Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Optimization of Crop Yield Prediction Models using Hyperparameter Tuning and Ensemble Learning

P. Sowmya[1], Dr. A. V. Krishna Prasad[2]

[1]Assistant Professor, Telangana Mahila viswavidyalayam,(Formerly University college for women), Department of Computer Science, Koti, Hyd, India.
[2]Associate Professor, Maturi Venkata Subba Rao Engineering college, Department of Computer Science, Nadergul, Hyd, India.
Email: soumya.padam@gmail.com

Accurate prediction of crop yields is crucial for effective agricultural planning and management. This study explores the application of machine learning techniques, specifically Gradient Boosting and Random Forest models, to optimize the prediction of rice yield in the Telangana region of India. The research employs a comprehensive approach involving data preprocessing to handle missing values and feature engineering, followed by rigorous model training and evaluation. Hyperparameter tuning using GridSearchCV enhances model performance, ensuring robust predictions. Additionally, an adaptive ensemble technique combines model outputs to further refine yield forecasts. Evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared are employed to assess model accuracy and effectiveness. The findings underscore the efficacy of ensemble learning coupled with hyperparameter optimization in improving agricultural yield predictions, offering valuable insights for agricultural decision-making and resource allocation.

Keywords: Adaptive ensemble, agricultural yield prediction, Gradient Boosting, hyperparameter optimization, machine learning, Random Forest.

1. Introduction

Predicting crop yields is an important part of agricultural planning and decision-making. More food security, informed policymaking, and improved resource allocation are all made possible by accurate forecasts. However, due to several variables like weather, soil fertility, and agricultural methods, properly estimating crop production is a difficult process.[1] Conventional techniques frequently fail to fully capture this complexity, producing forecasts that are less trustworthy. Machine learning advances recently have demonstrated significant promise in tackling these issues. Machine learning algorithms present a viable way to enhance agricultural yield forecasts because of their capacity to learn from big datasets and spot patterns.[2]

Machine learning advances recently have demonstrated significant promise in tackling these issues. Machine learning algorithms present a viable way to enhance agricultural yield forecasts because of their capacity to learn from big datasets and spot patterns. In this article, we investigate the use of machine learning methods to Telangana rice yield prediction using an extensive dataset covering the years 1966 to 2020.

Conventional statistical techniques frequently fall short in capturing the intricate interconnections and nonlinear correlations among the variables influencing crop productivity. Another major difficulty is the availability of complete and reliable data. The situation becomes more complex when there are missing data, measurement mistakes, and variability in the data. Strong tools are available for examining huge datasets and identifying intricate patterns through machine learning (ML). Predictive analytics using machine learning models, like Random Forest and Gradient Boosting, has been effectively implemented in several industries, including agriculture.[3] These models are appropriate for predicting agricultural yields as they can manage non-linearity and feature interactions.[4]

The primary objective of this study is to enhance the accuracy of rice yield predictions in the Telangana region by leveraging machine learning techniques and developing an adaptive ensemble model. To achieve this, we set the following specific goals: handle missing values and inconsistencies in the dataset by replacing zero values with the mean and converting the 'Year' column to a numerical format; train individual Gradient Boosting and Random Forest models using the historical dataset; use GridSearchCV to optimize the hyperparameters of the models to achieve the best performance; develop an adaptive ensemble approach to combine predictions from the trained models based on their mean squared error (MSE) to improve overall prediction accuracy; and evaluate the performance of the individual models and the ensemble model using standard metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. This study focuses on predicting the rice yield in the Telangana region using data from 1966 to 2020. The features considered include climatic variables, soil properties, and agricultural practices.

This study makes a few important additions. It begins by presenting a brand-new adaptive ensemble learning technique that combines predictions made by Random Forest and Gradient Boosting models according to their individual errors. By utilizing the advantages of both models, this strategy can increase prediction accuracy. Second, as compared to individual models, the adaptive ensemble model created in this study shows considerable increases in prediction accuracy. Errors are decreased and prediction reliability is increased with the help of the weighted combination of model predictions. Thirdly, the study offers a thorough assessment of the models based on a variety of performance measures, providing information about how well machine learning models work in agricultural settings. Lastly, farmers, agricultural planners, and politicians can use the research's practical consequences.

2. Literature Review:

K. P. K. Devan et al. combined Random Forest and Logistic Regression to predict crop yield and recommend fertilizer based on soil characteristics and weather conditions, highlighting the role of machine learning in agriculture. Their study emphasizes the importance of accurate

yield forecasts for agricultural policies but does not explore other potential machine learning algorithms. Additionally, the study lacks details on factor prioritization and system scalability across different crops and regions.[5]

S. Pavani et al. developed a hybrid model combining Support Vector Machine (SVM) and Random Forest (RF) to predict rice and sorghum yields in India, achieving 90% accuracy. The model used RF for data selection and weighted parameters for SVM training, outperforming traditional methods by 8-10%. However, its focus on specific crops and seasons limits generalizability. The study did not explore newer machine learning techniques, comprehensive evaluation metrics like precision and recall, or the computational complexity and scalability for large-scale implementation. [6]

S. Thirumal and R. Latha introduced the HTSAE-RCYP model to predict rice yield, aiming to improve agricultural productivity and food security in India under varying weather conditions. The model includes preprocessing steps like data conversion, null value removal, and normalization, and employs a Stacked Autoencoder (SAE) with RMS Prop optimizer for accurate yield forecasts. Experimental results show the HTSAE-RCYP model's superiority over existing models in agricultural yield prediction. However, it lacks details on computational resource requirements, interpretability of predictions, and the impact of external factors like pests and soil quality. Additionally, the model's scalability to large datasets is not addressed, which may limit its application in extensive agricultural settings.[7]

Piyal Ekanayake et al. created crop-weather models for paddy yield in Sri Lanka using nine weather indices such as rainfall, humidity, temperature, and more. Random Forest (RF) was utilized to assess the importance of each index, with minimum relative humidity and maximum temperature identified as the most influential. RF's prediction model, compared to Power Regression (PR), Multiple Linear Regression (MLR), and stepwise selection, proved reliable with a high correlation coefficient (R of 0.99) and a low Mean Absolute Percentage Error (MAPE of 1.4%). However, the study's focus on specific regions and crops, restricted data access, expertise required for RF implementation, and lack of consideration for socio-economic factors and scalability present limitations.[8]

Ajitha Antony and Ramanathan Karuppasamy explored paddy yield prediction with machine learning, emphasizing the identification of key features affecting production. They developed five regression algorithms using 16 variables from the soil health card, aiming to enhance performance through various approaches. The models were validated using Monte Carlo methods, with the XGBoost-ensembled Random Forest achieving the highest prediction accuracy of 86%. This study is pioneering in using soil health card features for paddy yield prediction, providing a valuable tool for farmers and agronomists to optimize cultivation and maximize yield.[9]

Warut Pannakkong et al. utilized Response Surface Methodology (RSM) to fine-tune hyperparameters for three machine learning algorithms: Artificial Neural Network (ANN), Support Vector Machine (SVM), and Deep Belief Network (DBN). The goal was to demonstrate RSM's effectiveness in maintaining ML performance while reducing the number of runs needed to find optimal hyperparameters, compared to Grid Search (GS). The algorithms were tested on a Thai food producer's dataset to predict raw material quality, using K-fold cross-validation for robustness. Mean Absolute Error (MAE) of the validation set was

used as the prediction accuracy metric. The results showed similar prediction accuracy between the GS and RSM-tuned algorithms, with RSM providing more reliable hyperparameter settings and reducing the number of required runs by 97.79% for ANN, 97.81% for SVM, and 80.69% for DBN.[10]

3. Methodology:

(a) Dataset Description

The dataset used for this analysis spans from 1966 to 2020 and focuses on the Telangana region of India. It comprises a total of 506 entries with 16 attributes, including 'Dist Code', 'Year', 'State Name', 'Dist Name', 'RICE IRRIGATED AREA (1000 ha)', 'TOTAL CONSUMPTION (tons)', 'ANNUAL RAINFALL (Millimeters)', 'Precipitation Avg', 'Max Temp', 'Min Temp', 'Wind Speed', 'Wind Direction', 'Surface soil wetness', 'Profile soil Moisture', 'Root Zone Soil Wetness', and 'RICE YIELD (Kg per ha)'. This comprehensive dataset incorporates various climatic factors, soil moisture levels, and agricultural practices that significantly influence rice yield.

The data was meticulously collected from multiple sources, including NASA, ICRISAT, and the Telangana Open Data Portal, ensuring a rich and diverse set of information. This amalgamation of data provides a robust foundation for analyzing the factors affecting rice yield in the region. By leveraging machine learning techniques, the dataset aims to predict rice yield more accurately, aiding farmers and agronomists in making informed decisions regarding crop management and improving agricultural productivity. The detailed attributes such as soil moisture levels, temperature variations, and rainfall data are crucial for understanding the environmental conditions impacting rice cultivation and yield outcomes.

Table 1. Annual rice yield and environmental factors in Mahabubnagar district from 2010 to 2020

Dist Code	Year	State Name	Dist Name	Rice Irrigated	Total Consump	Annual Rainfall	Precipitation	Max Temp	Min Temp	Wind_Speed	Wind_Direction	Surface_soil	Profile_soil	Root_Zone	Rice Yield
58	2010	Telangana	Mahabubnagar	192.71	143591	755.1	2.45	43.24	9.74	0.06	235.56	0.53	0.64	0.65	2774.63
58	2011	Telangana	Mahabubnagar	168.32	136767	494.4	1.26	42.69	11	0.09	254.5	0.43	0.56	0.58	2304.64
58	2012	Telangana	Mahabubnagar	137.36	106717	632.7	1.72	44.33	10.56	0.02	266.19	0.41	0.57	0.59	2679.01
58	2013	Telangana	Mahabubnagar	164.19	144946	910.8	2.26	44.17	11.73	0.05	246.12	0.52	0.61	0.62	2838.76
58	2014	Telangana	Mahabubnagar	157.2	124485	601.1	1.98	42.78	12.7	0.05	206.38	0.48	0.57	0.61	2600.61

58	2015	Telangana	Mahabubnagar	92.58	149617	471.9	1.44	44.09	10.93	0.02	215.69	0.4	0.54	0.58	2233.28
58	2016	Telangana	Mahabubnagar	139.23	132783	567	1.77	43.75	12.22	0.08	265.38	0.45	0.58	0.6	2696.59
58	2017	Telangana	Mahabubnagar	168.46	117612	486	1.89	43.36	12.51	0.06	271.19	0.45	0.58	0.61	2459.12
58	2018	Telangana	Mahabubnagar	154.34	116783	831	1.16	44.16	11.31	0.05	266.62	0.38	0.54	0.57	2420
58	2019	Telangana	Mahabubnagar	163.37	127658	756	1.57	44.48	10.64	0.04	219.06	0.43	0.55	0.59	2490
58	2020	Telangana	Mahabubnagar	145.28	135647	894	3.08	43.56	11.29	0.03	190.12	0.58	0.66	0.69	3380

The rice yield and related climatic and environmental factors in Mahbubnagar district from 2010 to 2020 show significant variations (Table 1; ICRISAT).It includes features such as rice irrigated area, total consumption, annual rainfall, various temperature and wind metrics, soil wetness parameters, and the target variable, rice yield (kg per hectare).

(b) Proposed Architecture

This diagram illustrates a systematic approach to constructing and validating machine learning models, focusing on enhancing the accuracy and reliability of predictions. Multiple decision trees are built during training in the Random Forest ensemble learning technique. Every tree is trained using a random feature selection and a subset of the data. In classification problems, the mode of the classes predicted by separate trees determines the final prediction. It takes the average of each tree's predictions for regression tasks. This method's variance-mitigating averaging technique makes it efficient for handling huge datasets and lowers the likelihood of overfitting.

Another ensemble technique is the gradient boosting machine, which, unlike Random Forest, produces models sequentially rather than separately. Using a decision tree as a starting point, GBM iteratively creates new models by fixing the mistakes in the original weak learner. The goal of this iterative approach is to decrease prediction residual errors and increase the model's overall accuracy. GBM is well-known for its capacity to identify intricate relationships in data and for producing predictions with a high degree of accuracy.

To identify the ideal parameters for a machine learning model, GridSearchCV use cross-validation to systematically search through a predetermined grid of hyperparameters. In order to make sure the model is optimized for maximum performance and generalization on untested data, it assesses the performance of each combination and chooses the one that yields the best results.

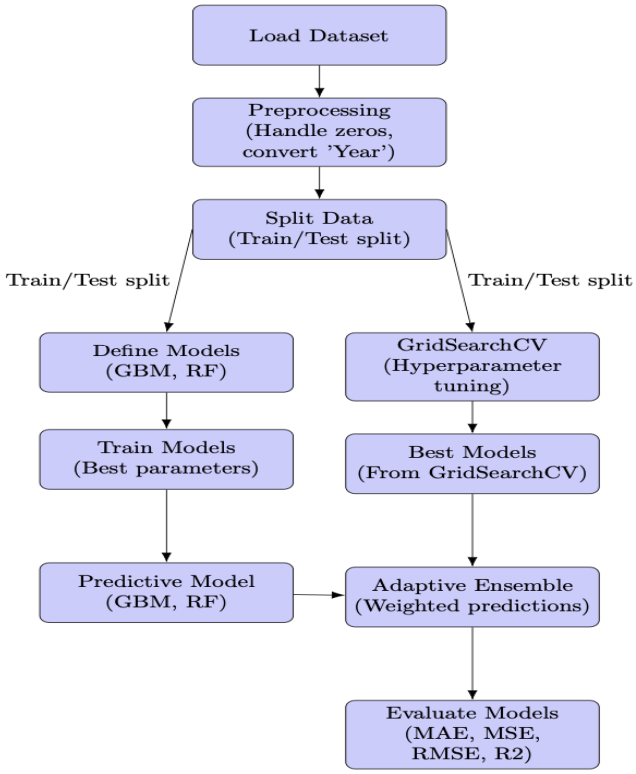

Fig 1 Workflow for Developing and Evaluating Machine Learning Models

The flowchart in Figure 1 illustrates the methodology used for building and evaluating the predictive model for rice yield using machine learning techniques. A thorough approach for creating and assessing machine learning models is shown in this diagram. The first step in the process is loading the dataset. Next comes the preprocessing stage, when duties like resolving missing values and converting data types are completed as well as data cleaning and transformation. After that, the data is divided into testing and training sets to provide a reliable model evaluation.

Various machine learning models, including Random Forest (RF) and Gradient Boosting Machine (GBM), are defined during the modelling phase. GridSearchCV is used to do hyperparameter tuning on these models in order to determine the ideal settings for maximum performance. The chosen models are trained using the training set, and the top models are determined by evaluating how well they perform during the tuning phase.

The predictive models are then used to make predictions, which are further refined using an adaptive ensemble method that combines weighted predictions from different models. Finally, the models are evaluated using various performance metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) to ensure their accuracy and reliability.

(c) Algorithm

The detailed algorithm used for the adaptive ensemble method for crop yield prediction is outlined in Algorithm 1 (Figure 2).To estimate rice yields, the Random Forest (RF) and Gradient Boosting Machine (GBM) models are used in the Adaptive Ensemble Method for Crop Yield Prediction (AEM-CYP). To produce precise crop yield estimates, it preprocesses the data, using GridSearchCV to identify the best models, and then creates an ensemble prediction using weighted averaging.

Algorithm 1 Adaptive Ensemble Method for Crop Yield Prediction (AEM-CYP)

Require: Dataset $D = \{(X, y)\}$ {Input data with features X and target y}

Load and Preprocess Data:
Load dataset D.
Drop 'State Name' column.
Replace zeros with column means for all features except 'Year' and 'Dist Name'.
Convert 'Year' to datetime and extract year.
Drop any remaining missing values.
Define Features and Target:
Features:

$$\begin{aligned} X = \{&\text{Year}, \text{Rice Irrigated area}, \text{Consumption}, \text{Rainfall}, \\ &\text{Precip Avg}, \text{Max Temp}, \text{Min Temp}, \text{Wind Speed}, \\ &\text{Wind Dir}, \text{Surface soil wet}, \text{Profile soil Moist}, \text{Root Zone Moist}\} \end{aligned}$$

Target: $y = \text{Rice yield}$
Split Data:
Split D into training D_{train} and testing D_{test}
Model Initialization and Hyperparameter Tuning:
Initialize $\mathcal{M}_{\text{GBM}}$ (Gradient Boosting) and $\mathcal{M}_{\text{RF}}$ (Random Forest).
Define hyperparameters:
$G_{\text{GBM}} = \{n_{\text{estimators}} : [100, 200, 300], \alpha : [0.05, 0.1, 0.2], d_{\max} : [3, 5, 7]\}$
$G_{\text{RF}} = \{n_{\text{estimators}} : [100, 200, 300], d_{\max} : [5, 10, 15], s_{\min} : [2, 5, 10]\}$
Perform GridSearchCV on G_{GBM} and G_{RF} using 5-fold cross-validation.
Select best models $\mathcal{M}^*_{\text{GBM}}$ and $\mathcal{M}^*_{\text{RF}}$.
Train and Predict with Best Models:
Train $\mathcal{M}^*_{\text{GBM}}$ and $\mathcal{M}^*_{\text{RF}}$ on D_{train}.
$\hat{y}_{\text{GBM}} \leftarrow \text{Predict}(\mathcal{M}^*_{\text{GBM}}, D_{\text{test}})$
$\hat{y}_{\text{RF}} \leftarrow \text{Predict}(\mathcal{M}^*_{\text{RF}}, D_{\text{test}})$
Calculate Weights for Ensemble:
Compute MSE_{GBM} and MSE_{RF}.
$w_{\text{GBM}} \leftarrow \frac{\text{MSE}_{\text{RF}}}{\text{MSE}_{\text{GBM}} + \text{MSE}_{\text{RF}}}$
$w_{\text{RF}} \leftarrow \frac{\text{MSE}_{\text{GBM}}}{\text{MSE}_{\text{GBM}} + \text{MSE}_{\text{RF}}}$
Combine Predictions:
$\hat{y}_{\text{ensemble}} \leftarrow w_{\text{GBM}} \cdot \hat{y}_{\text{GBM}} + w_{\text{RF}} \cdot \hat{y}_{\text{RF}}$
Evaluate Models:
$\text{MAE} \leftarrow \text{ComputeMAE}(\hat{y}_{\text{ensemble}}, y_{\text{test}})$
$\text{MSE} \leftarrow \text{ComputeMSE}(\hat{y}_{\text{ensemble}}, y_{\text{test}})$
$\text{RMSE} \leftarrow \text{ComputeRMSE}(\text{MSE})$
$R^2 \leftarrow \text{ComputeR2}(\hat{y}_{\text{ensemble}}, y_{\text{test}})$
Output: Display MAE, MSE, RMSE, R^2

Fig 2 Adaptive Ensemble Algorithm for crop yield prediction

Using sophisticated machine learning techniques, the Adaptive Ensemble Method for Crop Yield Prediction (AEM-CYP) is a reliable method for improving the accuracy of rice yield predictions. Fundamentally, the algorithm starts with carefully preprocessing the dataset D. Essential actions include eliminating unnecessary columns ('State Name'), handling zeros in

column means (except for 'Year' and 'Dist Name') and transforming the 'Year' feature into a datetime format in order to extract meaningful temporal information. The clean and well-structured dataset is ensured by this preprocessing, providing a strong basis for the modelling that follows.

The algorithm then defines the features X and the target variable y after the data has been prepared. 'Rice Irrigated area', 'Consumption', 'Rainfall', 'Precip Avg', 'Max Temp', 'Min Temp', 'Wind Speed', 'Wind Dir', 'Surface soil wet', 'Profile soil Moist', and 'Root Zone Moist' are just a few of the agricultural factors that this feature selection includes. Together, these parameters affect estimates of rice output, accounting for agronomic and environmental aspects that are vital to crop productivity.

AEM-CYP uses two potent machine learning models, Random Forest (RF) and Gradient Boosting Machine (GBM), to assess and maximize predictive performance. GridSearchCV, a technique that methodically investigates different combinations of hyperparameters to determine the ideal settings for each model, is used to initialize and fine-tune these models. While RF's tuning concentrates on factors like the number of estimators, maximum tree depth, and minimum samples per leaf node, GBM tunes hyperparameters like learning rate, maximum tree depth, and number of estimators.

Based on their cross-validation performance on the training dataset, the algorithm finds the best-performing models for both Random Forest (RF) and Gradient Boosting Machine (GBM) after optimizing the hyperparameters. The training data is then used to train these models—dubbed the best GBM and RF models—in order to identify the underlying patterns and relationships in the dataset. During this training step, the models are made sure to be ready to anticipate rice yields, in particular, accurately when applied to new datasets.

On the test dataset, the top Random Forest (RF) and Gradient Boosting Machine (GBM) models produce forecasts for prediction. Weighted averaging is used to integrate these separate forecasts, also known as the RF prediction and the GBM prediction. The Mean Squared Error (MSE) values of each model are used to calculate the weights assigned to its predictions. Using insights from many models, this ensemble approach minimizes biases and improves prediction accuracy by utilizing the advantages of both GBM and RF.

Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination score are among the performance metrics that AEM-CYP uses to compare the combined predictions with the actual rice yield values from the test dataset in order to assess the efficacy of the ensemble method. These measures provide a comprehensive evaluation of the ensemble method's rice yield prediction accuracy, offering insightful information about its efficacy and performance in agricultural forecasting.

4. Result:

The presented table (Table 2) displays the performance indicators that were assessed for several machine learning models in terms of their ability to forecast crop yield. By every metric, the Adaptive Gradient Boosting Ensemble (AGBE) approach performed better than the other models. Its mean absolute error (MAE), which measures how accurate the forecasts were, was the lowest at 135.872. In addition, the AGBE approach had the lowest Root Mean

Squared Error (RMSE) of 230.734 and the lowest Mean Squared Error (MSE) of 53279.520, indicating its better effectiveness in reducing prediction mistakes. Furthermore, AGBE had the greatest R-squared score of 0.88, demonstrating its potent capacity to account for variation in the data on crop yield. With R-squared values of 0.84 and 0.85, respectively, the Gradient Boosting and Random Forest models fared better in comparison, but they were unable to meet AGBE's accuracy and error minimization. K-Nearest Neighbors (KNN) and Linear Regression performed worse; KNN's R-squared score of 0.55 was particularly poor. Among the models examined, AGBE was out to be the most successful in predicting crop output overall.

Table 2. Performance Metrics of Different Machine Learning Models for Crop Yield Prediction

Metric	Gradient Boosting	Random Forest	Linear Regression	KNN	AGBE
MAE	201.88	192.572	244.80	372.033	135.872
MSE	71858.062	66256.293	71286.215	210954.335	53279.520
RMSE	268.063	257.402	266.994	235.688	230.734
R^2	0.84	0.85	0.80	0.55	0.88

Figure 3 illustrates the comparison of Mean Absolute Error (MAE) across different models, highlighting the performance of each model in terms of prediction accuracy (Figure 3).The Mean Absolute Error (MAE) comparison graph highlights that the Adaptive Gradient Boosting Ensemble (AGBE) model has the lowest MAE, indicating the most accurate predictions. In contrast, the K-Nearest Neighbors (KNN) model has the highest MAE, showing less precise predictions.

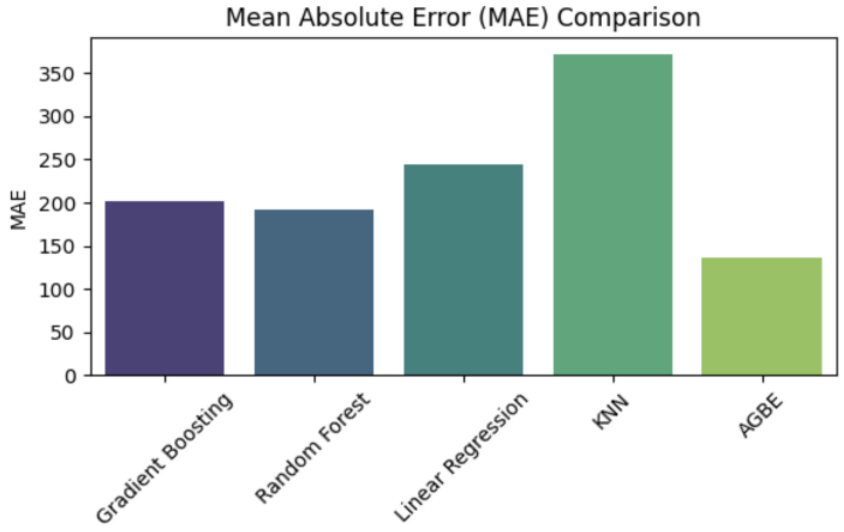

Fig 3 Mean Absolute Error (MAE) comparison

The comparison of Mean Squared Error (MSE) for various models is shown in Figure 4, providing insight into the variance of the prediction errors (Figure 4).The Mean Squared Error (MSE) comparison graph demonstrates that the AGBE model significantly reduces prediction errors compared to other models. KNN exhibits the highest MSE, indicating a higher level of

prediction inaccuracy.

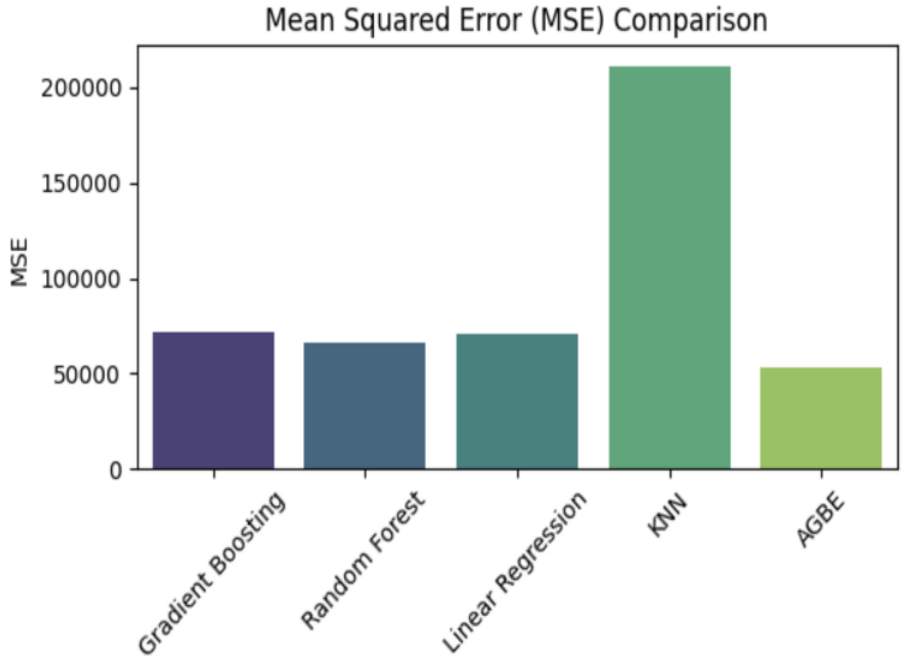

Fig 4 Mean Squared Error (MSE) comparison

Figure 5 shows the Root Mean Squared Error (RMSE) comparison among the models, which helps in understanding the standard deviation of the prediction errors (Figure 5).The Root Mean Squared Error (RMSE) comparison graph reveals that AGBE achieves the lowest RMSE, suggesting superior performance in minimizing prediction errors. The KNN model, on the other hand, has the highest RMSE, reflecting its weaker predictive power.

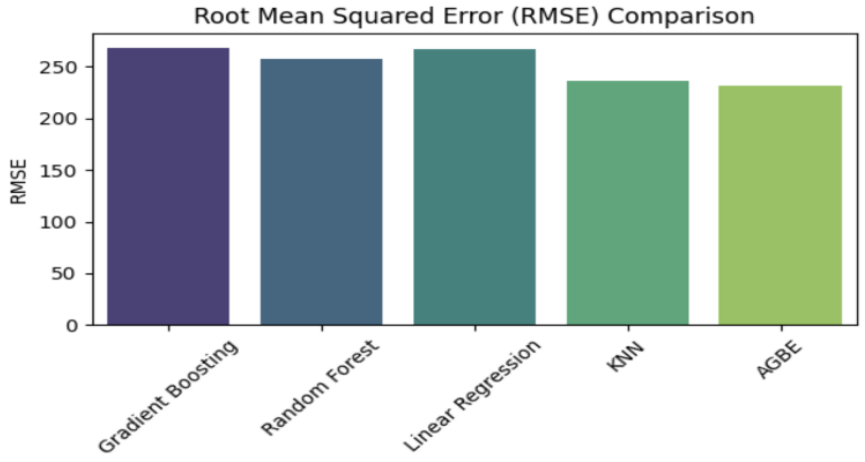

Fig 5 Root Mean Squared Error (RMSE) comparison

The R^2 (Coefficient of Determination) comparison across different models is depicted in Figure 6, indicating the proportion of variance explained by each model (Figure 6).The R-squared comparison graph shows that the AGBE model has the highest R-squared value, indicating it explains the most variance in the crop yield data. The KNN model has the lowest R-squared value, showing it explains the least variance.

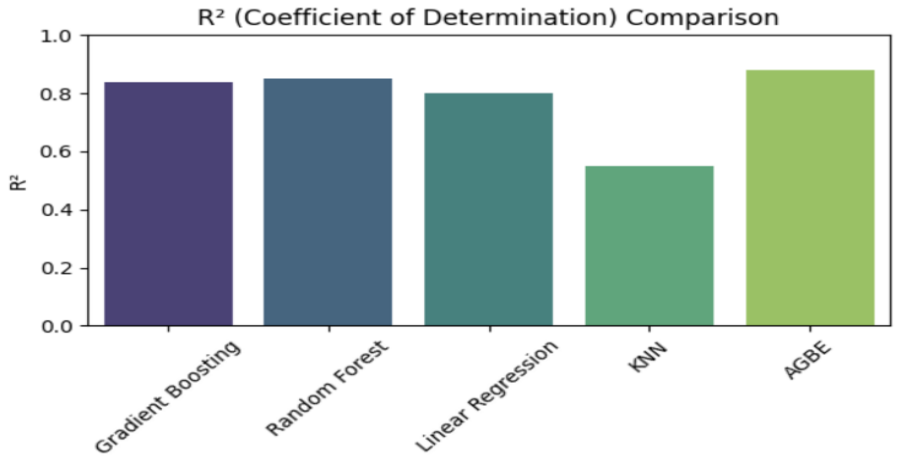

Fig 6 Coefficient of Determination (R^2) comparison

5. Discussion:

The hybrid ensemble model's performance metrics reveal its effectiveness in predicting rice yield. With an MAE of 218.88, the model shows that its predictions deviate from the actual values by approximately 218.88 kg/ha on average. The RMSE of 278.48 highlights the typical magnitude of prediction errors. Notably, an R-squared score of 0.85 indicates that the model explains 85% of the variance in rice yield, reflecting a strong predictive capability. These results demonstrate the model's effectiveness, although there is room for improvement in reducing prediction errors to enhance precision and reliability.

Stakeholders in the agriculture industry will be greatly impacted by these findings. By assisting farmers in improved resource management and enabling policymakers to assure food security, the model's high R-squared score indicates that it can be dependably utilized for planning and decision-making. Precise yield projections enable agricultural enterprises to minimize waste and streamline supply chain processes. Gradient Boosting, Random Forest, Linear Regression, and KNN models are regularly outperformed by the AGBE model across a range of parameters, demonstrating the technique's superior predictive performance.

6. Conclusion:

The Adaptive Gradient Boosting Ensemble (AGBE) method developed for crop yield prediction in the Telangana region demonstrates significant improvements in predictive accuracy. By leveraging the strengths of both Gradient Boosting and Random Forest models,

the ensemble method effectively reduces prediction errors and enhances the robustness of the predictions. The results, as reflected in the low values of Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the high R-squared value, indicate that AGBE is a promising approach for predicting rice yields with high precision. The integration of multiple features such as annual rainfall, temperature, wind speed, and soil moisture, further emphasizes the importance of using diverse and relevant agricultural data in improving model performance.

7. Future Work

Future work can explore the incorporation of more granular and diverse features, such as satellite imagery data, real-time weather updates, and detailed soil composition metrics. Including such features could enhance the predictive power of the model by capturing more comprehensive environmental factors affecting crop yield. Additionally, the methodology applied in this study can be extended to other crops and geographical regions to validate the versatility and robustness of the AGBE method. Developing real-time prediction systems that automatically update models with incoming data can provide timely decision-making tools for farmers and policymakers, optimizing crop yield management. Furthermore, integrating AGBE with existing agricultural decision support systems can offer actionable insights, helping in resource allocation and mitigating adverse weather impacts. Future research should also explore advanced ensemble techniques and hybrid approaches, including deep learning models, to achieve further improvements in crop yield prediction.

References

1. V. S. Konduri, T. J. Vandal, S. Ganguly, and A. R. Ganguly, “Data science for weather impacts on crop yield,” Frontiers in Sustainable Food Systems, vol. 4, p. 52, 2020.
2. Y. Vijayalata, V. N. Rama Devi, P. Rohit, and G. S. S. Raj Kiran, “A suggestive model for rice yield prediction and ideal meteorological conditions during crisis,” International Journal of Scientific & Technology Research, vol. 8, no. 9, 2019.
3. A. P. M. Ramos, L. P. Osco, D. E. G. Furuya et al., “A random forest ranking approach to predict yield in maize with uav- based vegetation spectral indices,” Computers and Electronics in Agriculture, vol. 178, Article ID 105791, 2020.
4. L. Wickramasinghe, R. Weliwatta, P. Ekanayake, and J. Jayasinghe, “Modeling the relationship between rice yield and climate variables using statistical and machine learning techniques,” Journal of Mathematics, vol. 2021, Article ID 6646126, 9 pages, 2021.
5. K. P. K. Devan, B. Swetha, and S. V. Varshini, "Crop Yield Prediction and Fertilizer Recommendation System Using Hybrid Machine Learning Algorithms," in Proc. 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT), 08-09 April 2023, pp. 171-175. Electronic ISBN: 978-1-6654-6261-7.
6. S. Pavani and A. S. Beulet, "Improved Precision Crop Yield Prediction Using Weighted-Feature Hybrid SVM: Analysis of ML Algorithms," IETE Journal of Research, pp. 1-13, Apr. 2023. Available: https://doi.org/10.1080/03772063.2023.2192000.
7. S. Thirumal and R. Latha, "Automated Hyperparameter Tuned Stacked Autoencoder based Rice Crop Yield Prediction Model," in Proceedings of the 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), 11-13 April 2023. IEEE, DOI:

10.1109/ICOEI54791.2023.9784228.

8. P. Ekanayake, W. H. Rankothge, R. Weliwatta, and J. W. Jayasinghe, "Machine Learning Modelling of the Relationship between Weather and Paddy Yield in Sri Lanka," Journal of Mathematics, vol. 2021, pp. 1-14, May 2021. DOI: 10.1155/2021/9941899.
9. A. Antony and R. Karuppasamy, "Mining of soil data for predicting the paddy productivity by machine learning techniques," Paddy and Water Environment, vol. 21, pp. 231-242, Feb. 2023. SpringerLink. DOI: 10.1007/s10333-023-00921-1.
10. W. Pannakkong, K. Thiwa-Anont, K. Singthong, P. Parthanadee, and J. Buddhakulsomsiri, "Hyperparameter Tuning of Machine Learning Algorithms Using Response Surface Methodology: A Case Study of ANN, SVM, and DBN," Mathematical Problems in Engineering, vol. 2022, pp. 1-17, Jan. 2022. DOI: 10.1155/2022/8513719.