Journal of Information Systems Engineering and Management
2025, 10(27s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Ensemble Learning Framework for Crop Yield Prediction with Optuna Hyperparameter Tuning

Dr. S. Jayanthi[1*], Dr. M. A. Josephine Sathya[2], Dr. B. Nathan[3], Dr. Karthik Karmakonda[4], Dr. Muthuvel Laxmikanthan[5], Dr. Manojkumar V[6]

[1*]Senior Assistant Professor, Department of AI & DS, Faculty of Science and Technology (IcfaiTech), The ICFAI Foundation for Higher Education (IFHE), Hyderabad, Telangana – 501503, India.

[2]Assistant Professor, Department of Computer Science and Applications, Christ Academy Institute for Advanced Studies, Bangalore, Karnataka– 560 083, India.

[3]HOD, Department of Computer Science Engineering, Dhaanish Ahmed Institute of Technology, Coimbatore, TamilNadu - 641105, India.

[4]Associate Professor, Department of CSE, CVR College of Engineering, Hyderabad, Telangana 501510.

[5]HOD, Department of Artificial Intelligence and Data Science, Dhaanish Ahmed Institute of Technology, Coimbatore, TamilNadu - 641105, India.

[6]Department of Computer Science & Engineering, School of Engineering and Applied Sciences (SEAS), SRM University-AP, Amaravati - 522502, Andhra Pradesh, India.

ARTICLE INFO

Received: 21 Dec 2024

Revised: 20 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

The growing risk of food scarcity, along with climate change induced shifts in agriculture, demands precise crop yield predictions (CYP). Most existing machine learning (ML) and deep learning (DL) methods face challenges of integrating complex models from diverse data sources and accommodating different agro-ecological regions. Existing solutions do not offer a fully automated and explainable ensemble approach at this scale. This research proposes an automated and explainable ensemble learning framework, using Optuna for hyper-parameter optimization to tune eight regressor models, Gradient Boosting, XGBoost, LightGBM, CatBoost, Random Forest, Bagging Regressor, and KNN, for improved accuracy and generalization. Through the use of multi-source agricultural data and Explainable AI (XAI), our approach seeks to achieve high performance while retaining interpretability. The traditional Gradient Boosting model outperformed other classical ML models achieving $R2$ = 0.999 and RMSE=3298.326. Other traditional ML models could not match the performance of the ptoposed optimized models in this study. Important explanatory factors such as amount of pesticide applied, Temperature, and Rainfall were identified through SHAP analyses to underpin yield variability, enabling precise farming. By integrating automation, optimization, and advanced algorithms, the work enables more intelligent agricultural forecasting that allows farmers to make better data driven decisions.

Keywords: Crop Yield Prediction (CYP), Gradient Boosting (Gbst), Xgboost (Xbst), Random Forest (RF) Optuna, Explainable AI (XAI), SHAP Values, Precision Agriculture, Ensemble Learning, Sustainable Agriculture, Smart Farming, Data-Driven Decision Making.

INTRODUCTION

Precise CYP is the intersection of global food security and effective agricultural management [1, 2]. Traditional statistical models very often do not include all the subtle relations among yield-affecting factors especially when the climate is changing unpredictably [3]. On the other hand, the recent leaps in ML and DL make use of piles of data such as remote sensing, meteorological data, and soil characteristics in the development of substantially more precise and robust predictive models [4, 5]. These computerized data tools are the key to the present-day agriculture, making proactive techniques of decision-making possible that are focused on productivity optimization [6].

Meanwhile, there are also many very serious difficulties that still need to be resolved. Despite being able to reach high prediction accuracy, DL models are difficult to understand due to their black-box characteristics and thus very hard actionable insights to be extracted for agricultural decision makers [7, 8]. The connection of multi-modal data and the adaption of models to diverse agro-ecological settings are areas of future inquiry [9, 10]. Closing these gaps demands not only highly accurate models but also watertight and adaptable ones.

To address these issues, this study presents a fully automated, explainable ensemble learning (EL) framework that can be used to improve accuracy and interpretability.

The major contributions of this study are:

- Systematic hyperparameter tuning using Optuna across eight ensemble classifiers to enhance prediction performance and generalization.
- Utilization of multi-source agricultural data that combines remote sensing, climate conditions, and soil characteristics to seize intricate yield-controlling factors.
- Use of XAI techniques, especially Shapley Additive Explanations (SHAP) values, to increase transparency and provide useful explanations of the model outcomes.
- Creation of an ML-based agricultural forecasting model that is accurate and easy to modify for different agro-ecological regions, which is a primary limitation of most existing models.
- Facilitation of as well as the provision of an automated CYP model that is clear and easy to interpret in order to stimulate agricultural policies based on data.

Integrating Optuna-based hyperparameter tuning and XAI in this model, enables maintenance of explainability and adaptibility while improving accuracy, modifying it for widespread usability, and enhancing the quality of yield prediction models to foster intelligent agronomy.

RELATED WORK

The CYP is one of the primary activities of precision agriculture aimed at safeguarding food production and improving the use of resources. There is more than sufficient evidence that predictive accuracy can be enhanced with the use of ML and DL techniques, but there are still important predictive challenges of building accurate models that are interpretable and generalizable to other agro-ecological regions. Feature selection, EL, automated DL architectures, and even Automated ML have been investigated by existing studies. Yet, there still remains a wide gap towards effective integration of multiple data sources, model explainability, and Real World application context adjustability. The process of feature selection is one of the most critical tasks in ML as filters. Any predictor that does not contribute useful information can be removed, and as a result, it becomes easier to enhance model generalization. Lan et al. [10] incorporated mutual information and genetic algorithms with gradient boosting regression, thereby establishing the requirement for strong feature selection in addressing high-dimensional agriculture datasets. Abdel-Salam [11] similarly outlined a hybrid scheme integrating K-means clustering with correlation based filtering and subsequent FMIG-RFE feature selection. An enhanced Crayfish Optimization Algorithm used Support Vector Regression hyperparameter optimization to increase accuracy and improve computational efficiency. Tree models have consistently exhibited strong predictive ability in agricultural prediction. According to Jhajharia et al. [12], RF was identified (R^2 = 0.963, RMSE = 0.035, MAE = 0.0251) as the most accurate model which outperformed other ML

methods. In the same way, Burdett et al. [13] compared the performance of different ML models in precision agriculture, and found the R2 score for corn and soybeans as 0.85, and 0.94 respectively, using RF. The seasonal information is necessary, Filippi et al. [14] argued along with other models trying to estimate wheat, barley and canola yields. The implementation of DL frameworks has also been done in the recent past. Oliveira et al. [15] studied transformer-based models capable of capturing long-term dependencies in time series data related to crops and achieved remarkable results. With the increasing application of sophisticated black-box DL algorithms, however, the issue of their interpretability becomes very problematic which is the strong argument for using XAI technologies. The monitoring of agriculture in real-time using IoT technology has changed dramatically the CYP. Talaat et al. [16] developed the CYP Algorithm which merges IoT features with DT, RF, and Extra Tree (ET) Regressors. It provided $R^2 = 0.9933$ with Extra Tree Regressor. Their system also employed active learning to reduce the amount of labeled data needed. EL methods have surfaced as powerful models for reinforcing predictive accuracy. Kuppan et al. [17] is the one who showed DT, Extra Tree (ET), and CatBoost (CatBst) classifiers performed well, with Extra Tree providing 99.15% accuracy. Ramesh et al. [6], the other author, built a stacked ensemble model with six base learners and a Decision Tree meta-model, which achieved $R^2 = 0.98$ with reduced RMSE. AutoML frameworks have further automated the processes of model selection and hyperparameter tuning. Kheir [18] examined the performance of 22 ML models for wheat yield prediction by stacking the models and got $R^2 = 0.7$. The growing gap regarding the implementation of XAI in agriculture was addressed by Mariadass et al. [19] who created a multi-objective AutoML framework focusing on accuracy and interpretability.

Although the literature presents advances towards feature selection, EL, DL, and IoT-based methodologies [20-25], the following notable gaps persist:

- Most studies do not systematically combine remote sensing, meteorological, and soil data for yield prediction. They did not explore multi-source data integration.
- The black-box nature of ML models renders them useless in agricultural settings despite their higher accuracy because stakeholders fail to make sense of the results.
- Many models are trained on specific regional datasets to increase accuracy but this greatly reduces adaptability to other agricultural environments.
- Most studies use only a limited number of models for evaluation which are unable to depict the reality of yield prediction.

To fill these leaps, this research proposes the fully automated, explainable EL framework that will adjust all the hyperparameters in ensemble models and will become more interpretable. Our central achievements are as follows:

- Systematic Hyperparameter Optimization with Optuna: We utilize Optuna for optimization of hyperparameters through eight different diverse models, and we obtain an optimal model performance.
- Multi-Source Agricultural Data Integration: We bring together multiple data sets including soil, climate, and remote sensing data, for developing a more accurate prediction model.
- Improved Model Transparency via XAI: SHAP values are used to explain the features that the model is using and makes better decisions for alerting the users about how to take care of the farm land and also the responsibility of the government to the farmers.
- Evaluation across Diverse Agro-Ecological Conditions: We conduct our study on multiple datasets to determine the generalization and the differences of agriculture environments in which case our approach is validated.
- Comprehensive Benchmarking against Existing Models: We present an investigation comparing multiple ML models and suggest an efficient model in terms of both the accuracy of prediction and computational efficiency, which is often overlooked by the researchers.

By leveraging state-of-the-art hyperparameter tuning, XAI, and multi-data sources, this work provides an interpretable and scalable method to improve the accuracy of crop yield prediction.

METHODOLOGY

A. Overview of the Proposed Framework

The process diagram of the proposed research for CYP is shown in Fig.1.

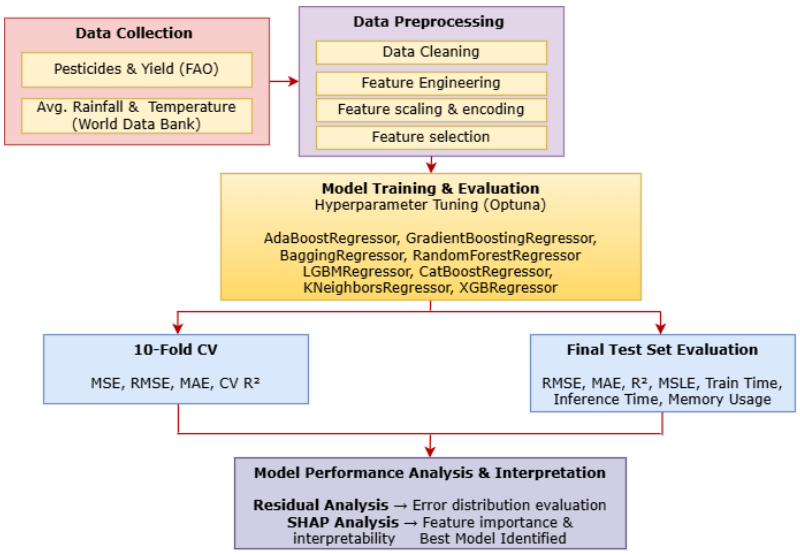

Figure 2: The Process Diagram of the Proposed Research

B. Dataset Description

This analysis makes use of reliable sources which provide quality agricultural and climate data. Pesticide usage and crop yield data was collected from the FAO provided FAOSTAT database [25], whereas rainfall and temperature data were obtained from the World Bank Climate Change Indicators database [26]. These databases are very rich to enable integrative studies of crop productivity in different agro-ecological zones and regions. The integrated dataset contains 4349 individual observations from 168 countries over the period 1990 to 2016. It has predefined important agricultural and environmental parameters such as crop yield (hg/ha), average rainfall in mm per year, pesticide used in tonnes, and average temperature yearly in degrees Celsius. The pesticide's variability is high; it ranges from 0 to 1.81 million tonnes with an average of 20,300 and standard deviation of 118,000 tonnes, which indicates

diverse global farming practices. The study focus was on several crops such as potatoes, sorghum, soybeans, and several others to enhance the applicability scope. The dataset was cleaned, merged and missing values were dealt to provide complete datasets without bias. Table 1 shows a subset of the dataset which reflects the geographical diversity in rainfall, temperature, pesticide application, and crop yield for various crops across regions. For example, Qatar's potato yield with low rainfall (74mm) is very low when compared with South African sorghum yield which receives significantly higher rainfall of 495mm. This scenario variability is extreme, which speaks to the depth of the dataset and its capture of many different agricultural and climatic conditions for the purposes of predictive modeling. This comprehensive dataset provides a robust foundation for developing accurate and interpretable CYP models.

Table 1: Sample Records from the Final Integrated Dataset

Unnamed	Area	Item	Year	Crop yield (hg/ha)	avg_rain_fall (mm)	Pesticides (Tonnes)	Avg temp
24811	South Africa	Potatoes	1999	315545	495.0	26098.80	21.64
24933	South Africa	Sorghum	2004	28692	495.0	26857.00	18.75
4877	Bulgaria	Soybeans	1999	8333	608.0	3004.75	9.67
23775	Qatar	Potatoes	2004	80000	74.0	68.00	28.20
22182	Pakistan	Soybeans	1997	12942	494.0	16936.00	22.14

C. Exploratory Data Analysis

A comprehensive exploratory data analysis (EDA) was performed to examine temporal trends, feature distributions, and inter-variable correlations within the dataset, thereby informing our data preprocessing and model-building strategies.

Figure 2 presents normalized time-series data (1990–2016) for crop yield, rainfall, pesticide usage, and temperature across all regions. Normalization was performed by scaling each variable to the range [0,1], thereby enabling direct visual comparisons despite inherent differences in their units of measurement.

The analysis reveals some important aspects:

1. The agricultural output increases slowly with improved farming technologies and jumps sharply in the 2000s.
2. Rainfall is evidently cyclical, mirroring seasons, with some rains occurring when yields are higher.
3. Pesticide use is correlated with agricultural expansion and likely increased because of pest pressures or as a rationale to increase yields.
4. And most notably, there is an observable increase in temperature over the years which raises concerns about the thermal sensitivity of crops as time goes by.

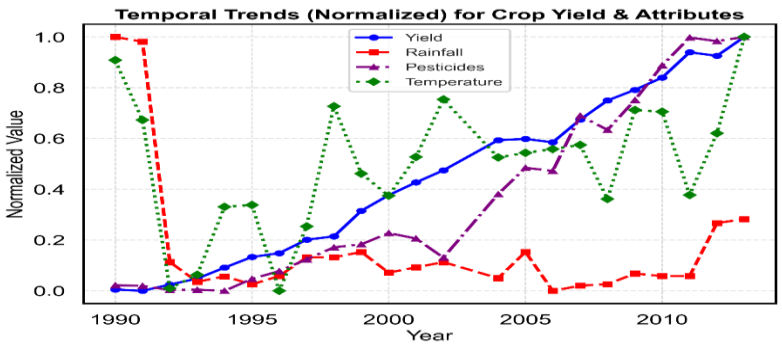

Figure 2: Temporal Trends for Crop Yield and Attributes

In Figure 3, average crop yield, rainfall, and temperature are illustrated for each group of crops according to their agronomic features. The high-yield crops (for instance, potato and cassava) outperform the drought-tolerant cereals (millet and sorghum).
Rice and maize have better yields in regions with medium to high rainfall due to the constant water supply. Yield variability (10-20% standard deviation) shows the variation between regions' farming practices and climates.

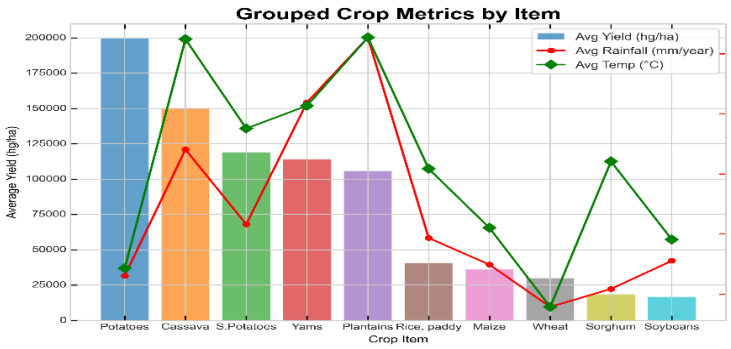

Figure 3: Grouped Crop Metrics by Item

The matrices of correlation (Pearson's, Spearman's, Kendall's) of the main variables are presented in Figure 4. Rainfall and yield have a medium positive correlation in all the plots which serves to substantiate the necessity of water. In regard to yield, temperature is poorly correlated which imply more complex issue, perhaps due to the presence of thresholds or interactions with other factors. At the same time usages of pesticides slightly positively correlates with yield which suggests that there may be some underlying supressing factors, but a more elaborate integration between the diverse influencing factors is essential. These results encouraged the development of more sophisticated models able to explain the intricate agronomic patterns in the EDA.

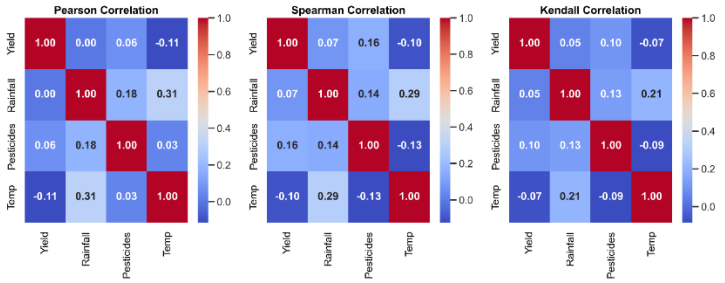

Figure 4: Correlation Analysis using Pearson, Spearman and Kendall

D. Data Preprocessing

In this study, an effective data preprocessing was performed through a methodical pipeline at data integration, missing value computation, outlier processing, feature value standardization, categorical data transformation, and sample data partitioning stages was conducted which emerged with useful models. The analyzed outlier datasets integrated agricultural yield data from different regions, metrological information (rainfall, temperature), and data on pesticide usage was done by linking the records with the unique IDs (location, time). The datasets (yield_df, rainfall_df, temp_df, and pesticides_df) were merged and required string matching, metadata verification, data cleansing in order to remove duplicates.
Missing numeric values were replaced with the median, which eliminates bias from extreme values, whereas missing categorical values were substituted with the descriptor "Unknown" in order to safeguard context. Total changes, discovered through boxplots and whittling the interquartile range (IQR), were winsorized to the 5th and 95th percentiles to lower their effects without altering relative distribution.

To scale measurements, we implemented Min-Max Scaling first on the numeric features as defined in Eq(1):

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

This form of normalization improves performance using neural networks that are sensitive to scale. Subsequently, Z-score standardization was performed as in Eq.(2):

$$x = \frac{X - \mu}{\sigma} \quad (2)$$

ensuring features are filtered on average at zero, which accelerates convergence towards a solution, while standard deviation is set to one. For enhanced interpretability and reduced stability Min-Max Scaling was applied for gradient based models while Z-score standardization was utilized for distance-based models. Furthermore, log transformation was used on features such as pesticides, avg_rainfall, and avg_temp in order to reduce the amount of bias a skew and normal distribution approximates.
Tree based models require label encoded categorical variables in order lessen dimensionality while maintaining compatibility and finally the dataset was split stratified into training and test sets, 80%, 20% respectively to maintain bias and proportion of classes.

E. Model Selection and Optimization

Finding a suitable predictive model is important as it needs to achieve adequate accuracy, robustness, and generalizability, which dictates the comprehensive evaluation of different ensemble model techniques, boosting techniques, and distance based models, amongst other machine learning methods. The criteria used for selection were based on predictive accuracy and computational time.
RF which reduces variance reduction through the combination of various decision trees and is calculated as in Eq(3):

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \quad (3)$$

N, as noted in the equation, is the number of trees in the ensemble.
XGBoost, LightGBM, and CatBoost, structured data regressors, integrate weak learners in a boosting iterative and optimize through gradient boosting methods.
XGBoost utilizes a second order Taylor approximation in the optimization of a loss function with L1 and L2 regularization for generalization. The objective function for XGBoost is given as in Eq.(4):

$$L(\theta) = \sum_{i=1}^{n} l(y_i - \hat{y}_i) - \sum_{k=1}^{T} \Omega(f_k) \quad (4)$$

where $l(y_i - \hat{y}_i)$ and $\Omega(f_k)$ represent loss, and complexity functions, respectively. LightGBM uses histogram-based splitting to reduce computational complexity. CatBoost can use the ordered boosting technology to solve target leakage for the categorical features. Also, outcome prediction by a weighted average of the k nearest neighbors is performed by the k-NN regressor and it defined as in Eq.(5):

$$f(x) = \sum_{i=1}^{n} w_i y_i, \quad w_i = \frac{1}{d(x_i, x)} \quad (5)$$

Where d(xi, x) is the chosen distance metric.
Hyperparameter tuning was executed using Optuna, a Bayesian optimization which relies on Tree-structured Parzen Estimator (TPE) for search space particulalry modification. This approach was preferred to traditional ones (e.g., GridSearchCV, Randomized Search) because of the lower computational cost and better scalability. In order to trust these performance improvements, models were reinitialized before every cross validation fold to block state leakage. Joblib was used to control model persistence because it is efficient with large NumPy arrays. Model robustness can be observed as the performance metrics remained consistent across folds. The best hyperparameters identified for each model are summarized in Table 2 below.

Table.2. Best Hyperparameters for Model Optimization through Optuna

Regressor Models	Hyperparameters
XGBoost (XBst)	n_estimators=400, max_depth=11, learning_rate=0.1887, subsample=0.7787, colsample_bytree=0.8937

LightGBM (LGBMR)	n_estimators=700, num_leaves=130, learning_rate=0.2612, feature_fraction=0.8667, bagging_fraction=0.8862
CatBoost(CatBst)	iterations=1000, depth=8, learning_rate=0.2644, l2_leaf_reg=7.7061, border_count=206
Random Forest (RF)	n_estimators=100, max_depth=16, min_samples_split=6, min_samples_leaf=4
Gradient Boosting (GBst)	n_estimators=700, learning_rate=0.0861, max_depth=9
AdaBoost (AdaBst)	n_estimators=50, learning_rate=0.0253
Bagging Regressor (BR)	n_estimators=160, max_samples=0.9974, max_features=0.9135
KNN	n_neighbors=3, weights='distance', metric='manhattan'

F. Environmental Setup

The experiments were performed in a high-performance computing setup to support effective processing of large-scale agricultural data. The hardware configuration included an an Intel Core i7 processor and 8GB RAM. The experiments were run on Windows 11with Python 3.10, using major libraries like Scikit-learn, TensorFlow, PyTorch, Pandas, and NumPy for model training and testing.

G. Model Evaluation and Performance Analysis

After the model selection and hyperparameter optimization processes, we verifiably analyzed the predictive accuracy of the final model with a 10-fold CV strategy. In each iteration, 90% of the data was utilized for training purposes while the remaining 10% was used for validation. This method effectively managed the bias variance trade off without sacrificing computational efficiency.

Model performance was measured quantitatively through a number of key measures. The models were evaluated through Mean Squared Error (MSE) as defined in Eq.(6)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (6)$$

and its square root, the Root Mean Squared Error (RMSE) is defined as in Eq.(7)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (7)$$

Both introduced a direct measure of our predicted mean error magnitudes. Other measures such as mean absolute error (MAE) also αn insight on the average scatter of the predicted and actual values. It is defined as in Eq.(8)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)| \qquad (8)$$

In addition, the Coefficient of Determination(R^2) is calculated and defined as in Eq.(9)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \qquad (9)$$

calculated how much the model explains the variance of the target variable. In order to capture the relative error for situations that can lead to exponential growth of the dependent variable, we computed Mean Squared Log Error (MSLE). It defined as in Eq.(10)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}(\log(1 + y_i) + \log(i + \hat{y}_i)^2 \qquad (10)$$

RESULTS AND DISCUSSION

The results regarding 10 fold CV and testing of the model are shown in table 3 which include the consolidated training time, test inference time, and memory consumption.

Table 3: Summary of Performance Metrics and Computational Resource Utilization for the Evaluated Models.

Regressor	CV MSE (avg)	CV RMS	CV MAE (avg)	CV R2	Test RMSE	Test MAE	Test R2	Test	Train	Test Inference	Memory Usag

Models		E (avg)		(avg)				MSLE	Time (s)	Time (s)	e (MB)
GBst	1316277 0	3614. 445	661.5 30	0.9 98	3298. 326	626.6 36	0.9 99	0.0 26	122. 164	0.510	38.48 0
LGBM R	144360 10	3786. 217	760.0 88	0.9 98	3746.3 05	736.53 1	0.9 98	0.0 83	2.05 7	0.192	8.848
XBst	149249 50	3853. 321	614.2 55	0.9 98	3681.5 40	597.61 5	0.9 98	0.0 25	2.94 5	0.040	0.000
CatBst	165228 50	4057. 427	1432. 074	0.9 98	3950.1 78	1389.8 00	0.9 98	0.0 49	8.62 2	0.006	5.324
RF	205977 30	4526. 439	947.2 19	0.9 97	4164.6 44	889.6 81	0.9 98	0.0 07	1.79 1	0.062	32.84 0
KNN	414790 40	6417. 398	834.3 90	0.9 95	5590.1 07	733.63 9	0.9 96	0.01 2	0.18 0	0.226	21.58 2
BR	802991 90	8960. 127	6874. 625	0.9 90	8837. 208	6841.2 76	0.9 90	0.0 70	3.13 6	1.020	128.7 23
AdaBst	158888 7000	39857 .121	25175. 059	0.8 01	39969 .798	25230 .546	0.8 00	0.27 5	5.07 2	0.083	0.000

Finding the best model is a matter of trade-offs in predictive power, processing time, and interpretability. As per the results obtained, GBst is the best model as it achieved the highest Test R^2 of 0.999 and the lowest Test RMSE (3298.326) as well as Test MAE (626.636). These results show that GB generalizes the best out of all the models tested for unseen data. This shows that GB is the best and most trustworthy model that can be used for CYP since it captures the complex phenomena of agricultural data supremely well.

The RF is an exceptionally solid and graspable approach, yet is lagging in exactness and efficiency. In spite of a high-value of Test R^2, (0.998), its Test RMSE (4164.644) and MAE (889.681) are lower than XGbst and LGBMR. Even though RF is a quick trainer (1.791s), the inference speed (0.062s) is slower than XGBoost (0.040s) but it only uses little memory (32.840 MB) which makes it acceptable for many applications. The RF is the most appropriate algorithm when models are to be fitted very quickly and they come out as well and they can be tuned quite easy. However, for the most accurate and fast execution, it is better to use XGBst and LGBMR in the case of CYP in the real world. The bar plots depicting the regressor performance comparison **are** shown in Fig. 5.

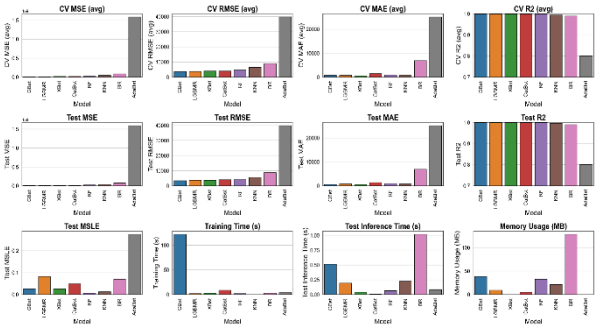

Figure 5: Regressor Models performance analysis CV & Test Set

In terms of efficiency, XBst and LGBMR present viable options. For absolute error minimization, XGBst provides the lowest MAE of 597.615 while achieving a Test R^2 of 0.998. Furthermore, LGBMR offers real-time accuracy for extremely large datasets since it balances high accuracy with the fastest training time at 2.057s. On the other hand, AdaBst is highly inefficient, with Test R^2 = 0.800 and Test RMSE = 39,969.798 which clearly demonstrates her inability in dealing with the complex feature interactions

presented in agricultural data. Furthermore, BR also performed poorly confirming the increasable errors exhibited by simple ensemble methods which are not suitable for yield forecasting.

The predicted vs. actual charts are shown in Figure 6. The most effective models are positioned near the line for actual values as predicted, while the BR and AdaBst deviate more. Residual plots showing an error distribution is presented in Figure 7. For GBst, XGBst, and LGBMR, the residuals are scattered around zero which means there is hardly any bias. AdaBst and BR, on the other hand, have distinct residuals which indicate they have higher error rates.

These results are a powerful testament to the performance of the gradient boosting models and their competency in real world yield prediction. While GBst had so much training time that it was comparatively slower (122.16 s), XGBst and LGBMR provided the best trade-off between speed, memory consumption, and precision. BR and AdaBst, on the other was too inefficient in the more complex regions of feature space, which validates the need for careful model selection. The approach we propose, based on Optuna hyperparameter tuning with strong minima validation, provides an accurate and computationally efficient, which is needed in large-scale agricultural prediction, predictive framework.

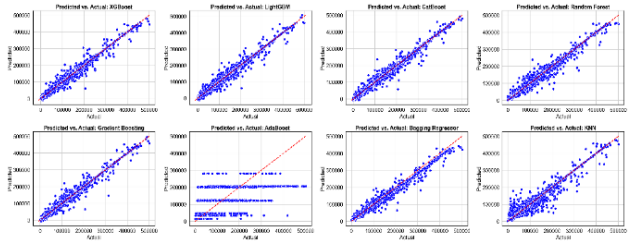

Figure 6: Model Performance Comparison – Predicted vs. Actual Values

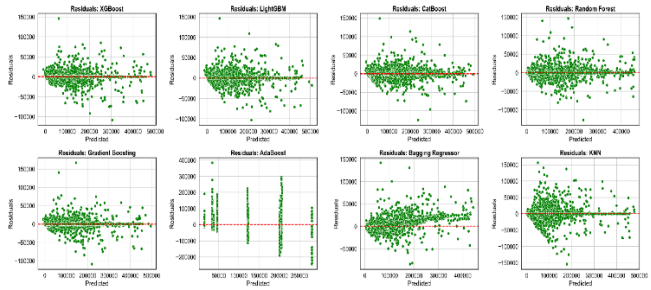

Figure 7: Residual Plots for Error Distribution in CYP

Feature Importance Analysis Using Shap

SHAP were used to identify the feature contributions to the model's predictions. Here XAI is developed with XBst model. Three visuals were used, and the SHAP summary plot (Figure 8) ranks the features based on importance, while each point represents an observation. The x-axis displays SHap values, which represent the effect on the model's output, and the color gradient characterizes feature values (lower values are on the left side of the color scale and higher values are on the right). Pesticides had the most significant influence, which is followed by temperature demonstrating a strongly non-linear effect. Rainfall exhibited a moderate influence though it was consistent and the cultivated area in fact appeared to be the most important in determining predictions. In the other hand, the least important factor was the year which indicates that it does not change much with time. The analysis of SHAP reaffirmed reliability of the model from the domain knowledge and confirmed that overfitting is not an issue that needs to be worried about. These findings show the significance of environmental factors in relationship to prediction phenomena, which emphasizes the case for the need to conduct agricultural modeling based on actual data instead of using assumptions.

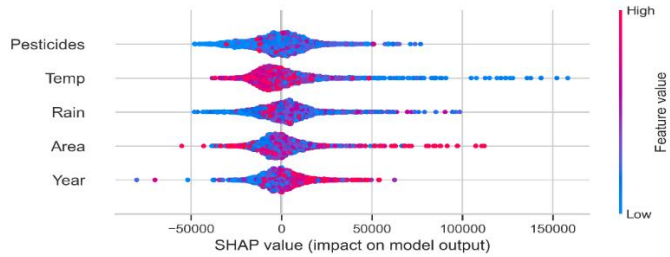

Figure 8: SHAP Summary Plot Illustrating Feature Importance and the Impact of Features on Model Predictions in CYP.

CONCLUSION AND FUTURE WORK

This work provides a thorough examination of ML models for CYP, utilizing sophisticated ensemble methods and hyperparameter tuning by Optuna. The research results showed that GBst could make the most accurate predictions by a Test R^2 of 0.999, with the smallest Test RMSE (3298.326), and the lowest Test MAE (626.636), thereby it is the best model for CYP estimation. XBst and LGBMR also demonstrated competitive performance, achieving a good balance between accuracy and efficiency, and are thus apt alternatives for practical deployment. RF provided strong and interpretable results with mediocre computational efficiency but with larger prediction errors compared to GBst, XBst, and LGBMR. On the other hand, AdaBst and BR showed much weaker predictive power, highlighting the weaknesses of dealing with sophisticated agricultural data. The paper also presents the role of hyperparameter tuning through Optuna in optimizing model performance while ensuring computational efficiency.

Future Research Directions

While the proposed framework greatly improves CYP, there are still CYP gaps that require further work. (1) Improving Spatiotemporal Generalization: The integration of satellite images and climate forecasts in the model's adaptation for various agro-ecological zones could be improved. (2) Hybrid Learning Architectures: Pytorch-based deep neural networks like transformers and LSTMs in an ensemble are good choices to improve the robustness of predictions further. (3) Real-Time Implementation: The optimization of inference pipelines for edge computing and the IoT would enable real time decision support systems to farmers. (4) Fairness and Bias Mitigation: Analyzing model biases among different crop types and regions address equitable agricultural intelligence. (5) Uncertainty Quantification: The application of probabilistic models can provide confidence intervals which would help in making decisions towards deeply sensitive agricultural structures. This multidisciplinary work will help inform agricultural policy as well as assist in crop yield forecasting models which ensures sustainability and leads to automated agriculture as well as improves broader planning.

REFRENCES

[1] Lobell, David & Field, Christopher. (2007). Global scale climate–Crop yield relationships and the impacts of recent warming. Environmental Research Letters. 2. 014002. 10.1088/1748-9326/2/1/014002.

[2] Sun, Zhilu, and Teng Fu. 2022. "The Evolutionary Trends and Convergence of Cereal Yield in Europe and Central Asia" Agriculture 12, no. 7: 1009. https://doi.org/10.3390/agriculture12071009

[3] Safdar, Muhammad & Shahid, Muhammad Adnan & Yang, Ce & Rasul, Fahd & Tahir, Muhammad & Raza, Aamir & Sabir, Rehan Mehmood. (2024). Climate Smart Agriculture and Resilience. 10.4018/979-8-3693-4864-2.ch002.

[4] Mengjia Qiao, Xiaohui He, Xijie Cheng, Panle Li, Haotian Luo, Lehan Zhang, Zhihui Tian, Crop yield prediction from multi-spectral, multi-temporal remotely sensed imagery using recurrent 3D

convolutional neural networks, International Journal of Applied Earth Observation and Geoinformation, Volume 102, 2021, 102436, ISSN 1569-8432, https://doi.org/10.1016/j.jag.2021.102436.

[5] Fengwei Guo, Pengxin Wang, Kevin Tansey, Yue Zhang, Mingqi Li, Junming Liu, Shuyu Zhang, A novel transformer-based neural network under model interpretability for improving wheat yield estimation using remotely sensed multi-variables, Computers and Electronics in Agriculture, Volume 223, 2024, 109111, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2024.109111.

[6] V, Ramesh & P, Kumaresan. (2025). Stacked ensemble model for accurate crop yield prediction using machine learning techniques. Environmental Research Communications. 7. 10.1088/2515-7620/adb9c0.

[7] Dilli Paudel, Allard de Wit, Hendrik Boogaard, Diego Marcos, Sjoukje Osinga, Ioannis N. Athanasiadis, “Interpretability of deep learning models for crop yield forecasting,” Computers and Electronics in Agriculture, Volume 206, 2023, 107663, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2023.107663.

[8] Molnar, C. (2020). Interpretable machine learning. A guide for making black box models explainable. arXiv preprint arXiv:2012.12218.

[9] G, Manju, Syam Kishor K S, and Binson V A. 2024. "An IoT-Enabled Real-Time Crop Prediction System Using Soil Fertility Analysis" Eng 5, no. 4: 2496-2510. https://doi.org/10.3390/eng5040130

[10] Lan, Yuan-Dong. (2017). A Hybrid Feature Selection Based on Mutual Information and Genetic Algorithm. Indonesian Journal of Electrical Engineering and Computer Science. 7. 214-225. 10.11591/ijeecs.v7.i1.pp214-225.

[11] Abdel-salam, M., Kumar, N. & Mahajan, S. A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning. Neural Comput & Applic 36, 20723–20750 (2024). https://doi.org/10.1007/s00521-024-10226-x

[12] Jhajharia, Dr & Mathur, Pratistha & Jain, Sanchit & Nijhawan, Sukriti. (2023). Crop Yield Prediction using Machine Learning and Deep Learning Techniques. Procedia Computer Science. 218. 406-417. 10.1016/j.procs.2023.01.023.

[13] Burdett, Hannah & Wellen, Christopher. (2022). Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. Precision Agriculture. 23. 10.1007/s11119-022-09897-0.

[14] Filippi, Patrick & Jones, Edward & Wimalathunge, Niranjan & Somarathna, Sanjeewani & Pozza, Liana & Ugbaje, Sabastine & Jephcott, Thomas & Paterson, Stacey & Whelan, B. & Bishop, Thomas. (2019). An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. Precision Agriculture. 20. 1-15. 10.1007/s11119-018-09628-4.

[15] Oliveira, J. M., & Ramos, P. (2024). Evaluating the Effectiveness of Time Series Transformers for Demand Forecasting in Retail. Mathematics, 12(17), 2728. https://doi.org/10.3390/math12172728

[16] Talaat, F.M. Crop yield prediction algorithm (CYPA) in precision agriculture based on IoT techniques and climate changes. Neural Comput & Applic 35, 17281–17292 (2023). https://doi.org/10.1007/s00521-023-08619-5

[17] Kuppan, P. & Priya, V.. (2024). Crop Yield Prediction Using Ensemble Machine Learning Techniques. SN Computer Science. 5. 10.1007/s42979-024-03536-3.

[18] Ahmed M S Kheir, Ajit Govind, Vinay Nangia, Mina Devkota, Abdelrazek Elnashar, Mohie El Din Omar and Til Feike, “Developing Automated Machine Learning Approach for Fast and Robust Crop Yield Prediction Using a Fusion of Remote Sensing, Soil, and Weather Dataset.” Environmental Research Communications, Volume 6, Number 4, DOI 10.1088/2515-7620/ad2d02

[19] Mariadass, D. A. L., Moung, E. G., Sufian, M. M., & Farzamnia, A. (2022). Extreme Gradient Boosting (XGBoost) Regressor and Shapley Additive Explanation for Crop Yield Prediction in Agriculture. In 2022 12th International Conference on Computer and Knowledge Engineering, ICCKE 2022 (pp. 219-224). Inc.. https://doi.org/10.1109/ICCKE57176.2022.9960069

[20] Bogireddy, Srinivasa Rao & Murari, Haritha. (2024). Enhancing Crop Yield Prediction through Random Forest Classifier: A Comprehensive Approach. 1663-1668. 10.1109/ICOSEC61587.2024.10722249.

[21] Sharma, Shubham & Walia, Gurleen & Singh, Kanwalpreet & Batra, Vanshika & Sekhon, Amandeep & Kumar, Aniket & Rawal, Kirti & Ghai, Deepika. (2024). Comparative Analysis on Crop Yield Forecasting using Machine Learning Techniques. Rural Sustainability Research. 52. 63-77. 10.2478/plua-2024-0015.

[22] Islam, Mohammad & Alharthi, Majed & Alkadi, Rotana & Islam, Rafiqul & Masum, Abdul. (2024). Crop yield prediction through machine learning: A path towards sustainable agriculture and climate resilience in Saudi Arabia. AIMS Agriculture and Food. 9. 980-1003. 10.3934/agrfood.2024053.

[23] Mahesh P, Soundrapandiyan R (2024) Yield prediction for crops by gradient-based algorithms. PLoS ONE 19(8): e0291928. https://doi.org/10.1371/journal.pone.0291928

[24] Yan, Yueru & Wang, Yue & Li, Jialin & Zhang, Jingwei & Mo, Xingye. (2025). Crop Yield Time-Series Data Prediction Based on Multiple Hybrid Machine Learning Models. Applied and Computational Engineering. 133. 217-223. 10.54254/2755-2721/2025.20800.

[25] Jayanthi, S., Rajkumar, K., Shaheen, Shrivastava, S., Herman, I.A. (2022). Design and Development of Framework for Big Data Based Smart Farming System. In: Saini, H.S., Sayal, R., Govardhan, A., Buyya, R. (eds) Innovations in Computer Science and Engineering. Lecture Notes in Networks and Systems, vol 385. Springer, Singapore. https://doi.org/10.1007/978-981-16-8987-1_27

[26] http://www.fao.org/home/en/

[27] https://data.worldbank.org/