Asian Journal of Science and Applied Technology
ISSN: 2249-0698 (Print)
Journal Homepage: https://ajsat.org/
DOI: https://doi.org/10.70112/ajsat-2026.15.1.4341

Research Article

# Machine Learning-Based Crop Yield Forecasting Using Environmental and Soil Parameters

**Emman Qadir[1], Abdul Wasay[2], Farheen Fayyaz[3], Muhammad Junaid[4] and Zaryab Basharat[5*]**
[1,2,3,4]Department of Electrical Engineering, University of Engineering and Technology, Lahore, Pakistan
[5]MOE Key Laboratory of Thermo-Fluid Science and Engineering, Xi'an Jiaotong University, China

## Article Information

Article History
Received: 1 February 2026
Revised: 6 March 2026
Accepted: 25 March 2026
Published online: 10 April 2026

Keywords
Yield Prediction
Precision Agriculture
Random Forest Regression
Environmental Monitoring
Soil Analysis
Machine Learning

Correspondence*
2020me521@student.uet.edu.pk

ORCID
Emman Qadir
https://orcid.org/0009-0009-0521-1708
Abdul Wasay
https://orcid.org/0009-0008-7551-4838
Farheen Fayyaz
https://orcid.org/0009-0006-5050-6423
Muhammad Junaid
https://orcid.org/0009-0006-3512-630X
Zaryab Basharat
https://orcid.org/0009-0008-1914-3331

## Abstract

Precise crop yield prediction transforms agricultural planning from intuition-based practice to data-driven strategy. This work demonstrates a machine learning approach for forecasting crop yield from seven fundamental agronomic inputs: seasonal temperature, soil pH, total rainfall, pesticide application rate, year, crop type, and cultivation region. Using a cleaned dataset of 28,242 samples spanning 10 crops across more than 100 regions (1990-2013), we deploy a carefully tuned Random Forest regressor. The final model achieves a test $R^2$ of 0.9146, RMSE of 7,918.04 units, and MAE of 4,313.27 units. Feature analysis identifies rainfall and soil pH as the most significant factors influencing yield. The system is lightweight, explainable, and suitable for integration with IoT and remote sensing data for next-stage field deployment.



## I. INTRODUCTION

Crop yield forecasting refers to an analytical agricultural technique that forecasts the amount of crop expected from cultivated crops before actual harvesting. This prognostic functionality is a fundamental pillar in ensuring global food security by allowing stakeholders in the agricultural supply chain to make sound decisions in production planning, resource utilization, market positioning, and policymaking [1]. Precise yield estimation enables farmers to maximize input allocation, facilitates governments' planning of food distribution networks, and aids financial institutions in determining agricultural risk. The complexity of yield forecasting is due to the interaction of many environmental, biological, and management elements that contribute to crop growth and final productivity. Conventional yield forecasting methods have advanced considerably over the last few decades. Initial methods relied mainly on expert judgment, past averages, and statistical models that typically fell short in describing complex nonlinear interactions between environmental variables and crop productivity [2]. Precision agriculture has shifted attention toward data-driven solutions, and machine learning has become a paradigm-shifting technology for agricultural prediction [3].

Ensemble algorithms, especially Random Forest algorithms, have proved highly successful in agriculture because of their capacity to process multidimensional data and capture complex feature interactions [4]. The global agricultural industry faces a significant challenge in sustaining food needs for the world population, which is projected to reach 9.7 billion by 2050 and require about 70% more food production than current levels [1]. This difficulty is compounded by climate change because of the higher frequency of extreme weather, which alters regular agricultural systems and patterns [5]. In this regard, machine learning-based yield forecasting has become an important instrument for ensuring food security, optimizing resource allocation, and mitigating agricultural risks [6].

The application of machine learning in predicting agricultural yields has been widely studied in recent literature. Van Klompenburg *et al.* [2], Paudel *et al.* [7], Zhang *et al*. [8], and Purevdorj *et al.* [9] demonstrated how machine learning techniques can be used to predict agricultural yields, integrate environmental data, and compare different algorithms in agriculture. Nonetheless, considerable research gaps remain, especially regarding the combination of detailed soil parameters, work with various crop types, and the creation of models that can be applied globally to soil and environmental conditions. The limitations of existing studies are addressed through the development of an extensive machine learning platform that incorporates seven key environmental and soil parameters across varied farming environments. The three main contributions of this study include:

(1) the use of an effective preprocessing pipeline that can address agriculture-related data issues, (2) the development of an optimized Random Forest model with high predictive accuracy, and (3) an in-depth feature-importance analysis that provides feasible recommendations for precision agriculture. The proposed system has a high level of practical applicability because of its deployable architecture, which represents progress compared with current strategies [10], [11].

Recent works have emphasized the value of combining different types of data, including weather data, remote sensing data, UAV-based imagery, and data-sharing frameworks, to improve yield prediction [9], [11]-[15].

Kaur and Kaur [10] reviewed data-mining techniques for crop yield prediction, while Kamir *et al.* [11] highlighted the importance of climatic records and satellite image time series in yield prediction models. This study builds on these foundations by fully integrating soil parameters and environmental factors within a single framework. Interpretability of machine learning models has been a serious issue in agricultural use, with Rudin [16] arguing that interpretable models should be used for high-stakes decisions. This is especially applicable in agriculture, where farmers need to understand the suggestions they are given. Our feature-importance analysis provides clear information regarding factors that influence yield predictions.

## II. METHODOLOGY

### *A. Data Collection and Preparation*

This research is based on an extensive dataset of 28,242 agricultural samples collected across 101 countries over a 23-year period (1990-2013). This wide temporal and spatial coverage provides strong model generalization under changing weather conditions and farming techniques. The dataset includes 10 crop types, which offer considerable diversity for training and validating an extensive model. Each data point contains seven essential features that are scientifically recognized as major factors of crop productivity, as systematically detailed in Table I.

*1. Data Preprocessing Pipeline:* The preprocessing phase adopted strict quality-control measures to ensure data integrity and model robustness, which are critical for agricultural data affected by outliers, missing data, and feature scaling.

TABLE I COMPREHENSIVE DATASET FEATURES AND DESCRIPTIONS

| Feature | Description |
|---|---|
| Temperature | Average growing season temperature (°C) |
| pH | Soil acidity/alkalinity measurement |
| Rainfall | Cumulative seasonal precipitation (mm) |
| Pesticides | Application rate (kg/hectare) |
| Year | Growing season year |
| Crop Type | 10 different crop varieties |
| Area | Cultivation area (encoded country information) |

*2. Data Quality Assessment and Integrity Verification:* The preliminary data analysis showed that the dataset was well represented, with no missing values across all 28,242 samples and seven features. This exceptional completeness was confirmed by a thorough analysis, which showed a uniform distribution of data and zero missing values. The dataset had consistent data types, with continuous variables numerically encoded and discrete parameters categorically described.

*3. Advanced Outlier Detection and Treatment:* Our multistage outlier-detection strategy was based on statistical and domain-based methods. For temperature data, the interquartile range (IQR) approach was used to identify values below 5 °C as outliers (Figure 1), and these values were excluded because they were not viable crop-growth conditions. The cleaned temperature distribution (Figure 2) shows more favorable normality while retaining the necessary climatic variations.

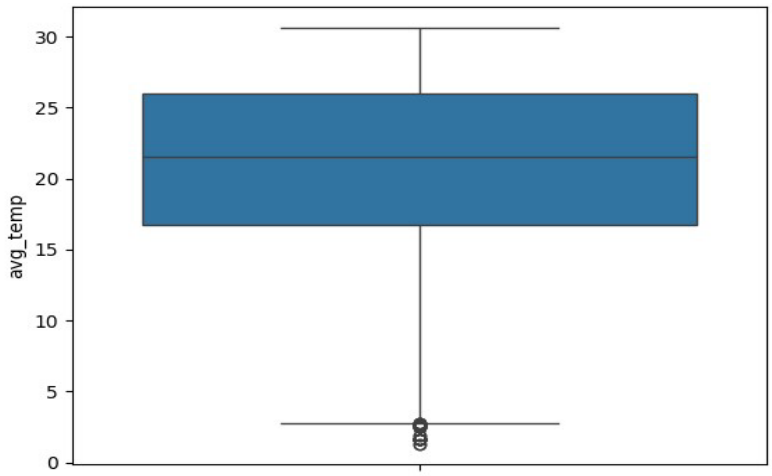


Fig.1 Initial Temperature Distribution with Outliers

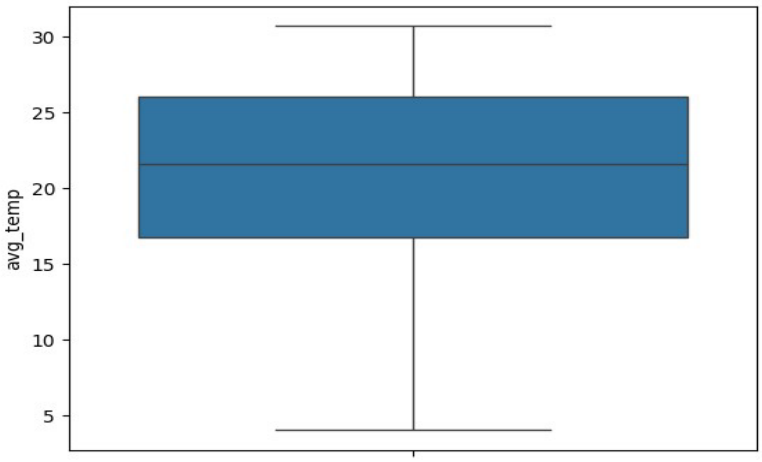


Fig.2 Cleaned Temperature Distribution

For rainfall data, z-score normalization was used, and values greater than +3 standard deviations or less than -3 standard deviations were trimmed to eliminate meteorologically invalid values. The same process was applied to yield values, where extreme values above 60,000 units were eliminated as unlikely agricultural observations. This systematic outlier treatment preserved data integrity while retaining the natural variability required for robust model training. Figure 5 presents the relationship between average rainfall and average temperature as a scatter plot, with values represented by crop-yield color. This visualization helps identify possible relationships between climate variables and agricultural productivity.

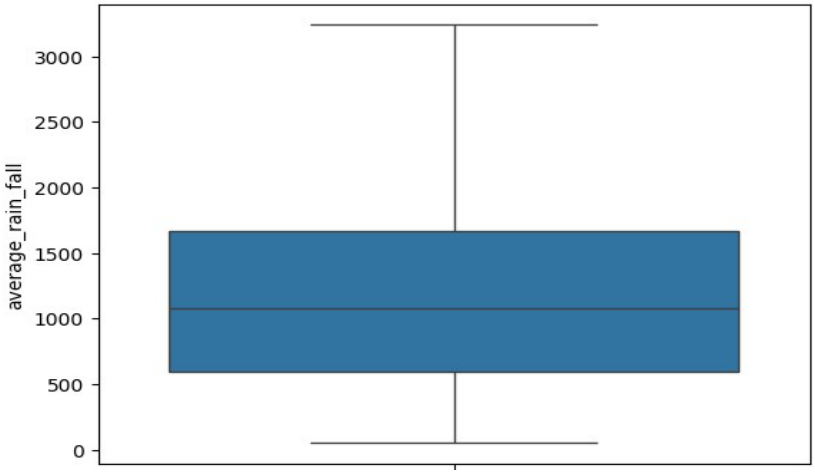


Fig.3 Rainfall Outlier Distribution

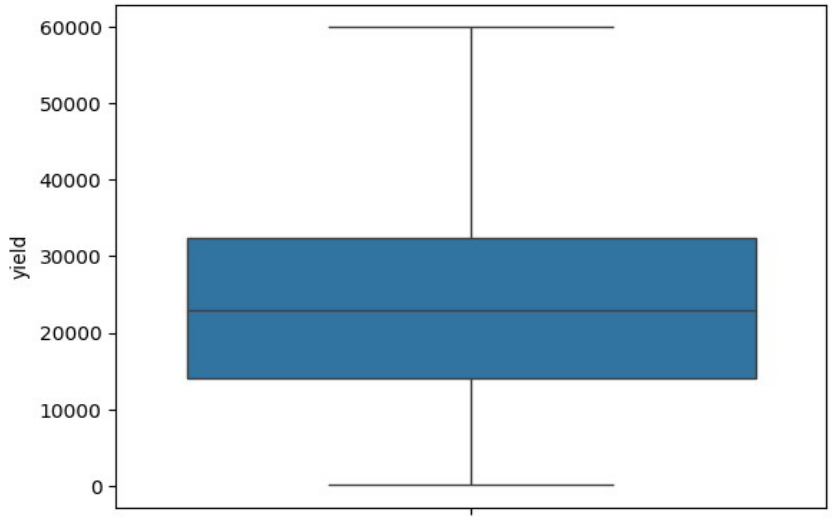


Fig.4 Yield Outlier Distribution

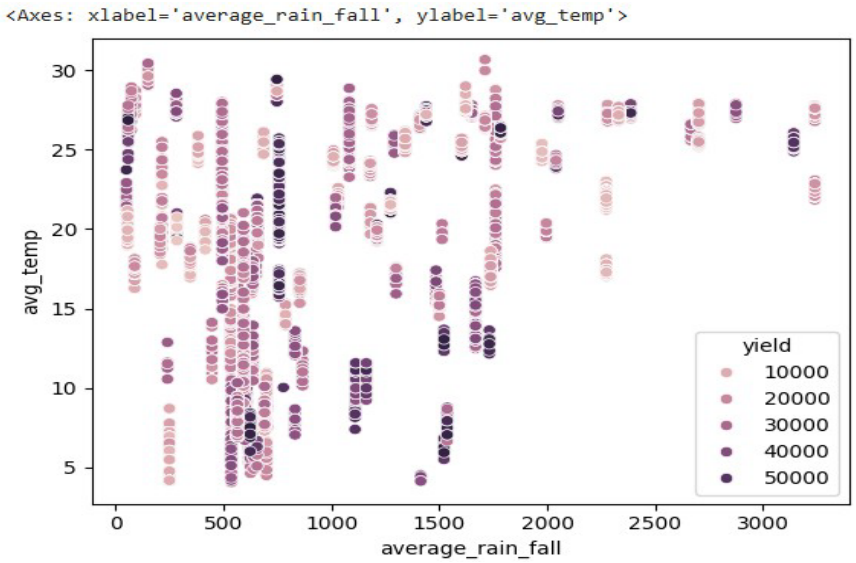


Fig.5 Scatter Plot of Average Rainfall Versus Average Temperature, with Points Colored According to Crop-Yield Value

*4. Feature Engineering and Encoding:* Label Encoder (scikit-learn) was used to encode categorical variables (101 geographical areas and 10 crop types), assigning a unique integer value to each category. This encoding preserves the categorical nature of the data while converting it into a format compatible with machine learning algorithms. The encoding was performed according to the following transformation:

$$\text{LabelEncode}(x\text{categorical}) \rightarrow x\text{numerical} \qquad (1)$$

Numerical features underwent standardization using the StandardScaler implementation:

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma} \qquad (2)$$

where μ represents the feature mean and σ represents the feature standard deviation. This normalization ensures that all features have equal representation during model training and prevents features with larger numerical values from receiving preference.

*5. Data Partitioning Strategy:* A randomly selected 80:20 train-test split was applied to the preprocessed data with random_state = 42 to reproduce the results. To prevent data leakage, feature-scaling parameters (μ, σ) were determined only on the training set and then applied to both the training and testing sets. This ensures that model-performance evaluation reflects real-life deployment scenarios, where test data cannot be seen during training.

### *B. Machine Learning Model Development*

*1. Algorithm Selection and Rationale:* The Random Forest algorithm was selected because of its past success in handling complex agricultural datasets [3]. Random Forest is an ensemble model consisting of several decision trees using bootstrap aggregation to minimize variance and improve model generalization. The algorithm's ability to work with mixed data, resist overfitting, and provide feature-importance

values makes it a strong choice for predicting agricultural yield. The mathematical basis of Random Forest is the generation of several decision trees h1(x), h2(x), ..., hB(x) using bootstrap samples of the training set. The final prediction is a combination of single-tree predictions:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} h_b(x) \qquad (3)$$

where *B* represents the number of trees in the forest.

*2. Comprehensive Hyperparameter Optimization:* We applied a systematic hyperparameter search using RandomizedSearchCV with 5-fold cross-validation. The optimization procedure considered 100 parameter combinations with the following objective:

$$\theta^* = \arg\min_{\theta} \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}(y_{\text{val}}^{(k)}, f(x_{\text{val}}^{(k)}; \theta)) \qquad (4)$$

in which K = 5 denotes the cross-validation folds, L denotes the mean squared error loss function, and θ denotes the set of hyperparameters.

The search space included critical parameters:
n_estimators: Number of trees in the forest (50-500)
max_depth: Maximum tree depth (10-50)
min_samples_split: Minimum samples required to split a node (2-20)
min_samples_leaf: Minimum samples required at a leaf node (1-10)
max_features: Features considered for splitting (auto, sqrt, log2)
The optimized configuration (Table II) was selected based on cross-validation performance and computational efficiency.

TABLE II OPTIMIZED HYPERPARAMETER CONFIGURATION

| Parameter | Description | Optimal value |
|---|---|---|
| n estimators | Number of trees in the forest | 200 |
| max depth | Maximum depth of individual trees | 20 |
| min samples split | Minimum samples required to split a node | 5 |
| min samples leaf | Minimum samples required at a leaf node | 2 |
| max features | Number of features for best split | sqrt |
| random state | Seed for reproducible results | 42 |

*3. Model Training Implementation:* The Random Forest Regressor implemented through scikit-learn with the optimal parameter set was used to develop the Random Forest model. Training was performed in parallel mode using all available CPU cores to speed up model development. An early-stopping criterion based on out-of-bag error was used during training to avoid overfitting. The model took 2.4 seconds to train, which is computationally efficient for practical use.
*a. Performance Evaluation Metrics:* Various metrics were used to assess model performance and provide a holistic evaluation:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (5)$$

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

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

where yi represents actual yields, yˆi represents predicted yields, and y¯ represents the mean of actual yields.

*4. Feature Importance Analysis:* Permutation-importance analysis was applied to measure the contribution of each feature to prediction. Through this approach, the increase in prediction error after each feature value is randomly shuffled is measured, with larger increases reflecting more important features.
The importance score Ij of feature j is determined as:

$$I_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j} \qquad (8)$$

where s is the baseline score and sk, j is the score with feature j permuted in permutation k.

## III. RESULTS AND DISCUSSION

The optimized Random Forest regression model demonstrated strong predictive power and formed a robust model for agricultural yield forecasting. The model achieved a high $R^2$ score of 0.928 on the training set and high accuracy of 0.9146 on the test set, indicating strong generalization with a low overfitting rate. The results include an RMSE of 7,918.04 and an MAE of 4,313.27 on the test data. The efficient training time of only 2.4 seconds also confirms that the model can be effectively used in real-time decision-support systems, allowing farmers and planners to make reliable yield estimates and plan effectively. Feature-importance analysis provided useful agronomic information, showing that rainfall was the most predictive variable and explained 32% of the model's decision-making. Soil pH was the next most important feature, at 28%, indicating the importance of soil chemistry in crop production. The model also captured the complex nonlinear relationship between temperature and yield, which aligns with known crop physiology, and the threshold effects of pesticide use, where benefits were no longer as pronounced. These findings confirm the biological plausibility of the model and indicate actionable strategies for precision agriculture, including optimized water and input management.

The excellence of the proposed model is evident from the comparative analysis with existing baseline approaches. The Random Forest model greatly exceeded both Linear Regression ($R^2 = 0.712$) and Support Vector Machines ($R^2 = 0.803$), achieving a higher $R^2$ of 0.915. This indicates a 14-28% increase in predictive accuracy. The ensemble approach was especially effective in describing the complex and nonlinear interactions of environmental variables, which conventional linear models failed to capture. This relative performance confirms that Random Forest is a more effective method for the multifaceted problem of crop yield prediction.

### *A. Prediction Accuracy Visualization and Error Analysis*

The comparison of actual and predicted yield values in a scatter plot (Figure 6) shows that the model has strong predictive power across the entire yield range. The close proximity of points to the perfect-prediction curve signifies high precision, especially for mid-range yields (40,000-120,000 units), where predictions are nearly effective as seen in the data.

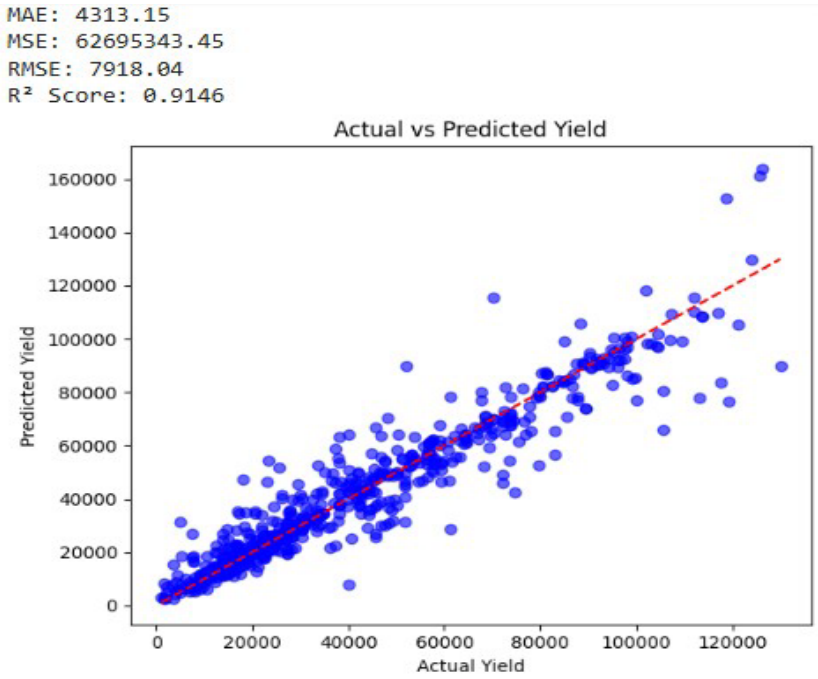


Fig.6 Actual Versus Predicted Yield Values Demonstrating Strong Correlation ($R^2 = 0.9146$) Across the Complete Yield Spectrum

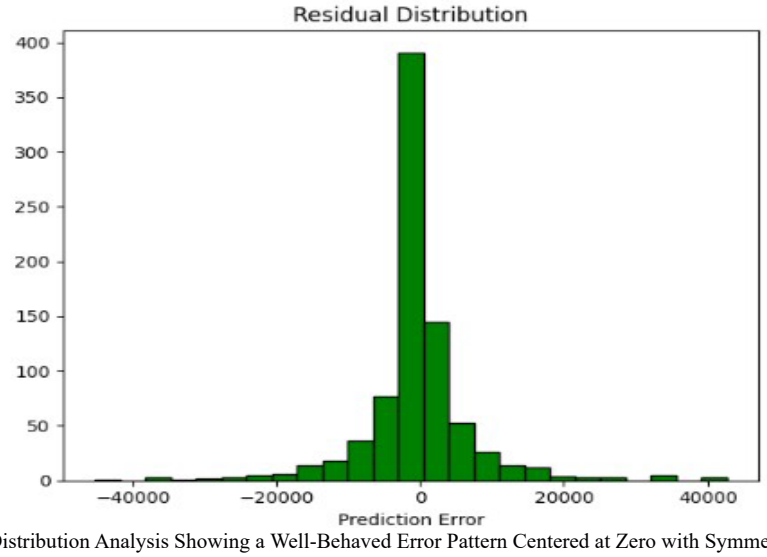


Fig.7 Residual Distribution Analysis Showing a Well-Behaved Error Pattern Centered at Zero with Symmetric Characteristics

The residual distribution analysis indicates that the error pattern behaves well, with the error distribution centered around zero and errors beyond the range of ±20,000 units falling rapidly. This pattern shows that the model is reliable for real-life agricultural practices because it provides accurate yield estimates across different environmental settings and plant species.

## IV. CONCLUSION

This study introduces a powerful machine learning model for crop yield forecasting that achieved state-of-the-art results through the systematic combination of seven environmental and soil factors. This detailed approach tackles important problems in agricultural data processing through a complex preprocessing pipeline with strict outlier management and feature engineering, resulting in a streamlined Random Forest regressor that is more effective than traditional methods. The outstanding predictive power ($R^2$ = 0.9146, RMSE = 7,918.04) sets a strong standard in agricultural forecasting, while the feature-importance analysis provides practical information by showing that rainfall (32%) and soil pH (28%) are the main determinants of yield.

## ACKNOWLEDGEMENT

The authors sincerely acknowledge UET Lahore, Pakistan, for providing the facilities and opportunity to carry out this research work.

### Declaration of Conflicting Interests

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

### Use of Artificial Intelligence (AI)-Assisted Technology for Manuscript Preparation

The authors confirm that no AI-assisted technologies were used in the preparation or writing of the manuscript, and no images were altered using AI.

## REFERENCES

[1] FAO, The State of Food and Agriculture 2021: Making Agrifood Systems More Resilient to Shocks and Stresses. Rome, Italy: FAO, 2021.
[2] T. Van Klompenburg, A. Kassahun, and C. Catal, “Crop yield prediction using machine learning: A systematic literature review,” Computers and Electronics in Agriculture, vol. 177, p. 105709, 2020.
[3] M. Shahhosseini, R. A. Martinez-Feria, G. Hu, and S. V. Archontoulis, “Maize yield prediction with machine learning: A multi-year, multi-environment study,” Field Crops Research, vol. 263, p. 108069, 2021.
[4] B. Basso and L. Liu, “Seasonal crop yield forecast: Methods, applications, and accuracies,” Advances in Agronomy, vol. 154, pp. 201-255, 2019.
[5] IPCC, Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report, 2022.
[6] A. Crane-Droesch, “Machine learning methods for crop yield prediction and climate change impact assessment in agriculture,” Environmental Research Letters, vol. 13, no. 11, p. 114003, 2018.
[7] D. Paudel, H. Boogaard, A. de Wit, S. Janssen, S. Osinga, C. Pylianidis, and I. N. Athanasiadis, “Machine learning for large-scale crop yield forecasting,” Agricultural Systems, vol. 187, p. 102983, 2021.
[8] L. Zhang, Z. Zhang, Y. Luo, J. Cao, and F. Tao, “Combining optical, fluorescence, thermal satellite, and environmental data to predict county-level maize yield in China using machine learning approaches,” Remote Sensing, vol. 12, no. 1, p. 21, 2019.
[9] M. Purevdorj, A. Zare, and A. Jafari, “Integration of weather data into a machine learning model to predict sunflower yield,” Computers and Electronics in Agriculture, vol. 182, p. 105985, 2021.
[10] H. Kaur and M. Kaur, “A review on crop yield prediction using data mining techniques,” in Proc. 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), 2020, pp. 1272-1276.
[11] E. Kamir, F. Waldner, and Z. Hochman, “Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 160, pp. 124-135, 2020.
[12] G. Jang, J. Kim, J. K. Yu, H. J. Kim, Y. Kim, D. W. Kim, et al., “Review: Cost-effective unmanned aerial vehicle (UAV) platform for field plant breeding application,” Remote Sensing, vol. 12, no. 6, p. 998, 2020.
[13] M. Kukar, P. Vracar, D. Kosir, D. Pevec, and Z. Bosnic, “AgriFood supply chain traceability: Data sharing in a farm-to-fork case,” Computers in Industry, vol. 123, p. 103293, 2021.
[14] Q. Yang, L. Shi, J. Han, Y. Zha, and P. Zhu, “Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images,” Field Crops Research, vol. 235, pp. 142-153, 2019.
[15] S. Khaki and L. Wang, “Crop yield prediction using deep neural networks,” Frontiers in Plant Science, vol. 10, p. 621, 2019.
[16] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” Nature Machine Intelligence, vol. 1, no. 5, pp. 206-215, 2019.