Hindawi
Scientifica
Volume 2022, Article ID 7271293, 11 pages
https://doi.org/10.1155/2022/7271293

Research Article

A Prediction Model Optimization Critiques through Centroid Clustering by Reducing the Sample Size, Integrating Statistical and Machine Learning Techniques for Wheat Productivity

Muhammad Islam and Farrukh Shehzad

Department of Statistics, The Islamia University of Bahawalpur, Bahawalpur, Pakistan

Correspondence should be addressed to Muhammad Islam; mislam6667@gmail.com

Received 20 August 2021; Accepted 12 January 2022; Published 11 March 2022

Academic Editor: Mehdi Rahimi

Copyright © 2022 Muhammad Islam and Farrukh Shehzad. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Machine learning algorithms are rapidly deploying and have made manifold breakthroughs in various fields. The optimization of algorithms got abundant attention of researchers being a core component for deploying the machine learning model (MLM) abled to learn the parameters in significant ways for the given data. Modeling crop productivity through innumerable agronomical constraints has become a crucial task for evolving sustainable agricultural policies. The cross-sectional datasets of 26430 (D1) crop-cut experiments are taken by 2nd-stage area frame sampling, collected from crop reporting service. This research is taken as follows: firstly three more effective numerical optimized datasets are generated (D1, D2, and D3) from D1 by taking the centroid points of features which decrease the sample size; secondly MLM is integrated with the traditional statistical models (TSMs) for multiple linear regression (MLR), and thirdly decision tree regression (DTR) and random forest regression (RFR) are deployed to get the optimized models able to predict the wheat productivity well with 75% datasets to train and 25% to test the model using the evaluation metrics (R^2, RMSE), information criterion (AIC) with weights (AIC_W), evidence ration (E.R), and decompositions of prediction error. The MLR outperformed for MLM than TSM. The performance capability of MLM and TSM got upswing for generated datasets. RFR got optimized and superperformed for D1, D2, D3, and D4. This study demonstrated strong evidences for deploying MLM for prediction of wheat productivity as an alternative of traditional statistical modeling.

1. Introduction

1.1. Significances, Motivations, and Objectives of the Study. Producing enough food for evolving population explosion has become the major concerns for the global world. Agriculture in aspect of core contributor in food production is ensuring to meet the sustainable food availability [1]. Food security has been considered as the foremost global threat, and therefore, it is essential to steer strategies to determine policies for future food security and sustainable food availability [2, 3]. Food and agricultural organization, international food policy research institute, and many other international organizations deem their great concerns on this converted threat to attain sustainable food availability [4–6]. Modeling crop productivity through innumerable agronomical constraints has become a crucial task to attain sustainable agriculture and for evolving effective agricultural strategies [7]. A precise crop model based on certain conditions is a foremost need of time to evoke to handle the prevailing food trepidations [8, 9]. Wheat being a 3rd largest food crop is playing a vital role for assuring the food supply in the world [4, 10–12]. Developing food prediction models, capable for true estimation of food availability, can assure veracious policy decision for managing national action plans for food security [13]. Pakistan stood 6th for wheat production, 8th for cultivated area under wheat crop, and 59th for wheat productivity [14]. Its exigent need of era is to develop accurate and precise wheat productivity model capable to predict the production on the reliable statistics which would help us to attain the assurance or nonassurance

of future food demand [15]. Islam et al. [2] presented the study on the large datasets for building the statistical prediction model for the wheat productivity in Pakistan using hierarchical regression approach for selecting the features to address food security threat for the global concerns based on cross-sectional record. This study presented the tradition statistical modeling and introduced the theory of centroid clustering used to generate the three more datasets from the original datasets. Generated datasets enhanced the model prediction capability with the reduction of sample size. They applied different evaluation metrics, adjusted R^2, ΔR^2, MSE, and information criterion approaches such as Akaike information criteria (AIC), Schwarz information criterion (SIC), and weighted information criterion (Akaike weight "Wi") with evidence ratio "E.R," etc. The normality analysis and constant error variance are done by graphical presentation. The VIF is applied for multicollinearity, and nonconstant error variance is checked by Breusch–Pagan test which is developed in 1979 by Trevor Breusch and Adrian Pagan. The reliably analysis is performed by Cronbach's alpha test.

Machine learning algorithms widely develop and deploy rapidly and have made manifold breakthroughs in various fields. The advancement in science, technologies, and implementations of innumerable agronomical constraints in various fields of agriculture leads to immense volume of data [1, 8, 16]. The optimization of algorithms has become a significant part of machine learning and got abundant attention of researchers, and significance proficiency of numerical optimized algorithms of datasets affectedly influenced the machine learning model performance capability for the massive amount of data [17]. In this research, firstly the effective numerical optimized datasets are developed by taking the centroid points of features abled to enhance the machine learning model performance by decreasing the sample size, secondly machine learning models are integrated with the traditional statistical models, and thirdly different machine learning models are deployed to get the optimized models able to predict the wheat productivity well. This study designed to apply the supervise machine learning techniques, i.e., multiple linear regression model (MLRM), decision tree regression model (DTRM), and ensemble learning random forest regression model (RFRM) on the same datasets with the aim to enhance the model performance by reducing the sample size through centroid clustering. This study integrates the efficacies of machine learning algorithms with benchmark traditional statistical models for wheat productivity.

2. Material and Methods

2.1. Data Collection, Sampling Method, and Important Features Selection. Punjab is the 2nd largest province of Pakistan which accounted 76% share in total wheat cultivation area. The administrative setup of Punjab comprises upon nine divisions, thirty-six districts, and one hundred and forty-five tehsils. The 26,430 field of wheat crop-cut experiments (C.C.E) is taken from crop reporting service (CRS), Punjab, for the year comprised from 2016-17 to 2019-2020. The list frame sampling (LFS) technique using systematic random sampling (S_yRS) in which complete village (sample unit) was selected as basic unit was remained in practice in CRS, but after 2018-19, 2nd-stage area frame sampling (AFS) is applied to select the sample for C.C.E [18].

$$\mathrm{AFS}(P_i) = \frac{Z_i}{\sum_{i=1}^{N} Z_i}, \tag{1}$$

where Z_i = cropped area of i^{th} village in j^{th} union councils of district, $\sum_{i=1}^{N} Z_i$ = total crop area of village in j^{th} union councils of district, and P_i = probability of selecting the i^{th} village as sample. Qayyum and Shera [18] reported, at stage I, union councils are considered as population and village as sampling units using probability proportion to size (PPS), while at stage II, the selected sample village is considered as population and the land segment area is considered as sampling unit using the simple random sampling (SRS) techniques. The C.C.E is selected in land area segments. The wheat productivity with measuring scale munds/acre along with seven agronomical quantitative variables, i.e., fertilizer urea kg/acre, fertilizer DAP kg/acre, other fertilizers kg/acre, no. of water, seed quantity used kg/acre, no. of pest spry, no. of weeds spry, and eight binary categorical (0 for absence and 1 for presence) agronomical features, i.e., seed treatment, soil-type chikny loom, varieties adoption, harvest period April (1-20), planting in November, land irrigated, farmers' area >25 acres, and seed type, is used in the current study. Experiment is performed using *Python*'s key library called scikit-learn (Sklearn) by Jupyter Notebook as https://scikit-learn.org/stable/supervised_learning.html. Sklearn offers various prominent features for data processing, classification, clustering evaluation, and model selection. Model_selection is Sklearn method used for setting to analyze datasets and then using it on unseen datasets for evaluation purpose.

2.2. Supervised Machine Learning Technique. Machine learning is viewed as innovative extension of statistics capable of dealing with the massive datasets by adding the methods from computer science to the repertoire of statistics [19]. Machine learning is categorized as advanced tools applied for the prediction of agricultural production [20–23]. According to Jeong et al. [9], machine learning used latest process-based techniques as an alternative to traditional statistical modeling. Machine learning is viewed as assumption-free methods for correct data structure of model, and it is applied in complex projection concerns, i.e., function form for crop yield prediction [8, 24]. Arthur Samuel (1901–1990), a pioneer in artificial intelligence (AI), coined the term machine learning in 1959 as "Field of study that gives computers the capability to learn without being explicitly programmed" [25, 26]. The prominent layout of machine learning process is narrated as follows:

(i) Data gathering
(ii) Data preparations
(iii) Selection of machine learning model

(iv) Data partitions into train and test split datasets

(v) Model evaluations for train model and for test model

(vi) Hyperparametric tuning of machine learning models

(vii) Deployment of ML model or prediction

2.2.1. Multiple Linear Regression Models (MLRMs). MLR is used to endeavor the relationship of feature with wheat productivity for prediction for both statistical and machine learning modeling as $Y_i = X_i\beta_i + \varepsilon$, where Y_i = wheat productivity munds/acre, X_j = features, and β_j = features coefficients.

2.2.2. Decision Tree Regression Model (DTRM). The decision tree regression model (DTRM) used the flowchart structure to predict the response. DTR-built internal node signifies a test, branches signify the outcome for test, and each leaf node signifies the final decision [27, 28]. In contemporary speech, leaf nodes reproduce the outcomes of prediction after getting hierarchal representation of leaf and branch structure for root-to-leaf direction. DTRM with depths ranging from 1 to 20 is plotted for training and test performance to determine the optimum DTRM capable to predict the wheat productivity well. The cross-sectional hyperparametric tuning is exercised using the GridSearchCV. The GridSearchCV is scikit-learn library applied to find out the optimum number for min_sample_split and max_depth (tree depth). Figure 1 shows the structural flow for the decision tree model.

2.2.3. Random Forest Regression Model (RFRM). The RFRM almost consists of the same set of hyperparameter tuning as DTRM except random forest (RF). RF used the additional randomness for the predicted made while growing the regression trees instead of pointing the important features to split the node. RFRM searches the best set of features and averaged multiple regression decision trees to avoid overfitting problem, and parameter no. of trees (n_sample) in the forest has been used which ranged from 10–100 [29, 30]. RFRM used precision to build up the forest random and search the best feature [31]. RFRM uses bootstrap aggregating for agricultural decision related to crop productivity prediction [21, 30, 31]. Figure 2 depicts the structural flow for RFR.

2.3. Preparation of Datasets. Data preprocessing is a technique used as a branch of data miming applied to search out accurate dataset from large dataset-based identifying, classification, clustering, and regression [32–34]. Three new datasets are generated from original 26,430 C.C.E by data preprocessing using centroid point clustering to increase the prediction interpretability and capability of models by reducing the sample size based at villages, tehsils, and district-level datasets [2].

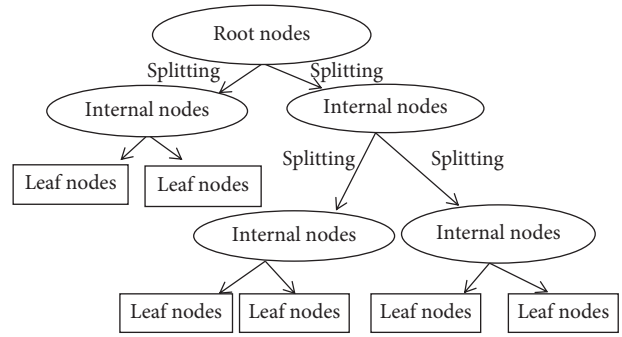

Figure 1: Structural flow of decision tree regression.

$$\overline{x}_{i_{c_m}} = \frac{\sum_{j=1}^{N_{jn_m}} x_{i_{c_m}}}{N_{jn_m}},$$
$$\overline{q}_{i_{c_m}} = \frac{\sum_{j=1}^{N_{jn_m}} q_{i_{c_m}}}{N_{jn_m}}. \quad (2)$$

For 1st subsets, $i = 1, 2, \ldots, 7$ (quantitative variables), $j = 1, 2, \ldots, N_{jn_m}$ (j^{th} observation of i^{th} predictors in m^{th} cluster ($m = 1, 2, 3$)), N_{jn_m} = total no. of j^{th} observation of i^{th} predictor in m^{th} cluster, and $\overline{x}_{i_{c_m}}$ = average of the i^{th} quantitative variable in m^{th} cluster. For 2nd subsets, $i = 1, 2, \ldots, 8$ (categorical binary variables), $j = 1, 2, \ldots, N_{jn_m}$ (j^{th} observation of the presence of i^{th} binary variable in m^{th} cluster ($m = 1, 2, 3$)), N_{jn_m} = total no. of j^{th} observation of i^{th} binary variable in m^{th} cluster, and $\overline{q}_{i_{c_m}}$ = proportion of the i^{th} binary predictor in m^{th} cluster. The original datasets (D_1) comprise 26,430 rows/records/samples points of features, and the following three datasets are generated.

(i) Cluster-1 (D_2) comprises 6034 rows/records/sample points taken by village centroid point of features

(ii) Cluster-2 (D_3) comprises 145 rows/records/sample points taken by tehsils centroid point of features

(iii) Cluster-3 (D_4) comprises 36 rows/records/sample points taken by district centroid point of features

2.4. Data Partition. Sklearn provides a way to generate accurate results abled to make true prediction, and for that, it is needed to train your model using train datasets and then test on unseen datasets using Sklearn train_test_split function. The train_test_split function is used for splitting a single dataset into two different subsets using random partitions called training subsets and testing subsets. The training subset is used to learn or to build model, and testing subset is used to evaluate the model performance for unseen datasets. For the current study, data partition is carried out using randomization train-test split and capability performance of models is investigated based on four types of datasets taking the 75% data as training subsets and 25% dataset for testing/validation subsets as follows.

(i) D_1 consists of 19822 sample points as training subsets and 6608 subsets as testing subsets

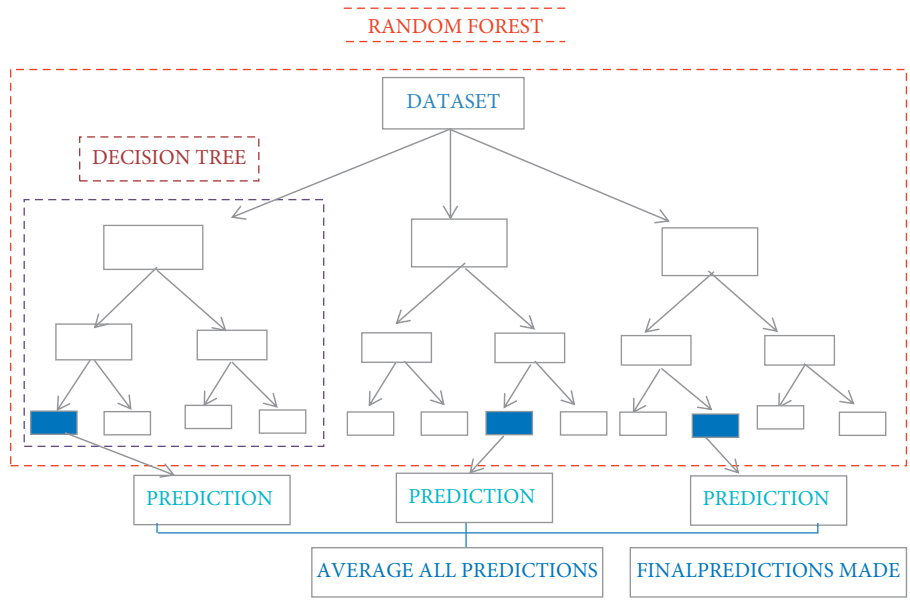

Figure 2: Structural flow of random forest regression.

(ii) Cluster-1(D_2) uses 4525 sample points as training and 1509 for testing subsets

(iii) Cluster-2 (D_3) uses 108 sample points as training and 37 for testing subsets

(iv) Cluster-3(D_4) uses 27 sample points as training subsets and 09 for testing subsets

2.5. Hyperparametric Tuning of Machine Learning Models. While applying the machine learning algorithms to predict the response variable (wheat productivity), the datasets split into two parts named training and testing datasets (Section 2.4). Two types of error are reported in prediction of response using machine algorithms [35], the error reported during training phase is called training error or bias, and this error is measured from overall observed data samples in the training phase, while the out-of-sample error (generalization error) measures the expected error on testing phase or in unseen datasets called variance. Both the underfit (high bias and high variance) and overfit (low bias and high variance) algorithms mislead the machine learning model prediction capability, and the bias-variance trade-off is common property in application of machine learning model building. The decomposition of prediction error is comprised as the sum of three components, bias, variance, and irreducible error [25, 36]. The mathematical illustration of bias and variance is presented as the target variable (wheat yield) is going to be predicted by machine learning model taking the covariates (15 features) by the relation as $y = g(x) + e$ where "e" is supposed to be the error term fallow normality. Using machine learning modeling technique, the estimated model of $g(x)$ is $\widehat{g}(x)$ and the expected squared prediction error at "x" is found as follows:

$$P.\mathrm{E}(x) = E\left[y - g(x)^2\right]. \tag{3}$$

Prediction error is decomposed into categories as bias and variance components as follows:

$$\begin{aligned} E[y - \widehat{g}(x)]^2 &= [E\{\widehat{g}(x)\} - g(x)]^2 + E\left[\{\widehat{g}(x) - E[\widehat{g}(x)]\}^2\right] + \sigma_e, \\ P.\mathrm{E}(x) &= \mathrm{Var}[\widehat{g}(x)] + [\mathrm{Bias}\{\widehat{g}(x)\}]^2 + \mathrm{Var}(e), \\ \text{prediction error} &= \text{variance} + \text{bias}^2 \\ &\quad + \text{irreducible error term}. \end{aligned} \tag{4}$$

That irreducible error term may be known as noise term which exists in the true relationship between the feature and response in model prediction and in machine learning model; the aim is to decrease both the bias and variance terms. However, in machine learning model prediction, there exists a bias-variance trade-off and the optimum model complexity means a situation where the model predicted well with low variance and low bias and is free from overfit and underfit model [37]. Figure 3 elaborates the condition of overfitting and underfitting at lower and higher model complexity, while at ideal range of model complexity, the MLM predicted well.

2.6. Evaluation Metrics and Information Criterion. The evaluation metrics using the performance score ($R2$) and root mean square error (RMSE) are applied to measure the accuracies of regression models. Lower the value of RMSE and higher the performance score lead to support the good fit.

$$\begin{aligned} \mathrm{RMSE} &= \sqrt{\frac{\sum_i^n (y_i - \overline{y}_i)^2}{n}}, \\ R^2 &= \frac{\sum (\widehat{Y} - \overline{Y})^2}{\sum (Y - \overline{Y})^2}. \end{aligned} \tag{5}$$

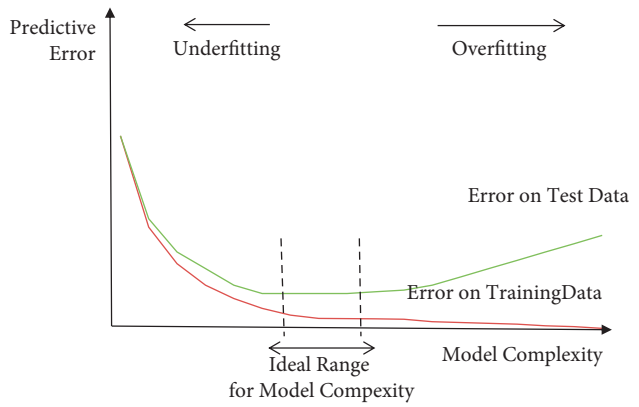

Figure 3: Structure of MLM complexity for over- and underfitting.

2.6.1. Akaike Information Criterion, AIC Weights, Evidence Ratio, and Reliability Analysis. The Akaike information criterion (AIC) using the log-likelihood functions with simple penalties is applied to determine the theoretical and logical relevance of the predictors to the response and their statistical significance in model. Lower the value of AIC leads to conclude that the fitted regression model is good [38–40].

$$\begin{aligned} \mathrm{AIC} &= e^{2k/n}\frac{\sum \widehat{u}_i^2}{n} \\ &= e^{2k/n}\frac{\mathrm{RSS}}{n}, \end{aligned} \tag{6}$$

where k = no. of features and intercept, n = sample size, and $2k/n$ = penalty factor.

One of the key objectives of driving the AIC is to determine the range of models with their relative AIC value. For comparing the multiple models, we can measure how much better the best candidate model is to be compared with next best models, and the easiest way to determine the comparison is to measure the change in of AIC values for the best model with the i^{th} other models $\Delta\mathrm{AIC}_i = \mathrm{AIC}_i - \mathrm{AIC}_{\min}$. $\Delta\mathrm{AIC}_i$ is also used to measure the relative strengths of best models with other models. $\Delta\mathrm{AIC}_i$ is used to determine the level of empirical support of model comparisons for quick strength of evidence, and lower the difference leads to support the model. Burnham and Anderson [41] defined the evidence ratio "E.R" used to compare the efficiencies of various models and depicted the measure of how much more likely the best model is than other models [42].

$$\mathrm{ER} = \frac{\exp\left(-(1/2)\Delta_{\mathrm{best}}\right)}{\exp\left(-(1/2)\Delta_i\right)} = \frac{\mathrm{Weight}_{\mathrm{best}}(\mathrm{AIC})}{W_j(\mathrm{AIC})}. \tag{7}$$

Akaike weight is used to determine the probability of model having good prediction capability or not to predict the wheat productivity and summing to unity $[\sum W_i(\mathrm{AIC}) = 1]$. The higher weights lead to model having relatively good prediction capability and vice versa [38, 43]. Cronbach's alpha "α" and reliability analysis are applied to determine the degree of consistency and relevance of predictors with reference to the measure of response [44, 45].

$$\begin{aligned} W_i(\mathrm{AIC}) &= \frac{\exp\{-(1/2)\Delta_i(\mathrm{AIC})\}}{\sum_{k=1}^{n}\exp\{-(1/2)\Delta_i(\mathrm{AIC})\}}, \\ \text{Corn bach's } \alpha &== \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_T^2}\right), \end{aligned} \tag{8}$$

where k = no. of items, s_i^2 = variance of i^{th} item, and s_T^2 = aggregate item variance.

Reliability coefficient ranging from 0 to 1 and its values near to 0 indicate poor reliability while near to 1 depict strong reliability. The prediction capabilities of models are integrated by using the four different sample size datasets generated through centroid clustering scheme. This study integrates the efficacies of machine learning models with benchmark traditional statistical models to select the most optimum model that follows the evaluation metrics and information criteria.

3. Data Analysis

3.1. Importance of Agronomical Features and Reliability of Datasets. Feature importance refers to techniques that ascribe importance score to input variables which are useful to investigate that how useful the features are to predict the response. Feature importance scores provide the view insight datasets as well as inside the model and improved the efficiency, predictability, and effectiveness of a predictive machine learning model. Before deployment of machine approaches to different datasets, the variations of agronomical features prevailed in simultaneous order for the importance of usefulness in the current study are particularized in Figure 4 for D1, Figure 5 for D2, Figure 6 for D3, and Figure 7 for D4. Table 1 shows the values of Cronbach's alpha for the reliability measure and reports the reliability coefficients as 0.35 for D1, 0.39 for D2, 0.63 for D3, and 0.64 for D4. The reliability of datasets has become strong and strongest as we advanced from D1 to D4.

3.2. Performance Measures of Multiple Linear Regression Models. The performance for the prediction capability of multiple linear regression for the generated different size datasets is evaluated and integrated for both the traditional statistical models and machine learning approaches.

3.3. Machine Learning Models. Multiple linear regression models (MLRMs) are constructed using the machine learning approach and integrated with benchmark traditional statistical models. For MLM, Table 2 shows the performance score 0.266, 0.289, 0.838, and 0.932 for the training datasets and 0.264, 0.285, 0.834, and 0.655 for testing/validated datasets, respectively, for D1, D2, D3, and D4. The R^2 has become strong and strongest as we advance from D1 to D4 for train datasets ($R^2_{\mathrm{Dtrain}(i)} < R^2_{\mathrm{Dtrain}(i+1)}$) and de novo the same for test data except for D4. The RMSE found 9.14 and 9.21 for D1, 7.65 and 8.09 for D2, 3.15 and 3.34 for D3, 1.95 and 3.31 for D4, respectively, for train and test models. The RMSE decreases as we advanced from D1 to

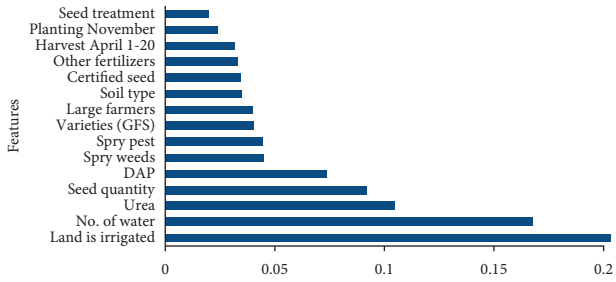

Figure 4: Feature importance for D1.

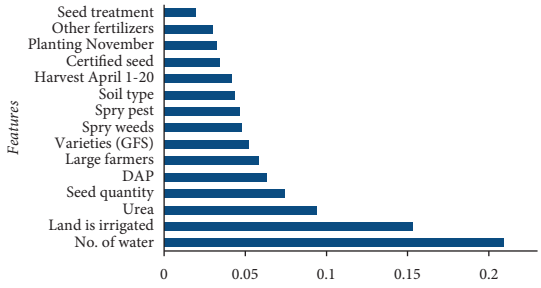

Figure 5: Feature importance for D2.

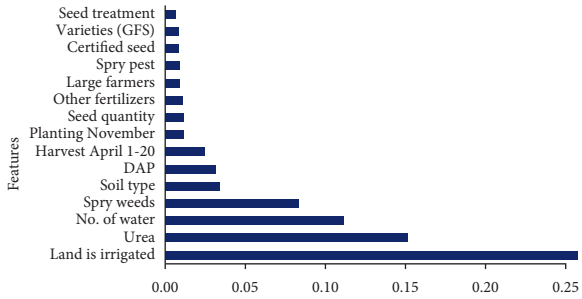

Figure 6: Feature importance for D3.

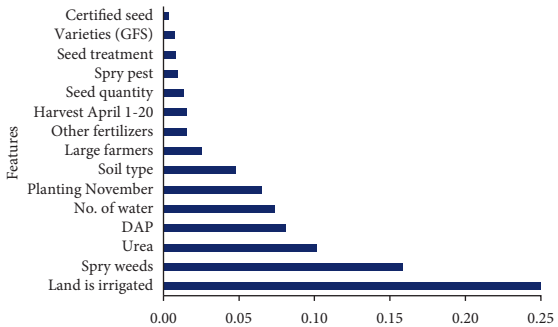

Figure 7: Feature importance for D4.

D4 ($\text{RMSE}_{D(i)} > \text{RMSE}_{D(i+1)}$) for both train and test datasets. The model is train and deployed for the training datasets using 75% train subsets. The D4 shows lowest AIC as 1.62 with highest Akaike weights (AIC_W) as 0.45 followed by AIC as 2.43 and AIC_W as 0.30 for D3, AIC as 4.07 and AIC_W as 0.13 for D2, and AIC as 4.43 and AIC_W as 0.11 for D1. The Akaike weights are increasing ($\text{AIC}_{w(i)} < \text{AIC}_{w(i+1)}$), and AIC is decreasing ($\text{AIC}_{(i)} > \text{AIC}_{(i+1)}$) as we advance from D1 to D4. The evidence ratio justifies the results as D4 model is 4.06, 3.41, and 1.50 more likely to D1, D2, and D3 models, respectively.

Table 1: Cronbach's alpha reliability coefficients for various datasets.

Cronbach's alpha coefficients	D1	D2	D3	D4
	0.35	0.39	0.63	0.64

3.3.1. Integrating Machine Learning and Traditions Statistics Modeling for MLR. Table 2 shows the comparisons of model performance for MLM with benchmarks TSM. For TSM, the performance score is found as 0.265, 0.287, 0.823, and 0.862 and RMSE as 9.17, 7.77, 3.35, and 2.66, respectively, for D1, D2, D3, and D4. It is evident that the highest values of performance score and lowest value of RMSE are found for MLM comparing with benchmark TSM as we advanced from D1 to D4 ($R^2_{\text{TSM}} < R^2_{\text{MLM}}$, $\text{RMSE}_{\text{TSM}} > \text{RMSE}_{\text{MLM}}$). The lowest value of AIC is obtained from MLM comparing with TSM for all the datasets as $\text{AIC}_{\text{TSM}} > \text{AIC}_{\text{MLM}}$. The AIC weight reported 0.45 for MLM, while it is 0.38 for TSM for D4 which elaborated as MLM has high probability for selecting the best model. The evidence ratio for TSM based on D4 is 2.96, 2.51, and 1.14 more likely to D1, D2, and D3 and integrated that E.R is found better in MLM comparing with TSM for all datasets ($\text{E.}R_{\text{TSM}} < \text{E.R}_{\text{MLM}}$). All the performance measure optimized well in ML models clarified that MLM has good prediction capability for prediction of the wheat productivity based on agronomical features. Figure 8 clarifies that the graphical relations exist for learning points of the models for evaluation metrics and information criterion for both MLM and TSM and shows that machine learning performed well for all the datasets and D4 optimized the machine learning multiple regression models.

3.4. Decision Tree and Random Forest Regression Models. The machine learning models are trained and deployed for multiple linear regression models, and predicted well is further trained and deployed for the important and most prominent machine algorithms, i.e., decision tree regression models (DTRMs) and random forest regression models (RFRMs) with the aim to get the most optimized models able to predict the wheat productive well using 75% data to learn the model and 25% as validated datasets to evaluate the model capability on unseen datasets.

3.4.1. Hyperparametric Tuning of DTRM and RFRM. Hackeling [46] reported hyperparametric tuning of DTRM models applied to avoid over and underfitting using the scikit-learn's library GridSearchCV to find out the optimum value of min_sample_split and max_depth (tree depth). Figure 9 shows DTR for D1 having 19822 samples point for training and 6608 sample points for testing phase and

Table 2: Integrating machine learning and tradition statistics modeling for MLR.

Datasets	Machine learning MLRM						Statistical MLRM					
	P.Score	RMSE	AIC	ΔAIC_i	AIC_{Wi}	E.R	P.Score	RMSE	AIC	ΔAIC_i	AIC_{Wi}	E.R
D_1	0.266 (0.264)	9.14 (9.21)	4.43	2.80	0.11	4.06	0.265	9.17	4.43	2.17	0.13	2.96
D_2	0.289 (0.285)	7.65 (8.09)	4.07	2.45	0.13	3.41	0.287	7.77	4.10	1.84	0.15	2.51
D_3	0.838 (0.834)	3.15 (3.34)	2.43	0.81	0.30	1.50	0.823	3.35	2.52	0.26	0.34	1.14
D_4	0.932 (0.655)	1.95 (3.31)	1.62		0.45		0.862	2.66	2.26		0.38	

Testing dataset values are shown in parenthesis.

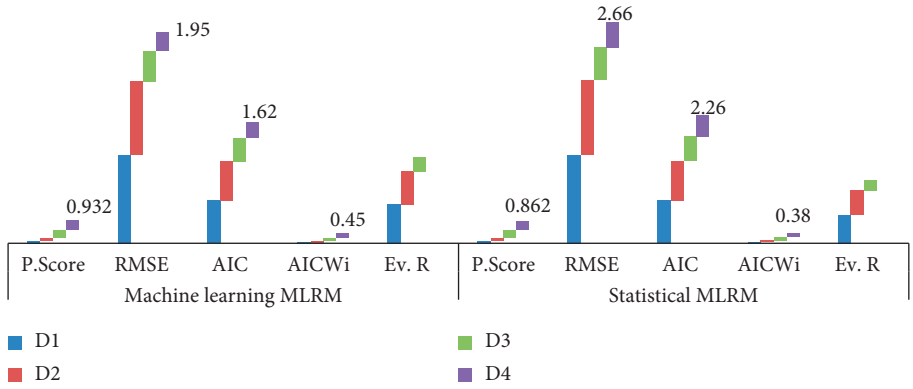

Figure 8: Integrating/comparisons of machine learning and statistical models.

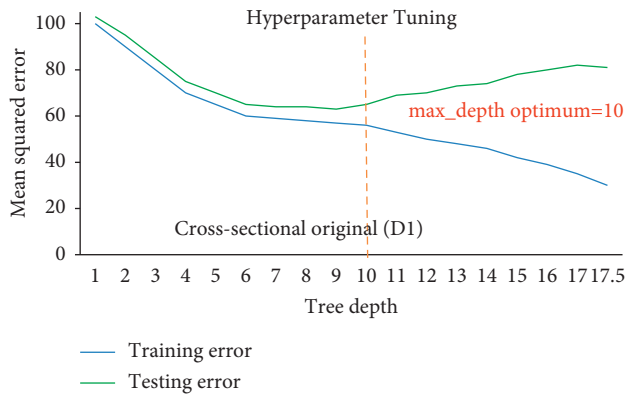

Figure 9: Hyperparameter tuning of DTR for D1.

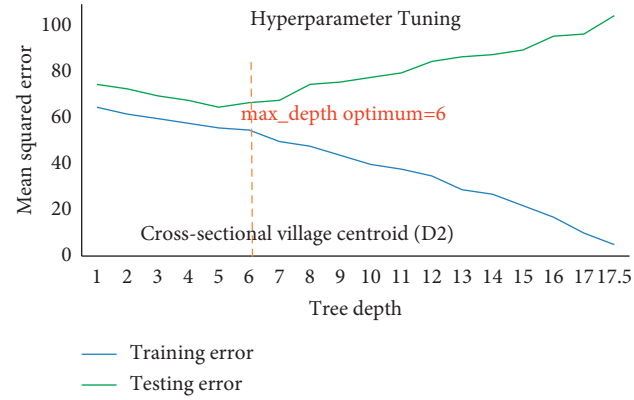

Figure 10: Hyperparameter tuning of DTR for D2.

illustrates that at lower model complexity the model is underfit (high bias and high variance) and the error curve for testing set raises again after tree depth 10 which leads to overfit the model, while for Figure 10, DTR for D2 has 4525 sample points for training and 1509 sample points for testing phase, and the same prevails after tree depth 06, indicating that optimum hyperparameter for tree depth is found 10 and 06 for DTR model based on D1 and D2. The tree depth values got optimized at 05 and 04 for models based on D3 having 108 sample points for training and 37 sample points for testing phase and D4 having 27 sample points for training and 09 sample points for testing phase (Figures 11 and 12). The min_sample_split value found optimized at 29, 28, 6, and 2, respectively, for D1, D2, D3, and D4. The RFR and DTR consist of the same set hyperparameters except random forest called no. of trees in the forest (*n*_sample) and its default value ranged from 10-100. The D1 optimized at no. of tree 10, D2 and D3 at no. of tree 50, and D4 optimized at no. of tree 100 for the prediction model for wheat productivity.

3.4.2. Decision Tree Regression Models. For the DTRM, Table 3 shows the performance score and RMSE as 0.364, 0.366, 0.940, and 0.987 and 8.51, 7.22, 1.92, and 0.828 for train models, while for test model the performance scores are 0.323, 0.331, 0.731, and 0.741 and RMSEs are 8.82, 7.82, 4.26, and 2.87. R^2 is increasing, and RMSE is decreasing ($R^2_{D(i)} < R^2_{D(i+1)}$, $\text{RMSE}_{D(i)} > \text{RMSE}_{D(i+1)}$) for train and test models as we advanced from D1 to D4. The DTR model is trained and deployed for the training datasets using 75% train subsets. The AIC reported diminishing trend as 4.28, 3.96, 1.44, and 0.29 for D1 to D4 ($\text{AIC}_{(i)} > \text{AIC}_{(i+1)}$). The AIC_W of models based on D4 is highest with probability 0.54

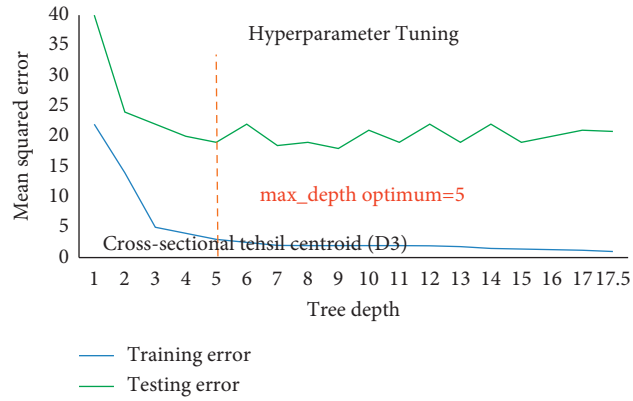

Figure 11: Hyperparametric tuning of DTR for D3.

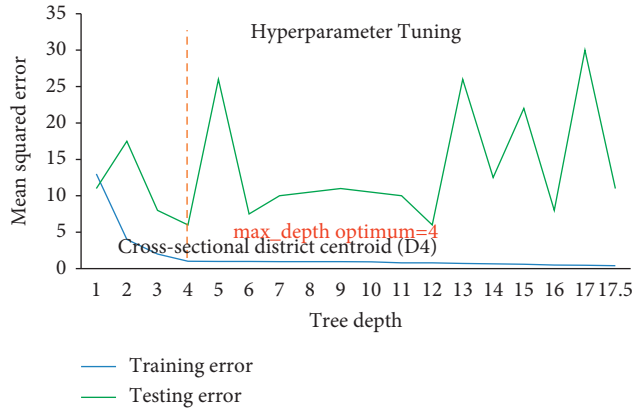

Figure 12: Hyperparametric tuning of DTR for D4.

Table 3: Integrating the DTR and RFR with evaluation metric and information criteria.

Datasets	Machine learning DTR						Machine learning RFR					
	P.Score	RMSE	AIC	ΔAIC_i	AIC_{Wi}	E.R	P.Score	RMSE	AIC	ΔAIC_i	AIC_{Wi}	E.R
D_1	0.364 (0.323)	8.51 (8.82)	4.28	4.00	0.07	7.37	0.380 (0.345)	8.40 (8.68)	4.26	3.56	0.09	5.92
D_2	0.366 (0.331)	7.22 (7.82)	3.96	3.67	0.13	6.27	0.388 (0.362)	7.09 (7.64)	3.92	3.22	0.11	5.00
D_3	0.940 (0.731)	1.92 (4.26)	1.44	1.15	0.30	1.78	0.948 (0.786)	1.78 (3.79)	2.18	1.48	0.26	2.10
D_4	0.987 (0.741)	0.828 (2.87)	0.29		0.54		0.973 (0.877)	1.23 (1.97)	0.70		0.54	

Testing dataset values are shown in parenthesis.

followed by D3 = 0.30, D2 = 0.13, and D1 = 0.07 ($AIC_{w(i)} < AIC_{w(i+1)}$). The E.R values of DTR models show that the model learns from D4 is 7.37, 6.27, and 1.78 more likely to models learns from D1, D2, and D3.

3.4.3. Random Forest Regression Models. For the RFR, Table 3 shows the performance score and RMSE as 0.380, 0.388, 0.948, and 0.973 and 8.40, 7.09, 1.78, and 1.23 for train sets, while the performance score and RMSE is reported as 0.345, 0.362, 0.786, and 0.877 and 8.68, 7.64, 3.79, and 1.97 for test models. R^2 shows the increasing, and RMSE shows the diminishing relation as we advanced from D1 to D4 ($R^2_{D(i)} < R^2_{D(i+1)}$, $RMSE_{D(i)} > RMSE_{D(i+1)}$). The RFR model is trained and deployed for the training datasets using 75% train subsets. The AIC reported diminishing trend 4.26, 3.92, 2.18, and 0.70 with increasing AIC_W as 0.09, 0.11, 0.26, and 0.54 for D1, D2, D3, and D4 models ($AIC_{w(i)} < AIC_{w(i+1)}$ and $AIC_{(i)} > AIC_{(i+1)}$). The highest values of AIC weight reported from model learn from D4 followed by models learn from D3, D2, and D1. The E.R values of RFR models show that the models learn from D4 and are 5.92, 5.0, and 2.10 more likely to models learn from D1, D2, and D3.

3.5. Comparative Quantification of Machine Learning Models for Different Datasets. Section 3.3.1 depicts that machine learning performed well comparing with traditional statistical approaches for multiple regression models. Section 3.3 presents models further trained and deployed for machine learning algorithms, i.e., decision tree regression models (DTRMs) and random forest regression models (RFRMs) with the aim to get the most optimized models able to predict the wheat productive well.

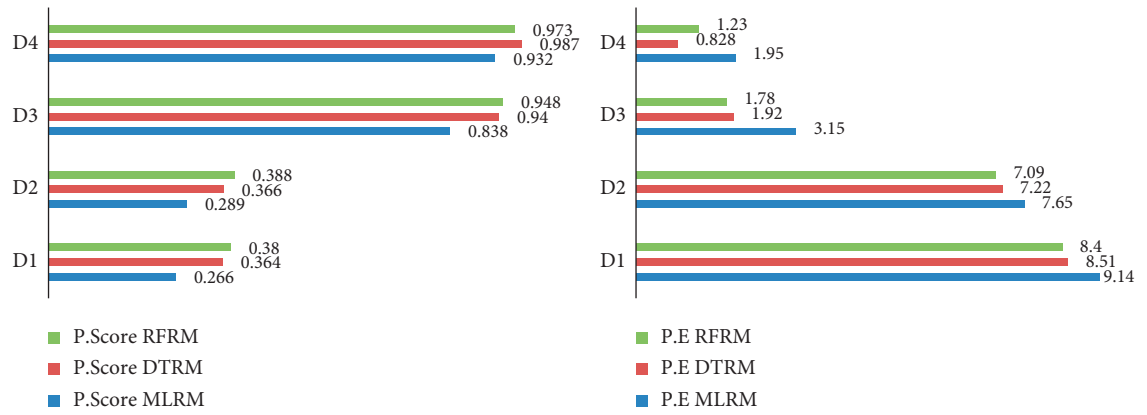

Figure 13: The performance measure of MLMLR, DTR, and RFR machine learning models.

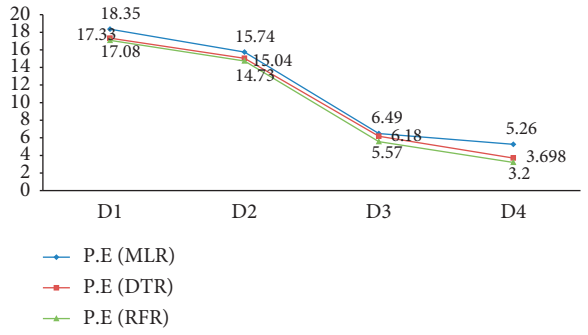

Figure 14: Integrating decomposition prediction error for MLR, DTR and RFR.

In Tables 2 and 3 and Figure 13, the performance score of RFR models is reported well for all training and testing datasets followed by DTR and MLR for D1 and D2. The performance score of RFR is found high for D3 training set, while little bit variation is found for testing sets, and for D4 all models show performance above 90% for training sets and only RFR approach to 0.877% for testing/validation datasets, while DTR has 0.741 and MLR has 0.655. The RMSE of RFRM reported low for D1 and D2 for train and de novo the same for test models. The RFRM shows good for D3 train models, while MLR supersedes on slight extent for test models. The DTR performed well for D4 train model, while for test model RFR supersedes the DTR. The MLM train and deployed for training datasets revealed the relation as ($R^2_{\mathrm{MLRM}} < R^2_{\mathrm{DTRM}} < R^2_{\mathrm{RFRM}}$), ($\mathrm{RMSE}_{\mathrm{MLRM}} > \mathrm{RMSE}_{\mathrm{DTRM}} > \mathrm{RMSE}_{\mathrm{RFRM}}$)for D1, D2 and D3, while for D4, all models show high performance score as MLR = 0 .932, DTR = 0.987 and RFR = 0.973. Data preprocessing optimized the model predictability well for all datasets as all models upswing the performance from original datasets (D1) to generated datasets (D2, D3, D4) for MLM. In Figure 14, learning curves (L.E) demonstrate the comparison for decomposition of prediction error (P.E), and it is validated that RFRM revealed lower prediction error simultaneously for D1, D2, D3, and D4 prediction models as 17.08, 14.73, 5.57, and 3.2 followed by DTR as 17.33, 15.04, 6.18, and 3.698 and MLR 18.35, 15.74, 6.49, and 5.26.

(P.EMLRMDi) >(P.EDTRMDi) > (P.ERFRMDi). RFRM revealed good performance score and bottommost decomposition prediction error as we advanced from D1 to D4. RFRM successfully predicted the wheat productivity when compared against other models using the original and generated datasets.

4. Conclusions

This study integrated the efficacies of machine learning regression algorithms using multiple linear regression models (MLRMs), decision tree regression models (DTRMs), and random forest regression models (RFRMs) with benchmark traditional statistical models to converge the optimization capability of prediction models for wheat productivity. The original dataset of 26430 (D1) crop-cut experiment along with fifteen features is collected from the crop reporting service. The 2nd-stage area frame sampling is applied to select the sample. The new approach of centroid clustering scheme is introduced which can enhance the model performance by reducing the sample size. Three more datasets are generated to optimize the model performance for both the machine learning models (MLMs) and traditional statistical models (TSMs). The generated datasets comprise from 6034, 145, and 36 sample points generated from village, tehsil, and district-level centroid clusters. The 75% dataset is used as training and 25% as testing subsets. Evaluation metrics approach (R^2, RMSE), Akaike information criterion (AIC) with weights ($\mathrm{AIC}_{\mathrm{W}}$), evidence ration (E.R), reliability analysis, and decomposition prediction error (P.E) are applied to compare the performance of models. The performance score (P.S) increased, while the RMSE and AIC decreased for both MLM and TSM as we advanced from D1 to D4 for MLRM. The P.S and E.R reported high ($\mathrm{E.R}_{\mathrm{TSM}} < \mathrm{E.R}_{\mathrm{MLM}}$ & $R^2_{\mathrm{TSM}} < R^2_{\mathrm{MLM}}$), while RMSE and AIC reported low ($\mathrm{RMSE}_{\mathrm{TSM}} > \mathrm{RMSE}_{\mathrm{MLM}}$ & $\mathrm{AIC}_{\mathrm{TSM}} > \mathrm{AIC}_{\mathrm{MLM}}$) for MLM comparing with benchmark TSM as we proceed from D1 to D4 for MLRM. The MLM based on MLRM has good prediction capability for all the datasets, and D4 optimized the MLM. The MLM trained and deployed for MLRM is further trained and deployed for DTRM and RFRM with the aim to get the most optimized

model. RFRM revealed good P.S, bottommost P.E for all the datasets. The RFRM successfully predicted the wheat productivity followed by DTRM and MLRM for D1, D2, D3, and D4. It is demonstrated that machine learning models provide superior performance by centroid clustering even for sample size as we advanced from D1 to D4. This study demonstrated strong evidences for the implementation of machine learning models as an alternative of traditional statistical models for future research direction and correct policy decisions regarding wheat productivity. The advancement in science, technologies, and implementations of innumerable agronomical constraints in various fields of agriculture leads to immense volume of data, and this study provides the detailed hierarchy of centroid clustering which leads to increase the model performance by reducing the sample size. This hierarchy of centroid clustering could also be extended to multistage centroid clustering for future research, and it could also be applied for all supervised machine learning algorithms to enhance the model performances.

Data Availability

The cross-sectional original datasets and generated datasets used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors' Contributions

Muhammad Islam performed descriptions, data preparations, methodologies, data analysis, and conclusion. Farrukh Shehzad contributed to supervision, preparations, data analysis, and descriptions.

Acknowledgments

The authors would like to thank Dr. Muhammad Omar, Assistant Professor, Department of Computer Science, the Islamia University of Bahawalpur, Pakistan, for their appreciable directions regarding implementations of machine learning techniques. The authors are very grateful to Dr. Abdul Qayyum, Director of Agriculture, Crop Reporting Service, Govt. of the Punjab, Pakistan, who provide us valuable statistical datasets and good directions for scaling and categorization of data levels. The assistance provided by Mrs. Rabia Siddiqui, Statistical Officer, CRS, regarding data handling is appreciable for us. The very strong data collection mechanisms and efforts of all the team of Crop Reporting Service, Agriculture Department, Punjab, are considerable good asset for us and for our sweet homeland Pakistan.

References

[1] D. Elavarasan and P. D. R. Vincent, "A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 11, pp. 10009–10022, 2021.

[2] M. Islam, F. Shehzad, and M. Omar, "Modeling wheat productivity using hierarchical regression: a way to address food security concerns," *Ilköğretim Online*, vol. 20, no. 2, 2021.

[3] S. A. Shoaib, M. Z. K. Khan, N. Sultana, and T. H. Mahmood, "Quantifying uncertainty in food security modeling," *Agriculture*, vol. 11, no. 1, p. 33, 2021.

[4] M. Islam, "Factors affecting major food crops production a case study of District Bahawalpur," 2017.

[5] G. C. Nelson, M. W. Rosegrant, A. Palazzo et al., *Food Security, Farming, and Climate Change to 2050: Scenarios, Results, Policy Options*, Vol. 172, International Food Policy Research Institute, , Washington, DC, USA, 2010.

[6] C. P. Timmer, "Food security in asia and the pacific: the rapidly changing role of rice," *Asia & the Pacific Policy Studies*, vol. 1, no. 1, pp. 73–90, 2014.

[7] D. Elavarasan and D. R. Vincent, "Reinforced XGBoost machine learning model for sustainable intelligent agrarian applications," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 5, pp. 7605–7620, 2020.

[8] D. Elavarasan and P. D. R. Vincent, "Fuzzy deep learning-based crop yield prediction model for sustainable agronomical frameworks," *Neural Computing & Applications*, vol. 33, no. 20, pp. 13205–13224, 2021.

[9] J. H. Jeong, J. P. Resop, N. D. Mueller et al., "Random forests for global and regional crop yield predictions," *PLoS One*, vol. 11, no. 6, Article ID e0156571, 2016.

[10] A. Enghiad, *Examining the Response of World Wheat Prices to Climatic and Market Dynamics*, Colorado State University, Fort Collins, CO, USA, 2015.

[11] I. Kiss, "Significance of wheat production in world economy and position of Hungary in it," *APSTRACT: Applied Studies in Agribusiness and Commerce*, vol. 5, pp. 115–120, 2011.

[12] C. Ramesh, "Challenges to ensuring food security through wheat," *CAB reviews: Perspectives in agriculture, veterinary science, nutrition and natural resources*, vol. 4, no. 65, pp. 1–13, 2009.

[13] S. Nosratabadi, S. Ardabili, Z. Lakner, C. Mako, and A. Mosavi, "Prediction of food production using machine learning algorithms of multilayer perceptron and ANFIS," *Agriculture*, vol. 11, no. 5, p. 408, 2021.

[14] F. Zulfiqar and A. Hussain, "Forecasting wheat production gaps to assess the state of future food security in Pakistan," *Journal of Food and Nutritional Disorders*, vol. 3, no. 3, p. 2, 2014.

[15] I. Sharma, B. Tyagi, G. Singh, K. Venkatesh, and O. Gupta, "Enhancing wheat production—a global perspective," *Indian Journal of Agricultural Sciences*, vol. 85, no. 1, pp. 3–13, 2015.

[16] D. Elavarasan and P. D. Vincent, "Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications," *IEEE Access*, vol. 8, pp. 86886–86901, 2020b.

[17] S. Sun, Z. Cao, H. Zhu, and J. Zhao, "A survey of optimization methods from a machine learning perspective," *IEEE Transactions on Cybernetics*, vol. 50, no. 8, pp. 3668–3681, 2019.

[18] A. Qayyum and H. M. J. Shera, "Method of area frame sampling using probability proportional to size sampling technique for crops' surveys: a case study in Pakistan," *Journal of Experimental Agriculture International*, vol. 41, no. 2, pp. 1–10, 2019.

[19] D. Cielen, A. Meysman, and M. Ali, *Introducing Data Science: Big Data, Machine Learning, and More, Using Python Tools*, Manning Publications Co, Shelter Island, NY, USA, 2016.

[20] M. Alagurajan and C. Vijayakumaran, "ML methods for crop yield prediction and estimation: an exploration," 2020.
[21] D. Elavarasan, D. R. Vincent, V. Sharma, A. Y. Zomaya, and K. Srinivasan, "Forecasting yield by integrating agrarian factors and machine learning models: a survey," *Computers and Electronics in Agriculture*, vol. 155, pp. 257–282, 2018.
[22] S. Mishra, D. Mishra, and G. H. Santra, "Applications of machine learning techniques in agricultural crop production: a review paper," *Indian Journal of Science and Technology*, vol. 9, no. 38, pp. 1–14, 2016.
[23] N. Yadav, "Machine learning in agriculture: techniques and applications," 2020.
[24] P. Priya, U. Muthaiah, and M. Balamurugan, "Predicting yield of the crop using machine learning algorithm," *International Journal of Engineering Sciences & Research Technology*, vol. 7, no. 1, pp. 1–7, 2018.
[25] P. Dangeti, *Statistics for Machine Learning*, Packt Publishing Ltd, Birmingham, UK, 2017.
[26] J. McCarthy and E. A. Feigenbaum, "In memoriam: Arthur Samuel: pioneer in machine learning," *AI Magazine*, vol. 11, no. 3, p. 10, 1990.
[27] J. R. Quinlan, "Simplifying decision trees," *International Journal of Man-Machine Studies*, vol. 27, no. 3, pp. 221–234, 1987.
[28] P. E. Utgoff, "Incremental induction of decision trees," *Machine Learning*, vol. 4, no. 2, pp. 161–186, 1989.
[29] A. A. Alif, I. F. Shukanya, and T. N. Afee, *Crop Prediction Based on Geographical and Climatic Data Using Machine Learning and Deep Learning*, BRAC University, Dhaka, Bangladesh, 2018.
[30] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[31] S. V. Venishetty, "Machine learning approach for forecasting the sales of truck components," 2019.
[32] D. Elavarasan and D. R. Vincent, "Effective mining approach to produce quality search results using proposed approach," *International Journal of Intelligent Engineering and Systems*, vol. 10, no. 3, pp. 435–443, 2017.
[33] J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques," *The Morgan Kaufmann Series in Data Management Systems*, vol. 54, pp. 83–124, 3rd edition, 2011.
[34] A. Rahman, "Statistics-based data preprocessing methods and machine learning algorithms for big data analysis," *International Journal of Artificial Intelligence*, vol. 17, no. 2, pp. 44–65, 2019.
[35] L. Igual and S. Seguí, *Introduction to Data Science*, pp. 1–4, Springer, Berlin, Germany, 2017.
[36] P. Geurts, "Bias vs variance decomposition for regression and classification," *Data Mining and Knowledge Discovery Handbook*, Springer, Berlin, Germany, pp. 733–746, 2009.
[37] A. Jain, "Complete guide to parameter tuning in XGBoost (with codes in Python)," 2016, https://www.analyticsvidhya.com/blog/2016/03/completeguide-parameter-tuning-xgboost-with-codes-python.
[38] H. T. Banks and M. L. Joyner, "AIC under the framework of least squares estimation," *Applied Mathematics Letters*, vol. 74, pp. 33–45, 2017.
[39] J. J. Dziak, D. L. Coffman, S. T. Lanza, and R. Li, "Sensitivity and specificity of information criteria," 2012.
[40] D. N. Gujarati, *Basic Econometrics*, McGraw-Hill, New York, NY, USA, 4th edition, 2003.
[41] K. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference*, Springer, New York, NY, USA, 2nd edition, 2002.
[42] M. R. Symonds and A. Moussalli, "A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion," *Behavioral Ecology and Sociobiology*, vol. 65, no. 1, pp. 13–21, 2011.
[43] E.-J. Wagenmakers and S. Farrell, "AIC model selection using Akaike weights," *Psychonomic Bulletin & Review*, vol. 11, no. 1, pp. 192–196, 2004.
[44] J. M. Bland and D. G. Altman, "Statistics notes: Cronbach's alpha," *BMJ*, vol. 314, no. 7080, p. 572, 1997.
[45] U. Sekaran and R. Bougie, *Research Methods for Business: A Skill Building Approach*, John Wiley & Sons, Hoboken, NJ, USA, 2016.
[46] G. Hackeling, *Mastering Machine Learning with Scikit-Learn*, Packt Publishing Ltd, Birmingham, UK, 2017.