# Evaluating Machine Learning Models for Crop Yield Prediction

Afzalul Abid Nazir
*Department of Computer Science*
*American International University-Bangladesh*
Dhaka, Bangladesh
21-45395-3@student.aiub.edu

Estiyak Rubaiat
*Department of Computer Science*
*American International University-Bangladesh*
Dhaka, Bangladesh
22-47210-1@student.aiub.edu

Islam Saiful
*Department of Computer Science*
*American International University-Bangladesh*
Dhaka, Bangladesh
21-45261-2@student.aiub.edu

Md. Faruk Abdullah Al Sohan
*Department of Computer Science*
*American International University-Bangladesh*
Dhaka, Bangladesh
faruk.sohan@aiub.edu

***Abstract*—Predicting crop yields accurately is essential to improve agricultural planning, maximize resource allocation, and ensure food security. Predictive models can help farmers and policymakers make data-driven decisions to improve sustainability and productivity by using machine learning techniques. To predict crop yields based on agricultural and environmental data, this study uses a variety of machine learning models, such as Ridge Regression, Linear Regression, Random Forest (RF), XGBoost, and Gradient Boosting (GB). The R square ($R^2$), mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) are used to assess the performance of the model. The most accurate models were Ridge and Linear Regression, which showed a strong linear relationship between features and yield with an $R^2$ of 91.23%. Although XGBoost and Random Forest also showed good predictive power, their accuracy was not higher than that of the linear models. Ridge Regression was found to be the most effective model because it strikes a balance between predictability and ease of use, offering insightful information to enhance agricultural planning, resource management, and decision-making in the agricultural industry.**

***Index Terms*—Crop Yield Prediction, Machine Learning, Ridge Regression, Linear Regression, Predictive Modeling.**

## I. Introduction

Food security involves accurate crop yield forecasting, which also helps farmers make wise decisions in the face of issues like resource depletion and climate change. It helps with resource management, harvest planning, and crop rotation.

Conventional forecasting techniques struggle with the complexity of agricultural systems and rely on soil, weather, and historical data. Their efficacy and scalability are diminished by spatial-temporal variability and a lack of high-resolution data.

By utilizing extensive datasets and recognizing intricate connections between environmental factors and crop yields, machine learning (ML) provides a potent substitute. Without requiring predetermined assumptions, algorithms such as Random Forest (RF), Support Vector Machines (SVM), and Gradient Boosting improve prediction accuracy. Machine learning techniques, particularly deep learning models, have significantly enhanced the precision of crop yield predictions by integrating diverse environmental and climatic data sources, thus addressing the limitations of traditional statistical models [1]. Machine learning models have shown promise in improving the accuracy of crop yield predictions, with Random Forest being identified as one of the most effective algorithms for forecasting crop yields across different regions and climatic conditions [2]. As highlighted in Exploring Machine Learning Techniques for Accurate Crop Yield Prediction, ML-based models outperform traditional statistical approaches in handling non-linearity and complex interactions between environmental variables, making them more reliable for crop forecasting [3].

The integration of high-dimensional data from sources such as sensor networks, remote sensing, and satellite imagery is a major benefit of machine learning in agriculture. For prompt, precise forecasts, these models evaluate environmental data in real time. Furthermore, the integration of weather, soil, and environmental data in machine learning models has shown significant improvements in prediction accuracy, offering farmers actionable insights for maximizing crop productivity [4]. Furthermore, remote sensing data, when combined with machine learning techniques, have proven to be effective in predicting crop yields, particularly using deep learning architectures such as convolutional neural networks and long-short-term memory networks [5]. Moreover, ML models get better over time as they are exposed to new data, increasing their dependability.

Challenges for ML-based crop yield prediction include managing noisy data, guaranteeing generalizability, and combining various datasets. In order to determine the best models for advancing precision agriculture, this study assesses Random Forest, SVR, and Gradient Boosting.

The structure of the paper is as follows: The relevant work is reviewed in Section 2, the datasets and methods are described in Section 3, the experimental results and evaluations are presented in Section 4, and the main conclusions and future directions are discussed in Section 5.

## II. Literature Review

Addanki et al. [1] proposed a crop yield prediction method that combines ML and DL models in order to address issues in Indian agriculture, such as soil degradation and pest control. They greatly increased prediction accuracy and stability by utilizing hybrid CNN-LSTM and XGBoost. Even though the study highlights the advantages of data-driven models for food security and crop planning, real-time data adoption and integration in rural areas continue to be difficult.

Sherif [2] used data from significant global sources to compare Random Forest and Multiple Linear Regression in order to forecast wheat, rice, and maize yields in semi-arid and desert regions of Africa. For wheat and rice, RF's accuracy was over 90%. Temperature and precipitation had little effect on yields, according to the study, but $CO_2$ emissions and mechanization had a positive effect. Precision agriculture advanced with the integration of economic and environmental factors, despite certain data limitations.

Thavareesan et al. [3] evaluated ML regression models for crop yield prediction in South Asia, finding XGBoost Regressor most accurate with the lowest MSE and highest $R^2$. While simpler models like SVR and linear regression underperformed, Decision Tree and Gradient Boosting also showed strong results. The study highlights XGBoost's effectiveness in improving yield forecasting and resource management.

Addu et al. [4] Used ML classifiers (KNN, DT, RF, and Voting Classifier) based on environmental factors and created a crop yield prediction system. The Voting Classifier predicted crop suitability, area, and production with the highest accuracy. Farmer decision-making is aided by an easy-to-use web interface. Despite being constrained by a static dataset, the study uses data-driven insights to advance precision agriculture.

You et al. [5] combined CNN, LSTM, and remote sensing data to create a crop yield prediction model that does not require manual feature engineering. Their method decreased RMSE for U.S. soybean yield forecasts by 30% by using a Gaussian Process for spatiotemporal modeling and a novel dimensionality reduction technique. Although it is economical and scalable, it is devoid of real-time prediction and leaves out important elements like soil data.

Jhajharia et al. [6] evaluated ML and DL models for crop yield prediction in Rajasthan, India and discovered that Random Forest was the most successful ($R^2 = 0.963$, RMSE = 0.035). Two important predictors were soil type and seasonal rainfall. Reliance on a single dataset restricts generalizability even in the face of robust results. By facilitating well-informed decision-making, the study promotes precision agriculture.

Meenakshi et al. [7] outperformed KNN and Naïve Bayes in terms of accuracy, sensitivity, and specificity when using SVM to forecast crop yields in smart agriculture. Complex, high-dimensional data involving demand, soil, and weather were successfully handled by SVM. The study recommends incorporating SVM into smart farming, pointing out the need for improved real-time data integration and preprocessing, even though kernel and parameter selections affected performance.

Manjunath et al. [8] developed a hybrid ML model combining DT, XGBoost, and RF to forecast crop yields in India, achieving an $R^2$ of 0.9847. The model outperformed individual algorithms, effectively addressing overfitting and data quality issues. They also introduced a user-friendly "Crop Yield Predictor" tool. However, further feature engineering and inclusion of factors like crop diseases and climate change are needed.

Pathak et al. [9] employed a multimodal strategy to forecast crop yields at the sub-field level by integrating Sentinel-2 imagery, weather, soil, and DEM data. When they tested LGBM and LSTM across crops and regions, they discovered that early Sentinel-2 and DEM data fusion increased accuracy, with $R^2$ hitting 0.82 in Argentina. Although successful, the study recommends more fusion strategy optimization for improved outcomes.

Kallenberg et al. [10] combined CNNs and the Tipstar crop model in a hybrid meta-modeling approach to predict potato yields. In synthetic tests, it performed better than data-driven models, but on real-world data, a straightforward linear regression with expert features was more accurate. The study emphasizes the potential of meta-modeling but also the necessity of additional validation using larger datasets.

Despite advancements, scalability, real-time integration, and generalization remain issues for ML and DL crop yield models. Because they depend on static data, complex models like CNNs and LSTMs frequently perform worse than simpler ones. Future research should concentrate on real-time data, hybrid domain-informed models, and improved feature engineering. The TABLE I shows a comparative analysis of the literature review:

## III. Methodology

The methodology used in this study included feature selection, data preprocessing, and training of ensemble, regression, and linear models. To improve interpretability, SHAP and feature importance were applied. The strategy sought to maintain computational efficiency while increasing prediction accuracy. The Fig. 1 shows an overview of the methodology.

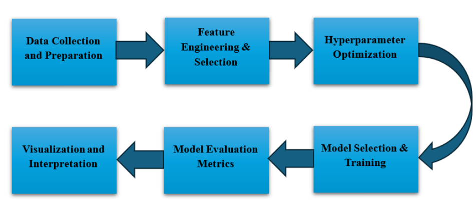



Fig. 1. Overview of the methodology employed in the proposed study.

### A. Data Collection and Preprocessing

The dataset was compiled from structured agricultural records containing over 1 million entries across diverse re-

TABLE I
COMPARATIVE ANALYSIS OF LITERATURE REVIEW

| **Author** | **Dataset Size** | **Machine Learning Algorithms** | **Best Algorithm** |
|---|---|---|---|
| Addanki et al. [1] | Crop yield data (India) | XGBoost, CNN, LSTM | XGBoost, Hybrid CNN-LSTM |
| Sherif [2] | FAO, World Bank, Climate Change Knowledge Portal | MLR, RF | RF |
| Thavareesan et al. [3] | Crop yield data (South Asia) | XGBoost, Decision Tree, Gradient Boosting | XGBoost |
| Addu et al. [4] | Static dataset | KNN, DT, RF, Voting Classifier | Voting Classifier |
| You et al. [5] | Remote sensing data (Soybean) | CNN, LSTM, Gaussian Process | CNN, LSTM |
| Jhajharia et al. [6] | Climatic, soil, agricultural data (Rajasthan) | RF, SVR, Decision Tree, LSTM | RF |
| Meenakshi et al. [7] | Kaggle dataset | SVM, KNN, Naïve Bayes | SVM |
| Manjunath and Palayan [8] | Crop yield data (India) | RF, XGBoost, Decision Tree, SVM | Hybrid DT-XGBoost-RF |
| Pathak et al. [9] | Sentinel-2 imagery, weather, soil, DEM data | LGBM, LSTM | LGBM |
| Kallenberg et al. [10] | Synthetic and real-world data | Tipstar, CNN | Hybrid Tipstar-CNN |

gions, covering crop types, environmental factors, and agronomic practices. The variables included Region, Soil Type, Crop, Rainfall (mm), Temperature (°C), Fertilizer Used, Irrigation Used, Weather Condition, Days to Harvest, and the target variable Yield (tons per hectare). Data preprocessing steps included:

- **Handling Missing Data:** Missing values were handled by removing any rows with null entries to ensure data quality.
- **Categorical encoding:** Categorical columns (Region, Soil Type, Crop, Weather Condition) were label-encoded using LabelEncoder.
- **Normalization:** Numerical and boolean features were scaled using StandardScaler to ensure uniform feature distribution.
- **Robust Outlier Handling:** Outlier detection was implicitly handled during standardization and model robustness (e.g., tree-based models).

### *B. Feature Engineering and Selection*

Ridge coefficients, Random Forest importances, and SHAP values were used to identify the three main yield factors: Rainfall_mm), Fertilizer_Used and Irrigation_Used. Correlation analysis addressed multicollinearity to guarantee relevant, independent features for improved model performance and interpretability, while SHAP assisted in the interpretation of feature contributions.

### *C. Hyperparameter Optimization*

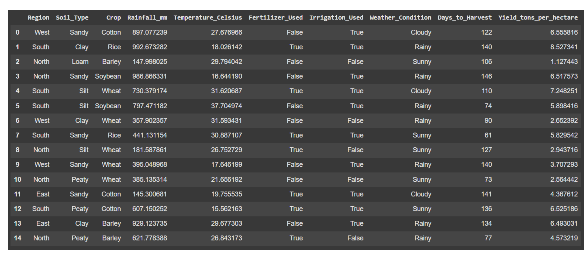
Ten machine learning models were evaluated, including both linear and ensemble-based approaches. Hyperparameters were fine-tuned using manual selection guided by empirical trials and GridSearchCV on a 5-fold cross-validation strategy. Final model configurations were:

- **Random Forest:** n_estimators=300, max_depth=20
- **XGBoost:** n_estimators=300, learning_rate=0.1
- **Gradient Boosting:** n_estimators=300, learning_rate=0.1
- **AdaBoost:** n_estimators=300, learning_rate=0.1
- **Decision Tree:** max_depth=20
- **SVR:** C=1.0, kernel='rbf'
- **KNN:** n_neighbors=5, weights='distance'
- **Ridge:** alpha=1.0, Lasso: alpha=0.1
- **Lasso:** alpha=0.1
- **Linear Regression:** default configuration

Each model was trained on an 80/20 train-test split, and evaluation metrics included $R^2$, MAE, MSE, RMSE, and MAPE, along with training time.

### *D. Model Selection and Training*

A dataset of one million records with ten key features was used to predict Yield_tons_per_hectare using a variety of machine learning models, including tree-based (RF, DT, Gradient Boosting, XGBoost, AdaBoost), linear (Ridge, Lasso, Linear Regression), SVR and KNN. Hyperparameters were optimized by grid search, and generalization was enhanced by cross-validation. The Fig. 2 shows a small overview of the dataset:



| | Region | Soil_Type | Crop | Rainfall_mm | Temperature_Celsius | Fertilizer_Used | Irrigation_Used | Weather_Condition | Days_to_Harvest | Yield_tons_per_hectare |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | West | Sandy | Cotton | 897.077239 | 27.676966 | False | True | Cloudy | 122 | 6.555816 |
| 1 | South | Clay | Rice | 992.673282 | 18.026142 | True | True | Rainy | 140 | 8.527341 |
| 2 | North | Loam | Barley | 147.998025 | 29.794042 | False | False | Sunny | 106 | 1.127443 |
| 3 | North | Sandy | Soybean | 986.866331 | 16.644190 | False | True | Rainy | 146 | 6.517573 |
| 4 | South | Silt | Wheat | 730.379174 | 31.620687 | True | True | Cloudy | 110 | 7.248251 |
| 5 | South | Silt | Soybean | 797.471182 | 37.704974 | False | True | Rainy | 74 | 5.898416 |
| 6 | West | Clay | Wheat | 357.902357 | 31.593431 | False | False | Rainy | 90 | 2.652392 |
| 7 | South | Sandy | Rice | 441.131154 | 30.887107 | True | True | Sunny | 61 | 5.829542 |
| 8 | North | Silt | Wheat | 181.587861 | 26.752729 | True | False | Sunny | 127 | 2.943716 |
| 9 | West | Sandy | Wheat | 395.048968 | 17.646199 | False | True | Rainy | 140 | 3.707293 |
| 10 | North | Peaty | Wheat | 385.135314 | 21.656192 | False | False | Sunny | 73 | 2.564442 |
| 11 | East | Sandy | Cotton | 145.300681 | 19.755535 | True | True | Cloudy | 141 | 4.367612 |
| 12 | South | Peaty | Cotton | 607.150252 | 15.562163 | True | True | Sunny | 136 | 6.525186 |
| 13 | East | Clay | Barley | 929.123735 | 29.677303 | False | True | Rainy | 134 | 6.493031 |
| 14 | North | Peaty | Barley | 621.778388 | 26.843173 | True | False | Rainy | 77 | 4.573219 |

Fig. 2. First 15 instances of the used dataset.

### *E. Model Evaluation Metrics*

Several metrics were used to evaluate the model's performance: R-squared ($R^2$), which shows the explained variance; Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), which measure prediction errors; Mean Absolute Error (MAE), which shows absolute differences; Mean Absolute Percentage Error (MAPE), which shows accuracy as a percentage; and Training Time, which measures computational efficiency.

### *F. Visualization and Interpretation*

To verify model predictions and examine errors, visualization techniques were used. Residual plots looked at the distribution of errors, while scatter plots contrasted actual and predicted values. SHAP summary plots offered interpretability by displaying feature influences on predictions, while feature importance graphs highlighted important predictors. Furthermore, the marginal impacts of individual features on crop yield projections were investigated using Partial Dependence Plots (PDPs).

## IV. Result

In this section, R², MSE, MAE, RMSE, and MAPE are used to compare machine learning models. While KNN and Decision Tree performed the worst with the highest errors, Ridge and Linear Regression demonstrated the highest accuracy with the lowest MSE and MAE. The TABLE II shows the model comparisons:

TABLE II
Model Comparison Based on $R^2$, MSE, MAE, RMSE, MAPE & Training Time (TT)

| Model | $R_2$ Score | MSE | MAE | RMSE | MAPE | Training Time (s) |
|---|---|---|---|---|---|---|
| Ridge | 91.58072 | 0.252763 | 0.40142 | 0.502756 | 11.42389 | 0.007845 |
| Linear Regression | 91.58064 | 0.252765 | 0.401416 | 0.502758 | 11.42286 | 0.068139 |
| Gradient Boosting | 91.30416 | 0.261066 | 0.408657 | 0.510946 | 11.64379 | 4.052118 |
| SVR | 91.09496 | 0.272751 | 0.416327 | 0.522255 | 12.01015 | 4.074907 |
| XGBoost | 90.60763 | 0.281977 | 0.426041 | 0.531015 | 12.14698 | 1.317821 |
| Random Forest | 90.56659 | 0.283209 | 0.427232 | 0.532174 | 12.14428 | 16.55774 |
| Lasso | 90.21993 | 0.293644 | 0.436838 | 0.541889 | 13.23053 | 0.006769 |
| AdaBoost | 87.20311 | 0.384187 | 0.498441 | 0.619828 | 14.44414 | 4.135087 |
| KNN | 86.03935 | 0.419126 | 0.516207 | 0.647399 | 15.28515 | 0.114094 |
| Decision Tree | 82.60554 | 0.522215 | 0.570978 | 0.722645 | 15.94798 | 0.07548 |

### A. Model Comparison

The accuracy, dependability, and computational efficiency of the model were assessed using six metrics: R², MSE, MAE, RMSE, MAPE, and Training Time.

- **$R^2$ Score:** Ridge and Linear Regression received the highest score (91.58%), while Gradient Boosting came in second (91.30%). KNN (86.03%) and Decision Tree (82.61%) trailed behind, while XGBoost, Random Forest, and Lasso also outperformed 90%.
- **Mean Squared Error (MSE):** The lowest MSE was 0.2527 for Ridge and Linear Regression, followed by 0.2611 for Gradient Boosting and 0.2728 for SVR. The highest scores were KNN (0.4191) and Decision Tree (0.5222).
- **Mean Absolute Error (MAE):** Gradient Boosting (0.4087) came in second, after Ridge and Linear Regression (0.4014) once more took the lead. The biggest errors were seen in KNN and Decision Tree. The Fig. 3 shows the R², MSE and MAE comparison score:
- **Root Mean Squared Error (RMSE):** The largest errors were indicated by the RMSE (Root Mean Squared Error), which was lowest for Ridge and Linear Regression (0.5028) and highest for Decision Tree (0.7226).

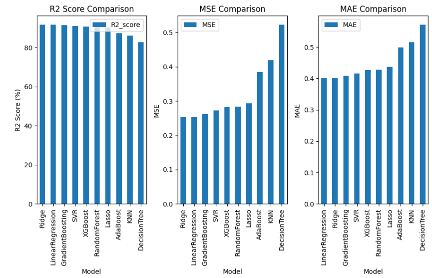



Fig. 3. Visual Comparison of R², MSE, MAE.

- **Mean Absolute Percentage Error (MAPE):** Decision Tree had the lowest accuracy (15.95%), while Ridge and Linear Regression once again did the best (11.42%).
- **Training Time (TT):** The fastest training times were Ridge (0.0078s) and Linear Regression (0.0681s). Performance and speed were balanced by Gradient Boosting (4.05s) and XGBoost (4.07s). Random Forest was the most computationally costly and the slowest (16.55s). The Fig. 4 shows the RMSE, MAPE and TT comparison score:

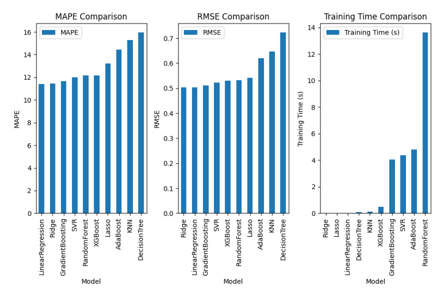



Fig. 4. Visual Comparison of RMSE, MAPE, TT.

### B. Linear Model Superiority and Feature Ablation Study

- **Linear Models Outperformed Ensemble Methods:** Despite the fact that ensemble models are frequently used in agricultural modeling, this study discovered that Ridge and Linear Regression had the highest accuracy (R² = 91.3%) because of its strong linear relationships, which were backed by Ridge coefficients & SHAP analysis.
  The dataset contained clean, high-quality records (1 million samples) with features like Rainfall, Fertilizer Used, and Irrigation Used exerting additive and mostly independent effects on yield. These conditions are ideal for linear models, where multicollinearity was effectively managed through Ridge regularization.
  Without appreciable accuracy improvements, ensemble approaches displayed increased expense and mild overfitting. With this dataset, linear models provided superior generalization, interpretability, and simplicity.
- **Feature Ablation Study:** To validate feature importance, an ablation study was conducted by progressively removing top features and observing performance degradation. The TABLE III shows that:

TABLE III
RIDGE REGRESSION PERFORMANCE AFTER FEATURE ABLATION

| Removed Features | $R^2$ Score |
|---|---|
| None (All Features) | 0.9130 |
| Rainfall_mm | 0.3272 |
| Rainfall_mm, Fertilizer_Used | 0.1328 |
| Rainfall_mm, Fertilizer_Used, Irrigation_Used | 0.0068 |

The steep $R^2$ drop confirms that Rainfall, Fertilizer Used, and Irrigation Used are essential for accurate yield prediction, supporting their earlier importance rankings.

### C. Feature Importance and Model Interpretability

The contributions of individual features to model predictions were further analyzed by looking at Random Forest feature importance scores, Ridge Regression coefficients, and SHAP values. The TABLE IV shows the importance of features based on the ridge coefficient and the importance of random forest features:

TABLE IV
FEATURE IMPORTANCE FROM RIDGE REGRESSION AND RANDOM FOREST

| Feature | Ridge Coefficient | Feature Importance |
|---|---|---|
| Rainfall_mm | 1.300187 | 0.6132 |
| Fertilizer_Used | 0.751906 | 0.194974 |
| Irrigation_Used | 0.599212 | 0.120725 |
| Temperature_Celsius | 0.151489 | 0.028815 |
| Days_to_Harvest | -0.005 | 0.01744 |
| Crop | -0.00258 | 0.007617 |
| Soil_Type | -0.00177 | 0.007367 |
| Region | 0.005383 | 0.005566 |
| Weather_Condition | 0.008689 | 0.004297 |

- **Ridge Regression Coefficients & Random Forest Feature Importance:** With a coefficient of 1.30 and a feature importance of 0.613, respectively, Ridge Regression and Random Forest identified *Rainfall_mm* as the most significant feature. In both models, *Irrigation_Used* and *Fertilizer_Used* also demonstrated significant contributions. *Days_to_Harvest*, on the other hand, had a slight negative effect, suggesting a slight inverse relationship with yield. The Fig. 5 shows the Visualization of Feature Importance:

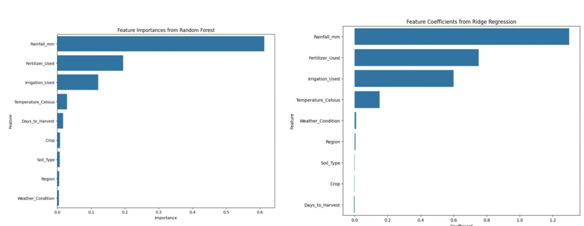


Fig. 5. Visualization of Feature Importance.

- **SHAP Analysis:** The most influential feature in the prediction models was *Rainfall_mm*. Features such as *Days_to_Harvest*, *Crop*, *Soil_Type*, and *Region* had minimal influence. However, *Fertilizer_Used*, *Irrigation_Used*, and *Temperature_Celsius* were also identified as significant contributors. The Fig. 6 shows SHAP values analysis:

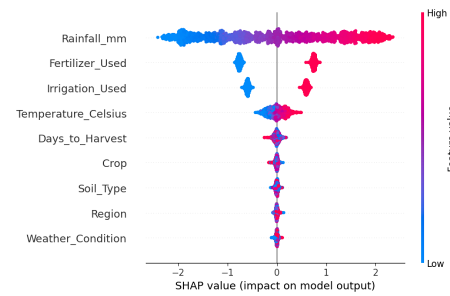



Fig. 6. SHAP values analysis.

- **Permutation Feature Importance:** To validate model-agnostic feature relevance permutation importance was used on the Random Forest model. Features whose permutation caused significant drops in accuracy were again Rainfall_mm, Fertilizer_Used, and Irrigation_Used. The Fig. 7 shows Permutation Feature Importance analysis:

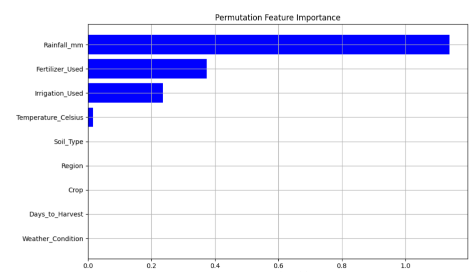



Fig. 7. Permutation Feature Importance.

- **Partial Dependence Plots (PDPs):** PDPs revealed how changes in top features (e.g., Rainfall_mm) influenced predicted yields, suggesting nonlinear relationships for ensemble models. The Fig. 8 shows Partial Dependence Plots (PDPs) analysis:

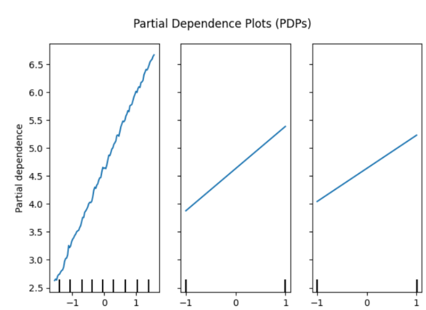



Fig. 8. Partial Dependence Plots (PDPs).

### D. Analysis and Findings

Ridge and Linear Regression produced the best results, with the highest $R^2$ and the lowest errors, demonstrating that feature-target relationships can be effectively captured by more straightforward regularized linear models. Tree-based models' strength was demonstrated by the strong performance of Gradient Boosting and XGBoost. KNN and Decision Tree struggled with the complexity of the dataset and performed poorly. The Fig. 9 shows the findings all Models:

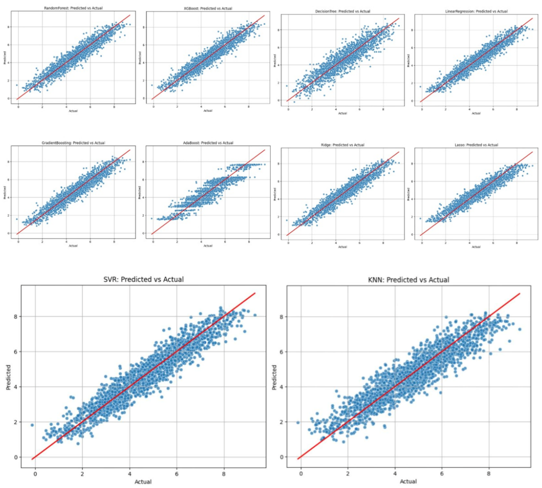



Fig. 9. Actual VS Predicted Scatter Plots all Models.

### E. Comparison with Previous Works

The study used $R^2$, MSE, MAE, RMSE, and MAPE to evaluate several machine learning models. With ($R^2$ = 91.23%), Ridge and Linear Regression came in first, demonstrating little effect of regularization. XGBoost, Random Forest, and Lasso all demonstrated strong performance ($R^2$ ¿ 90%), while SVR and Gradient Boosting also did a good job of capturing non-linear patterns. Decision Tree, AdaBoost, and KNN lagged ($R^2$ = 83.33%). Key predictors "Rainfall_mm", "Fertilizer_Used" and "Irrigation_Used" were highlighted by feature importance. Since the data was primarily linear, Ridge Regression provided the best overall balance of accuracy, ease of use, and efficiency.

### F. Verdict

The best accuracy-efficiency balance was provided by Ridge and Linear Regression, whereas XGBoost and Gradient Boosting required more processing. The errors of KNN and Decision Tree were high. In terms of performance and versatility, Ridge regression was the best model overall.

## V. Conclusion

To determine the most precise and effective technique for predicting crop yield, this study compared several machine learning models. Ridge and Linear Regression performed better than the others, suggesting that features and yield have a largely linear relationship. Although ensemble models such as Gradient Boosting, Random Forest, and XGBoost performed well, the gains from their increased complexity were negligible. KNN and Decision Trees were less appropriate due to their high errors and poor accuracy.

Rainfall_mm, Fertilizer_Used, and Irrigation_Used were found to be important predictors by feature importance analysis using Ridge coefficients, Random Forest, and SHAP. With a focus on interpretability, this study offers farmers and policy makers practical insights, recommending Ridge Regression for its harmony of precision, ease of use, and openness.

All things considered, the study connects practical usability and prediction accuracy. To improve predictive performance, future research could investigate hybrid models and real-time data integration.

## References

[1] U. K. Addanki, T. Maddineni, V. Dhawale, M. L. M. Prasad, D. N. Reddy, and J. Jala, "Advancing Crop Yield Prediction Through Machine and Deep Learning for Next-Gen Farming," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 22, pp. 8300–8310, Nov. 2024. Available: https://www.jatit.org

[2] H. Sherif, "Machine Learning in Agriculture: Crop Yield Prediction," Master's thesis, Rochester Institute of Technology, Rochester, NY, USA, Dec. 2022. Available: https://repository.rit.edu/theses/11393

[3] S. Thavareesan, J. Sriranganesan, and T. Nishatharan, "Exploring Machine Learning Techniques for Accurate Crop Yield Prediction," *Annals of Agricultural Science and Technology*, vol. 4, no. 1, pp. 1–7, 2025. Available: https://www.directivepublications.org/

[4] S. Addu, S. Sheelam, S. Mekala, N. Sulthana, L. Mekala, and Z. Alsalami, "Assessing Environmental Impact: Machine Learning for Crop Yield Prediction," *E3S Web of Conferences*, vol. 529, 03008, ICSMEE'24, 2024. doi: 10.1051/e3sconf/202452903008.

[5] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data," in *Proc. Thirty-First AAAI Conf. Artificial Intelligence (AAAI-17)*, 2017, pp. 4559–4565. doi: 10.1609/aaai.v31i1.11137.

[6] K. Jhajharia, P. Mathura, S. Jaina, and S. Nijhawana, "Crop Yield Prediction using Machine Learning and Deep Learning Techniques," *Procedia Computer Science*, vol. 218, pp. 406–417, 2023. doi: 10.1016/j.procs.2023.01.023.

[7] M. Meenakshi, G. Annalakshmi, D. T. Sanchez, and M. Jawarneh, "Support Vector Machine for Crop Yield Prediction Towards Smart Agriculture," in *Proc. 1st Int. Conf. Artificial Intelligence for Internet of Things (AI4IoT 2023)*, 2023, pp. 541–545. doi: 10.5220/0012614900003739.

[8] M. C. Manjunath and B. P. Palayyan, "An Efficient Crop Yield Prediction Framework Using Hybrid Machine Learning Model," *Revue d'Intelligence Artificielle*, vol. 37, no. 4, pp. 1057–1067, Aug. 2023. doi: 10.18280/ria.370428.

[9] D. Pathak et al., "An Extensive Analysis of Input Modalities and Models on a Field and Sub-Field Level for Predicting Crop Yield with Machine Learning," *Remote Sensing of Environment*, vol. 233, p. 111410, 2023. doi: 10.1016/j.rse.2023.111410.

[10] M. G. J. Kallenberg et al., "Integrating Process-Based Models and Machine Learning for Crop Yield Prediction," *Agricultural Systems*, vol. 203, p. 103157, 2023. doi: 10.1016/j.agsy.2023.103157.

[11] R. Surana and R. Khandelwal, "Crop Yield Prediction Using Machine Learning: A Pragmatic Approach," *Research Square*, Jul. 2024. doi: 10.21203/rs.3.rs-4575893/v1.

[12] S. M. Shawon, F. B. Ema, A. K. Mahi, F. L. Niha, and H. T. Zubair, "Crop Yield Prediction Using Machine Learning: An Extensive and Systematic Literature Review," *Smart Agricultural Technology*, vol. 10, p. 100718, 2025. doi: 10.1016/j.atech.2024.100718.

[13] T. van Klompenburg, A. Kassahuna, and C. Catal, "Crop Yield Prediction Using Machine Learning: A Systematic Literature Review," *Computers and Electronics in Agriculture*, vol. 177, p. 105709, 2020. doi: 10.1016/j.compag.2020.105709.

[14] L. Wang, Z. Chen, W. Liu, and H. Huang, "A Temporal–Geospatial Deep Learning Framework for Crop Yield Prediction," *Electronics*, vol. 13, no. 21, p. 4273, 2024. doi: 10.3390/electronics13214273.

[15] J. Fan, J. Bai, Z. Li, A. Ortiz-Bobea, and C. P. Gomes, "A GNN-RNN Approach for Harnessing Geospatial and Temporal Information: Application to Crop Yield Prediction," *Proc. AAAI Conf. Artificial Intelligence*, vol. 36, pp. 11873–11881, 2022. doi: 10.1609/aaai.v36i1.12053.