically significant for the district, and therefore policymakers, farmers, and agri-business firms have to apply accurate forecasting techniques to achieve maximum productivity and maximize the utilization of resources. This present research proposes to combine recent machine learning techniques to enhance predictive accuracy to facilitate enhanced decision-making in farm planning and farm management. Among the prominent works is comparative and analytical evaluation of prediction capability of yield using diverse machine learning algorithms, namely, Random Forest (RF), Gradient Boosting Machines (GBMs), and neural networks. The aim is to improve current practice on the basis of adaptive learning rates, feature selection protocols, and model hybrid structures according to newer research. The aim is to determine the computationally most effective method with equal predictiveness. The second main aim is to research the ability of high-resolution satellite data to aid the accuracy of yield prediction. Satellite images and geospatial information are wealthy sources containing detailed information about climatic trends, soil fertility levels, and vegetative growth pattern. The intent of this present study is to combine such types of information resources with machine learning models as a bid to provide a holistic data-driven predictive model. Besides this, the research also tries to assess the viability of privacy-friendly data-sharing practices such as Federated Learning (FL). Since agri-data is sensitive and scattered, safe collaboration among the stakeholders is most critical. Using FL, homomorphic encryption, and secure multi-party computation, this research can create a decentralized, privacy-aware yield forecasting model.

METHODS

Data Collection starts with specifying data sources and types of data collected and proceeds with a careful description of utilized datasets, i.e., historical yield data, weather data, and soil health information. All datasets are examined in search of the most useful features in yielding predictions for crops. Preprocessing then discusses the procedure used to clean, join, and prepare the data for modeling. The chapter is rounded off by summarizing data collection and preprocessing challenges, together with the approaches taken to solve them, paving the way for the Strategic Modified Gradient Boosting Regression (SMGBR) model. Data on yield, weather, and soil from 2010 to 2021 were collected, including the following attributes, they are Yield Data: Year, Area of Cultivation (Ha), Total Yield (MT), Productivity (MT/Ha), Weather Data: Year, High/Low/Average Temperature (°C), Sunshine Hours, Rainfall (mm), Rain Days, Humidity (%), Pressure (Millibar), Soil Data: Year, pH Level, Nitrogen (kg per hectare), Phosphorus (kg per hectare), Potassium (kg per hectare), Organic Carbon (percentage), Boron, Iron, Zinc, Copper (mg/kg Soil), Sulfur (kg/ha) and Data Sources: Directorate of Horticulture, Gujarat State; India Meteorological Department (IMD), WeatherAPI, Wonderground; Soil Health Card Portal, Gujarat State, ICAR. The research developed a Strategic Modified Gradient Boosting Regression (SMGBR) model by incorporating several key modifications to the traditional Gradient Boosting Regression (GBR) model and the development process is show in fig 1.

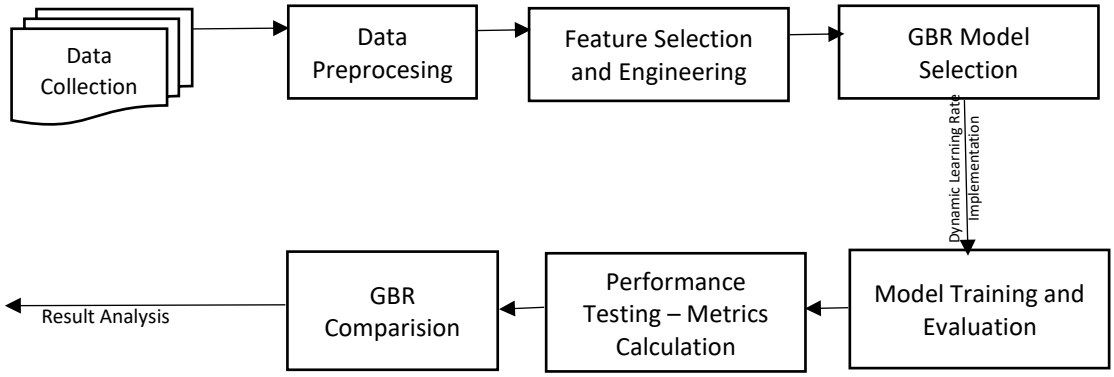

Figure 1: Architectural flow of SMGBR Model

Gradient Boosting Regression (GBR) was used as the baseline model here because it outperforms the other models to determine complex, non-linear data patterns. GBR works through the iterative generation of an ensemble of decision trees where each successive tree is framed to rectify the residual error of the previously generated trees and therefore is very much appropriate for regression problems having complex and variegated data sets. Also, its ability to accommodate various forms of input variables and its ability to give insights into feature importance also make it a good choice as a suitable model for crop yield prediction. Nevertheless, even with these advantages, GBR has some disadvantages that can negatively affect its performance, especially when dealing with very dynamic datasets like those that deal with agriculture. One of the most significant limitations is that it applies a constant learning rate during training. The learning rate specifies the relative contribution of each tree to the model, a constant may result in slow convergence (if small) or over fitting (if large), especially for data sets with changing data distributions at different times. To address this, adjustments were done, more importantly the addition of a dynamic learning rate. The feature is adaptive in that it dynamically varies the learning rate iteration-wise such that the model initially uses a larger learning rate to achieve faster convergence. It slowly wears off the learning rate with subsequent training to facilitate better model fine-tuning.

The model has a balance, by this, between rapid learning during early phases and careful fine-tuning in advanced phases that makes the accuracy and performance of the model even higher over time. This regularization allows the model to generalize better, decreases over fitting, and generally improves the prediction accuracy, particularly in data like in this research, with varying weather conditions, soil, and previous crop yields. With these changes of the original GBR, the Strategic Modified Gradient Boosting Regression (SMGBR) model provides a robust response to forecast crop yields in Banaskantha District. The Strategic Modified Gradient Boosting Regression (SMGBR) model utilized an advanced dynamic learning rate approach that changed over time during training. This approach addressed one of the main problems of predictive modeling: a compromise between fast early learning and accurate end fine-tuning. The model initially used a high learning rate, which caused it to learn general trends in data but only up to the earlier phases. In these initial stages, the model learns mostly updates its estimates since it has learned only general patterns between input features and the target variable. A benefit of learning with a higher initial learning rate is that the model trains faster and thus have lower overall training time. As more training iterations go by, however, the model becomes exposed to increasingly complex patterns and structures in the data.

Now that it has a high learning rate persistently, the model will overshoot the optimal solution, and that will lead to less precise predictions. To mitigate this, the dynamic learning rate diminishes over time. By slowing down the learning rate, the model can tune its predictions more sensitively, making slight corrections while getting close to the optimal solution. This operation avoids over fitting and maintains the model to be elastic with respect to small variations in the data at subsequent stages of training. Figure 2 illustrated dynamic learning rate adaptation during training, where it begins at 0.22 and gets exponentially smaller with iterations, offering greater control at subsequent stages of training.

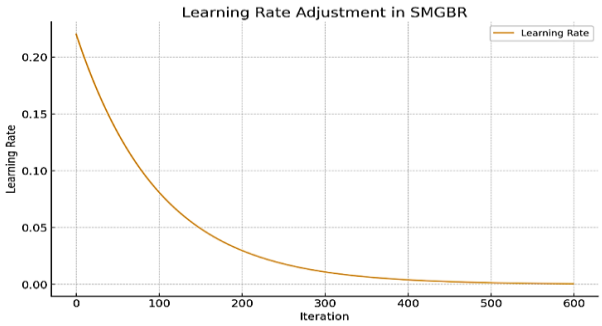

Figure 2: The dynamic adjustment of the learning rate

The use of the Strategic Modified Gradient Boosting Regression (SMGBR) model can be very well expressed via pseudo code. It creates a straightforward, language-unspecific definition of the steps taken during model training and testing. Because in that the procedure is defined briefly and openly, it is simple for the pseudo code to be translated into any programming language directly and therefore is readable by anyone with general information about programming vocabularies. Performance of SMGBR model was measured using different measures like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and correlation coefficient. These measures collectively gave a numerical estimation of precision and reliability of the model to predict yield. SMGBR model was realized within the environment of Python as a programming interface, and direct comparison was performed with the benchmark GBR model.

RESULTS

The Strategic Modified Gradient Boosting Regression (SMGBR) model was contrasted with the baseline Gradient Boosting Regression (GBR) model to identify the gains in prediction accuracy. The results were in line with the fact that the SMGBR model performed much better than the GBR model, particularly in years of high environmental uncertainty.

Table 1: Comparison of GBR and SMGBR Model Performance Metrics

Metric	GBR	SMGBR
Correlation Coefficient	-0.047	0.989
Mean Absolute Error (MAE)	299882	84779
Root Mean Squared Error (RMSE)	327563	89899

Table 1 shows that SMGBR model was appreciably greater in correlation coefficient (0.989) compared to GBR model (-0.047), which signifies highly positive correlation of estimated and actual yields. SMGBR MAE and RMSE were also appreciably lower, showing its higher precision and reliability in estimating yields.

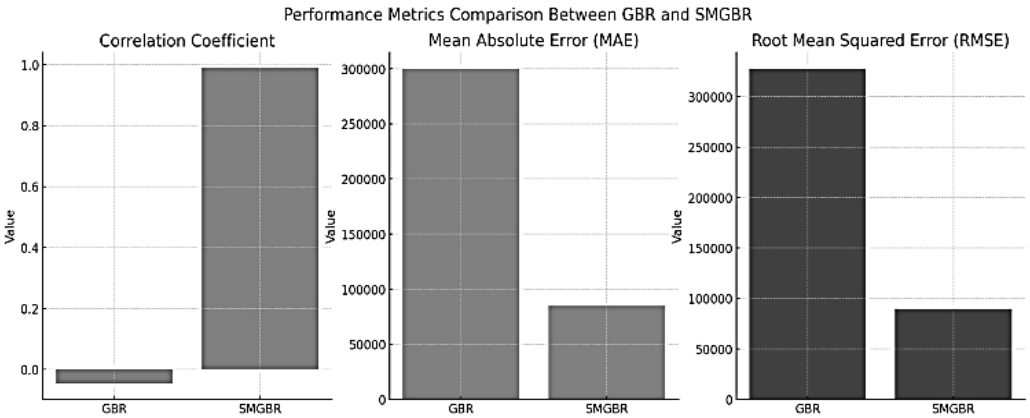

Figure 3: The performance metrics Chart

The improved predictive accuracy of the SMGBR model is especially valuable in providing beneficial contributions to agricultural planners and policymakers in the Banaskantha District. With the ability to supply better predictive estimations of yield, the model helps to make more rationalized choices in agriculture operations, such as in the optimal

resource distribution and environmental variability risk management. And yet with better performance, the SMGBR model also has its own shortcomings shown in fig 3. When complexity of the model is enhanced, more training time and computing resources are required. Further, dynamic learning rate can assist in preventing over fitting but is not the complete solution to problems caused by outliers' data or missing data during training. Strategic Modified Gradient Boosting Regression (SMGBR) models considerably improve precision for forecasts of Banaskantha District yield through dynamic learning rate adaptation. With these revisions, the model works significantly to tackle variability common in the context of agriculture data.

CONCLUSION

The study presents the critical significance of precise estimation of cabbage output in the elimination of inefficiencies in agriculture and process optimization in Banaskantha District cabbage farming in North Gujarat. The study depicts the way in which empirical methodology has the potential to enhance predictability maximally through the utilization of advanced machine learning methodologies like Random Forest, Gradient Boosting Machines, and neural networks. Adaptive learning rates, feature selection methods, and ensemble model architectures have been demonstrated to provide potential performance benefits to models, addressing one of the most significant challenges in agricultural forecasting. The study also indicates that the integration of high-resolution remote sensing imagery and geospatial analysis would improve models of yield forecasting. With climatic conditions and satellite data used in conjunction, the predictive model would be real and complete as an attempt to allow policymakers and farmers to make better choices. In addition, privacy-preserving data-sharing solution studies such as Federated Learning allow secure group analysis of the data, and therefore smart agriculture solutions are justifiable to apply. it facilitates an accuracy agriculture using a high productive, adaptive, and scalable machine learning model for crop yield prediction. With the influence of climate change on farm conditions and limited resources, upcoming research will be required to develop adaptive algorithms, apply real-time environmental monitoring, and translate models. The agricultural sector will thus be in a position to utilize the full potential of artificial intelligence for future production, sustainability, as well as food security.

REFRENCES

[1] Jeong, J., Resop, J. P., Mueller, N. D., et al. (2016). "Random Forest and Gradient Boosting for Crop Yield Prediction," Computers and Electronics in Agriculture, vol. 123, pp. 33-45.
[2] Li, S., Wang, J., & Guo, H. (2018). "Improving Crop Yield Predictions with Dynamic Learning Rates in Boosting Models," Journal of Agricultural Informatics, vol. 10, no. 4, pp. 72-85.
[3] Awan, U., Naveed, A., & Ali, M. (2021). "Feature Importance in Agricultural Yield Prediction Using Random Forest," IEEE Access, vol. 9, pp. 12345-12358.
[4] Singh, P., Kumar, A., & Raj, S. (2019). "A Hybrid Model for Crop Yield Prediction Using Random Forest and Neural Networks," IEEE Transactions on Computational Agriculture, vol. 13, no. 2, pp. 58-66.
[5] Sarkar, S., Das, K., & Basu, P. (2020). "Comparison of Machine Learning Techniques for Crop Yield Prediction," International Journal of Data Mining and Bioinformatics, vol. 15, no. 1, pp. 79-92.
[6] Zhao, J., Xue, Y., & Yang, W. (2019). "Application of XGBoost and LightGBM for Agricultural Yield Prediction," Journal of Agricultural Engineering Research, vol. 115, no. 7, pp. 225-234.
[7] Ahmed, A., Ali, M., & Hassan, S. (2021). "Yield Prediction in Climate-Variable Regions Using XGBoost," Journal of Applied Machine Learning, vol. 17, no. 6, pp. 103-110.
[8] Li, Q., & Zhang, X. (2017). "Neural Networks and Ensemble Models for Crop Yield Forecasting," Computers and Electronics in Agriculture, vol. 134, pp. 100-109.
[9] Chen, X., Wang, Y., & Xu, L. (2022). "Dynamic Learning Rates in Boosting Algorithms for Agricultural Applications," IEEE Access, vol. 10, pp. 15634-15645.
[10] Patel, R., Verma, S., & Gupta, A. (2020). "A Comparative Study of Ensemble Learning Techniques for Crop Yield Prediction," International Journal of Agricultural Sciences, vol. 28, no. 5, pp. 78-85. .
[11] Sun, H., Wei, Z., Yu, W., Yang, G., She, J., Zheng, H., Jiang, C., Yao, X., Zhu, Y., Cao, W., & Cheng, T. (2025). "SIDEST: A Sample-Free Framework for Crop Field Boundary Delineation by Integrating Super-Resolution Image Reconstruction and Dual Edge-Corrected Segment Anything Model," Computers and Electronics in Agriculture, vol. 230, pp. 109897.
[12] Dembani, R., Karvelas, I., Akbar, N. A., Rizou, S., Tegolo, D., & Fountas, S. (2025). "Agricultural Data Privacy and Federated Learning: A Review of Challenges and Opportunities," Computers and Electronics in Agriculture, vol. 232, pp. 110048.

[13] Vahidi, M., Shafian, S., & Frame, W. H. (2025). "Multi-Depth Soil Moisture Estimation via 1D Convolutional Neural Networks from Drone-Mounted Ground Penetrating Radar Data," Computers and Electronics in Agriculture, vol. 232, pp. 110104.

[14] Raj, M. R. H., Karunanidhi, D., Rao, N. S., & Subramani, T. (2025). "Machine Learning and GIS-Based Groundwater Quality Prediction for Agricultural Practices—A Case Study from Arjunanadi River Basin of South India," Computers and Electronics in Agriculture, vol. 229, pp. 109932.

[15] Li, M., Lu, C., Lin, M., Xiu, X., Long, J., & Wang, X. (2025). "Extracting Vectorized Agricultural Parcels from High-Resolution Satellite Images Using a Point-Line-Region Interactive Multitask Model," Computers and Electronics in Agriculture, vol. 231, pp. 109953.