IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 4, August 2026, pp. 3411~3420
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3411-3420

SHAP-enhanced ensemble learning for yield prediction: insights from CY-Bench data on Indian wheat

Soma Gupta[1], Dayal Kumar Behera[2], Satarupa Mohanty[2], Subhra Swetanisha[3], Ritik Mallik[2], Namita Panda[2]
[1]Department of Computer Science, Ravenshaw University, Cuttack, India
[2]School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, India
[3]Department of Computer Science and Engineering, Silicon University, Bhubaneswar, India

Article Info

Article history:

Received Apr 25, 2025
Revised May 19, 2026
Accepted Jun 19, 2026

Keywords:

Boosting
Ensemble models
Feature selection
Normalized difference vegetation index
Remote sensing
Shapley additive explanations
Yield prediction

ABSTRACT

Predicting crop yield accurately is essential for providing food security and improving agricultural practices. This study examines the use of ensemble machine learning models combined with Shapley additive explanations (SHAP) feature selection to improve wheat yield prediction in India. The study uses CY-Bench data, incorporating normalized difference vegetation index (NDVI), meteorological data, and soil moisture data for yield prediction. Various ensemble techniques, including voting, stacking, and boosting are evaluated. Boost m1 ensemble model consistently outperforms other models in the prediction. Additionally, the integration of SHAP-based feature selection with the best ensemble model significantly improves the model accuracy and interpretability by identifying the most influential features affecting yield. The results show the effectiveness of ensemble boosting model, in capturing the complex relationships within agricultural data particularly when combined with feature selection. This method improves the transparency and actionability of machine learning models for agronomists, policymakers, and farmers.

Corresponding Author:

Soma Gupta
Department of Computer Science, Ravenshaw University
Cuttack, India
Email: somagupta29@gmail.com

1. INTRODUCTION

Wheat is an important agricultural crop of India and is extremely important for its economy, culture, and food security. Despite its importance wheat production in India faces several challenges, including climate change, water scarcity, soil degradation, pests, and diseases. Several steps are being taken to address these challenges and ensure the future of wheat in India, including developing climate-resilient varieties, promoting sustainable farming practices, and investing in research and technology that are essential for increasing yields and address new problems. Addressing these problems and investing in sustainable practices, it will be possible to ensure the importance of wheat to support India's growing population and its agricultural economy. Previously historical data and expert knowledge were used by farmers for predicting crop yield. Crop yield prediction aims to contribute to a more sustainable, resilient, and food-secure future by informed decision-making supporting individual farmers to policymakers.

Yield prediction still faces challenges despite tremendous advancements. For large-scale agricultural operations, field surveys and manual data collection techniques are labor-intensive, time-consuming, and costly [1]. The studies show a shift from traditional statistical models to machine learning and deep learning approaches for yield prediction. According to Sharma *et al.* [2], autoregressive integrated moving average

(ARIMA) with differencing for stationarity and quantile regression forests is applied for uncertainty estimations. Numerous studies highlight the importance of meteorological factors [3]–[5], soil moisture [6], multispectral remote sensing data, vegetation indices [7], [8], and hyperspectral bands [9] to improve yield prediction accuracy. Agricultural ERA5 (AgERA5) [10] support agricultural analysis by providing ready to use meteorological data. Deep learning models dominates the prediction of wheat, maize, corn, soybean, tea, and sugarcane. Sun *et al.* [11] uses a multilevel deep learning deep learning network with time series satellite data for corn yield prediction while other studies employ convolutional neural network (CNN)-transformer [12], hybrid CNN-recurrent neural network (RNN) models [13], Bayesian optimized deep neural network (DNN) [14], and convolutional long short-term memory – vision transformer (Conv LSTM-ViT) [15] to capture spatial and temporal patterns. Sharma *et al.* [16] compares regression and deep learning models using multiple evaluation metrics, and Gupta *et al.* [17] confirms deep learning models outperform traditional approaches in learning complex nonlinear relationships.

Numerous studies discussed ensemble models [18], [19], principal component analysis (PCA) and Shapley additive explanations (SHAP) [20], Bayesian hyperparameter optimization [21] to improve robustness and interpretability. These studies show ensemble models achieve better accuracy than single models. The authors [22], [23] applies multiple machine learning models for yield prediction and identifies the key factors affecting it. However, while deep learning dominates maize and wheat prediction, interpretable ensemble models remain less explored for wheat yield forecasting, motivating the need for hybrid and explainable frameworks.

To address these, this study proposes the development of a reliable, interpretable yield prediction framework using normalized difference vegetation index (NDVI), meteorological, and soil moisture data by combining ensemble learning methods with SHAP-based feature selection. The use of SHAP feature selection and model interpretability adds an important degree of novelty. It improves model transparency by identifying most influential environmental factors affecting yield and helps to bridge the gap between machine learning and domain-specific decision-making. This study distinguishes itself by improving explainability along with accuracy, making the results more actionable and reliable for agronomists, policymakers, and farmers.

The structure of this paper is organized as follows. Section 1 introduces the importance of accurate crop yield prediction and the limitations of existing methods. It highlights the use of ensemble machine learning methods and SHAP feature selection to address these limitations. Methodology used in the paper is explained in section 2. Section 3 discusses the performance metrics of various ensemble methods along with SHAP feature selection for the best ensemble model. Section 4 concludes the paper.

2. METHOD

2.1. Dataset description

In this study, the wheat yield dataset is obtained for India using CY-Bench data [24] that covers wheat and maize crops for 29 and 42 countries. A benchmark dataset for subnational crop yield predictions, CY-Bench described in Table 1 that facilitates model intercomparison across diverse agricultural system worldwide. CY-Bench dataset is organized as a collection of CSV files, where each file represents a specific category of variable for a specific country. The CSV file is named based on the category and country, helping in easy identification and retrieval. CY-Bench offers a robust, multi-country dataset for yield prediction but faces limitations regarding spatial-temporal resolution mismatch and data correlations. It focuses on subnational but not field-level data that requires preprocessing to align the varied predictor resolutions that may introduce uncertainty.

Table 1. CY-Bench summary table

Category	Descriptions
Primary crops	Maize, wheat
Spatial coverage	Maize (38 countries), wheat (29 countries)
Spatial resolution	Subnational (highest available resolution at administrative level)
Temporal coverage	Varies
No. of records	57,000 rows × 12 columns

Table 2 describes the data used in this study, which focuses on the Indian wheat crop. NDVI data obtained from MOD09CMG [25] has a spatial resolution of 500 meters × 500 meters with a weekly temporal resolution. Meteorological data collected from AgERA5 [26], provides a spatial resolution of 0.1°×0.1° with daily temporal resolution. Soil moisture obtained from GLDAS [27] has a spatial resolution 1° and 1/4° with daily temporal resolution.

Table 2. Detailed data description

Data	Description	Variables (units)	Temporal resolution
ndvi	Normalized difference vegetation index	-	weekly
meteo	Temperature, precipitation, radiation, potential evapotranspiration, climatic water balance	tmin (˚C), tmax (˚C), tavg (˚C), prec (mm), et0 (mm), cwb (mm), rad (J m-2 day-1)	daily
soil moisture	Surface and rootzone soil moisture	ssm (kg m^{-2}), rsm (kg m^{-2})	daily
yield	End-of-season yield	yield (t ha^{-1})	yearly

2.2. Data preprocessing and yield prediction framework

The yield prediction study uses a standalone computer system with Jupyter notebook, Python 3.12 using a set of Python libraries for data preprocessing, modeling, evaluation, and interpretability described in Table 3. A crucial step in the machine learning pipeline is preprocessing the data because raw data collected from real-world sources is often messy and not immediately suitable for building effective models. Before training the proposed models, a series of feature engineering and preprocessing procedures were applied to the acquired data set. In this study soil moisture and metrological datasets were grouped on admin id with sorted date and amalgamated these datasets having daily temporal resolution. Various aggregate functions such as max(), min(), and mean() were applied on the columns of the formed dataset to calculate the weekly values. Mean aggregation captured the temporal trends, reduced noise in daily observations, provided more stable performance, lower prediction error compared to max()/min() and is thus chosen over max()/min(). Weekly formed dataset having soil and metrological features are then merged with NDVI dataset having weekly temporal resolution.

NDVI is temporally aggregated using summary statistics and used as an input for yield prediction as this approach reduces dimensionality and noise. Finally, the resulting dataset was merged with yield dataset. On amalgamating columns such as country code, season name, planting, year, planted area were identified as irrelevant to the research and is eliminated as the study domain is India. The identification and handling of null values is a crucial step in the preparation and analysis of data. Following a thorough analysis of null values, the dataset was determined to be null value-free, enabling us to move on with this null-value-free data. An essential preprocessing step in data analysis and machine learning is identifying and eliminating duplicate records. In this study no duplicate records were found. Normalization preprocessing approach guarantees that every feature makes an equal contribution to the model's learning process. The resultant dataset comprises twelve columns, with 'Yield' as the target and the remaining eleven as features. Figure 1 shows the pairwise correlations among the agroclimatic variables, soil variables, vegetation indices, and yield showing the key relationship relevant to yield prediction.

Table 3. Software environment

System specification	Descriptions
Operating system	Windows 11
Processor	Intel Core i5
IDE	PyCharm professional 2024.3.4
Python version	3.12
Data processing	pandas, NumPy, sklearn
Visualization	matplotlib, seaborn

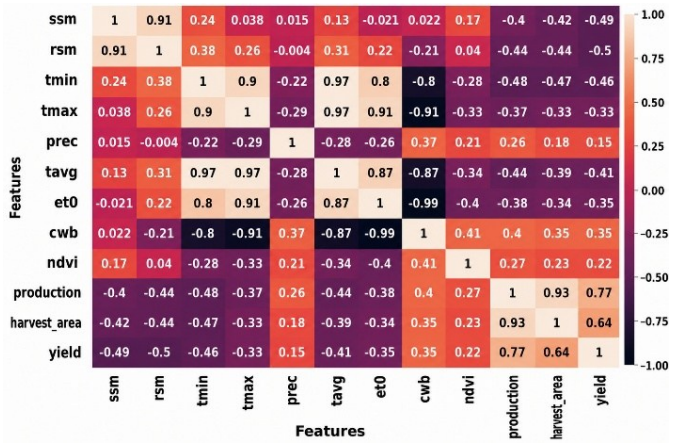


Figure 1. Feature correlation heatmap

Figure 2 illustrates the comprehensive workflow adopted for yield prediction using the CY-Bench dataset. The process begins with data acquisition from the CY-Bench dataset, which comprises four key components meteorological data, soil moisture data, NDVI, and crop yield data. After preprocessing of several partitioning ratios considered, partitioning ratio of 60%, 20%, and 20 % for training, validation, and testing is taken, as it resulted in low variance on 10 repetitions. Partitioning ensures the models are trained, validated, and tested on separate subsets. Here, 5-fold cross validation technique is employed, where the data is partitioned into five disjoint folds. In each iteration, one-fold is used as test and the remaining four are used for training. The preprocessing steps are carried out within each training fold only, thereby ensuring the test fold remains completely unseen during training. This approach strictly ensures that there is no data leakage and the integrity of evaluation process is maintained.

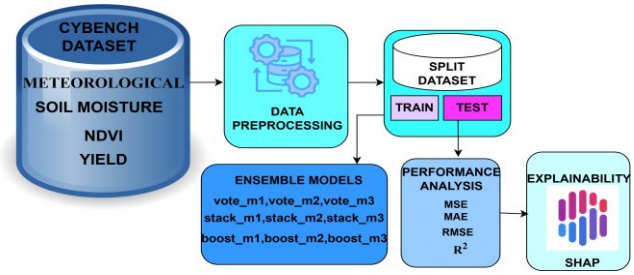


Figure 2. Comprehensive framework for yield prediction

Several baseline models, such as multiple linear regression (LR), decision tree (DT), random forest (RF), k-nearest neighbors (KNN), support vector regression (SVR), and extreme gradient boosting (XGBoost), are used to develop ensemble models. Various ensemble machine learning techniques such as voting, stacking and, boosting ensemble are used. The goal of using an ensemble is to improve performance (like accuracy or robustness) compared to any single model. The performance of the model is evaluated using the performance metrics, namely mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). Transparency and interpretability of the model predictions is enhanced using SHAP in the explainability phase [28]. The dataset is divided into training and test data to avoid data leakage. The models are trained using training data, and SHAP values are computed using only trained models and training samples. Feature importance and selection is done using the training SHAP values. The selected features are then applied to the test data for model evaluation ensuring no information from the test data influences feature selection or model training. This step provides insights into the contribution of individual features to model outputs, supporting informed decision-making.

The detailed composition of base learners, ensemble strategies, and the parameter setting used in the study is described as follows. Voting models use soft voting with combinations of LR, DT, RF, KNN, SVR, and XGBoost. Key parameters for RF and XGBoost include 100 estimators, DT depth of 4, KNN (k =5), and SVR with radial basis function (RBF) kernel (C =100, γ =0.1). Stacking models uses similar base learners with meta-model of 100 estimators, learning rate of 0.1, and maximum depth of 3. Boosting models applies sequential boosting with selected learners configured with 100 estimators, learning rate of 0.05, and maximum depth of 3.

Algorithm 1 to 3 describes the ensemble regression techniques [29] that improve the performance by combining the outputs of multiple base learners. Voting generates final prediction by averaging the outputs of several independent regressors thus reducing variance and improving robustness. Stacking employs a meta regressor to learn the optimal combination of predictions from multiple base models. Boosting sequentially constructs a strong predictive model by sequentially training weak learners where each successive learner focuses on correcting the errors made by its predecessors.

Algorithm 1. Voting regressor algorithm
Input: given M regression models M_1, M_2, …, M_k; dataset D. $y^{(j)} = M_i(x_i)$ is model M_i prediction of on input x_i.
Output: average of predictions across all k models.
Step 1: for each $j = 1$ to k, train model M_j on D.
Step 2: predict using each model. For new input x, compute predictions $y^{(j)} = M_j(x)$ for $j = 1,2, \ldots, k$.
Step 3: compute the final prediction using average (simple voting): $y^{(voting)} = \frac{1}{k}\sum_{j=1}^{k} y^{(j)}$

Algorithm 2. Stacking regressor algorithm
Input: $d = \{(x_i, y_i)\}_{i=1}^{n}$ be the dataset. M_1, M_2, …, M_k: base regressors (level 0). M_{meta}: meta regressor (level 1). $y_i^{(j)}$: model M_j prediction for input x_i.
Output: prediction generated by the meta-regressor using combined predictions of all base regressors.
Step 1: train the base models. Each base model M_k is trained on the training dataset: $y^{(k)} = M_k(x_i)$ for all $k = 1, \dots, K$ and $i = 1, \dots, n$.
Step 2: for each input x_i, collect prediction from all base models: $z_i = \left[y_i^{(1)}, y_i^{(2)}, \dots, y_i^{(k)}\right]$ which forms a new dataset $Z = \{(z_i, y_i)\}_{i=1}^{n}$
Step 3: train the meta regressor on dataset Z.
Step 4: make final predictions:

i) Get base model predictions for x_{new}: $y_{new}^{(k)} = M_k(x_{new}), k = 1, \dots, k.$
ii) Create meta feature vector $z_{new} = [y_{new}^{(1)}, \dots, y_{new}^{(k)}]$.
iii) Predict with meta model: $y_{final} = M_{meta}(Z_{new})$.

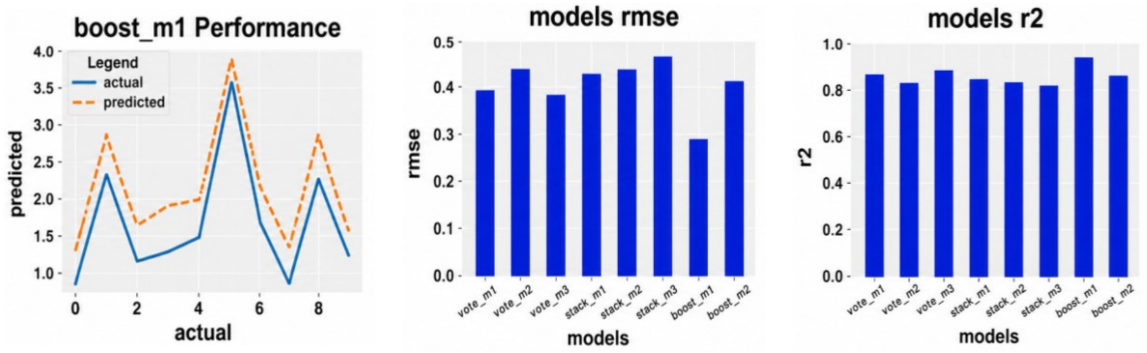
Algorithm 3. Boosting regressor algorithm
Input: training data $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. Weak learner $h(x, \theta)$. Number of boosting rounds M. Learning rate v.
Output: a strong predictive model $F_m(x)$ formed by sequentially combining all the weak learners with optimized weights.
Step 1: initialize the model with constant predictor: $F_0(x) = arg\min_c \sum_{i=1}^{n} L(y_i, c)$.
Step 2: for m = 1 to M do:

i) Compute residuals: $r_i = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}$.
ii) Fit a weak learner $h_m(x)$ to predict residuals: $h_m(x) \sim r_i$
iii) Compute step size: $\gamma_m = \min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$.
iv) Update the model: $F_m(x) = F_{m-1}(x) + v\,\gamma_m h_m(x)$.

Step 3: output final boosted model $F_M(x)$.

3. RESULTS AND DISCUSSION

Figure 3 illustrate the performance of the ensemble models. Figure 3(a) presents the actual and predicted yield values obtained using the boost m1 model, while Figures 3(b) and 3(c) compare the RMSE and R^2 values of all ensemble models, respectively. Boost m1 achieved the lowest RMSE (0.289) and MAE (0.21), with a high R^2 score (0.92), indicating strong accuracy and generalization ability of the model. Among all the voting models, vote m3 has the best (RMSE =0.37, R^2 =0.862). stack m1 shows the best (RMSE =0.41, R^2 =0.83), while stack m3 exhibited the worst (RMSE =0.44, R^2 =0.81) of all the stacking models. The results show voting and stacking ensembles provide competitive performance, boosting will provide more reliable solution in complex agricultural prediction scenarios. Table 4 provides RMSE, MAE, R^2, and MSE performance metrics for various ensemble model applied for yield prediction. 5-fold cross validation is performed and mean ± standard deviation for different evaluation metrices is highlighted in Table 5. These tests were applied on fold-wise results so that each comparison uses paired observations (same folds).



(a) (b) (c)

Figure 3. Performance metrices plots of (a) actual v/s predicted yield, (b) RMSE, and (c) R^2

Table 4. Performance of ensemble methods

Models	RMSE	MAE	R^2	MSE
vote m1	0.39	0.29	0.85	0.15
vote m2	0.42	0.31	0.82	0.18
vote m3	0.37	0.29	0.86	0.14
stack m1	0.41	0.33	0.83	0.17
stack m2	0.42	0.32	0.82	0.17
stack m3	0.44	0.35	0.81	0.19
boost m1	0.28	0.21	0.92	0.08
boost m2	0.4	0.3	0.84	0.16

Table 5. 5-fold CV results (mean ± standard deviation)

Models	RMSE	MAE	R^2
vote m1	0.39±0.03	0.29±0.02	0.85±0.02
vote m2	0.42±0.02	0.31±0.03	0.82±0.02
vote m3	0.37±0.02	0.29±0.01	0.86±0.02
stack m1	0.41±0.03	0.33±0.02	0.83±0.02
stack m2	0.42±0.02	0.32±0.02	0.82±0.02
stack m3	0.44±0.03	0.35±0.03	0.81±0.02
boost m1	0.28±0.01	0.21±0.01	0.92±0.01
boost m2	0.40±0.03	0.30±0.01	0.84±0.02

To observe and determine the impacts of explainable feature selection, the best ensemble model (boost m1) is enhanced by adding a feature F1-score, obtained through SHAP-based feature selection. The resulting model, boost m1 + F1, shows notable improvements across all the evaluation metrics, as summarized in Table 6. The performance evaluation of the proposed models shows the impact of feature selection on predictive accuracy. The baseline boost m1 model exhibited strong performance, achieving a RMSE of 0.28, MAE of 0.21, and R^2 of 0.92. These results indicate that the ensemble boosting approach is effective in capturing the nonlinear relationships within the input features comprising NDVI, meteorological, and soil moisture data. Following the application of F1-based feature selection, the enhanced model (boost m1 + F1) showed a marked improvement across all evaluation metrics. RMSE and MAE were significantly reduced to 0.17 and 0.15, respectively, while the MSE decreased from 0.08 to 0.03. The R^2 value also increased to 0.96, indicating a stronger fit and improved generalization capability. Figure 4 illustrates how SHAP tree explainer is used to compute SHAP values from model predictions enabling global and local feature attribution and summarizing feature importance.

Table 6. Performance metrics for best ensemble model with and without feature selection

Metric	boost m1	boost m1 + F1 (F1=selected features using SHAP: production, harvest area, tmin, NDVI, tavg)
RMSE	0.28	0.17
MAE	0.21	0.15
MSE	0.08	0.03
R2	0.92	0.96

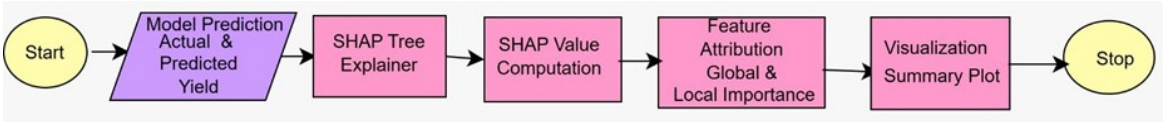
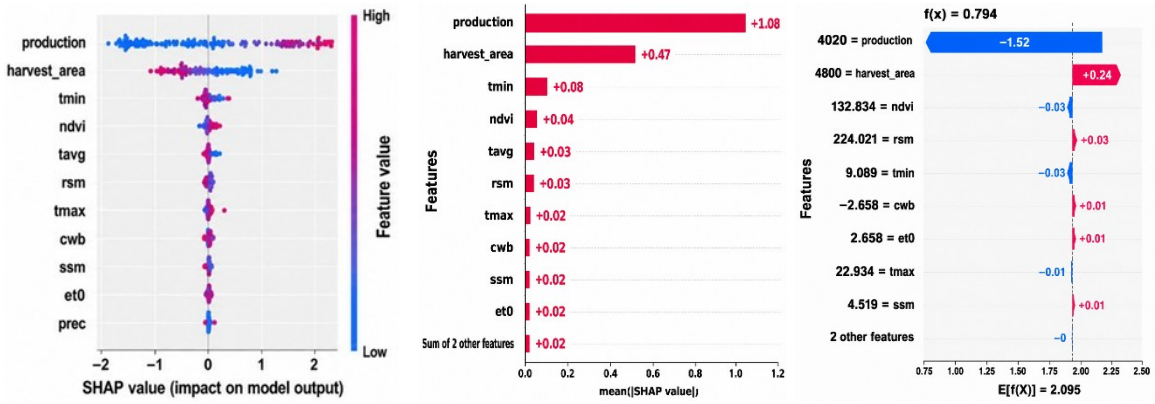


Figure 4. SHAP-based model interpretability workflow

Figure 5 presents the SHAP-based explainability results for the proposed model. Summary plot shown in Figure 5(a) interprets the impact of features on model output. Bar plot shown in Figure 5(b) describes the average impact (mean absolute SHAP value) of each feature on model predictions with production the most impactful with a SHAP value of +1.08. Waterfall plot in Figure 5(c) shows how individual features contributes to a single prediction. These results indicates that F1-based feature selection effectively filtered out less relevant features, allowing the model to focus on the most informative features.



(a) (b) (c)

Figure 5. Various SHAP plots of (a) summary plot, (b) bar plot, and (c) waterfall plot

This resulted in better predictive accuracy and model efficiency. The results show the integration of feature selection approaches, such as SHAP-based analysis, with ensemble learning enhances the performance and interpretability of yield prediction models. The null hypothesis (H0) is set as: there is no significant difference between boost m1 and other ensemble methods. Paired t-tests and Wilcoxon signed-rank tests confirmed that the reduction in RMSE and the increase in R^2 after SHAP-based feature selection were significant at the 95% confidence level ($p < 0.05$). A comparison with previous studies is also highlighted in Table 7.

Table 7. Quantitative comparison with previous studies

Study	Crop / Region	Dataset type	Models used	Explainability	Best performance
Seireg *et al.* [23]	Wild Blueberry / USA	Simulation + weather	Light gradient boosting machine (LGBM), gradient boost regression (GBR), XGBoost + stacking regression	SHAP (interpretation only)	R^2 =0.985, RMSE =179.9
Waqar *et al.* [21]	Multi-crop / Global	Real + synthetic (Prophet)	RF, DT, XGB, KNN + optimized stacking ensemble model (OSEM)	Not reported	R^2 =0.996, MAE =0.185 t/ha
Proposed study	Wheat / India	CY-Bench (real, sub-national)	Voting, stacking, boosting	SHAP-based feature selection	R^2 =0.92, RMSE =0.28

4. CONCLUSION

Ensemble models combined with SHAP feature selection could be a powerful approach for increasing accuracy and interpretability in yield prediction models. These methods combine the strengths of different learning algorithms to improve prediction accuracy and generalization. SHAP values identifies the most influential features for yield prediction and create more efficient and accurate models. Boost m1 performs better using F1-based feature selection with improved evaluation metrics, thus confirming feature selection filters out less relevant features effectively. Combining SHAP-based analysis with ensemble learning agricultural yield prediction models can achieve better performance and interpretability. Feature selection improved the performance, significantly reducing RMSE, MAE, MSE to 0.17, 0.15, 0.03 and increasing R^2 to 0.96. These findings indicate the effectiveness of ensemble boosting approach, in capturing the nonlinear relationships within the input features comprising NDVI, meteorological, and soil moisture data when combined with feature selection. Unlike prior studies that apply SHAP solely for post-hoc interpretability, this study is the first attempt to apply SHAP-driven feature selection on CY-Bench Indian wheat data in an ensemble boosting framework. NDVI, AgERA5 meteorology, and soil moisture at district level have also been integrated. This dual focus on predictive performance enhancement and explainability oriented feature selection differentiate the proposed framework from available literatures. Performance improvements are statistically validated through cross-validation and hypothesis testing. Although this study shows the effectiveness of ensemble models combined with SHAP-based feature selection for crop yield prediction, but there are several areas yet to be explored for future research. A temporal extension of SHAP

values could be explored to understand how feature importance evolves during the growing season, offering more actionable insights for dynamic decision-making.

FUNDING INFORMATION

This research received no grant from any funding agency.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Soma Gupta	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Dayal Kumar Behera		✓		✓	✓		✓	✓		✓	✓	✓		
Satarupa Mohanty		✓		✓	✓		✓	✓		✓	✓	✓		
Subhra Swetanisha				✓	✓	✓	✓	✓		✓	✓			
Ritik Mallik				✓	✓	✓	✓	✓		✓	✓			
Namita Panda				✓	✓	✓	✓	✓		✓			✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available with the AgML (https://www.agml.org/) at https://doi.org/10.5281/zenodo.13838912, reference [24]. The full dataset can be downloaded directly from Zenodo or using the "zenodo get" library.

REFERENCES

[1] X. Zhu, R. Guo, T. Liu, and K. Xu, "Crop yield prediction based on agrometeorological indexes and remote sensing data," *Remote Sensing*, vol. 13, no. 10, May 2021, doi: 10.3390/rs13102016.

[2] C. Sharma, P. Jha, P. Anand, and M. Sharma, "Prediction and time series analysis of wheat, rice and maize yields using ARIMA models," in *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, Nov. 2023, pp. 1055–1061, doi: 10.1109/ICTACS59847.2023.10389919.

[3] M. Ashfaq, I. Khan, A. Alzahrani, M. U. Tariq, H. Khan, and A. Ghani, "Accurate wheat yield prediction using machine learning and climate-NDVI data fusion," *IEEE Access*, vol. 12, pp. 40947–40961, 2024, doi: 10.1109/ACCESS.2024.3376735.

[4] M. J. Hoque *et al.*, "Incorporating meteorological data and pesticide information to forecast crop yields using machine learning," *IEEE Access*, vol. 12, pp. 47768–47786, 2024, doi: 10.1109/ACCESS.2024.3383309.

[5] M. J. Hoque, M. S. Islam, A. A. Noman, M. A. Hoque, I. A. Chowdhury, and M. Saifuddin, "Optimizing potato crop productivity: a meteorological analysis and machine learning approach," *IAES International Journal of Artificial Intelligence*, vol. 14, no. 2, Apr. 2025, doi: 10.11591/ijai.v14.i2.pp1116-1129.

[6] R. Mai, Q. Xin, J. Qiu, Q. Wang, and P. Zhu, "High spatial resolution soil moisture improves crop yield estimation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 19067–19077, 2024, doi: 10.1109/JSTARS.2024.3417424.

[7] U. Shafi *et al.*, "Tackling food insecurity using remote sensing and machine learning-based crop yield prediction," *IEEE Access*, vol. 11, pp. 108640–108657, 2023, doi: 10.1109/ACCESS.2023.3321020.

[8] N. Ali *et al.*, "Field-scale precision: predicting grain yield of diverse wheat breeding lines using high-throughput UAV multispectral imaging," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 11419–11433, 2024, doi: 10.1109/JSTARS.2024.3411994.

[9] E. Cheng *et al.*, "A GT-LSTM spatio-temporal approach for winter wheat yield prediction: from the field scale to county scale," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–18, 2024, doi: 10.1109/TGRS.2024.3418046.

[10] E. Choi, V. C. D. Souza, J. A. Dillon, E. Kebreab, and N. D. Mueller, "Comparative analysis of thermal indices for modeling cold and heat stress in US dairy systems," *Journal of Dairy Science*, vol. 107, no. 8, pp. 5817–5832, Aug. 2024, doi: 10.3168/jds.2023-24412.

[11] J. Sun, Z. Lai, L. Di, Z. Sun, J. Tao, and Y. Shen, "Multilevel deep learning network for county-level corn yield estimation in the U.S. corn belt," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5048–5060, 2020, doi: 10.1109/JSTARS.2020.3019046.

[12] J. Du, Y. Zhang, P. Wang, K. Tansey, J. Liu, and S. Zhang, "Enhancing winter wheat yield estimation with a CNN-transformer hybrid framework utilizing multiple remotely sensed parameters," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–13, 2025, doi: 10.1109/TGRS.2025.3544434.
[13] S. Khaki, L. Wang, and S. V. Archontoulis, "A CNN-RNN framework for crop yield prediction," *Frontiers in Plant Science*, vol. 10, Jan. 2020, doi: 10.3389/fpls.2019.01750.
[14] Z. Ramzan, H. M. S. Asif, I. Yousuf, and M. Shahbaz, "A multimodal data fusion and deep neural networks based technique for tea yield estimation in Pakistan using satellite imagery," *IEEE Access*, vol. 11, pp. 42578–42594, 2023, doi: 10.1109/ACCESS.2023.3271410.
[15] S. M. M. Nejad, D. A. -Moghadam, and A. Sharifi, "ConvLSTM-ViT: a deep neural network for crop yield prediction using earth observations and remotely sensed data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 17489–17502, 2024, doi: 10.1109/JSTARS.2024.3464411.
[16] P. Sharma, P. Dadheech, N. Aneja, and S. Aneja, "Predicting agriculture yields based on machine learning using regression and deep learning," *IEEE Access*, vol. 11, pp. 111255–111264, 2023, doi: 10.1109/ACCESS.2023.3321861.
[17] S. Gupta, S. Mohanty, and D. K. Behera, "AI-based yield prediction: a thorough review," *Indian Journal of Science And Technology*, vol. 18, no. 10, pp. 822–838, Mar. 2025, doi: 10.17485/IJST/v18i10.175.
[18] R. Shamim and T. Agarwal, "Optimizing crop yield prediction using machine learning algorithms," in *Smart Agritech: Robotics, AI, and Internet of Things (IoT) in Agriculture*, Wiley, 2024, pp. 443–487, doi: 10.1002/9781394302994.ch17.
[19] L. S. Cedric *et al.*, "Crops yield prediction based on machine learning models: case of West African countries," *Smart Agricultural Technology*, vol. 2, Dec. 2022, doi: 10.1016/j.atech.2022.100049.
[20] S. Kumar, M. Pant, and A. Nagar, "Forecasting the sugarcane yields based on meteorological data through ensemble learning," *IEEE Access*, vol. 12, pp. 176539–176553, 2024, doi: 10.1109/ACCESS.2024.3502547.
[21] M. Waqar, Y.-W. Kim, and Y.-C. Byun, "A stacking ensemble framework leveraging synthetic data for accurate and stable crop yield forecasting," *IEEE Access*, vol. 13, pp. 136909–136926, 2025, doi: 10.1109/ACCESS.2025.3591802.
[22] S. P. Raja, B. Sawicka, Z. Stamenkovic, and G. Mariammal, "Crop prediction based on characteristics of the agricultural environment using various feature selection techniques and classifiers," *IEEE Access*, vol. 10, pp. 23625–23641, 2022, doi: 10.1109/ACCESS.2022.3154350.
[23] H. R. Seireg, Y. M. K. Omar, F. E. A. El-Samie, A. S. El-Fishawy, and A. Elmahalawy, "Ensemble machine learning techniques using computer simulation data for wild blueberry yield prediction," *IEEE Access*, vol. 10, pp. 64671–64687, 2022, doi: 10.1109/ACCESS.2022.3181970.
[24] M. Kallenberg *et al.*, "CY-Bench: a comprehensive benchmark dataset for sub-national crop yield forecasting (1.10)," *AgML*, Jun. 2026, doi: 10.5281/zenodo.13838912.
[25] E. Vermote, "MODIS/terra surface reflectance 8-day L3 global 500m SIN grid V006," *NASA Land Processes Distributed Active Archive Center*, 2015, doi: 10.5067/MODIS/MOD09A1.006.
[26] H. Boogaard, J. Schubert, A. D. Wit, J. Lazebnik, R. Hutjes, and G. V. D. Grijn, "Agrometeorological indicators from 1979 to present derived from reanalysis," *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*, 2020, doi: 10.24381/cds.6c68c9bb.
[27] M. Rodell *et al.*, "The global land data assimilation system," *Bulletin of the American Meteorological Society*, vol. 85, no. 3, pp. 381–394, Mar. 2004, doi: 10.1175/BAMS-85-3-381.
[28] Y. Wang, P. Wang, K. Tansey, J. Liu, B. Delaney, and W. Quan, "An interpretable approach combining Shapley additive explanations and LightGBM based on data augmentation for improving wheat yield estimates," *Computers and Electronics in Agriculture*, vol. 229, Feb. 2025, doi: 10.1016/j.compag.2024.109758.
[29] D. Banik, R. Paul, R. S. Rathore, and R. H. Jhaveri, "Improved regression analysis with ensemble pipeline approach for applications across multiple domains," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 3, pp. 1–13, Mar. 2024, doi: 10.1145/3645110.

BIOGRAPHIES OF AUTHORS

Dr. Soma Gupta is presently working in the Department of Computer Science, Ravenshaw University, Cuttack, India. She has over 12 years of experience in teaching undergraduate and postgraduate students. She has one patent and authored several research publications in reputed journals and international conferences. Her research interests include machine learning, data science, design and analysis of algorithms, and database management system. She can be contacted at email: somagupta29@gmail.com.

Dr. Dayal Kumar Behera is working as an associate professor in the School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, India. He is an enthusiastic, proactive faculty member and researcher with more than 20 years of experience teaching courses at the undergraduate and postgraduate levels. Supervised many IEDC-funded projects, Ph.D. and M.Tech. thesis, and B.Tech. projects. He has two patents and published more than 40 research papers in various reputed, peer-reviewed refereed journals, and international conferences. He can be contacted at email: dayalbehera@gmail.com.

Dr. Satarupa Mohanty is currently working as an associate professor in School of Computer Engineering, KIIT deemed to be University, Bhubaneswar, India and having more than 20 years of teaching experience. She has authored more than 50 research papers, edited several books, and guided multiple Ph.D. students. Her research interests include machine learning, bioinformatics, and computer graphics. She can be contacted at email: satarupafcs@kiit.ac.in.

Dr. Subhra Swetanisha received the Ph.D. degree in Computer Science Engineering from KIIT deemed to be University, Bhubaneswar, India. She has been working as an associate professor in the Department of Computer Science and Engineering at Silicon University, Odisha. She has completed M.Tech. degree in Computer Science and Engineering from KIIT Deemed to be University. Her research interests include machine learning, data science, image processing, and remote sensing. She has nineteen years of teaching experience and published more than fifteen Scopus/SCI indexed research articles. She is a life member of the ISTE and IAENG. She can be contacted at email: sswetanisha@gmail.com.

Ritik Mallik is currently a research scholar in School of Computer Engineering, KIIT deemed to be University, Bhubaneswar, India. He has so far published 3 conference paper's with keen interest in machine learning, artificial intelligence, pattern recognition, and data minning. He can be contacted at email: ritikmallik@gmail.com.

Dr. Namita Panda is currently working as associate professor in School of Computer Engineering, KIIT deemed to be University, Bhubaneswar, India. There are more than 30 number of publications to her credit in many international and national level journals and conference proceedings. She has 20 years of experience in teaching undergraduate and postgraduate courses. Her research interests include object-oriented systems, software testing, security testing, data mining, operating systems, and computer architecture. She is a member of ISTE and a reviewer of many conferences and journals. She can be contacted at email: npandafcs@kiit.ac.in.