International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Optimizing crop yield forecasting with ensemble machine learning techniques

Bushara A. R*, Adnan Zaman K. T and Fathima Misriya P. S

Department of ECE, KMEA Engineering College, Aluva, India.

International Journal of Science and Research Archive, 2025, 14(01), 1456-1467

Publication history: Received on 09 December 2024; revised on 18 January 2025; accepted on 21 January 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.14.1.0189

Abstract

Accurate crop yield prediction is critical for ensuring food security and efficient agricultural management, particu- larly in the face of climate change and rising global populations. Current predictive models often fall short in generalizing across diverse agricultural contexts due to their inability to capture complex interactions between various climatic and soil variables effectively. This study addresses these gaps by proposing a com- prehensive machine-learning framework that integrates ensemble methods to enhance crop yield prediction accuracy. Using a dataset enriched with climatic and agricultural features, we evaluated multiple models, including Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, Bagging Regressor, and K-nearest neighbors. The Random Forest model emerged as the top performer, achieving an accuracy of 0.985 and a Mean Squared Error (MSE) of 1.08e+08. At the same time, the Bagging Regressor closely followed with an accuracy of 0.984 and comparable MSE. Gradient Boosting and XGBoost models also demonstrated robust performance, with accuracies ranging from 0.865 to 0.974 and MSE values between 9.60e+08 and 1.89e+08. Our approach includes extensive hyperparameter tuning and k-fold cross-validation to ensure model generalizability and robustness across agricultural scenarios. These findings highlight the effectiveness of ensemble methods in capturing complex data relationships and their superiority over traditional models in predicting crop yields. Our work sets the stage for future research into integrating real-time data and advanced hybrid models, aiming to refine predictive accuracy further and support sustainable agricultural practices.

Keywords: Machine Learning; XGBoost; Gradient Boosting; Crop yield prediction; Regression

1. Introduction

The accurate prediction of crop yields is a crucial aspect of agricultural planning and management, directly influencing food security, economic stability, and sustainable resource utilization. As global populations continue to rise and climate change exacerbates agricultural challenges, the demand for reliable crop yield forecasting has never been more critical. Traditional methods, which often rely on historical data and simplistic statistical models, fail to account for the com- plex interactions between numerous variables such as soil characteristics, weather patterns, and farming practices. This complexity necessitates more sophisticated approaches to in- tegrate and analyze diverse datasets to provide accurate and timely predictions, ultimately aiding farmers, policymakers, and stakeholders in making informed decisions.

In recent years, machine learning (ML) advancements have opened new avenues for enhancing crop yield predictions. Current research predominantly focuses on leveraging various ML algorithms, including regression models, decision trees,

ensemble methods, and neural networks, to improve the ac- curacy of predictions. Studies have shown that models such as Random Forest, Gradient Boosting, and Support Vector Machines (SVM) can effectively handle agricultural datasets' non-linear relationships and high-dimensionality characteris- tics. Despite these advancements, many models still

* Corresponding author: Bushara A. R

struggle with generalizability across different regions and crop types, and there is a significant need for models that can maintain high predictive accuracy in diverse agricultural contexts.

The primary challenge identified in existing research is the inadequacy of many ML models for handling the complex, non-linear relationships inherent in agricultural data. Tradi- tional models often fail to capture the nuanced interactions between climatic conditions, soil properties, and crop yields, leading to suboptimal performance. To address these challenges, our research proposes applying ensemble methods, specifically random forest and bagging regression, which have shown promise in previous studies for their robustness and ability to generalize across various datasets. These models are designed to integrate multiple weak learners to form a strong predictor, enhancing accuracy and reducing the risk of overfitting. Our proposed solution involves a comprehensive evaluation of these models alongside other ML techniques, utilizing a dataset rich in climatic and agricultural variables to identify the most effective approach for crop yield prediction. The primary contributions of our research are as follows:

- We conduct an extensive comparative analysis of multi- ple ML models, highlighting each model's strengths and weaknesses in predicting crop yields based on a diverse set of agricultural and climatic features.
- Our research demonstrates the superior performance of ensemble methods, specifically hyperparameter-tuned Random Forest and Bagging Regressor, which consis- tently outperform other models in accuracy and robust- ness.
- We emphasize the critical role of k-fold cross-validation in ensuring the generalizability and reliability of predictive models in agricultural applications. By system- atically evaluating model performance across multiple folds, we assess each model's ability to generalize to new data more rigorously.
- Our study integrates various agricultural and climatic features, providing a comprehensive approach to crop

yield prediction. This integration enables the models to account for the multifaceted influences on crop yields, thereby enhancing the accuracy and applicability of the predictions in real-world conditions.

The remainder of this paper is structured as follows: Sec- tion 2 reviews related work and the state-of-the-art in crop yield prediction using ML. Section 3 describes the dataset, detailing the features and preprocessing steps involved and our methodology, including the models evaluated and the criteria for their selection. Section 4 discusses the experimental results, highlighting the performance of each model. Finally, Section 5 concludes with our findings, their implications for agricultural planning, and potential directions for future research.

2. Related Works

Recent advancements in ML have significantly enhanced the accuracy of crop yield prediction models, leveraging diverse techniques and datasets. Jovanovic et al. [6] employed metaheuristic-tuned weight-agnostic neural networks to predict crop yields, demonstrating how metaheuristic approaches can optimize neural network architectures to improve prediction performance. Similarly, Kolipaka and Namburu [7] proposed a two-stage classifier framework incorporating meta-heuristics to refine crop yield predictions, illustrating the effectiveness of combining multiple classification stages with optimization techniques for higher accuracy. Zare et al. [8] explored within- season crop yield prediction using a multi-model ensemble approach, integrating data assimilation to enhance model ro- bustness and predictive capability across different growing conditions. Their work highlights the advantages of ensemble methods to capture diverse patterns and trends within agricul- tural data.

Gopi and Karthikeyan [10] introduced an innovative crop rec- ommendation and yield prediction model using an ensemble of recurrent neural networks optimized by Red Fox optimiza- tion. Their approach underscores the potential of combining evolutionary algorithms with neural network ensembles for superior performance in agricultural applications. Chaudhary and Pathak [11] developed a crop yield prediction model using a bi-directional LSTM under the PySpark framework, showcasing how advanced deep learning (DL) techniques and big data platforms can facilitate the processing of large-scale agricultural datasets for accurate yield forecasts. Additionally, Bhadra et al. [13] utilized a 3D CNN for plot-scale soy- bean yield prediction, integrating multitemporal UAV-based RGB images to capture temporal changes in crop growth, demonstrating the efficacy of convolutional neural networks in processing and analyzing remote sensing data for precise yield estimates. Wang et al. [15] combined CNN and GRU in a DL framework to improve wheat yield estimates using time-series remotely sensed multi-variables, highlighting the potential of hybrid models in capturing both spatial and temporal dynamics for crop yield prediction. learning network that advances plant and leaf classification, emphasizing the utility of multitask learning in improving model performance across related tasks. Vardhan and Sharma

[12] further explored the application of hierarchical convolu- tional neural networks for plant pathology, demonstrating sig- nificant improvements in disease identification accuracy. Saini et al. [14] combined CNN and Bi-LSTM in their DL approach for sugarcane yield prediction, illustrating how integrating con- volutional layers with extended short-term memory networks can effectively model both spatial and sequential data for yield forecasting. These studies underscore the benefits of using DL and ensemble methods to capture complex relationships within agricultural datasets, leading to more accurate and reliable crop yield predictions.

Despite the advancements in crop yield prediction models, sev- eral gaps remain in the current research. Many existing models, including those developed by Wang et al. [15] and Bhadra et al. [13], are limited by their specific focus on single crop types or lack of generalizability across different agricultural contexts. Furthermore, the complexity of integrating diverse climatic and soil variables into predictive models is often un- derexplored. Our research addresses these gaps by proposing a comprehensive approach that integrates multiple ensemble methods, such as Random Forest and Bagging Regressor, which have demonstrated superior performance in various settings. Additionally, using a diverse dataset encompassing a wide range of agricultural and climatic features ensures the models' applicability across different crop types and growing conditions, enhancing their generalizability and robustness. Future work could build on our findings by exploring more advanced DL techniques and real-time data integration to refine crop yield predictions further and support sustainable agricultural practices.

3. Method

3.1. Dataset

In this study, we have utilized the Crop Yield Prediction Dataset from FAO (Food and Agriculture Organization) [16] and World Data Bank [17], which offers a comprehensive set of agricultural data essential for analyzing and predicting crop yields across different regions. This dataset includes five dis- tinct files, each providing crucial information: the Yield Data file consists of 56,717 entries covering crop yields for various crops across different countries and years, capturing key details such as area, crop type, year, and yield in hectograms per hectare. The Temperature Data file contains 71,311 entries, providing average temperature data for 137 countries over 271 years, with some missing values in the temperature column. The Rainfall Data file includes 6,727 entries detailing average annual rainfall across 217 areas over 31 years, with some entries missing rainfall data. The Pesticide Usage Data file has 4,349 entries covering pesticide usage for various crops across 168 regions over 27 years. Finally, the Comprehensive Crop Yield Data file integrates all the information above, containing 28,242 entries with no missing values. This file provides a holistic view of crop yield data, encompassing details such as average temperature, rainfall, pesticide usage, and yield metrics across 101 areas and 10 different crop types over 23 years. The dataset's richness and variety make it a valuable resource for in-depth analysis and robust prediction of crop yields by integrating various environmental and agricultural factors.

3.2. Proposed Work

This study proposes a comprehensive approach to predict crop yields by leveraging advanced ML techniques. Our methodology involves a detailed data preprocessing phase, the application of various ML models, and subsequent hyperpa- rameter tuning to optimize the performance of the models. The goal is to identify the most effective model for accurately pre- dicting crop yields based on various agricultural and climatic factors.

Algorithm 1 focuses on developing a robust model for pre- dicting crop yields using diverse features. The primary steps include data loading, preprocessing, EDA, data preparation, and model training.

We start by importing the necessary libraries, such as numpy and pandas, and then load the dataset D from vari- ous files (yield.csv, temp.csv, rainfall.csv, pesticides.csv, and yield_df.csv). The dataset comprises features like temperature, rainfall, pesticides, and other agricultural parameters essential for predicting crop yields. We handle any missing values and duplicates to ensure data integrity.

Basic EDA is performed to understand feature distributions and relationships. This includes visualizing distributions and correlations among the features to uncover potential patterns and insights.

The dataset is split into features X and the target variable y, where X contains all columns except for the crop yield (hg/ha_yield), and y represents the crop yield. We convert categorical variables into dummy variables using:

$$X \leftarrow \text{pd.get\_dummies}(X) \ldots\ldots\ldots \quad (1)$$

to facilitate numerical analysis. The data is then split into training and testing sets using a train-test split with 20% of the data reserved for testing:

$$\text{Xtrain, Xtest, ytrain, ytest} \leftarrow \text{train\_test\_split (X, y, 0.2, 1)} \ldots\ldots\ldots(2)$$

where 0.2 denotes the proportion of the dataset to include in the test split, and 1 sets the random seed for reproducibility.

We define a list of ML models including Linear Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), XGBoost (XGB), Bagging Regressor (BR), and K-Nearest Neighbors (KNN), where each model is initialized with specific parameters.

We train each model on the training data Xtrain and ytrain and evaluate their performance on the testing data Xtest and ytest. The model training process involves fitting the model:

$$\text{model.fit(Xtrain, ytrain)} \ldots\ldots \quad (3)$$

and making predictions:

$$\hat{y}_{test} \leftarrow \text{model.predict}(X_{test}) \ldots\ldots\ldots\ldots(4)$$

The predictions $\hat{y}_{test}$ are compared against the actual test labels y_{test} using various evaluation metrics

Algorithm 1 Crop Yield Prediction Model

Require: Dataset D with features: temperature, rainfall, pes- ticides, etc.

Ensure: Preprocessed dataset for crop yield prediction.

- **Import Libraries:** numpy, pandas, sklearn, xgboost
- **Data Loading and Preprocessing:**
- Load D from: *yield.csv*, *temp.csv*, *rainfall.csv*, *pesticides.csv*, *yield_df.csv*
- Clean data: handle missing values, duplicates
- **EDA:** Perform basic exploratory data analysis
- **Data Preparation:**
- Split D: $X, y \leftarrow D.drop('hg/ha_yield', axis = 1), D['hg/ha_yield']$
- Convert categorical to dummy variables: $X \leftarrow$ pd.get_dummies(X)
- Split into train/test: $X_{train}, X_{test}, y_{train}, y_{test} \leftarrow$train_test_split($X, y,$ 0.2, 1)
- **Model Training:**
- Define models:
  - models $\leftarrow$ {('LR', LinearRegression()),
  - ('DT', DecisionTreeRegressor(1)),
  - ('RF', RandomForestRegressor(1)),
  - ('GB', GradientBoostingRegressor(100, 0.1, 3, 1)), ('XGB', XGBRegressor(1)),
  - ('BR', BaggingRegressor(100, 1)), ('KNN', KNeighborsRegressor(10))}
- Train and evaluate models on X_{train}, y_{train}:
  - for (m _ name, model) in models: model.fit(Xtrain, ytrain) yˆtest $\leftarrow$ model.predict(Xtest) $\epsilon \leftarrow$ Evaluate(yˆtest, ytest) results.append((m_name, ϵ))
- Model Evaluation:
- Evaluate models using MAE, MSE, and R^2
- Return Best model and metrics

Algorithm 2 focuses on optimizing the hyperparameters of the top-performing models from the first algorithm to enhance their predictive performance. The models considered for hy- perparameter tuning are Random Forest, Bagging Regressor, and XGBoost.

Random Forest Tuning: For the Random Forest model, we explore a grid of hyperparameters, including the number of es- timators (n_estimators), maximum tree depth (max_depth), minimum samples for splitting a node (min_samples_ split),

Algorithm 2 Hyperparameter Tuning for Crop Yield Prediction Models Require: Training dataset Xtrain, ytrain.

Ensure: Optimized models with tuned hyperparameters.

- Random Forest Tuning:
- Define parameter grid:
  - param_grid_rf ← {′n_estimators′ : [100, 200, 300],
  - ′max_depth′ : [10, 20, 30],
  - ′min_samples_split′ : [2, 5, 10],
  - ′min_samples_leaf′ : [1, 2, 4]}
- Use GridSearchCV: rf _best ←
  - GridSearchCV(RandomForestRegressor(1), param_grid_rf)
- Fit rf _best: rf _best.fit(Xtrain, ytrain)
- Bagging Regressor Tuning:
- Define parameter grid:
  - param_grid_br ← {′n_estimators′ : [50, 100, 150],
  - ′max_samples′ : [0.5, 0.7, 1.0],
  - ′max_features′ : [0.5, 0.7, 1.0]}
- Use GridSearchCV: br_best ←
  - GridSearchCV(BaggingRegressor(1), param_grid_br)
- Fit br_best: br_best.fit(Xtrain, ytrain)
- XGBoost Tuning:
- Define parameter grid:
  - param_grid_xgb ← {′n_estimators′ : [100, 200, 300],
  - ′learning_rate′ : [0.01, 0.1, 0.2],
  - ′max_depth′ : [3, 6, 9],
  - ′subsample′ : [0.8, 1.0]}
- Use GridSearchCV: xgb_best ←
  - GridSearchCV(XGBRegressor(1), param_grid_xgb)
- Fit xgb_best: xgb_best.fit(Xtrain, ytrain)
- Model Comparison:
- Compare models using MAE, MSE, and R2 score.
- Return Best tuned model and evaluation metrics.

And minimum samples at a leaf node (min_samples_leaf). The grid is defined as follows:

param_grid_rf ← {′n_estimators′ : [100, 200, 300],

′max_depth′ : [10, 20, 30],

′min_samples_split′ : [2, 5, 10],

′min_samples_leaf′ : [1, 2, 4]}

We employ GridSearchCV for exhaustive search over the parameter grid to find optimal parameters:

$$rf\_best \leftarrow GridSearchCV(RFRegressor(1), param\_grid\_rf)..(5)$$

The best parameters are identified by fitting the model to the training data:

$$rf\_best.fit(Xtrain, ytrain)\ldots\ldots \quad (6)$$

Bagging Regressor Tuning: Similarly, for the Bagging Re- gressor, we define a grid of hyperparameters, including the number of estimators (n_estimators), maximum samples per base estimator (max_samples), and maximum features per base estimator (max_features). The parameter grid is:

param_grid_br ← {′n_estimators′ : [50, 100, 150],

′max_samples′ : [0.5, 0.7, 1.0],

′max_features′ : [0.5, 0.7, 1.0]}

GridSearchCV is utilized to find the best parameters:

br_best ← GridSearchCV(BR(1), param_grid_br) (7) The model is fitted as follows:

br_best.fit(Xtrain, ytrain)……..(8)

XGBoost Tuning: For XGBoost, the hyperparameters tuned include the number of estimators (n_estimators), learning rate (learning_rate), maximum tree depth (max_depth), and subsample ratio (subsample). The parameter grid is defined as:

param_grid_xgb ← {′n_estimators′ : [100, 200, 300],

′learning_rate′ : [0.01, 0.1, 0.2],

′max_depth′ : [3, 6, 9],

′subsample′ : [0.8, 1.0]}

GridSearchCV is employed to find the optimal parameters:

xgb_best ← GridSearchCV(XGBRegressor(1), param_grid_xgb)……..(9)

The model is fitted as:

xgb_best.fit(Xtrain, ytrain)…….(10)

We have implemented a comprehensive approach for pre- dicting crop yields using advanced ML techniques. The first part of our methodology involves preprocessing a diverse crop yield dataset and training multiple ML models, including Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, Bagging Regressor, and K-nearest neighbors. We performed data cleaning, conversion of categorical variables, and a train-test split to ensure robust training and evaluation. Each model was evaluated using key performance metrics to identify the best-performing model. In the sec- ond part, we focused on hyperparameter tuning for the top- performing models—Random Forest, Bagging Regressor, and XGBoost—using GridSearchCV to optimize their parameters and enhance their predictive accuracy. We selected these three models for tuning due to their initial solid performance and ability to handle complex interactions within the data. This approach ensures that our model is accurate and generaliz- able across different agricultural conditions. We chose this comprehensive strategy to leverage the strengths of various algorithms, ensuring that the final model can effectively predict crop yields by integrating key agricultural and climatic factors.

3.2.1. Model Evaluation

The performance of each model is evaluated using three key metrics: Accuracy, Mean Squared Error (MSE), and the coefficient of determination (R2). These metrics provide a comprehensive assessment of the model's predictive performance and accuracy.

Accuracy: In regression tasks, accuracy can be interpreted as how close the predicted values are to the actual values. Although more commonly used in classification, for regression purposes, accuracy is generally assessed through proximity measures like MSE and R2. High accuracy indicates that the model's predictions are very close to the actual

values, reflecting the model's reliability in making precise predictions. Mean Squared Error (MSE): MSE measures the average of the squared differences between predicted values (yˆi) and

actual values (yi). It is given by:

$$f(z) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y_i)^2 \qquad (11)$$

where n is the number of observations. MSE emphasizes larger errors more than smaller ones, providing a sensitive metric to significant deviations in predictions. Lower MSE values indicate better model performance as the model errors are minimal.

Coefficient of Determination (R2): The R2 score measures the proportion of variance in the dependent variable that is predictable from the independent variables. It is computed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}-y_i)^2}{(y_i-\bar{y})^2} \qquad (12)$$

that its predictions are closely aligned with the actual values, reflecting minimal deviation. Similarly, the Bagging Regressor demonstrates comparable performance with an accuracy of 0.984792 and an MSE of 1.08e+08. These results underscore the effectiveness of ensemble methods in handling complex interactions within the dataset. The Decision Tree model also

Table 1 Model Performance Metrics

Model	Accuracy	MSE	R2 Score
Linear Regression	0.751364	1.77e+09	0.751364
Decision Tree	0.978228	1.55e+08	0.978228
Random Forest	0.984811	1.08e+08	0.984811
Gradient Boost	0.865138	9.60e+08	0.865138
XGBoost	0.973514	1.89e+08	0.973514
Bagging Regressor	0.984792	1.08e+08	0.984792
KNN	0.332706	4.75e+09	0.332706

performs remarkably well, with an accuracy of 0.978228 and an MSE of 1.55e+08, showcasing its robustness in making precise predictions despite its simplicity. Gradient Boosting and XGBoost models show strong performance with accuracies of 0.865138 and 0.973514, respectively, highlighting their capability to capture non-linear relationships within the data. However, the KNN model exhibits significantly lower accuracy at 0.332706 and a high MSE of 4.75e+09, indicating that it is not well-suited for this task.

Figures 1 visually compare each model's actual versus predicted values. The scatter plots illustrate the correlation between actual and predicted crop yields, with the red trend- ine indicating the line of perfect prediction. These plots help visually assess each model's performance and understand the accuracy achieved in predictions. Each subplot in Figure 1 where $\bar{y}$ is the mean of the actual values. An R2 value close to 1 indicates that the model explains a large portion of the variance in the dependent variable, while a value near 0 suggests poor explanatory power. High R2 values reflect the model's ability to capture the variability in the data effectively. By assessing the models using these metrics, we identify the best-performing model that achieves high accuracy, low MSE, and a high R2 score, ensuring robust and reliable predictions of crop yield.

4. Results and Discussion

We present and discuss the results of the ML models applied to predict crop yields. Each model's performance is assessed using Accuracy, MSE, and the coefficient of determination (R2). The comprehensive evaluation of these models high- lights their strengths and limitations in predicting crop yields based on various agricultural and climatic features.

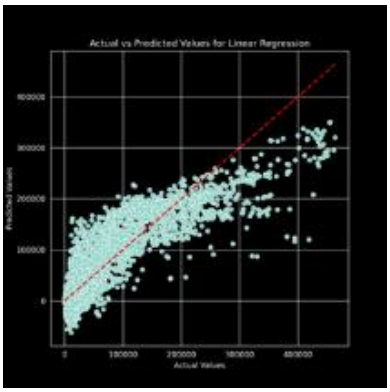
The evaluation metrics for each model are summarized in Table I. The table provides a comparative overview of the mod- els' predictive performance, which is critical for understanding their reliability and effectiveness. From Table I, we observe that the Random Forest model achieved the highest accuracy of 0.984811, indicating its superior ability to predict crop yields accurately. The model's Mean Squared Error (MSE) is also the lowest among all models at 1.08e+08, which shows

shows the actual crop yields plotted against the predicted yields for the corresponding model. The closer the data points align with the trendline, the better the model's predictive performance—models like Random Forest and Bagging Re- gressor display data points closely following the trendline, sig- nifying high accuracy. On the contrary, the KNN model shows significant dispersion, indicating poorer predictive capability.

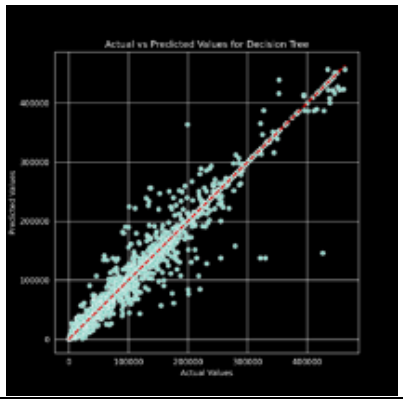
4.1. Discussion of K-Fold Validation Results

Figure 2 presents the accuracy scores for each model across 10 folds. The plots provide insight into the stability and consistency of each model's performance.

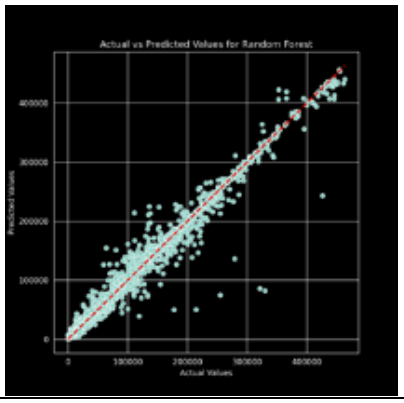
In Figure 2(a), the Linear Regression model shows relatively stable accuracy scores with minor fluctuations between 0.735 and 0.770 across the folds, indicating consistent performance. Figure 2(b) illustrates the Decision Tree model, which achieves high accuracy scores ranging from 0.980 to 0.986. The variability observed in the accuracy scores across folds highlights the model's sensitivity to the training data varia- tions.



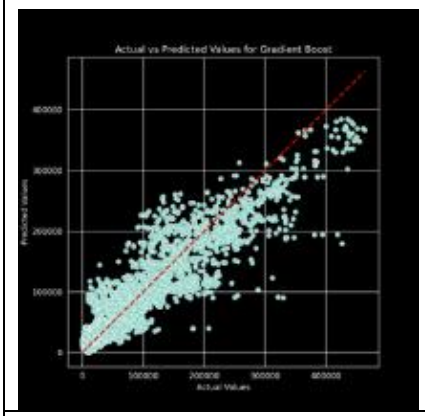
a Linear Regression



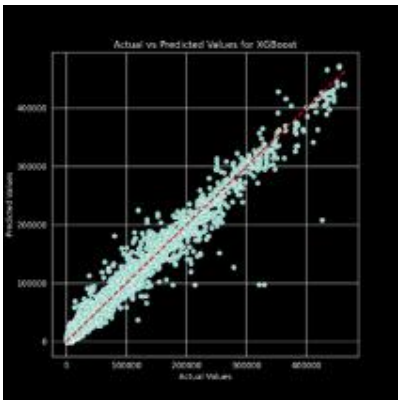
b Decision Tree



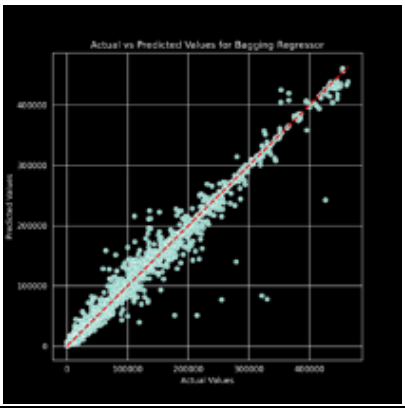
c Random Forest



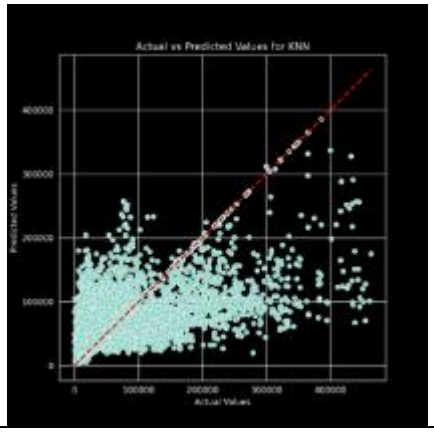
d Gradient Boost



e XGBoost



f Bagging Regressor



g KNN

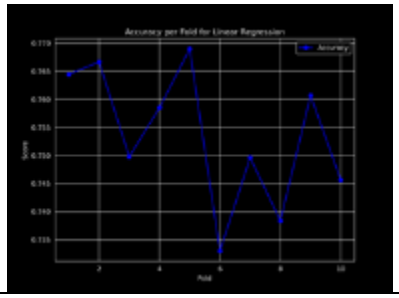
Figure 1 Actual vs Predicted Values for Different Models

In Figure 2(d), the Gradient Boost model exhibits fluctu- ations in accuracy between 0.862 and 0.880. This variability suggests the model's moderate sensitivity to the folds, which could be attributed to its iterative learning process that may overfit on certain subsets of the data.

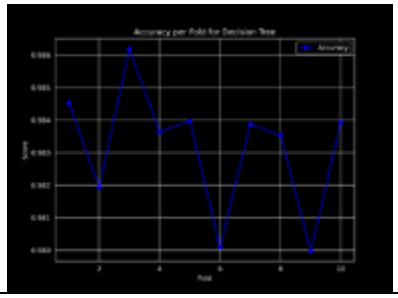
The XGBoost model, depicted in Figure 2(e), maintains high accuracy scores ranging from 0.973 to 0.979. The slight variations across folds indicate a strong generalization capa- bility with minimal overfitting.

Figure 2(f) shows the Bagging Regressor model, which maintains consistently high accuracy scores between 0.988 and 0.991. The stable performance across folds demonstrates the model's ability to aggregate multiple predictions effectively, leading to robust results.

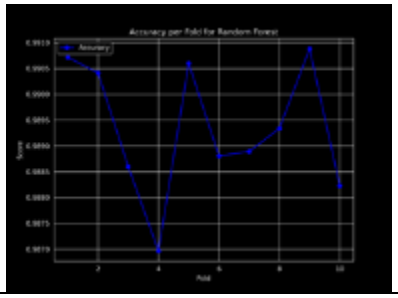
Lastly, Figure 2(g) displays the KNN model, which shows significant variation in accuracy scores ranging from 0.310 to 0.360. This variability reflects the model's high sensitivity to the different folds, suggesting it may not generalize well across varying subsets of the data.



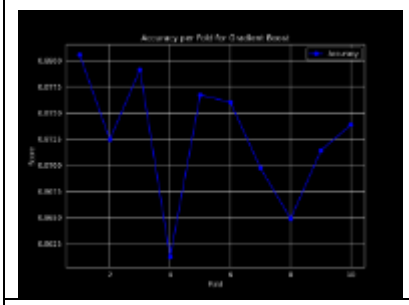
A Linear Regression



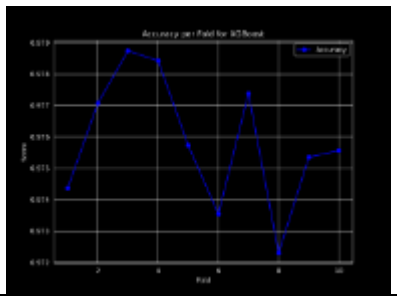
B Decision Tree



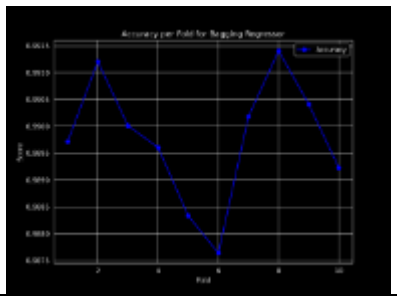
C Random Forest



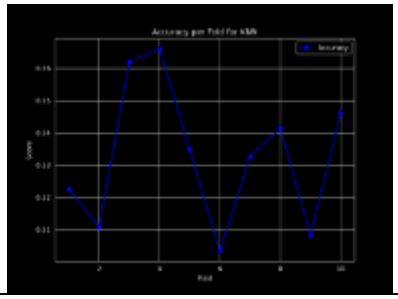
D Gradient Boost



E XGBoost



F Bagging Regressor



(g) KNN

Figure 2 K-Fold Validation Accuracy for Different Models. Each plot represents the accuracy scores across 10 folds for the respective model: (a) Linear Regression, (b) Decision Tree, (c) Random Forest, (d) Gradient Boost, (e) XGBoost, (f) Bagging Regressor, and (g) KNN

Evaluating various ML models for crop yield prediction reveals significant differences in their performance and robustness. The Random Forest model demonstrated superior performance with an accuracy consistently around 0.985 and an MSE of approximately 1.08e+08, making it the most reliable model across multiple metrics. The Bagging Regressor closely followed, with an accuracy of 0.984 and a similar MSE. The Decision Tree model also performed well, achieving an accuracy of 0.978 and an MSE of 1.55e+08, highlighting its simplicity yet effective predictive capability. Ensemble methods like Gradient Boost and XGBoost exhibited robust performance with accuracies of 0.865 and 0.974, respectively, and MSE values around 9.60e+08 and 1.89e+08, indicating their strong generalization capabilities. In contrast, the KNN model showed lower accuracy at 0.333 and a high MSE of 4.75e+09, reflecting its inadequacy for this task. The Linear Regression model, while straightforward, managed an accuracy of 0.751 and an MSE of 1.77e+09, demonstrating moderate effectiveness. K-fold validation further confirmed these findings, showing consistent performance across folds for Random Forest and Bagging Regressor, with accuracy ranges of 0.988 to 0.991 and minimal fluctuation. Conversely, models like KNN and Gradient Boost displayed more signif- icant variability in accuracy across folds, ranging from 0.310 to 0.360 and 0.862 to 0.880, respectively, indicating their sensitivity to the data splits. In conclusion, ensemble methods, particularly Random Forest and Bagging Regressor, proved the most effective and reliable for crop yield prediction, offering high accuracy, low error rates, and consistent performance across different data subsets.

5. Conclusion

This research thoroughly evaluates ML models for crop yield prediction using a dataset with various agricultural and climatic features. Ensemble methods, particularly Random Forest and Bagging Regressor, showed superior performance with accuracies around 0.985 and MSE values near 1.08e+08, making them highly reliable for this task. The Decision Tree model also performed well with an accuracy of 0.978 and an MSE of 1.55e+08, proving effective for more straightfor- ward scenarios. Gradient Boosting and XGBoost demonstrated strong capabilities with accuracies of 0.865 and 0.974 and MSE values ranging from 9.60e+08 to 1.89e+08. In contrast, the KNN model showed the lowest performance, with an accuracy of 0.333 and a high MSE of 4.75e+09, indicating its unsuitability for this application. The k-fold validation confirmed the consistency of Random Forest and Bagging Regressor, highlighting their robustness with accuracy scores between 0.988 and 0.991 across folds. This study underscores the effectiveness of ensemble methods for accurate and con- sistent crop yield predictions, offering valuable insights for agricultural planning and management. Future work could further explore advanced DL techniques, hybrid models, and real-time data integration to improve predictive accuracy and adaptability in agricultural contexts.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Oikonomidis, Alexandros, Cagatay Catal, and Ayalew Kassahun. ”Deep learning for crop yield prediction: a systematic literature review.” New Zealand Journal of Crop and Horticultural Science 51, no. 1 (2023): 1-26.

[2] Sharma, Shubham, and Manu Vardhan. ”Self-attention vision transformer with transfer learning for efficient crops and weeds classification.” In 2023 6th International Conference on Information Systems and Computer Networks (ISCON), pp. 1-6. IEEE, 2023.

[3] Attri, Ishana, Lalit Kumar Awasthi, Teek Parval Sharma, and Priyanka Rathee. ”A review of deep learning techniques used in agriculture.” Ecological Informatics (2023): 102217.

[4] Sharma, Shubham, and Manu Vardhan. ”Hyperparameter Tuned Hybrid Convolutional Neural Network (H-CNN) for Accurate Plant Disease Classification.” In 2023 International Conference on Communication, Circuits, and Systems (IC3S), pp. 1-6. IEEE, 2023.

[5] Cheung, Liege, Yun Wang, Adela SM Lau, and Rogers MC Chan. ”Using a novel clustered 3D-CNN model for improving crop future price prediction.” Knowledge-Based Systems 260 (2023): 110133.

[6] Babu Nuthalapati, S., & Nuthalapati, A., "Accurate Weather Forecasting with Dominant Gradient Boosting Using Machine Learning," Int. J. Sci. Res. Arch., vol. 12, no. 2, pp. 408-422, 2024, doi:10.30574/ijsra.2024.12.2.1246.

[7] Jovanovic, Luka, Miodrag Zivkovic, Nebojsa Bacanin, Milos Dobro- jevic, Vladimir Simic, Kishor Kumar Sadasivuni, and Erfan Babaee Tirkolaee. ”Evaluating the performance of metaheuristic-tuned weight agnostic neural networks for crop yield prediction.” Neural Computing and Applications (2024): 1-30.

[8] AR, B., & RS, V. K. (2022). A deep learning-based lung cancer classification of CT images using augmented convolutional neural networks. ELCVIA. Electronic letters on computer vision and image analysis, 21(1), 0130-142.

[9] Kolipaka, Venkata Rama Rao, and Anupama Namburu. ”An automatic crop yield prediction framework designed with two-stage classifiers: a meta-heuristic approach.” Multimedia Tools and Applications 83, no. 10 (2024): 28969-28992.

[10] Nuthalapati, S. B., Arun, M., Prajitha, C., Rinesh, S., & Abubeker, K. M. (2024, September). Computer Vision Assisted Deep Learning Enabled Gas Pipeline Leak Detection Framework. In 2024 5th International Conference on Smart Electronics and Communication (ICOSEC) (pp. 950-957). IEEE.

[11] Zare, Hossein, Tobias KD Weber, Joachim Ingwersen, Wolfgang Nowak, Sebastian Gayler, and Thilo Streck. ”Within-season crop yield prediction by a multi-model ensemble with integrated data assimilation.” Field Crops Research 308 (2024): 109293.

[12] Sharma, Shubham, and Manu Vardhan. ”MTJNet: Multi-task joint learn- ing network for advancing medicinal plant and leaf classification.” Knowledge-Based Systems (2024): 112147.

[13] AR, B., & RS, V. K. (2022). A deep learning-based lung cancer classification of CT images using augmented convolutional neural networks. ELCVIA. Electronic letters on computer vision and image analysis, 21(1), 0130-142.

[14] Gopi, P. S. S., and M. Karthikeyan. ”Red fox optimization with ensemble recurrent neural network for crop recommendation and yield prediction model.” Multimedia Tools and Applications 83, no. 5 (2024): 13159-13179.

[15] S. B. Nuthalapati, "Advancements in Generative AI: Applications and Challenges in the Modern Era," Int. J. Sci. Eng. Appl., vol. 13, no. 8, pp. 106-111, 2024, doi:10.7753/IJSEA1308.1023.

[16] Chaudhary, Yashi, and Heman Pathak. ”CYPBL: Crop Yield Prediction using Bi-Directional LSTM under PySpark interface.” Multimedia Tools and Applications (2024): 1-20.

[17] Suri Babu Nuthalapati, "AI-Enhanced Detection and Mitigation of Cybersecurity Threats in Digital Banking," Educational Administration: Theory and Practice, vol. 29, no. 1, pp. 357–368, 2023, doi:10.53555/kuey.v29i1.6908.

[18] Vardhan, Manu, and Shubham Sharma. ”Enhancing Plant Pathology with CNNs: A Hierarchical Approach for Accurate Disease Identification.” In Proceedings of the 2024 13th International Conference on Software and Computer Applications, pp. 159-164. 2024.

[19] Muhammed Kunju, A. K., Baskar, S., Zafar, S., & AR, B. (2024). A transformer based real-time photo captioning framework for visually impaired people with visual attention. Multimedia Tools and Applications, 1-20.

[20] Bhadra, Sourav, Vasit Sagan, Juan Skobalski, Fernando Grignola, Supria Sarkar, and Justin Vilbig. ”End-to-end 3D CNN for plot-scale soybean yield prediction using multitemporal UAV-based RGB images.” Precision Agriculture 25, no. 2 (2024): 834-864.

[21] Suri Babu Nuthalapati and Aravind Nuthalapati, "Transforming Healthcare Delivery via IoT-Driven Big Data Analytics in a Cloud-Based Platform," J. Pop. Ther. Clin. Pharm., vol. 31, no. 6, pp. 2559–2569, Jun. 2024, doi:10.53555/jptcp.v31i6.6975.

[22] Saini, Preeti, Bharti Nagpal, Puneet Garg, and Sachin Kumar. ”CNN-BI- LSTM-CYP: A deep learning approach for sugarcane yield prediction.” Sustainable Energy Technologies and Assessments 57 (2023): 103263. Wang, Jie, Pengxin Wang, Huiren Tian, Kevin Tansey, Junming Liu, and Wenting Quan. ”A deep learning framework combining CNN and GRU for improving wheat yield estimates using time series remotely sensed multi-variables.” Computers and Electronics in Agriculture 206 (2023): 107705.