Mapping Wheat Crop Phenology and the Yield using Machine Learning (ML)

Muhammad Adnan[1]
Institute of Manufacturing Information and Systems, Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan City 701, Taiwan

Abaid-ur-Rehman[2], M. Ahsan Latif[3], Naseer Ahmad[4]
Department of Computer Science,
University of Agriculture, Faisalabad, Pakistan

Maria Nazir[5]
Department of Computer Science,
COMSATS University Islamabad, Lahore, Pakistan

Naheed Akhter[6]
Department of Computer Science,
GC University Faisalabad, Pakistan

***Abstract*—Wheat has been a prime source of food for the mankind for centuries. The final wheat grain yield is the multitude of the complex interaction among the various yield attributes such as kernel per plant, Spike per plant, NSpt/s, Spike Dry Weight (SDW), etc. Different approaches have been followed to understand the non-linear relationship between the attributes and the yield to manage the crop better in the context of precision agriculture. In this study, Principle Component analysis (PCA) and Stepwise regression used to reduce the dimension of the original data to get the critical attributes under study. The reduced dataset is then modeled using the Radial Basis neural network. RBNN provides the regression value more than 0.95 which indicates the strong dependence of the yield on the critical traits.**

***Keywords*—RBNN; PCA; stepwise regression; attributes; yield**

I. Introduction

Wheat is the major agriculture crop of the Pakistan. It acts as a back bone of agriculture for the food security throughout the world. The prediction of wheat yield is too much important. The demand of the wheat has been doubled from last many decades. The demand is increasing day by day due to many factors. This may be due to many climate changes in environment. Increasing population also affect the growth and demand of the wheat. Therefore, wheat is becoming very important crop from last many years. Importance of wheat in the economy of the world is clearly reflected by its share of 15 % to the total arable land in the world for the year 2011-12 [3].

Now a days, the thing that is alarming and challenging for scientists is gap between production and the demand of wheat. Most of the people demand wheat as food .The reason is that need of wheat has been increased and it is becoming difficult to fulfill the demands. To meet such kind of challenges, it is needed to increase the total area for the agriculture land. With increasing the land area, wheat at huge quantity can be produced. There is another effort to increase the production of wheat from the present growing area.

The approach used to analyze the relation among yield and traits was machine learning (ML). Machine Learning technique has an ability to deal with high dimension problem by using less computational power. Apply machine learning in order to analyze the high number of trades to find the most relevant crop for better agriculture production. Apply machine learning algorithm for the classification of yield component in order to get high wheat yield. Principle Component Analysis (PCA) and step wise regression techniques applied on data to get the reduced dimensional data. Both techniques analyzed the data as according to its nature of effectiveness. As a result, the trimmed and the most dependent data set is achieved. New Data set collected and applied Radial Basis Neural Network (RBNN) on reduced data and got significant results. A work related to the yield measurement was conducted in which estimation of seed and grain corn yield was done on the basis of input data. ANN model with back propagation algorithm was used. The ANN model worked best with 6-4-8-1 and 6-3-9-1 structure for prediction of the yield. The result of this model is compared with multiple linear regression model. The result was approximately 95%.

The distribution of paper is following, Section 2: Related work, Section 3: Material and method, Section 4: Result and discussion, Section 5: Conclusions following with References.

II. Related Works

Adnan (2018) [1] studied the impact of water supply on wheat yield with the help of Lasso and Radial; machine learning techniques and the result of lasso Radial technique accuracy was 89% corresponding other machine learning techniques. In this study Relative water contents, waxiness, grain per spike and plant height used for experiments. Different techniques used in this study and result is clearly show that growth of wheat is highly affected by water stress. Normal-values. “Awnlength”, “pendulacnelength”, “extractionlength” and “noofdaysheading” variation is low in water stress condition as compared to Yield and TGW. Wheat yield and growth affected by water condition. In this study different techniques used for find the relationship of yield of

wheat and other variables and neural network gave the best result.

Adnan et al. (2017) [2] used the machine learning method for observation the evapotranspiration rate in Faisalabad region. In this study PCA techniques used for reduced the data set dimension because the information lost minimum with this technique. PCA gave the new variable after reduction of data set, the value of regression is 0.83426. A time series Neural Network used after getting reduced data set from PCA technique. Time series give the accurate result as compared to other Neural Network techniques. The accuracy of this model is 83%.

Awan et al., (2015) [4] described that 176 different types of genetic wheat traits were used to evaluate variety of traits practically multivariable analysis. Analysis revealed a simple correlation that indicated that there was major positive relationship of yield weight with cell membrane solidity, osmotic modification and transpiration and adverse relationship leaf area. Study also revealed significance of physical traits and their effect on grain mass. Multivariable analysis which included factor and cluster analysis showed that variable genetic pool was sufficient for breeding design. Wheat yield of each plant was strictly correlated to water substances, cell membrane solidity and leaf area. For more deep analysis, eight groups of different traits were made and study revealed that groups with smaller genetic distance were effective for breeding.

Mukhtar (2015) [5] stated that areas where major source of water is rain, has significant effect on wheat grain quality and yield because weather circumstances are randomly vary. This climate fluctuation provides chance of improving wheat grain yield production. These fluctuations and variability were studied years after year w.r.t. regions and sowing techniques and then wheat grain yield was analyzed. For this study, field tests were practiced by using three genetic traits, three different locations for the period of two years in rain source of water. Under these variations and conditions, wheat grain quality and mass resulted significant change. In region where sowing was delayed, temperature was high and water was stressed, show increased grain yield quality. However contradictory results were observed in opposing climate and water absorbing conditions. Fluctuation and variability in climate conditions had significant influence on wheat grain yield and inverse relationship was experienced among climate conditions, wheat yield and grain quality. Hence we can conclude that weather conditions, area of cultivation, sowing techniques, temperature and water can effectively alter the quality of wheat grain yield.

Emamgholizadeh et al., (2015) [6] described that in the agricultural research the most vital purpose of breeding is production of seed yield. In account to this research two techniques were used, artificial neural network and multiple regression model. Both methods were used to predict the same seed yield on the basis of premeasured features of plant like, maximum flowering days, height of the plant in centimeters, numbers of capsules of each plant and weight of seed and seed numbers. Results were tested by using both MLR and ANN techniques and it resulted that ANN was more accurate w.r.t. root mean square error and founded coefficient. It was found also that ANN technique was better than MLR. At the end it was examined that this analysis had large and small significant effects on the same w.r.t. numbers of capsule and flower time for each plant. So in result ANN method is better for predicting seed yield than MLR and it predicts more accurately.

Khoshnevisan et al., (2014) [12] described the relation between energy consumption and crop yield in order to get sustainable agriculture they develop adaptive neuron-fuzzy system to predict wheat grain yield on the basis of energy input. The developed ANN was MLP with 11 neuron in input layer and 32 and 10 neuron in hidden layer. The result showed that ANFIS gave more accurate result than other ANN.

Paswan and Begum (2014) [7] described that how important it is for the policy maker to know about the approximate yield of crop. Computer scientists are working for making exact prediction about the yield. The crop area and crop production (maize) of Assam using ANN. They used MLP with radial bias function network which has been trained with metrological data and maize production data from various sources. The accuracy of this model was measured by using RMSE and correlation coefficient. It was observed that this model had performed better as compared to other statistical model.

Bagheri et al., (2015) [8] stated that land survey is important for crop yield prediction. Comprehensive survey may be expensive and time consuming. Since soil survey is important as it provides information for agricultural needs. Hence, there must be rapid and precise soil survey map. For this ANNs perceptron were purposed to survey map soil elements Digital Evaluation Model (DEM) features. Various multilayer ANNs were developed having input dataset and hidden element layers. This technique is implemented and tested to cumulate accuracy of interposed and inferred data. From result it was obvious that soil organization had a direct influence on accuracy of results. Errors were very small and low. Almost all techniques of ANN methods training errors were less than 11 percent. While testing and certifying, errors were 50 and 70 percent respectively. To attain superior predictions, in addition with DEM features, dataset related to lands in term of soil-farming elements must be given to ANNs perceptron as well.

Kogan et al., (2013) [9] stated that Ukraine is the biggest agricultural production country around the globe. Time management and production estimate are main elements of yield security. This study reveals wheat yield proficiency using oblast management with satellite resolution. Oblast is multinational statistics study division in European Union. Observations were made in rain fed region and average data were collected from MODIS sensing device at 250 m spatial resolution and used in a regression technique for estimating wheat yield. For reliable wheat yield projection root mean square error was acknowledged. In case of many oblasts, values which were taken in April to May using NDVI, when matched with official statistics it gave minimum root mean square error. The NDVI technique was matched with empirical model and WOFOST growth simulation applied in

CGMS, all these comparisons provided minimum RMSE. This study and comparison of wheat yield production was done totally on independent values for the period of 2010 to 2011. The most accurate forecast was predicted in 2010 via CGMS which provided root mean square error value 0.3 t ha^{-1} in June and 0.4 t ha^{-1} in April. So, it was concluded that empirical NDVI based regression was parallel to CGMS when forecasting wheat yield at oblast level.

Hung et al., (2013) [10] described that in this paper forecasting the fruit yield, multi-scale machine learning technique was used at different divisions. In this learning technique, algorithm is so flexible and usable that it can be applied at various divisions of problems. So, this approach was applied to large variety trees for fruit yield forecast. A comprehensive test was conducted on apple orchard which consisted of eight thousand images for learning. This test showed that algorithm was most fit to apple segment of various colors and sizes. Segmentation outcomes were used to count fruit and then to compare with manual counting. Squared correlation coefficient resulted from this study was R^2=0.81.

Alvarez (2009) [11] described that to get a model for reasonable yield prediction and grain production estimate, an analysis was conducted in Argentine grasslands in terms of soil characteristics and climate features on wheat yield. Data record collected from soil and climate analysis were implemented. Data of wheat yield production from all over the country was tested at geomorphological level. Grasslands were divided into 10 sub-regions units and from these sub-regions 10 growth seasons were recorded from 1995-2004. For data analysis, surface regression (SR) and artificial neural networks (ANNs) techniques were implemented. Yield of wheat was concentrated with water holding capacity of soil and organic corban of soil. Climate features on yield was strong rainfall over crop potential evapotranspiration (R/CPET). Surface regression design was implemented on 64 percent of model to predict yield variance, however this design has not performed better prediction of yield. Then ANN design was tested and it gave 76 percent of yield prediction variability. So, ANN developed model was good to predict wheat yield production in Argentine grasslands.

III. Material and Methods

A field experiment was conducted in the University of Agriculture Faisalabad, where the growth of wheat yield was observed. That experiment was completed in two years. Where the yield was classified according to its trait values are shown in Table I. These traits were Grain Yield (GY), Kernels Per Plant (K/P), Weight/Kernel Size (KS), Number of Spikes per Plant (S/P), Number of Fertile Spikelet's per Spike (NSpt/S), Maximum Fertile Loret per Spikelet (MFFI/Spt), Spike Dry Weight (SDW), Plant Height (PH), Spike Length (SL), Awn Length (AL), Spikelets Density (SD) and Chlorophyll Contents (CC). The value of each trait was saved and was processed to find out the relation with respect to yield. In machine learning, used different approaches for classification and to find the relation of yield and variables.

TABLE I. List of Variables with Acronyms

GY	Grain Yield
K/P	Kernel per plant
K/S	Kernel size
NSpt / S	Number of spikes per plant
MFFI / Spt	Number of fertile spikelet per spike
SDW	Spike dry weight
PH	Plant height
SL	Spike length
AL	Awn length
SD	Spikelet's density
CC	chlorophyll contents

A. Principal Component Analysis (PCA)

PCA is a quantitatively rigorous method for achieving problem of relation. This method generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. The outcome data is thus divided according to the variation in data set. The variation is done as according to the level of affecting of the data. The higher data variation is plotted first in the list of the graph. The very first plotted line represents the most variation variable. Same other plotted lines represent the gradual decreased variation of data set.

$$\text{Coeff} = \text{PCA}(X) \quad (1)$$

1) Standardize: We also remove unwanted data from the spread sheet data. Since PCA yields a characteristic subspace that maximizes the difference along the axes, assuming that it might have been measured on diverse scales. The conversion of the information into unit scale will be a pre requisite for those ideal executions of many machine learning algorithms.

2) Calculate covariance: Covariance is a measure of the degree with which components comparison is initiated for requested information move in those same heading. We find the covariance among the wheat yield and other traits. Actually we want to measure how different trait and yield depend upon each other. The formula for calculating the covariance of the variables X and Y is

$$\sum ni = 1\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)n - 1 \quad (2)$$

With $\bar{x}$ and $\bar{y}$ denoting the means of X and Y, respectively. X denotes the input variable and Y denotes the output variable. Equation (2) helping in measure, how wheat yield depends upon the important variable like Kernels per Plant (K/P), Number of Spikes per plant (S/P), Number of Fertile Spikelet's per Spike (NSpt/S), Spike Dry Weight (SDW), spike length (SL), Spikelet's density (SD) that was used as input. These all above following variables are treated as Y.

3) Selecting principal components: That ordinary objective of a PCA is to decrease that dimensionality of the first characteristic space by projecting it onto a more abstract subspace, the place the eigenvectors will appear on those axes. However, those eigenvectors best define the directions of the new axis, since they have all the same unit length.

In this step, the PCA processes the data. The resulting value or the set of outcome which we have derived from PCA appeared in the form of eigenvector. Here, our new data after processing will be represented as eigenvector. Each principle component is a different eigenvector. The PCA has reduced the data dimension on the basis of dependency, i.e., from eleven traits into six traits. The following relation reveals the eigenvalue of an eigenvector.

$$\Sigma v = \lambda v \quad (3)$$

In equation (3):

Σ=Covariance matrix v=Eigenvector λ=Eigenvalue

To choose which eigenvector we need to drop from our lower-dimensional subspace, we must examine the relating eigenvalues of the eigenvectors. Approximately speaking, the eigenvectors for the least eigenvalues bear the slightest majority of the data over those distribution of the data and those need to be dropped. PCA provides the most significant traits by reducing the dimension of data.

4) Transforming the samples into new subspace: In the last step, we use-dimensional matrix W that is computed to transform our samples into the new subspace as per the following equation. The transformed new traits are used for estimating the wheat yield and equation is given below:

$$Y=W^t \times X \quad (4)$$

B. Stepwise Regression

Stepwise regression includes regression models in which the choice of predictive variables is carried out by an automatic procedure. Stepwise regression creates a linear model and automatically adds to or trims the model. The priority in the regression model is measured according to the significant importance of the data. The data has more impact as it is added to the regression model. The data with lesser effectiveness is trimmed from the model. Only the data that is most relevant has produced targeted values whereas all the other data which is not relevant to the targeted values is discarded. This technique actually has reduced the data dimension and has given low dimensional data but highly correlated.

The stepwise model performs a multilinear regression of the response values in the n-by-1 vector y on the p predictive terms in the n-by-p matrix X. Distinct predictive terms should appear in the different columns of X with b as a p-by-1 vector of estimated coefficients for all of the terms in X. If a term is in the final model, the coefficient estimated in b for that term is a result of the final model.

C. Data Modeling

The radial basis function network is a viable alternative approach in machine learning for regression measurement in data dependency relation. A common learning algorithm for radial basis function networks is based on first choosing randomly some data points as radial basis function centers and then using singular value decomposition to solve the weights of the network. The procedure chooses radial basis function centers one by one in a rational way until an adequate network has been constructed. Here, this approach is applied on data set that was collected after obtaining the result of the PCA and the step wise regression.

In the field of mathematical modeling, a radial basis function network is an artificial neural network that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. Radial basis function networks have many uses, function approximation, time series prediction, classification, and system control. Radial basis networks can require more neurons than standard feed-forward back-propagation networks, but often they can be designed in a fraction of the time it takes to train standard feed-forward networks. They work best when many training vectors are available. Here this network output layer consists of a single neuron.

IV. Result and Discussion

A. PCA Result

The dataset processed using PCA technique consists of 12 variables. Each variable is of different characteristics with respect to the yield production. The PCA result is shown in the Fig. 1 as each bar represents a specific principal component. The height of each bar represents the level of variation. The first component in the graph has more than 28% of the total variation in the dataset. The higher variation in a principal component reflects its significant relation with the outcome variable. The first eight principal components contribute 85% of the total variation. We can take into account this as PCA has reduced the dimension of the data by neglecting other four variables because of their least impact and variation in the graph.

Here, in Fig. 2 the graph shows PC_1 along x-axis and PC_2 along y-axis. Dependency of variables can be found out if its coefficient value is definable. From the figure, it is clear that "GY" coefficient value is higher among all other components and that is 0.46 that defines its significant role in defining the variation for the very first principal component. The trait "SDW" shows 0.41 values on the graph. Same as other variables contributing in the first component reflect their behavior from coefficient values. In component 2, which is along the y-axis the "KP" has 0.49 higher values which is

higher among all other traits. The coefficient value of "KP" in 2nd component has significant importance in 2nd principle component behavior. Similarly, it has been observed that the major contributions for PC_3 and PC_4 come from "NSpt/S" and "KS". So, from the PCA based analysis, we concluded that the variables "GY", "KP", "KS", "NSpt/S", etc. play critical role in the final yield production.

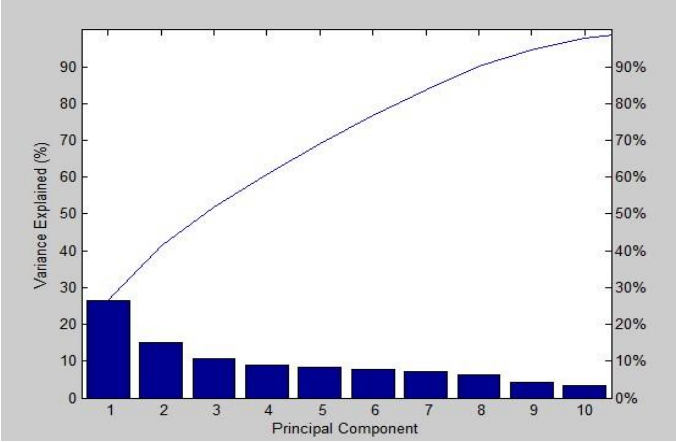

Fig. 1. PCA vs Variance.

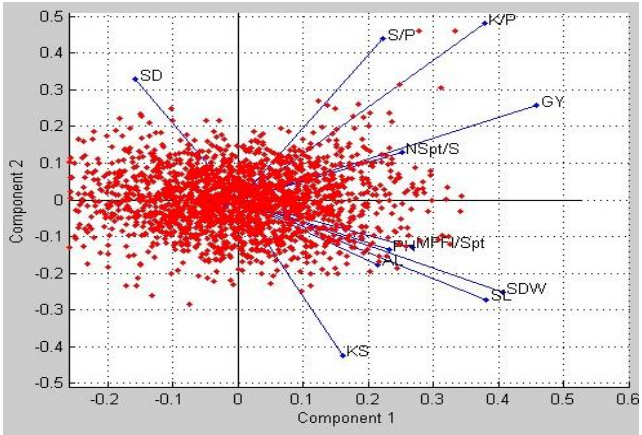

Fig. 2. PCA of the Yield and Traits.

B. Stepwise Regression Result

The stepwise regression works automatically to add or remove the predictive variables. It works best where there is large space of search. Final columns included: 1 2 3 4 6 9.

Table II explains that the model starts working when no column is added there. In the first step, it adds the kernel per plant and the value of predictive terms p is zero. In the second step, it adds the kernel size variable and the value of p =0. In the 3rd step, it adds the spike per plant variable in the stepwise model then the value of p = 2.1620e-09. In step 4, it adds the number of fertile spikelet's per spike variable and the value of p= 0.0076. In step number six it adds the spike dry weight variable into model and the value of p= 5.2410e-06. Finally model adds Awn length, Spikelets density and Chlorophyll contents. In this process, six variables have been considered out of the total eleven variables. These six variable are "GY", "KP", "KS", "SP", "NSpt/S" and "SL". The same variables also have been recognized by the PCA. So by the stepwise regression we have concluded that "GY", "KP" "KS", "NSpt/S" show closer relation and dependency to yield in our data set. Yield production is highly dependent on these traits. This shows that stepwise also reduced the dimension of variables.

TABLE II. STEPWISE PREDICTIVE VARIABLES

'Coeff'	'Std.Err.'	'Status'	'P'
[0.0375]	[3.3429e-04]	'In'	[0]
[0.7521]	[0.0119]	'In'	[0]
[0.1126]	[0.0187]	'In'	[2.1620e-09]
[0.1044]	[0.0391]	'In'	[0.0076]
[-0.0636]	[0.1616]	'Out'	[0.6941]
[0.5198]	[0.1138]	'In'	[5.2410e-06]
[0.0113]	[0.0071]	'Out'	[0.1103]
[0.0313]	[0.0575]	'Out'	[0.5860]
[0.1517]	[0.0543]	'In'	[0.0053]
[-0.1938]	[0.2978]	'Out'	[0.5153]
[0.0042]	[0.0048]	'Out'	[0.3841]

In RBNN model trained the neural network under the data set of total eleven traits. Here, a single layered architecture is used. This model consists of 100 numbers of neurons in hidden layer and has one output layer which conventionally contained single neuron. Here radial basis function (RBF) was used as an activation function. Number of epoch in that model was 100. The obtained result from this experiment is shown in Fig. 3, the regression graph the value of regression R is 0.97695. Regression value indicate that traits and yield dependency is greater than 95%.

The model performance is best for validation value 2.6538 at Epoch number 44 which is shown below in Fig. 4. The total 44 Epoch is run by the model. The dotted line indicates the best mapping found.

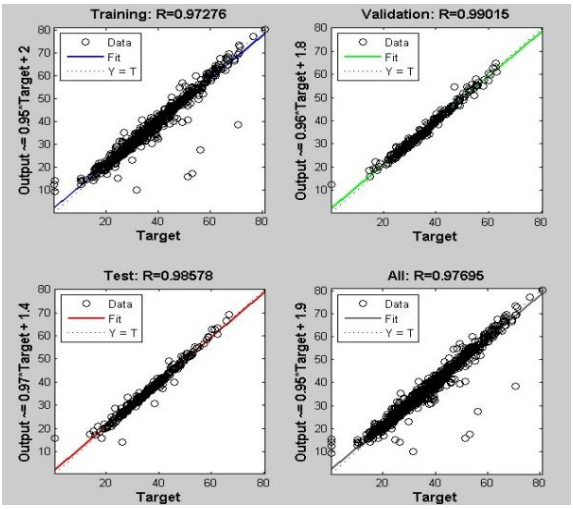

Fig. 3. Validation of Data.

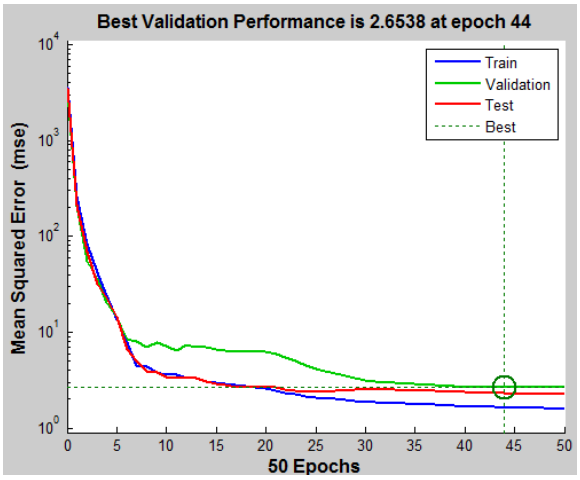

Fig. 4. Performance of Data.

It has been observed that by applying PCA, we have got the results with greater accuracy. In this way, we have reduced the computational time and power by using reduced and new variables provided by the PCA. The reduced variables provide almost same results as we have got from all the available variables to map the yield.

We also have applied some different techniques and methods in this neural network model to have different regression values as shown in Table III that help to find out the best relation among the traits and yield.

TABLE III. DATA MODELING ANALYSIS

	Sr#	Input			Output
		No. of Neuron	Activation Function	Training Function	
Stepwise Regression	1	30	Logsig	trainlm	0.967
	2	40	Tansig	trainlm	0.949
PCA	3	30	Logsig	trainlm	0.955
	4	40	Tansig	trainlm	0.943

V. CONCLUSION

Crop modeling is an active research area which finds its roots in the dire need to understand the mutual relationships within the crop variables. These mutual relations either linear, nonlinear or stiff in nature govern the overall crop progress and hence the yield. In this study, some of the machine learning techniques used to understand and model these relationships. The results found in this research are positive as these are highly correlated with the field results. In future work, specifically focus on the nonlinear relations which exist within these crop variables and the machine learning approaches to control that.

REFERENCES

[1] Adnan, M., Akhter, N., Abid, M., Latif, M.A., Abaid-ur-Rehman and Kashif, M., 2018. Studying the Impact of Water Supply on Wheat Yield by using Principle Lasso Radial Machine Learning Model. *International journal of advanced computer science and applications*, 9(2): 229-235

[2] Adnan, M., Latif, M.A. and Nazir, M., 2017. Estimating Evapotranspiration using Machine Learning Techniques. *International journal of advanced computer science and applications*, 8(9): 108-113.

[3] Khan, M.U., Malik, R.N. and Muhammad, S., 2013. Human health risk from heavy metal via food crops consumption with wastewater irrigation practices in Pakistan. *Chemosphere*, 93(10):1-8.

[4] Awan, S.I., Ahmad, S.D., Ali, M.A., Ahmed, M.S. and Rao, A., 2015. Use of multivariate analysis in determining characteristics for grain yield selection in wheat. Sarhad *J. of Agric*., 31: 139-150.

[5] Ahmed, M., 2015. Response of spring wheat (Triticum aestivum L.) quality traits and yield to sowing date. *PloS one* 10(4):40-56.

[6] Emamgholizadeh, S., Parsaeian, M. and Baradaran, M., 2015. Seed yield prediction of sesame using artificial neural network. *European Journal of Agronomy*, 68: 89-96.

[7] Paswan, R.P. and Begum, S.A., 2014, February. ANN for prediction of Area and Production of Maize crop for Upper Brahmaputra Valley Zone of Assam. *In Advance Computing Conference (IACC), 2014 IEEE International* : 1286-1295.

[8] Bagheri Bodaghabadi, M., Martínez Casasnovas, J.A., Salehi, M.H., Mohammadi, J., Esfandiarpoor Borujeni, I., Toomanian, N. and Gandomkar, A., 2015. Digital soil mapping using artificial neural networks and terrain-related attributes. *Pedosphere* 25 (4): 580-591.

[9] Kogan, F., Kussul, N., Adamenko, T., Skakun, S., Kravchenko, O., Kryvobok, O., Shelestov, A., Kolotii, A., Kussul, O. and Lavrenyuk, A., 2013. Winter wheat yield forecasting in Ukraine based on Earth observation, meteorological data and biophysical models. *International Journal of Applied Earth Observation and Geoinformation*, 23: 192-203.

[10] Hung, C., Underwood, J., Nieto, J. and Sukkarieh, S., 2015. A feature learning based approach for automated fruit yield estimation. *In Field and Service Robotics, Springer* : 485-498.

[11] Alvarez, R., 2009. Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach. *European Journal of Agronomy* 30(2): 70-77.

[12] Khoshnevisan, B., Rafiee, S., Omid, M. and Mousazadeh, H., 2014. Development of an intelligent system based on ANFIS for predicting wheat grain yield on the basis of energy inputs. *Information processing in agriculture* 1(1): 14-22.