Journal of Animal & Plant Sciences, 36(3): 2026, Page: 754-765
ISSN (print): 1018-7081; ISSN (online): 2309-8694
https://doi.org/10.36899/JAPS.2026.3.0062

INETERPRETABLE MACHINE LEARNING FOR SOYBEAN YIELD PREDICTION WITH SHAP-BASED INSIGHTS

I. A. Cheema, M. K. Hanif*, M. U. Sarwar and M. I. Khan

Department of Computer Science, Government College University, Faisalabad, Pakistan
*Corresponding Author's e-mail: mkashifhanif@gcuf.edu.pk

ABSTRACT

Accurate crop yield prediction is important to minimize uncertainty for informed decision- making and resource allocation. A variety of machine learning models are used in yield prediction; however, the available benchmarking literature offers limited insight to achieve a balance between predictive accuracy and model interpretability of different models. Therefore, this study was conducted to evaluates popular machine learning models for U.S. soybean yield prediction using a multi-source spatiotemporal dataset comprising weather, soil, and management features. The model performance was evaluated using root mean squared error (RMSE) metric, and feature impact was explained using Shapley Additive Explanations (SHAP) for interpretability. The findings indicate that Random Forest is the best model that achieved least RMSE of 5.07 and highest correlation coefficient of 90.36% on test set. SHAP results revealed that precipitation and solar radiation are leading yield determinants, while soil properties, such as soil pH and bulk density, exerted moderate effects. The contribution of this work is fourfold: (i) a rigorous benchmarking of ML models using accuracy metrics for yield prediction, (ii) evidence based affirming the model superiority for complex agronomic dataset, (iii) systematic assessment of global feature importance connecting yield affecting climatic and edaphic factors, and (iv) application of SHAP as a means for interpretation and explainability. The results bring together predictive performance and explanation, providing insights into the advancement of smart agriculture through informed decision-making for irrigation planning, efficient input application, and climate-resilient strategy formulation.

Key words: Crop Yield Prediction, Informed decision-making, Machine Learning, Smart Agriculture, SHAP Interpretability, Explainable AI

Published first online February 14, 2026 Published final May 05, 2026

INTRODUCTION

Accurate and timely forecasting of crop yield in large territories at an affordable cost is crucial for global food security (Ahmad *et al*., 2024), especially amid population surge, reducing arable land and climatic shifts (Junaid and Gokce, 2024; Yuan *et al*., 2024). Smart farming and precision agriculture enable proactive monitoring and data analytic that prescribe data-driven irrigation, fertilization and pest management on site (Karunathilake *et al.,* 2023). The major hurdle in adoption of smart agricultural practices is setting realistic and accurate targets for each field segment, especially crop yield before planting begins (Cai *et al.,* 2019; Han *et al.,* 2020; Wakweya, 2023). Reliable forecasting supports even better grain policy formulation and plant breeding programs by detecting high yielding genotypes faster, rather than using traditional methods (Maimaitijiang *et al.,* 2020; van Dijk *et al.,* 2021). The application of traditional statistical approaches using field surveys is often limited in precision, adaptability and scalability to the diverse surrounding due to complex interactions of features impacting the crop yield (Basso and Liu, 2019; Xiong *et al*., 2022).

The advancement of Internet of Things (IoT), remote sensing, Unmanned Aerial vehicle (UAVs) and the availability of multi-source datasets have provided better opportunities to employ machine learning (ML) techniques for crop yield forecasting (Shafi *et al*., 2020; Van Klompenburg *et al*., 2020). Recent studies employed spatiotemporal data having climate records, soil characteristics, remotely sensed vegetation indices such as NDVI and EVI, UAV imagery, thermal imagery, solar-induced chlorophyll fluorescence (SIF), and crop management features in yield prediction tasks (Rashid *et al.,* 2021; Cedric *et al*., 2022; Metin *et al.,* 2023). Several ML models ranging from Random Forest (RF), Gradient Boosting, Support Vector Regression to XGBoost, ensemble learning, hybrid frameworks and deep learning (DL) have been employed and tested with promising results in enhancing the predictive accuracy of the crop yield (Morales and Villalobos, 2023; Zhou *et al*., 2023; Jabed and Murad, 2024). The use of ML in agriculture has grown significantly for tasks such as predicting crop

yield, classifying crop types, detecting diseases and recommending crops. While DL models are also popular choices in these areas but those are particularly considered "black boxes" lacking interpretability and transparency. Furthermore, the agricultural domain often lacks the vast datasets and computational power required by DL implementations (Batool *et al.,* 2022; Oikonomidis *et al.,* 2022). Comparatively, ML models are less data extensive, simpler to train and interpret, and provide competitive accuracy along-with practical advantages in explainability and deployment for smallholder farmers to strengthen smart agriculture practices on large scales (Bazrafshan *et al.,* 2022; Jabed and Murad, 2024).

Even though accuracy is considered as important benchmark for evaluating models, the sole reliance on predictive performance limits the real-world adoption. Despite the advancements and growing use of ML models in predicting crop yield, several critical gaps remain (Elavarasan *et al.,* 2020; Prasad *et al.,* 2021). These gaps mainly consist of single source data, limited spatial scope, missing stakeholder-oriented analysis, methodological constraints, sole focus on accuracy without explainability, inadequate benchmarking, lack of interpretability and poor generalization across datasets and geolocations (Abbas *et al.,* 2020; Shahhosseini *et al.,* 2020; Ahmed, 2023). For transparent and data driven decision-making, this is not sufficient to merely predict what will happen and ignore the equally important aspect of why it will happen. To address these gaps, there is pressing need for a comprehensive methodology that uses ML models trained on high-quality spatiotemporal, multi-source dataset integrating weather, soil, and management features. There must be rigorous accuracy benchmarking with interpretability in explainable artificial intelligence (XAI) to promote adoption of smart agricultural practices. This research aimed (I) to evaluate multiple ML models for yield prediction of soybean (*Glycine max* L. using comprehensive methodology; (II) to compare their performance by using multi-source spatiotemporal dataset; (III) to analyze feature importance using ML-based ranking; and (IV) to employ SHAP analysis for interpretability and agronomically significant insights By accuracy benchmarking along-with explainability, this study contributes in advancing ML-based yield forecasting towards stakeholder-oriented smart agriculture practices for informed decision-making and resource optimization.

MATERIALS AND METHODS

This section describes the methodological framework to carry out a comprehensive comparison using ML models for crop yield prediction, by employing a feature-rich spatiotemporal dataset, while also providing a rigorous assessment of accuracy and interpretability. These steps include data preprocessing, hyperparameter tuning, model training, model evaluation and interpretability analysis as shown in Fig. 1.

The US soybean dataset used in this paper is obtained from publicly available source (Ansarifar, 2020). The United States is the main soybean producer with a cultivation area of approximately 33.45 million hectares distributed among 31 states (USDA, 2022). The Selected dataset covers 12 states and 1045 counties, with a total cultivated area equal to 29.69 million hectares, which represents approximately 75% of the total planted soybeans in the country (NASS, 2024).

The yield, a target feature, is represented by bushels per acre (bu/a) and reflects ground truth values ranging from 1980 to 2018 (39 years). There are six weather components, each having 52 weeks of data points, are precipitation (mm), solar radiation (MJ/m^2), minimum temperatures (°C), snow water equivalent (mm), maximum temperatures (°C) and vapor pressure (kPa). Measured at six depths, eleven soil variables include Bulk density (g/cm^3), coarse fragments (%), cation exchange capacity (cmol), total nitrogen ($kg\ ha^1$), clay content (%), organic carbon density ($g\ cm^3$), soil pH (pH), sand content (%), organic carbon stock $kg\ ha^1$, silt content (%) and soil organic carbon ($kg\ ha^1$). The management practices features are the accumulated percentages of planted fields represented as of weekly basis, starting from April every year. Fig. 2a shows the US soybean production area map and Fig. 2b illustrates the historical soybean yield where the significantly increasing yield trend over the last few decades is obvious. The rising trend can mainly be credited to consistent improvements in seed quality and management methods because of investments in breeding techniques' research and development.

The dataset required specific preprocessing steps suitable for ML models. The set of categories that include location descriptors were integer-encoded for their location relevance, and time-series variables were time-ordered. Missing time-series were dealt with like this: short gaps were linearly interpolated, and long gaps were filled with location-wise means to avoid bias. Feature engineering captured the time-varying structure of yield with moving averages that are relevant to crop growth and an impactful feature for prediction. Despite their prevalence, common agroclimatic indices like Growing Degree Days (GDD) and drought indices were not used to avoid redundancy and collinearity with the weather variables already present. Also, using easily measurable variables improves the ease of SHAP analysis to delineate clearer actionable recommendations for stakeholders.

We used several techniques of data preprocessing to ensure that the model training and evaluation is performed using high quality inputs. The mean imputation was used to address the missing feature

values in the management practices (3%) and the soil (7%). The values of yield that are below 5 bu/a dropped due to the consideration of being outliers. To have comparable scales for all the features, we used StandardScaler that is available in python library as scikit-learn (Hao and Ho, 2019). The feature scaling process subtracts the average and then divides the result by unit variance so that every attribute having zero mean and the standard deviation that is equal to one. We implemented several popular ML models for soybean yield prediction; random forest (RF), multi-layer perceptron (MLP), extreme gradient boosting (XGB), gradient boosting (GB), LASSO regression (LASSO), LightGBM, ElasticNet, support vector regression (SVR), AdaBoost, and Decision Tree (DT) (Huang *et al*., 2020). MLP was used as a DL baseline, that is much simpler than a Convolutional Neural Network (CNN) or a Recurrent Neural Network (RNN) so an MLP is appropriate for the kind of structured, tabular data used in this study. More complex architectures were excluded because the compute intensive and lack of transparency associated with black-box systems would be counterproductive to the aim of providing explainable, decision-supportive models for the agricultural stakeholders. The trainable dataset was divided into two parts, 80:20 (train:test) split, to ensure an unbiased evaluation where training part used to construct the models and the test part helped understand the performance outside of known values. Each split is based upon the temporal (years) and spatial (locations) aspects, observing their variations over time and space. This ensures the strength and wider reach of the analysis, where the ML models are assessed under varying conditions of seasonality and locations. This also avoids leakage and emulates the task of predicting yields based on temporal patterns while considering spatial variations, generalization across time and location. A rigorous hyperparameter tuning was done for every model to enhance performance using a cross-validation method with 10-folds where the training part is split into ten different folds. The model training was performed using nine folds whereas validation on the remaining fold. The process was iterated for the folds, making it a more reliable method of best hyperparameter settings. We used the grid search technique and systematically scanned the range of hyperparameters to locate the best. In Table 1, the selected best hyperparameters of each ML model are listed.

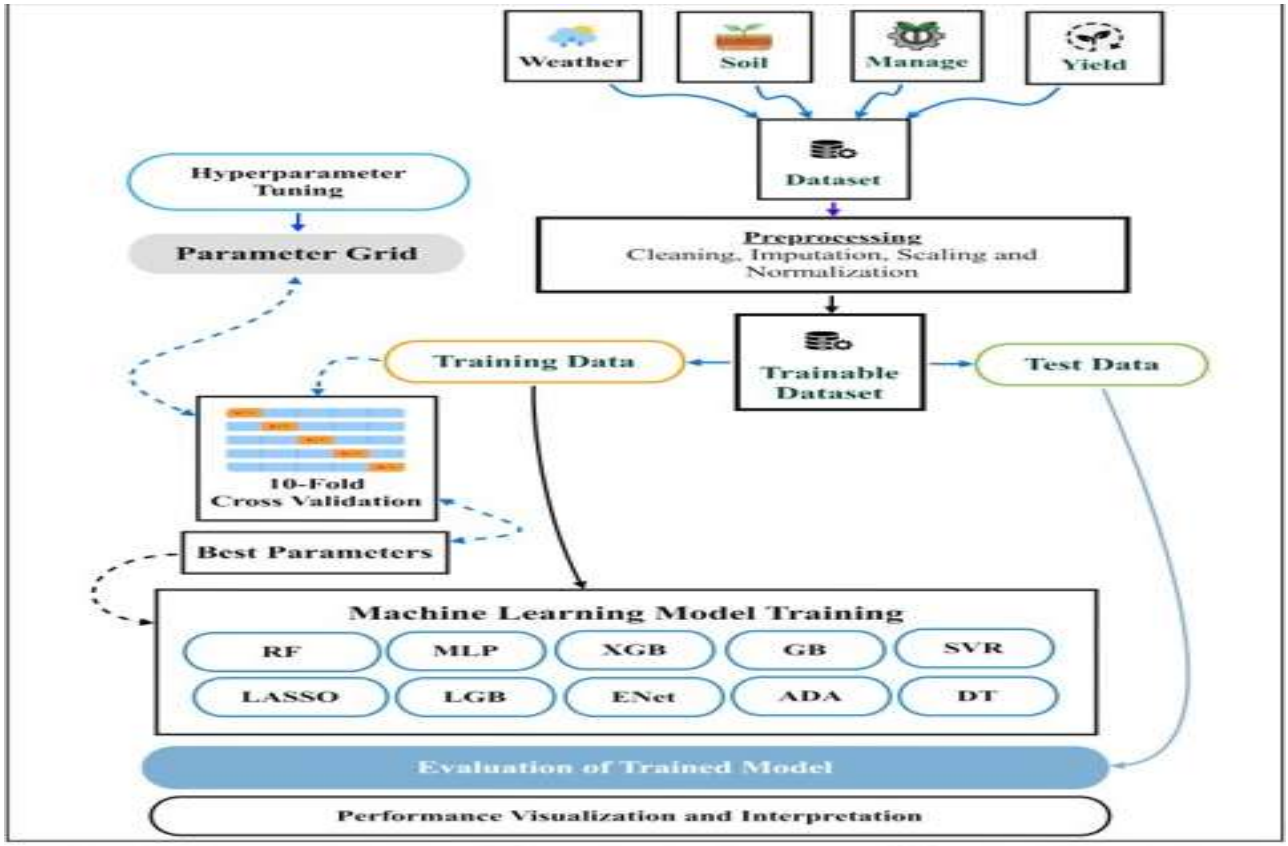
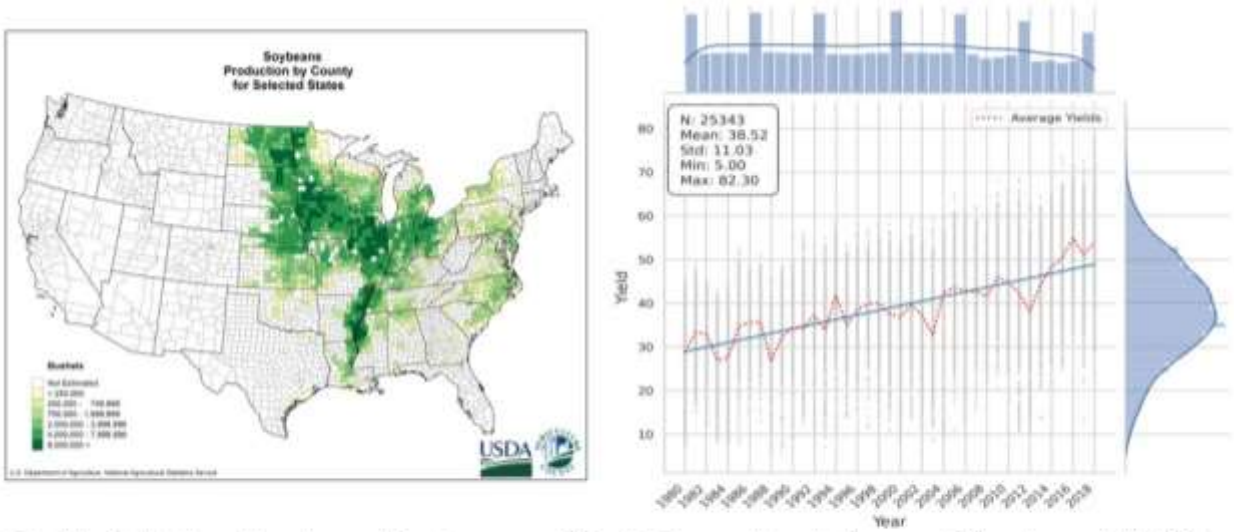


Fig. 1: Workflow of the interpretable machine learning framework for crop yield prediction.

[1]Here RF = Random Forest; MLP = Multi-Layer Perceptron; XGB = Extreme Gradient Boosting; GB = Gradient Boosting; SVR = Support Vector Regression; LASSO = Least Absolute Shrinkage and Selection Operator; LGB= Light Gradient Boosting; ENet= ElasticNET; ADA= Adaptive Boosting; DT = Decision Tree.



(a) Spatial distribution of soybean cultivation across US counties included in the study, representing approximately 75% of total U.S. soybean planted area. Counties are shaded according to their relative production intensity.

(b) Temporal trend of average U.S. soybean yield (1980–2018) with a linear regression fit (R^2=0.87), illustrating the steady increase in yield attributable to genetic improvements, advanced management practices, and technological adoption.

Fig. 2: Spatial distribution and temporal yield trend of US soybean production. (a) Geographic distribution of soybean cultivation across US counties included in the study. (b) Historical trend of average soybean yield (bushels per acre) from 1980 to 2018.

As an evaluation metric, RMSE (Hodson, 2022) and *r* (Akoglu, 2018) were employed. RMSE quantifies the errors a model typically makes during the predictions and computes an average error value between predicted and ground truth points, with smaller values indicating better performance. On the other hand, *r* is used to measure the direction and strength of the relation between the ground truth values and predicted values. The ranges are from −1 to 1, where the number that is near to 1 or −1 shows a strong relationship either positive or negative respectively.

Table 1: Optimized hyperparameters configurations for all evaluated ML models.

RF	n_estimators = 200, max_depth = 20, min_samples_leaf = 5, bootstrap = True, max_features = 'sqrt', min_samples_split = 10, ccp_alpha=0.06
MLP	hidden_layer_sizes= (50,1), solver = 'adam', activation = 'relu', alpha = 0.1, max_iter = 500, early_stopping = True, learning_rate = 'constant', validation_fraction = 0.1
XGB	n_estimators = 50, max_depth = 3, learning_rate = 0.2, subsample = 0.8, objective = 'reg:squarederror', colsample_bytree = 0.8
GB	n_estimators = 50, max_depth = 3, learning_rate = 0.1, min_samples_split = 10, subsample = 0.8, min_samples_leaf = 5
LASSO	alpha=0.1
LightGBM	n_estimators = 50, max_depth = 3, learning_rate = 0.08, min_child_samples = 5, subsample = 0.8, colsample_bytree = 0.8
ElasticNET	alpha = 0.5, l1_ratio = 0.7
SVR	kernel = 'rbf', C = 0.05, epsilon = 0.5, gamma = 'auto'
ADABoost	n_estimators = 10, learning_rate = 0.05
DT	max_depth = 50, min_samples_leaf=5, min_samples_split=10, max_leaf_nodes=10, max_features = 'sqrt'

Equations 1 and 2 show formula to calculate RMSE and *r*, where y_j denotes the ground truth value, $\hat{y}_j$ is the predicted value, and the overall observations are designated as *n*. Where y and $\hat{y}$ representing the mean actual and predicted data, respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2} \quad (1)$$

$$r = \frac{\sum_{j=1}^{n}(y_j - \bar{y})(\hat{y}_j - \bar{\hat{y}})}{\sqrt{\sum_{j=1}^{n}(y_j - \bar{y})^2 \cdot \sum_{j=1}^{n}(\hat{y}_j - \bar{\hat{y}})^2}} \quad (2)$$

We used the feature importance technique to find the most important predictors in our model. This technique quantifies how much each input variable can contribute to a model's accuracy, allowing to prioritize features with the highest impact on crop yield prediction that is a useful tool for dimensionality reduction. We employed SHAP analysis for explainable insights to obtain the understanding that a particular variable is affecting the model's predictive power. SHAP interpretability explains the direction of the relationship between input features and crop yield. These insights are valuable as they allow agricultural stakeholders to implement optimization strategies, allocate resources efficiently, and make informed decisions. We performed the experiments using Python 3 Google Compute Engine back-end on Google Collab Notebook with the scikit-learn library (Hao and Ho, 2019).

RESULTS

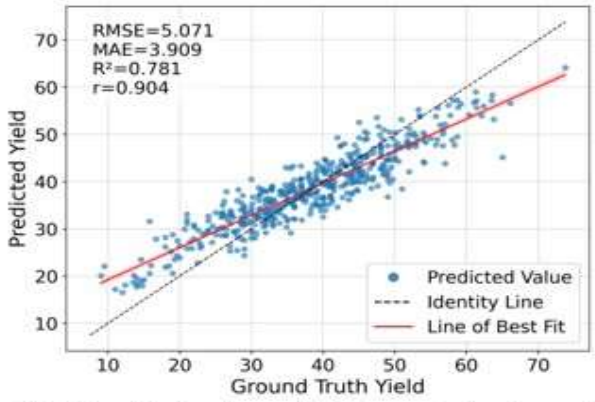
The relative performance of ten ML models for prediction of soybean yield is presented in Table 2. The RF model outperformed with the lowest test RMSE (5.07) and highest correlation coefficient (r = 90.36%) of all candidates, closely followed by the MLP with an RMSE of 5.25 and r of 87.54%. Linear ensemble learners (GB, XGB) also showed similar good performance, compared to simple linear models (LASSO, ElasticNet) and SVR. DT and AdaBoost showed lowest predictive performance with a test RMSE over 8.5 and correlation less than 61%. These findings suggest that nonlinear ensemble methods are superior because they can account for complex features interactions, non-linear relationships and heterogeneities in agricultural data which cannot be well described by linear models and single learners.

The parity plot (Fig. 3a) showing a reasonable alignment between predicted and actual yield, with correlation (r = 90.36%) and coefficient of determination (R^2 = 0.781), indicating model robustness. The residual spread analysis (Fig. 3b) shows that prediction errors are mostly concentrated around zero and the underestimation slightly increases for higher yield values. In Fig. 3c, Error Distribution histogram shows an almost Gaussian pattern, confirming the presence of unbiased residuals. The RMSE value estimated at 5 bu/a leads to a quite moderate error at farm level. For example, in production fields with average yields of approximately 50–60 bu/a, this uncertainty translates to an 8–10% error, which is generally acceptable for operational decision making at the field management level. The learning curve (Fig. 3d) shows that RMSE improves with increasing size of the training set, though this improvement becomes smaller from around 12,000 training examples on-wards, past this point adding more data would have diminishing returns under current feature availability. For further examination of statistical stability, RMSE distributions from cross-validation (Fig. 3e) indicate consistency of RF model performance with a mean RMSE 5.11 and small standard deviation across cross-validation folds. This finding validates about the stability and lower uncertainty of the model in multiple sampling. The line plot of predicted values vs test samples (Fig. 3f) verifies that the RF model accounts for temporal and spatial fluctuation in the yield trends across test samples, except occasional discrepancies where ground truth yield is largely deviating from average.

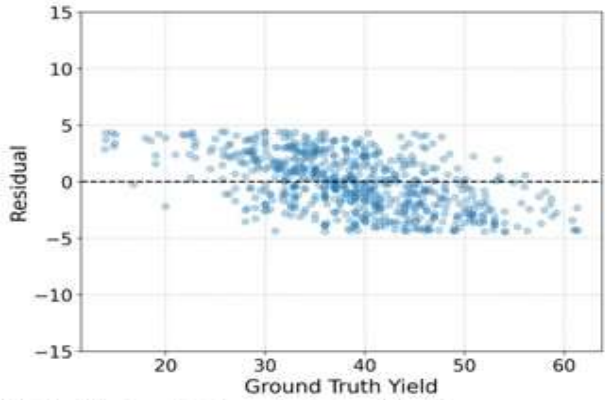
Table 2: Performance comparison of ML models based on RMSE and correlation coefficient for training and test datasets.

Model	Type	RMSE (Training)	Correlation Coefficient (Training)	RMSE (Test)	Correlation Coefficient (Test)
RF	**ensemble**	**4.56**	**92.99**	**5.07**	**90.36**
MLP	Neural	4.96	89.46	5.25	87.54
XGB	ensemble	5.66	86.73	5.84	84.97
GB	ensemble	5.73	82.17	5.92	81.01
LASSO	Linear	5.95	81.56	6.02	79.39
LightGBM	ensemble	6.18	77.48	6.29	75.90
ElasticNET	Linear	6.89	72.30	6.91	71.92
SVR	SVM	7.65	67.05	7.77	66.80
ADABoost	ensemble	8.59	63.50	8.63	60.74
DT	Tree	9.22	52.62	9.27	54.72[1]

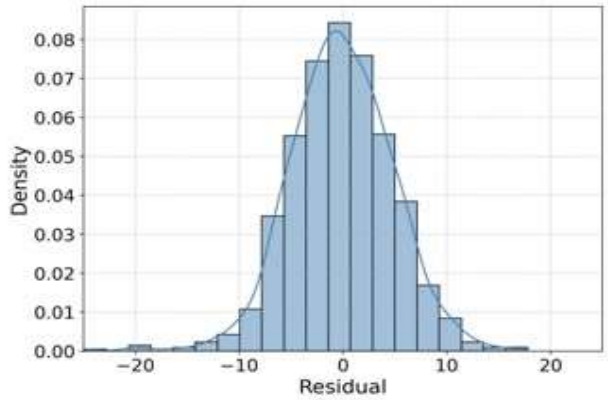
Here RMSE = Root Mean Square Error; RF = Random Forest; MLP = Multi-Layer Perceptron; XGB = Extreme Gradient Boosting; GB = Gradient Boosting; LASSO = Least Absolute Shrinkage and Selection Operator; SVR = Support Vector Regression; DT = Decision Tree.



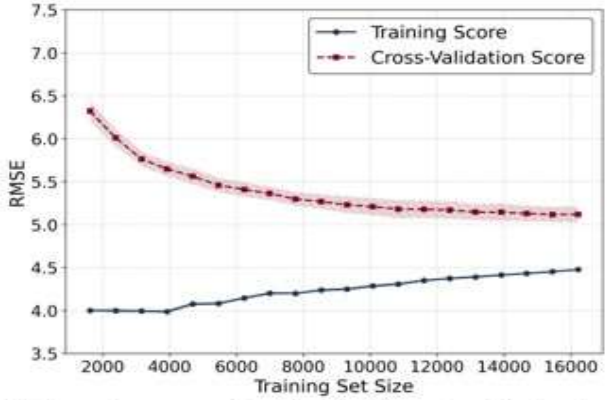
(a) *Parity plot showing predicted vs. actual soybean yields with identity and best-fit lines*



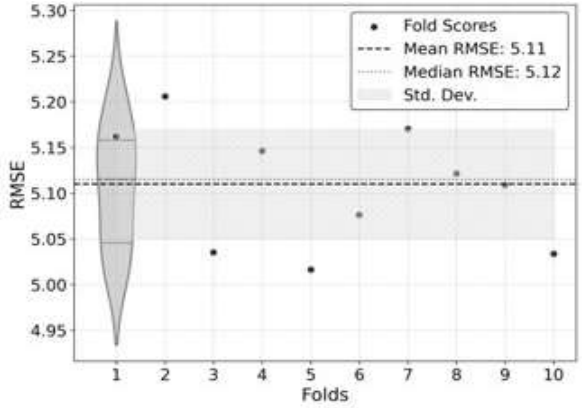
(b) *Residuals plotted against actual yield to assess error patterns.*



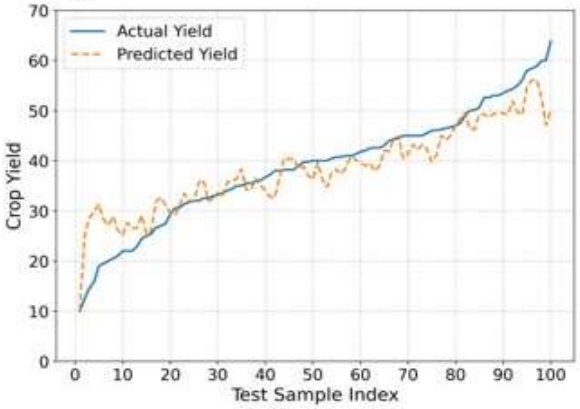
(c) *Distribution of residual errors illustrating model bias and variance.*



(d) *Learning curve of the Random Forest model showing training and cross-validation RMSE trends.*



(e) *Cross-validation RMSE distribution across folds with summary statistics.*

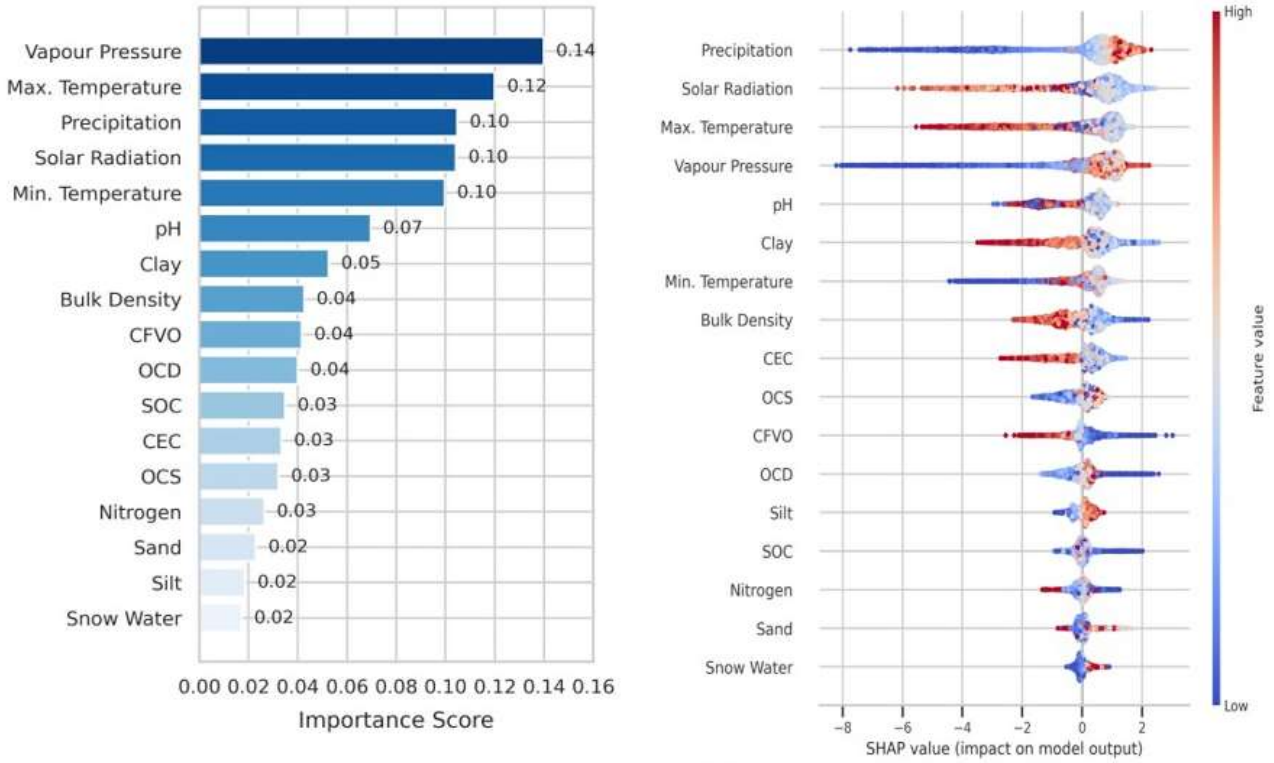


(f) *Line plot comparing actual and predicted yields for the test samples.*

Fig. 3: Performance evaluation of the best-performing RF model. (a) parity plot of predicted vs. actual yields, (b) residuals vs. actual yield, (c) Histogram of prediction errors, (d) learning curve for training and cross-validation, (e) Distribution of RMSE values across 10 cross-validation folds, and (f) comparison of actual and predicted yields on the test set.

Fig. 4a shows a bar plot illustrating the feature importance rankings produced by RF model. Here, y-axis denotes features and the x-axis about the scores pertaining to importance value. The Vapour Pressure

feature is considered the most influencing having highest score of 0.14, followed closely by maximum temperature, precipitation, and solar radiation. This demonstrates that climate factors are higher in their impact on determining the accuracy of model. Vapour Pressure and Precipitation have a strong positive correlation, with similar impact of Solar Radiation and Maximum Temperature. These findings highlight key attributes in understanding complex agronomic traits. Among the six weather components examined here, the Vapour Pressure and Maximum Temperature are among the most sensitive factors, while the equivalent of Snow Water is the less sensitive to soybean crop yield prediction accuracy. This makes sense from an agronomic perspective, as Solar Radiation and Precipitation are the important drivers for photosynthesis, biomass, and improved grain yield (Wadgymar *et al.*, 2018). The feature importance technique demonstrated how input features influenced the estimation of soybean crop yield at global level but not in feature level about what direction these impacts would go and their magnitude. Moreover, there could be possible interactions and nonlinear impact among features. This gap was tackled through the SHAP interpretability analysis so that comprehensive insights are extracted having significant agronomic value to support informed decision-making and promote smart agricultural practices.



(a) *Feature importance ranking showing the relative contribution of climatic and soil variables to the model's predictive performance.*

(b) *SHAP summary plot illustrating the impact and direction of each feature on model output, with color indicating feature value magnitude.*

Fig. 4: Feature importance and SHAP-based interpretability analysis showing the relative influence and directional effects of key climatic and soil predictors on model performance.

SHAP interpretability analysis defines how much each feature adds to or subtracts from one's yield prediction and how much they influence and to what degree (Lundberg and Lee, 2017). The SHAP summary plot in Fig. 4b measures the importance of the input features towards predicting soybean yields using the RF model. Rainfall, maximum temperature, and solar radiation were the most influential predictors, with vapor pressure, soil pH, and clay content as of secondary impact. These predictors have large and consistent SHAP values which were much higher than the average value and thus exerted strong nonlinear impact on crop yield predictions. Secondary variables including the soil nitrogen, sand content and snow water had weaker impact on the prediction accuracy. The direction and magnitude of impact of SHAP values have revealed insightful agronomic relationships. For example, most of the time larger rainfall values were associated with positive contributions to yield prediction, reflecting the need for water availability throughout soybean development.

Likewise, favorable solar radiation and maximum temperature ranges contributed to a positive direction, but not extreme heat. Soil attributes such as pH, BD and clay content also showed high variability with deviation from agronomically optimal values that detracted the predicted yield. The interpretability results establish that the soil properties are secondary to the acute climate variability in determining the yield. Jointly, these findings illustrate the role of climatic and soil variables in structuring soybean productivity, and that the model accounts for this interplay in an understandable way.

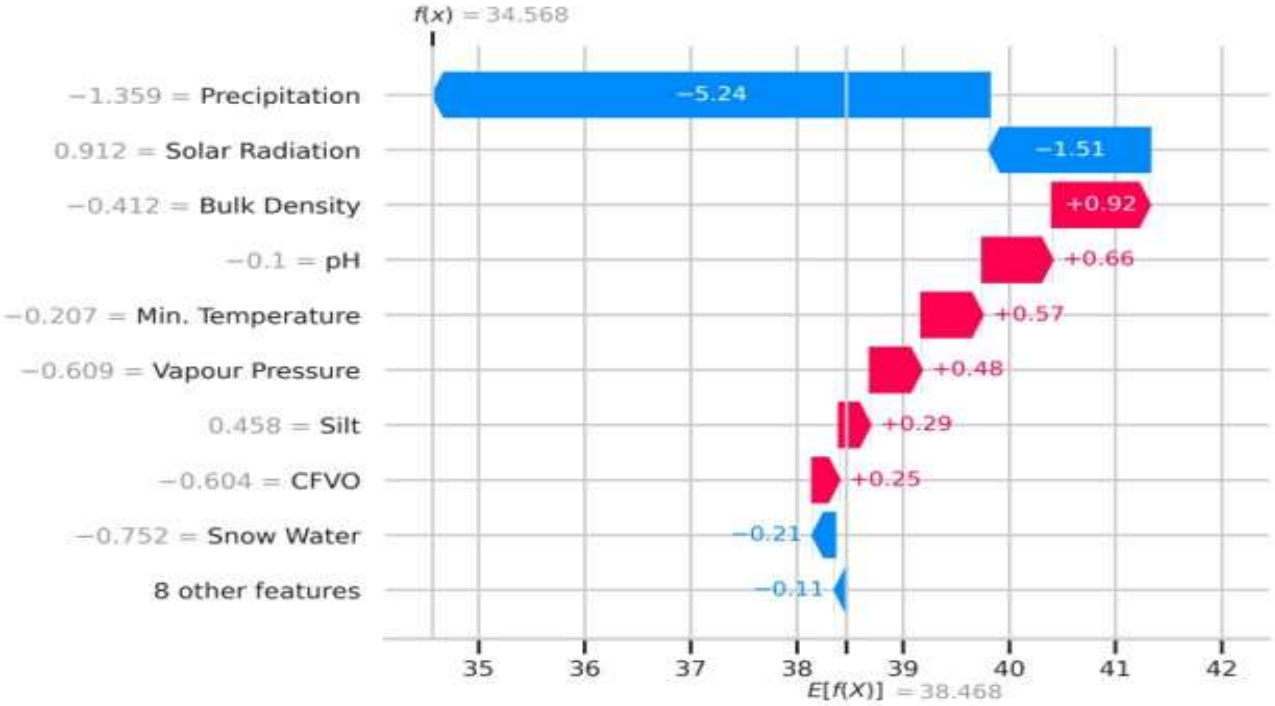


Fig. 5: SHAP waterfall plot illustrates how each feature increases or decreases the model's predicted yield for a specific sample.

SHAP waterfall plot in Fig. 5 explains the significance of contributions by various features to a particular chosen instance. The x-axis shows the yield values whereas the respective SHAP values are marked at the y-axis as horizontal bars. The bars indicate the direction of contribution, whereas the length of bar shows the contribution magnitude. The model predicts a value of $f(x) = 34.568$ for this instance, starting from a base value (the expected prediction) of $E[f(X)] = 38.468$. To obtain the final prediction (Equation 3), all these SHAP values are added to the base prediction: $SHAP_i$ is called the SHAP value of the i-th input variable. According to this instance, *Precipitation* is the most powerful variable with a feature value of −1.359, resulting in a significant effect of −5.24, which reduced the forecast sharply downward. Another important factor is *Solar Radiation*, having feature value equals 0.912 and SHAP value -1.51, decreased the final predicted. The other variables such as *Bulk Density* and *pH* contributed towards higher prediction with positive SHAP values of +0.92 and +0.66 correspondingly.

$$f(x) = E[f(X)] + \sum_{i=1}^{n} SHAP_i \quad (3)$$

Features like *Silt* and *CFVO* contribute positively towards increasing the prediction, having SHAP values of +0.48 and +0.25 respectively. In general, the sum of individual feature contributions resulted in a decrease from the expected prediction of 38.47 to the final prediction of 34.57, primarily dominated by the significant negative contributions of *Precipitation and Solar Radiation.*

DISCUSSION

This study demonstrates that an interpretable ML framework, combining predictive modeling with SHAP, can provide both accurate forecasts and agronomically intelligible insights for crop yield prediction. This integrated approach addresses a critical gap in agricultural data science, where high predictive accuracy alone is insufficient for stakeholder adoption

and informed decision-making (Wakweya, 2023; Ahmad *et al.*, 2024).

Our results align with the broader consensus in recent crop yield prediction literature, where ensemble tree-based methods consistently rank among top performers. RF model achieved a test RMSE of 5.07 bu/a and an *r* of 0.904, outperforming other candidates. This finding is similar with studies on wheat (*Triticum aestivum* L.) and maize (*Zea mays* L.) where RF and GB variants often surpass simpler linear models and single decision trees due to their inherent ability to model complex, nonlinear interactions within heterogeneous agricultural data (Van Klompenburg *et al*., 2020; Burdett and Wellen, 2022). For instance, Shahhosseini *et al*. (2021) reported RF as a robust performer for corn yield prediction across the US Corn Belt. XGBoost and GB performed admirably (RMSE ~5.8-5.9). However, they did not surpass RF in this specific spatiotemporal soybean dataset. This contrasts with some studies where boosting algorithms marginally outperform bagging ensembles, suggesting feature-target relationships or dataset characteristics influence the optimal model choice (Liaw and Wiener, 2002; Jabed and Murad, 2024). MLP performed competitively (RMSE = 5.25), serving as a capable deep learning baseline. Yet, more complex deep learning architectures (e.g., CNNs, RNNs), while powerful for image or sequence data (Khaki *et al.*, 2020; Zhou *et al.*, 2023), were not considered here, prioritizing model transparency and computational efficiency for tabular data which is a common trade-off discussed in agricultural XAI literature (Bazrafshan *et al.*, 2022). Our findings on model performance extend previous benchmarking efforts in crop yield modeling, reinforcing the need for context-specific model selection (Junaid and Gokce, 2024).

A key contribution of this work is the systematic application of SHAP for yield model interpretation, an approach gaining traction in agricultural sciences. Recent studies have employed SHAP to demystify crop yield models for wheat (*Triticum aestivum*) (Cao *et al.*, 2020), maize (*Zea mays*) (Feng *et al.*, 2019), and oil palm (*Elaeis guineensis*) (Rashid *et al.*, 2021). Our SHAP analysis reaffirms the primacy of climate variables—particularly precipitation and solar radiation—as dominant yield predictors, a conclusion consistent with these studies across different crops and geographies. This cross-crop consistency underscores the fundamental role of these abiotic factors in crop productivity (Xiong *et al.*, 2022; Yuan *et al.*, 2024). Our work extends this discourse by applying SHAP to a high-dimensional, multi-source US soybean dataset, explicitly linking global feature rankings with local, instance-level explanations to reveal contextual dependencies (Karunathilake *et al.*, 2023).

The SHAP-based identification of precipitation and solar radiation as primary drivers is deeply rooted in soybean physiology. Soybean is particularly sensitive to water deficit during critical reproductive stages, specifically flowering and pod-filling (R1–R5). Insufficient precipitation during these windows drastically reduces pod set and seed size, directly impacting final yield (Chen *et al.*, 2016). Our model has captured this sensitivity, with lower precipitation values frequently associating with negative SHAP values. Solar radiation drives photosynthetic rate and biomass accumulation. The positive association of solar radiation with yield in our SHAP summary plot reflects the crop's radiation-use efficiency. However, the model also learned subtleties, as extremely high radiation values coupled with water stress can lead to photoinhibition and reduced efficiency, a nuance visible in the distribution of SHAP values (Wadgymar *et al.*, 2018).

The significant roles of soil pH and bulk density highlight the importance of edaphic conditions. Soil pH critically governs the availability of essential nutrients like phosphorus, molybdenum, and rhizobia symbiont vitality for nitrogen fixation (Fageria and Baligar, 2008). Suboptimal pH limits these processes, which our model interprets as a yield constraint. Bulk density, an indicator of soil compaction, affects root penetration, water infiltration, and aeration. High bulk density restricts root exploration, limiting access to water and nutrients, particularly in subsurface layers, thereby exacerbating drought stress (Lipiec *et al.*, 2012). The model's identification of these factors underscores its ability to integrate climate-soil interactions that define crop-growing conditions (Basso and Liu, 2019).

This study discusses the trade-off between model complexity and interpretability. While deep learning models can achieve high accuracy, they often function as "black boxes," hindering trust and adoption by farmers and agronomists. As emphasized by Rudin (2019) and echoed in agricultural AI reviews (Jabed and Murad, 2024), interpretable models are crucial for high-stakes domains like agriculture. Our approach demonstrates that highly interpretable models like RF, when coupled with post-hoc explainability tools like SHAP, can achieve better accuracy while providing the explanations necessary for actionable insight (van Dijk *et al.*, 2021).

These insights have direct implications for farm management. The prominence of precipitation supports data-driven irrigation scheduling, encouraging interventions during identified critical dry periods. The importance of solar radiation can inform planting date selection to align peak canopy demand with periods of high radiation. Soil property insights advocate for site-specific management: fields with suboptimal pH may be prioritized for liming, while areas with high bulk density may benefit from subsoiling or controlled traffic to alleviate compaction. For policymakers and breeders, identifying these key stress factors (water, heat, soil constraints) provides empirical targets for investing in

resilient infrastructure, developing drought-tolerant varieties, or promoting soil health programs (Wakweya, 2023).

This study has several limitations that provide avenues for future research. First, the dataset, while extensive, is region-specific to the US and may not generalize directly to other soybean-growing regions with different climatic regimes or management practices. Second, the temporal resolution is annual, missing within-season dynamics that could improve in-season forecasting. Third, management features are aggregated and lack field-specific detail on cultivars, fertilizer rates, or pest control.

Future work should focus on: (1) Multi-modal data fusion: Integrating high-resolution remote sensing (e.g., Sentinel-2, UAV) and IoT sensor data can capture within-field variability and phenological dynamics, as demonstrated in maize and wheat studies (Maimaitijiang *et al.,* 2020; Shafi *et al.,* 2020; Zhou *et al.,* 2023). (2) Temporal deep learning with explainability: Exploring interpretable hybrid models or using tools like Temporal Fusion Transformers (TFTs) with integrated attention mechanisms could capture temporal dependencies while maintaining some interpretability (Metin *et al.,* 2023). (3) Causal analysis: Moving beyond correlation via causal inference methods could help distinguish direct drivers of yield from confounding factors, strengthening the agronomic basis of recommendations. (4) Cross-crop and global validation: Applying this interpretable ML framework to other staple crops and diverse agro-ecologies will test its robustness and expand its utility for global food security (Ahmad *et al.,* 2024).

In summary, this research reinforces the value of RF for accurate soybean yield prediction while championing SHAP as an essential tool for transforming predictions into agronomic intelligence. By bridging the gap between statistical performance and biological interpretability, we provide a transparent, trustworthy framework that can inform decisions from the field to the policy level. As climate volatility intensifies, such explainable AI frameworks will be indispensable for developing adaptive, resilient, and sustainable agricultural systems.

Conclusion: This work showcases the potential of interpretable ML models to improve crop yield prediction with explainable insights, taking soybean as a test case. The RF model was superior in accuracy and stability; the prediction errors were acceptable from an agronomical point of view. Most importantly, SHAP-based explanations unveiled the hidden factors influencing yield variability and thereby connected black-box prediction with actionable agronomic understanding. Such findings are useful for farmers to calibrate practices at field-level, for breeders in focusing traits and for policy-makers in planning targeted interventions. By placing the results in a larger climate resilience and food security context, this paper demonstrates how transparent ML models can be predictive as well as prescriptive. Ground truth experiments across crops, regions, and data types will be essential to validate these findings and fully unlock the potential of explainable AI for sustainable smart agriculture.

Conflict of interest: The authors declare no conflicts of interest.

Author's contribution: IAC and MKH conceptualization, methodology, and writing original draft; MUS and MIK formal analysis, validation, and investigation; IAC data curation and visualization; MKH supervision.

Funding statement: The authors received no specific funding for this study.

REFERENCES

Abbas, F., H. Afzaal, A.A. Farooque and S. Tang (2020). Crop yield prediction through proximal sensing and machine learning algorithms. Agronomy. 10(7): 1046. https://doi.org/10.3390/agronomy10071046

Ahmad, A., A. X. Liew, F. Venturini, A. Kalogeras, A. Candiani, G. Di Benedetto, S. Ajibola, P. Cartujo, P. Romero, A. Lykoudi, M. M. De Grandis, C. Xouris, R. Lo Bianco, I. Doddy, I. Elegbede, G. F. D'Urso Labate, L. F. García del Moral and V. Martos (2024). AI can empower agriculture for global food security: challenges and prospects in developing nations. Front. Artif. Intell. 7: 1328530. https://doi.org/10.3389/frai.2024.1328530

Ahmed, S. (2023). A software framework for predicting the maize yield using modified multi-layer perceptron. Sustainability. 15(4): 3017. https://doi.org/10.3390/su15043017

Akoglu, H. (2018). User's guide to correlation coefficients. Turk. J. Emerg. Med. 18(3): 91-93. https://doi.org/10.1016/j.tjem.2018.08.001

Ansarifar, J. (2020). An explainable model for crop yield prediction: US soybean dataset. Available online at: https://github.com/ansarifar/An-Explainable-Model-for-Crop-Yield-Prediction

Basso, B. and L. Liu (2019). Seasonal crop yield forecast: methods, applications, and accuracies. Adv. Agron. 154: 201-255. https://doi.org/10.1016/bs.agron.2018.11.002

Batool, D., M. Shahbaz, H. Shahzad Asif, K. Shaukat, T.M. Alam, I.A. Hameed, Z. Ramzan, A. Waheed, H. Aljuaid and S. Luo (2022). A hybrid approach to tea crop yield prediction using simulation models and machine learning. Plants.

11(15): 1925. https://doi.org/10.3390/plants11151925

Bazrafshan, O., M. Ehteram, Z.G. Moshizi and S. Jamshidi (2022). Evaluation and uncertainty assessment of wheat yield prediction by multilayer perceptron model with Bayesian and copula Bayesian approaches. Agric. Water Manag. 273: 107881. https://doi.org/10.1016/j.agwat.2022.107881

Burdett, H. and C. Wellen (2022). Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. Precis. Agric. 23(5): 1553-1574. https://doi.org/10.1007/s11119-022-09934-6

Cai, Y., K. Guan, D. Lobell, A.B. Potgieter, S. Wang, J. Peng, T. Xu, S. Asseng, Y. Zhang, L. You et al. (2019). Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agric. For. Meteorol. 274: 144-159. https://doi.org/10.1016/ j.agrformet. 2019.03.010

Cao, J., Z. Zhang, F. Tao, L. Zhang, Y. Luo, J. Han and Z. Li (2020). Identifying the contributions of multi-source data for winter wheat yield prediction in China. Remote Sens. 12(5): 750. https://doi.org/10.3390/rs12050750

Cedric, L.S., W.Y.H. Adoni, R. Aworka, J.T. Zoueu, F.K. Mutombo, M. Krichen and C.L.M. Kimpolo (2022). Crops yield prediction based on machine learning models: case of West African countries. Smart Agric. Technol. 2: 100049. https://doi.org/10.1016/j.atech.2022.100049

Chen D., S. Wang, B. Cao, D. Cao, G. Leng, H. Li, L. Yin, L. Shan and X. Deng (2016). Genotypic variation in growth and physiological response to drought stress and re-watering reveals the critical role of recovery in drought adaptation in maize seedlings. Front. Plant Sci. 6:1241. https://doi.org/10.3389/fpls.2015.01241

Elavarasan, D., D.R. Vincent, K. Srinivasan and C.Y. Chang (2020). A hybrid CFS filter and RF-RFE wrapper-based feature extraction for enhanced agricultural crop yield prediction modeling. Agriculture. 10(9): 400. https://doi.org/10.3390/agriculture10090400

Fageria, N. K. and V. C. Baligar (2008). Ameliorating soil acidity of tropical Oxisols by liming for sustainable crop production. Advances in Agronomy. 99: 345-399. https://doi.org/10.1016/ S0065-2113(08)00407-0

Feng, P., B. Wang, D.L. Liu, C. Waters and Q. Yu (2019). Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern Australia. Agric. For. Meteorol. 275: 100–113. https://doi.org/10.1016/j.agrformet.2019.05.018

Han, J., Z. Zhang, J. Cao, Y. Luo, L. Zhang, Z. Li and J. Zhang (2020). Prediction of winter wheat yield based on multi-source data and machine learning in China. Remote Sens. 12(2): 236. https://doi.org/10.3390/rs12020236

Hao, J. and T. K. Ho (2019). Machine learning made easy: a review of scikit-learn package in Python programming language. J. Educ. Behav. Stat. 44(3): 348–361. https://doi.org/10.3102/1076998619832248

Hodson, T.O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): when to use them or not. Geosci. Model Dev. 15(14): 5481-5487. https://doi.org/10.5194/gmd-15-5481-2022

Huang, J. C., K. M. Ko, M. H. Shu and B. M. Hsu (2020). Application and comparison of several machine learning algorithms and their integration models in regression problems. Neural Comput. Appl. 32(10): 5461–5469. https://doi.org/10.1007/s00521-019-04644-5

Jabed, M. A. and M. A. A. Murad (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. Heliyon. 10(4): e25807. https://doi.org/10.1016/j.heliyon.2024.e25807

Junaid, M. and A. Gokce (2024). Global agricultural losses and their causes. Bull. Biol. Allied Sci. Res. 2024(1): 66. https://doi.org/10.54112/bbasr.v2024i1.66

Karunathilake, E., A. T. Le, S. Heo, Y. S. Chung and S. Mansoor (2023). The path to smart farming: innovations and opportunities in precision agriculture. Agriculture. 13(8): 1593. https://doi.org/10.3390/agriculture13081593

Khaki, S., L. Wang and S. V. Archontoulis (2020). A CNN-RNN framework for crop yield prediction. Front. Plant Sci. 10: 1750. https://doi.org/10.3389/ fpls.2019.01750

Liaw, A. and M. Wiener (2002). Classification and regression by randomForest. R News. 2(3): 18–22. https://doi.org/10.1021/ci034160g

Lipiec, J., R. Hatano and A. Słowińska-Jurkiewicz (2012). Effects of soil compaction on root growth and crop yield in Central and Eastern Europe. Int. Agrophysics. 26(2): 143–156. https://doi.org/10.2478/v10247-012-0021-y

Lundberg, S. M. and S. I. Lee (2017). A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30: 4765-4774.

Maimaitijiang, M., V. Sagan, P. Sidike, S. Hartling, F. Esposito and F. B. Fritschi (2020). Soybean yield prediction from UAV using multimodal data fusion and deep learning. Remote Sens.

Environ. 237: 111599. https://doi.org/10.1016/j.rse.2019.111599

Metin, A., A. Kasif and C. Catal (2023). Temporal fusion transformer-based prediction in aquaponics. J. Supercomput. 79: 14946-14970. https://doi.org/10.1007/ s11227-023-05223-9

Morales, A. and F. J. Villalobos (2023). Using machine learning for crop yield prediction in the past or the future. Front. Plant Sci. 14: 1128388. https://doi.org/10.3389/fpls.2023.1128388

NASS (2024). USDA national agricultural statistics service. USDA NASS. Available online at: https://www.nass.usda.gov/Quick_Stats/index.php

Oikonomidis, A., C. Catal and A. Kassahun (2022). Hybrid deep learning-based models for crop yield prediction. Appl. Artif. Intell. 36(1): 2031822. https://doi.org/10.1080/ 08839514. 2022.2031823

Prasad, N., N. Patel and A. Danodia (2021). Crop yield prediction in cotton for regional level using random forest approach. Spat. Inf. Res. 29(2): 195–206. https://doi.org/10.1007/s41324-020-00346-6

Rashid, M., B. S. Bari, Y. Yusup, M. A. Kamaruddin and N. Khan (2021). A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. IEEE Access 9: 63406–63439. https://doi.org/10.1109/ACCESS.2021.3075159

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 1(5): 206–215. https://doi.org/10.1038/s42256-019-0048-x

Shafi, U., R. Mumtaz, N. Iqbal, S. M. H. Zaidi, S. A. R. Zaidi, I. Hussain and Z. Mahmood (2020). A multi-modal approach for crop health mapping using low altitude remote sensing, internet of things (IoT) and machine learning. IEEE Access 8: 112708–112724. https://doi.org/10.1109/access.2020.3002948

Shahhosseini, M., R. A. Martinez-Feria, G. Hu and S. V. Archontoulis (2021). Maize yield and nitrate loss prediction with machine learning algorithms. Environmental Research Letters. 16(12):124026. https://doi.org/10.1088/1748-9326/ac2fc4

Shahhosseini, M., G. Hu and S. V. Archontoulis (2020). Forecasting corn yield with machine learning ensembles. Front. Plant Sci. 11: 527890. https://doi.org/10.3389/fpls.2020.01120

USDA (2022). United States Department of Agriculture soybeans: production by county. Available online at: https://www.nass.usdagov/Charts_and_Maps/Crops_County/sb-pr.php

Van Dijk, A.D.J., G. Kootstra, W. Kruijer and D. de Ridder (2021). Machine learning in plant science and plant breeding. iScience 24(1): 101922. https://doi.org/10.1016/j.isci.2020.101890

Van Klompenburg, T., A. Kassahun and C. Catal (2020). Crop yield prediction using machine learning: A systematic literature review. Comput. Electron. Agric. 177: 105709. https://doi.org/10.1016/j.compag.2020.105709

Wadgymar, S. M., J. E. Ogilvie, D. W. Inouye, A. E. Weis and J. T. Anderson (2018). Phenological responses to multiple environmental drivers under climate change: insights from a long-term observational study and a manipulative field experiment. New Phytol. 218(2): 517–529. https://doi.org/10.1111/nph.14961

Wakweya, R. B. (2023). Challenges and prospects of adopting climate-smart agricultural practices and technologies: Implications for food security. J. Agric. Food Res. 14: 100698. https://doi.org/10.1016/j.jafr.2023.100698

Xiong, W., M. Reynolds and Y. Xu (2022). Climate change challenges plant breeding. Curr. Opin. Plant Biol. 70: 102308. https://doi.org/10.1016/j.pbi.2022.102308

Yuan, X., S. Li, J. Chen, H. Yu, T. Yang, C. Wang, S. Huang, H. Chen and X. Ao (2024). Impacts of global climate change on agricultural production: a comprehensive review. Agronomy. 14(7): 1360. https://doi.org/10.3390/agronomy14071360

Zhou, H., J. Yang and D. Li (2023). Improving grain yield prediction through fusion of multi-temporal spectral features and agronomic trait parameters derived from UAV imagery. Front. Plant Sci. 14: 1217448. https://doi.org/10.3389/fpls.2023.1217448.