International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6, 2025
https://theaspd.com/index.php

Stnbl-A Crop Yield Prediction Framework for Assam Using Stacking and Blending Methodology

Pinky Saikia Dutta[1] ,Amrita Bose Paul[2] ,Subhrajyoti Bordoloi[3]
[1,2,3]Assam Engineering College, Assam Science and Technology University, Assam,India.
pinkysdutta78@gmail.com , amrita.ca@aec.ac.in, subhra.ca@aec.ac.in

Abstract *Agriculture forms the backbone of Assam's economy, employing nearly 60% of the state's workforce. One of the enduring challenges faced by the Assam agricultural sector is accurately predicting crop yields. Generally, the local farmers decide the cropping pattern based on their past experiences, thus avoiding the adoption of alternative crops due to economic uncertainties. In the recent literature, machine learning (ML) techniques for crop yield predictions have been applied to a few states across India. However, limited studies have considered the unique climatic conditions specific to Assam. To address this issue, this work introduces a structured, region-specific dataset and a crop yield prediction framework that is built on top of the newly generated dataset. This crop yield prediction framework, named as STnBL (Stacking and Blending) is developed considering specific climatic conditions of Assam and thus enabling more accurate and locally relevant outcomes. The performance of STnBL is analyzed using the coefficient of determination (R^2) and runtime efficiency. The results highlighted that integrating individual models within an averaging ensemble model improved prediction accuracy from 60 % to 96%. Furthermore, the comparative analysis demonstrated that heterogeneous ensemble models, using stacking and blending, exhibited better performance compared to homogeneous ensemble approaches for crop yield prediction.*
Keywords: *crop yield prediction; machine learning; Blending*

1. INTRODUCTION

Agriculture is crucial to the growth and sustainability of the Indian economy, employing nearly 60% of the population and accounting for approximately 18% of the national GDP[1]. Assam, being predominantly an agrarian state, relies heavily on its agricultural productivity for economic development. Approximately 54.11%[2] of Assam's total land area is under cultivation, resulting in agriculture becoming the primary land use. Moreover, about 80% of the state's population, including those employed in plantation industries, are either indirectly or directly reliant on agriculture [2]. Therefore, the adoption of modern technologies for crop yield prediction by farmers is crucial to fulfill the growing demand for food and to ensure food security. Traditionally, rice yield prediction in Assam mostly depends on observations of rainfall and temperature trends[3]. Furthermore, it has also become increasingly challenging to anticipate crop yield that is solely based on climatic changes such as gradually decreasing subsurface water levels, irregular rainfall, temperature, use of pesticides, and more [4]. As an agricultural state, the economy of Assam is significantly influenced by its crop yields [5]. Agriculture-related machine learning applications [6] are still in their initial stages of development, and thus, further research is required to fully utilize this technology. The prediction model developed for other states of India may not be directly relevant to Assam due to its distinct climatic conditions, which differ greatly from those observed in other regions of the country. Moreover, the data needed to implement various machine learning models for crop yield prediction for Assam is not easily accessible in a structured manner.

The authors of the work in [7], used several classification algorithms viz, k-Nearest Neighbour (kNN), SVM, Decision Tree Classifier (DTC), Gradient Boosting Classifier (GBC), Random Forest Classifier (RFC), and Artificial Neural Network (ANN) for crop yield prediction. The results reveal that the performance of the ANN outperformed other models. Although the models are compared based on K-fold cross-validation accuracy, other critical aspects such as testing and training times, as well as the impact of altering hyperparameters, have not been included for a thorough assessment.

The authors in [8] present the results of crop yield prediction for 15 districts of Assam using ANN techniques and Stepwise Multiple Linear Regression (SMLR). The ANN models outperformed the SMLR models in terms of prediction, as evidenced by their higher R^2 values. However, their analysis is limited to two particular crops, viz, rapeseed and mustard, despite the presence of a wide range of crops cultivated in Assam.

The authors of the work in [9], have assessed different weight initialization techniques for the ANN-Multilayer Perceptron (ANN-MLP) model for predicting rice crop production in the Barak Valley Zone

(BVZ) of Assam. To enhance predictive performance, a hybrid GA-ANN model is developed by optimizing the connection weights of the ANN-MLP using a Genetic Algorithm (GA). Empirical results suggested that the proposed GA-ANN hybrid model significantly outperforms the conventional ANN-MLP model in terms of prediction accuracy for both area and crop production in the BVZ. However, the work compares the GA-ANN model with a standard ANN-MLP, but it does not evaluate its performance against other available advanced or hybrid optimization techniques[10][11]. Their study is limited to a specific cultivation zone, but not in general. Furthermore, the work has been carried out only for rice, despite the availability of various other crops in BVZ.

Using weekly weather data and historical yield data, the authors in [12] evaluated district-level rice yield forecasting models for 13 districts in the Brahmaputra Valley of Assam. The models were developed at two key crop growth stages: the vegetative stage (F1) and the mid-season stage (F2), using a modified Hendrick and Scholl technique. Model performance is evaluated using R-squared (R2) and the lowest percentage error. Stepwise regression is employed for model fitting. The analysis revealed that rainfall, in combination with relative humidity and maximum temperature, played a more crucial role in yield prediction for districts located in the lower Brahmaputra Valley. Furthermore, the model's applicability and efficiency in predicting the yield of other significant crops in the region are not taken into consideration. These restrict the generalizability of the findings across diverse agricultural systems.

The Authors in [13] present ensemble approaches, including bagging, boosting, and stacking, to integrate multiple models to improve agricultural yield prediction accuracy. The study found that an ensemble of Random Forest and Gradient Boosting Regressors outperformed individual models, achieving higher accuracy and lower prediction errors.

The authors in Paper [14] analyze the effectiveness of several ensemble techniques for crop yield prediction, specifically Random Forest, XGB Regression, and a Stacking Regressor that combines the two. The efficiency of these models is evaluated using L1 (LASSO) and L2 (Ridge) regularization metrics. Among the approaches tested, XGB Regression demonstrated the highest predictive accuracy. While model performance has been primarily evaluated using R^2, root mean squared error (RMSE), and mean squared error (MSE). Computational efficiency, an equally important criterion for comparing ML and stacking models, has not been taken into consideration.

To predict the yield of potatoes, rice, wheat, and maize, the authors of [15] developed four different machine learning models: Random Forest Regressor, Support Vector Machine (SVM), Gradient Boosting Regressor, and Decision Tree Regressor. Among these, the Decision Tree Regressor exhibited the highest predictive accuracy. Their work has been focused solely on specific crops, namely potatoes, rice, wheat, and maize, with a focus on yield prediction.

In [16], the authors used deep learning models such as Long Short-Term Memory (LSTM) and Gradient Descent-based techniques, together with machine learning models like RF, SVM, and Lasso Regression. According to their findings, machine learning models outperformed deep learning models, which they attributed to the high data requirements typically associated with deep learning models. While the current research provides valuable insights, it does not include a comprehensive analysis or forecasting of the various factors influencing crop yield. This limits the ability to fully understand the dynamic relationships between environmental, agronomic, and socioeconomic factors that affect agricultural productivity.

To summarize the shortcomings/limitations of the recent work, as mentioned in [7][8][9][12][13][14][15][16] are as follows:

- A limited scope has been considered for the regional crops of Assam, such as lentil, wheat, Jute, sugarcane, cotton, black gram, caster, and summer rice.
- Although a few hybrid or ensemble models have been developed for crop yield prediction, their performance has not been rigorously evaluated against other advanced or hybrid techniques in the literature, particularly in terms of computational efficiency.
- The selection of appropriate base learners and their optimization for model evaluation has not been considered during performance evaluation.
- Furthermore, the performance of the models available in the literature is not evaluated for their prediction purposes on local crop varieties within the context of Assam.

To address these issues/shortcomings, our work proposes the development of a hybrid framework that integrates linear, tree-based, homogeneous, and heterogeneous approaches using multiple ensemble techniques such as bagging, boosting, stacking, and blending. The newly proposed model is named as

Stacking and Blending (STnBL) framework for crop yield prediction in Assam.

The contributions of this paper are as follows:

- Generation of crop yield prediction dataset by combining raw data collected from Krishi Bhawan, Assam, and climate data from the Meteorological Department of Assam.
- Implementation and analysis of different machine learning models viz, linear model, homogeneous model, heterogeneous models on top of generated data sets to predict crop yield production for the state of Assam.
- Design of a novel crop yield prediction framework named STnBL that integrates linear, tree-based, homogeneous, and heterogeneous models using bagging, boosting, stacking, and blending techniques. Homogeneous ensembles include bagging and boosting state-of-the-art algorithms such as Random Forest (RF), Gradient Boosting, LightGBM, XGBoost, and CatBoost. For novel heterogeneous ensembles, a diverse set of base learners, including Multiple Linear Regression (MLR), Ridge Regression (RR), Random Forest (RF), and Bayesian Ridge Regression (BRR), is integrated using ensemble techniques, such as blending and stacking. This comprehensive approach facilitates a robust comparison of ensemble strategies and their impact on predictive accuracy.
- Explore suitable individual machine learning or ensemble techniques for accurately predicting specific local crop yields, viz, lentil, wheat, Jute, sugarcane, cotton, black gram, caster, and summer rice in Assam.
- A comparative performance analysis of different models, viz., SM-XGRR, SB-XGRR, S0-XGRR, SB-XGR, BM-XGRR, BB-XGRR, B0-XGRR, and BB-XGR, has been carried out to prove the efficiency and efficacy of the proposed STnBL framework.

2. MATERIALS AND METHODS

The experiments were performed in Jupiter Notebook 6.4.5 in an Anaconda environment and were conducted in an Intel(R) Core (TM) i7-1065G7 (4C / 8T, 1.3 / 3.9GHz, 8MB). The libraries used for our experimental implementation are: 1) Numpy, 2) Matplotlib, 3) Sci-kit Learn, 4) Pandas. After reviewing the relevant studies, we selected Random Forest, Multiple linear regression, Ridge Regression, XGBoost, Bayesian Ridge Regression, CatBoost Regressor, and LightGBM as the machine learning algorithms for our initial investigation.

2.1 Data collection

The crop yield datasets used in this analysis were obtained from the Government Office of Krishi Bhawan, Guwahati, Assam. The dataset covers the period from 2016 to 2022 for 27 districts of Assam. It comprises variables such as district, year, crop type, production (in metric tons), yield (in kilograms per hectare), and cultivated area (in hectares). For 15 major crops, data were collected, including Arhar, Black gram, Castor, Cotton, Gram, Green gram, Jute, Lentil, Linseed, Maize, Autumn rice, Winter rice, Summer rice, Wheat, and Sugarcane. The data collection process was particularly challenging due to inconsistencies in data availability across districts. In addition to agricultural records, meteorological data, specifically average annual rainfall (in millimeters) for each district, were sourced from the Indian Meteorological Department, Azara, Assam. To enhance model performance, rainfall data, which is crucial for agricultural productivity, was integrated with crop data. The work in [17], Employed Support Vector Machines (SVM) to predict rainfall, incorporating variables like humidity and temperature to enhance the precision of crop forecasting. The authors in [2] observed that while rainfall variability negatively impacts autumn and winter rice productivity, increased temperature fluctuations were found to be beneficial for Assam. Similarly, the authors in [18] reported that the yield of potatoes, winter rice, and summer rice was nonlinearly affected by the daily average mean temperature.

Machine learning applications need an enormous dataset. Trees serve as the raw material used in the production of paper. Similarly, data can be viewed as the raw material used to create knowledge[19]. Therefore, we need to process the data into the required form, a step known as pre-processing. The input features from various sources are combined into a single dataset. This final dataset contains the District name, Year, Item, Average rainfall per year, Area, and Production, as shown in Table 1.

Table 1: Data Set Description

Attributes	Description
Year	Data collected from the years 2016-2022.
District	Various Districts of Assam.

Item	The crops collected are Arhar, Castor, Cotton, Green gram, Jute, Lentil, Linseeds, Maize , Autumn Rice, Summer Rice, Wheat and Sugarcane.
Area	The total area of each crop in Hectares.
Average rainfall per year	The average annual rainfall in each district per Year(in mm).
Production	The production of each crop in metric tons(MT).

The following are the preprocessing stages that were carried out:

- Data Conversion to CSV Format: The initial raw data collected was transformed into CSV format for easier processing and analysis.
- Converting Object Data Types to Integer or Float: Since Python cannot handle object data types for mathematical operations, we need to convert any columns with object data types into integer (int) or floating-point (float) types, depending on the data.
- Handling Missing Data: Handling missing data appropriately is essential because it can produce erroneous results. To begin, we check for missing values using df.isnull().sum() to identify columns with null values. Then, to replace these missing values, we calculate the median of the respective feature and use it to fill in the missing data, ensuring that the dataset remains complete for further analysis.
- Normalization: Normalization is performed using a Min-Max Scaler to transform the data into a range between 0 and 1. This process is applied to features to ensure that their values are scaled to fall within the range of 0 to 1.

2.2 Train and test model

After preprocessing, the dataset is divided into the training dataset (70%) and the testing dataset (30%). To learn patterns and relationships within the data, a training set is provided to train the model. Once the model is trained, it is evaluated using the testing dataset to assess its performance and generalization ability, ensuring that the model does not overfit the training data**.** Once the model is trained and tested, the strengths of different approaches are utilized to create an ensemble model to improve performance. Figure 1 shows the STnBL framework for crop yield prediction.

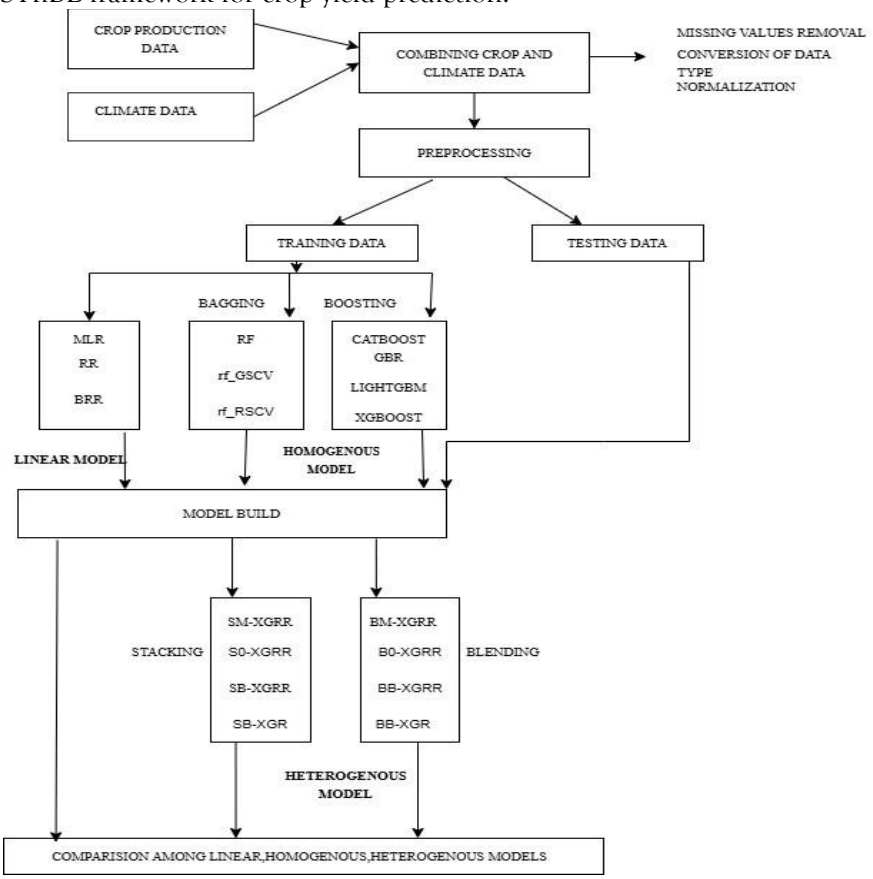


Figure 1: STnBL Framework for Crop Yield Prediction Model

2.3 Heterogeneous Ensemble Model:

A machine learning approach that combines multiple models to produce a stronger, more accurate prediction model is called an ensemble model**.** In this model, the views of several models are aggregated to improve overall performance instead of relying on one. Two common techniques to build a heterogeneous model are stacking and blending. Stacking is an ensemble learning technique that combines various heterogeneous models to enhance predictive performance. Stacking involves using the same training set of data to train several base models, sometimes referred to as first-level models [20]. The base model predictions are fed as input to a higher-level model known as a meta model. As a result, the meta-model typically performs better overall and frequently produces a superior model than all of the intermediate models combined [**13**].

An ensemble learning strategy that enhances overall predictive performance by combining the predictions of several base models is a blending technique. Blending typically employs a hold-out validation set to train the meta-model, whereas stacking trains the meta-learner on out-of-fold predictions from the base learners to prevent data leakage. In this configuration, the meta-learner is trained using the base models' predictions on a different validation set after they have been trained on the training set.

2.3.1 Development of Heterogeneous Models

In the development of heterogeneous ensemble models, base learners comprise linear models, such as MLR, RR, and BRR, as well as tree-based models like XGBoost and RF. A combination of a linear algorithm and a boosting algorithm is used to construct stacked models. XGBoost demonstrated the highest coefficient of determination among various evaluated boosting methods. Thus, in the final ensemble framework, XGBoost was selected as the boosting component. XGBoost outperforms traditional algorithms such as linear regression, support vector machines, and random forests, as supported by the authors in [21]. Eight distinct heterogeneous models were developed using the techniques outlined above and detailed in **Table 2.**

Table 2. Heterogeneous ensemble regressor model

	Base Learner				Meta Learner	
Model	MLR	XGBOOST	RR	BRR	RF	
SM-XGRR	√	√	√		√	STACKING
S0-XGRR		√	√		√	STACKING
SB-XGRR		√	√	√	√	STACKING
SB-XGR		√		√	√	STACKING
BM-XGRR	√	√	√		√	BLENDING
B0-XGRR		√	√		√	BLENDING
BB-XGRR		√	√	√	√	BLENDING
BB-XGR		√		√	√	BLENDING

- .Stacking Configurations:

Training several base models and using their predictions as parameters for a meta-model that generates the final prediction is known as stacking. The following stacking configurations were implemented:

Case 1: MLR, RR with XGBoost as base learner and random forest (RF) as meta learner termed SM-XGRR

Case 2: XGBoost and RR as base learners, with RF as the meta-learner termed S0-XGRR

Case 3: XGBoost, RR, and BRR as base learners, with RF as the meta-learner termed SB-XGRR

Case 4: XGBoost and BRR as base learners, with RF as the meta-learner termed SB-XGR

- Blending Configurations:

Blending is a simpler alternative to stacking, where base model predictions are combined, typically through averaging, to form the final prediction. The following blending configurations were tested:

Case 1: MLR, RR with XGBoost as the base-learner and RF as meta-learner termed BM-XGRR

Case 2: XGBoost and RR as base learners, with RF as the meta-learner termed B0-XGRR

Case 3: XGBoost, RR, and BRR as base learners, with RF as the meta-learner termed BB-XGRR

Case 4: XGBoost and BRR as base learners, with RF as the meta-learner termed BB-XGR

These configurations aimed to leverage the strengths of diverse models, enhancing the ensemble's ability to generalize and improve predictive accuracy. Assessing stacking and blending methods helps recognize the most effective ensemble strategy, achieving an optimal balance between model complexity and

performance. This novel ensemble configuration integrates linear regression and boosting techniques at the base level and utilizes a bagging algorithm as the meta-model. To the best of our knowledge, this combination has not been used in the domain of crop yield prediction before, representing a new direction in the application of ensemble methods in agricultural data science. Figure 2 illustrates the SB-XGRR and S0-XGRR models, a heterogeneous ensemble model composed of linear and boosting models as base learners and bagging as a meta learner.

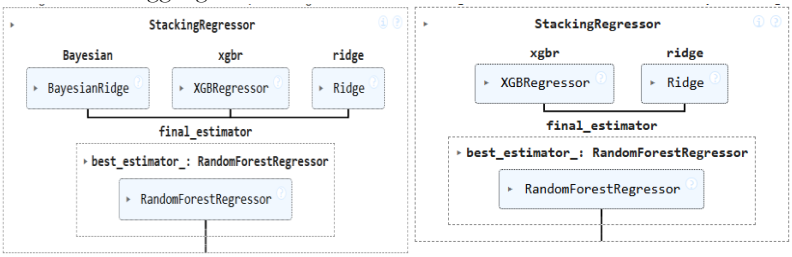


Figure 2. Stacking model SB-XGRR and S0-XGRR

Similarly, the blending technique has been used to achieve several heterogeneous models. These models, built using linear regression, bagging, and boosting methods, and integrated through stacking and blending ensemble approaches, are summarized in Table 2. Stacking and blending are employed to enhance predictive performance by leveraging the complementary strengths of multiple base learners.

2.4 Homogeneous model

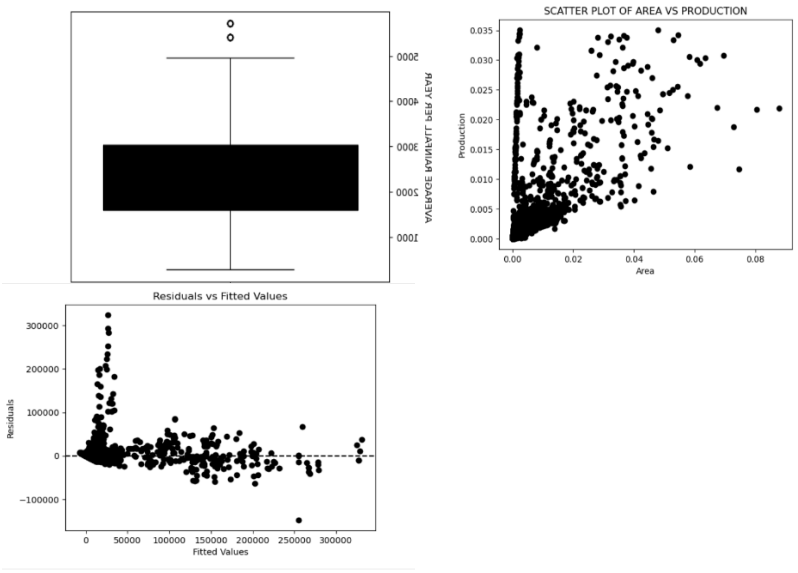
We employed Bagging and Boosting ensemble techniques to construct homogeneous machine-learning models and compare them with heterogeneous ensemble models. Bagging trains several models concurrently using distinct bootstrap samples of the data. It prevents overfitting and lowers variance. In contrast, boosting builds models sequentially, one after the other, with each model attempting to correct the errors of the previous model. Its primary goals are to reduce bias and enhance model accuracy [24]. We employed Random Forest for bagging and Gradient Boosting, CatBoost, XGBoost, and LightGBM for boosting to develop robust predictive models. A hybrid model was built that integrated XGBoost, Random Forest,` and a Decision tree [25]. The hybrid model was built using a voting technique, and it outperformed the individual models. To enhance agricultural productivity prediction, a statistical model that combines MLR and Artificial Neural Networks (ANN) is proposed.

3. RESULTS AND DISCUSSION:

3.1 Exploratory Data Analysis:

We perform exploratory data analysis (EDA) to gain clear insights and values for every numerical trait, especially focusing on each numerical attribute. Figure 3(a) illustrates the average yearly rainfall for the various regions studied using a boxplot. There are two outliers situated above the upper whisker in the boxplot, showing that some areas experience extraordinarily high rainfall, exceeding 5000 mm annually. The distribution displays a slight right skew, as indicated by the longer upper whisker and the presence of elevated outliers, suggesting that while the majority of regions receive moderate rainfall, some receive unusually high levels. A scatter plot is generated (Figure 3b) to examine the relationship between area under cultivation and production. The graph exhibits an overall upward trend, suggesting that larger areas are typically associated with increased production. However, the relationship is not entirely linear, and there is significant variation in production for comparable area values. A dense cluster of points near the origin reflects small-scale farming, while increasing spread at higher area values suggests heteroscedasticity, violating the constant variance assumption of linear regression. To check the presence of heteroscedasticity, we perform Breusch–Pagan test and we achieved a p-value: 4.174e-11, which is less than ($p < 0.05$), indicating that the assumption of constant variance in residuals is violated. This suggests that a linear regression may be inadequate. Therefore, the need for more flexible, robust, tree-based algorithms, such as hybrid or ensemble modeling strategies like stacking or blending multiple base models, could better capture the underlying data structure and improve prediction accuracy. Figure 3c shows the residuals vs. fitted values plot, and a clear funnel-shaped pattern, indicating increasing variance at lower fitted values. This visual evidence confirms heteroscedasticity, violating the assumption of constant variance in linear regression and necessitating more robust modeling techniques. Table 3 presents a statistical analysis of the numeric dataset.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6, 2025
https://theaspd.com/index.php



(a) (b) (c)

Figure 3. Boxplot of average yearly rainfall, scatter plot of area vs. production, and residuals vs. fitted values plot indicating heteroscedasticity in the regression model.

Table 3. Statistics of Numeric Data

	Year	Item	AVERAGE RAINFALL PER YEAR	Area	Production
count	2318.000000	2318.000000	2318.000000	2318.000000	2318.000000
mean	2019.010354	9.569888	2379.366316	0.004819	0.003368
std	1.998894	5.586924	1062.192862	0.009741	0.006587
min	2016.000000	0.000000	280.400000	0.000000	0.000000
25%	2017.000000	5.000000	1599.150000	0.000127	0.000038
50%	2019.000000	10.000000	2099.700000	0.001052	0.000449
75%	2021.000000	14.000000	3041.600000	0.004274	0.002817
max	2022.000000	18.000000	5715.600000	0.087760	0.035067

3.2 Model evaluation parameter

We conducted a comparative analysis based on the coefficient of determination(R^2) and runtime performance. R^2 is regarded as an assessment criterion for crop yield production forecasts. When compared to Symmetric Mean Absolute Percentage Error(SMAPE), Mean Absolute Percentage Error(MAPE), Mean Absolute Error(MAE), Mean Squared Error(MSE), and Root Mean Squared Error(RMSE), R^2 appears to be the most informative rate in many situations. For this reason, we recommend that R^2 be used as the industry standard statistical metric for assessing regression analyses across all scientific domains [26].

We apply the following equation to determine R^2:

$$R^2 = 1-(SSE/SST) \quad (2)$$

where SST represents the total amount of squared differences, while SSE is the sum of the squared residuals. SSE stands for the variability that the independent variables are unable to explain. We conduct a comparative analysis of linear regression variants and ensemble models for crop yield prediction. The outputs are presented in Table 4.

Table 4. Analysis of Linear regression and Heterogeneous models

Sl No	Model	Training time	Evaluation time	R2
1	Multiple Linear Regression	1.71	0.95	0.602
2	Ridge Regression	1.12	0.90	0.801
3	Bayesian Ridge Regression	1.21	0.95	0.435
4	SM-XGRR	3.72	1.17	0.879
5	S0-XGRR	2.65	1.15	0.891
5	SB-XGRR	2.12	1.93	0.956
6	SB-XGR	2.23	1.56	0.946
7	BM-XGRR	1.87	1.13	0.901
8	B0-XGRR	1.50	1.12	0.891

9	BB-XGRR	0.99	1.13	0.912

3.3 Comparative results on run time performance

An analysis of heterogeneous ensemble models in Table 4 reveals that the training and evaluation times for stacking-based approaches are notably higher compared to those of their blending-based counterparts. This difference highlights a trade-off between computational efficiency and predictive accuracy. From a computational standpoint, blending models are more suitable due to their reduced processing time. However, when model performance is prioritized, particularly in terms of predictive accuracy, the stacking model SB-XGRR demonstrates superior results, making it a more appropriate choice for applications that require accuracy**.** Figure 5 shows the output visualization of the stacking model SB-XGRR, the blending model BB-XGRR, the tree-based model XGBoost, and a linear model RR.

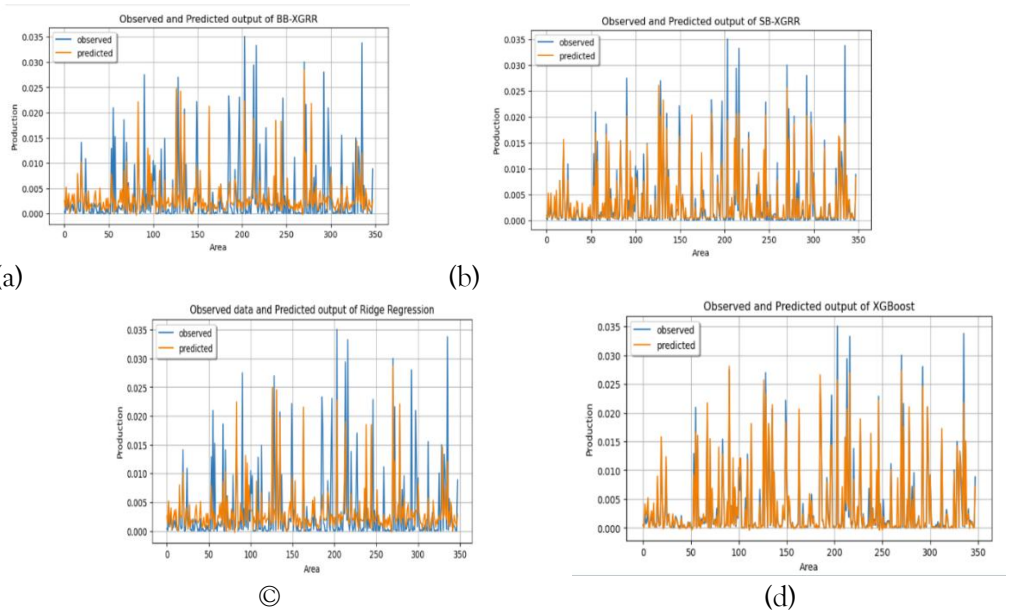


Figure 4. Observed and Predicted Plot of model (a)BB-XGRR, (b)SB-XGRR,(c)RR, and (d) XGBoost

As observed in Table 4, the performance of heterogeneous ensemble models varies based on the choice of base learners:

- When the base learners are MLR, XGBR, and RR, the blending model BM-XGRR outperforms the stacking model SM-XGRR.
- With BRR, XGBR, and RR as base learners, the stacking model SB-XGRR yields better results than BB-XGRR.
- In the case where XGBR and RR are used as base learners, both blending and stacking approaches exhibit comparable performance.
- When the base learners are XGBR and BRR, the stacking model SB-XGR demonstrates superior performance to BB-XGR.

In the evaluation of homogeneous ensemble methods, a comparative analysis between bagging and boosting models indicates that boosting is preferable in terms of both computational efficiency and predictive accuracy. Empirical results depicted in Table 5 show that the training time required for all boosting models is consistently lower than that of their bagging counterparts. The XGBoost model demonstrates the highest accuracy among the evaluated models in terms of predictive performance, as measured by R^2 metric, suggesting that the boosting technique XGBoost, in particular, offers a more optimal balance between computational cost and model performance in homogeneous ensemble settings.

Table 5. Analysis of the Homogeneous Model

Model	Bagging/Boosting	No of estimators	Training time	Testing time	R2
RF	Bagging	10	1.21	0.82	0.428
		20	1.24	0.84	0.430
		30	1.33	0.86	0.431
		40	1.32	0.87	0.320

		50	1.37	0.88	0.433
		60	2.11	0.91	0.432
		70	2.32	0.06	0.435
		80	2.45	0.96	0.435
		90	2.61	0.98	0.541
		100	2.62	0.99	0.602
CatBoost	Boosting	10	1.15	0.77	0.652
		20	1.0	0.49	0.723
		30	0.60	0.35	0.740
		40	0.89	0.72	0.753
		50	0.88	0.88	0.790
		60	1.14	1.09	0.812
		70	0.78	0.45	0.824
		80	1.01	0.84	0.833
		90	0.98	1.86	0.843
		100	1.07	1.01	0.844
LightGBM	Boosting	10	0.59	0.34	0.782
		20	0.57	0.35	0.786
		30	0.72	0.43	0.791
		40	0.96	0.42	0.560
		50	0.95	0.60	0.799
		60	0.85	0.50	0.802
		70	0.69	0.36	0.60
		80	0.84	0.53	0.853
		90	1.28	0.60	0.864
		100	1.07	1.95	0.864
GBR	Boosting	10	0.67	0.54	0.617
		20	0.71	0.55	0.766
		30	0.76	0.57	0.799
		40	0.88	0.55	0.815
		50	0.97	0.47	0.828
		60	0.97	0.46	0.602
		70	0.92	0.59	0.846
		80	0.98	0.54	0.852
		90	0.99	0.38	0.857
		100	0.99	0.54	0.865
XGBR	Boosting	10	0.63	0.40	0.781
		20	0.65	0.51	0.784
		30	0.51	0.37	0.793
		40	0.88	0.55	0.795
		50	0.83	0.50	0.823
		60	0.81	0.63	0.828
		70	0.73	0.58	0.834
		80	0.84	0.50	0.880
		90	0.97	0.84	0.882
		100	0.96	0.77	0.884

Observations on the effect of increasing the value of estimators from 10 to 100 on R^2 values of some of the models are as follows:

With a significant increase in R^2 from 62% to 86%, the Gradient Boosting Regressor showed the most substantial improvement among the models assessed. Strong performance is also shown by CatBoost from 65% to 84% and XGBoost from 78% to 88% , which demonstrated moderate to high increases in predicted accuracy. Despite starting from a lower baseline, Random Forest demonstrated a moderate improvement from 43% to 60%. The LightGBM model, on the other hand, only showed a slight

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6, 2025
https://theaspd.com/index.php

improvement, with R^2 slightly rising from a baseline that is already high, from 78% to 86% as shown in Figure 5.

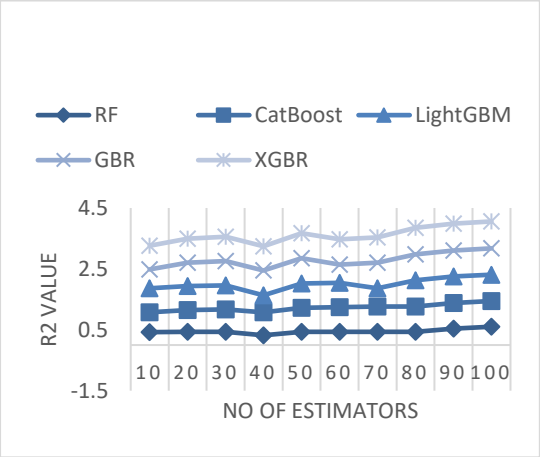


Figure 5. Bagging and Boosting accuracy bagging,

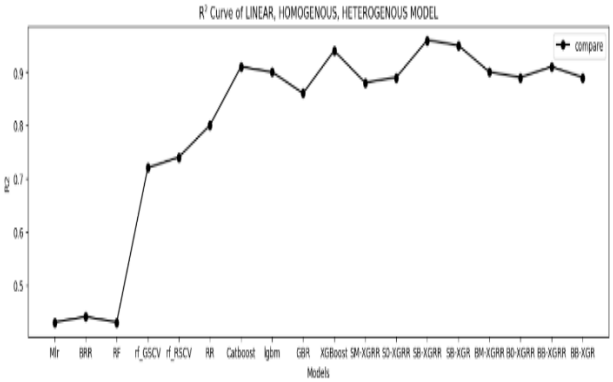


Figure 6. R2 Comparison plot of linear, boosting, stacking and blending Models

Figure 6, shows the R2 value of all the models, the linear model, and ensemble models, including bagging, boosting, stacking, and blending. The highest value is achieved by SB-XGRR with an R^2 value of 96%.

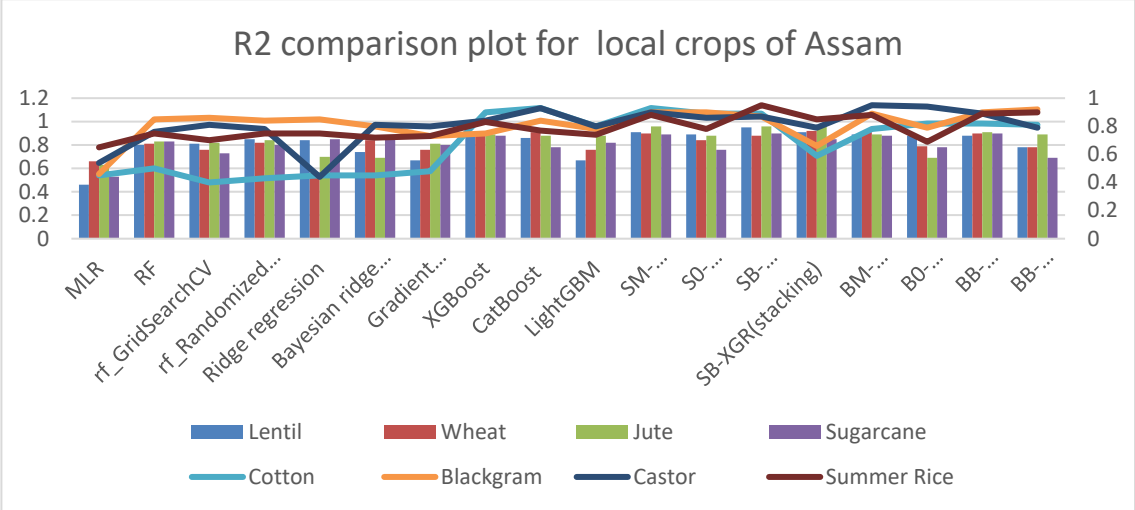


Figure 7. ML and Ensemble model results based on R2 for the local crop of Assam

Figure 7 presents the R^2 values for a few local crops grown in Assam, specifically targeting crops such as lentil, wheat, Jute, sugarcane, cotton, black gram, castor, and summer rice using linear machine learning, homogeneous, and heterogeneous machine learning models. The findings suggest that:

- Multiple Linear regression demonstrated better performance for wheat yield prediction, achieving an accuracy of 0.66. In contrast, its lowest accuracy was found for cotton, with a value of 0.45. Random forest demonstrated strong performance for blackgram yield prediction, achieving a high accuracy of 0.85. In contrast, its lowest accuracy was found for cotton, with a value of 0.50. Random Forest algorithm, when hyperparameter-tuned using GridSearchCV, performed well for blackgram yield prediction, achieving an R^2 score of 0.86. However, its performance was lower for cotton, with an R^2 of 0.40. Alternatively, when hypertuned using RandomizedSearchCV, the model showed high predictive accuracy for lentil with R^2 scores of 0.85 and lowest for cotton with R^2 score of 0.43. Ridge regression demonstrated strong performance for sugarcane, blackgram yield prediction, achieving a high accuracy of 0.85. In contrast, its lowest accuracy was noted for castor, with a value of 0.44. Bayesian ridge regression demonstrated strong performance for wheat yield prediction, achieving a high accuracy of 0.84. In contrast, its lowest accuracy was noted for cotton, with a value of 0.45. Gradient boosting demonstrated strong performance for jute yield prediction, achieving a high accuracy of 0.81. In contrast, its lowest accuracy was observed for cotton, with a value of 0.48. XGBoost demonstrated strong performance for

wheat, cotton, achieving a high accuracy of 0.94. In contrast, its lowest accuracy was recorded for sugarcane, with a value of 0.88. CatBoost demonstrated strong performance for cotton and castor, achieving a high accuracy of 0.93. In contrast, its low accuracy was observed for summer rice with a value of 0.77. LightGBM demonstrated strong performance for jute, achieving a high accuracy of 0.88. In contrast, its low accuracy was observed for lentil, with a value of 0.67. SM-XGRR demonstrated strong performance for jute yield prediction, achieving a high accuracy of 0.96. In contrast, its lowest accuracy was observed for summer rice, with a value of 0.88. SB-XGRR demonstrated strong performance for blackgram yield prediction, achieving a high accuracy of 0.90. In contrast, its lowest accuracy was observed for summer rice, with a value of 0.78. SB-XGRR achieved high performance in predicting Jute yield, reaching an accuracy of 0.96. However, its performance was weakest for black gram and castor, where the accuracy dropped to 0.87. SB-XGR performed notably well in predicting jute yield, attaining an accuracy of 0.96. In contrast, it showed the poorest performance for cotton, with a significantly lower accuracy of 0.59. While BM-XGRR achieved high accuracy in predicting castor yield with an accuracy of 0.95, it performed considerably worse for cotton, registering the lowest accuracy among the crops at 0.78. B0-XGRR performed notably well in predicting castor yield, with an accuracy score of 0.94. However, its performance dropped significantly for jute and summer rice, where it recorded the lowest accuracy of 0.69 among all the crops. BB-XGRR was highly accurate in predicting jute yield, with an accuracy score of 0.91. However, its performance dropped significantly for cotton, where it recorded the lowest accuracy of 0.82 among all the crops. BB-XGR performed notably well in predicting blackgram yield, attaining an accuracy of 0.92. In contrast, it showed the poorest performance for sugarcane, with a significantly lower accuracy of 0.69.

4. CONCLUSION :

This study addresses the challenge of crop yield prediction in Assam,a region with diverse agro-climatic conditions, by evaluating the performance of linear models and tree-based machine learning approaches. To enhance predictive accuracy and generalization, a novel ensemble framework, STnBL, is proposed, combining homogeneous and heterogeneous strategies through stacking and blending. The framework integrates optimized linear base learners and a boosting model, with a bagging-based meta-learner, leveraging the strengths of each: linear models capture underlying trends, while boosting and bagging effectively reduce bias and variance. The optimal tree-based learner is selected through hyperparameter tuning of the number of estimators. This optimized model is subsequently integrated with a linear regressor to construct heterogeneous ensemble variants, enabling a comprehensive comparison of model configurations within the STnBL framework. The STnBL framework demonstrated strong predictive performance across multiple crops in Assam, including lentil, wheat, jute, sugarcane, cotton, blackgram, castor, and summer rice, as evaluated using R^2 scores and computational efficiency. Notably, these crops have remained largely unexplored in the context of crop yield prediction within the agricultural literature of Assam. Furthermore, the framework exhibits promising scalability and potential for national-level application. The key findings are:

- When evaluated primarily on predictive accuracy, the stacking ensemble model denoted as SB-XGRR demonstrated superior performance compared to other ensemble approaches, including traditional bagging, boosting, and blending techniques. The next highest accuracy was achieved by the stacking-based model, SB-XGR. The third performance attained by the ensemble model BB-XGRR, which is based on blending.
- A homogeneous model, such as XGBoost, is efficient in a situation where computation efficiency is a priority. When the number of estimators is between 80 and 100, optimal result is obtained by XGBoost.
- During the training phase on crop-specific datasets to identify the most efficient predictive model for local crop yield estimation in Assam, the SB-XGRR model demonstrated superior performance for lentil, jute, rice, and sugarcane. SB-XGR yielded the best results for wheat and jute, BB-XGRR also performed well for sugarcane; CatBoost and SM-XGRR are found to be effective for both cotton and blackgram. BM-XGRR showed strong performance for castor, and SB-XGRR proved to be the best performer for summer rice.

In the future, additional agronomic variables, such as soil type, nutrient composition, and pH levels, will be the focus to strengthen predictive accuracy further.

Acknowledgements

The authors extend their sincere acknowledgment to the Regional Meteorological Center, Guwahati, Assam, India, and the Krishi Bhavan, Assam, India, for their coordination during the data collection process.

REFERENCES

1. S. Bhuyan et al., Bhuyan, S., JYOTI Medhi, S. & Abonmai, T., (2023) Jan '"Machine Learning-based Crop Recommendation System in Biswanath District of Assam, Biological Forum." Biological Forum – An International Journal, 417–421.
2. Upadhyaya, T.P. (2022) 'Role of Agriculture in Economic Development of Assam', Cognizance Journal of Multidisciplinary Studies, 2(6), pp. 10–21. Available at: https://doi.org/10.47760/cognizance.2022.v02i06.002.
3. Goswami,B ., Hussain,R. Rao,V, U M., & Saikia, U,S., " Impact of climate change on rice yield at Jorhat, Assam", Journal of Agrometeorology 18 (2) : 252-257, December 2016. https:// doi.org/10.54386/ jam. v18i2.944
4. J. Mahajan Mahajan, J., Banal, K. and Mahajan, S. (2021) 'Estimation of crop production using machine learning techniques: a case study of J&K', International Journal of Information Technology, 13(4), pp. 1441–1448. Available at: https://doi.org/10.1007/s41870-021-00653-7.
5. Dutt,P.S., and Tahbilder,H., "Prediction of Rainfall using data mining technique over Assam." Indian Journal of Computer Science and Engineering, May 05, 2014. [Online]. Available: https://www.ijcse.com/docs/INDJCSE14-05-02-081.pdf
6. Katharria, A., Rajwar, K., Pant, M., Velásquez, J.D., Snášel, V., Deep, K., 2024. Information Fusion in Smart Agriculture: Machine Learning Applications and Future Research Directions. https://doi.org/10.48550/ARXIV.2405.17465
7. Mali, B., Saha, S., Brahma, D., Singh, P.K., Nandi, S., 2021. Alternate Crop Prediction using Artificial Intelligence: A Case Study in Assam, in: 2021 IEEE International Symposium on Smart Electronic Systems (iSES). Presented at the 2021 IEEE International Symposium on Smart Electronic Systems (iSES), IEEE, Jaipur, India, pp. 267–270. https://doi.org/10.1109/ises52644.2021.00067
8. Kakati, N., Deka, R.L., Das, P., Goswami, J., Khanikar, P.G., Saikia, H., 2022. Forecasting yield of rapeseed and mustard using multiple linear regression and ANN techniques in the Brahmaputra valley of Assam, North East India. Theor Appl Climatol 150, 1201–1215. https://doi.org/10.1007/s00704-022-04220-3
9. Paswan,R P., Begum, S,A., Hemochandran,L.,2018 GA-ANN Hybrid model for prediction of Area and crop production,Int. J. Agricult. Stat. Sci. Vol. 14, Supplement 1, pp. 15-26, 2018.
10. Yenkikar, A., Mishra, V.P., Bali, M., Ara, T., 2025. An explainable AI-based hybrid machine learning model for interpretability and enhanced crop yield prediction. MethodsX 15, 103442. https://doi.org/10.1016/j.mex.2025.103442
11. Bazrafshan, O., Ehteram, M., Dashti Latif, S., Feng Huang, Y., Yenn Teo, F., Najah Ahmed, A., El-Shafie, A., 2022. Predicting crop yields using a new robust Bayesian averaging model based on multiple hybrid ANFIS and MLP models. Ain Shams Engineering Journal 13, 101724. https://doi.org/10.1016/j.asej.2022.101724
12. Chutia,S., Deka,R.L., Goswami,J., Phukon, M,H., (2021). Forecasting rice yield through modified Hendrick and Scholl technique in the Brahmaputra valley of Assam. J. Agrometeorol. 23, 106–112. https://doi.org/10.54386/jam.v23i1.95
13. Sankareswari,K and Sujatha, G.,"Evaluation of an Ensemble Technique for Prediction of Crop Yield," in Proceedings of the 5th International Conference on Information Management & Machine Intelligence, Jaipur India: ACM, Nov. 2023, pp. 1–9. doi: 10.1145/3647444.3647833
14. Hariyani, G., Singh, A., Patil, P., Kothari, V., Javale, D., 2024. Analysis on Crop Yield Prediction using various Ensemble Methods, in: 2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA). Presented at the 2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA), IEEE, Pune, India, pp. 1–6. https://doi.org/10.1109/ICCUBEA61740.2024.10775263
15. Pant, J., Pant, R.P., Kumar Singh, M., Pratap Singh, D., Pant, H., 2021. Analysis of agricultural crop yield prediction using statistical techniques of machine learning. Materials Today: Proceedings 46, 10922–10926. https://doi.org/10.1016/j.matpr.2021.01.948
16. Jhajharia K, Mathur P, Jain S, Nijhawan S, 2023. Crop Yield Prediction using Machine learning and Deep learning Technique. ELSEVIER, Procedia Computer Science Volume 218, 2023, 406–417. https://doi.org/10.1016/j.procs.2023.01.023
17. Mahendra N, Vishwakarma, D., Mahendra, N., 2020. Crop Prediction using Machine Learning Approaches. IJERT V9, IJERTV9IS080029. https://doi.org/10.17577/IJERTV9IS080029
18. Nath,H, and Mandal,R.," Heterogeneous Climatic Impacts on Agricultural Production: Evidence from Rice Yield in Assam", India, Asian Journal of Agriculture and Development, Southeast Asian Regional Center for Graduate Study and Research in Agriculture (SEARCA), vol. 15(1), June.doi: 10.22004/ag.econ.275687
19. .Oluwatosin Ajayi, V., 2023. A review on Primary sources of data and Secondary sources of Data., European Journal of Education and Pedagogy.2023 ELSEVIER, Jan. 31, 2023. https://doi.org/10.5281/zenodo.15328023
20. Khan, A.A., Chaudhari, O., Chandra, R., 2024. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. Expert Systems with Applications 244, 122778. https://doi.org/10.1016/j.eswa.2023.122778
21. Ravi,R., & Baranidharan, "Crop Yield Prediction using XG Boost Algorithm". Int. J. Recent Technol. Eng. IJRTE. 2020 Jan; vol. 8, no. 5, pp. 3516–3520 :http:// doi 10.35940/ijrte.D9547.018520
22. Nirajan Khadka," Machine Learning Model's Performance", https:// dataaspirant.com/ stacking-technique/
23. Parth Shukhla," How Blending Technique Improves Machine Learning Model's Performance ",https://dataaspirant.com/blending-technique-machine-learning/
24. Ribeiro, M.H.D.M., Dos Santos Coelho, L., 2020. Ensemble approach based on bagging, boosting and stacking for short-

term prediction in agribusiness time series. Applied Soft Computing 86, 105837. https://doi.org/10.1016/j.asoc.2019.105837
25. Manjunath, M.C., Palayyan, B.P., 2023. An Efficient Crop Yield Prediction Framework Using Hybrid Machine Learning Model. RIA 37, 1157–1167. https://doi.org/10.18280/ria.370428
26. Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science 7, e623. https://doi.org/10.7717/peerj-cs.623