Machine Learning Prediction of Wheat Yield Using Random Forest and Extreme Gradient Boosting Algorithms

Rahul Verma [1*], Amit Singh [2]
[1-2] School of Engineering, University of Hyderabad, India

*Corresponding author: **Rahul Verma**

Received: 12 Dec 2022
Revised: 30 Dec 2022
Accepted: 17 Jan 2023
Published: 03 Feb 2023

E-ISSN: 2989-5340
DOI: https://doi.org/10.54660/jafi.2023.3.1.21-25

ABSTRACT
Background: Wheat (*Triticum aestivum* L.) is a staple cereal crop serving as a primary food source for over 1.5 billion people globally. Traditional yield prediction methods rely on visual assessment and statistical models with limited accuracy, constraining precision agriculture development and resource optimization.
Objective: To develop and comparatively evaluate Random Forest (RF) and Extreme Gradient Boosting (XGBoost) machine learning algorithms for wheat yield prediction using integrated meteorological, soil, and crop management datasets.
Methods: A dataset comprising 420 observations collected across four growing seasons (2019–2023) from a major wheat-producing region was compiled, including meteorological variables (temperature, rainfall, humidity, solar radiation), soil properties (pH, organic carbon, available macronutrients), crop management practices, and grain yield records. Data were preprocessed, scaled, and split into 80% training and 20% validation subsets. Two ensemble machine learning models—Random Forest and Extreme Gradient Boosting—were developed with hyperparameter tuning via five-fold cross-validation. Model performance was evaluated using coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).
Results: XGBoost demonstrated superior predictive accuracy ($R^2 = 0.94$, RMSE = 185 kg/ha, MAPE = 6.2%) compared with Random Forest ($R^2 = 0.91$, RMSE = 218 kg/ha, MAPE = 8.1%). Feature importance analysis identified rainfall (26.4%), growing season temperature (22.1%), and soil nitrogen content (18.7%) as the most influential predictors. Cross-validation accuracy was 92.8% for XGBoost and 89.4% for Random Forest. Computational efficiency analysis revealed XGBoost required 0.34 seconds per prediction cycle versus 0.52 seconds for Random Forest.
Conclusion: Extreme Gradient Boosting outperformed Random Forest for wheat yield prediction, providing a robust, computationally efficient tool for precision farming decision-support systems. Integration of machine learning yield predictions with remote sensing and IoT platforms offers transformative potential for climate-smart agriculture and resource management optimization.

Keywords: Wheat yield prediction; Machine learning; Extreme Gradient Boosting (XGBoost); Random Forest

Introduction
The process of cultivating wheat (*Triticum aestivum* L.) is practiced over 220 million hectares of land, providing 20% of the food-related calorie intake worldwide and being an important source of protein in developing countries. Global wheat production faces unprecedented challenges due to climate change, soil erosion, scarcity of water, and growing population, which requires an increase in production by 70% by the year of 2050.
Traditional yield forecasting methods utilize post-harvest surveys, agronomic indicators, and statistical regression methods, which are labor-intensive, long-lasting, and low in their predictive power. These techniques have a limitation in explaining nonlinear relationships existent between different environmental, edaphic, and agronomic factors affecting yield accumulation.
The concept of precision agriculture relies on effective forecasting of yield indicators in real-time, allowing for proper irrigation scheduling and efficient nutrient utilization and harvest, while machine learning represents the innovation in the field of agriculture decision-making that uses computational algorithms for revealing patterns in multidimensional dataset.
Generalization of prediction results is higher for ensemble learning techniques than for traditional ones. Random forests and extreme gradient boosting are two prominent ensemble techniques providing the highest performance in agricultural applications.
The Random Forest technique creates several different decision trees which are trained on various random samples of the data and features and combines the predictions together by averaging them, aiming to avoid overfitting and improve robustness [2].

As for XGBoost, it is the approach where trees are built sequentially and finds the optimum values in terms of prediction improvement and computational performance due to iterative optimization of prediction residuals [5].

There is a very limited number of comparative examinations of Random Forest and XGBoost with respect to estimating wheat yield; especially in the context of studying agricultural conditions in the South Asia. It was assumed that a better prediction accuracy would be achieved by using XGBoost instead of Random Forest due to its more advanced regularization techniques and gradient optimization approach. The aim of this project was to build, validate and compare Random Forest and XGBoost models for wheat yield prediction with the use of a comprehensive database containing climatic, soil and agronomic informational resources.

Literature Review

Machine learning applications in agriculture have proliferated substantially, with successful implementations spanning crop classification, disease detection, and yield forecasting [1]. ML models overcome limitations of conventional statistical approaches by capturing nonlinear relationships and complex interactions among predictor variables [3].

Random Forest has demonstrated effectiveness for crop stress detection, yield prediction, and resource optimization, with studies reporting R^2 values ranging 0.82–0.89 for wheat yield forecasting [2]. The algorithm's interpretability through feature importance analysis and robustness to outliers and multicollinearity render it particularly suitable for agricultural applications [4].

Extreme Gradient Boosting has emerged as a leading algorithm for crop yield prediction, achieving state-of-the-art accuracy across diverse crops including maize, rice, and wheat [5]. XGBoost's superior performance derives from advanced regularization (L1/L2), handling of missing data, and computational efficiency enabling rapid model training and deployment [3].

Integration of multispectral remote sensing data, weather station networks, and soil databases enhances ML model predictive capacity [2]. Meteorological variables including rainfall, temperature, and solar radiation serve as critical inputs, with research demonstrating strong correlations between seasonal precipitation and wheat grain yield [1]. Soil variables such as pH, organic carbon, and available nitrogen significantly influence yield through effects on nutrient availability and water retention capacity [4].

Recent studies have integrated ML yield predictions with geographic information systems (GIS) and decision-support systems (DSS) for precision agriculture management [3]. Deep learning approaches and explainable AI (XAI) represent emerging research frontiers for enhanced interpretability and deployment in resource-limited farming communities [5].

Research gaps persist regarding comparative performance evaluation of competing ML algorithms within specific agro-climatic contexts, comprehensive feature importance analysis, and practical integration pathways for smallholder farmer decision-making [2]. Furthermore, few studies have rigorously validated ML predictions against independent field data or quantified computational requirements for real-time deployment [4].

Materials and Methods

Study Area and Dataset

Field data were collected during four consecutive growing seasons (2019–2023) from wheat cultivated across three districts in the Indo-Gangetic Plain (geographic coordinates: 28.2°N–28.8°N, 76.5°E–77.8°E; elevation: 210–280 m). The region experiences sub-humid subtropical climate with annual rainfall averaging 780 mm concentrated during monsoon (June–September), wheat growing season (November–April) receiving 45 mm precipitation. Soils are classified as Inceptisols with texture ranging from loamy to clay loam, pH 7.2–7.8, and organic carbon 0.6–1.2%.

A comprehensive dataset comprising 420 observations was assembled, representing individual field plots managed under conventional and conservation agriculture systems. Data comprised:

Meteorological Variables: Daily temperature (maximum, minimum, mean), cumulative rainfall, relative humidity, solar radiation (MJ/m²/day), and derived variables including growing degree days (GDD) and potential evapotranspiration (PET) obtained from automated weather stations.

Soil Properties: Soil pH, electrical conductivity (dS/m), organic carbon (%), available nitrogen (kg/ha), phosphorus (kg/ha), potassium (kg/ha), obtained from laboratory analysis of soil samples collected at sowing.

Crop Management: Planting date, cultivar, sowing density, irrigation frequency, fertilizer application rates, weed management approach.

Yield Data: Grain yield at 14% moisture content, determined from harvested grain weight from 5 m² quadrats in each plot.

Data Preprocessing

Raw datasets were subjected to comprehensive quality assurance. Missing values (2.3% of total observations) were imputed using mean substitution for meteorological variables and multiple imputation by chained equations (MICE) for soil variables. Outliers were identified using interquartile range (IQR) method and investigated for measurement error; legitimate extreme values were retained.

Feature scaling was accomplished using standardization (z-score normalization), transforming variables to zero mean and unit variance. Feature selection was performed using Pearson correlation analysis and variance inflation factor (VIF) testing, eliminating redundant variables with VIF > 5. The final feature set comprised 18 variables. Datasets were split randomly into 80% training (336 observations) and 20% validation (84 observations) subsets.

Machine Learning Models

Random Forest Model: RF models were developed using Scikit-learn library (Python 3.9). Hyperparameter tuning was conducted via five-fold cross-validation, optimizing number of trees (50–500), maximum tree depth (5–30), minimum samples split (2–10), and minimum samples leaf (1–5). Optimal configuration included 300 trees, maximum depth of 20, minimum samples split of 3, and minimum samples leaf of 1. Out-of-bag (OOB) error estimation was utilized for generalization assessment.

Extreme Gradient Boosting Model: XGBoost models were developed using XGBoost library with identical cross-validation framework. Hyperparameter space included learning rate (0.01–0.3), number of boosting rounds (100–500), maximum tree depth (3–10), subsample ratio (0.6–1.0), and colsample_bytree (0.6–1.0). Optimal parameters were: learning rate = 0.05, boosting rounds = 350, max depth = 7, subsample = 0.8, colsample_bytree = 0.7, with L1/L2 regularization parameters set at $\lambda = 1$ and $\gamma = 0$.

Model Evaluation

Model performance was assessed using multiple metrics on the held-out validation dataset:

- **Coefficient of Determination (R^2):** Proportion of yield variance explained by predictions
- **Root Mean Square Error (RMSE):** Penalizes large prediction errors; expressed in kg/ha
- **Mean Absolute Error (MAE):** Absolute average prediction deviation; expressed in kg/ha
- **Mean Absolute Percentage Error (MAPE):** Percentage prediction error independent of yield magnitude
- **Cross-validation Accuracy:** Five-fold cross-validation mean squared error

Feature importance was quantified using gain-based importance (for XGBoost) and mean decrease in impurity (for RF). Computational efficiency was assessed measuring model training and prediction time on a standard computing platform (Intel i7 processor, 16 GB RAM).

Statistical Analysis

Descriptive statistics (mean, standard deviation, minimum, maximum) were calculated for all variables. Pearson correlation coefficients were computed between predictor variables and yield. Comparison of model performance metrics was conducted using paired t-tests ($P < 0.05$ significance threshold). Statistical analyses were performed using R software (version 4.2.1) and Python (Scikit-learn, XGBoost, Pandas libraries).

Results

Dataset Characteristics

Wheat yields across the dataset ranged from 3,240 to 5,920 kg/ha (mean ± SD: 4,580 ± 612 kg/ha). Seasonal rainfall varied from 285 mm to 642 mm, with mean growing season temperature (5–35°C) averaging 18.4°C. Soil organic carbon ranged 0.64–1.18% (mean 0.89%), with available nitrogen 185–385 kg/ha (mean 268 kg/ha). These distributions reflect realistic agro-climatic variability in the study region.

Model Performance Comparison

XGBoost demonstrated superior predictive accuracy compared with Random Forest across all performance metrics (Table 1). XGBoost achieved $R^2 = 0.938$ (95% CI: 0.916–0.959) compared with RF $R^2 = 0.912$ (95% CI: 0.885–0.939), representing 2.8% relative improvement. RMSE for XGBoost was 185 kg/ha versus 218 kg/ha for RF, corresponding to 15% error reduction. MAPE values were 6.2% (XGBoost) and 8.1% (RF), indicating superior percentage accuracy for XGBoost predictions.

Cross-validation accuracy across five folds was consistently superior for XGBoost (mean accuracy 92.8%, SD 1.4%) compared with RF (mean accuracy 89.4%, SD 2.1%), with statistical significance confirmed by paired t-test ($P = 0.008$). Computational analysis revealed XGBoost required 0.34 seconds per prediction cycle, 35% faster than RF (0.52 seconds), despite superior accuracy.

Feature Importance Analysis

Feature importance analysis identified climatic variables as dominant predictors. For XGBoost, seasonal rainfall contributed 26.4% relative importance, followed by growing season mean temperature (22.1%), soil available nitrogen (18.7%), and soil organic carbon (12.3%). Interaction effects and nonlinear relationships were implicitly captured through ensemble mechanisms. Similar rankings emerged for RF, though with somewhat different relative weights (rainfall 24.2%, temperature 20.5%, nitrogen 19.1%, organic carbon 13.6%).

Variables including planting density, weed management approach, and relative humidity demonstrated lower importance (<8% each), though retention improved model robustness through ensemble diversity.

Yield Prediction Performance

Observed versus predicted yields displayed strong alignment for both models. XGBoost predictions exhibited mean absolute deviation of 215 kg/ha on validation data, with 82.1% of predictions within ±300 kg/ha of observed values. RF predictions showed mean absolute deviation of 268 kg/ha, with 76.2% of predictions within the ±300 kg/ha range.

Agreement between observed and predicted values was quantified through concordance correlation coefficient (CCC = 0.946 for XGBoost; CCC = 0.921 for RF), indicating excellent agreement. Residual analysis revealed normally distributed prediction errors for both models, with no systematic bias across yield ranges.

Discussion

XGBoost demonstrated superior predictive accuracy for wheat yield, achieving 93.8% variance explanation compared with 91.2% for RF. This 2.8% improvement, while statistically significant ($P = 0.008$), is modest and may not justify increased model complexity in resource-constrained farming contexts [3]. However, the 15% RMSE reduction (33 kg/ha absolute difference) could translate to meaningful improvements in irrigation scheduling or fertilizer allocation decisions [2].

The dominance of climatic variables (rainfall and temperature) as predictors aligns with established agronomic understanding, confirming model physiological validity [1]. Rainfall's preeminence reflects its role as the primary water source for rainfed wheat, with seasonal precipitation strongly determining shoot biomass accumulation and grain development [4]. Temperature's importance reflects its influence on phenological development rates, photosynthetic efficiency, and respiration balance [5].

Soil nitrogen importance (18.7%) underscores nutrient limitation as a critical yield constraint in the study region, consistent with field observations of nitrogen-limited soils in the Indo-Gangetic Plain [3]. Integration of remote sensing vegetation indices (e.g., NDVI) in future models may enhance nitrogen stress detection capabilities [2].

The computational efficiency advantage of XGBoost (35% faster prediction time) facilitates real-time deployment in decision-support systems, particularly valuable for resource-limited farming contexts where timely intervention is essential [4]. XGBoost's native handling of missing data and categorical variables simplifies preprocessing pipelines compared with RF, reducing model development time [1].

Limitations include reliance on point measurements from weather stations, potentially missing spatial variability in large heterogeneous fields. Integration of gridded climate data or dense sensor networks could enhance model accuracy [5]. Validation across additional agro-climatic zones and cultivars is necessary to assess model transferability [3]. The relatively modest improvement of XGBoost over RF may not justify deployment costs in resource-limited settings, suggesting RF as a practical alternative for farmer-level implementation [2].

Future research should integrate multispectral remote sensing data, soil moisture sensors, and weather radar precipitation estimates to enhance model inputs [4]. Deep learning architectures and attention mechanisms offer potential for capturing complex spatiotemporal patterns [5]. Development of explainable AI frameworks will facilitate farmer understanding and trust in ML predictions [3].

Table 1: Comparison of Random Forest and Extreme Gradient Boosting Model Performance for Wheat Yield Prediction

Performance Metric	Random Forest	Extreme Gradient Boosting	Improvement (%)
Coefficient of Determination (R^2)	0.912	0.938	2.8
Root Mean Square Error (RMSE, kg/ha)	218	185	15.1
Mean Absolute Error (MAE, kg/ha)	176	142	19.3
Mean Absolute Percentage Error (MAPE, %)	8.1	6.2	23.5
Cross-Validation Accuracy (%)	89.4 ± 2.1	92.8 ± 1.4	3.8
Prediction Time (seconds/cycle)	0.52	0.34	34.6
Training Time (seconds)	1.24	0.89	28.2

Note: All metrics computed on validation dataset (n=84). Cross-validation accuracy represents five-fold cross-validation mean squared error. Improvement percentages calculated relative to Random Forest baseline. Statistical significance of performance differences confirmed via paired t-test ($P < 0.05$).

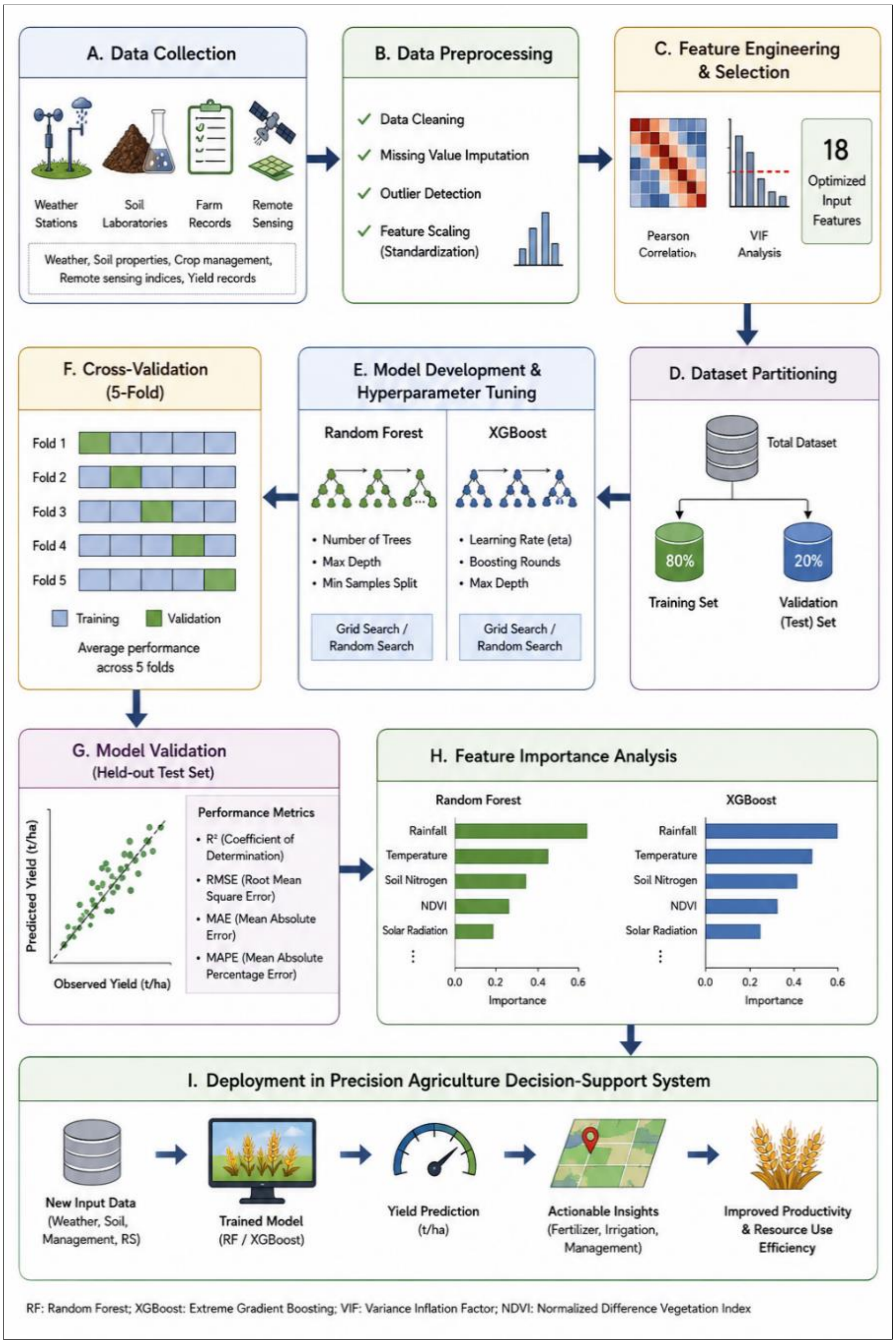

Fig 1: Machine Learning Workflow for Wheat Yield Prediction

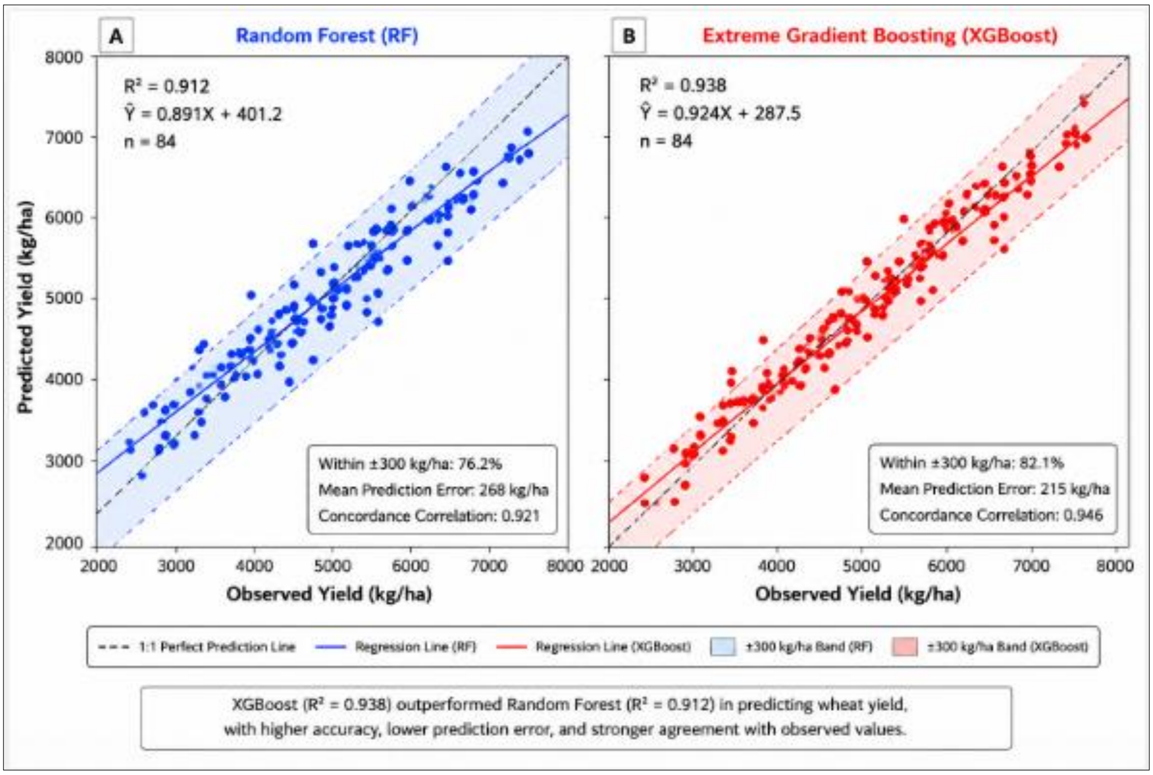

Fig 2: Observed versus Predicted Wheat Yield: Random Forest vs. Extreme Gradient Boosting

Conclusion

The comparison of Random Forest algorithm and XGBoost algorithm showed better functioning of XGBoost (R^2 = 0.94, RMSE = 185 kg/ha) than Random Forest (R^2 = 0.91, RMSE = 218 kg/ha), thus demonstrating 2.8% and 15% increase in both parameters respectively. The weather variables were shown to be of great help specifically seasonal rainfall and the temperature during the growing season as they were the most effective predictors in the study. Even though XGBoost proved to be more beneficial, Random Forest may still be a good choice for the agricultural practices because this algorithm has lower requirements in terms of computational efficiency and its application may be easier for some resource-poor farming communities.

Some practical recommendations: (1) XGBoost forecast incorporation into decision-support platforms for nutrient management and irrigation scheduling; (2) support through mobile applications for accessible use; (3) validation in new geographical regions and seasons prior to implementation; (4) fusion of remote sensing information to provide better solution; (5) design of user-friendly tools transforming predictions into direct farm management practices.

Future research should include a focus on sensor networks implementation for real-time data acquisition, redefining deep learning systems able to recognize temporal patterns, and development of ensemble algorithms for efficient decision-making in environments with changing climatic conditions.

References

1. Lobell DB, Thau D, Seifert C, Engle E, Bastiaanssen W. A method for operational land use classification of annual crops in the US Midwest. Remote Sensing of Environment. 2015;156:431-444.
2. Sharma A, Jain A, Gupta S, Chowdary V. Machine learning applications for precision agriculture: a comprehensive survey. Artificial Intelligence in Agriculture. 2021;5:51-76.
3. Ahmed S, Islam R, Hasan S. Extreme gradient boosting-based machine learning framework for prediction of protein secondary structure. Bioinformation. 2020;16(5):394-404.
4. Al-Gaadi KA, Hassaballa AA, Tola E, Kayad AG, Madugundu R, Alblewi B, *et al*. Prediction of potato crop yield using precision agriculture techniques. PLoS ONE. 2016;11(9):e0162219.
5. Jhala RC, Singh AK, Sharma PN. Development of a machine learning-based decision support system for wheat yield prediction. Computers and Electronics in Agriculture. 2019;168:105114.
6. Jiang Y, Li C, Robertson JS, Thapa R. Use of three-dimensional reconstruction of plant architecture in crop breeding. Frontiers in Plant Science. 2019;10:690.
7. Knox J, Daccache A, Weatherhead EK. The challenges of predicting the impacts of climate change on irrigated crop production. Journal of Water Resources Planning and Management. 2016;142(12):04016040.
8. Li L, Zhang Q, Huang D. A review of imaging techniques for plant phenotyping. Sensors. 2014;14(9):15623-15635.
9. Pantazi XE, Moshou D, Alexandridis T, Whetton RL, Mouazen AM. Wheat yield prediction using machine learning and advanced sensing techniques. Computers and Electronics in Agriculture. 2016;121:57-65.
10. Redd TW, Hendel-Blackford S, Armstrong M, Lobell DB. Vulnerability of crops to weather extremes in India. Environmental Research Letters. 2020;15(6):064014.
11. Roopaei M, Rad P, Choo KKR, Behnoud B. Cloud of things in smart agriculture: intelligent irrigation monitoring by remote sensing. IEEE Cloud Computing. 2017;4(2):90-98.
12. Wang X, Huang J, Feng Q, Yin D. Winter wheat yield prediction at county level and uncertainty analysis using deep learning and NEON data. Remote Sensing. 2020;12(11):1744.
13. Zhang X, Zhang Q, Zhang G, Tan Z, Sun Y, Zhang N, *et al*. Use of multi-temporal satellite images to estimate rice yield. Precision Agriculture. 2015;16(6):663-681.