ANN-Based Wheat Crop Yield Prediction Technique for Punjab Region

Nishu Bali and Anshu Singla

1 Introduction

Agriculture is an important occupation in almost all the countries of today. The ever growing population of the world and unexpected changes in climatic trends is mounting pressure on global food supply to assure food for everyone. Thus, efforts are made to find ways of increasing the production which in turn requires correct and timely estimation of crop yield.

Crop yield is dependent on different environmental parameters which vary nonlinearly and in turn make the estimation of yield a complex procedure. Two general approaches are generally followed for crop yield estimation–crop growth model and statistical models. Crop growth models are quite efficient in estimation of yield, but they require thorough knowledge of crop physiological behaviour at different stages of growth, and their results are sometimes not transferrable or reusable on the fields due to varying environmental conditions. In statistical models, empirical equations are formulated with yield as dependent variable and factors affecting the yield as independent variables. The emergence of machine learning in the field of agriculture has given strength to the statistical method of yield estimation.

N. Bali · A. Singla (✉)
Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab, India
e-mail: anshu.singla@chitkara.edu.in

N. Bali
e-mail: nishu.bali@chitkara.edu.in

2   Machine Learning in Agriculture Domain

In machine learning techniques, machine is made to learn through the given data. On the basis of learnings attained by the machine, it predicts or classifies the unseen data. Machine learning has shown its role in various fields of agriculture such as crop selection, in assessing the effects of various climatic and soil parameters, crop disease detection and prediction, precise and only required use of water for irrigation and many more. Among these, yield estimation is one of the most important field in which machine learning has shown an appreciable contribution. There are broadly two types of machine learning algorithms: supervised and unsupervised. In supervised machine learning techniques, the machine is trained for already known outputs based on some inputs. The examples are linear regression, logistic regression, support vector machine, decision tree, Bayesian logic and neural networks. In unsupervised techniques, the machine is given raw data and is made to identify the patterns and classify the data according to the identified patterns or trends. These include clustering, KNN and Apriori algorithm. Whether supervised or unsupervised, the machine is first trained on some training data and is made to use that learning for prediction or classification on unseen data. Both types of techniques have contributed in the study of crop yield estimation. The present study is done with an objective to use ANN technique for wheat crop yield estimation for specific region of Punjab and to compare its efficiency with multivariate regression technique.

In next section, we present a brief literature on the works already done in the field by the researchers. This will be followed by dataset and methodology and experimental results in fourth and fifth section, respectively.

3   Related Work

The effect of climatic variations on yield of wheat crop was studied using various machine learning techniques. Support vector regression was compared with other approaches used with NDVI index and results proved that SVR outperformed the latter approach with $R^2 < 0.46$ [1]. In another study, linear regression model was used for quantifying the effect of different meteorological parameters on the rice yield in district Raipur, India. It was found that maximum temperature increase had not much detrimental effect at tillering stage of plant growth but had widespread effect at flowering stage. Minimum temperature was within the cardinal limits so was not much affecting the yield. Rainfall and sunshine were found to be prominent parameters affecting the yield [2]. Remote sensing data was used for finding the efficiency of machine learning methods for predicting yields and results concluded that a combination of sensor technology along with machine learning techniques can give even better results [3]. The use of deep learning techniques like CNN was explored in a study on orchards where fruit bearing capability of bitter melon crop was analysed based on the leaves of plant gathered from Ampalaya farms [4]. In yet another study

on orchards, two BPNN models were explored for two phases of season, opening and ripening period, for estimation of yield of fruit crops based on image analysis. Satisfactory results obtained proved the efficiency of proposed approach in area of yield estimation [5]. A comparative analysis of four machine learning techniques for corn yield estimation was done in Iowa State. Results gave good results especially for deep learning which showed most stable results [6]. In another study, spiking neural network technique was used for spatio temporal analysis of data for crop yield evaluation. The study made pre-harvest yield prediction six weeks prior to harvest with an accuracy of 95.4% and average error of prediction of 0.236 t/ha and correlation coefficient of 0.801 using a nine-feature model [7]. In another study, three machine learning techniques, counter-propagation artificial neural networks (CP-ANNs), XY-fused networks (XY-Fs) and Supervised Kohonen Networks (SKNs), were compared in performance for finding the variations of wheat yield based on multilayer soil data and satellite imagery crop growth characteristics. Results showed that in low yield class varieties, the accuracies obtained were 91%; whereas for average and high pitched yield varieties, the accuracies were 70% and 83%, respectively. Among the three machine learning models, SKN showed highest accuracy of 81.65% proving it to be the best model [8]. ANN was explored in another study for efficiency in rice yield prediction for the years 1998–2002. The results gave high accuracies of 97.5% with a sensitivity of 96.3 and specificity of 98.1 [9]. The effect of customization of ANN model on its efficiency for wheat yield estimation was studied. The customized model was compared with default ANN model and MLR technique. Significant improvement in efficiency was found in customized ANN model with higher R2 statistics and lower percentage errors [10]. Another extensive study was done to compare various machine learning techniques for crop yield estimation of multiple crops. Results favored M5-Prime and KNN techniques with lowest error values [11]. In yet another study, the architecture of ANN model was varied by wavering number of hidden layers used, and the effect of variations on efficiency of model was evaluated for finding the effect of various predictor variables related to soil and climate on yields of various crops [12]. A new hybrid approach based on modern representation learning ideas was proposed to predict county-level soybean crop yield. A new dimensionality reduction technique was used to compensate for lack of sufficient training data. Deep learning architectures like CNNs and LSTMs were used to predict the crop yield. Experimental results showed that proposed model had outperformed the customary remote sensing centered techniques in efficiency [13]. Crop yield prediction in area of greenhouse operations was studied using an intelligent system called EFuNN (Evolving Fuzzy Neural Network) for yield estimation of tomato crop. Results gave weekly prediction with an accuracy of 90% [14]. Customization of ANN models was explored in yet another study in which 11 varied ANN models with different number of neurons in hidden layers were tried, and optimum model was selected. ANN-MLP model based on conjugate gradient back propagation algorithm reported lowest MAPE making it the preferred or optimum model [15].

Fig. 1 Map showing Ludhiana Region of Punjab [16]

4 Dataset and Methodology

4.1 Study Area

The present study is focused on one of the main agriculture-based districts of Punjab, Ludhiana. The district is spread across a geographical area of 3767 km^2 and has 3 lakh ha of net sown area out of which almost 100% is doubly cropped and in some cases, three crops are sown in a year (Fig. 1).

Ludhiana has always been a role model for other districts as far as adoption of advanced techniques in agriculture is concerned. Wheat is one of the most important crop sown in the area. Around 2.57 lac ha of the area is devoted to wheat cultivation which contributes to 50.26 qt/ha of productivity of the crop from district. The data used in the study is mainly collected from statistical abstract of Punjab issued by Economic advisor to Government, Punjab. An extensive data of 43 years from 1970 to 2010 has been used for the study. The climatic data was obtained from meteorological department of Punjab.

4.2 Methodology

The data obtained from various sources was pre-processed. The processed data was then partitioned into train and test data. Although there is no exact protocol to divide the data into test and train, but we used 1:9 ratios, i.e., 10% data was taken as test data and 90% was taken as train data. Model after being trained with the training data was tested for the accuracies of prediction.

4.2.1 Data Pre-processing

The data obtained from various sources was scrutinized to find the occurrence of any null values in the data which were substituted with appropriate statistical values. As machine learning techniques can only work on numeric data, the features selected for the study were examined to find any non-numeric parameters in the study. Out of the annual data obtained from reports, the data for specific months actually used for the wheat cultivation ranging from October (sowing period) to April (harvest period) was selected and compiled. Environmental parameter values (maximum and minimum temperature, maximum and minimum relative humidity, rainfall and evaporation) pertaining to this period were selected and stored in an excel sheet. The data was normalized and scaled.

4.3 Machine Learning Techniques

The crop yield of wheat in Punjab region has been performed by employing artificial neural network. Present study compared the results obtained with the machine learning technique-multivariate linear regression. Sections 4.3.1 and 4.3.2 briefly describe both the techniques.

4.3.1 Multivariate Linear Regression

Linear regression is a supervised machine learning technique in which target value is determined based on some independent variables related to the target variable. Regression technique is mostly used for finding the relationships between the target and independent variables. As this technique deals with linear relationships between the variables, it is called linear regression. The function of a linear regression is defined as:

$$y = \theta_0 + \sum_{i=1}^{n} \theta_i * x_i \tag{1}$$

where x is input variable and y is output or target variable.

θ_0: intercept.

θ_i: Coefficient of x_i.

The model is first trained using training data and during training, the best line that fits the data values is accepted. The model gets the best regression line by varying the values of θ_0 and θ_i.

4.3.2 Neural Networks

Artificial neural networks is a machine learning technique in which machine is made to behave and think like a human brain. Like human brains, an artificial neural network consists of neurons which are spread out in different layers. Broadly, any ANN consists of an input layer through which data is fed to the network, an output layer at which the output in the form of prediction or classification is obtained and a hidden layer which is may or may not be the part of network. The data is fed on input layer where each input is given some weight that signifies the importance of that input parameter to the study. The weighted mean of all the inputs is passed to the next layer. Here comes the task of activation function. An activation function acts like a filter to remove unnecessary information from the previous layer and pass on only the necessary or required part to the next layer for further processing. It can be taken as a simple step function to switch on or off a neuron output. Mostly, nonlinear activation functions are used in neural networks so that they can deal with complex problems and data such as images, voice and data with high dimensionality. Also, nonlinear activation functions can deal with backpropagation which is important for the improvement of the network and is difficult to be dealt by linear activation function. Finally, the output is generated at the output layer. Number of neurons in each layer and number of hidden layers used in the network are decided on the basis of number of inputs and the type of problem need to be solved (Fig. 2).
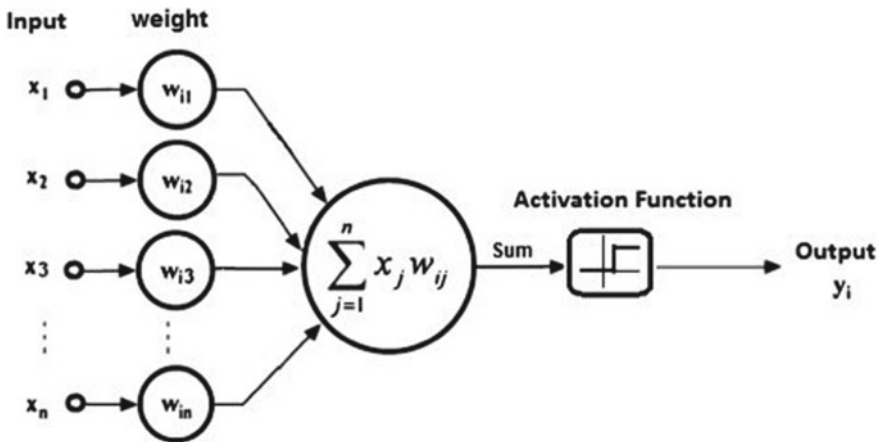

Fig. 2 Layout of a neural network [17]

5 Experimental Results

In the present study, ANN was employed on the data obtained from various sources of Punjab for Wheat crop yield prediction. For comparison purposes, another machine learning technique, multivariate linear regression was applied on the same data. The results obtained on applying both the models are discussed in Sects. 5.1 and 5.2.

5.1 Multivariate Linear Regression Model

In linear regression model, environmental parameters, were taken as independent variables, whereas yield obtained was the dependent variable. 43 climatic features were considered as the independent variables, whereas yield to be determined was taken as the dependent variable. The data was randomly selected by the model during training and testing, and the predicted and actual values obtained for the test data for various years are as shown in Table 1 and Fig. 3.

Table 1 Predicted and actual yield values for linear regression model

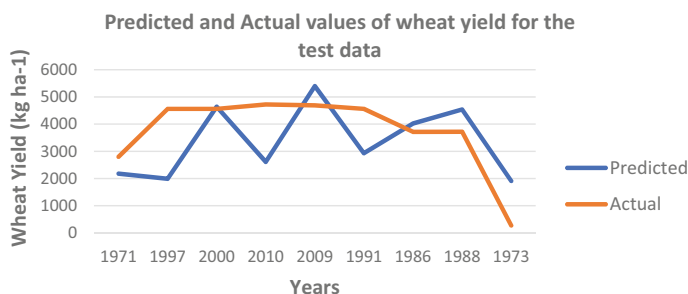| Years	Predicted	Actual
1971	2178.697	2800
1997	1987.084	4563
2000	4640.336	4563
2010	2610.989	4724
2009	5396.709	4689
1991	2931.266	4563
1986	4026.307	3715
1988	4537.306	3720
1973	1910.199	273

Fig. 3 Graphical variation of predicted and actual values for regression model

Table 2 Values of evaluation metrics for linear regression model

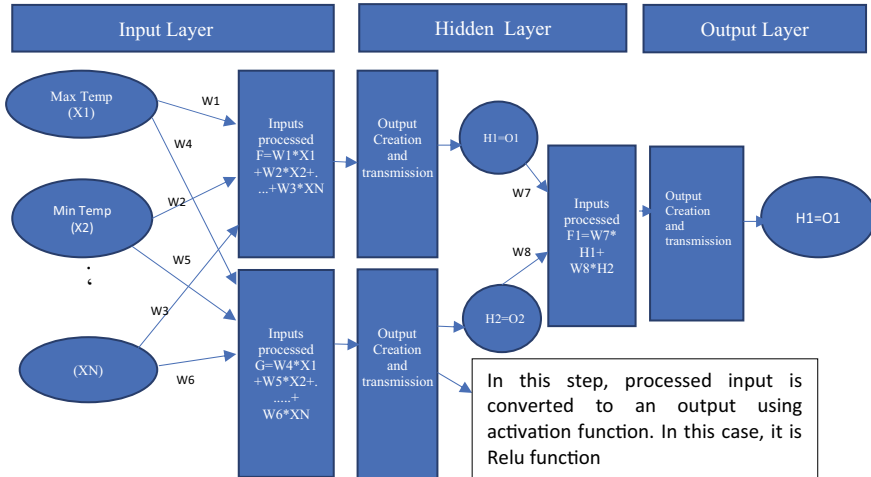| Technique	R-Square	Adjusted R-Square	RMSE	MAE
Linear regression	1.0	1.0	1337	1075

Fig. 4 Artificial neural network model used in study

The values of various evaluation metrics like R-square, adjusted R-square, RMSE and MAE are as shown in Table 2.

## 5.2   Artificial Neural Network

Artificial neural network model used in the study has been shown in Fig. 4. There are two ways of initializing a neural network model–defining each layer one by one or defining a graph. We used the sequential function of python library with no parameters to design the model layer by layer manually. The model was designed with three layers, one input layer, hidden layer and an output layer. Stochastic gradient descent algorithm was used for training the model and rectified activation function (Relu), and one of the most widely used activation function for nonlinear problems was used in all the layers. As the number of features was 43, so the input layer was fed with 43 neurons. The number of nodes in hidden layer was calculated as a mean of neurons in input and output layer and was taken as 19. Model was made to run for 2500 epochs in a batch size of 10.

The predicted and actual values of wheat yield as obtained from the ANN model are as shown in Table 3 and Fig. 5.

The values of various evaluation metrics obtained are as shown in Table 4.

Table 3 Predicted and actual values of yields for ANN model

Years	Predicted	Actual
2000	3506.285	4563
2006	5505.499	4690
1997	5041.7134	4563
1974	3569.511	2720
1980	2664.287	2730
1995	3315.5383	4600
1998	5961.345	4562
1981	4357.9854	3716
2007	4127.3887	4688
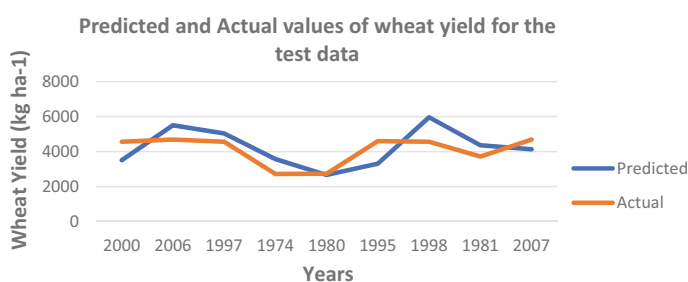
Fig. 5 Graphical variation of predicted and actual values for ANN model

Table 4 Values of evaluation metrics for ANN model

Technique	RMSE	MAE
ANN	819	744.3

Results obtained clearly indicate that ANN technique has shown much closer predictions as compared to multivariate linear regression technique. Also, the values of evaluation metrics have shown that the RMSE values obtained in case of ANN technique are quite less than those in linear regression.

6   Conclusions

The present study employed ANN technique for wheat crop yield prediction in an area of Punjab. The closeness of predicted values obtained in results to actual yield values have shown good prospects for ANN as a crop yield prediction model. For comparison purposes, on examining the values of evaluation metrics, RMSE and MAE, obtained in ANN and linear regression technique, it is clearly visible that ANN has shown much less error as compared to regression technique. This further

proves that neural networks can be a better choice when dealing with nonlinear behaviours which are inherent in the study. As this study pertains to the areas of Punjab, application of various machine learning techniques in this area still needs to be explored. For future scope, many other climatic parameters related to wind and soil are not included in study which can be further investigated in future studies. Also the advanced techniques of ANN in the form of deep learning can be explored in hybridization with other techniques of AI in the said area.

References

1. Kamir E, Waldner F, Hochman Z (2020) Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. ISPRS J Photogramm Remote Sens 160:124–135
2. Jain A, Chaudhary JL, Beck MK, Kumar L (2019) Developing regression model to forecast the rice yield at Raipur condition. J Pharmacogn Phytochem 8(1):72–76
3. Chlingaryan A, Sukkarieh S, Whelan B (2018) Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. Comput Electron Agric 151:61–69
4. Villanueva BM, Salenga MLM (2018) Bitter melon crop yield prediction using machine learning algorithm. Int J Adv Comput Sci Appl 9:1–6
5. Cheng H, Damerow L, Sun Y, Blanke M (2017) Early yield prediction using image analysis of apple fruit and tree canopy features with neural networks. J Imaging 3(1):6
6. Kim N, Lee YW (2016) Machine learning approaches to corn yield estimation using satellite images and climate data: a case of Iowa State. J Korean Soc Surv Geod Photogramm Cartogr 34(4):383–390
7. Bose P, Kasabov NK, Bruzzone L, Hartono RN (2016) Spiking neural networks for crop yield estimation based on spatiotemporal analysis of image time series. IEEE Trans Geosci Remote Sens 54(11):6563–6573
8. Pantazi XE, Moshou D, Alexandridis T, Whetton RL, Mouazen AM (2016) Wheat yield prediction using machine learning and advanced sensing techniques. Comput Electron Agric 121:57–65
9. Gandhi N, Petkar O, Armstrong LJ (2016) Rice crop yield prediction using artificial neural networks. In: 2016 IEEE technological innovations in ICT for agriculture and rural development (TIAR), July 2016. IEEE, pp 105–110
10. Shastry KA, Sanjay HA, Deshmukh A (2016) A parameter based customized artificial neural network model for crop yield prediction. J Artif Intell 9:23–32
11. González Sánchez A, Frausto Solís J, Ojeda Bustamante W (2014) Predictive ability of machine learning methods for massive crop yield prediction
12. Dahikar SS, Rode SV (2014) Agricultural crop yield prediction using artificial neural network approach. Int J Innov Res Electr, Electron, Instrum Control Eng 2(1):683–686
13. You J, Li X, Low M, Lobell D, Ermon S (2017) Deep Gaussian process for crop yield prediction based on remote sensing data. In: Thirty-first AAAI conference on artificial intelligence, Feb 2017
14. Qaddoum K, Hines EL, Iliescu DD (2013) Yield prediction for tomato greenhouse using EFuNN. ISRN Artificial Intelligence
15. Ghodsi R, Yani RM, Jalali R, Ruzbahman M (2012) Predicting wheat production in Iran using an artificial neural networks approach. Int J Acad Res Bus Soc Sci 2(2):34

16. Ludhiana District (2020, June 29). Retrieved from https://en.wikipedia.org/wiki/Ludhiana_district
17. Activation Function (2020, June 29). Retrieved from https://www.quora.com/What-is-meant-by-activation-function