Pak. J. Agri. Sci., Vol. 58(2), 461-470; 2021
ISSN (Print) 0552-9034, ISSN (Online) 2076-0906
DOI: 10.21162/PAKJAS/21.661
http://www.pakjas.com.pk

YIELD FORECASTING AND ASSESSMENT OF INTERANNUAL WHEAT YIELD VARIABILITY USING MACHINE LEARNING APPROACH IN SEMI-ARID ENVIRONMENT

Hafiza Hamrah Kanwal[1,*], Ishfaq Ahmad[2], Ashfaq Ahmad[3] and Yongfu Li[4]

[1]School of Computer Science, Chongqing University of Posts and Telecommunications, Chongqing, China; [2]Climate Resilience Department, Asian Disaster Preparedness Center, Islamabad Pakistan; [3]Department of Agronomy, University of Agriculture, Faisalabad, Pakistan; [4]Key Laboratory of Intelligent Air Ground Cooperative Control for Universities in Chongqing, College of Automation, Chongqing, China
***Corresponding author's email: hamrah.kanwal@yahoo.com**

Accurate and timely information about production estimates of wheat is useful for policymakers and government planners. The traditional methods for yield forecasting are labor insensitive and time-consuming therefore remote sensing is an effective approach for precise yield forecasting. The study was planned to develop a comprehensive framework for yield forecasting and to assess interannual yield variability in semi-arid regions. For wheat area classification, the peak season Landsat-8 satellite images were acquired, and Top of Atmospheric (TOA) correction was performed. The ground-truthing points of 100 farms were collected from the study area for the training of algorithms. The eight machine learning algorithms were used tune and tested using 10-k fold cross-validation and the best model was used for land cover classification of wheat. For yield forecasting, the temporal normalized difference vegetation index (NDVI) and land surface temperature (LST) were derived for the wheat-growing season from November to April. A Principal Component Analysis (PCA) was used to variable selection and then Least Absolute Shrinkage Selection Operator (LASSO) analysis was performed to develop coefficients of the yield forecasting model. The developed model was further used in yield forecasting of 10 years (2008-2018) in four semi-arid regions. The predicted yield was compared with Crop Reporting Service (CRS), Pakistan department. The results of all machine learning algorithms showed an accuracy of 88% to 96%, however, the Random forest algorithm showed higher accuracy, which was further used for classification. The wheat estimated area of 6.9% was less than reported by CRS. For interannual variability, the relationship of observed (CRS) and predicted yield of 10 years showed a close relation with R^2 ranged from 0.69 to 0.75 in the semi-arid region of Punjab, Pakistan. It was concluded that machine learning algorithms can be used as novel tools for yield forecasting and assessment of interannual yield variability.
Keywords: Image classification, Machine learning algorithms, Yield forecasting, Interannual variability.

INTRODUCTION

Precise and timely yield prediction of wheat is valuable for farmers, researchers, and policymakers in decision making and devising the import, export policies (Lobell *et al.,* 2003; Dempewolf *et al.,* 2014; Ahmad *et al.,* 2018a). The population growth of Pakistan is increasing and expected to rise by 271 million up to 2050 (Kirby *et al.,* 2017), while food production is not increasing at the same rate that will affect food security (Cheeseman, 2016). Food availability is also affected by global climatic changes like floods and droughts (Funk *et al.,* 2019). The accurate and real-time yield prediction at a larger scale is useful to address these concerns. Wheat is the most important cereal crop in Pakistan and is cultivated in the winter season on an area of 8.7 million hectares with a production of 25 million tons (Government of Pakistan, 2018). Timely and accurate wheat yield prediction helps the decision-makers in deciding the import or export of grains to maintain the national reserve for food security and to set support prices(Nagy *et al.,* 2018; Fahad *et al.,* 2019; Roell *et al.,* 2020). Thus, there is a need to develop a comprehensive framework of yield forecasting that helps policy planner in decision making.

Conventional methods such as opinion surveys, crop cut area frame sampling, and the girdawari system are used by the provincial government for agriculture statistics (Dempewolf *et al.,* 2014), which are labor-intensive and time-consuming. A few sample villages are selected for crop cuts which are not representative of all agricultural areas. However, the collected data is available three months after crop harvest, which is not useful for the policymaker, resulted in limited or surplus of wheat (Akhtar, 2014). For this remote sensing is a useful tool in assessing the accurate yield prediction (Dubey *et al.,* 2018; Franch *et al.,* 2018; Funk *et al.,* 2019). Remote sensing used satellite observation such as normalized difference vegetation index (NDVI) and Land surface temperature (LST) for crop monitoring and yield forecasting (Ahmad *et al.,* 2018a; Neinavaz *et al.,* 2020). A peak seasons image was used for

land cover classification of wheat to avoid false land cover change detection due to phenology (Clark and Pellikka, 2009). It is reported that images close to the peak of the growing season or time of maximum vegetation ‘‘greenness’’ should be preferred (Kim *et al.*, 2011). A similar approach was used by Ahmad *et al.* (2020).

Land cover classification, which is a description of what tangible material is covering the Earth’s surface, is of enormous value to society (di Gregorio, 2005). Determining accurate land cover classification is critical, required expert judgment and selection of algorithms (Punn and Bhalla, 2013). The reliability and use of one algorithm for land cover is a tremendous task, thus For accurate land cover classification, the use of various statistical learning methods such as discriminant analysis, support vector machines, and decision trees produce more accuracy (Johnson *et al.*, 2012). Various approaches have been used for landcover classification, for example, random forest (Saeed *et al.*, 2017), semi-automatic classification (Bouaziz *et al.*, 2017), simple linear method (Hughes *et al.*, 2019), and logistic regression (Das and Pandey, 2019). The accuracy and reliability in using an algorithm for landcover classification is a challenge, further, the factor used in classification often results in the mixing of landcover classes. But machine learning is an emerging approach in data science, the algorithms used in classification showed high accuracies (Abdi, 2020). Several parameters can be tuned for a linear SVM while including a cost term that adds a penalty to the slack variables. Parameters include the tolerance which represents an optimization termination criterion and epsilon that is an insensitive-loss function. Non-linear decision boundaries can be determined by projecting p-dimensional variables to an infinite dimensional space. Each parameter combination was tested using 10 K fold cross-validation, in which data are divided into 10 parts, and validation was executed by K-1. One part was used for calibration and other folds for validation. The accuracy was calculated by the average of all folds. The best classifier was used for the classification of the wheat area, the same methods was used by Ahmad *et al.* (2020).

The commonly used machine learning algorithm is a random forest, which is widely used in earth science for land cover classification (Breiman, 2008). Random forest (RF)was compared with outperforming decision tree classifier and found that RF showed high accuracy of 92% by Rodriguez-Galiano *et al.* (2012). Support vector machine (SVM) can generalize the complex feature and showed high accuracy of 89% in the classification of Landsat-8 images (Goodin *et al.*, 2015). Boosting is another effective algorithm that produces an accurate prediction rule by combing rough and moderate rules(Man *et al.*, 2018). Keeping in view, the current study was planned to use a variety of machine learning algorithms for landcover classification and to develop yield forecasting model for wheat in semi-arid environments.

MATERIAL AND METHODS

***Description of Study Site and collection of ground-truthing data*:** The studies were conducted in Faisalabad (31.25 N, 73.06 E) Punjab, Pakistan. It has a semi-arid climate, where in annual temperature of about 24.2°C and rainfall of 346 mm is recorded (Ahmad *et al.*, 2019). The soil of Faisalabad is silt loam or very fine sandy loam (Ahmad *et al.*, 2018b). Faisalabad is a mixed cropping zone at which wheat, rice, maize, sugarcane, and cotton are cultivated. Wheat is normally grown in Rabi season (November to mid-April), while other crops grown in rabi are clover (berseem), sugarcane, orchards, canola, and potato (Fahad *et al.*, 2019).

To examine the crop classification, an extensive field survey was conducted to collect georeferenced field samples data of 100 farms in 2018. A stratified random sampling technique was used to collect. The data of latitude, longitude, and crop type at each sample field were recorded as shown in Figure 1.

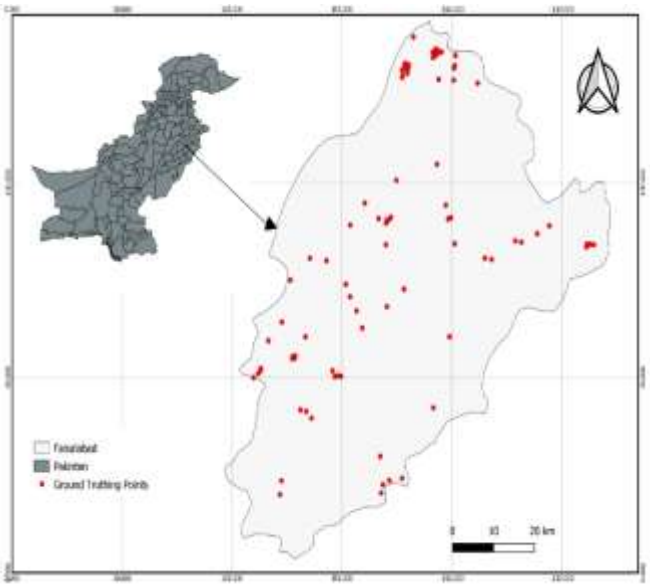


Figure 1. Collection of georeferenced ground-truthing farms in Faisalabad, Pakistan

Acquisition of Satellite data and calculation of temporal NDVI and LST: For classification, three Landsat L8 OLI/TIRS satellite images were acquired from the United State Geological Survey (USGS) portal (https://earthexplorer.usgs.gov/). The study area was covered by three images, with path row of P150-R038, P150-R039, and P149-R038. Top of atmospheric corrections (TOA) was applied by converting the digital number to absolute values of TOA reflectance, by following the method described by De Keukelaere *et al.* (2018).

For the development of the yield forecasting model, the temporal images of Landsat L8 OLI/TIRS for the wheat-growing season (November to April) were acquired with 16

days interval. Cloud contaminated pixels were removed from each image using a cloud mask provided by NASA, using the method described by Gao and Li (2017).

Where NIR is the near-infrared and red is the visible light. Land surface temperature (LST) was calculated using Band-10 of the Thermal Infrared sensor from the Landsat 8. The digital numbers of the sensor were converted into Top of Atmosphere (TOA) Reflectance (Masek *et al.*, 2006) and then reflectance values were converted into Satellite Brightness Temperature by using equation 2. The derived LST was in kelvin, which was further converted into degree centigrade (°C), subtracting the Kelvin temperature by 273.15. The temporal Normalized Difference Vegetation Index (NDVI) of 100 farms was calculated by using the following formula

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}} \quad (1)$$

Where T is the brightness temperature in kelvin, while Lλ is Spectral radiance in watts/ (m^2 * sr * micrometer) and K1 and K2 are the thermal conversion for the band.

$$\text{T} = \frac{\text{K2}}{\ln\frac{\text{K1}}{\text{L}\lambda}+1} \quad (2)$$

***Imaging Classification using Machine Learning*:** For classification, eight machine learning algorithms were trained and tested to select the best classifier for landcover classification of wheat. The eight algorithms were; linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k-nearest neighbor (KNN), Support Vector Machine (SVM) with linear kernel, SVM with Radial Basis Kernel, decision trees, boosting, and random forests. All statistical computing and training of algorithms were conducted in R (RCore, 2016). The algorithms were tested with ground-truthing data using10-k fold cross-validation and the best classifier was selected classification of wheat area. In 10-k fold cross validation data is divided into 10 equal parts and validation is executed by k-1(Anguita *et al.*, 2012).

The LDA methods are supervised method of classification which acquire multiple distinctive class feature from the available pixel (Ye *et al.*, 2005). LDA increase the inter-class variance, while reduced the intra-class variance that leads to generate new feature of data and provide distinctive features of classified data. In LDA the measurement and probability of landcover class are computed from the Bayes Theorem (Lindley, 1958).

Where "f_i" is a linear discriminate function, "μ" is a mean of class, "$\mu_i C^{-1} \mu_i^T$" is Mahalanobis distance (a distance used to measure the dissimilarity in classes). The calculation of probability is unpractical so the use of relative frequencies of each class is calculated by using equation 3.

$$f_i = \mu_i C^{-1} X_k^T - \frac{1}{2}\mu_i C^{-1} \mu_i^T + ln\,(\rho_i) \quad (3)$$

In QDA the decision boundaries are in a quadratic curve and each landcover class has an individual covariance matrix (Tharwat, 2016), as shown in equation 4

$$\partial_k(x) = -\tfrac{1}{2}\, log|\Sigma_k| - \tfrac{1}{2}(x-\mu_k)^T \textstyle\sum_k^{-1}(x-\mu_k) + log\pi_k \quad (4)$$

In QDA there is a need to calculate the Σ_k for each class $k \in (1, \ldots, K)$ rather than assuming $\sum_k = \Sigma$

Random forests construct multiple classification trees with training data (Liaw and Wiener, 2002). To classify the individual feature of the class, the input class is classified with each tree in the forest. The prediction from each tree is pooled to get the final prediction (Bosch *et al.*, 2007). The decision tree classifier also builds the classification in the form of a tree. The data is split into smaller subsets and the topmost decision nodes in the tree are assumed to be the best predictor (Bertsimas and Dunn, 2017).Each node has a decision based on *binary* whether $x_i < a$ or not for a fixed a. The diversity is measured through a Gini criterion using equation 5

$$Gini = 1 - \textstyle\sum_{i=1}^{c}(p_i)^2 \quad (5)$$

Where p*i* is the probability of an object being classified to a particular class.

K-Nearest neighbor (KNN) is a classification algorithm that estimates the landcover class which is nearest to the training data. Where "*n*" is dimensional space, "*q* and *p"* representing the Euclidean vector which starts from initial to terminal points. The KNN calculates the Euclidean distance (d) between training and landcover class as given in equation 6.

$$d(q,p) = \sqrt{\textstyle\sum_{i=1}^{n}(q_i - p_i)^2} \quad (6)$$

Support vector machine is algorithms that find the hyperline in N-dimensional (N is the number of feature) that distinctly classify the landcover class(Heumann, 2011). The boosting is a meta algorithm which improved the classification through training the sequence of weak model and convert into strong learner and each compensate the weaknesses of its predecessors(Liu *et al.*, 2005). Where "f_m" is the weak classifier and "θ_m" is the corresponding weight. The equation for boosting classification is given in Equation 7.

$$F(x) = sign\,(\textstyle\sum_{m=1}^{M} \theta_m f_m(x)), \quad (7)$$

All machine learning algorithms were used for landcover classification of wheat and accuracy was test by comparing with ground-truthing data. The 10 k-fold cross-validation of all algorithms was carried to select the best model for final wheat classification.

***Principle Component Analysis (PCA) for assessing variable importance*:** A PCA was conducted to assess the highly correlated variables with wheat yield. PCA is multivariate statistical techniques that emphasize the variation in data and find a strong pattern in a data set. PCA extracts the information of different variables and expressed them in a new set of orthogonal variables which are called principle components (Pacheco *et al.*, 2013). PCA transforms the data and explorer's the interrelation between the variables (Jackson, 2005). Where "W" is the matrix of the coefficient that is determined through PCA and "*X*" is the variable under study. The basic equation of PCA with matrix notation is given in equation 8.

$$y = W'X \quad (8)$$

Temporal NDVIs and LSTs were derived from 100 farms during the wheat season with 16 days intervals. A total of eight times NDVIs and LSTs related to yield of 100 farms using PCA analysis. The highly correlate eight times NDVIs and LSTs with farms yield was derived by calculating the standard deviation, proportion of variance, and cumulative proportion through PCA. Biplots were also drawn to show which variable is corrected with farm yield.

***Development of yield forecasting model using the Least Absolute Shrinkage and Selection Operator (LASSO) Analysis*:** The selected correlated variables with yield in PCA analysis were used in the LASSO regression to derive the coefficients of the yield forecasting model. LASSO is a type of linear regression that uses the shrinkage, where data values are shrunk to a central point (Jackson, 2005). LASSO regression is used to develop parsimonious models from a large number of variables. LASSO computes the regression coefficient through a ℓ_1-norm penalized least squares. It minimized the residual sum of squares by adding ℓ_1 penalty on the coefficient (Saporta and Niang, 2009) as shown in equation 9.

$$\sum_{i=1}^{n}(y_i - \sum x_{ij}\beta_j)^2 + \lambda \sum_{j-1}^{p}|\beta_j| \qquad (9)$$

Where λ showed the shrinkage amount, the λ=0 indicates that all features are considered, while $\lambda = \propto$ implies no features are considered

In the current study, LASSO was performed using caret and glmnet package in the R statistical program. (Friedman *et al.*, 2010). The selected correlated NDVIs and LSTs of 100 farms with yield through PCA were used in LASSO regression to develop the coefficient of the yield forecasting model. The bootstrapping method was used to develop the yield forecasting model. It is a statistical method that relies on random sampling with replacement (Holmes, 2003). 70% of data were used to train the model while 30% of data were used for testing.

Assessing the interannual yield variability of wheat: The developed model was used to predict the wheat yield of four semi-arid regions for 10 years (2008-2018). The predicted yield of each year is compared with the observed yield reported by Crop Reporting Services (CRS) Punjab, Pakistan. For this purpose, historical 10 years (2008-2018) satellite images of Landsat 7 ETM were downloaded from the USGS website and were atmospherically corrected using methods described by Flood (2014). Cloud masking and scan-line correction were also applied to all images by following the protocol (Scaramuzza and Barsi, 2005). After mosaicking, the images of each year the wheat area were extracted using a landcover map of wheat, developed from the best machine learning algorithm. The selected NDVIs and LSTs used in the yield forecasting model were derived from images of each year (2008-2018). The derived mean NDVIs and LSTs of a particular period were used in the developed model to predict the regional yield for 10 years, which was compared with the CRS yield to assess the interannual variability. The detailed methodological framework is given in Figure 2.

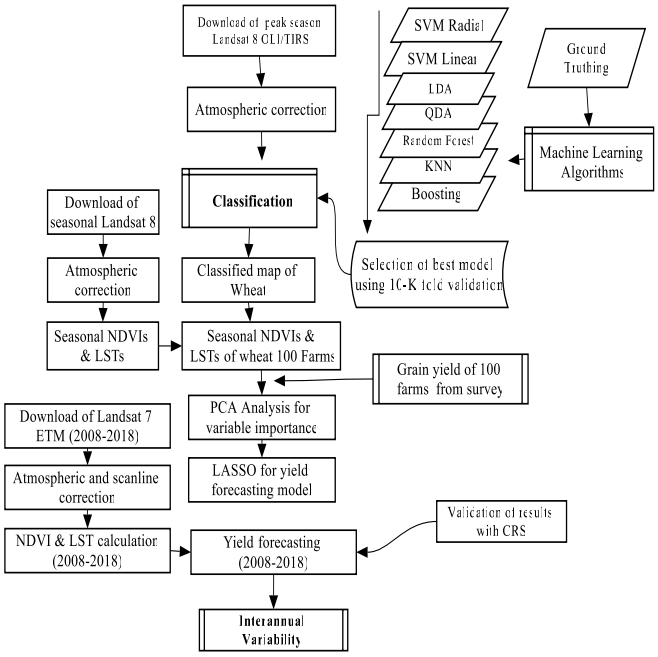


Figure 2. Methodological framework for yield forecasting and assessing interannual wheat yield variability.

RESULTS

***Landcover classification of wheat using machine learning algorithms*:** Machine learning algorithms used for landcover classification of wheat showed a accuracy between 88- 96% as shown in table 1. The hyper-parameter i.e. known as tuning was also selected for all algorithms, which was based on iterative and grid search approach. Optimization of hyper parameter was done to compare the algorithms in fair methods without any prior knowledge. The parameters were tuned to optimize the performance of each machine learning algorithm as shown in Table 1. The results showed that the values for optimum termination criteria cost parameters in SVM were 30.3 for radial and 0.933 for linear. In the random forest, the number of tresses was 125 with a node size of 3, while for boosting the number of tresses was 100, and the shrinkage value was 0.20 (Table 1). The 10-k fold cross-validation and hyper parameter showed a higher accuracy of 96% in a random forest, while the booting algorithm performed relatively poor which showed an accuracy of 88%. The best-selected model (Random Forest) was used for final landcover classification.

The ground-truthing data were used as training of wheat class and then Random forest algorithm was used for classification, the classified map of wheat is given in Figure 3. The wheat classified area predicted by random forest was 6.18 million-

acre in the Faisalabad district during 2018, while the CRS department reported a wheat area of 6.64 million-acre. The estimated area of 6.9% was less than reported by CRS. It could be due to fact that the CRS department harvests the wheat sample on a small village level and convert into an acre and also count the small path and water channel, resulted an increase in wheat area.

Table 1. Accuracy of Machine Learning Algorithms and selection of parameters

Model	Accuracy	Parameters
SVM-Radial	0.93	cost=30.4, tolerance=1e-5, epsilon=0.1
QDA	0.91	N/A
Random Forests	0.96	mtry=4, node size=3, no. of trees=125
Trees	0.90	size=3
SVM-Linear	0.93	cost=0.933, tolerance=1 epsilon=0.1
KNN	0.89	k = 4
LDA	0.90	N/A
Boosting	0.88	shrinkage=0.20, number of trees=100, depth=1

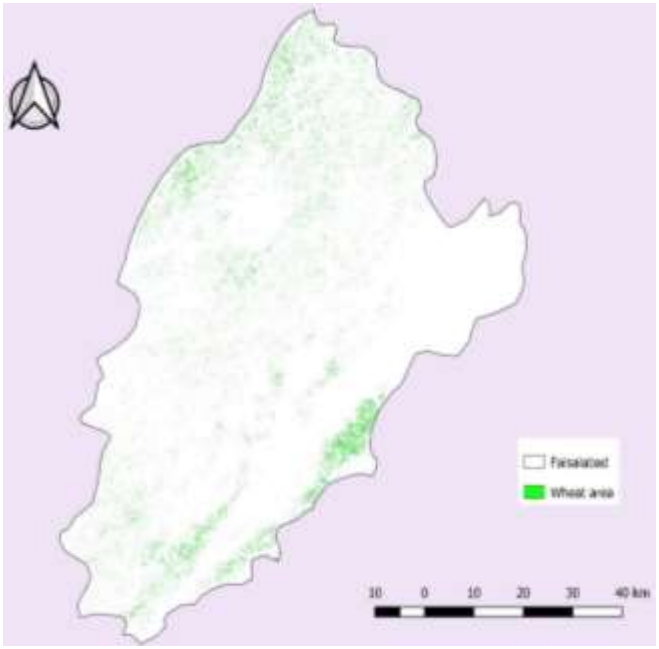
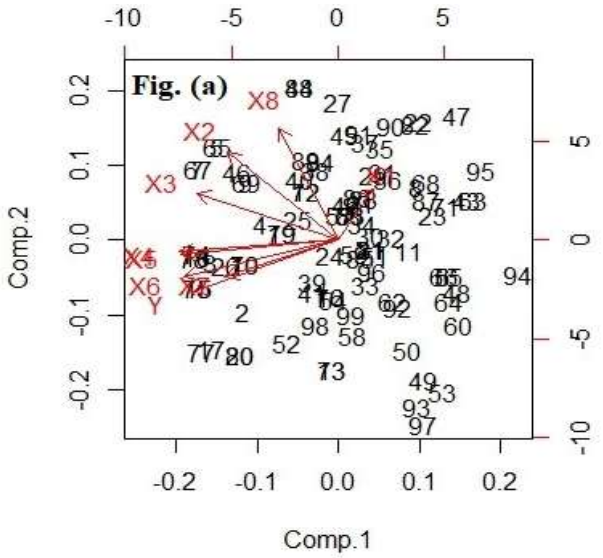


Figure 3. Classified area of wheat in Faisalabad during the year 2018

***Selection of variable through Principle Component Analysis*:** PCA analysis was carried out on temporal NDVIs and LSTs with a yield of 100 farms. The results indicated that out of eight temporal NDVIs, The NDVI4, NDVI5, and NDVI6 are closely related to farm yield (Figure 4a). The NDVI4 was derived ~90 days after planting i.e. before anthesis of wheat; NDVI5 was derived ~105 days after plating *i.e.*, anthesis stage of wheat, while NDVI6 was calculated ~120 days after planting i.e. after anthesis stage.

The PCA results of temporal LSTs indicated that LST3 and LST4 showed a close association with wheat yield. The LST3 was calculated ~45 days after planting, while LST4 was derived ~60 days after planting (Figure 4b). Further results indicated that LST after 60 days of planting showed a negative relation with yield, an increase in temperature around the anthesis stage (~100-120 days after planting) could reduce wheat yield.

a)



b)

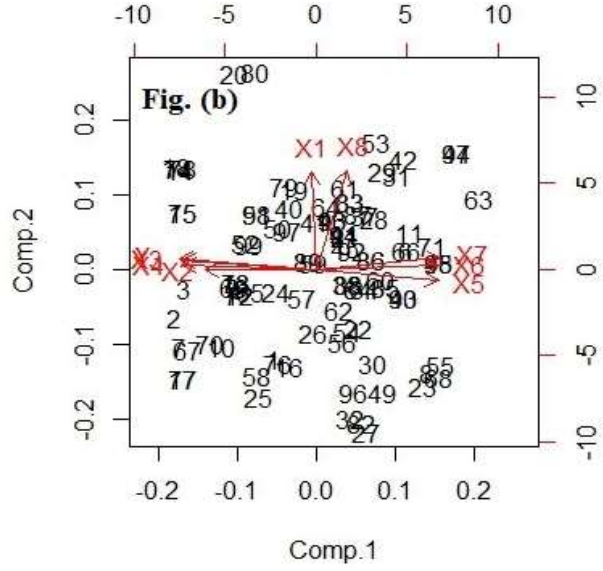


Figure 4. Principle Component Analysis of NDVIs and LSTs during the year 2018 in figure 4a the "$x1, x2, x3..x8$" represents the temporal NDVIs, while Figure 4b shows the temporal LSTs. The "Y" indicates the wheat yield of 100 farms.

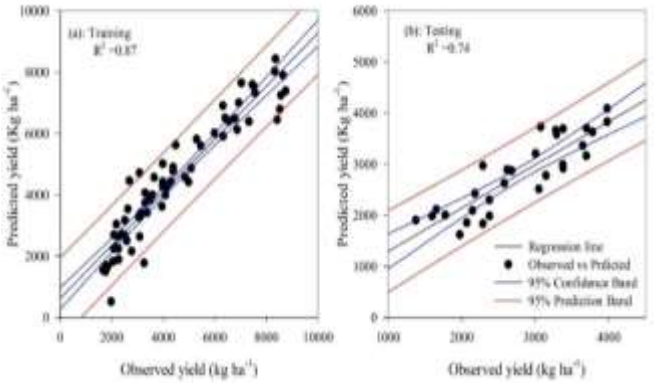
***Development of yield forecasting model*:** The selected variables through PCA analysis were used in LASSO regression to develop the coefficients of the yield forecasting model as shown in table 2. The selected NDVIs and LSTs were significantly (*Pr*>0.001) correlated with wheat yield. Results showed that LSTs which are acquired during early plant growth stage showed a positive relation with yield. The LST3 which was acquired at 45 days after planting (DAP) of wheat crop, while LST4 were calculated at 60 DAP are significantly correlated with wheat yield. While the peak seasons NDVIs were closely related to yield. The NDVI4, which was calculated at 90 days after planting (before anthesis); the NDVI which was derived at 105 days after planting (anthesis or peak stage) and NDVI6 which was calculated at 120 days after planting (after anthesis) showed close association with yield(Table 2).

The developed statistical model was developed with70% of data and tested with 30% data. The results showed a close relationship between observed and predicted wheat yield with R^2 of 0.87 (Figure 5a). The testing of the model was carried out with 30% data which also showed a close match between observed and predicted wheat yield with R2 of 0.74 as shown in Figure 5b.

Table 2. Developed yield forecasting model through LASSO regression.

	Estimate	Std. Error	T value	Pr(>|t|)
Intercept	16193.0	2988.1	5.419	5.74e-07***
LST3	851.7	170.3	5.001	3.14e-06***
LST4	-1005.4	215.3	-4.670	1.15e-05***
NDVI4	10573.6	1942.8	5.442	5.22e-07***
NDVI5	7356.5	2499.4	2.943	0.004210**
NDVI6	-22592.9	2413.7	-9.360	1.24e-14***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



LST3: 45 days after planting; LST4: 60 days after planting; NDVI4: 90 days after planting (before anthesis); NDVI5: 105 days after planting (anthesis or peak stage); NDVI6: 120 days after planting (after anthesis)

Figure 5. Training and testing of developed yield forecasting model through LASSO regression.

***Assessment of interannual yield variability*:** The developed model was used to predict the historical wheat yield of 10 years (2008-2018) in semi-arid regions, which was compared with CRS yield to assess the interannual variability. The relationship of observed (CRS) and predicted yield of 10 years showed a close relation with R^2 ranged from 0.69 to 0.75 in the semi-arid region of Punjab, Pakistan (Figure 6). The high association (R^2=0.75) was recorded in the Faisalabad region, followed by Jhang (R^2=0.73) and Chiniot (R^2=0.72) region. The Sargodha region showed relatively less agreement between observed and predicted yield (R^2=0.69) as compared to Faisalabad, Jhang, and Chiniot regions, which could be due to fact that citrus is cultivated in the Sargodha region and wheat is intercropped win the citrus area, resulting in a mixed pixel which can create an error in the derivation of NDVI and LST.

The developed yield forecasting model also predicted the interannual variation in wheat yield (Figure 7). In Faisalabad and Chiniot regions, the years 2013 and 2014 showed higher observed yield, the developed yield forecasting model also predicted similar variations in both years (Figure 7a, b). In the Chiniot region, the predicted yield was higher than observed, which could be due to some terminal stress in crops that reduced the observed yield. The model predicted a similar yield in the Jhang region during the year 2008, while the model under predicted the yield in 2009, 2011, and 2013 as compared to the observed yield (Figure 7c). The underproduction in yields could be due to biotic and abiotic stress by plants when satellite data were collected. The model under-predicted the yield in the Sargodha district in all years (Figure 7d), however, a close association was recorded between observed and predicted yield.

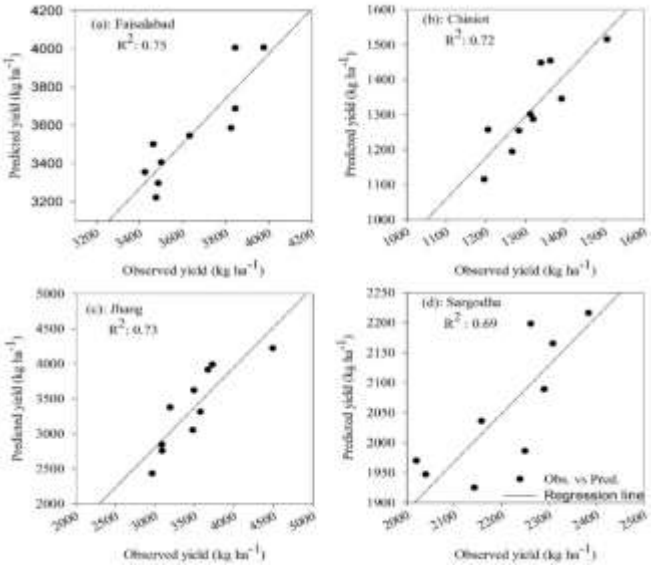


Figure 6. Relationship of observed and predicted wheat yield in semi-arid regions of Punjab Pakistan.

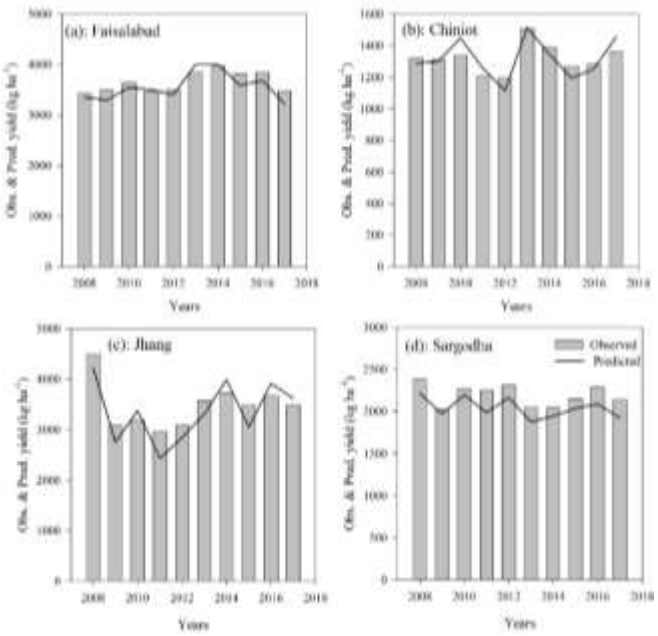


Figure 7. Interannual yield variability of 10 years (2008-2018) in semi-arid regions of Punjab Pakistan

DISCUSSION

The objective of this study was to develop a comprehensive framework of machine learning algorithms for yield forecasting of wheat in Punjab, Pakistan. Various algorithms are available for image classification, but it is not yet defined which algorithm or technique is best enough to generate accurate results for landcover classification especially in mixed cropping zone like Faisalabad. Thus, machine learning algorithms are useful, cost-effective, and achieving high quality and accuracy in landcover classification (Lary *et al.*, 2018).McIver and Friedl, (2001)found that machine learning algorithms like SVM, QDA, KNN, LDA, and random forest are reliable and shows greater accuracy if the training sample is large in landcover classification. In the current study, a more training sample was collected for landcover classification as shown in figure 1. Breiman (2008) also found that the RF classifier performs better in greater training samples. The random forest was also used by Saeed *et al.*, (2017) for yield forecasting of wheat using weather data in semi-arid environment which showed a greater accuracy with R^2 of 0.95.

The performance of algorithms depends upon the iteration of hyperparameters, in the current study various parameters were tuned as shown in table 1. Hyperparameters is a set of function arguments for which has a range of value, in modern machine learning the parameters is tuned to get optimal predicted performance. A similar approach of tunning the machine learning algorithms was also used by Vanli *et al.* (2020)

PCA analysis was used in the current study for the selection of important variables such as NDVIs and LSTs yield. PCA is a statistical algorithm that is used to find out the correlated variables from the set of values. Bro and Smilde (2014) reported that PCA is a multivariate and dimension reducing technique that is significantly used to describe the inter-correlated dependent variables. In the current study, the LASSO regression was used to derive the coefficients of the model (Table 2). The PCA and LASSO analysis was used by Ahmad *et al.* (2020) for the selection of NDVIs, LSTs and further development of yield forecasting model. He further found that early stage LSTs and Peak seasons are closely related with Maize yield. Kuhn and Johnson (2013) reported that LASSO is a useful method in the estimation of predictor parameters with low biased. Ahmad et al. (2018) developed a yield forecasting model for maize in the semi-arid region and used the PCA and LASSO regression to find out the coefficient for the model.

High spatial resolution Landsat satellite images were used in the current study for landcover classification (Figure 3). The study area has a mixed cropping zone and farmers have small landholdings and they mostly grow wheat, maize, sugarcane, and fodder. Due to crop diversification and small-sized field, however, the use of Landsat-8 data improved the accuracy of classification. Fahad *et al.* (2019) also used Landsat 8 imagery for landcover classification of wheat at Faisalabad under semi-arid environment. The study results showed that the estimated area though remote sensing was 6.9% was less than reported by CRS in Faisalabad (Figure 3). The overestimation of areas by CRS is due to them that they harvest small areas and count the number of acres under cultivation, but on ground area cultivated by wheat is not equal to one acre. The reasons could be poor patches in the field for wheat, field borders, temporary paths within the fields for transportation, water channels. However, in remote sensing, these areas are excluded by the algorithm in land cover classification.

In the current study, interannual yield variability was assessed for 10 years by predicting the yield from the yield forecasting model, which showed a good relationship between observed and predicted yield (Figure 6 & 7). The accuracy of production forecast depends upon the satellite-derived statistical indices. The combination used NDVI and LST in the empirical model, improved the accuracy in predicting the year-to-year variability (Leroux *et al.*, 2015). The LST is a fundamental parameter that affected the crop yield, the LST during the vegetative stage of the crop, increases the growth and development of the crop (Li *et al.*, 2013), while NDVI is a strong satellite-derived vegetative index used for seasonal yield forecasting of The variation in observed and predicted yield is due to a change in management practices. In case of under prediction, the crop might be in stress condition when satellite data were collected, after that farmer applied fertilizers and pesticides to get more yield. Similarly, in over prediction, the crop could be in good condition and later

stages faced biotic and abiotic stress, resulted in less yield. Similar was discussed by various researcher (Ahmad *et al.*, 2017, 2018, 2019; Waqas *et al.* 2019, 2020; Ullah *et al.*, 2018, 2019; Vanli *et al.*, 2019).

The proposed method for yield prediction is effective for semi-arid regions of Pakistan; however, a few points can be improved in this study. The current study used top-of-atmosphere reflectance for Landsat, which is fine for Pakistani conduction where atmospheric conditions are stable during the January to April months. However, the other regions might have more possibilities of cloud contamination thus, the gap-filling technique should be employed as reported by Scaramuzza and Barsi (2005).

Conclusion: Machine learning algorithms were used for assessing the spatial distribution of wheat area and the development of yield forecasting models in semi-arid regions. The algorithms showed an accuracy of more than 86%. The highest accuracy of 0.96% was recorded in the random forest which was further used for classification. The wheat estimated area of 6.9% was less than reported by CRS. The yield forecasting model was developed by calculating the NDVIs and LSTs values of 100 farms. The developed model was used to predict the wheat yield of 10 years (2008-2018). The relationship of observed (CRS) and predicted yield of 10 years showed a close relation with R^2 ranged from 0.69 to 0.75 in the semi-arid region of Punjab, Pakistan. The developed yield forecasting model is useful for policymakers in decision making.

REFERENCES

Abdi, A.M. 2020. Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. GIScience Remote Sens. 57: 1-20.

Ahmad, A., M. Ashfaq, A. Wajid, T. Khaliq, I. Ahmad and F. Rasul. 2017. Food and Income Security in Punjab, Pakistan Adapting cotton-wheat farming systems to climate change. Agric. Model Intercomp. Improv. Proj. 2:25-35.

Ahmad, I., U. Saeed, M. Fahad, A. Ullah, M. ur Rahman, A. Ahmadand J. Judge. 2018a. Yield Forecasting of Spring Maize Using Remote Sensing and Crop Modeling in Faisalabad-Punjab Pakistan. J. Indian Soc. Remote Sens. 46:1701-1711.

Ahmad, I., A. Singh, M. Fahadand M.M. Waqas. 2020. Remote sensing-based framework to predict and assess the interannual variability of maize yields in Pakistan using Landsat imagery. Comput. Electron. Agric. 178:.

Ahmad, I., S.A. Wajid, A. Ahmad, M.J.M. Cheemaand J. Judge. 2018b. Assessing the Impact of Thermo-temporal Changes on the Productivity of Spring Maize under Semi-arid Environment. Int. J. Agric. Biol. 20: 2203-2210.

Ahmad, I., S.A. Wajid, A. Ahmad, M.J.M. Cheemaand J. Judge. 2019. Optimizing irrigation and nitrogen requirements for maize through empirical modeling in semi-arid environment. Environ. Sci. Pollut. Res. 26:1227-1237

Ahmed, I., A. Ullah, M.H. ur Rahman, B. Ahmad, S.A. Wajid, A. Ahmadand S. Ahmed. 2019. Climate Change Impacts and Adaptation Strategies for Agronomic Crops.. *In* Hussain, S. (ed.), Climate Change and Agricultur. 5th ed. IntechOpen, London.pp. 1–15

Ahmed, I., M.H. ur Rahman, S. Ahmed, J. Hussain, A. Ullahand J. Judge. 2018. Assessing the impact of climate variability on maize using simulation modeling under semi-arid environment of Punjab, Pakistan. Environ. Sci. Pollut. Res. 25:28413-28430.

Akhtar, I.U.2014. Pakistan Needs a New Crop Forecasting System. Available online: http://www.scidev.net/en/new-technologies/space-technology/opinions/pakistan-needs-a-new-crop - forecasting-system.html (accessed on 05 May 2020).

Anguita, D., L. Ghelardoni, A. Ghio, L. Onetoand S. Ridella. 2012. The'K'in K-fold Cross Validation. *In* ESANN.Proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, April.pp. 25-27.

Bertsimas, D. and J. Dunn. 2017. Optimal classification trees. Mach. Learn. 106:1039-1082.

Bosch, A., A. Zissermanand X. Munoz. 2007. Image classification using random forests and ferns. *In* 2007 IEEE 11th international conference on computer vision. Ieee.pp. 1-8.

Bouaziz, M., S. Eisoldand E. Guermazi. 2017. Semiautomatic approach for land cover classification: a remote sensing study for arid climate in southeastern Tunisia. Euro-Mediterranean J. Environ. Integr. 2: 24-29.

Breiman, L. 2008. Random forests. : 45–60Available at doi:10.1023/a:1010933404324.

Bro, Rand A.K. Smilde. 2014. Principal component analysis. Anal. Methods. 6:2812–2831.

Cheeseman, J. 2016. Food security in the face of salinity, drought, climate change, and population growth. pp. 111-123. *In* Halophytes for food security in dry lands. Elsevier.

Clark, B.J. Fand P.K.E. Pellikka. 2009. Landscape analysis using multiscale segmentation and object orientated classification. Recent Adv. Remote Sens. Geoinf. Process. L. Degrad. Assess. 8:323.

Das, P. and V. Pandey. 2019. Use of Logistic Regression in Land-Cover Classification with Moderate-Resolution Multispectral Data. J. Indian Soc. Remote Sens. 47:1443–1454.

Dempewolf, J., B. Adusei, I. Becker-Reshef, M. Hansen, P. Potapov, A. Khan and B. Barker. 2014. Wheat yield forecasting for Punjab Province from vegetation index time series and historic crop statistics. Remote Sens. 6:9653-9675.

Dubey, S.K., A.S. Gavli, S.K. Yadav, S. Sehgal and S.S. Ray. 2018. Remote Sensing-Based Yield Forecasting for Sugarcane (*Saccharum officinarum* L.) Crop in India. J. Indian Soc. Remote Sens. 46:1823-1833.

Fahad, M., I. Ahmad, M. Rehman, M.M. Waqas and F. Gul. 2019. Regional Wheat Yield Estimation by Integration of Remotely Sensed Soil Moisture into a Crop Model. Can. J. Remote Sens. 45: 770–781Available at https://doi.org/10.1080/07038992.2019.1692651.

Flood, N. 2014. Continuity of reflectance data between Landsat-7 ETM$^+$ and Landsat-8 OLI, for both top-of-atmosphere and surface reflectance: A study in the Australian landscape. Remote Sens. 6:7952-7970.

Franch, B., E. Vermote, S. Skakun, J.-C. Roger, I. Becker-Reshef and C. Justice. 2018. Enhancing Remote Sensing Based Yield Forecasting: Application to Winter Wheat in United States.. *In* IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE.pp. 8177–8180

Friedman, J., T. Hastieand R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33:1-20.

Funk, C., S. Shukla, W.M. Thiaw, J. Rowland, A. Hoell, A. McNally, G. Husak, N. Novella, M. Buddeand C. Peters-Lidard. 2019. Recognizing the famine early warning systems network: over 30 years of drought early warning science advances and partnerships promoting global food security. Bull. Am. Meteorol. Soc. 100:1011-1027.

Gao, B.-Cand R.-R. Li. 2017. Removal of thin cirrus scattering effects in Landsat 8 OLI images using the cirrus detecting channel. Remote Sens. 9: 834-845.

Goodin, D.G., K.L. Anibasand M. Bezymennyi. 2015. Mapping land cover and land use from object-based classification: an example from a complex agricultural landscape. Int. J. Remote Sens. 36:4702-4723.

Government of Pakistan. 2018. Economic survey of Pakistan. Econ. Advis. Wing, Financ. Div. Govt. Pakistan:pp. 29-30.

di Gregorio, A. 2005. Land cover classification system: classification concepts and user manual: LCCS. Food & Agriculture Org. Romevol? pp?

Heumann, B.W. 2011. An object-based classification of mangroves using a hybrid decision tree - Support vector machine approach. Remote Sens. 3:2440-2460.

Holmes, S. 2003. Bootstrapping phylogenetic trees: theory and methods. Stat. Sci.: 241-255.

Hughes, L.H., S. Streicher, E. Chuprikovaand J. du Preez. 2019. A cluster graph approach to land cover classification boosting. Data 4:10-20.

Jackson, J.E. 2005. A user's guide to principal components. John Wiley & Sons. New York.

Johnson, B.A., R. Tateishiand Z. Xie. 2012. Using geographically weighted variables for image classification. Remote Sens. Lett. 3:491-499.

de Keukelaere, L., S. Sterckx, S. Adriaensen, E. Knaeps, I. Reusen, C. Giardino, M. Bresciani, P. Hunter, C. Neiland D. Van der Zande. 2018. Atmospheric correction of Landsat-8/OLI and Sentinel-2/MSI data using iCOR algorithm: validation for coastal and inland waters. Eur. J. Remote Sens. 51:525-542.

Kim, D.-H., R. Narashiman, J.O. Sexton, C. Huangand J.R. Townshend. 2011. Methodology to select phenologically suitable Landsat scenes for forest change detection. p. 2613–2616. *In* 2011 IEEE International Geoscience and Remote Sensing Symposium. IEEE.

Kirby, M., M. Mainuddin, T. Khaliqand M.J.M. Cheema. 2017. Agricultural production, water use and food availability in Pakistan: Historical trends, and projections to 2050. Agric. Water Manag. 179:34-46.

Kuhn, M.and K. Johnson. 2013. Applied predictive modeling. Springer.New York

Lary, D.J., G.K. Zewdie, X. Liu, D. Wu, E. Levetin, R.J. Allee, N. Malakar, A. Walker, H. Mussa and A. Mannino. 2018. Machine Learning Applications for Earth Observation. p. 165–218. *In* Earth Observation Open Science and Innovation. Springer.

Leroux, L., C. Baron, B. Zoungrana, S.B. Traoré, D. Lo Seenand A. Bégué. 2015. Crop monitoring using vegetation and thermal indices for yield estimates: case study of a rainfed cereal in semi-arid West Africa. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9:347-362.

Li, Z.-L., B.-H. Tang, H. Wu, H. Ren, G. Yan, Z. Wan, I.F. Trigoand J.A. Sobrino. 2013. Satellite-derived land surface temperature: Current status and perspectives. Remote Sens. Environ. 131:14-37.

Liaw, A. and M. Wiener. 2002. Classification and regression by randomForest. R news 2:18-22.

Lindley, D.V. 1958. Fiducial distributions and Bayes' theorem. J. R. Stat. Soc. Ser. B: 102–107.

Liu, X., L. Zhang, M. Li, H. Zhangand D. Wang. 2005. Boosting image classification with LDA-based feature combination for digital photograph management. Pattern Recognit. 38: 887–901.

Lobell, D.B., G.P. Asner, J.I. Ortiz-Monasterioand T.L. Benning. 2003. Remote sensing of regional crop production in the Yaqui Valley, Mexico: estimates and uncertainties. Agric. Ecosyst. Environ. 94:205-220.

Man, C.D., T.T. Nguyen, H.Q. Bui, K. Laskoand T.N.T. Nguyen. 2018. Improvement of land-cover classification over frequently cloud-covered areas using Landsat 8 time-series composites and an ensemble of supervised classifiers. Int. J. Remote Sens. 39:1243-1255.

Masek, J.G., E.F. Vermote, N.E. Saleous, R. Wolfe, F.G. Hall, K.F. Huemmrich, F. Gao, J. Kutlerand T.-K. Lim. 2006. A Landsat surface reflectance dataset for North America, 1990-2000. IEEE Geosci. Remote Sens. Lett. 3:68-72.

McIver, D.K. and M.A. Friedl. 2001. Estimating pixel-scale land cover classification confidence using nonparametric machine learning methods. IEEE Trans. Geosci. Remote Sens. 39: 1959-1968.

Mohsin Waqas, M., E. Yasir Niaz, H. Rashid, M. Adnan Bodlah, Y. Niaz, S. Ali, H. Raza, M. Fahad, I. Ahmadand H. H. Shah. 2019. Impact of Climate Change on Rainfall in the Irrigated Indus Basin: a Case Study in the Lower Chenab Canal System. Big Data Agric. 2: 04-05Available at http://doi.org/10.26480/bda.01.2020.04.05.

Nagy, A., J. Fehérand J. Tamás. 2018. Wheat and maize yield forecasting for the Tisza river catchment using MODIS NDVI time series and reported crop statistics. Comput. Electron. Agric. 151:41-49.

Neinavaz, E., A.K. Skidmoreand R. Darvishzadeh. 2020. Effects of prediction accuracy of the proportion of vegetation cover on land surface emissivity and temperature using the NDVI threshold method. Int. J. Appl. Earth Obs. Geoinf. 85:101984-101999.

Pacheco, J., S. Casadoand S. Porras. 2013. Exact methods for variable selection in principal component analysis: Guide functions and pre-selection. Comput. Stat. Data Anal. 57: 95-111.

Punn, M. and N. Bhalla. 2013. Classification of wheat grains using machine algorithms. Int. J. Sci. Res. 2:363-366.

RCore, T. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http//www. R-project. org.

Rodriguez-Galiano, V.F., B. Ghimire, J. Rogan, M. Chica-Olmoand J.P. Rigol-Sanchez. 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J. Photogramm. Remote Sens. 67:93-104.

Roell, Y.E., A. Beucher, P.G. Møller, M.B. Greveand M.H. Greve. 2020. Comparing a Random-Forest-Based Prediction of Winter Wheat Yield to Historical Yield Potential. Agronomy 10:395-405.

Saeed, U., J. Dempewolf, I. Becker-Reshef, A. Khan, A. Ahmadand S.A. Wajid. 2017. Forecasting wheat yield from weather data and MODIS NDVI using Random Forests for Punjab province, Pakistan. Int. J. Remote Sens. 38:4831-4854.

Saporta, Gand N. Niang. 2009. Principal component analysis: application to statistical process control. Data Anal.Vol: 1-23-29.

Scaramuzza, P. and J. Barsi. 2005. Landsat 7 scan line corrector-off gap-filled product development.. *In* Proceeding of Pecora.pp. 23–27

Tharwat, A. 2016. Linear vs. quadratic discriminant analysis classifier: a tutorial. Int. J. Appl. Pattern Recognit. 3:145-180.

Ullah, A., I. Ahmad, A. Ahmad, T. Khaliq, U. Saeed, M. Habib-ur-Rahman, J. Hussain, S. Ullahand G. Hoogenboom. 2019. Assessing climate change impacts on pearl millet under arid and semi-arid environments using CSM-CERES-Millet model. Environ. Sci. Pollut. Res. 26: 6745–6757Available at https://doi.org/10.1007/s11356-018-3925-7.

Ullah, A., A. Ahmad, T. Khaliq and U. Saeed. 2018. Optimizing nitrogen rate of Pearl Millet under arid and semi-arid. p. 287–288. *In* 2oth Nitrogen workshop coupling C-N-P-S cycles. Couvent des Jaccobins Conference Centre, Rennes, France.

Vanli, Ö., I. Ahmadand B.B. Ustundag. 2020. Area Estimation and Yield Forecasting of Wheat in Southeastern Turkey Using a Machine Learning Approach. J. Indian Soc. Remote Sens.: 1-10-21.

Vanli, Ö., B.B. Ustundag, I. Ahmad, I.M. Hernandez-Ochoaand G. Hoogenboom. 2019. Using crop modeling to evaluate the impacts of climate change on wheat in southeastern turkey. Environ. Sci. Pollut. Res. 26:29397-29408.

Waqas, M.M., U.K. Awan, M.J.M. Cheema, I. Ahmad, M. Ahmad, S. Ali, S.H.H. Shah, A. Bakhshand M. Iqbal. 2019. Estimation of canal water deficit using satellite remote sensing and GIS: A case study in lower chenab canal system. J. Indian Soc. Remote Sens. 47: 1153–1162.

Waqas, M.M., S.H.H. Shah, U.K. Awan, M. Waseem, I. Ahmad, M. Fahad, Y. Niazand S. Ali. 2020. Evaluating the Impact of Climate Change on Water Productivity of Maize in the Semi-Arid Environment of Punjab, Pakistan. Sustainability. 12:3905.

Ye, J., R. Janardaand Q. Li. 2005. Two-dimensional linear discriminant analysis. *In* Advances in neural information processing systems.pp.1569-1576.

[Received 10 Jun 2020; Accepted 20 Nov 2020; Published (online) 18 April 2021]