

Regression and Neural Networks Models for Prediction of Crop Production

¹Raju Prasad Paswan, ²Shahin Ara Begum

Abstract- Neural networks have been gaining a great deal of importance and are used in the areas of prediction and classification; the areas where regression and other statistical models are traditionally being used. In this paper, a comprehensive review of literature comparing feedforward neural networks and traditional statistical methods viz. linear regression with respect to prediction of agricultural crop production has been carried out. This study presents a useful insight into the capabilities of neural networks and their statistical counterparts used in the area of prediction of crop yield.

Keywords: Linear regression, Neural networks, Crop Production, Prediction

1 INTRODUCTION

In agriculture, decision-making processes often require reliable crop response models. Agricultural management specialists need simple and accurate estimation techniques to predict crop yields in the planning process. Over the last few decades, statistical methods have traditionally been used for predictions and classifications. Some of the common traditional statistical techniques used for predictions and classifications are multiple regression, discriminant analysis, logistic regression etc. Most of the researchers have employed regression models for prediction purposes in various disciplines. Due to the nature of linear relationship in the parameters, regression models may not provide accurate predictions in some complex situations such as non linear data and extreme values data. As regression models need to fulfill the regression assumptions and multiple co-linearity between independent and dependent variables, it causes regression models to be inefficient. (Molazem *et al.* 2002[1], Zaefizadah *et al.* 2011[2]). In agricultural practices, crop production is influenced by a great variety of interrelated factors and it is difficult to describe their relationships by conventional methods. Thus, artificial neural network (ANN) is highly suggested to present the complicated relations and strong nonlinearity between different parameters and crop production. It is considered to be one of the best techniques for extracting information from imprecise and non-linear data (Caselli *et al.* 2009[3]). Hence, Neural networks (NNs) methods have become a very important tool for a wide variety of applications across many disciplines including prediction of crop production where traditional statistical techniques were used. This has led to a number of studies comparing the traditional statistical techniques with neural networks in a variety of applications. It has been recognized in the literature that regression and neural network methods have become competing model-building methods (Smith *et al.*, 1997[4]). Nowadays, NNs methods have been largely used in the areas of prediction and classification (Warner *et al.*, 1996[5]). NNs models are also preferred in the area of pattern recognition (Setyawati *et al.* 2002[6]). Many researchers have shown the relationship between neural networks and statistical models.

(Buntine and Weigend, 1991[7]); Ripley, 1992[8]); Sarle, 1994[9]); Werbos, 1991[10]) Cheng and Titterington (1994) [11]) showed a complete analysis and comparison of different network techniques with traditional statistical techniques. The strong association of the feedforward neural networks with discriminant analysis was also shown by the authors. Schumacher *et al.* (1996[12]) have shown a comparison between feedforward neural networks and logistic regression. The similarities and dissimilarities were also analyzed. Sarle (1994[9]) presented a neural network terminology into statistical terminology and showed the relationship between neural networks and statistical techniques. Warner *et al.* (1996[5]) compared the performances of regression analysis and neural networks using simulated data from known functions and also using real world data. The authors discussed the situations where it would be advantageous to use NNs models in place of parametric regression models. They also compared regression analysis with neural networks in terms of notation and implementation. Ripley (1994[8]) presented the statistical aspects of neural networks and classified neural networks as one of the flexible non linear regression methods. Thus, a good number of multidisciplinary studies including prediction of agricultural crop production have been carried out to compare the traditional statistical techniques with neural networks. The purpose of this paper is to present a review of articles that compare neural networks with standard statistical methods mainly regression techniques used for prediction of agricultural crop production. Many authors have attempted a comprehensive survey of articles involving neural networks in different field of applications but a very few works have been done on review of articles using neural networks for prediction of agricultural crop production and similar areas. Pande *et al.* (2008[13]) provided a comprehensive survey on crop yield estimation with emphasis on neural networks. Bouten *et al.* (2005[14]) presented a critical review of the used techniques on applications of Artificial Neural Networks (ANN) in Ecology. Maier *et al.* (2000[15]) provided a review on neural networks for prediction and forecasting of water

resources variables. The performance of a particular technique in comparison to other techniques depend on a number of factors like the volume of the data, selection of model or technique, the methods of validation of results, the measure used for comparison and whether significant difference exists in the results etc.. Therefore, in the present study, attempt has been made to critically assess the literatures in relation to the criteria stated above. The rest of the paper is structured as follows: In section 2, a brief introduction to the terminologies used in Neural Networks literature and statistics has been presented. Section 3 provides a comparison of neural networks diagram with statistical models. Section 4 presents the survey of the relevant papers in the area of crop yield prediction. Section 5 presents summary and findings in tabular form and finally section 6 concludes the paper alongwith a brief discussion of some issues relating to neural networks and statistical techniques.

2 TERMINOLOGIES IN NEURAL NETWORKS LITERATURE AND STATISTICS

Neural networks models are though similar to some extent with statistical models, the terminologies used in Neural Networks literature is totally different from that used in statistics (Sarle 1994[9]). These are shown in Table 1.

TABLE 1
 TERMINOLOGIES IN NEURAL NETWORKS LITERATURE AND STATISTICS

Neural Network	Statistics
Features	Variables
Inputs	Independent variables
Outputs	Predicted variables
Targets or training values	Dependent variables
Errors	Residuals
Training, learning, adaptation or self organization	Estimation
Error function, or cost function, or Lyapunov function	Estimation criteria
Patterns or training pairs	Observations
Weights (synaptic)	Parameter estimate
Higher order neurons	Interactions
Functional links	Transformations
Supervised learning or heteroassociation	Regression or discriminant analysis
Unsupervised learning, encoding or autoassociation	Data reduction
Competitive learning or adaptive vector quantization	Cluster analysis
Generalization	Interpolation

3 COMPARISONS OF NEURAL NETWORKS STRUCTURE WITH STATISTICAL MODELS

Here, different statistical models are compared with network structures (Sarle 1994[9]). Fig.1 shows neural networks and statistical terminology for a simple linear regression model. Neurons are represented by circles and boxes, while the connections between neurons are shown as arrows: Circles with the name inside indicate observed variables.

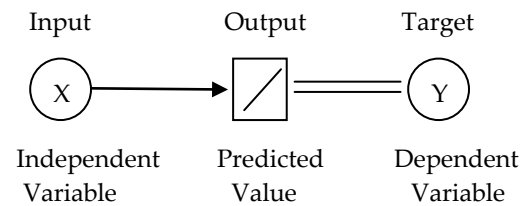


Fig. 1 Simple Linear Regression

Boxes represent values computed as a function of one or more arguments. The symbol inside the box denotes the type of function. Most boxes also have a corresponding parameter called a *bias*. Each arrow usually indicates a *weight* or parameter to be estimated. Two long parallel lines represent that the values at each end are to be fitted by least squares, maximum likelihood, or some other estimation criterion.

3.1 Perceptrons

A simple perceptron computes a linear combination of the inputs with a bias called the net input. Then, the output is produced by applying an activation function to the net input. Some common activation functions are: linear or identity function, hyperbolic tangent, logistic or sigmoidal, threshold and Gaussian function. Usually a perceptron may be of one or more outputs. Each output has a separate bias and set of weights. Generally the same activation function is used for each output, though it is possible to use different activation functions. Very often, perceptrons are generally trained by least squares, i.e. by minimizing $\sum \sum r_j^2$, where the summation is over all outputs and over the training sets. Thus, a perceptron with linear activation function is a linear regression model (Weisberg 1985[16]; Myers 1986[17]). A perceptron with a logistic activation function is a logistic regression model (Hosmer *et al.* 1989[18]) which is shown in Fig. 3

With a threshold activation function, a perceptron is known as a linear discriminant function (Hand 1981[19]; McLachlan 1992[20]; Weiss *et al.* 1991[21]). With only one output, a perceptron is also called an *adaline*, which is shown in Fig. 4. With multiple outputs, the threshold perceptron is a multiple discriminant functions.

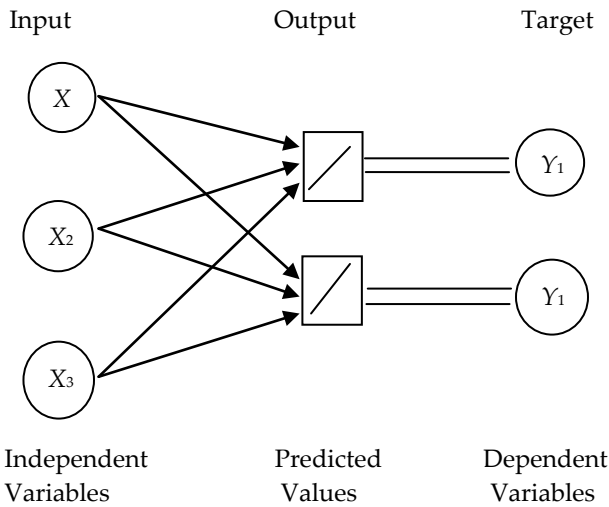


Fig. 2 Simple Linear Perceptron = Multivariate Multiple Linear Regression

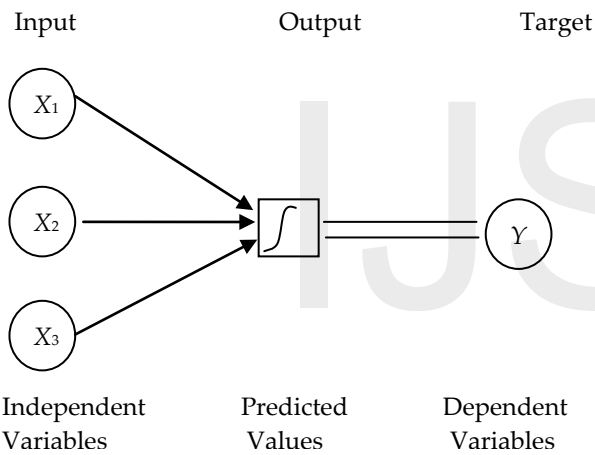


Fig. 3 Simple Nonlinear Perceptron = Logistic Regression

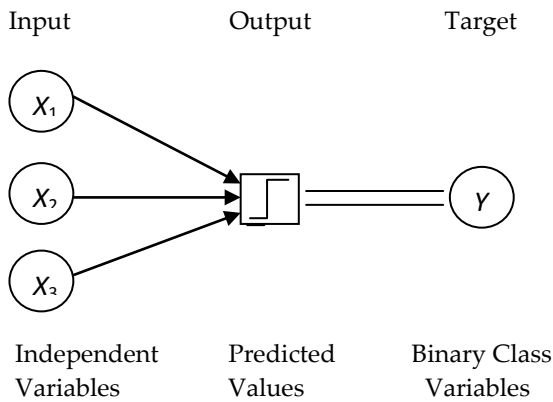


Fig. 4 Adaline = Linear Discriminant Function

3.2 Multilayer Perceptrons

A model becomes nonlinear, if it considers estimated weights between inputs and the hidden layer, and the hidden layer uses nonlinear activation function like logistic function. The resulting model is known as multilayer perceptron or MLP. An MLP for simple nonlinear regression is shown in Fig. 5.

An MLP may have multiple inputs and outputs (Fig. 6). The number of hidden neurons can be less than the number of inputs and outputs. It can also have direct connection from the input layer to the output layer. In statistical terminology, this is known as main effects.

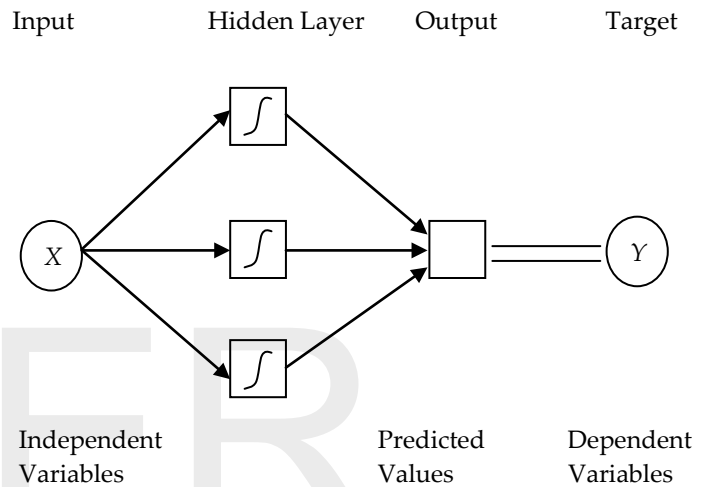


Fig. 5. Multilayer Perceptron = Simple Nonlinear Regression

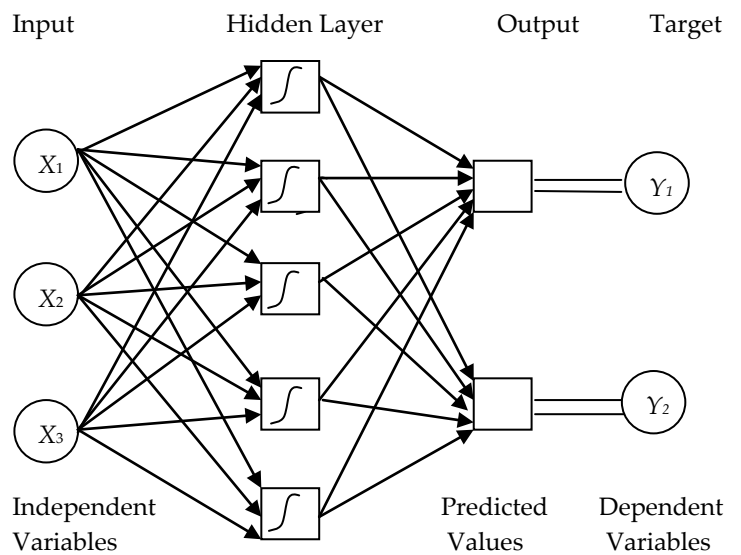


Fig. 6. Multilayer Perceptron = Multiple Nonlinear Regression

MLPs are universal approximator (White 1992[22]). When we know little about the relationship between the dependent and independent variables, in that case MLPs can be used.

The complexity of the MLP model can be varied by varying the number of hidden layers and the number of hidden neurons in each hidden layer. When there contains a small number of hidden neurons in an MLP, then the MLP is known as parametric model that acts as an alternative to polynomial regression. An MLP can be considered a quasi-parametric model similar to projection pursuit regression (Friedman *et al.* 1981[23]) when there contains a moderate number of hidden neurons. Generally, an MLP with one hidden layer is same as the projection pursuit regression model. The only difference is that except that an MLP uses a predetermined functional form for the activation function in the hidden layer, whereas projection pursuit uses a flexible nonlinear smoother.

3.3 Radial Basis Functions (RBF)

In a RBF network, the hidden neuron compute radial basis functions of the inputs, which are similar to kernel functions in kernel regression (Hardle 1990[24]). The net input to the hidden layer is the distance from the input vector to the weight vector. The weight vectors are also called centers. The distance is usually computed in the Euclidean metric, although it is sometimes a weighted Euclidean distance or an inner product metric. The activation function can be any of a variety of functions on the nonnegative real numbers with a maximum at zero, approaching zero at infinity. The outputs are computed as linear combinations of the hidden values with an identity activation function.

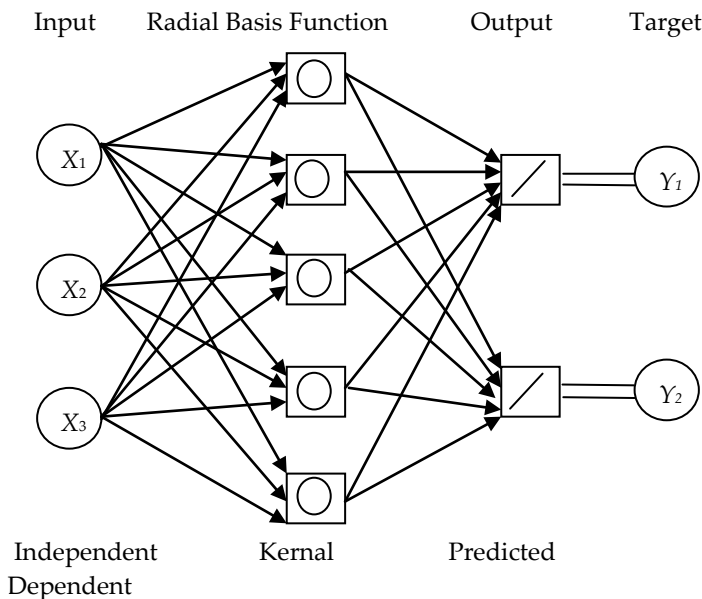







Fig. 7 Radial Basis Function Network

TABLE 2

MEANING OF THE SYMBOLS FOR NEURONS USED IN THE FIGURES FROM FIG.1 TO FIG.7

	=	Observed Variables
	=	Linear Combination of Inputs
	=	Logistic Function of Linear Combination of Inputs
	=	Threshold Function of Linear Combination of Inputs
	=	Radial Basis Function of Inputs

Often, values of the hidden layer are normalized to sum to 1 as is commonly done in kernel regression (Nadaraya *et al.* 1964[25]). Then if each observation is taken as an RBF center, and if the weights are taken to be the target values, the outputs are simply weighted averages of the target values, and the network is identical to the well-known Nadaraya-Watson kernel regression estimator. This method has been reinvented twice in the NN literature (Specht 1991[26]; Schioler *et al.* 1992[27]). Since an RBF network can be viewed as a nonlinear regression model, the weights can be estimated by any of the usual methods for nonlinear least squares or maximum likelihood. Generally, RBF networks are treated as hybrid networks. The inputs are clustered, and the RBF centers are set equal to the cluster means. The bandwidths are often set to the nearest-neighbor distance from the center. It works better to determine the bandwidths from the cluster variances. Once the centers and bandwidths are determined, estimating the weights from the hidden layer to the outputs reduces to linear least squares.

4 REVIEW OF LITERATURE

In recent times, the development and application of neural networks is not limited to a certain specific area. Rather, it has been widely used in most of the areas for predictions and classifications. Some of the areas include accounting and finance, health and medicine, engineering and manufacturing, marketing, agriculture, general applications etc. As the aim of this paper is to present an overview of the research papers involving comparison of neural networks with traditional statistical techniques namely regression techniques used for

prediction of agricultural crop production, therefore discussion will be restricted to this area i.e. prediction of crop production only.

4.1 Prediction of crop Production using Regression Model and ANN

Recently, Researchers have developed several forecasting and prediction models of various crop yields in relation to different parameters as influencing factors by applications of artificial neural networks and by combining ANN and statistical techniques such as linear regression technique. In this section, a number of related works dealing with the applications of neural network models, comparison with linear regression techniques and some combined models for the prediction and forecasting of crop yields has been reviewed. Since, the performance of a particular technique in comparison to other techniques, depends on a number of factors like the volume of the data, selection of model or technique, the methods of validation of results, the measure used for comparison and whether significant difference exists in the results etc., therefore, attempt has been made to carry out the review on these points.

The ANNs have largely impressed the agricultural researchers, as they are able to overcome the difficulties to many extents of traditional statistical approaches. In last few decades, researchers have examined ANN models from a statistical point of view (e.g. White, 1989[22a]; Cheng and Titterington, 1994[11]; Hill *et al.*, 1994[28]; Ripley, 1994[8]; Sarle, 1994[9]; Warner and Misra, 1996[5] and Maier *et al.* 2000[15]). Statistical models that can be expressed in neural network form are regression, discriminant, density estimation and graphical interaction models such as simple linear regression, projection pursuit regression, polynomial regression, non-parametric regression, logistic regression, linear discriminant functions, classification trees, finite mixture models, kernel regression and smoothing splines (Cheng and Titterington, 1994[11]; Sarle, 1994[9]).

Drummond *et al.* (1995[29]) compared several methods for predicting crop yield based on soil properties. They noted that the process of understanding yield variability is made extremely difficult by the number of factors that affect yield. They used several multiple linear regression methods, such as multiple linear regression, $r^2=0.42$; stepwise multiple linear regression, $r^2=0.43$; partial least squares regression, $r^2=0.43$; projection pursuit regression, $r^2=0.73$; and BP neural network, $r^2=0.67$ for modeling the relationship between corn yield or soybean yield and soil properties. They concluded that less-complex statistical methods, such as standard correlation, did not seem to be particularly useful in understanding yield variability.

Sudduth *et al.* (1996[30]) used neural network to predict soy bean yield based on soil parameters and achieved a testing error of 17.3%.

Tourenq *et al.* (1999[31]) used a classic multilayer feed-forward neural network with back-propagation algorithm throughout their experiments to assess the performances of artificial neural networks (ANNs) in predicting presence or absence of flamingo damages from 11 variables describing landscape features of rice paddy. They compared the performances to the results obtained by previous authors and found to be more accurate. They concluded that ANN can be an alternative tool for prediction of rice yield damage.

Liu *et al.* (2001[32]) used NN to predict maize yield based on rainfall, soil and other parameters and obtained a testing error of 14.8%.

Safa *et al.* (2002[33]) worked for prediction of wheat yield using ANN in Iran. The climatic observation data in different phenological stages of wheat crop were used in this study. Data were arranged in two matrix form. The study showed that the most important effective meteorological factors on crop yield, was quantity and quality of rainfall, but the most sensitivity stages relative to rainfall are flowering and heading. After them are primary stages of growth. So rainfall quantity after sowing and also the first two months of spring are very important to the crop production.

Boonprasom *et al.* (2002[34]) carried out a study by using MLFANN using BP learning algorithm of ANN on Prediction of Tangerine yield. In this study, weather parameters were considered as influencing factors and 9 years data relating to yield and weather parameters were collected. The study indicated that the ANN had high potential and ability to forecast tangerine yield accurately despite small set of data available. It was reported that the amount of rainfall had strong influence on yield of tangerine while average temperature had less influence.

Drummond *et al.* (2003 [35]), carried out another study on Site-Specific Yield Prediction using Statistical and Neural methods. They tried to understand the relationship between yield and soil properties and topographic characteristics. Stepwise multiple linear regression (SMLR), projection pursuit regression (PPR), and several types of supervised feed-forward neural networks were investigated in an attempt to identify methods able to relate soil properties and grain yields on a point-by-point basis within ten individual site-years. To avoid overfitting, evaluations were based on predictive ability using a 5-fold cross-validation technique. The neural techniques consistently outperformed both SMLR and PPR and provided minimal prediction errors in every site-year. A second phase of the experiment involved estimation of crop

yield across multiple site-years by including climatological data. The ten site-years of data were appended with climatological variables, and prediction errors were computed. The results showed that significant overfitting had occurred and indicated that a much larger number of climatologically unique site-years would be required in this type of analysis.

Paul *et al.* (2004[36]) used backpropagation neural network structure to develop model by combining regression and artificial neural network to severity of the gray leaf spot in maize caused by *Cercospora zea-maydis*. Model performance was evaluated based on r^2 and Mean Square Error for the validation of the complete data set. They reported that the best model had r^2 ranging from 0.70 to 0.75 and MSE ranging from 174.7 to 202.8. The model is found to be very useful for prediction of disease of Maize crop.

Jiang *et al.* (2004[37]) successfully applied ANN in developing model for crop yield forecasting using back-propagation algorithms at He Nan province of China. The model had adapted and calibrated using on ground survey and statistical data, and proved to be stable and highly accurate. The authors used sunlight supply, temperature, water stress and soil conditions and average yield as input parameters, they selected 5-8-1 neurons in the input layer, hidden layer and output layer respectively in the structure of the model. In the study, they divided the whole province into eight sub-regions and two to three sample counties were selected for yield data in each sub regions. Out of 30 counties selected, 20 were used for model training and 10 of them were used for validation data. Multi-regressing linear model (MR model). They found that ANN model performed better than MR model. The average relative error (absolute value) of the ANN model was 3.5% compared to the 11.5% error of MR model.

Kaul *et al.* (2004[38]) worked on feed-forward back-propagating ANN structure for yield prediction of corn and soya bean for typical climatic conditions. They evaluated ANN model performance and compared the effectiveness of multiple linear regression models to ANN models. It was reported that adjusting ANN parameters such as learning rate and number of hidden nodes affected the accuracy of crop yield predictions. Optimal learning rates fell between 0.77 and 0.90. Smaller data sets required fewer hidden nodes and lower learning rates in model optimization. ANN models consistently produced more accurate yield predictions than regression models. ANN corn yield models resulted in r^2 and RMSEs of 0.77 and 1036 versus 0.42 and 1356 for linear regression, respectively. ANN soybean yield models for Maryland resulted in r^2 and RMSEs of 0.81 and 214 versus 0.46 and 312 for linear regression, respectively. Although it is more time consuming to develop an ANN model as compared to linear regression models, ANN models proved to be a

superior methodology for accurate prediction of corn and soybean yields under typical Maryland climatic conditions.

Yong *et al.* (2005[39]) adopted back propagation neural network structure to carry out experiment on relationship analysis between wheat yield and soil nutrients by application of ANN. By training 50 tested soil samples in back-propagation neural network of topological structure 6:9:1, the model of analyzing the relation between the crop yield and those 6 soil characteristics was established to validate the remaining 13 samples. The results show that the soil water content and alkali-hydrolysable nitrogen are linear to the crop yield, the total nitrogen, organic matter and rapidly available potassium are respectively multinomial to it and that the rapidly available phosphorous is of the exponential relationship with the crop yield.

Ji *et al.* (2007[40]) investigated the performance of ANN model to see whether ANN models could effectively predict Fujian rice yield for typical climatic conditions of the mountainous region. They also compared the effectiveness of multiple linear regression models with ANN models. They reported that adjusting ANN parameters such as learning rate and number of hidden nodes affected the accuracy of rice yield predictions. Optimal learning rates were between 0.71 and 0.90. The study revealed that ANN models consistently produced more accurate yield predictions than regression models. ANN rice grain yield models for Fujian resulted in r^2 and RMSE of 0.67 and 891 vs. 0.52 and 1977 for linear regression, respectively. It is found that ANN models are more accurate than linear regression model for prediction of rice yields under typical Fujian climatic conditions.

Li *et al.* (2007[41]) carried out a study to develop a new methodology using an artificial neural network (ANN) to estimate and predict corn and soybean yields on a county-by-county basis, in the "corn belt" area in the Midwestern and Great Plains regions of the United States. The historical yield data and long time-series NDVI derived from AVHRR and MODIS were used to develop the models. A new procedure was developed to train the ANN model using the SCE-UA optimization algorithm. The performance of ANN models was compared with multivariate linear regression (MLR) models and validation was made on the model's stability and forecasting ability. The new algorithms effectively trained ANN models, and the prediction accuracy was as high as 85 percent. Three statistical parameters were used for performance analysis: correlation coefficient (r), root mean square error (RMSE), and average difference (AVDIF).

Sing *et al.* (2008[42]) worked on maize crop forecasting using multilayered feedforward network (MFN) of ANN. They considered maize crop yield data as response variable and total human labour, farm power, fertilizer consumption, and

pesticide consumption as predictors and found that a three-layered feed-forward network with (11,16) units in the two hidden layers performs best in terms of having minimum mean square errors (MSE) for training, validation, and test sets. Superiority of this MFN over multiple linear regression (MLR) analysis had also been demonstrated for the maize data considered in the study. It was concluded that the ANN is the most efficient tool for successfully tackling the realistic situation in which exact nonlinear functional relationship between response variable and a set of predictors is not known.

Khazaei *et al.* (2008[43]) used the backpropagation ANN method in regression for modeling between crop yield components of chickpea. Recently, Wen *et al.* (2010[44]) combined forecasting model by using multi-indicator for grain yield in China is based on BP network and grey system, which was named as GM(1,1)-BP model. Empirical results showed that the combined model had higher precision and training efficiency than the models based on GM(1,1), BP network or GM(1,N) alone. The results revealed that the grain yield can be accurately predicted by this model through small scale of requirement on samples and information. It is concluded that the GM (1, 1) - BP model is effective with the advantages of high precision, less requirement of samples and simple calculation.

Miaoguang *et al.* (2008[45]) used Generalized Regression Neural Networks (GRNN) for forecasting of agricultural crop production. They found GRNN to be a good technique for prediction grain production in rural areas. It was reported that GRNN model is suitable for non-linear, multi-objectives and multivariate forecasting.

Saad *et al.* (2009[46]) for rice yield prediction used ANN in precision farming and evaluated two models, Back Propagation Network and RBF network. The study showed that Radial Basis Network performed better than Back Propagation Network in terms of training time, accuracy and number of nodes in the hidden layer. It was also seen that training the MLP network was often too slow especially in the case of large size problems. Since RBF network can establish its parameters for hidden neurons directly from the input data and train the network parameters, it is generally much faster compared to MLP network.

Heninzow, (2009[47]) in his work, used four-layer back propagation network with two hidden layers and trained the network by resilient propagation algorithm to show the ability of artificial neural network technology for prediction of crop yields in different climatic zones based on reported daily weather data. The final neural networks was trained with data sets of three climate zones and tested against an independent northern zone which had high predictive power. The ANN

technology was found to be useful tool to investigate, approximate and predict spring crop yields in a heterogeneous climate region with wide ranges of temperature.

Mehnatkesh *et al.* [48] conducted a study to evaluate the efficacy of artificial neural network and multiple linear regression tools to predict biomass and grain yield of winter wheat (cv. Sadri). Another objective of the study was to identify the most important edaphic factors (soil, precipitation, topographic, and management factors) that influence yield production in the hilly regions of central Zagros, west of Iran. A total of 404 sampling points were chosen on the landscape covering summit, shoulder, backslope, footslope, and toeslope at two sites with varying climatic conditions. Surface (0-30 cm) soil samples and data on wheat yield were collected at two sites in Koohrang and Ardal districts. Four parameter groups including terrain attributes, soil physical and chemical properties, precipitation, and weed biomass, including 57 factors were used as the inputs, and wheat grain and total biomass yield as the targets for ANN and MLR models. Predictor ANN and MLR models resulted in R² values of 0.84 and 0.53 for grain yield, respectively; and 0.69 and 0.26 for total biomass, respectively. These models resulted in RMSE values of 0.033 and 0.055 for grain yield, and 0.038 and 0.070 for total biomass, respectively.

Mwasiagi *et al.* (2010[49]) designed an ANN model by selecting cotton-growing cost factors to predict cotton yield in Kenya. They found that this neural network model was able to predict cotton yield with a satisfactory performance error of 0.204 kg/ha and a regression correlation coefficient between network output and actual yield of 0.945.

Obe *et al.* (2010[50]) studied on forecasting of sugarcane production by application ANN based model. The performances of ANN models were measured using Mean Square Error (MSE), Normalized Mean Square Error (NMSE), correlation coefficient and Minimum Description Length (MDL). They found the result of prediction of the ANN model to be 85.70%.

Ayoubi *et al.* (2011[51]) In their study, designed artificial neural network (ANN) models to predict the biomass and grain yield of barley from soil properties; and they compared the performance of ANN models with earlier tested statistical models based on multivariate regression. Barley yield data and surface soil samples (0-30 cm depth) were collected from 1 m² plots at 112 selected points in the arid region of northern Iran. ANN yield models gave higher coefficient of determination (R²) and lower root mean square error compared to the multivariate regression, indicating that ANN is a more powerful tool than multivariate regression. Overall results indicated that the ANN models could explain 93 and

89% of the total variability in barley biomass and grain yield, respectively. The performance of the ANN models as compared to multivariate regression has better chance for predicting yield.

Laxmi *et al.* (2011[52]) worked on Neural Networks for crop yields forecasting using MLP with different learning algorithms at Uttar Pradesh. They considered crop productivity, maximum and minimum temperature, relative humidity morning and rainfall as input variables. They used stepwise regression techniques significant variables for selecting significant variables. They concluded that ANN models produced better results than statistical model. MAPE was used for performance evaluation of models.

Thongboonak *et al.* (2011[53]) carried out a study to develop the Artificial Neural Network (ANN) modules for agricultural yield prediction. The ANN modules developed were tested with longan yield prediction in Chiang Mai and Lamphun provinces. The ANN input data were soil group and climate data for the years 2006 – 2008, which related to longan yield in 2007 and 2008. All data were normalized in the same range of 0-1 to be suitable as the input of the ANN model. The normalized weekly highest, lowest, and average temperature, average sunlight, and rainfall were interpolated. They were then averaged to spatially represent districts in the study area, which corresponded to the longan yield districts. These data were varied with several input variations. The cross validation process was applied to each variation. The optimal parameters including learning rate, number of nodes in the hidden layer, and number of iterations obtained from testing were 0.4, 6, and 3,000 respectively. These parameters were applied for all training and testing processes. The best accuracy achieved is 99%. The ANN modules developed for the ArcMap environment worked well for longan yield prediction with accurate results despite the limitations of the data set.

Zaefizadeh, M. *et al* (2011[2]) in their work, compared Multiple Linear Regressions (MLR) and Artificial Neural Network (ANN) in Predicting the Yield Using its Components in the Hulless Barley in Iran. In this study 40 genotypes in a randomized complete block design with three replications for two years were planted in the region of Ardabil. The yield

related data and its components over the years of the analysis of variance were combined. Results showed that there was a significant difference between genotypes and genotype interaction in the environment. MLR and ANN methods were used to predict yield in barley. Also, yield prediction based on multi-layer neural network (ANN) using the Matlab Perceptron type software with one hidden layer including 15 neurons and using algorithm after error propagation learning method and hyperbolic tangent function was implemented. In both the methods, absolute values of relative error as a deviation index in order to estimate and using t test of mean deviation index of the two estimates was examined. Results showed that in the ANN technique the mean deviation index of estimation significantly was one-third (1/3) of its rate in the MLR. They recommended neural network approach due to high yield and more velocity in the estimation to be used instead of regression approach.

5 SUMMARY AND FINDINGS

A review in the application area of crop yield prediction focusing on the comparison of multilayered feedforward network to the traditional statistical techniques namely regression analysis, logistic regression and discriminant analysis have been presented in this paper. In all, 24 papers have been reviewed out of which most of the papers have used multilayered feedforward neural network with one of the statistical techniques as stated above to view if any particular method outperforms the other which has been presented in Table 3. Care has been taken to tabulate different criteria such as number of variables used in the study, the validation technique and measure used for comparing performance of various techniques.

Table 3 gives a summary of some of the study of comparison of ANNs and Statistical models discussed in this survey for prediction of agricultural crop yield production. The table consists of six columns. Column 1 provides references, Column 2 represents statistical model to which NN model is benchmarked, column 3 represents the number of variables used in the study, column 4 gives the validation method used, and column 5 gives the error measure used for comparison purposes and column

TABLE 3
 APPLICATIONS CROP YIELD PREDICTION

Reference	Statistical model	No. of Variables	Validation Method	Error Measure	Findings
Warner and Misra (1996)	LR	12	Tr-Ts (70-30)	C-index and Goodness of fit test	[C]
Drummond <i>et al.</i> (2003)	SMLR, PPR	11	Tr-Ts CV	CVMSE	[A]
Paul <i>et al.</i> (2004)	LRM	11	Tr-Ts CV	R ² , MSE	[A]
Jiang <i>et al.</i> (2004)	MLR	5	Tr-Ts CV	AV	[A]
Kaul <i>et al.</i> (2004)	MLR	20	Tr-Ts	R ² , RMSE	[A]
Yong <i>et al.</i> (2005)	MLR	6	Tr-Ts-CV	Goodness of fit test, CC and R ²	[A]
Ji <i>et al.</i> (2007)	MLR	60	Tr-V	R ² , RMSE	[A]
Li <i>et al.</i> (2007)	MLR	15	Tr-CV	R ² , RMSE, AVDIF	[A]
Sing <i>et al.</i> (2008)	MLR	5	Tr-Ts-V	MSE	[A]
Mehnatkesh <i>et al.</i>	MLR	57	Tr-Ts	R ² and RMSE	[A]
Ayoubi <i>et al.</i> (2011)	PCA	14	Tr-Ts (80-20)	R ² and RMSE	[A]
Zaefizadeh, M. <i>et al.</i> (2011)	MLR	5	Tr-Ts	t-test of mean deviation index	[A]
Laxmi <i>et al.</i> (2011)	SMLP	5	Tr-Ts-V	MAPE	[A]

AVDIF : Average Difference, SMLR : Stepwise multiple linear regression, PPR : Projection pursuit Regression, CVMSE : Cross-validated mean squared error, LRM: Logistic regression model, AV: Absolute Value, PCA : Principal Component Analysis, Tr-Ts : Training Testing, Tr-Ts-V : Training Testing Validation, Tr-Ts-CV : Training Testing Cross-Validation, MAPE: Mean Absolute Percentage Error.

6 gives the findings of the corresponding research paper which can be classified into 3 categories namely, neural networks outperforming statistical techniques, neural networks and statistical techniques being comparable and statistical techniques outperforming neural networks. These three classifications have been denoted as [A], [B] and [C] respectively in the 6th column of the tables in order to enhance the readability of the paper.

CONCLUSIONS

A literature review of comparative studies on artificial neural networks and traditional statistical techniques used for prediction of agricultural crop production has been carried out in the study. It is clear from the literature that ANN can automatically approximate any nonlinear mathematical function. This aspect of neural networks is particularly useful when the relationship between the variables is not known or is complex and hence it is difficult to handle statistically. However, the determination of various parameters like the number of hidden layers, number of nodes in the hidden layer etc. associated with neural networks is not straightforward and finding the optimal configuration of neural networks is a very time consuming process. In this respect, statistical model

clearly stands out as it allows interpretation of coefficients of the individual variables and due to the parametric assumptions of these models, inferences can also be drawn regarding the significance of certain variables in prediction or classification problems.

Neural networks and statistical models are not competing methodologies for data analysis. There is an overlap between the two fields. Neural networks include several models, such as MLPs that are useful for statistical applications. Statistical methodology is directly applicable to neural networks in a variety of ways, including estimation criteria, optimization algorithms, confidence intervals and graphical methods. Better communication between the fields of statistics and neural networks would benefit both.

REFERENCES

- [1] D. Molazem, M. Valizadeh and, M. Zaefizadeh, "North West of genetic diversity of wheat," *J. Agricultural Sciences.*, 20; 353-43, 1, 2002.
- [2] M. Zaefizadeh, M. Khayatnezhad and R. Gholamin, "Comparison of Multiple Linear regressions and Artificial Neural Network in Predicting the Yield Using its Components in the Hassle Barley,"

- American-Eurasian J. Agric. & Environ. Sci.*, 10 (1): 60-64, 2011, ISSN 1818-6769.
- [3] M. Caselli, L. Trizio, G. de Gennaro and P. Ielpo, "A simple feedforward neural network for the PM₁₀ forecasting: comparison with a radial basis function network and a multivariate linear regression model," *Water Air Soil Pollution*. 201, 365-377, 2009.
- [4] A. E. Smith and K. Mason, "Cost estimation predictive modeling: Regression versus neural network," *The Engineering Economist*, 42(2), 137-161, 1997.
- [5] B. Warner and M. Misra, "Understanding neural networks as statistical tools," *The American Statistician*, 50(4), 284-293, 1996.
- [6] B. R. Setyawati, S. Sahirman and R. C. Creese, "Neural networks for cost estimation" *AACE International Transactions EST13*, 13.1-13.8, 2002.
- [7] W. L. Buntine and A. S. Weigend, "Bayesian Back-propagation," *Complex Systems*, 5, pp 603-643, 1991.
- [8] B.D. Ripley, "Neural networks and related methods of classification," *Journal of the Royal Statistical Society*, 56 (3), 409-456, 1994.
- [9] W.S. Sarle, "Neural networks and statistical models," In: *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, pp 1538-1550, SAS Institute, 1994.
- [10] P. J. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," *Ph.D. Thesis*, Harvard University, Cambridge, MA, 1974.
- [11] B. Cheng and D.M. Titterington, "Neural Networks: A Review from Statistical Perspective," *Statistical Science*, Vol. 9, No.1, 2-54, 1994.
- [12] M. Schumacher, R. Robner and W. Vach, "Neural networks and logistic regression: Part I," *Computational statistics and data analysis*, Vol. 21, pp 661-682, 1999.
- [13] J. Pande and B. Varma, "Survey of Crop Yield Estimation Models with Emphasis on Artificial Neural Network Model," *Proceedings of the 2nd National Conference ; INDIA com-2008*.
- [14] I.W. Bouten and F. Arbelaez, "Working paper on Applications of Artificial Neural Networks in Ecology – A critical review of the used techniques, 2005.
- [15] H. R. Maier and G. C. Dandy, "Neural Networks for prediction and forecasting of water resources variables: a review of modeling and applications," *Environmental Modelling & Software*, 15, 101-124, 2000.
- [16] S. Weisberg, *Applied Linear Regression*, New York: John Wiley & Sons, 1985.
- [17] R.H. Myers, *Classical and Modern Regression with Applications*, Boston: Duxbury Press, 1986.
- [18] D.W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, New York: John Wiley & Sons, 1989.
- [19] D. J. Hand, *Discrimination and Classification*, New York: John Wiley & Sons, 1981.
- [20] G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, New York: John Wiley & Sons, 1992.
- [21] Weiss, S.M. and C.A. Kulikowski, *Computer Systems That Learn*, Morgan Kaufmann Publishers, San Mateo, California, 1991.
- [22] H. White, *Artificial Neural Networks: Approximation and Learning Theory*, Oxford, UK: Blackwell, 1992.
- [22a] H. White, "Learning artificial neural networks: A statistical perspective," *Neural Computation* 1, 425-464, 1989.
- [23] J.H. Friedman and W. Stuetzle, "Projection pursuit regression," *Journal of the American Statistical Association*, 76, 817-823, 1981.
- [24] W. Hardle, *Applied Nonparametric Regression*, Cambridge, UK: Cambridge University Press, 1990.
- [25] E.A. Nadaraya, "On estimating regression: Theory of Probability and its Applications," 9 (1), 55-59, 1964.
- [26] D. F. Specht, "A Generalized Regression Neural Network," *IEEE Transactions on Neural Networks*, 2, 568-576, 1991.
- [27] H. Schioler and U. Hartmann, "Mapping Neural Network Derived from the Parzen Window Estimator," *Neural Networks*, 5, 903-909, 1992.
- [28] T. Hill, L. Marquez, M. O'Connor and W. Remus, "Artificial neural network models for forecasting and decision making," *International Journal of Forecasting*, 10, 5-15, 1994.
- [29] S. T. Drummond, K. A. Sudduth and S. J. Birrell, "Analysis and correlation methods for spatial data," *ASAE Paper No 95-1335*, St. Joseph, Mich.: ASAE, 1995.
- [30] K. Sudduth, C. Fraisse, S. Drummond and N. Kitchen, "Analysis of Spatial Factors Influencing Crop Yield," in *Proc of Int Conf on Precision Agriculture*, 129-140, 1996.
- [31] C. Tourenq, A. Stephane, D. Laurent and S. Lek, "Use of artificial neural networks for predicting rice crop damage by greater flamingos in the Camargue, France," *Ecological Modelling*. 120, 349-358, 1999.
- [32] J. Liu, C. E. Goering and L. Tian, "A Neural Network for Setting Target Corn Yields," *transaction of the ASAE*, Vol 44(3): 705-713, 2001.
- [33] B. Safa, A. Khalili, A. M. Teshnehlab and A. M. Liaghat, "Prediction of Wheat Yield using Artificial Neural Networks" *25th Confon Agricultural and Forest Meteorology*, 2002.
- [34] P. Boonprasom and G. Bumroongitt, "Prediction of Tangerine Yield Using Artificial Neural Network," 2002, Chiang Mai University, Chiang Mai 50200.
- [35] S. T. Drummond, K. A. Sudduth, A. Joshi and S. J. Birrell, "Statistical and Neural Methods for Site-Specific Yield Prediction," *American Society of Agricultural Engineers*, ISSN 0001-2351, Vol. 46(1), 2003.
- [36] P. A. Paul, and Munkvold, "Regression and Artificial Neural Network Modelling for the Prediction of Gray Leaf Spot of Maize," *International Journal of Ecology and Environment*, Accepted for publication 14th December, 2004.
- [37] D. Jiang, X. Yang, N. Clinton and N. Wang, "An Artificial Neural Network Model for Estimating Crop Yield using Remotely Sensed Information," *INT. J. REMOTE SENSING*, VOL. 25, 1723-1732, 2004.
- [38] M. Kaul, R. L. Hill and C. Walthall, "Artificial neural networks for corn and soybean yield prediction," *Agricultural Systems* 85, 1-18, 2005.
- [39] H. Yong, Y. Zhang, S. Zhang and H. Fang, "Application of Artificial Neural Network on Relationship Analysis between Wheat Yield and Soil Nutrients," *Proceeding of IEEE*, 0-7803-8740-6, 2005.
- [40] B. Ji, Y. Sun, S. Yang and J. Wan, "Artificial neural networks for rice yield prediction in mountainous regions," *Journal of Agricultural Science*, 145, 249-261, 2007.
- [41] A. Li, S. Liang, A. Wang and J. Qin, "Estimating Crop Yield from Multi-temporal Satellite Data Using Multivariate Regression and Neural Network Techniques," *Photogrammetric Engineering & Remote Sensin*, Vol. 73, No. 10, pp. 1149-1157. 2007.
- [42] R. K. Singh and Prajneshu, "Artificial Neural Network Methodology for Modelling and Forecasting Maize Crop Yield," *Agricultural Economics Research Review*, Vol 21 January-June 2008, 5-10.
- [43] J. Khazaei, M. R. Naghavi, M. R., Jahansouz and G. Salimi, "Yield estimation and clustering of chickpea genotypes using soft computing techniques," *Agron J* 2008, 100, 1077-1087.
- [44] J. Wen and L. Lei, "A Combined Forecasting Method of Grain Yield in China based on GM (1,1) and BP network," *Third International Conference on Information and Computing*, 2010.

- [45] J. Miaoguang and J. Chaochong "Forecasting Agricultural Production via Generalized Regression Neural Network," 978-1-4244-2972-1/08/ IEEE, 2008.
- [46] P. Saad and N. Ismail, "Artificial Neural Network Modelling of Rice Yield Prediction in Precision Farming," Artificial Intelligence and Software Engineering Research Lab, School of Computer & Communication Engineering, Northern University College of Engineering (KUKUM), Jejawi, Perlis, 2009.
- [47] T. Heinzow, "Prediction of Crop Yields Across Four Climate Zones In Germany: An Artificial Neural Network Approach," Working Paper FNU-34, Germany, 2009.
- [48] A. Mehnatkesh, S. Ayoubi, A. Jalalian and A. A. Dehghani, "Prediction of rainfed wheat Grain yield and biomass using Artificial Neural Networks and Multiple Linear Regressions and determination the most factors by sensitivity analysis," 2010.
- [49] Mwasiagi, J. Igadwa, Huang, B. Xiu and X. H. Wang, "Prediction of cotton yield in Kenya," *Academy of Science of South Africa*, 2010.
- [50] O.O. Obe and D. K. Shangodoyin, "Artificial Neural Network Based Model for Forecasting Sugarcane Production," *Journal of Computer Science* (4), 439-445, 2010.
- [51] S. Ayoubi and K. L. Sahrawat, "Comparing multivariate regression and artificial neural network to predict barley production from soil characteristics in northern Iran," *Archives of Agronomy and Soil Science*, Vol. 57, No. 5, 549-565, 2011.
- [52] R.R. Laxmi and A. Kumar, "Weather based forecasting for crops yield using neural network approach," *Statistics and Application*, Vol. 9, Nos. 1&2, 2011, pp 55-59.
- [53] K. Thongboonnak and S. Sarapirome, Integration of Artificial Neural Network and Geographic Information System for Agricultural Yield Prediction, *Suranaree J. Sci. Technol.* 18(1):71-80, 2011.

Author 1:

Raju Prasad Paswan
Deptt. of Computer Science,
Assam University, Silchar, Assam, India
E-mail: raju.pas66@gmail.com

Author 2:

Shahin Ara Begum
Associate Professor, Deptt. of Computerr Science,
Assam University, Silchar, Assam, India
E-mail: shahin.begum.ara@gmail.com

IJSER