Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

A Predictive Model to Predict Seed Classes using Machine Learning

Tekalign Tujo G[1]., Dileep Kumar G.[2], Elifenesh Yitagesu D.[3], Meseret Girma B.[4]
[1,3,4](Lecturer, Madda Walabu University, Bale Robe, Ethiopia.
[2](Department of Computing, Adama Science and Technology University (ASTU), Adama, Ethiopia

Abstract: **- In Ethiopian history, agriculture has been the backbone of the economy. This agricultural activity remain undeveloped due to different factors. Most of the activities are done with a lack of modern technology. Currently, seed classification is done based on knowledge of human being. The current seed classification analysis is inefficient and has no validation mechanism. In this research, we have made an effort to present a predictive model to predict seed classes using machine learning algorithms which results in high crop production. For the development, this research machine learning algorithm is used to learn from data which can be used to make predictions, to make real-world simulations, for pattern recognitions and classifications of the input data. An artificial neural network is used for modelling complex relationships between inputs and outputs or to find patterns in data. The objective of this thesis is to understand the machine learning algorithm using neural networks and constructing model which predicts seed classes based on machine learning technique. The developed model is experimented using seed dataset and then seed classes are predicted using the developed model. Finally by using the developed model, determinant factors for classify seeds are identified and ranked.**

Keyword: Artificial Neural Network, Seed Classification, Machine Learning, Predictive Analysis

I. INTRODUCTION

The usage of technology in agriculture works has started since the early 20th century when the industry moved from the horse-drawn digger to mechanized tractors. The introduction of plant heredities, chemical inputs and crop management systems has transformed the industry into technology enabled and data-rich world [1].

The technological progress that make up the current computing environment have contributed to discussing about big data whereas data collecting is not new concepts especially in the context of public data collection. The only start of more efficient mobile technologies and the digitization of data have allowed large records be evaluated and analysed in a timely and more useful ways.

Agriculture is the most important economic sector of many developing countries. Ethiopian agricultural activities have continued underdeveloped ways because of lack of having sufficient technologies. The other cause of unproductive is drought, which has frequently affected the counter's agricultural activities since the early 1970s. This problem leads low productivity, weak infrastructure, low level of technology and overpopulation. For example, according to the World Bank 1980 and 1987 report, agricultural production dropped at an annual rate of 2.1 percent, while the population raised at an annual rate of 2.4 percent.

Therefore the country encountered a famine that resulted in the death of nearly 1 million people from 1984 to 1986 [2]. The Ethiopian farming community is facing different problems to maximize crop productivity. However, there is wide range of information gap are exists between research and existing practice. Due to this multitude problems, Ethiopian farmers need expert advice to have more productive.

These research is aiming to find solutions to problems in Melkassa research centre during seed classification and proposed to address research challenges in Agriculture sector. In order to take full advantage of the soil type, moisture, humidity, climate and etc. farmers need to know exactly the type of seeds for their cropping. Different districts in Ethiopia have varying climates and so it is very important to consider environmental factors of these separate areas. This helps to choose the best districts for farming of different type of seeds. Rainfall also varies from district to district and this has a huge impact on farming because while too little or too much rain can kill crops, the proper amount of rain leads to perfect crop yield. In today's conditions, agricultural enterprises are capable of generating large amounts of data. So this Growth in data size requires an automated method to extract and analysis necessary data. Machine learning algorithms, ANN are used to support agricultural centre experts. ANN holds one of the keys for farmers control centres to collect and process data in real time to help farmers that makes the best decisions with regard to planting, fertilizing and harvesting crops. In today's conditions, agricultural enterprises are capable of generating and collecting large amounts of data. So this Growth in data size requires an automated method to extract and analysis necessary data.

We proposed an automatic seed class predictor model which classifies seed dataset using ANN machine learning tool.

[3] Developed agricultural management for simple and precise estimation techniques to predict rice yields in the planning process. An ANN [4], is a form of artificial intelligence which is composed of a large number of simple processing components called artificial neurons or nodes that are interconnected by direct links, called connections, and which cooperate to perform parallel distributed processing (PDP) operation in order to solve a given problem. A subgroup of processing component[5] is called a layer in the network. The lowest layer is the input layer and the highest layer is the output layer. Between the lowest and highest layer, there may be an additional layer(s) of units, called hidden layer(s). The advantage of neural networks over conventional programming lies [5] in their capability to become

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

a solution for different problems that do not have an algorithmic solution or the available solution is too complex to be found. An ANN [6] is adjusted for a specific application, such as pattern recognition or data classification, through a training process.
The ANN modelling is becoming very popular in different areas of agriculture, specially, in the areas where straight statistical modelling becomes unsuccessful.
The ANN is using in the field of agriculture to predict the crop yield, biomass production, seeding dates, physical and physiological damaging of seeds, organic matter contents in the soils, soil moisture estimation, aerodynamic properties of crops, estimation of sugar content in fruits and characterization of crop varieties [7].
In our research, we have considered the effects of geometric parameters towards seed classification in Ethiopia. Taking these factors into consideration as datasets for various districts, then we applied suitable model with well- trained multilayer neural network classifiers for shapes, sizes and varietal type identification of irregular wheat grain samples grown in the various agro environmental zones in the country.

II. RELATED WORK

In this section, related works related to the study are reviewed and discussed. One of our objectives is to use the variety of data available in the agriculture domain to develop predictive model to predict seed classes using machine learning tools. In order to analyze this problem better, we focus our literature review on six aspects which is done by different researchers: 1) Autonomous Wheat Seed Type Classifier System 2) Classification of Rice Grains Using Neural Networks 3) Agricultural data prediction by means of neural network 4) Agricultural Crop Yield Prediction Using Artificial Neural Network Approach 5) Seed Classification using Machine Learning Techniques and finally 6) A Prediction Model Based on Big Data Analysis Using Hybrid FCM Clustering.
The first is [8], According to this article the researchers trying to use K-means clustering algorithm and the default Euclidean distance metric to cluster seed dataset. For this clustering, the researcher uses MATLAB as a programming environment. K-means function is used from statistics toolbox which is given two arguments. Those two arguments are the dataset and the number of the cluster the data going to be classified. Function k means can solve this problem by getting another argument called replicates; it is an integer number specifies how many times algorithm should be run with a new starting point. In this study, the authors propose the system which is capable of clustering approximately seeds and the profiting K-means algorithm leads to the operation.
The second is [9], this paper states Neural Networks to classify varieties of rice which contain a total of 9 different rice verities. To classify these varieties the authors uses image acquisition of seeds. They also developed to extract thirteen morphological features, six color features and fifteen texture features from color images of individual seed samples. Different neural network models were developed for individual feature sets and for the combined feature set.
Results of the paper is just designing and developing neural network models with two hidden layers in all networks using Matlab toolbox. Originally individual neural network models were created for each feature set (colour, morphology, and texture) separately. Then a combination of feature set model was implemented. In order to reduce the dimension of the input feature set, they applied principal component analysis. Finally, they combined feature model produced with an overall classification accuracy of 92%. The gap we found during the review of these paper is that they only focus on structured data with small data. The another is to identify the color or types of rice they used camera (Sony DSC-W270 digital camera) So it may have lack of quality when we compare it with X-ray technology which is new technology and best for classifying the color of seeds
The third related work is[10], according to this paper artificial intelligence approach and differentiability of the error function are used. The researcher focuses on studying the multi-layer neural network regressive model which has been used for solving the problem of the yield of onion and they recommended empirical non-linear regressive models to decide the relationship between the yield of the crop and the sowing density or the plantation density. The paper also presents a model with a multi-layer neural perceptron in the configuration (1-2-1), i.e. one neuron at the input, two in the hidden layer and one at the output, along with the non-linear activation function. For the learning itself, they used the Back Propagation algorithm with the implementation of the multi-layer neural network for the prediction of the crop yield, and the comparison of the accuracy of this approach with the accuracy of the well-known regression model designed for the prediction of empirical data. Empirical non-linear regressive models are used for determining the usefulness of a neural networks prediction approach, and the options of its implementation.
In this paper there three measures are used to predict agricultural data. These measures are: Non-linear regression, Multi-layer neural network and Back propagation algorithm
After comparing the above algorithms they found that the use of a multi-layer neural network has proved to be more accurate in the case of the given task than the published regressive model.
The fourth is[11]**,** These study put forwards crop prediction by identifying various parameter like the type of soil, PH, phosphate, potassium, calcium, nitrogen, magnesium, sulfur, manganese, copper, iron, organic carbon depth, temperature, rainfall, humidity and parameter associated to the atmosphere. The authors design a network which correctly learns associations of effective climatic factors on crop yield, it can be used to estimate crop production in long or short term and also with enough and useful data can get an ANNs. This paper shows the ability of artificial neural network technology for the approximation and prediction of crop yields at rural area. In this paper, we shall examine one of the most common neural network architectures. Lastly, they analyze the result by using feed forward back propagation ANN model for each area and finds the most effective factors on crop yield.
The fifth is [12], This paper presents the capability and potential of machine vision with the well- trained multilayer neural network classifiers for shapes, sizes, and varietal type identification of unequal rice grain samples grown in the assorted agro-environmental zones in the country. In order to classify the seeds, they used Weka classification tools; Function, Bayes, Meta and Lazy methods. Classifiers they used from these methods are Logistics, SMO, Naïve Bayes Updateable, Multilayer Perceptron, Naïve Bayes, Bayes Net and Classifier

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

Multi Class. According to this study, The classification seeds can be done based on three different fold cross validation i.e. 10 fold, 5 fold and 2 fold, as well as a training set method, are also used. After analysis of the data, they try to observe that, the overall performance measures decreases as we decrease the fold value except for the Multilayer Perceptron classifier that gives the highest accuracy value 97.6% using 5 fold Cross Validation. To measure the performance they used K-Fold cross-validations and Training set method. Multilayer Perceptron (MLP) using 5-Fold cross-validation gives the highest performance of 97.6% among all the Weka classifier. They also experiment that Multilayer Perceptron gives the highest accuracy value when we use Training Set method which is 99.5% and Logistics gives second highest accuracy value of 98.6%. Finally, they observed that Training Set method gives higher accuracy than Cross Validation during the classification process. This research is concluded as the unsupervised artificial neural network gives better performance with 79% accuracy as compared to the supervised artificial neural networks which give 73% accuracy.
The last one is [13], These study used FCM clustering. He shows the prediction models based on supervised learning have a high accuracy, but they have several problems such as requirement a vast amount of classified data, and difficulty in accepting the data with new patterns that wouldn't be used for learning. Another weakness of prediction models based on supervised learning is the difficulty in gradual learning for real-time input data. On the other hand, the prediction model based on unsupervised learning is fast and need not have labelled data. However, the analysis of the prediction result is difficult, since no information for the learning data is given to us for learning. In order to lower these weaknesses, he proposes a context-aware framework for business using the hybrid FCM clustering algorithm that is a kind of unsupervised learning with the feature of supervised learning. The implementation of conducted research combines the higher accuracy rates of supervised learning and the flexibility of unsupervised learning. Therefore, the researchers on the enhanced algorithm to improve the prediction accuracy of the system whose data set is small. It also, however, demonstrated that the proposed model is capable of taking advantage of the numerical prediction based on unsupervised learning, which can automatically categorize the input data without the manager's intervention. So in our paper, we propose "Big data analytics framework to predict determinant factors to seeds classification" which is used to predict seed classes for seed production. The ANN model is used for prediction and it can be implemented using different ANN algorithms.

III. METHODOLOGY

To experiment this research, we need data but when we came to developing country having the data in an electronic way is a difficult task and this make all research in Africa face more challenging. For the experimentation of the research, data was collected from https://archive.ics.uci.edu/ml/datasets/seeds. Primary data sources include information collected and processed directly by us, such as observations, interviews, and focus groups by working with the Agronomists and then we apply this method because the data is not organized. Secondary data source includes information that we retrieved through pre-existing sources such as related research articles from the Internet.
Information we gathered come from a range of sources as we explain before to have data for the experimentation. Likewise, there are a variety of techniques we used when we gathered the primary and secondary data. Listed below there are some of the most common data collection techniques we used for collecting data: Interviews, Observations, and Documents analysis.
In order to experiment this research, collecting and structuring the data is used for analysis and it is one of the tasks that needs close attention in the process of analyzing the data. Collecting and preparing sample data is the first step in designing ANN models. In this research Area, Perimeter, Compactness, Length of Kernel, the width of the kernel, Asymmetry of coefficient and length of kernel grove are independent variables of the model whereas Kama, Rosa and Canadian are dependent variables of the model. With difficulty of having machine learning model, the developed model is tested or experimented using datasets i.e. using 1882 records and 8 attributes (7 independent attributes and 1 dependent attribute).
The methodology we used for implementation are predictive modelling. In predictive modelling, first data was collected, then a statistical model is formulated, predictions are made, and the model is revised as additional data become available from other sources. The methodology followed here is shown on figure 1.

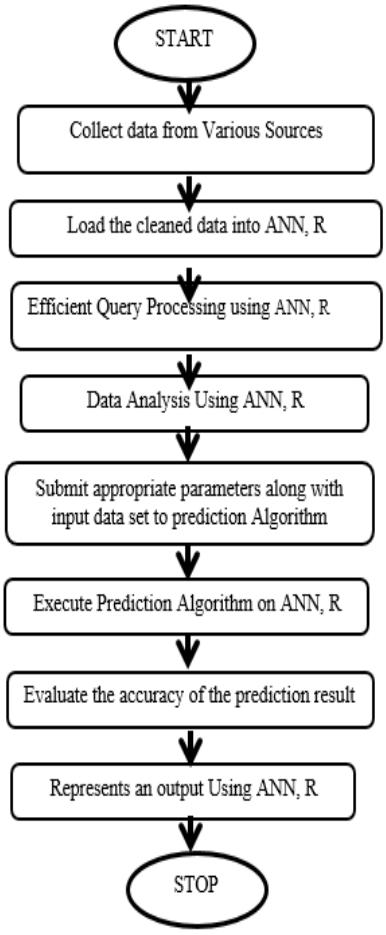

Fig 1: Work Flow of Proposed System using ANN and R [14]

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

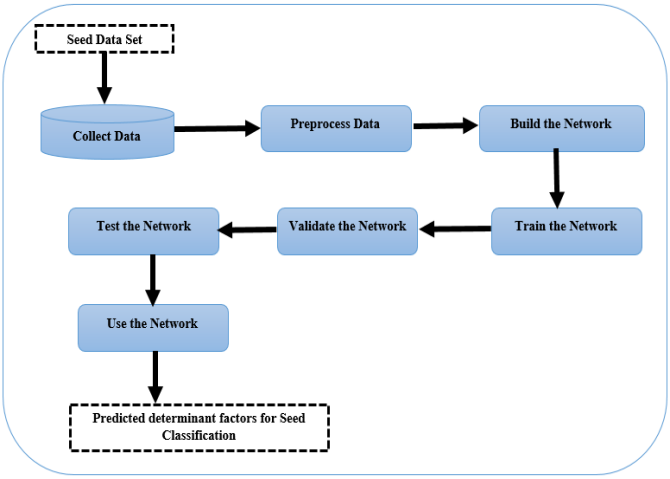

Fig 2: Proposed ANN model design flow basic steps

The proposed system workflow is shown in above figure 2. In the first step, seed datasets are collected from various sources and these data items are further pre-processed to make an effective input to prediction algorithm. After data cleaning, load into ANN and then apply query using math lab toolbox. We used ANN for prediction of seed classes and it can be implemented using learning algorithm. Then check the accuracy of the developed model.

The fundamental purpose of data preparation is to manipulate and transform raw data so that the information content is enclosed in the dataset can be exposed or made more easily accessible. In addition, data preparation involves enhancing and enriching the data in an attempt to improve knowledge discovery. There is recognition can be done by many arts as science in data preparation. Clearly, it takes additional effort for data preparation and hence, the question of the cost of doing it versus the benefits arises.

Data collection, data pre-processing (data cleaning, attribute selection, data formatting and transformation, dimensionality reduction and the like) are the most important activities under data preparation, which finally resulted in creating target data set. For the experimentation of the model, datasets are from UCI Machine learning repository were selected for the classification. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms [15]. UC has kept detailed information about the seed. This information is described through different attributes. The specific attributes that can describe the above-mentioned information are Area, Perimeter, Compactness, Length of Kernel, width of the kernel, Asymmetry of coefficient and length of kernel grove.

After data collection, data pre-processing procedures were conducted to train the ANNs more efficiently and then coordinate systems of the all the data were transformed to the same coordinate system so that all the data fit the model. These procedures might include solving the problem of missing data and normalizing datasets. Normalization procedure before presenting the input data to the network is generally a good practice, since mixing variables with large magnitudes and small magnitudes will confuse the learning algorithm on the importance of each variable and may force it to finally reject the variable with the smaller magnitude[16]. Normalize (pre-process) the dataset into input data, which is suitable to use in the neural network. Some variables such as year, month are excluded from the input data by discussing with domain experts. Some duplicate records are removed by using the technique "remove constant rows" which is available in MATLAB. The independent variables of the model are represented by a different number with different ranges. Dependent variables are also represented by 1, 2 and 3 for Kama, Rosa and Canadian of the wheat seed respectively.

IV. EXPERIMENTS AND RESULTS

Different experiments have been conducted to choose which model is appropriate for wheat seed classification. In this research, the artificial neural network model used is feed forward neural network with back propagation training algorithm. The model was constructed with an architecture having three layers (Input layer, one hidden layer, and an output layer). In this model, we have 7 input parameters so that the input layer have 7 input neurons. To determine the number of hidden neurons different rule of thumbs have been adopted, to include these rules the model was experimented starting from n/2 to 2n+1. So that the neurons in the hidden layer are varied from 4 up to 15. This model has three output neuron either Canadian, Rosa or Kama. Four different divide functions (dividerand, divideint, divideind and divideblock) are varied to see which divide function works optimally in this model. In this model, seven different training algorithm (trainlm, trainscg, trainrp, traincgf, traincgp, traincgb and trainoss) is used to train the network so that the network with the best result is chosen for the classification of seeds. Two different transfer function tan sigmoid, log sigmoid and purelin is used in the hidden layer and linear transfer function in the output layer. MSE and confusion matrix are used for performance evaluation of the training functions.

A. Network Configuration and Training

This section defines the steps deployed in determining critical RTA factors classification model using ANN back propagation algorithm. The data set file (*seedclassification.mat)* contains a predefined set of inputs and target vectors. The input vectors define data regarding different factors contributing to seed classification and the target values define the relative output of the input vectors to classify seeds. The input matrix consists of 1890 column vectors of 7 geometric variables for the corresponding 1890 target vectors. The next step in the experimentation is to create a neural network (using *wsc=feedforwardnet(1);* command and train it till it has learned the complex relationship between inputs and corresponding outputs. Note that the number of neurons in the output layer is automatically set to one. In this experiment, a feed-forward network with the default tan-sigmoid transfer function in the hidden layers and linear transfer function in the output layer was used.

Now, we can view the newly created network using *view(wsc*); command, where *wsc* is the name of the our network, and the following window is displayed.

To train the network (using *[wsc,tr]=train(wsc, i,o)* command are used using the data sets collected so far. The network uses the default Levenberg-Marquardt Algorithm to train the network.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

B. *Selection of hidden neurons to show model Performance*

The first experiment is conducted by varying the hidden layers from 4 to 15 neurons. The default training algorithm is "trainlm", the default transfer function in the hidden layer is tansig and default divide function dividerand. The performance result of each hidden neuron is shown in table 1 below.

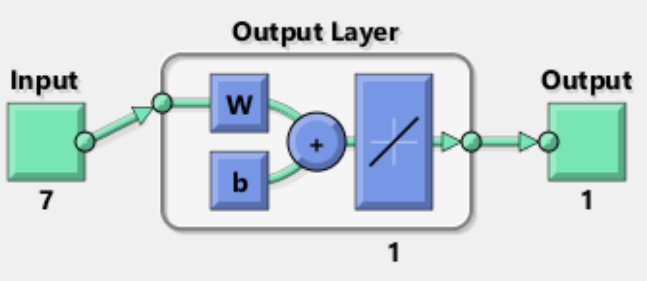

Fig 3: Un-configured neural network view

TABLE 1

Hidden Neurons Comparison

Number of neurons	Best validation performance (MSE)	Epoch
4	0.011086	21
5	0.0068835	16
6	0.010035	11
7	0.011934	12
8	0.014759	21
9	0.010481	21
10	0.011867	16
11	0.009037	20
12	0.0047947	32
13	0.010435	43
14	0.012928	26
15	0.0094526	6

From table 1 above five best results are selected. As we can see from the table an experiment with 5, 6, 11, 12 and 15 hidden neurons with best validation performance 0.010035, 0.010035, 0.009037, 0.0047947 and 0.0094526 respectively has minimum MSE, so these results are selected as best results. From now onwards these five hidden neurons are used to show the performance of divide functions, transfer functions, and training algorithms

C. *Selection of divide function to observe model performance*

The four standard divide functions are varied while keeping the default training algorithm trainlm and default transfer function in the hidden layer tan sigmoid. To select low MSE of the network, divide function (dividerand, divideint, divideblock and divideind) algorithms are used. And using 5, 6, 11, 12 and 15 hidden neurons we select the best performance. From the experiment, we observe that dividerand and divideint functions have minimum MSE. Table 2 below shows the performance of the four divide functions using 5, 6, 11, 12 and 15 hidden neurons.

TABLE 2

Comparison among dividerand and divideint with optimal MSE

Number of neurons	Divide function	Best validation Performance (MSE)	Epoch
5	dividerand	0.0068835	16
6	divideint	0.0070968	9
11	divideint	0.0060921	17
12	dividerand	0.0047947	32
15	divideint	0.0071707	34

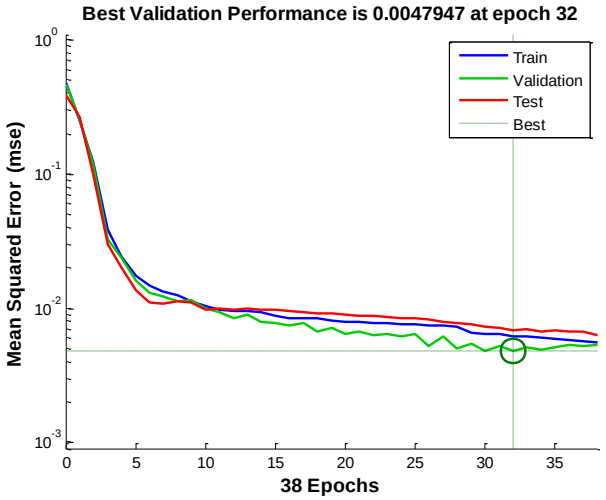

Fig 4: Best validation performance of divide functions using 12 hidden neurons

As we can see from table 2 and graphs using above the best validation performance 0.0047947 at epoch 32 is the optimal with the divide function divideirand with 12 hidden networks.
As we can show from the above tables and graphs using the five hidden neurons we get dividerand is the best in each of the hidden neurons. So that throughout this research to see the performance of the transfer functions and training algorithms dividerand is used as a best divide function.

D. *Experiments to observe the Performance of Transfer Function*

By setting the standard transfer functions in to tan sigmoid, log sigmoid and pure linear for the best five hidden neurons, this study has checked which transfer function works best for this model. The performance of each transfer function is checked by varying the number of hidden networks. To find the optimal transfer function experiments are performed by using default parameter (training function= "trainlm"). Here the performance result of the transfer functions are compared. Table 3 below shows the comparison among transfer functions using 5 hidden neurons.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

TABLE 3

Comparison among transfer functions with 5 hidden neurons

Number of neurons	Transfer function	Best validation performance (MSE)	Epoch
5	tansig	0.0068835	16
5	logsig	0.0081546	11
5	purelin	0.066506	4

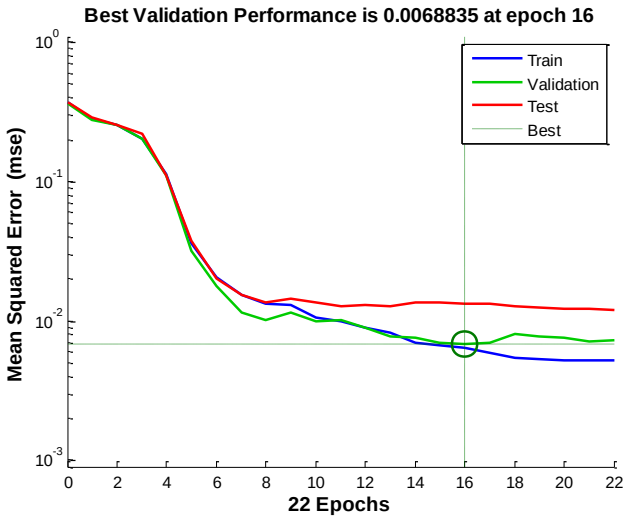

Fig 5: Best validation performance for tangsig with 5 hidden neurons

As we can see from table 3 above the best validation performance 0.0068835 at epoch 16 is the optimal with transfer function tansig. Figure 5 above shows the best validation performance for tansig. As we can see from Figure 5 above the best validation performance is 0.0068835 at iteration 16.
Table 4 below shows the comparison among transfer functions using 6 hidden neurons.

TABLE 4
Comparison among transfer functions with 6 hidden neurons

Number of neurons	Transfer function	Best validation performance (MSE)	Epoch
6	tansig	0.010035	11
6	logsig	0.018109	20
6	purelin	0.064346	5

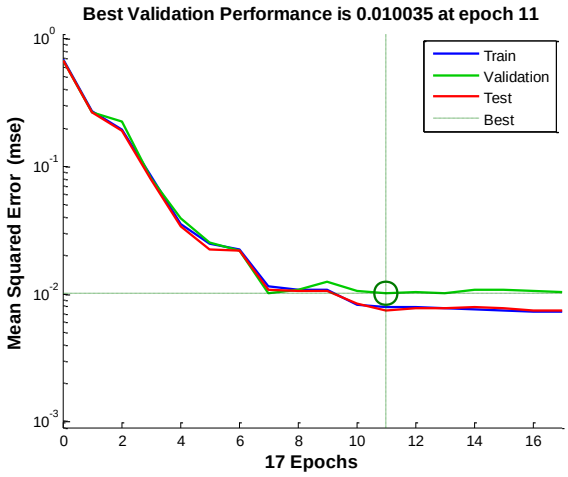

Fig 6: Best validation performance for logsig with 6 hidden neurons

As we can see from table 4 above the best validation performance 0.010035 at epoch 11 is the optimal with transfer function tansig.
Figure 7 below shows the best validation performance for tansig. As we can see from Figure 6 above the best validation performance is 0.010035 at iteration 11. Table 5 below shows the comparison among transfer functions using 11 hidden neurons.

TABLE 5

Comparison among transfer functions with 11 hidden neurons

Number of neurons	Transfer function	Best validation performance (MSE)	Epoch
11	tansig	0.009037	20
11	logsig	0.0094811	17
11	purelin	0.68338	4

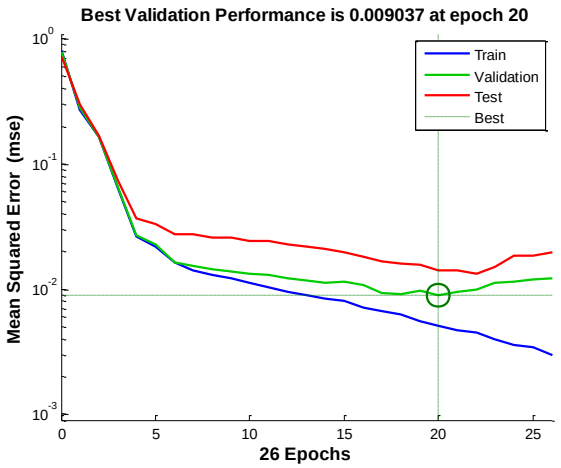

Fig 7: Best validation performance for tansig with 11 hidden neurons

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

As we can see from table 5 above the best validation performance is 0.009037 at epoch 20 is the optimal with transfer function tansig. Figure 8 below shows the best validation performance for tansig. As we can see from Figure 7 above the best validation performance is 0.009037 at iteration 20. Table 6 below shows the comparison among transfer functions using 12 hidden neurons.

TABLE 6
Comparison among transfer functions with 12 hidden neurons

Number of neurons	Transfer function	Best validation performance (MSE)	Epoch
12	tansig	0.0047947	32
12	logsig	0.011185	39
12	purelin	0.058283	3

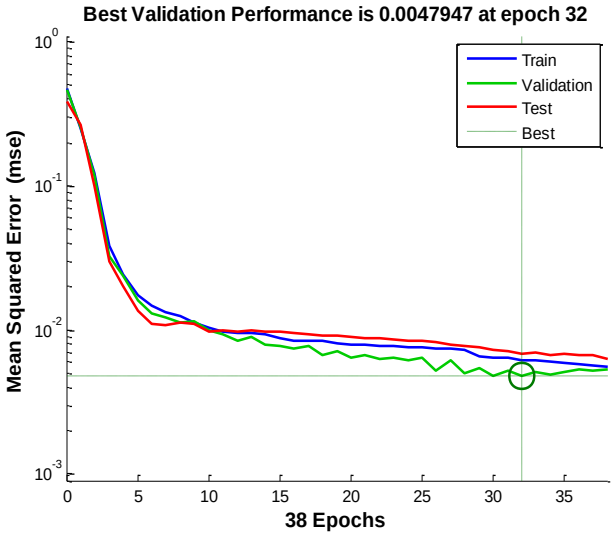

Fig 8: Best validation performance for tansig with 12 hidden neurons

As we can see from table 6 above the best validation performance 0.0047947 at epoch 32 is the optimal with transfer function tansig. Figure 9 below shows the best validation performance for tansig. As we can see from Figure 8 above the best validation performance is 0.0047947 at iteration 32. Table 7 below shows the comparison among transfer functions using 15 hidden neurons.

TABLE 7
Comparison among transfer functions with 15 hidden neurons

Number of neurons	Transfer function	Best validation performance (MSE)	Epoch
15	tansig	0.0094526	6
15	logsig	0.013807	15
15	purelin	0.058899	3

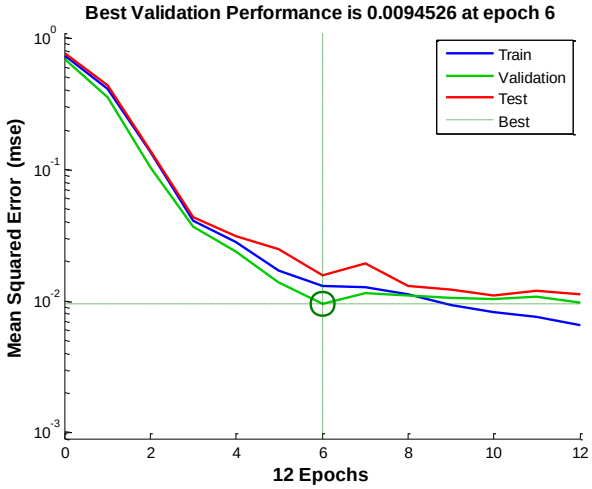

Fig 9: Best validation performance for tansig with 15 hidden neurons

As we can see from table 7 above the best validation performance 0.0094526 at epoch 6 is the optimal with transfer function tansig. Figure 9 above shows the best validation performance for tansig.

From all experiments using 5, 6, 11, 12 and 15 hidden neurons with tansig, logsig, and purelin algorithms we found that tansig have minimum MSE. So in the following table 9, we show that the result of tangsig using 5, 6, 11, 12 and 15 hidden neurons.

TABLE 1
The Comparison between tansig with optimal MSE at different number of Neurons

Number of neurons	Transfer function	Best validation performance (MSE)	Epoch
5	tansig	0.0068835	16
6	tansig	0.010035	11
11	tansig	0.009037	20
12	tansig	0.0047947	32
15	tansig	0.0094526	6

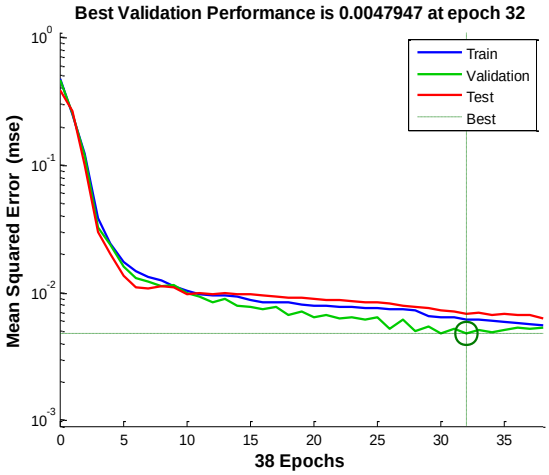

Fig 10: Best validation performance of transfer functions using 12 hidden neurons

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

From table 8 above we observe that the best validation performance 0.0047947 at epoch 32 is optimal with the transfer function tansig using 12 hidden networks. So tansig was applied for each of training functions in order to find the optimal training function for this model. The performance of this model is shown in figure 10 above.

E. *Experiments to show Performance of Training Functions*

Seven different training functions (trainlm, trainscg, trainrp, traincgb, traincgp, traincgf, and trainoss) are taken to see and compare the performance of each training function so that the training function with minimum mean squared error are taken as the best for the seed classification. The optimal training function experiments are performed by using the optimal hidden neurons (5, 6, 11, 12 and 15). Finally to show best MSE of the network we used that minimum divide function (dividerand) and minimum transfer function (tansig). The selected algorithm are the default, so using this default algorithm we experiment all training functions. Now we have selected the best result from each of the training functions. Then after the performance of these best results are compared to choose the training function with a minimum mean squared error. Table 9 below shows the performance comparison of the best results of each training function.

TABLE 9
Comparison of the best results of each training functions

Training function	Number of neurons	Best validation performance (MSE)	Epoch
trainlm	12	0.0047947	32
trainscg	6	0.0066237	83
trainrp	5	0.011786	55
traincgb	5	0.0099212	78
traincgp	12	0.0085349	142
trainoss	6	0.01444	6
traincgf	6	0.010227	79

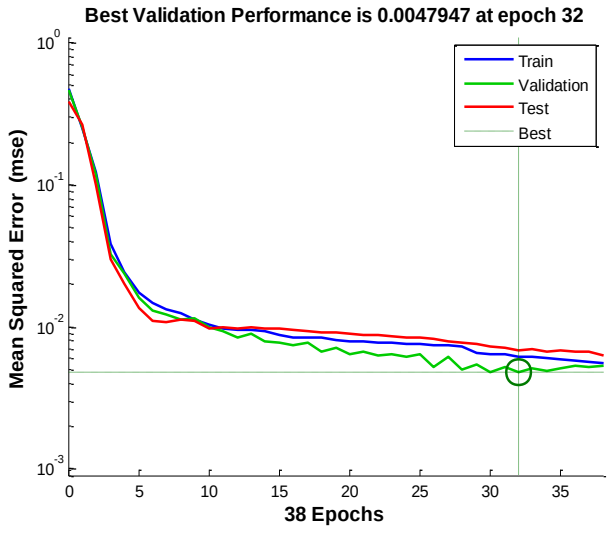

Fig 11: Best validation performance of the optimal training function

As we can observe from table 9 above a performance result with best validation performance 0.0047947 that occur at epoch 32 using 12 neurons in the hidden layer obtained by applying trainlm as a training function have a minimum mean squared error from the rest. So trainlm is the optimal training function for this study. Figure 12 below shows the performance result of the optimal training function. As it is shown from the experiments conducted so far a performance result with trainlm as its training function, with 12 neurons in the hidden layer, with tan sigmoid as transfer function in the hidden layer, dividerand as a divide function, with best validation performance 0.0047947 that occur at epoch 32 is the optimal model. Table 10 below indicates this model.

TABLE 10
The optimal model

Training function	Hidden neurons	Transfer function	Divide function	Best validation performance (MSE)	Epoch
trainlm	12	tansig	divide rand	0.0047947	32

The first experimentation answered the research question: *What is the most effective ANN architecture for determining Seed Classification factors in the research center?* To experiment this research question we train the network using different divide function, transfer function, and training function. After training the network using optimal hidden neuron we generate ANN architecture. The network architecture for the optimal model is depicted in figure 12 below.

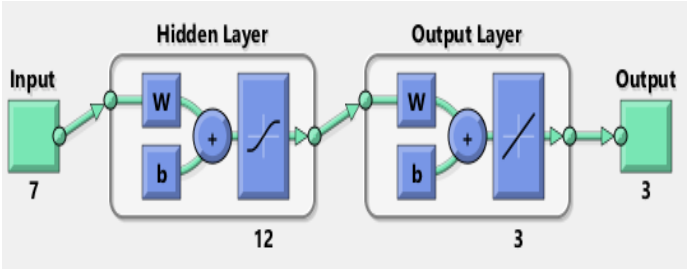

Fig 12: Neural network architecture of the selected model

As we can see from the above figure the network has 7 input neurons, 12 hidden neurons and three outputs and the above ANN architecture is implemented using 1882 data and with 7 independent attributes. Therefore if architecture is experimented using small data it also experiments big data files this means the architecture we developed is not limited to the small data file so within big data files the framework can also work.

The Second experimentation answered the research question: *How should predictor variables be represented to accurate predictive patterns for seed classes?* To answer this research question we used confusion matrix and regression plot after the optimal model is selected. The confusion matrix for the optimal model is shown in figure 13 below.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

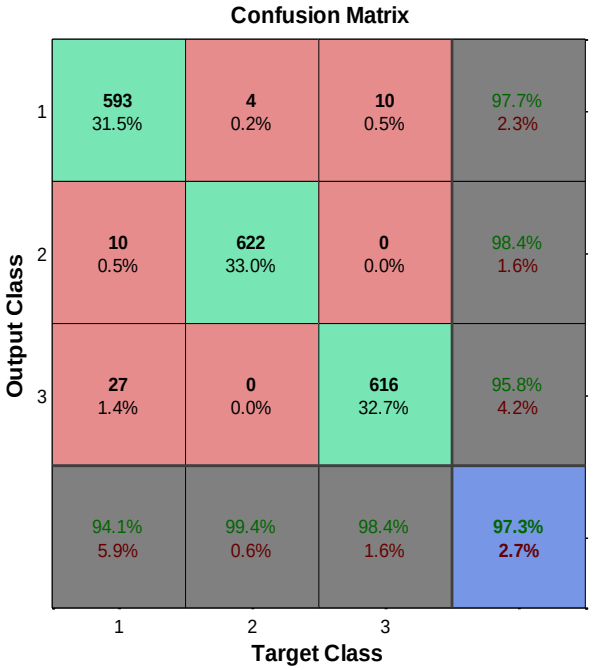

Fig 13: Confusion Matrix of the Selected Model

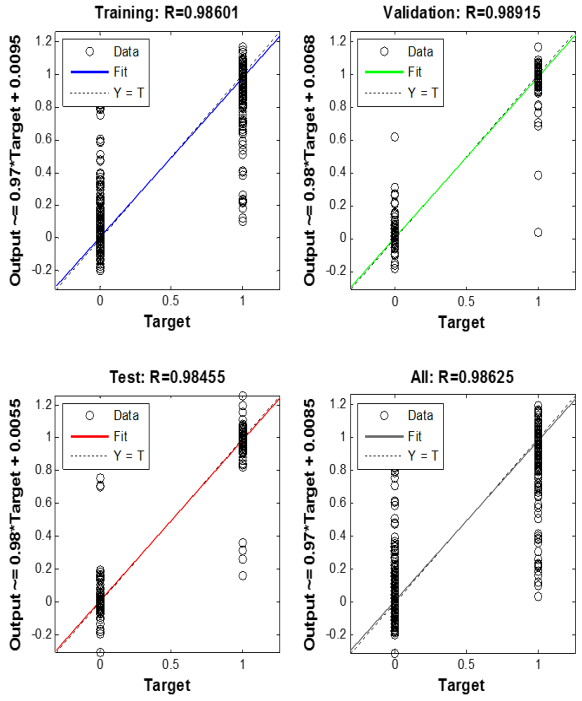

Fig 14: Regression plot of the Optimal Model

As we can observe from the above plot, the diagonal cells in each of the above table show the number of cases that were correctly classified, and the off-diagonal cells show the misclassified cases. The blue cell in the lower right illustrates the overall percent of correctly classified cases in green and the total percent of misclassified cases in red. From Figure 13 above for the total confusion matrix, 593 instances are classified as true negative but 37 instances are classified as false negative. And 622 instances are classified as true positive but 4 instances are classified as false positive. And the third column shows the third target values. As we can observe 616 instances are classified as true positive and 10 instances are classified as false positive. The overall confusion matrix indicated that 97.3% of the cases are correctly classified which is written by a green color and 2.7% cases are misclassified

In addition to confusion matrix to evaluate the network performance regression plot is created, which shows the relationship between the outputs of the network and the targets. If the training were perfect, the neural network outputs and targets would be exactly equal. Figure 14 above depicts the regression plot of the optimal model. The above regression plots display the network outputs with respect to targets for training, validation and test sets. For a perfect fit, the data should fall along a 45-degree line, where the network outputs are equal to the targets. The dashed line in each plot represents the perfect result – outputs = targets. The solid line represents the best fit linear regression line between outputs and targets. If R = 1, this indicates that there is an exact relationship between outputs and targets. If R is close to zero, then there is no linear relationship between outputs and targets. For this problem, the R values in training, validation, and test are close to 1 with the overall R-value of 0.98625. This indicates an optimal result. The last experimentation answered the research question: *What are the most determinant predictors to Seed Classification?* To answer this research question we conduct an experiment using the optimal model obtained from the experiment. So the researcher determined the relative importance of variables using [ranked, weights] = relieff (i, o, 1882) script. After running the script we found that relative importance of geometric parameters as shown in figure 15 below.

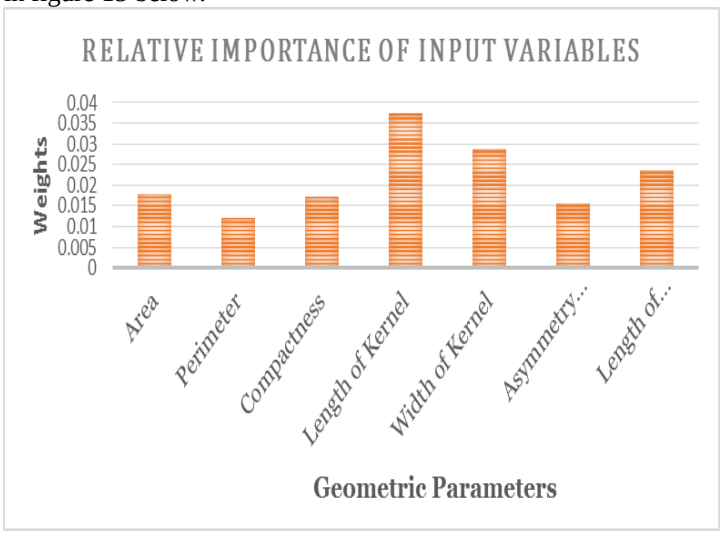

Fig 15: Relative importance of input variables

As we can see from Figure 15 above length of kernel is the first influential variable with the weight of 0.0373, the second influential variable is length of kernel grove with weight of 0.0236, the fourth influential variables is area with weight of 0.0179 and the fifth, the six and the seventh influential variables are Perimeter, Asymmetry of coefficient and compactness with weight of 0.0120, 0.0157, and 0.0172 respectively. Under this experimentation, we determined the relative importance of the geometric parameters during seed classification, because the determination of influential variables by agronomic experts is

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

subjective. In order to avoid the subjective determination of influential variables by experts, this study determined the significance level by using the weights of the variables. This shows the result obtained from this study is better than domain experts because there is no weight-based identification of input variables by the domain experts and significance of input variables is done based on their experience.

V. CONCLUSIONS

In this study, ANN method for seed classification and Variety types of seeds are has been implemented for Melkassa Research Center. Seven 7 independent variables were selected and examined in developing the model. The model is developed for prediction of determinant factors of seeds based on the artificial neural network. The researcher considers the appropriate neural network architecture such as hidden layer neurons, divide function, transfer function and training function, in order to achieve the optimal result by comparing their performance in terms of MSE to reach the optimal possible answer. By selecting trainlm training function, dividerand divide function, tansig transfer function and one hidden layer with 15 neurons, the model reached an optimal solution. The significance levels of input variables are done using the model which contains optimal hidden layer neurons, divide function, transfer function, and training function. From the input variables Length of Kernel is the first influential variable with weight of 0.0373, length of kernel grove is the second influential variable with weight of 0.02336, width of kernel is the third influential variable with weight of 0.0289, Area is the fourth influential variable with weight of 0.0179, the fifth influential variable is perimeter with weight of 0.0120, the six influential variable is Asymmetry of coefficient with the weight of 0.0157 and compactness is the last influential variable with weight of 0.0172. When we run confusion matrix we found that 97.3% accuracy with the trained network of ANN. More research could also still be done in making an empirical test of this model to extend this study and to come up with a generic model. Further research could also be done in developing ANN-based models for seed classification for the coming years in the region. Finally, the researcher tried to compare the result generated by this study with the domain experts and two international papers that are done for classifying seeds. This research is mostly conducted for an academic purpose. However, that the results of this study are applied to address practical problems of the research center. This research work can pay a lot towards solving research center problems. The results of this study have also shown that the data Analytics technology particularly ANN are appropriate in the determination of seed classification, data analysis, and decision-making process. Hence the researcher recommends that domain experts can use the model obtained from this study. Based on the results obtained from the study, the researcher makes the following recommendations:

- It is better to concentrate on the geometric parameters and factors gained using the proposed model in classifying seed because this result is obtained by following scientific research.
- It was found that Length of Kernel was the most critical factor in classifying seeds into its specified categories so that the researcher recommends that Length of Kernel should be given more emphasis in seed classification.
- It was found that compactness was the least critical factor in classifying seeds so that the researcher recommends that compactness should not be given more emphasis during the classification process.

Other researchers can extend this research by taking the following future research directions:

- This study used artificial neural network to model determinant factors in classifying seeds. Further study can be done using fuzzy based modeling for determining the critical factors of seed classification.
- Someone can develop a model for determining critical factors in seed classification using support vector machine.
- Someone can develop a graphical user interface as a prototype for this model.

ACKNOWLEDGMENT

I am grateful to express my deep appreciation and thanks to my God, for his overall support and Love.

I would like to express my sincere thanks to Dileep Kumar G., my research advisor, for providing me with all his support and expert guidance with constant encouragement throughout the research and Mr. Getinet Yilma for his advice starting from proposal preparation to till the end of these research. I would also like to thanks, all Data Science SIG (Special Interest Group) members for their important suggestions and giving me different workshops and resources. At last, I should not neglect to thank my dear beloved for her understanding and encouraging support she provided throughout my study.

REFERENCES

[1] M. Stubbs, "Big Data in U . S . Agriculture," 2016.
[2] Wikipedia, "Agriculture in Ethiopia." [Online]. Available: https://en.m.wikipedia.org/wiki/Agriculture_in_Ethiopia. [Accessed: 07-Oct-2015].
[3] B. J I ET AL, "Artificial neural networks for rice yield prediction in mountainous regions," *J. Agric. Sci.*, vol. 145, pp. 249–261, 2007.
[4] Prof. K.Suzuki, "Diagnosing Skin Diseases using an Artificial Neural Network," in *in Artificial Neural Networks-Methodological Advances and Biomedical Applications*, Shanghai, InTech, p. 374.
[5] Q. K. Al-Shayea and I.S. H. Bahia, "'Urinary System Diseases Diagnosis Using Artificial Neural Networks,'" *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 10 No.7.
[6] Mr. A.Vishwa, "pre-diagnosis of lung cancer using feed forward neural network and back propagation algorithm," *Int. J. Comput. Sci. Eng.*, vol. 3 No.9, pp. 3313–3319, 2013.
[7] F. Ghamari, S., Borghei, A. M., Rabbani, H., Khazaei, J., & Basati, "Modeling the terminal velocity of agricultural seeds with artificial neural networks," *Afr. J. Agric. Res*, vol. 5(5), pp. 389–398, 2010.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

[8] A. R. Parnian, "Autonomous Wheat Seed Type Classifier System," vol. 96, no. 12, pp. 14–17, 2014.
[9] C. S. Silva and U. Sonnadara, "Classification of Rice Grains Using Neural Networks," vol. 29, pp. 9–14, 2013.
[10] J. Šťastný, V. Konečný, and O. Trenz, "Agricultural data prediction by means of neural network," vol. 2011, no. 7, pp. 356–361, 2011.
[11] S. S. Dahikar and S. V Rode, "Agricultural Crop Yield Prediction Using Artificial Neural Network Approach," vol. 2, no. 1, pp. 683–686, 2014.
[12] R. H. Ajaz and I. Campus, "Seed Classification using Machine Learning Techniques," vol. 2, no. 5, pp. 1098–1102, 2015.
[13] J. Kim, "A Prediction Model based on Big Data Analysis Using Hybrid FCM Clustering," pp. 337–339, 2014.
[14] C. Balaji, "Environment Change Prediction to Adapt Climate- Smart Agriculture Using Big Data Analytics," vol. 4, no. 5, pp. 1995–2001, 2015.
[15] UCI, "UCI Machine Learning Repository," 2014.
[16] H. A. and H. A. A. H. Maitha, S. Al, "Using MATLAB to Develop Artificial Neural Network Models for Predicting Global Solar Radiation in AI Ain City-UAE," *Eng. Educ. Res. using MATLAB*, 2011.