



**ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY
SCHOOL OF GRADUATE STUDIES
DEPARTEMENT OF SOFTWARE ENGINEERING**

**A PREDICTIVE MODEL TO PREDICT SEED CLASSES USING
MACHINE LEARNING**

BY:

TEKALIGN TUJO GURMESSA

ID_No:- GSR/02989/07

February 2017

ADAMA, ETHIOPIA



**ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY
SCHOOL OF ELECTRICAL AND COMPUTING
ENGINEERING
DEPARTEMENT OF SOFTWARE ENGINEERING**

**A PREDICTIVE MODEL TO PREDICT SEED CLASSES USING
MACHINE LEARNING**

**A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF
ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY IN PARTIAL
FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER
OF SCIENCE IN SOFTWARE ENGINEERING**

BY:

TEKALIGN TUJO GURMESSA

Name and Signature of Members of the Examining Board

Name	Title	Signature	Date
Dileep Kumar G.	Advisor,	_____	_____
_____	Chairperson,	_____	_____
Dr. Wondwossen M.	External examiner's	_____	_____
_____	name,	_____	_____
____ Girma D. ____	Internal Examiner's name	_____	_____
_____	Department Head	_____	_____
_____	School Dean	_____	_____

February 2017

ADAMA, ETHIOPIA

DECLARATION

I declared that this thesis is my original work and has not been presented for a degree in any other Universities, and all sources of material used for the study have been accordingly acknowledged.

Tekalign Tujo Gurmessa
February 2017

This Thesis work has been submitted for examination with my approval as a University advisor.

Dileep Kumar G
February 2017

Dedication

This thesis is dedicated to my Families, Lemi Tujo, Abunesh Tujo, My Mom Motu Emiru and My beloved one Elifenes Yitagesu, *the most caring and loving.*

Acknowledgement

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.

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 modeling 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.

Table of Contents

DECLARATION	i
Dedication	ii
Acknowledgement.....	iii
Abstract	iv
List of Tables.....	viii
List of Figures	ix
List of Abbreviations	x
CHAPTER One.....	1
INTRODUCTION	1
1.1 Background and Introduction of Study.....	1
1.2 Motivation.....	3
1.3 Statement of Problems	4
1.4 Objectives of Research	5
1.4.1 General Objective.....	5
1.4.2 Specific Objectives.....	5
1.5 Research Question	6
1.6 Scope and limitation of Study	6
1.7 Significance of Study	7
1.8 Methodology of the Study	8
1.8.1 Research Framework.....	8
1.8.2 Design Science Research Methodology (DSRM) Process Model	9
1.9 Brief Survey of Thesis Chapters	11
CHAPTER Two	12
LITERATURE REVIEW	12
2.1 Overview of Predictive Data Mining.....	12
2.2 Machine Learning Algorithms	12
2.3 Predictive Modeling.....	13
2.4 Basics of Artificial Neural Networks.....	14
2.5 Types of Topologies in Artificial Neural Network.....	14
2.5.1 Feed-Forward Artificial Neural Networks	14
2.5.2 Recurrent Artificial Neural Networks	16
2.6 Artificial Neural Networks Types.....	17
2.6.1 Multilayer Perceptron.....	17

2.6.2 Radial Basis Function Networks	17
2.6.3 Kohonen Neural Network	17
2.7 Working Principles of ANN and Back Propagation Algorithm.....	18
2.8 Related Work	21
CHAPTER Three.....	26
METHODOLOGY	26
3.1 Research Design	26
3.2 Data Source.....	26
3.2.1 Data collection and Analysis.....	26
3.2.2 Data Cleaning Tools.....	28
3.3 Methodology for Implementation	28
3.4 Data Preparation	30
3.4.1 Data Understanding.....	30
3.4.2 Data Preprocessing.....	31
3.5 Process Model	32
3.6 The Predicting Model Platform	33
3.7 Data Design.....	34
CHAPTER Four	35
MODEL DEVELOPMENT	35
4.1 Designing Artificial Neural Network models	35
4.1.1 Building the Network Structure	35
4.1.2 Number of neurons in the hidden layer	37
4.1.3 Training Neural Network	39
4.1.4 Transfer Function	40
4.1.5 Divide functions	42
4.1.6 Overfitting	42
4.1.7 Training Stopping Criteria	43
4.1.8 Testing the Network.....	43
4.1.9 Performance Evaluation	43
4.2 Critical Seed Classifier Factors Determination.....	44
CHAPTER Five.....	45
EXPERIMENTAL RESULTS AND DISCUSSION	45
5.1 Network Configuration and Training.....	46
5.2 Selection of hidden neurons to show model Performance	48
5.3 Selection of divide function to observe model performance	49

5.4 Experiments to observe the Performance of Transfer Function.....	50
5.5 Experiments to show Performance of Training Functions.....	56
5.6 Determine Relative Importance of Input Variables	61
CHAPTER Six.....	64
CONCLUSION, RECOMMENDATION AND FUTURE WORKS	64
6.1 Conclusion	64
6.2 Recommendation	65
6.3 Future Work.....	66
BIBLIOGRAPHY	67
APPENDICES.....	69
Appendix A [50].....	69
Appendix B.....	71

List of Tables

Table 1: Comparison of Related Papers.....	25
Table 2: Attributes with their data types and description	30
Table 3: Comparison of hidden layers based on their task	37
Table 4: Training algorithms used in math lab neural network toolbox	39
Table 5: Divide Functions[47]	42
Table 6: Hidden Neurons Comparison.....	48
Table 7: Comparison among dividerand and divideint with optimal MSE.....	49
Table 8: Comparison among transfer functions with 5 hidden neurons	50
Table 9: Comparison among transfer functions with 6 hidden neurons	51
Table 10: Comparison among transfer functions with 11 hidden neurons	52
Table 11: Comparison among transfer functions with 12 hidden neurons	53
Table 12: Comparison among transfer functions with 15 hidden neurons	54
Table 13: The Comparison between tansig with optimal MSE at different number of Neurons.....	55
Table 14: Comparison of the best results of each training functions	57
Table 15: The optimal model	58
Table 16: Relative importance of symptoms according to experts and the proposed model.....	62

List of Figures

Figure 1: Information Systems Research Framework [11]	9
Figure 2: Design Science Research Methodology (DSRM) Process Model [12]	10
Figure 3: Predictive Analytics Process [18].....	13
Figure 4: Feed-forward Artificial neural network [20]	15
Figure 5: Fully recurrent artificial neural network [20]	16
Figure 6: Working principle of an artificial neuron [20]	19
Figure 7: Work Flow of Proposed System using ANN and R [35]	29
Figure 8: Proposed ANN model design flow basic steps.....	31
Figure 9: Process Methodology	32
Figure 10: Modified System Framework [35]	33
Figure 11: Basic Flow for Designing Artificial Neural Network Model [38]	35
Figure 12: Neural Network Architecture of the proposed model.....	38
Figure 13: Figure that illustrates Activation Function[39]	40
Figure 14: Log-Sigmoid Transfer Function [46]	41
Figure 15: Tan sigmoid Transfer Function [46].....	41
Figure 16: Linear transfer function [46]	41
Figure 17: Un-configured neural network view.....	46
Figure 18: Neural Network training progress.	47
Figure 19: Best validation performance of divide functions using 12 hidden neurons.....	49
Figure 20: Best validation performance for tangsig with 5 hidden neurons	51
Figure 21: Best validation performance for logsig with 6 hidden neurons.....	52
Figure 22: Best validation performance for tansig with 11 hidden neurons	53
Figure 23: Best validation performance for tansig with 12 hidden neurons	54
Figure 24: Best validation performance for tansig with 15 hidden neurons	55
Figure 25: Best validation performance of transfer functions using 12 hidden neurons.....	56
Figure 26: Best validation performance of the optimal training function.....	57
Figure 27: Neural network architecture of the selected model	58
Figure 28: Confusion Matrix of the Selected Model	59
Figure 29: Regression plot of the Optimal Model	60
Figure 30: Relative importance of input variables.....	61

List of Abbreviations

Abbreviations	Meanings
▶ ANN:-----	Artificial Neural Network
▶ BI&A:-----	Business Intelligence and Analytics
▶ CBR:-----	Case-based reasoning
▶ CM:-----	Confusion Matrix
▶ DA:-----	Development Agent
▶ DBMS:-----	Database Management System
▶ DSRM-----	Design Science Research Methodology Process Model
▶ FCM:-----	Fuzzy Clustering Means
▶ IS:-----	Information System
▶ ML:-----	Machine Learning
▶ MLP:-----	Multilayer Perceptron
▶ MSE:-----	Mean Square Error
▶ OLAP:-----	OnLine Analytical Processing
▶ PDP:-----	Parallel Distributed Processing

CHAPTER One

INTRODUCTION

1.1 Background and Introduction of Study

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 analyzed 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 center 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 center experts. ANN holds one of the keys for farmers control centers 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 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 modeling becoming very popular in different areas of agriculture, specially, in the areas where straight statistical modeling 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.

1.2 Motivation

The necessitated point to this research is to address challenges of food security and climate change that explicitly aims for the objectives of sustainably increasing agricultural productivity, to support agricultural research workers specially agronomist. The involvements of this research can benefits; managing the seed classification risks, understanding and predicting for suitable environment change, for example, appropriate soil for appropriate seed. These research also strengthening national and local institutions to enable farmer management of environment risks and adoption of context-suitable agricultural activities. So, if further research is conducted that can integrate agricultural activity with Information Technology; which can visualize the hidden pattern of attributes in multidimensional way, it is possible to identify the specific category of the weather, Temperature, humidity, land etc. and this is relevant for decision makers to take action based on the identified categories.

The impacts of unsuitable environment for chosen seed have major effects on agricultural production. So if there is a decrease of production in certain areas of the country it results in global imbalances. The solution to this problem can be solved by changing the agriculture in to more profits by using different technologies and tools by collecting real-time data on weather, soil and air quality, seed kind and even equipment and labor costs and availability, predictive analytics can be used to make smarter decisions.

1.3 Statement of Problems

One of the most difficulty jobs for agronomist in general is, subjective decision about seeds. Human beings always makes mistakes and because of his limitation errors occur during classification of seeds.

The crop productivities decreased due to various things such as climate change, soil degradation, water landscape, diseases, and biodiversity and etc. factors. Controlling agricultural problems are important to minimize crop losses and product reduction in the field.

Agricultural production has evolved into a complex business requiring the accumulation and integration of big data and information from many diverse sources. In order to remain competitive, the modern farmer often relies on agricultural specialists and advisors who provide information for decision making. Unfortunately, agricultural specialist assistance facing different problems while classify seed types. In order to solve this problem, knowledge based system has been identified as a powerful tool with extensive potential in improving agricultural problems[8].

Agriculture problems also requires a big data and application of big data from different interacting fields of data science and engineering to make a suitable analysis for agricultural products.

In Ethiopia agricultural development there is a problem of using different machine learning algorithm while working on experimentation of agricultural activities and they are not centrally connecting to each with automated system specially using big data analytics tools.

In Melkassa Research Center, identification of determinant factors for classification of seed is done based on subjective approach i.e. some agronomist assumes that length of kernel is the most influential variable for classifying seed, length of kernel grove is the second, width of kernel is the third, Area is the fourth, perimeter is the fifth, compactness is the six and asymmetry of coefficient is the last determinant factors.

Based on interview we made for the agronomists, they assumes that, width of kernel is the most influential variable, Area is the second, length of kernel grove is the third, length of kernel is the fourth, perimeter is the fifth, the Asymmetry of coefficient is the six and compactness is the last, There are also experts assumes that length of kernel is more determinant factor than length of kernel grove and so on. This shows there is no uniform or common guideline for the significance levels of the variables. Due to this reason domain, experts are exposed for inconsistent decision in the prediction of determinant factors for seed classification.

Research's conducted on application of machine learning in the area of agriculture are reviewed to justify the role of machine learning in solving agricultural seed class classification problems and to check the problems identified to be addressed in this research but not addressed by other researchers.

As the result, all these problems have prompted the researcher to develop predictive model to predict seed classes using machine learning tools which assists both research experts and agricultural analysts in analyzing and managing crops production rate to make quality decision on daily needs of farmers. So this study focused on developing a model based on ANN for the prediction of seed classes in order to avoid subjective approach.

1.4 Objectives of Research

1.4.1 General Objective

The general objective of the study is to develop a model which predicts the seed classes based on artificial neural network.

1.4.2 Specific Objectives

To achieve the general objective of the study the following specific objective of the study are identified:

- ▶ To review literatures on concept of ANN, big data analytics and framework for big data, that give deeper understanding to conduct this research work.
- ▶ To develop the appropriate neural network framework including number of hidden layers, number of hidden nodes, number of output nodes, transfer function, divide function and training function.
- ▶ To explore the major parameters that impact seed class prediction.
- ▶ To evaluate the performance of the developed model.
- ▶ To draw conclusions based on the findings and possible application of the developed framework and forward appropriate recommendations as future research directions in the area of machine learning.

1.5 Research Question

Based on the specific objectives indicated above the following questions were posed as a basis for this research are:-

- ▶ **RQ1.** What is the most effective ANN architecture for determining seed classification?
- ▶ **RQ3.** What are the most determinant predictors to seed classification?
- ▶ **RQ2.** How should predictor variables be represented to accurate predictive patterns for seed classes?

1.6 Scope and limitation of Study

The scope of this study is to develop a framework for the prediction of determinant factors of seeds based of an artificial neural network by using available training algorithm, transfer function, and divide function.

This research focuses on development of ANN model and to predict determinant factors of seed classes respectively. The reason of developing ANN model are the largest issue in terms of their importance to increase size of seed production in the country. These factors are highly affect the amount of crop production which lead to problem of food security.

Moreover, this study is intended to implement a model that predicts seed classes then correctly predicted seed classes can addresses the problem increasing crop productivity in the country.

The limitation of this research includes:

- ▶ Unorganized dataset.
 - The researcher tried to solve the problem of lack of data in organized forms. The files are available on different sheets, collecting the data from different sheet is time taking tasks.
- ▶ The developed framework only identifies determinant factors of seed classes; it doesn't determine plant heredities, chemical inputs, fertilizer, and tt does not predict soil type, moisture, humidity, and, climate changes.

1.7 Significance of Study

The proposed research aims at developing a computational model for the generation of algorithmic components and workflows, which support the generation of new processing blocks with emerging functionalities. The findings of this study will benefit agricultural researchers or experts in such a way that experts and researchers will have machine learning to predict seed classes. With the development of this model, the system can analysis the agricultural data for better prediction of suitable environment which increases production of crops in agricultural sectors through provision of adequate knowledge from stored attribute.

Therefore, those development agents who assist farmers can consult and access the system, analysis different geometric parameters like area, kernel groove and compactness etc. to make appropriate decisions on farmer's needs. This classification and prediction can be trained and consulted from by considering all elements of the geometric perimeters which can be implemented at different level of the zones.

Generally, the study has the following significances:

- ▶ Since the study is master thesis it has also learning out comes for the researcher. It helps the researcher to examine problems and solving it in scientific manner.
- ▶ The Analysis of this research benefits the farmers to increase their products through indirect means of selecting suitable type of seed.
- ▶ Improves the agricultural product development processes.
- ▶ Attract agricultural research centers to actively participate in use of machine learning tools.

The use of technology enabling farmers to head massive amounts of data collected through sensors to predict the best time to plant, what type of seed to use, and where to plant. We aim to help agronomist to have scientifically experimented so they can make better decision for the farmers. So if the farmers have enough knowledge regarding to selecting appropriate seed for their agricultural activities, they can maximize production and profits.

1.8 Methodology of the Study

1.8.1 Research Framework

In the information systems research framework[9] people, business organizations, and their existing or planned technologies are the components of the environment. In Melkassa Research Center, seeds agronomist and the available technology at the research center represent the environment for this study. Organizational strategies, structure, culture, and existing business processes [10] are important to evaluate and assess the business need or the problem (subjective determination of seed classes). Business needs are arranged or placed relative to the currently available technology infrastructure, applications, communication architectures, and development capabilities. Together these describe or explain the business need or “problem” as understood by the researcher. The importance or relevance of the research can be assured by framing research activities to address business needs. The design science addresses research through the building and evaluation of artifacts designed to satisfy the identified business need. The knowledge base provides the raw materials from and through which IS research is accomplished. Foundations and methodologies are components of the knowledge base. During the develop/build phase, models and methods are used for the foundations. The experimental approach is used in justify/evaluate phase to assess the performance of the model or the artifact and refine to improve the performance of the model by changing different parameters such as training function, transfer function, divide function and hidden layer neurons. Applying appropriate available foundations such models, methods and methodologies as data analysis techniques enable the researcher to achieve accuracy for this study. Both the environment and the knowledge base is the aim or objective of design science research. Figure 1 below shows the information system research framework.

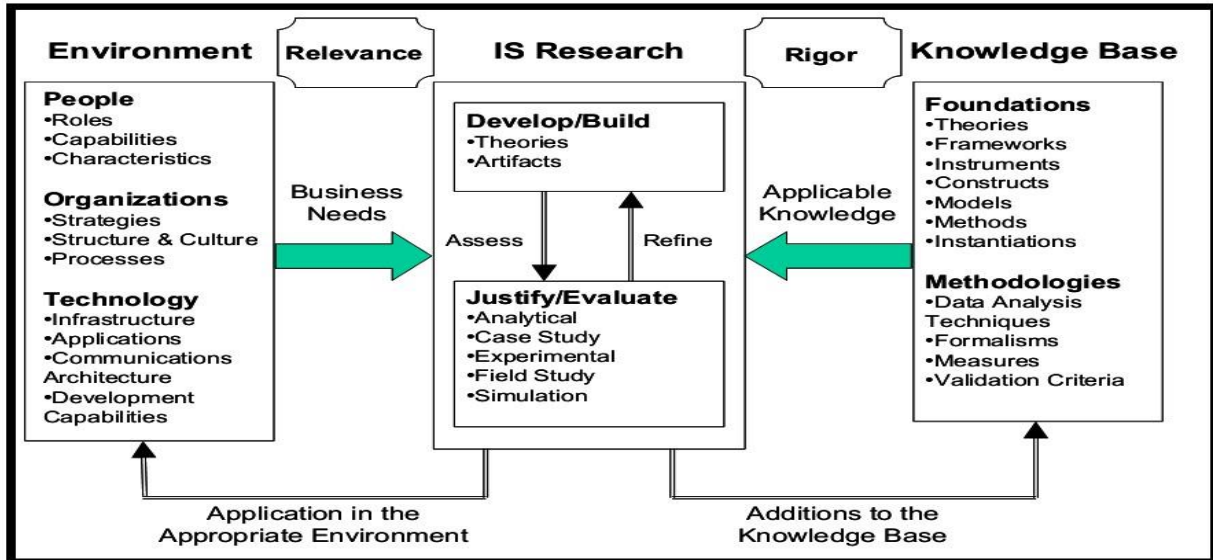


Figure 1: Information Systems Research Framework [11]

1.8.2 Design Science Research Methodology (DSRM) Process Model

Conceptual process and mental model might help the researcher[12] to produce and present high-quality design science research in IS that is accepted as valuable, rigorous, and publishable in the IS research literature. A process model consisting of activities in a nominal sequence. Each of the activities is described below:

1. Problem identification and motivation: - subjective determination of factors for seeds is the problem at Melkassa Research Center and this motivates the researcher to develop a model for the seed classes.
2. Objectives of a solution: - the model with appropriate framework is developed in order to identify determinant factors for seed.
3. Design and development: - different experiments are performed by changing number of hidden neurons starting from 4 up to 15, divide function such as dividerand, divideint, divideind and divideblock, transfer function such as tansig, logsig, purelin, and transfer function such as trainlm, trainscg, trainrp, traincgb, traincgp, traincgf, trainoss.
4. Demonstration: - the efficiency of the framework that is developed for seed classification is assured using experimentation.
5. Evaluation: - the framework is evaluated by using MSE and confusion matrix.

Figure 2 below shows design science research methodology (DSRM) process model which contains the activities described above.

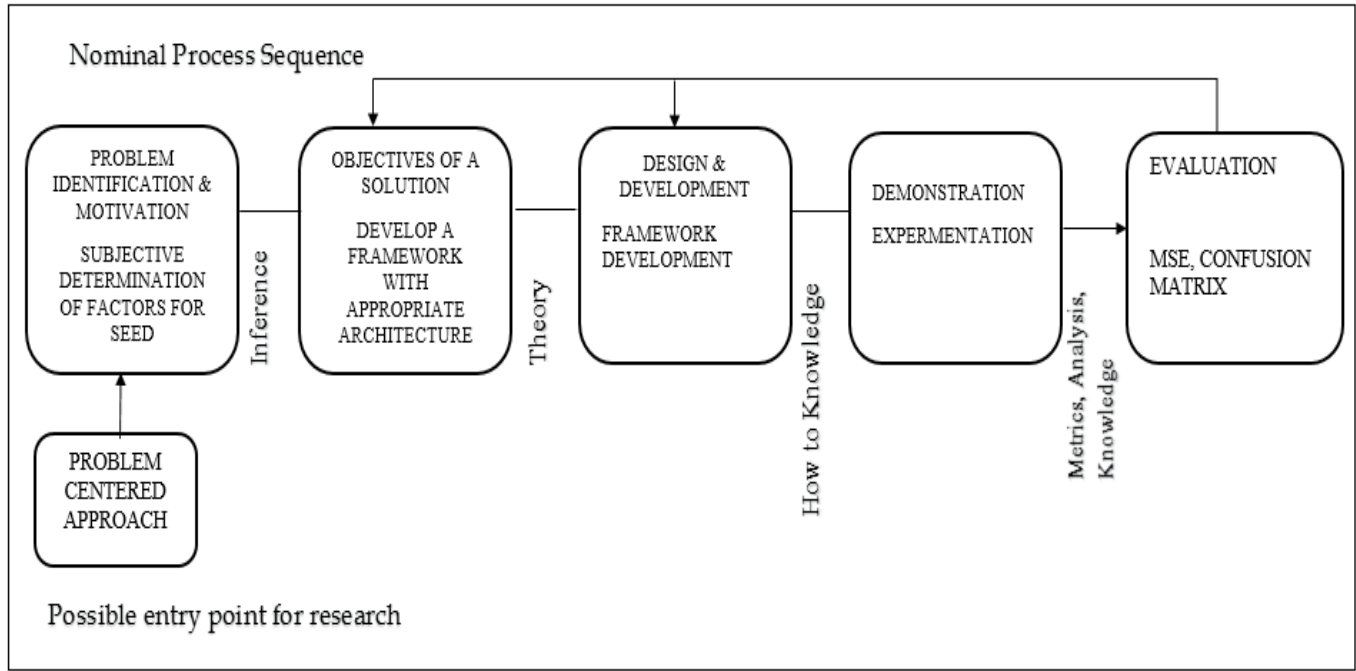


Figure 2: Design Science Research Methodology (DSRM) Process Model [12]

Implementation Tools

MATLAB is used for the implementation of this study. The main reason is MATLAB[13] is flexible, it has many inbuilt functions and suitable ANN simulation.

Approach to be used

ANN is used as an approach for this study. The main reason is according to [14][15], ANN is more efficient and accurate than SVM and other approaches.

Contribution of the Study

This work has presented a solution for the identification of significance levels of variables for seeds. The main contribution of this work include:

- ▶ Developing machine learning model to predict seed classes.

Application Results

This research was conducted for the following applications: It will:

- ▶ Enable agricultural researcher to predict adequate seed types.
- ▶ Control the problem of scarcity of agricultural researchers. Once the model is developed, even junior staffs can use it.
- ▶ Contribute towards the successful achievement of one of the programs of the government i.e. food security by increasing grain yield production per unit area through the appropriate use of inputs.
- ▶ Help researchers as a facilitator for further study.
- ▶ Enable farmers to get improved services and their profitability will increase.

1.9 Brief Survey of Thesis Chapters

This thesis contains six chapters. Chapter one has a primary focus on the background of the study, statement of the problem, objective of the study, scope, and limitations of the study, methodology of the study, the contribution of the study, and the outline of the thesis. In chapter two, we look at the Related Work in the field of machine learning and basics of an artificial neural networks, types of topologies in ANN, types of artificial neural networks, the working principles of artificial neural network and backpropagation algorithm, and applications of an artificial neural network in order to understand the subject matter of this thesis work and finally We shall provide an exhaustive coverage of prominent work conducted in these areas in the past decade, In Chapter three we look at Methodology like data source, a methodology for implementation, data preparation techniques, process models which are important to show the nature of the methodology and the predicted model platform.

Chapter four discusses model development, ANN architectures are determined, Number of input nodes, ANN training, transfer function and stopping criteria are identified. In chapter five we perform an experimental, results and discussion of study that deals with the design of class based ANN model is presented while conclusion, recommendation, and future works are forwarded in chapter six.

CHAPTER Two

LITERATURE REVIEW

In this chapter, overview of predictive data mining, machine learning algorithms, predictive modeling, basics of artificial neural networks, types of topologies in Artificial Neural Network, Types of Artificial Neural Networks, Working Principles of ANN and Back Propagation Algorithm, and related works are reviewed.

2.1 Overview of Predictive Data Mining

[16] On their part say that predictive data mining involves using some variables or fields in the database to predict unknown or future values of other variables of interest.

Similarly, [17] states that predictive methods aim to describe one or more of the variables in relation to all the others; they are also called asymmetrical, supervised or direct methods. This is done by looking for rules of classification or prediction based on the data. These rules help us to predict or classify the future result of one or more response or target variables in relation to what happens to the explanatory or input variables. The main methods of this type are those developed in the field of machine learning such as the neural networks (multilayer perceptions) and decision trees but also classic statistical models such as linear and logistic regression models.

2.2 Machine Learning Algorithms

Currently, non-cloud machine learning libraries lack scalability, not well verified and not built with production-quality, and they also lack open source community support. Therefore, Mahout has been created with scalable machine-learning algorithms all written using Map-Reduce paradigm, including 1) supervised learning algorithms such as neural networks, support vector machines, Naive Bayesian classifiers, decision trees, random forests, and logistic regression; 2) unsupervised learning algorithms including k-means, hierarchical clustering, self-organizing maps, fuzzy k-means, etc. Hundreds of ML algorithms have been tested by Mahout Libraries, Best ML algorithms are automatically selected by Algorithm Selection Optimizer to address specific needs of users.

2.3 Predictive Modeling

While we build our model we used predictive models method since it is an iterative process in which a model is created from an initial hypothesis and then refined until it produces a valuable business outcome or discarded in favor of another model with more potential [18]. The other advantage of using Predictive analytics is, it is also useful to incorporate the model for more analytical insights. Tasks we have performed for predictive model development is depicted on figure 3 below.

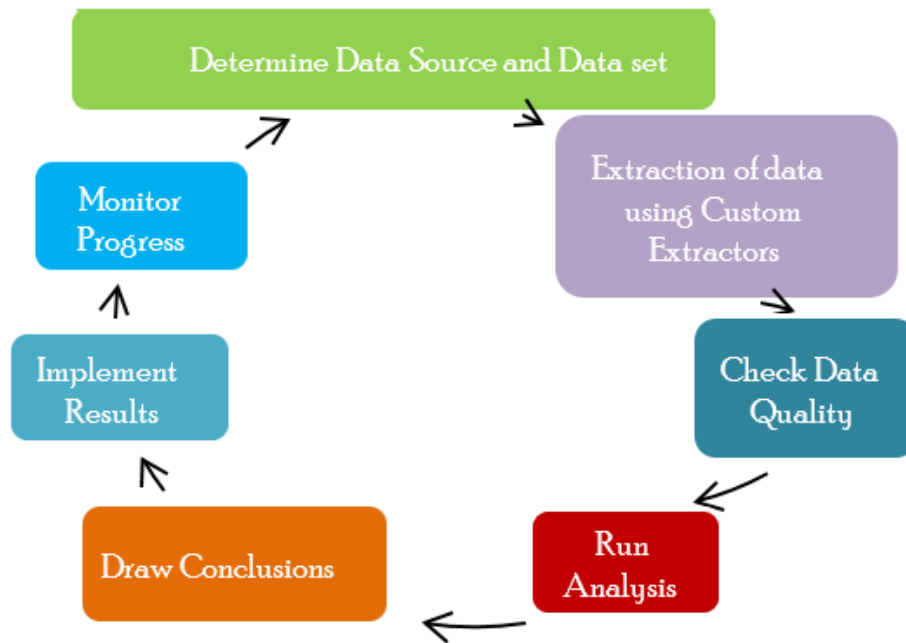


Figure 3: Predictive Analytics Process [18]

2.4 Basics of Artificial Neural Networks

The field of Artificial Neural Networks (ANNs) or Neurocomputing or Connectionists theory (CT) [4] was born out of a research done by McCulloch and Pitt in 1943 on a simple model of a neuron. Following this were further attempts to generate knowledge from the study of the central nervous system of a human being and application of such knowledge to the design and implementation of systems. Artificial neural networks [19] are networks based on mathematical calculations which try to imitate the working principles of the networks that are found in the nerve cell of the central nervous system of an animal or human being during the design and implementation of artificial systems. This imitation is a gross cell-by-cell (neuron-by-neuron, element-by-element) imitation. A neuron transmits input [4][19] to the cell body or soma of another neuron through the synaptic connections to the dendrites. The output signal or information (consisting of nerve impulses or electrical triggers) from a cell is transmitted through the connection wire (axon) to the synapses of other neurons.

2.5 Types of Topologies in Artificial Neural Network

Connecting two or more artificial neurons[20] getting an artificial neural network is useful to solve real life problems because single artificial neuron has almost no usefulness in solving real-life problems. Artificial neurons are the basic building blocks of an artificial neural network which are useful to solve complex real-life problems in a non-linear, distributed, parallel and local way. The way that individual artificial neurons are interconnected is called topology, architecture or graph of an artificial neural network. The architecture [21] is inspired by the working principle of the cerebral cortex of the brain. The two most widely used architectures of ANN are described below:

2.5.1 Feed-Forward Artificial Neural Networks

Artificial neural network with feed-forward topology[20] is called Feed-Forward artificial neural network and as such has only one condition: signal must flow from input to output in only one direction. However back loop is not allowed. Any number of layers, type of transfer function used in individual artificial neuron or number of connections between individual artificial neurons. The simplest feed-forward artificial neural network is a single perceptron that is only capable of learning linear separable problems. However simple multi-layer feed-forward artificial neural network for purpose of analytical description (sets of equations (3), (4) and (5)) is shown in Figure 4 below.

$$\left. \begin{aligned} n_1 &= F_1 (w_1 x_1 + b_1) \\ n_2 &= F_2 (w_2 x_2 + b_2) \\ n_3 &= F_2 (w_2 x_2 + b_2) \\ n_4 &= F_3 (w_3 x_3 + b_3) \end{aligned} \right\} \dots\dots\dots (1)$$

$$\left. \begin{aligned} m_1 &= F_4 (q_1 n_1 + q_2 n_2 + b_4) \\ m_2 &= F_5 (q_3 n_3 + q_4 n_4 + b_5) \\ y &= F_6 (r_1 m_1 + r_2 m_2 + b_6) \end{aligned} \right\} \dots\dots\dots (2)$$

$$y = F_6 r_1 (F_4 [q_1 F_1 (w_1 x_1 + b_1) + q_2 F_2 (w_2 x_2 + b_2) + b_4]) + r_2 (F_5 [q_3 F_2 (w_2 x_2 + b_2) + q_4 F_3 (w_3 x_3 + b_3) + b_5]) + b_6 \dots\dots\dots (3)$$

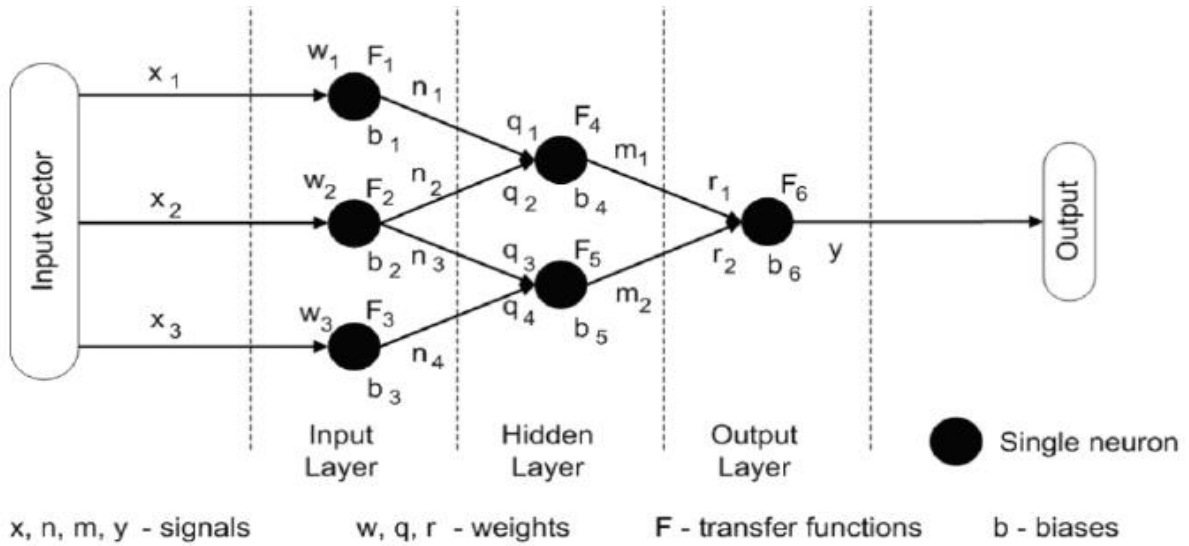


Figure 4: Feed-forward Artificial neural network [20]

As we can see in Figure 4 above and corresponding analytical description with sets of equations (1), (2) and (3) the simple feed-forward artificial neural network can lead to relatively long mathematical descriptions where artificial neural networks' parameters optimization problem solving by hand is impractical. Although analytical description can be used on any complex artificial neural network in practice we use computers and specialized software that can help us build, mathematically describe and optimize any type of artificial neural network. In addition to this [21] states that in feed forward ANN the output of any layer does not affect that same or preceding layer.

2.5.2 Recurrent Artificial Neural Networks

Artificial neural network with the recurrent topology[22], is called recurrent artificial neural network. It is similar to a feed-forward neural network, however, there is no limitations regarding back loops as well as the number of layers, type of transfer function used in individual artificial neuron or number of connections between individual artificial neurons. In these cases signal is no longer transmitted only in one direction but it is also transmitted backward. This creates an internal state of the network which allows it to exhibit dynamic temporal behavior. Recurrent artificial neural networks can use their internal memory to process any sequence of inputs. Figure 5 below shows small Fully Recurrent artificial neural network and complexity of its artificial neuron interconnections. The most basic topology of the recurrent artificial neural network is a fully recurrent artificial network where every basic building block (artificial neuron) is directly connected to every other basic building block in all direction.

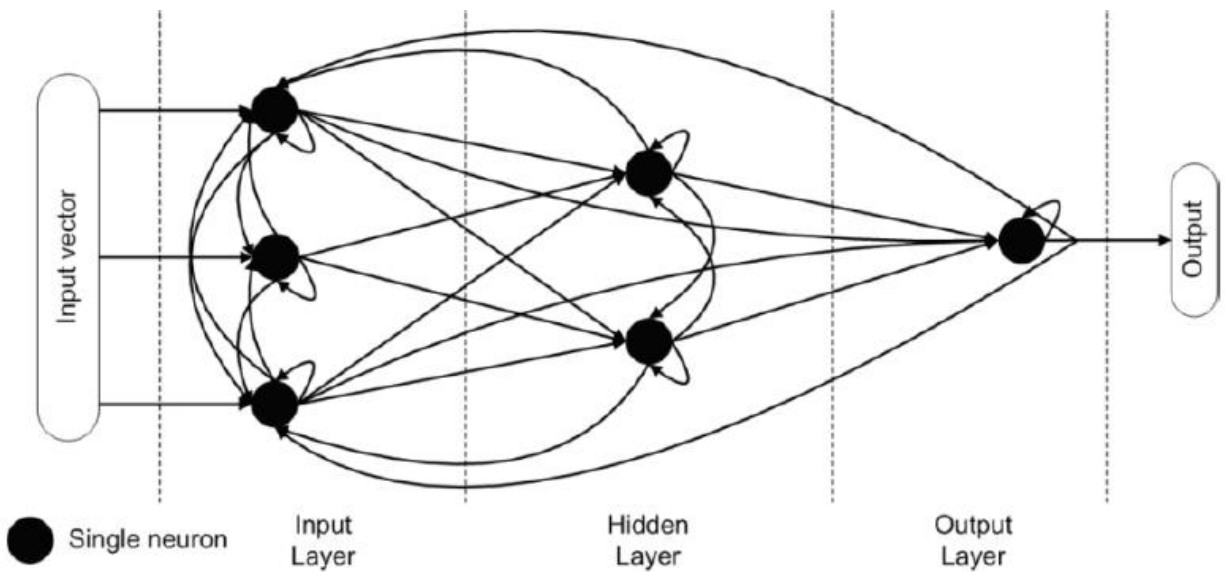


Figure 5: Fully recurrent artificial neural network [20]

In addition to this[21] states that in recurrent ANN the output of a neuron is fed back into itself as input.

In this thesis feed forward neural network is used. The main reason is to feed forward neural networks are widely and successfully used models for classification, forecasting and problem solving[23].

2.6 Artificial Neural Networks Types

The most significant class of neural networks for solving real world problems are[21] multilayer perceptron, radial basis function, and Self-Organizing Map. Each type of artificial neural network has their own unique characteristics as described below.

2.6.1 Multilayer Perceptron

The most popular form of neural network architecture is[21] the multilayer perceptron (MLP). MLP is [23] normally composed of several layers of nodes. In the MLP the first layer is an input layer where external information is received whereas the last layer is an output layer where the problem solution is obtained. The input layer and output layer are connected by one or more intermediate layers called the hidden layers. The nodes in adjacent layers are usually fully connected by acyclic arcs from the input layer to output layer.

2.6.2 Radial Basis Function Networks

RBF networks typically have[21] only one hidden layer with any number of units. The connections exist between the input layer and the hidden layer, and between the hidden layer and the output layer. It uses radial combination functions in the hidden layer, based on the squared Euclidean distance between the input vector and the weight vector. It also uses exponential or softmax activation functions in the hidden layer, in which case the network is a Gaussian RBF network. MLPs are said to be distributed processing networks because the effect of a hidden unit can be distributed over the entire input space. However Gaussian RBF networks are said to be local-processing networks because the effect of a hidden unit is usually concentrated in a local area centered at the weight vector. Radial basis networks can need[24] more neurons than standard feed forward back propagation networks, but often they can be designed in a fraction of the time it takes to train standard feed forward networks. They work best when many training vectors are available.

2.6.3 Kohonen Neural Network

The self-organizing map also called kohonen neural network [22] is an artificial neural network that is related to feed-forward networks but it needs to be told that this type of architecture is fundamentally different in arrangement of neurons and motivation. A common arrangement of neurons is in a hexagonal or rectangular grid. The self-organizing map is different in comparison to other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space. They use unsupervised learning paradigm to produce

a low-dimensional, discrete representation of the input space of the training samples, called a map what makes them especially useful for visualizing low-dimensional views of high-dimensional data. Such networks can learn to detect regularities and correlations in their input and adapt their future responses to that input accordingly.

In this thesis, Multilayer perceptron is used. The main reason is MLPs are the simplest and most commonly used neural network architecture due to their structural flexibility, good representational capabilities and a large number of programmable algorithms[25].

2.7 Working Principles of ANN and Back Propagation Algorithm

An artificial neuron is [16] a simple mathematical model (function) which is important for building ANN. Such a model has three simple sets of rules: multiplication, summation, and activation. At the entrance of artificial neuron, the inputs are weighted what means that every input value is multiplied by the individual weight. In the middle section of an artificial neuron, there is the sum function that sums all weighted inputs and bias. At the exit of artificial neuron, the sum of previously weighted inputs and bias is passing through activation function that is also called transfer function. Figure 6 below shows the three sets of rules which are important for the operation of ANN.

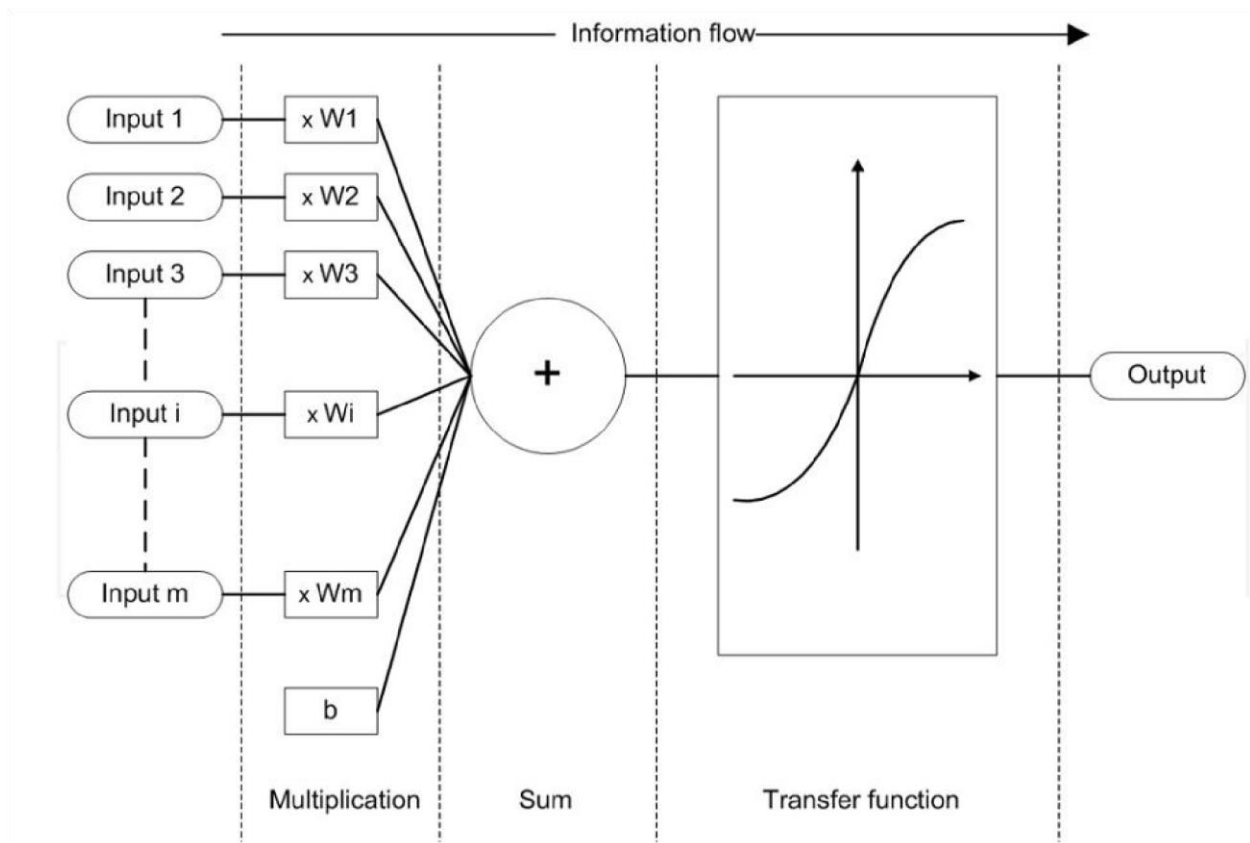


Figure 6: Working principle of an artificial neuron [20]

The artificial neural network has been trained [26] by exposing it to sets existing data where the outcome is known. Multi-layer networks use a variety of learning techniques; the most popular is back propagation algorithm. It is one of the most effective approaches to machine learning algorithm developed by David ruelhart and Robert McLelland (1994). The signal flows from the direction of the input layer towards the output layer. A network is trained rather than programmed. Learning in artificial neural networks is typically accomplished using examples. This is also called “training” in artificial neural networks because the learning is achieved by adjusting the connection weights in artificial neural networks iteratively so that trained. The number of iterations of the training algorithm and the convergence time is vary depending on the weight initialization.

The algorithm includes two passes: a forward pass and a backward pass.

- ▶ In forward pass [27] an activity pattern is applied to the input nodes and it propagates through the network layer by layer. As a result, the set of outputs is produced as the actual response of the network. The weights at the functional points of the network are fixed in the forward pass. In addition to this [28] also states that during the forward pass of the signal, according to the initial weights and activations function used, the network gives an output. That output is compared with the desired output. If both are not same, an error occurs.
- ▶ During backward pass[27] the synaptic weights are all adjusted in accordance with an error-correction rule. The actual response is subtracted from the desired output to produce an error signal. The error signal is propagated backward through the network. The synaptic weights are adjusted to have actual output nearer to the desired output. The weight adjustment is done according to the generalized delta rule to reduce the error. In addition to this [28] also states that during a reverse pass, the error is back propagated and weights of hidden and output layer are adjusted.

In this thesis back propagation algorithm is used. The main reason is back propagation algorithm is the most widely used method for determining the error[6].

2.8 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 [29], 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 [30], this paper states Neural Networks to classify varieties of rice which contain a total of 9 different rice varieties. 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 (color, 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[31], 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
- ▶ 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[32], 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 [33], 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 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 [34], 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 labeled 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.

Summary of Related works

The comparison between related papers is shown in the following table 1.

No	Authors	Related Paper	Method and Techniques	Key Findings
1.	▶ Ahmad Reza Parnian ▶ Reza Javidan	Autonomous Wheat Seed Type Classifier System,	▶ Used K-means clustering algorithm to categorize wheat seeds ▶ Matlab toolboxes like Bioinformatics, image processing, Neural Network, Data Acquisition, financial	▶ Analyze and classify wheat seeds faster with higher confidence level based on defined attributes.
2.	▶ Chathurika Sewwandi Silva ▶ Upul Sonnadara	Classification of Rice Grains Using Neural Networks	▶ Artificial neural network ▶ Matlab toolbox ▶ Sony DSC-W270 digital ▶ Two hidden layers were used to train all networks	▶ Developed different neural network models for individual feature sets and for the combined feature set ▶ Classifying local rice varieties using different feature sets. ▶ The model produced with an overall classification accuracy of 92%.

3.	▶ Jiří Šťastný, ▶ Vladimír Konečný, ▶ Oldřich Trenz	Agricultural data prediction by means of neural network	▶ Neural Network, ▶ Multi-Layer Perceptron, ▶ Back Propagation algorithm ▶ Non-linear Regression ▶ neural network simulation device	▶ Analyzed multi-layer neural network regressive model for solving the problem of the yield of onion, ▶ 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.
4.	▶ Miss.Snehal S.Dahikar, ▶ Dr.Sandeep V.Rode	Agricultural Crop Yield Prediction Using Artificial Neural Network Approach	▶ Artificial Neural networks ▶ Back Propagation algorithm	▶ They indicate that Crop Yield Prediction by using feed forward back propagation ANN is more efficient.
5.	▶ Raja Haroon Ajaz ▶ Lal Hussain	Seed Classification using Machine Learning Techniques	▶ Software used: ○ Weka ▶ Methods used for classification is: ○ Function, Bayes, Meta ,Lazy methods ▶ Classifiers used from these methods are: ○ Multilayer,Perceptro, Logistics, SMO, NaiveBayes Updateable, Naïve Bayes, Bayes Net, MultiClass Classifier, Regression , LWL	▶ The performance decreases as the number of folds decreases. ▶ Multilayer Perceptron classifier that gives the highest accuracy value 97.6% using 5 fold Cross Validation. ▶ Multilayer Perceptron gives the highest accuracy value when we use Training Set method which is 99.5% ▶ Logistics gives second highest accuracy value of 98.6%. ▶ Training Set method gives higher accuracy than Cross Validation during the classification process.
6.	▶ Seokhwan Yang ▶ Jaechun Kim	A Prediction Model Based On Big Data Analysis Using Hybrid FCM Clustering	▶ Big Data analysis using hybrid FCM clustering algorithm using R	▶ Integrates the higher accuracy rates of supervised learning and the flexibility of unsupervised learning. ▶ Improve the prediction accuracy of the system whose data set is small.

Table 1: Comparison of Related Papers

CHAPTER Three

METHODOLOGY

3.1 Research Design

This section describes overall experiments that define different research question with dependent and independent variables and apply data collection methods to do the experiment. If our data set is in the data warehouse and OLAP (Online Analytical Processing) we can work on aggregation and statistics or indexing, searching, and querying using keyword based search and pattern matching if it is knowledge discovery from dataset we can use data mining and statistical modeling. So in order to overcome the problem of data mining, using machine learning tool is the best way to analysis data. This tool has the capacity to handle large-scale computing or distributed computing than data mining tools.

In this research, machine learning tools are used to predict seed classes based on ANN. The reason why we choose machine learning tools is that, they can analysis data from a different sources like image, video, and text recognition, as well as serving as the power behind recommendation engines. Today, machine learning is being used to support crop health monitoring, face recognition systems for domestic cattle, CBR systems for fishing industry and etc.

3.2 Data Source

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, we collect the data from website: <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.

3.2.1 Data collection and Analysis

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).

In order to collect necessary research data, we used the following data collection methods:

Primary data collection

- ▶ *Interview:* interviews are used with the selected interviewer from the organization to know their attitude toward the proposed system. Therefore, we make an interview with the Development Agent, with an agronomist and with the staffs about how the seed classification is done.
- ▶ *Observation:* we conducted a strict observation in all the manual activities of the Agriculture starting from farmers, DA, and the laboratory work.

Secondary data collection

- ▶ To have a detailed knowledge and understanding of the current system, we referred previous work related to the proposed system that could help us as an input and facilitate the development of the new system. Different source books, Internet sources, Conferences, magazines etc. are analyzed and technical knowledge is reviewed. During document analysis, we gave a special consideration to those documents which can bring more features to our system.

3.2.2 Data Cleaning Tools

After collecting all data, we store the collected data and cleaned it using big data analytics tools.

For the purpose of cleaning the following tools are used:

- ▶ **Open Refine:** Using this tools we cleaned unstructured data and explore huge data sets easily and quickly even if the data is a little unstructured.
- ▶ **Data Cleaner:** This tool recognizes data manipulation as long and drawn out task we need. During data cleaning missing data and data instances that are incomplete needs to be removed from the data. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be or removed from the data entirely.

3.3 Methodology for Implementation

The methodology we used for implementation are predictive modeling. In predictive modeling, 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 7.

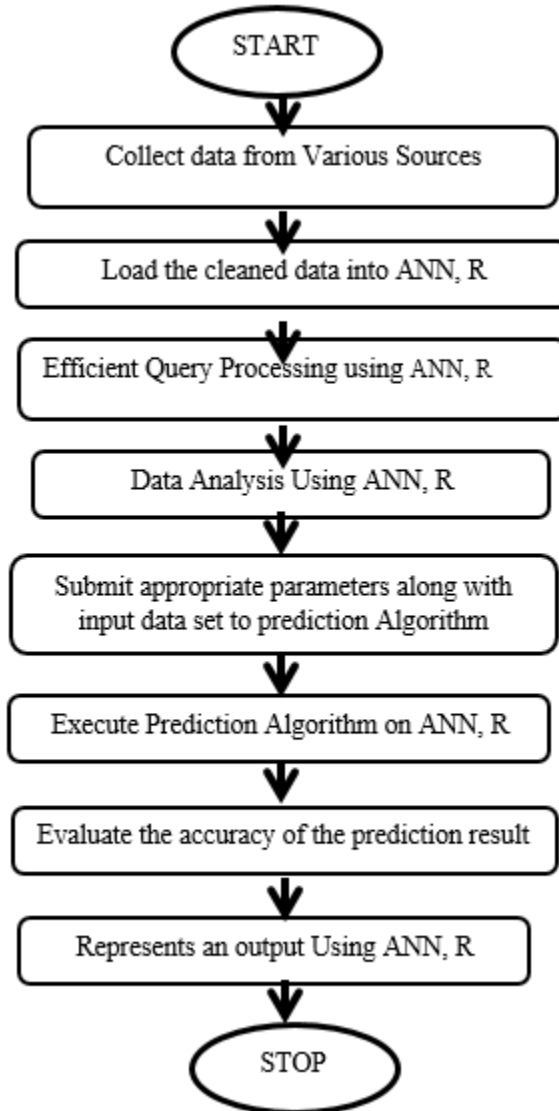


Figure 7: Work Flow of Proposed System using ANN and R [35]

The proposed system workflow is shown in above figure 7. In the first step, seed datasets are collected from various sources and these data items are further preprocessed 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.

3.4 Data Preparation

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 preprocessing (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.

3.4.1 Data Understanding

For the experimentation of the model, datasets 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 [36]. 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. The attributes with their description are shown in table 2 below.

No	Variable Name	Data Type	Descriptions
1.	Area	integer	The Surface range that is used during seed plantation
2.	Compactness	integer	$C = 4 \cdot \pi \cdot A / P^2$ or the amount of soil pressed together and reducing pore space between soil and seed
3.	Perimeter	integer	The shape of each log was combined which parameters indicated load quality
4.	Length of Kernel	integer	The distance of one seed have from the another seed
5.	width of kernel	integer	The width of one seed have from another seed
6.	Asymmetry of coefficient	integer	The proportion between seeds values that serve as measure of seeds property
7.	Length of kernel grove	integer	The distance of each seed has in the forest.
8.	Variety of wheat	character	The types of wheat chosen for the prediction

Table 2: Attributes with their data types and description

3.4.2 Data Preprocessing

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[37]. Normalize (preprocess) 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.

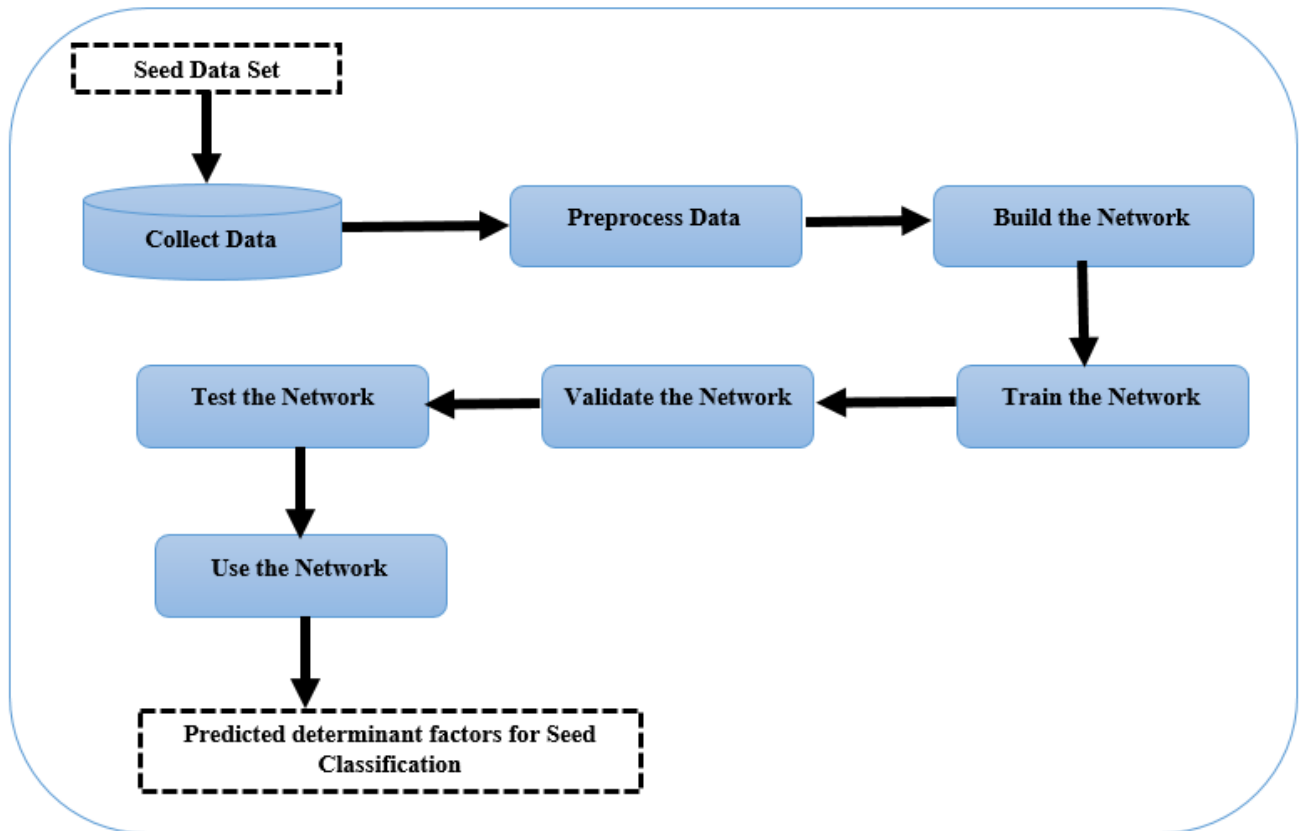


Figure 8: Proposed ANN model design flow basic steps

3.5 Process Model

Process models used for this thesis shows the nature that is used to classify together into a model. So the process model is an instantiation of it. The same process model is used repeatedly for the development of many applications and has many instantiations. The possible way of using process model is to show activities of the research work. The process model for this research is shown on figure 9 below.

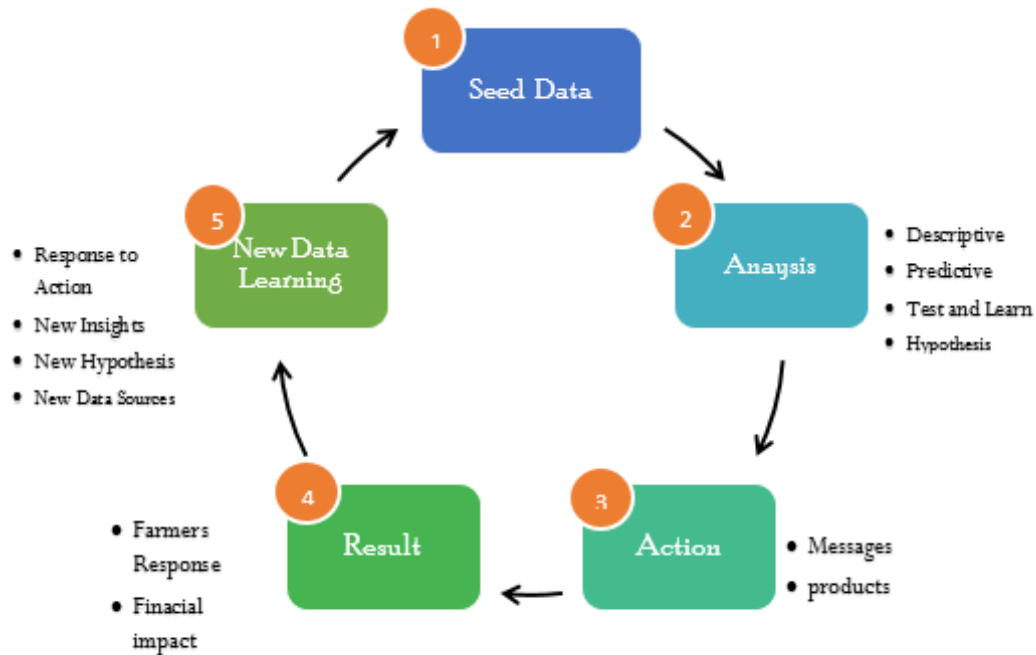


Figure 9: Process Methodology

The figure above shows the basic process of big data analysis. The main steps are collection of seed data → storage → Analysis → visualization and extraction of meaning through modeling → service action → collection of new data through feedback and update.

3.6 The Predicting Model Platform

The implementation of the system has five modules. As shown below in figure 10.

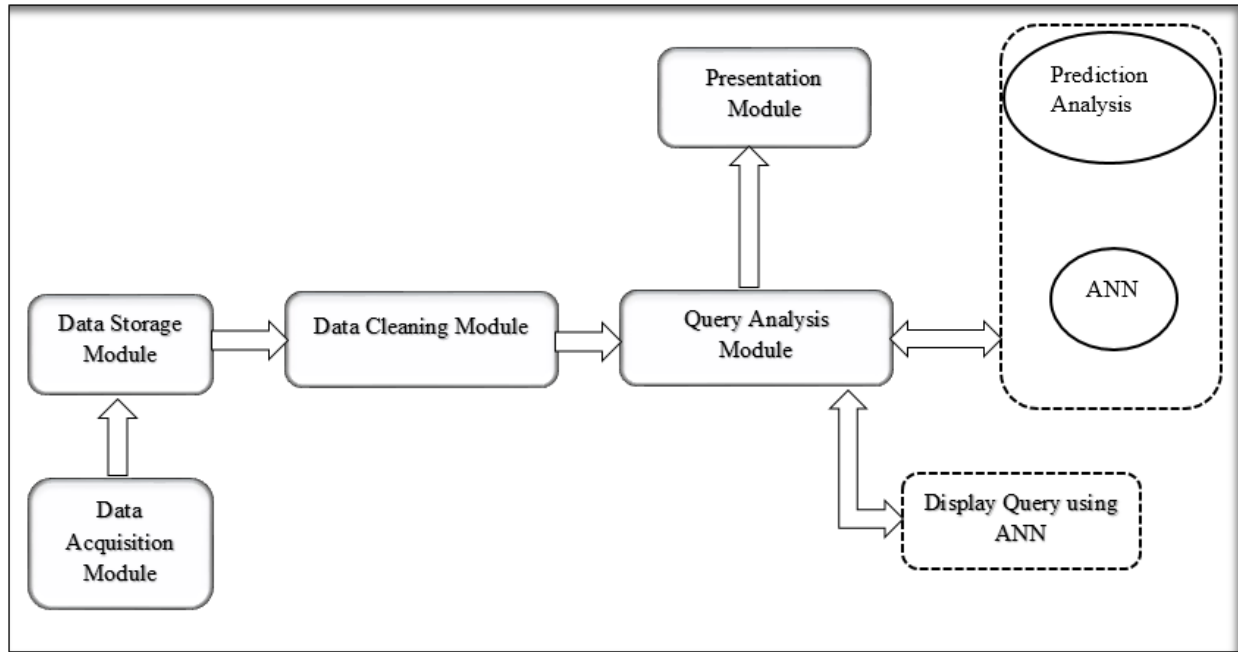


Figure 10: Modified System Framework [35]

To have a deep understanding of modified system framework, let us see each module of the system in the below section.

- ▶ Data Acquisition module.
- ▶ Data Storage module.
- ▶ Data Cleaning Module
- ▶ Query Analysis module.
- ▶ Presentation module.

A. Data Acquisition Module

This section collects data from various sources like sensor data, weather forecasting, and social media data. These data can be organized by hand, data can be acquired with access of data acquisition equipment, the received data are first stored in the form of excel database, when data are gathered to a certain number, the stored data will be transferred into the storage module and transferred data are stored in local server.

B. Data storage module

It is responsible for storage of metadata and data sets with replicated copy, which provides backup facility. HDFS is used as a storage container. Small data in the data acquisition module accumulated to a certain amount are placed in the storage module on a regular basis.

C. Data Cleaning Module

This module is used to clean unstructured data from the dataset. The cleaned data set is loaded into query analysis module for the analysis of data. Cleaning the data set can be done using Azure Machine Learning in the Machine Learning Studio.

D. Query Analysis Module

This section is processing phase includes two parts of data reading/analyzing and the establishment of prediction result. The data reading is done mainly by Hive. It was created to make it possible for analysts with strong SQL skills to run on the huge volumes of data stored in HDFS. Hive runs on workstations and changes SQL queries into a sequence of jobs for execution. In this section, we implement prediction function for establishing prediction of the data. The ANN provides an efficient way of implementing unsupervised machine learning algorithm.

E. Presentation Module

The results which are returned from the query analysis module are displayed in this module in a visualized mode. Representing complex data with charts and graphs is an essential part of the data analysis process, and we used ANN machine learning tool for prediction and have attractive data visualizations.

3.7 Data Design

The Datasets of these research contains the following two files:

► Training dataset:

The training dataset contains the data used to train the network. The Following are characteristics of this dataset.

1. Total number of rows are: 1882
2. Each row contains 7 input vector.

Dataset Evaluation: Dataset evaluation contains the data used to evaluate the machine which was trained using the training dataset earlier.

CHAPTER Four

MODEL DEVELOPMENT

4.1 Designing Artificial Neural Network models

Designing ANN models follow a number of systemic procedures. In general, there are five basics steps: collecting the data, preprocessing the data, choosing the network architecture (building the network), training the network and testing (evaluating) the performance of the network as it is shown in Fig 11.

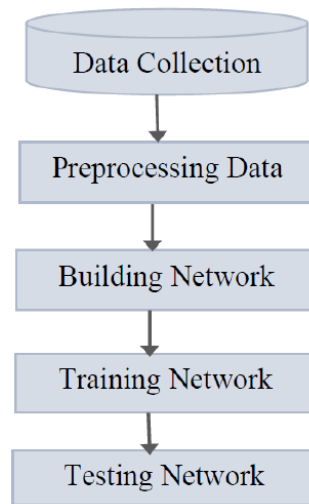


Figure 11: Basic Flow for Designing Artificial Neural Network Model [38]

The data collection phase and data preprocessing phase are discussed under chapter three and here we are trying to discuss the remaining phases.

4.1.1 Building the Network Structure

ANN is typically composed of layers of nodes. In the popular MLP, all the input nodes are in one input layer, all the output nodes are in one output layer and the hidden nodes are distributed into one or more hidden layers in between [39]. There are many ways that feed forward neural networks can be constructed. The user must decide how any neurons will be inside the input and output layers, and also decide how many hidden layers it is going to have, as well as how many neurons will be in each of these hidden layers. There are many techniques for choosing these parameters. There are some of the general rules of thumb that can be used to assist making these decisions. In nearly all cases some experimentation is required to determine the optimal structure for the feed forward neural network [40].

In the following section, the determination of ANN architecture is described.

4.1.1.1 Input Nodes

The external environment presents [40] the input pattern to the neural network through the input layer. Once the input pattern is presented to the input layer of the neural network the output layer will produce the output pattern. In essence, this is all the neural network does. The input layer should represent the condition for which the neural network is trained for. Every input neuron should represent some independent variable that has an influence on the output of the neural network. An important step in developing ANN models is [41] to select the independent variables that have the most significant impact on model performance. The number of nodes in the input layers is [42] limited by the number of model inputs Area, Perimeter, Compactness, Length of Kernel, width of the kernel, Asymmetry of coefficient and length of kernel grove are independent variables (inputs) of the model for this study.

4.1.1.2 Output Layer

The output layer of the neural network is [40] what actually presents an output of the neural network to the external environment. Whatever pattern is presented by the output layer can be directly traced back to the input layer. The number of output neurons should directly related to the type of work that the neural network is to perform. To consider the number of neurons to use in output layer one must consider the intended use of the neural network. If the neural network is to be used to classify items into groups, then it is often suitable to have one output neurons for each group that the item is to be assigned into The Kama, Rosa and Canadian are dependent variable (outputs) of the model for this study.

4.1.1.3 Hidden Layers and Nodes

The hidden layer and nodes [39] play very significant roles in many successful applications of neural networks. It is the hidden nodes in the hidden layer that allow neural networks to detect the feature, to capture the pattern in the data, and to perform the complicated nonlinear mapping between input and output variables. It is clear that without hidden nodes, simple perceptron's with linear output nodes are equivalent to linear statistical forecasting models [39]. There are really two decisions that must be made with regards to the hidden layers. The first is how many hidden layers to actually have in the neural network. Secondly, how many neurons will be in each of these layers? Neural networks with two hidden layers can represent functions with any kind of shape. There is currently no theoretical reason to use neural networks with any more than two hidden layers. Influenced by theoretical works

which show that a single hidden layer is sufficient for ANNs to approximate any complex nonlinear function with any desired accuracy [43]. Further for many practical problems, there's no reason to use any more than one hidden layer. Problems that require two hidden layers are rarely encountered [44]. Differences between the numbers of hidden layers are summarized in table 3 below:

Number of Hidden Layers	Result
none	Only capable of representing linear separable functions or decisions.
1	Can approximate arbitrarily while any functions which contain a continuous mapping from one finite space to another.
2	Represent an arbitrary decision boundary to arbitrary accuracy with rational activation functions and can approximate any smooth mapping to any accuracy.

Table 3: Comparison of hidden layers based on their task

4.1.2 Number of neurons in the hidden layer

Deciding the number of hidden neurons in layers is a very important part of deciding the overall neural network architecture. Though these layers do not directly interact with the external environment these layers have a tremendous influence on the final output. Both the number of hidden layers and number of neurons in each of these hidden layers must be considered. Using too few neurons in the hidden layers will result in something called under fitting. Underfitting occurs when there are too few neurons in the hidden layers to adequately detect the signals in a complicated data set. Using too many neurons in the hidden layers can result in several problems. First, too many neurons in the hidden layers may result in over-fitting. Overfitting occurs when the neural network has so much information processing capacity that the limited amount of information contained in the training set is not enough to train all of the neurons in the hidden layers. A second problem can occur even when there is sufficient training data. An inordinately large number of neurons in the hidden layers can increase the time it takes to train the network. The amount of training time can increase enough so that it is impossible to adequately train the neural network. Obviously, some compromise must be reached between too many and too few look neurons in the hidden layers[44].The most common way in determining the number of hidden nodes is via experiments or by trial and error [39]. But several rules of thumb have been proposed

to determining the correct number of neurons in the hidden layers. Some of them are summarized as follows:

- ▶ The number of hidden neurons should be in the range between the size of the input layer and the size of the output layer.
- ▶ The number of hidden neurons should be $\frac{2}{3}$ of the input layer size, plus the size of the output layer.
- ▶ The number of hidden neurons should be less than twice the input layer size[44].
- ▶ The number of hidden neurons should be equal to the input layer size.
- ▶ The number of hidden neurons should be half of the input layer size.
- ▶ The number of hidden neurons should be less than twice the input layer size plus one[39].

The above rules are only starting point to consider. In this study to incorporate the above rules hidden neurons starting from half the size of the input size to twice the size of the input layer plus one is varied to see the effect of each hidden neuron on the performance of the network.

4.1.2.1 Proposed model of Neural Network Architecture

Before experimenting the model the, the model have seven independent attributes on left side of the figure, hidden layer (black box) in middle, and three dependent attributes at right side of the figure. Figure 12 below shows the general architecture of the proposed artificial neural network model.

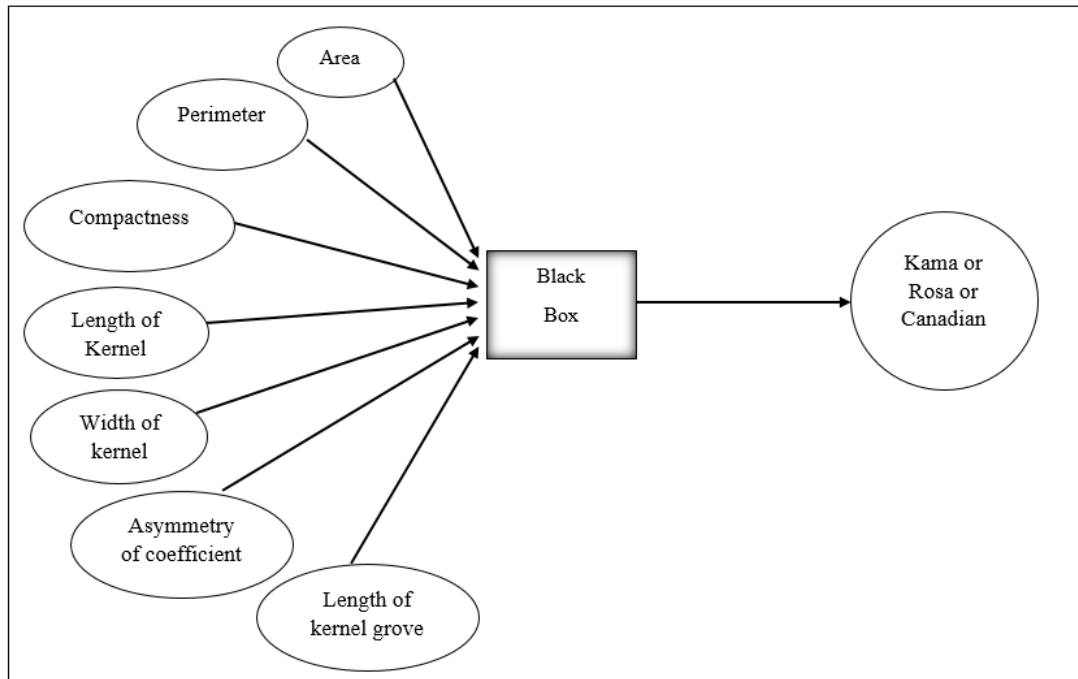


Figure 12: Neural Network Architecture of the proposed model

From above figure 12, we can observe that, the network is trained with dependent attributes until optimal MSE is obtained. Initially the hidden layer is unknown, so we call it black box hidden layer and we replace this black box with optimal hidden layer after completing the experiment.

4.1.3 Training Neural Network

The neural network training is an unconstrained nonlinear minimization problem in which arc weights of a network are iteratively modified to minimize the overall mean or total squared error between the desired and actual output values for all output nodes over all input patterns. Back propagation is the most commonly implemented training procedure for feed forward neural networks [45]. There are several different backpropagation training algorithms. They have a variety of different computation and storage requirements, and no one algorithm is best suited to all problems. The table 4 below shows the list of training algorithms included in MathLab.

Function	Description
Traingd	Basic gradient descent. Slow response, can be used in incremental mode training.
Traingdm	Gradient descent with momentum. Generally faster than traingd. Can be used in incremental mode training.
Traingdx	Adaptive learning rate. Faster training than traingd, but can only be used in batch mode training.
Trainrp	Resilient back propagation. Simple batch mode training algorithm with fast convergence and minimal storage requirements.
Traincgf	Fletcher-Reeves conjugate gradient algorithm. Has smallest storage requirements of the conjugate gradient algorithms.
Traincgp	Polak-Ribière conjugate gradient algorithm. Slightly larger storage requirements than traincgf. Faster convergence on some problems.
Traincgb	Powell-Beale conjugate gradient algorithm. Slightly larger storage requirements than traincgp. Generally faster convergence.
Trainscg	Scaled conjugate gradient algorithm. The only conjugate gradient algorithm that requires no line search. A very good general purpose training algorithm.
Trainbfg	BFGS quasi-Newton method. Requires storage of approximate Hessian matrix and has more computation in each iteration than conjugate gradient algorithms, but usually, converges in fewer iterations.
Trainoss	One step secant method. Compromise between conjugate gradient methods and quasi Newton methods.
Trainlm	Leven berg-Marquardt algorithm. Fastest training algorithm for networks of moderate size. Has memory reduction feature for use when the training set is large.
Trainbr	Bayesian regularization. Modification of the Leven berg-Marquardt training algorithm to produce networks that generalize well. Reduces the difficulty of determining the optimum network architecture.

Table 4: Training algorithms used in math lab neural network toolbox

From above back propagation training algorithms Trainlm, trainrp, trainscg, traincgb, traincgp, traincgf, trainoss training functions are used for this study. The main reason for selecting these training functions algorithm are since they are used for classification purpose [45].

4.1.4 Transfer Function

The activation or transfer function is [39] also called the transfer function. It determines the relationship between inputs (independent variables of the model) and outputs (dependent variables of the model) of a node and a network. An elementary neuron with R inputs is [24] shown below in Figure 13. Each input is weighted with an appropriate w . The sum of the weighted inputs and the bias forms the input to the transfer function f . Neurons may use any differentiable transfer function f to generate their output.

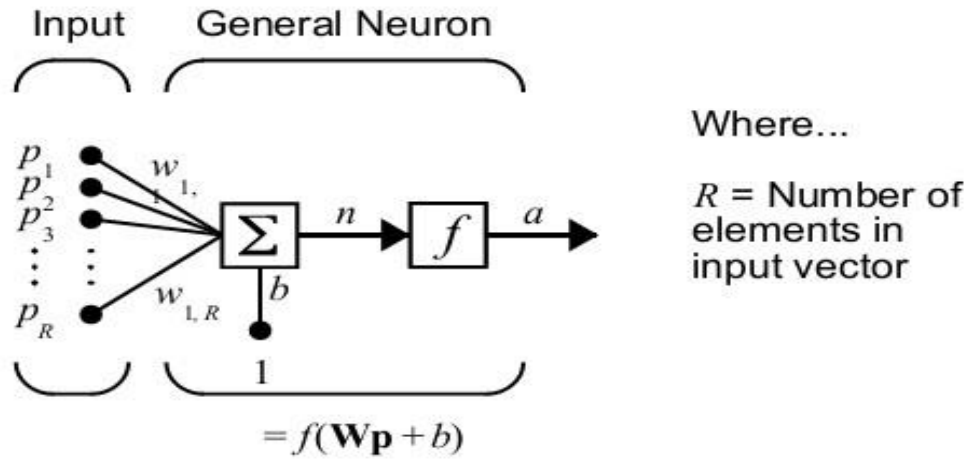


Figure 13: Figure that illustrates Activation Function[39]

For the experimentation of this research three sigmoid transfer functions are used in multilayer networks. This sigmoid transfer functions are log-sigmoid transfer function (logsig), tan-sigmoid transfer function (tansig) and linear transfer function (purelin).

Log-sigmoid transfer function

The function logsig generates outputs between 0 and 1 as the neuron's net input goes from negative to positive infinity. Figure 14 below shows logsig transfer function.

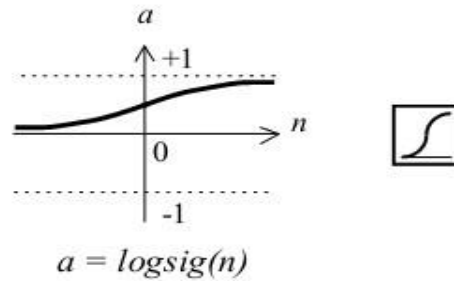


Figure 14: Log-Sigmoid Transfer Function [46]

Tan sigmoid transfer function

Multilayer networks may use the tan-sigmoid transfer function tansig alternatively. The function logsig generates outputs between -1 and 1 as the neuron's net input goes from negative to positive infinity. Figure 15 below shows tansig transfer function

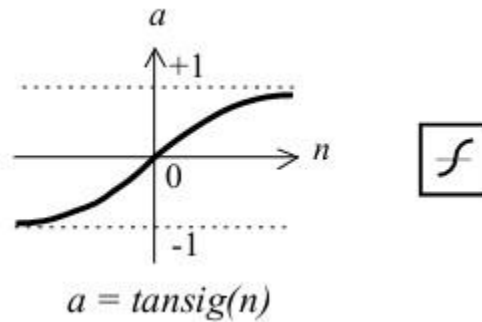


Figure 15: Tan sigmoid Transfer Function [46]

Infrequently, the linear transfer function purelin is used in back propagation networks.

Linear transfer function

Sometimes, the linear transfer function purelin is used in backpropagation networks. Figure 16 below indicates the linear transfer function.

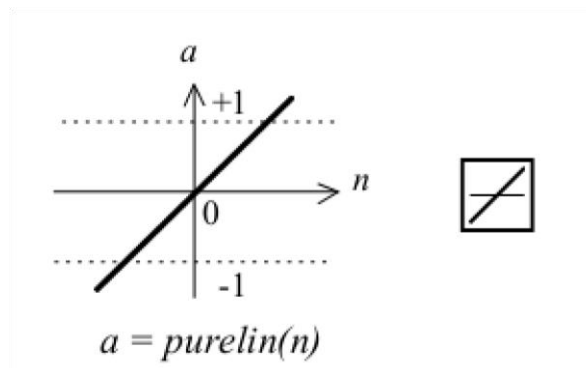


Figure 16: Linear transfer function [46]

In this study, the three sigmoid functions were used to see the performance result of the network by varying them.

4.1.5 Divide functions

There are four functions provided for dividing data into training, validation and test sets. They are `dividerand` (the default), `divideblock`, `divideint`, and `divideind`. The data division is normally performed automatically when you train the network. Table 5 below shows the divide functions.

Function	Algorithm
<code>dividerand</code>	Divide the data randomly (default)
<code>divideint</code>	Divide the data into contiguous blocks
<code>divideind</code>	Divide the data using an interleaved selection
<code>divideblock</code>	Divide the data by index

Table 5: Divide Functions[47]

4.1.6 Overfitting

One of the problems that occur during neural network training is called overfitting. The error on the training set is driven to a very small value, but when new data is presented to the network the error is large. The network has memorized the training examples, but it has not learned to generalize to new situations. Early stopping is the technique that is used to avoid this problem. In this technique of early stopping, the available data is randomly divided into three subsets. The first subset is the training set, which is used for computing the gradient and updating the network weights and biases. The second subset is the validation set. The error on the validation set is monitored during the training process. The validation error normally decreases during the initial phase of training, as does the training set error. However, when the network begins to overfit the data, the error on the validation set typically begins to rise. When the validation error increases for more than a specified number of iterations (the default value is used which is six iterations), the training is stopped, and the weights and biases at the minimum of the validation error are returned. The third subset is the test set, which is used to test the trained neural network to determine whether the neural network has generalized the training data set accurately. This subset is not used during training, but it is used to compare different models[48]. In this study, the dataset were divided into training, validation and test data sets.

4.1.7 Training Stopping Criteria

The networks are trained according to the following values of training parameters. The values of parameters that are listed below are default values in MATLAB used as stopping criteria.

Maximum number of epochs to train = 1000

Performance goal (MSE) = 0

Minimum performance gradient = 1e-5

Maximum validation failures = 6

Training stops when any of the following conditions occur:

- 1) The maximum number of epochs (repetitions) is reached.
- 2) Performance has been reduced to the goal.
- 3) The performance gradient moves below the minimum.
- 4) Validation performance has increased more than the maximum validation failures.

4.1.8 Testing the Network

The next phase in the development of ANN model is to test the performance of the proposed model. At this stage, unseen data are exposed to the model. The testing is used to evaluate the generalization ability of a supposedly trained network. In the case of seed, prediction was used for testing the ANN models. In order to evaluate the performance of the developed ANN models quantitatively and verify whether there is any underlying trend in performance of ANN models, statistical analysis involving the mean square error (MSE), the Regression (R) and the Confusion Matrix (CM) was conducted. Other error functions offered by different software's might include least absolute deviation, least fourth powers, asymmetric least squares and percentage differences.

4.1.9 Performance Evaluation

The performance of artificial neural network models can be evaluated using different standard parameters. In this study mean squared error (MSE), confusion matrix and regression analysis were considered to evaluate the performance of the model.

a. Mean Squared Error

Mean squared error is one of the most important parameters used during performance evaluation. It is used to determine how well the network output fits the desired output. The formula for MSE is shown below.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - o_i)^2 \dots\dots\dots (4)$$

Where y_i is the target and o_i is the observed output and N is the number of data set.

MSE is the average squared difference between outputs and targets. Lower values are better. Zero means no error.

b. Regression (R)

Regression values show the relationship between the outputs of the network and the targets. It measures the correlation between outputs and targets. If the training were perfect, the network outputs and the targets would be exactly equal, i.e. a regression value of 1 means a close relationship, 0 a random relationship. But the relationship is rarely perfect in practice.

c. Confusion Matrix

The confusion matrix gives the percentage of correct and incorrect classifications in the simulated feed forward back propagation networks.

The confusion matrix shows the number of cases that were correctly classified (diagonal cells) and misclassified cases (off-diagonal cells) during training, validation and testing phases.

In this study, MSE and confusion matrix are used for the performance evaluation of the model.

4.2 Critical Seed Classifier Factors Determination

$[RANKED, WEIGHT] = relief(X, Y, K)$ computes ranks and weights of attributes (predictors) for input data matrix X and response vector Y using the *Relieff* algorithm for classification or RRelieff for regression with K nearest neighbors. For classification, *relieff* uses K -nearest neighbors per class. *RANKED* are indices of columns in X ordered by attribute importance, meaning *RANKED* (1) is the index of the most important predictor. *WEIGHT* are attribute weights ranging from -1 to 1 with large positive weights assigned to important attributes. Attribute ranks and weights computed by *relieff* usually depend on K . If you set K to 1, the estimates computed by *relieff* can be unreliable for noisy data. If you set K to a value comparable with the number of observations (rows) in X , *relieff* can fail to find important attributes. You can start with $K = 10$ and investigate the stability and reliability of *relieff* ranks and weights for various values of K [49].

CHAPTER Five

EXPERIMENTAL RESULTS AND DISCUSSION

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.

Data Sources

For the experimentation of the study, the data was downloaded from the UCI machine learning repository (UCI machine learning repository:

Seeds data set <https://archive.ics.uci.edu/ml/datasets/seeds>) and loaded in to the MATHLAB and R Studio software for analysis. The download data was unstructured so it was structured using hadoop ecosystem which was configured in Data Science Special Interest Group Lab (SIG). After data preprocessing in the lab, the cleaned dataset i.e. Eight (8) attributes are used for the analysis of this experiment. The kernels of three varieties of wheat were examined under soft X-Ray technology and a number of arithmetical properties were recorded. The X- Ray technology is non-destructive and substantially cheaper than other imaging technologies and could possibly be used as a means to identify the variety of wheat.

In the following section, all experimentation are done using both big data analytics tool (ANN and R).

5.1 Network Configuration and Training

This section defines the steps deployed in determining critical RTA factors classification model using ANN backpropagation 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.

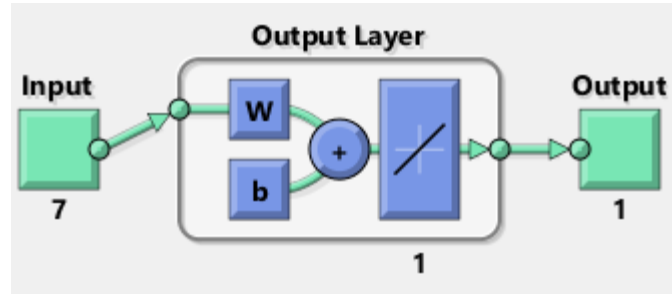


Figure 17: Un-configured neural network view

To train the network (using `[wsc,tr]=train(wsc, i,o)` command using the data sets collected so far. The network uses the default Levenberg-Marquardt Algorithm to train the network. The algorithm randomly divides input vectors and target e sets as follows:

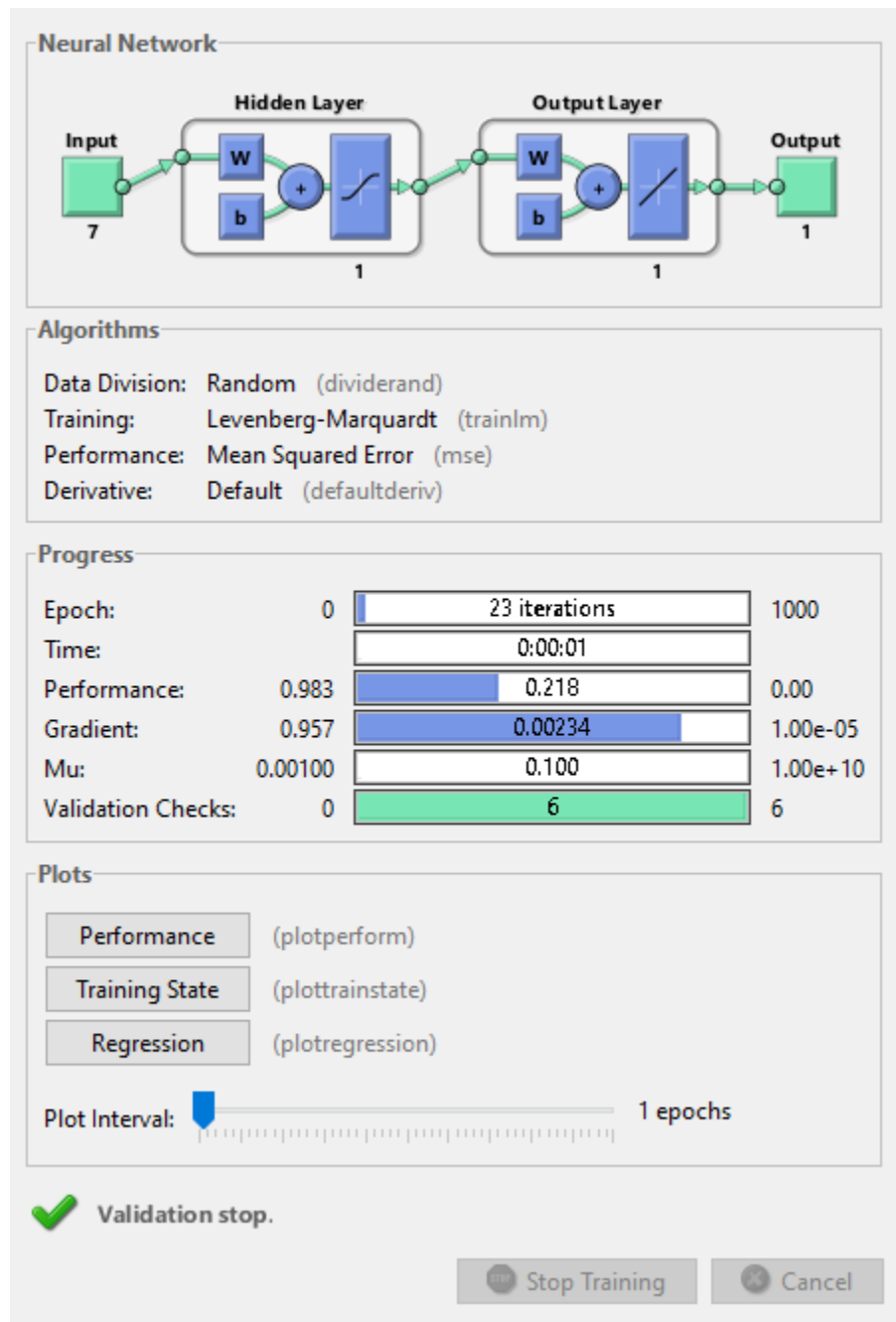


Figure 18: Neural Network training progress.

5.2 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 6 below.

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

Table 6: Hidden Neurons Comparison

From table 6 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.

5.3 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 7 below shows the performance of the four divide functions using 5, 6, 11, 12 and 15 hidden neurons.

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

Table 7: Comparison among dividerand and divideint with optimal MSE

As we can see from table 7 and graphs using above the best validation performance 0.0047947 at epoch 32 is the optimal with the divide function dividerand with 12 hidden networks. The performance of this model is depicted in figure 19 below.

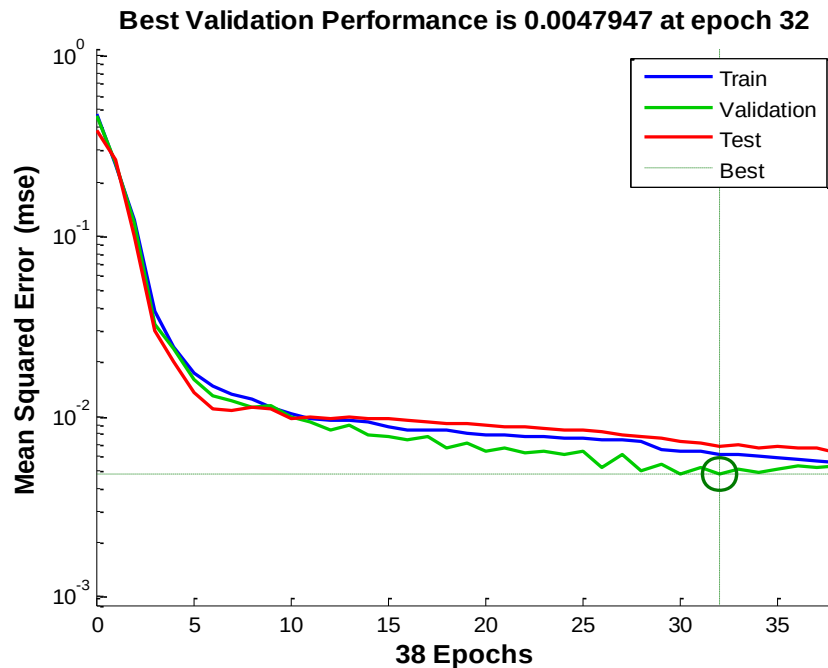


Figure 19: Best validation performance of divide functions using 12 hidden neurons

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.

5.4 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 8 below shows the comparison among transfer functions using 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

Table 8: Comparison among transfer functions with 5 hidden neurons

As we can see from table 8 above the best validation performance 0.0068835 at epoch 16 is the optimal with transfer function tansig. Figure 20 below shows the best validation performance for tansig.

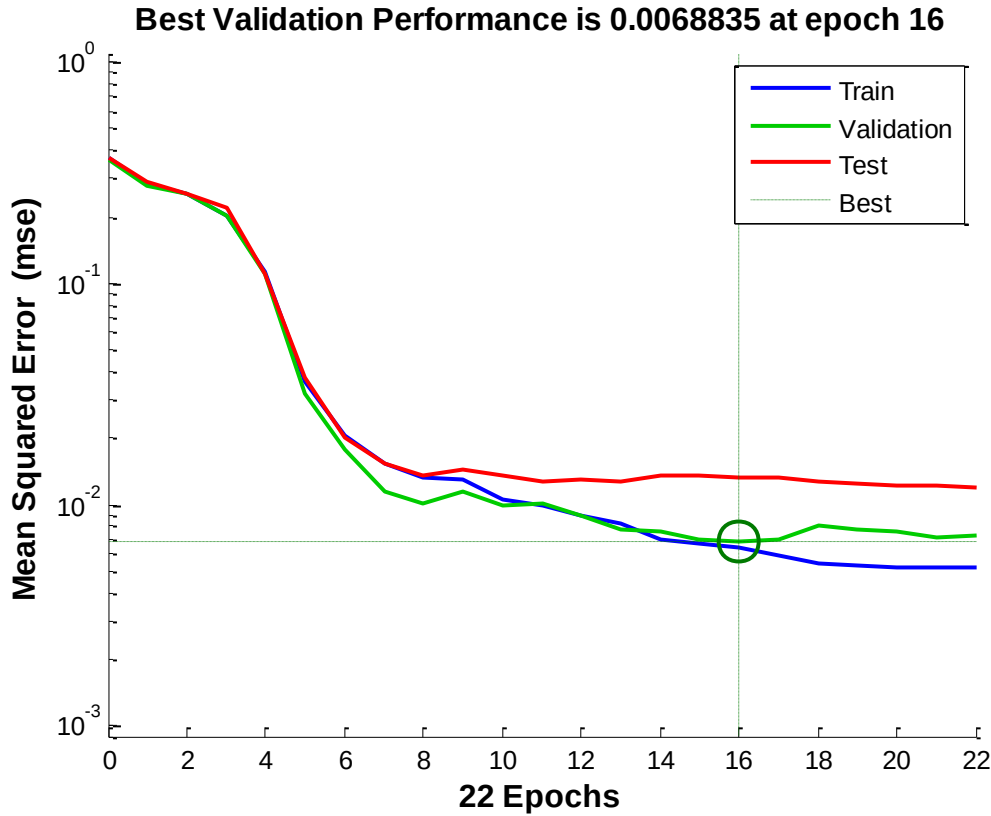


Figure 20: Best validation performance for tangsig with 5 hidden neurons

As we can see from Figure 20 above the best validation performance is 0.0068835 at iteration 16.

Table 9 below shows the comparison among transfer functions using 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

Table 9: Comparison among transfer functions with 6 hidden neurons

As we can see from table 9 above the best validation performance 0.010035 at epoch 11 is the optimal with transfer function tansig.

Figure 21 below shows the best validation performance for tansig.

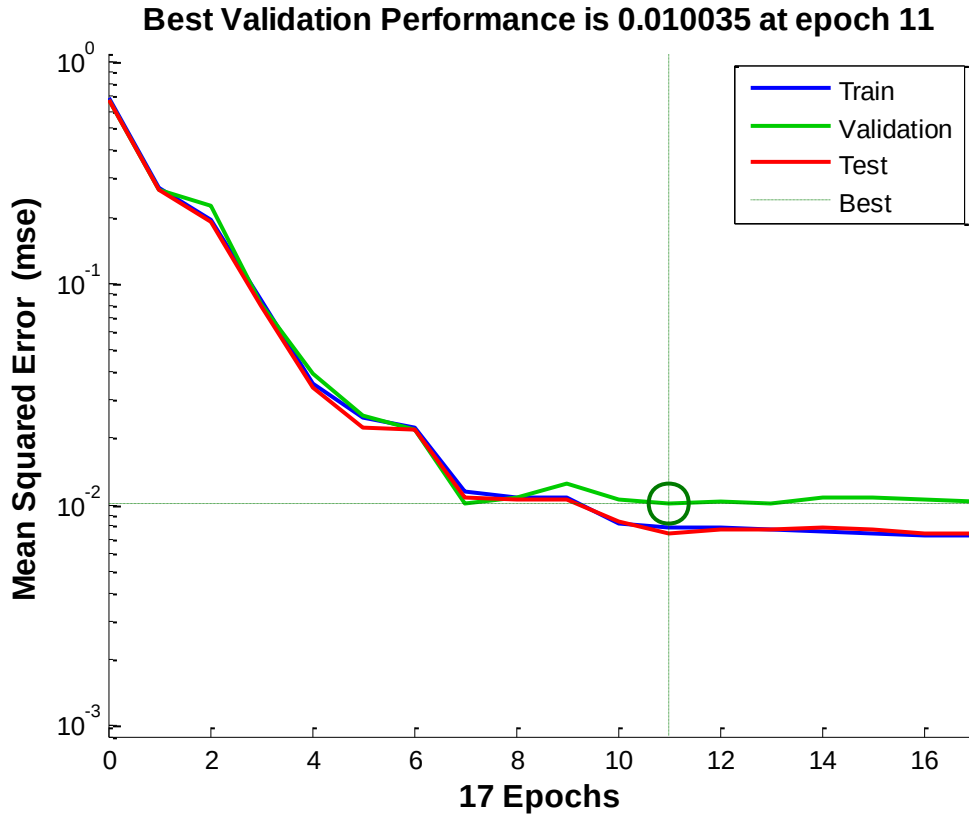


Figure 21: Best validation performance for logsig with 6 hidden neurons

As we can see from Figure 21 above the best validation performance is 0.010035 at iteration 11.

Table 10 below shows the comparison among transfer functions using 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

Table 10: Comparison among transfer functions with 11 hidden neurons

As we can see from table 10 above the best validation performance is 0.009037 at epoch 20 is the optimal with transfer function tansig. Figure 22 below shows the best validation performance for tansig.

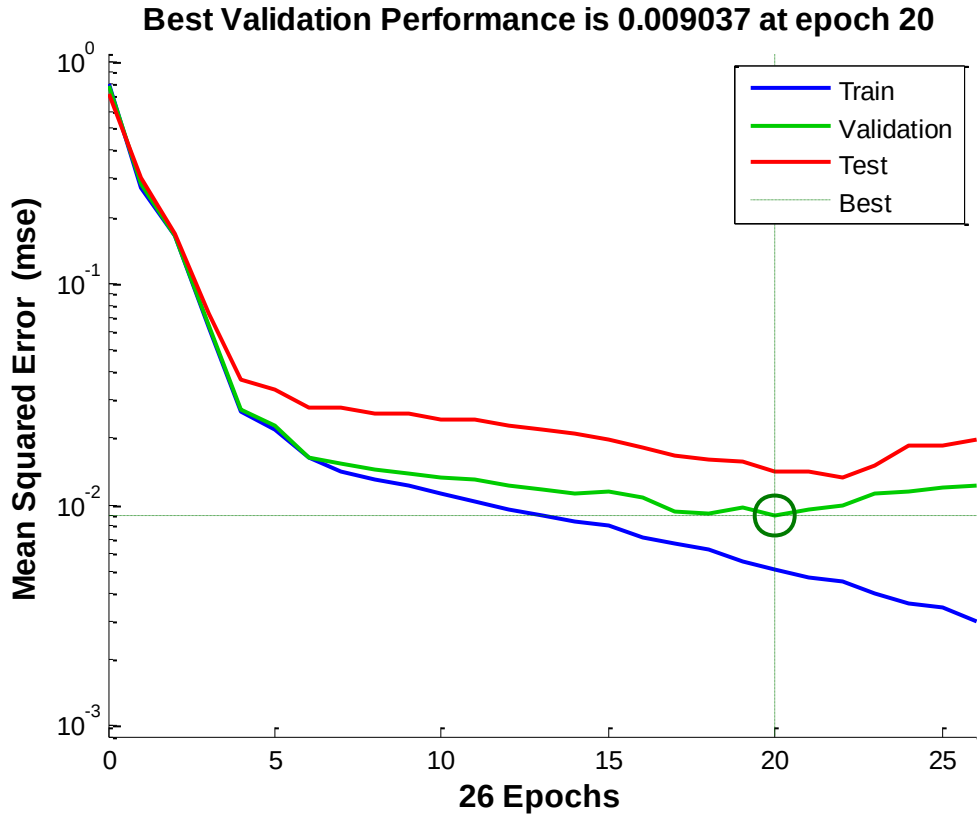


Figure 22: Best validation performance for tansig with 11 hidden neurons

As we can see from Figure 22 above the best validation performance is 0.009037 at iteration 20.

Table 11 below shows the comparison among transfer functions using 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

Table 11: Comparison among transfer functions with 12 hidden neurons

As we can see from table 11 above the best validation performance 0.0047947 at epoch 32 is the optimal with transfer function tansig. Figure 23 below shows the best validation performance for tansig.

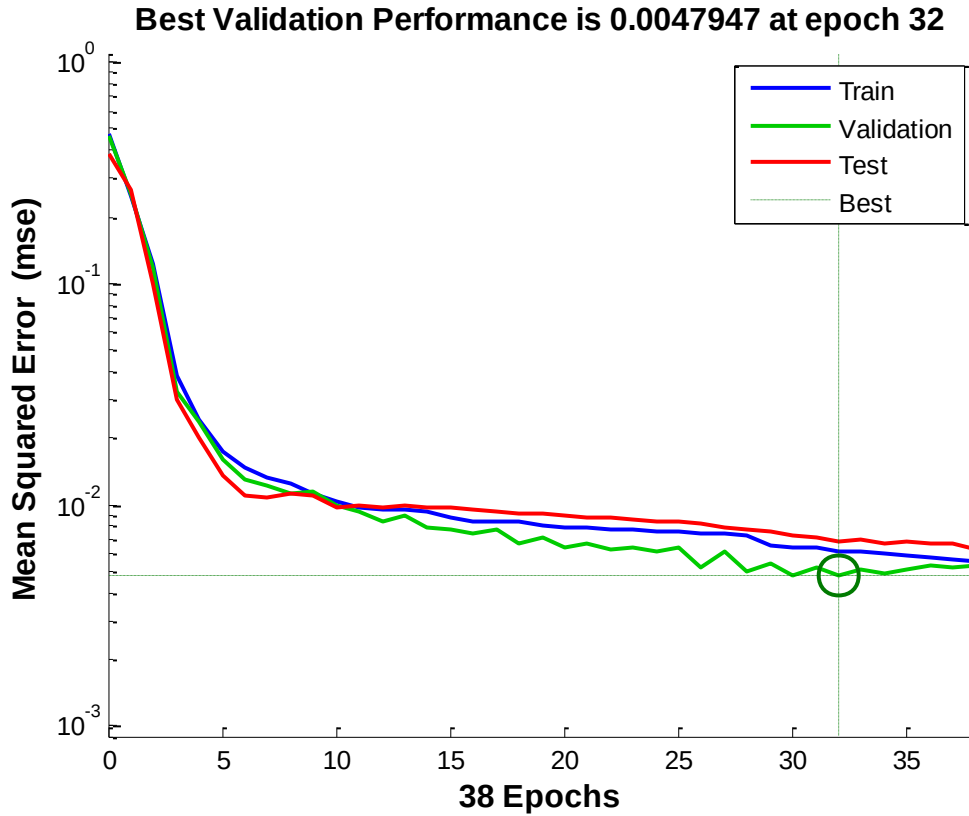


Figure 23: Best validation performance for tansig with 12 hidden neurons

As we can see from Figure 23 above the best validation performance is 0.0047947 at iteration 32.

Table 12 below shows the comparison among transfer functions using 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

Table 12: Comparison among transfer functions with 15 hidden neurons

As we can see from table 12 above the best validation performance 0.0094526 at epoch 6 is the optimal with transfer function tansig. Figure 24 below shows the best validation performance for tansig.

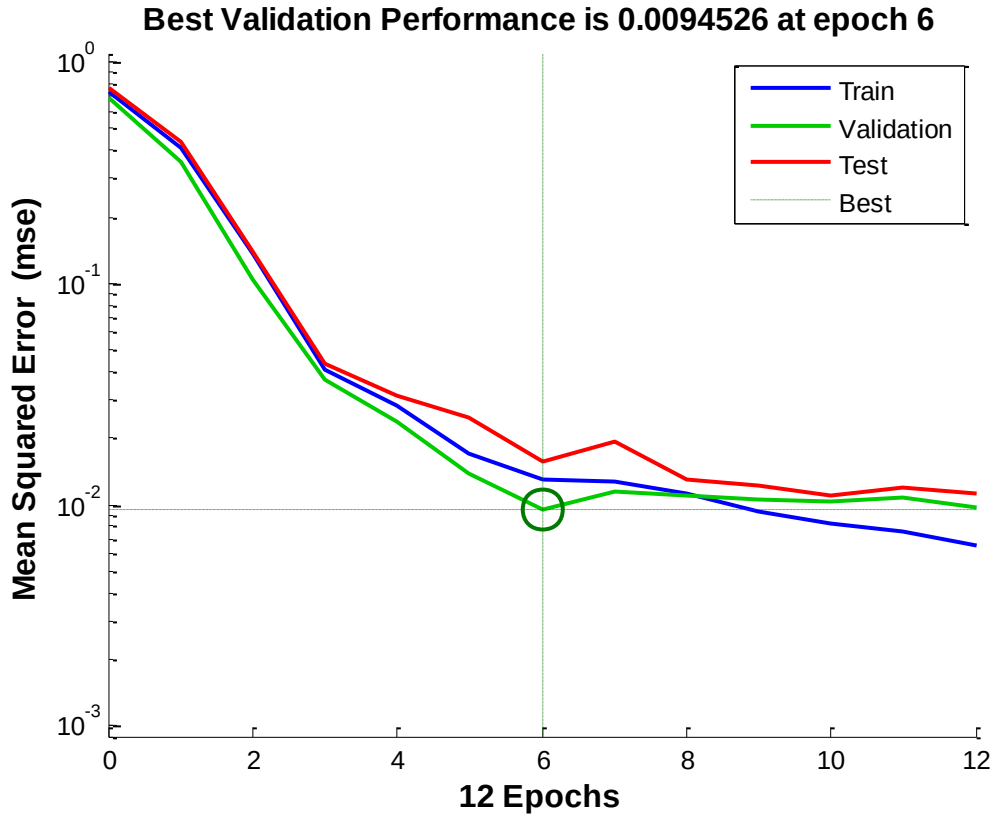


Figure 24: Best validation performance for tansig with 15 hidden neurons

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 13, we show that the result of tansig using 5, 6, 11, 12 and 15 hidden 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

Table 13: The Comparison between tansig with optimal MSE at different number of Neurons

From table 13 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 25 below.

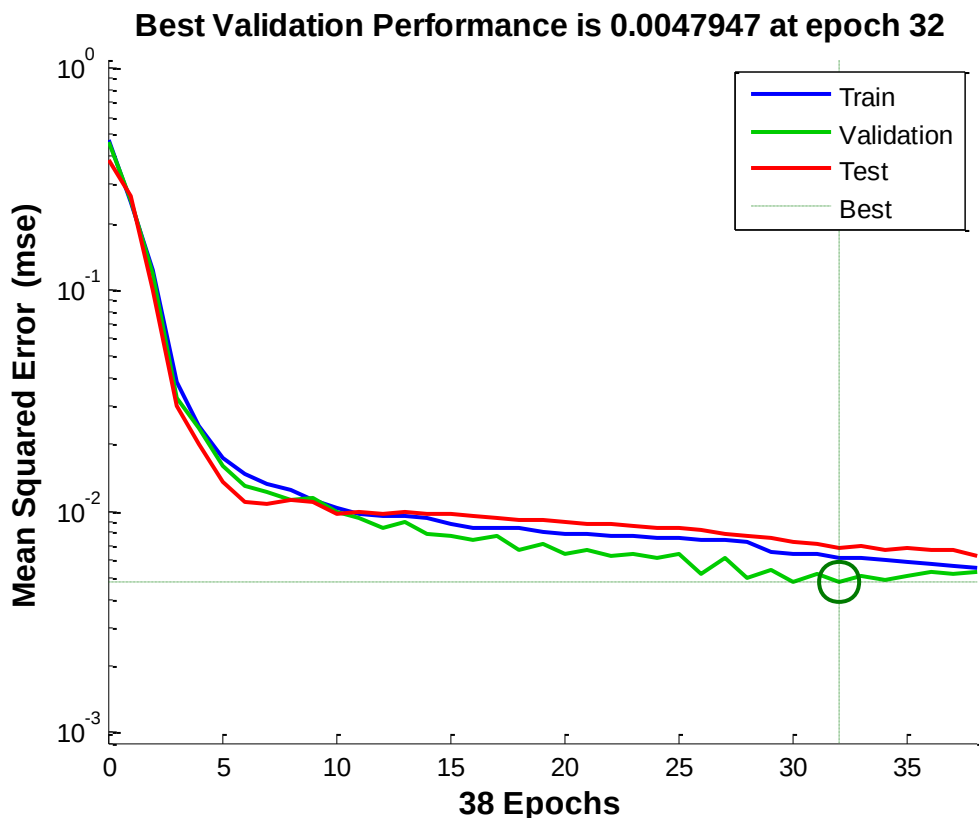


Figure 25: Best validation performance of transfer functions using 12 hidden neurons

5.5 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 14 below shows the performance comparison of the best results of each training function.

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

Table 14: Comparison of the best results of each training functions

As we can observe from table 14 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 26 below shows the performance result of the optimal training function.

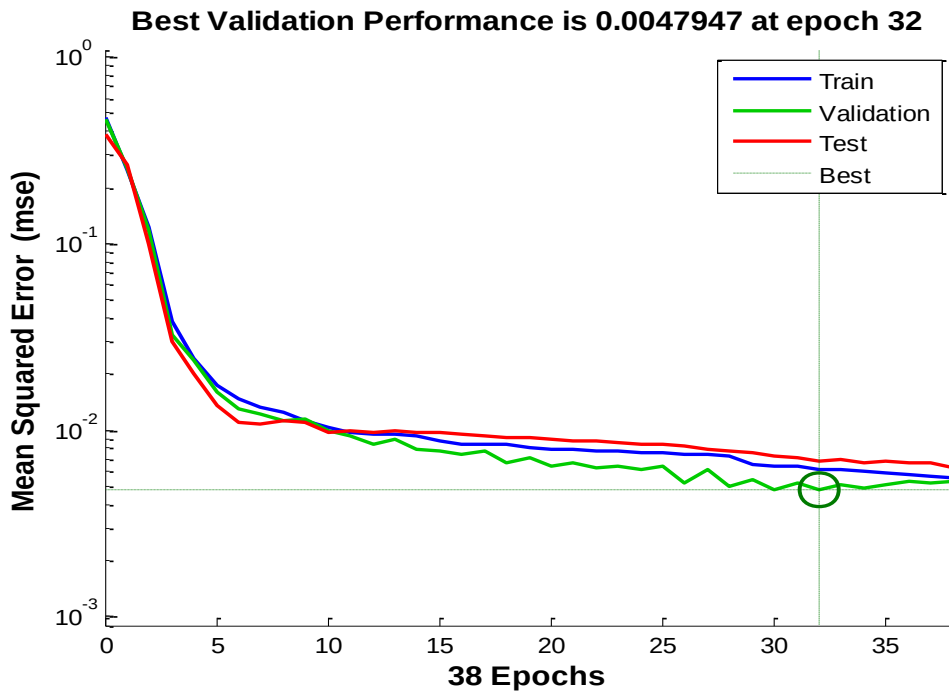


Figure 26: Best validation performance of the optimal training function

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 15 below indicates this model.

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

Table 15: The optimal model

Experimentation I

This Experimentation Result 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 27 below.

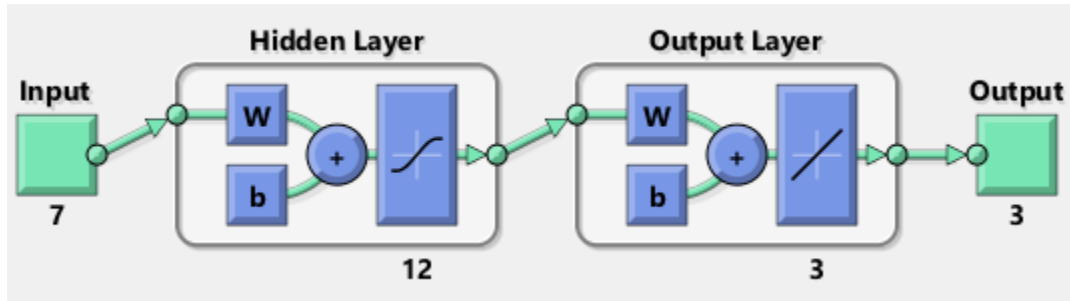


Figure 27: Neural network architecture of the selected model

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.

Experimentation II

This section of experiment 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 28 below.

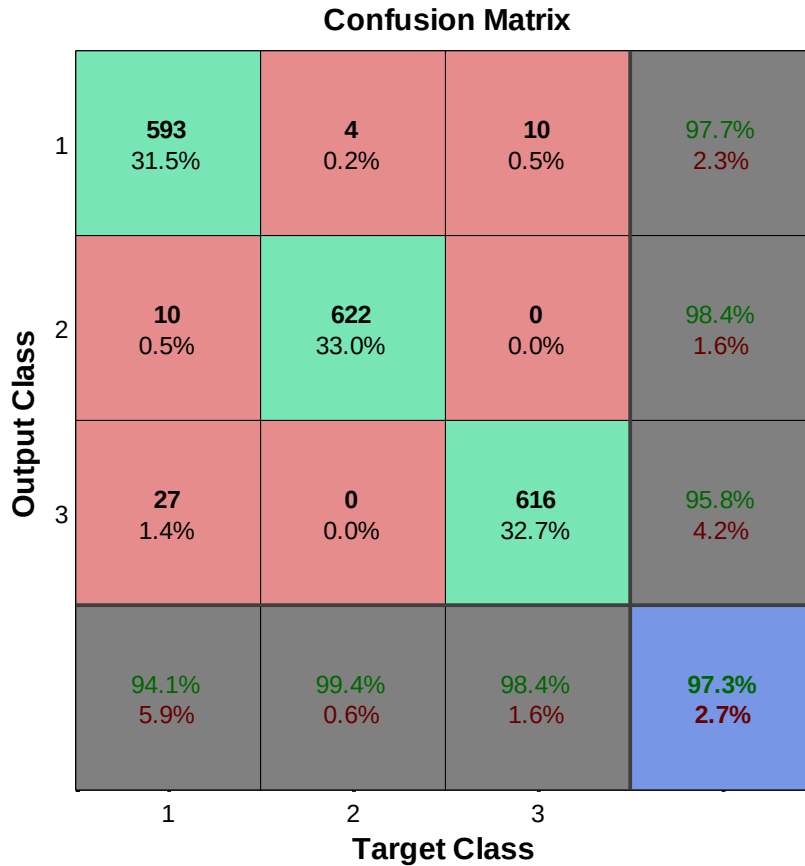


Figure 28: Confusion Matrix of the Selected Model

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 28 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 29 below depicts the regression plot of the optimal model.

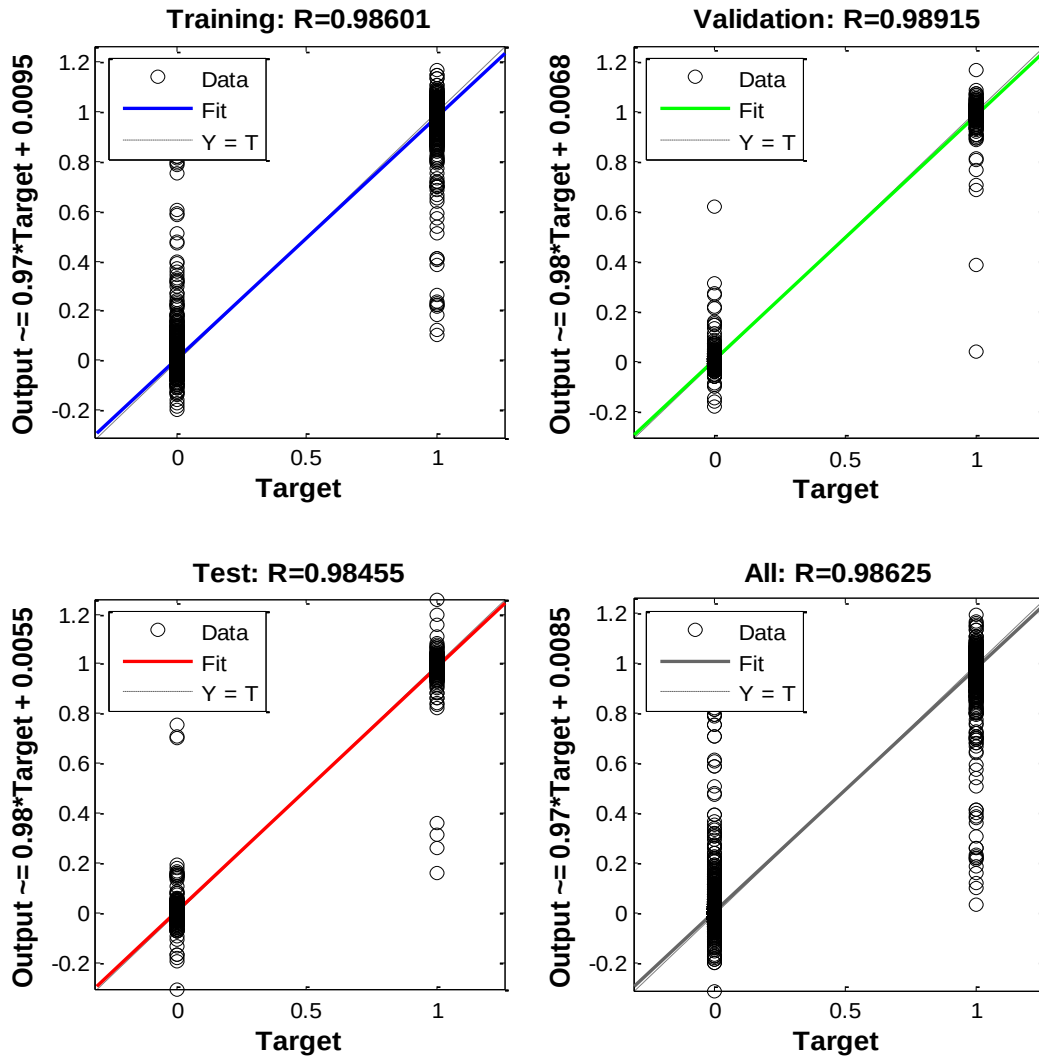


Figure 29: Regression plot of the Optimal Model

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.

5.6 Determine Relative Importance of Input Variables

Experimentation III

This section of experiment used to answer 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] = relief (i, o, 1882) script. After running the script we found that relative importance of geometric parameters as shown in figure 30 below.

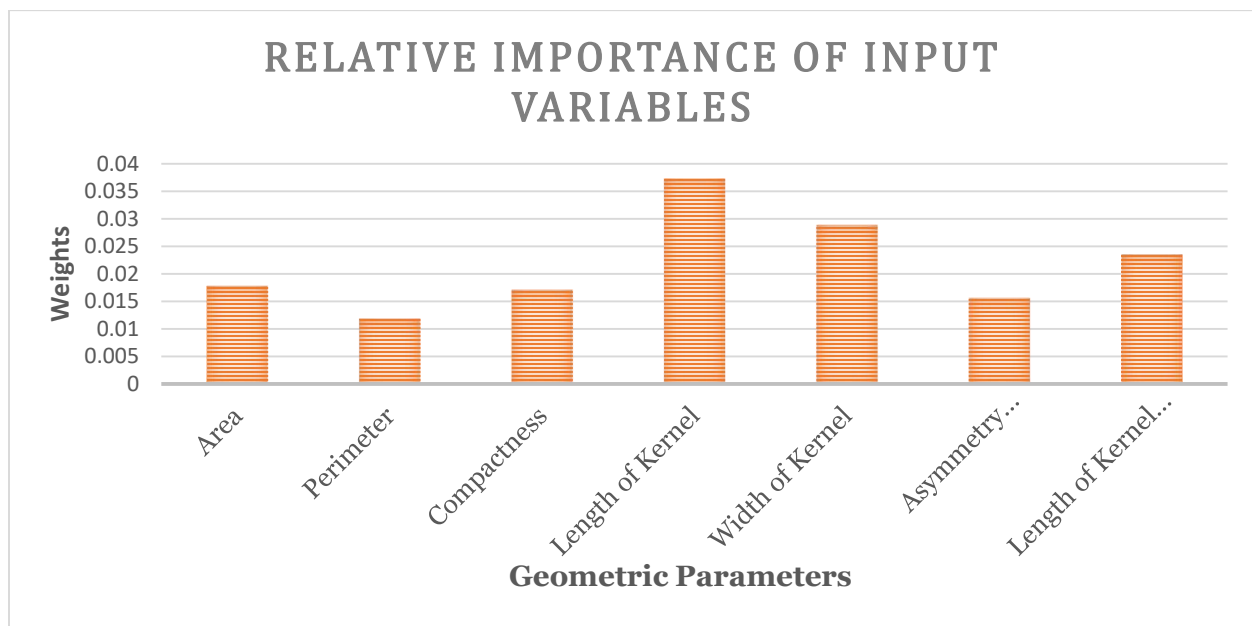


Figure 30: Relative importance of input variables

As we can see from Figure 30 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 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. To compare the significance level of the factors obtained using the proposed model with the domain experts, four experts were interviewed. The ranks of symptoms and factors shown below are the average rank of the four experts. The comparison of the relative importance of symptoms according to experts and proposed model is shown in table 11 below.

S.N	Geometric Parameters	Rank according to domain experts				to	Rank according to the selected model	Weight according to the selected model
1.	Area	4	2	2	4		4	0.0179
2.	Perimeter	5	5	4	5		5	0.0120
3.	Compactness	6	7	7	1		7	0.0172
4.	Length of Kernel	1	4	1	6		1	0.0373
5.	width of kernel	3	1	3	3		3	0.0289
6.	Asymmetry of coefficient	7	6	6	7		6	0.0157
7.	length of kernel grove	2	3	5	2		2	0.0236

Table 16: Relative importance of symptoms according to experts and the proposed model

As we can see from table 16 above, the geometrics parameters ranked according to experts and the proposed model are different. So the experimentation done by the proposed model ranks the geometric parameter as follows; the length of kernel is the first, length of kernel grove is the second, width of the kernel is the fourth, perimeter is the fifth, asymmetry of coefficient is the six and compactness is the last determinant factor for classifying the seeds.

In addition to comparing the result of the study with the domain experts, the researcher tried to compare with two international papers that are done for seed classification.

A. R. Parnian [29], built a connectionist expert system for Wheat Seed Type Classifier using ANN. Independent variables such Area, Perimeter, Compactness, Length of Kernel, width of kernel, Asymmetry of coefficient and length of kernel grove are used but significance levels of this variables are not identified by the researchers.

R. H. Ajaz and I. Campus,[33] presented Seed Classification using Machine Learning Techniques framework for classifying seed using artificial neural networks. The variables used for the classification are the Area, Perimeter, Compactness, Length of Kernel, width of kernel, Asymmetry of coefficient and length of kernel grove.

CHAPTER Six

CONCLUSION, RECOMMENDATION AND FUTURE WORKS

6.1 Conclusion

In this study, ANN method for seed classification critical factors identification and Variety types of seeds are has been implemented for Melkassa Research Center. 7 independent variables were selected and examined in developing the model. The model is developed for prediction of determinant factors for 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. Further investigation could be done on how each category of the selected seed would be affected by the given perimeters in the region. 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.

6.2 Recommendation

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.

6.3 Future Work

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.

BIBLIOGRAPHY

- [1] M. Stubbs, "Big Data in U . S . Agriculture," 2016.
- [2] Wikipedia, "Agriculture in Ethiopia." .
- [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 *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.
- [8] B. A. Prasad, G. N. R. and Vinaya, "Computer Science and Telecommunications," *A Study Var. Expert Syst. Agric. Georg. Electron. Sci. J.*, vol. 4(11), pp. 81–86, 2006.
- [9] S. Et al., "The Information Technology Interaction Model: A Foundation for the MBA Core Course," *MIS Q.*, vol. 19, No. 3, pp. 361–390.
- [10] H. Et al., "design science in information systems research," *MIS Q.*, vol. 28 No. 1, pp. 75–105.
- [11] R. S. Hevner A.R., March S.T., Park J., "Design science in information systems research," *MIS Q.*, vol. 28, no. 1, pp. 75–105, 2004.
- [12] P. Et al., "A Design Science Research Methodology for Information Systems Research," *Publ. J. Manag. Inf. Syst.*, vol. 24, no. 3, pp. 45–78.
- [13] F. S. Gharehchopogh et al., "using artificial neural network in diagnosis of thyroid disease: a case study," *Int. J. Comput. Sci. {{H&H}} Appl.*, vol. 3, No.4.
- [14] A. Et al., "Proposing Efficient Neural Network Training Model for Kidney Stone Diagnosis," *Int. J. Comput. Sci. Inf. Technol.*, vol. 3, No.3, pp. 3900–3904, 2012.
- [15] M. Durairaj and G. Kalaiselvi, "Prediction Of Diabetes Using Soft Computing Techniques- A Survey," *Int. J. Sci. {{H&H}} Technol. Res.*, vol. 4, no. 03.
- [16] P. Fayyad, Usma, Piatetsky-shpiro, G. and Smyth, "From Data Mining to Knowledge Discovery an overview.," in *In advances in knowledge discovery and data mining*, 1996.
- [17] Giudici P., *Applied Data mining Statistical Method for Business and Industry*. University of Pavia. Italy.
- [18] P. M. S. P. Babu, "Big Data and Predictive Analytics in ERP Systems for Automating Decision Making Process."
- [19] D. Graupe, "principles of artificial neural networks," *World Sci. Publ. Co. Pte. Ltd.*, 2007.
- [20] K. Suzuki, *ARTIFICIAL NEURAL NETWORKS I METHODOLOGICAL ADVANCES AND Edited by Kenji Suzuki.* .
- [21] K. Girish, "Artificial Neural networks and its applications," *IARI*, vol. 110, 2012.
- [22] Prof. K.Suzuki, "Introduction to Artificial Neural Networks," in *Artificial Neural Networks- Methodological Advances and Biomedical Applications*, p. 374.
- [23] 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.
- [24] M. Beale et al., "Neural Network Toolbox™ User's Guide," 2006.
- [25] D. M. M. H.S. N. Murthy, "ANN Model to Predict Coronary Heart Disease Based on Risk Factors," *Bonfring Int. J. Man Mach. Interface*, vol. 3, No.2.
- [26] C. V. S. D. vally, "Diagnosis Chest Diseases Using Neural Network and Genetic Hybrid Algorithm," *Int. J. Eng. Res. Appl. ISSN 2248*, vol. 5, no. 1(Part 2), pp. 20–26, 2015.
- [27] S.Radhimeenakshi and G.M.Nasira, "Prediction of Heart Disease using Neural Network with Back

- Propagation,” *Int. Conf. Comput. Intell. Syst.*, vol. 04, no. Special Issue, pp. 1166–1169, 2015.
- [28] Dr. T.P. Shah and P.J. Shah, “Connectionist Expert System for Medical Diagnosis using ANN– A case study of skin disease Scabies,” *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 8, pp. 227–230, 2013.
- [29] A. R. Parnian, “Autonomous Wheat Seed Type Classifier System,” vol. 96, no. 12, pp. 14–17, 2014.
- [30] C. S. Silva and U. Sonnadara, “Classification of Rice Grains Using Neural Networks,” vol. 29, pp. 9–14, 2013.
- [31] J. Šťastný, V. Konečný, and O. Trenz, “Agricultural data prediction by means of neural network,” vol. 2011, no. 7, pp. 356–361, 2011.
- [32] S. S. Dahikar and S. V Rode, “Agricultural Crop Yield Prediction Using Artificial Neural Network Approach,” vol. 2, no. 1, pp. 683–686, 2014.
- [33] R. H. Ajaz and I. Campus, “Seed Classification using Machine Learning Techniques,” vol. 2, no. 5, pp. 1098–1102, 2015.
- [34] J. Kim, “A Prediction Model based on Big Data Analysis Using Hybrid FCM Clustering,” pp. 337–339, 2014.
- [35] C. Balaji, “Environment Change Prediction to Adapt Climate- Smart Agriculture Using Big Data Analytics,” vol. 4, no. 5, pp. 1995–2001, 2015.
- [36] UCI, “UCI Machine Learning Repository,” 2014.
- [37] H. A. H. Maitha, S. Al and H. A. Assi, “Using MATLAB to Develop Artificial Neural Network Models for Predicting Global Solar Radiation in Al Ain City-UAE,” *Eng. Educ. Res. using MATLAB*, 2011.
- [38] A. H. A. M. H. Al Shamisi and H. A. N. Hejase, “Using MATLAB to Develop Artificial Neural Network Models for Predicting Global Solar Radiation in Al Ain City – UAE,” *Eng. Educ. Res. Using MATLAB*.
- [39] G. Zhang et al, “Forecasting with artificial neural networks: The state of the art,” *Int. J. Forecast.*, vol. 14, pp. 35–62, 1997.
- [40] P. Sharma et al., “feedforward neural network : A Review,” *Int. J. Adv. Res. Eng. Appl. Sci.*, vol. 2, no. 10, 2013.
- [41] C. Faraway, J., Chatfield, “Time series forecasting with neural networks: A comparative study using the airline data,” *Appl. Stat.*, vol. 47, no. 2, pp. 231–250, 1998.
- [42] M. A. Shahin et al., “State of the Art of Artificial Neural Networks in Geotechnical Engineering,” *J. Geotech. Eng.*, 2008.
- [43] Cybenko, G., “Approximation by super positions of a sigmoidal function,” *Math. Control. Signals, Syst.*, vol. 3, pp. 303–314, 1989.
- [44] B. Sharma, Prof. K. Venugopalan, “Comparison of Neural Network Training Functions for Hematoma Classification in Brain CT Images,” *IOSR J. Comput. Eng.*, vol. 16, Versio, no. 1, e-ISSN: 2278–0661, p- ISSN: 2278–8727, pp. 31–35, 2014.
- [45] H. Demuth and M. Beale, “Neural Network Toolbox, User’s Guide,” vol. 4, 2001.
- [46] H. Demuth, “Neural Network Toolbox.”
- [47] H. B. D. M. H. Beale, M. T. Hagan, “Neural Network Toolbox™ 7 User’s Guide.”
- [48] M. Swain et al., “An approach for iris plant classification using neural network,” *Int. J. Soft Comput. (IJSC)*, vol. 3, No.1, 2012.
- [49] D. Howard, B. Mark and H. Martin, “Neural Network Toolbox TM 6: User’s Guide,” *MathWorks TM Accel. pace Eng. Sci.*, 2009.
- [50] D. S. SIG, “Improving Predictive Data Analytics Methods in line with Aviation Industry,” Adama, 2016.

APPENDICES

Appendix A [50]

Interview questionnaires

Predictive Model to predict Seed Classes Using Machine Learning

Following the questionnaires are prepared to a targeted stakeholder in order to gather facts and views on the role and impacts of big data analytics in several Agriculture sectors.

This exercise has two main objectives:

1. To understand the use of machine learning in agriculture for classifying seed and predicting
2. To gather insights on the characteristics of a functioning data ecosystem in your specific domain and identify existing or potential barriers to the development of data-driven agriculture sectors in Ethiopia.

Part I: Organization Information

1. Sector:

☐ Agriculture

☐ Metrology

☐ Energy

☐ Public sector

☐ Research

☐ Maritime and Fisheries

☐ Academia

☐ IT Experts

Name: _____

City: _____

Phone: _____

2. Which departments in your organization are involved in using data technologies and data analytics?

Part II: Software and Hardware Configuration Details (Question for IT Experts)

3. Does your organization have a strategy on big data or data analytics?

☐ Yes ☐ No

4. What are the main sources of data for your database?

5. What analytical tools do you consider particularly important?

6. Does your organization have the right tools to data?

☐ Has now ☐ Expects to have them in 5 years ☐ No plans to have them ☐ Do not know

7. What tools do you consider particularly important to handle unstructured data?

8. Are you using any Predictive Machine Learning System for seed classification condition?

☐ Yes ☐ No

9. Give details on server and client systems configurations for your system.

a. Hardware Specifications:

➤ Processor's Model & Speed: _____

➤ Hard Disk's Model & Size: _____

➤ RAM Size: _____

b. Software Specifications:

➤ Operating System: _____

➤ Programming Languages: _____

➤ Application Software: _____

➤ Performance Monitoring Software: _____

10. Does your organization have the necessary skills to handle data?

☐ Yes ☐ No

11. Is senior management in your organization involved in big data-related projects?

☐ Yes ☐ No

12. In your organization, data collection is:
☐ In-house ☐ Outsourced ☐ Both
13. In your organization, data analytics is:
☐ In-house ☐ Outsourced ☐ Both
14. Does your organization buy datasets from other entities?
☐ Yes ☐ No

If yes, please specify

15. Which one is the deterrent to predict seed classes of wheat? (Question for Agronomists)

- ☐ Area ☐ width of kernel
☐ Perimeter ☐ Asymmetry of coefficient
☐ Compactness ☐ length of kernel grove
☐ Length of Kernel

16. What testing tools have you used for seed class prediction

Part III Challenges conditions for Seed Class Prediction (question for IT Experts)

17. How relevant are the following big data-related challenges for your organization?

For each row, please tick only one box. To suggest machine Learning challenges, please use the space provided

	Very Important	Important	Moderately Important	Of Little Importance	Unimportant (not a challenge)
Data quality	☐	☐	☐	☐	☐
Availability of data	☐	☐	☐	☐	☐
Access rights to data	☐	☐	☐	☐	☐
Data ownership issues	☐	☐	☐	☐	☐
Cost of data	☐	☐	☐	☐	☐
Lack of facilities, infrastructure	☐	☐	☐	☐	☐
Lack of pre-processing facilities	☐	☐	☐	☐	☐
Lack of technology	☐	☐	☐	☐	☐
Shortage of talent/skills	☐	☐	☐	☐	☐
Privacy concerns and regulatory risks	☐	☐	☐	☐	☐
Security	☐	☐	☐	☐	☐

Other, please specify:

Additional comments

If you have additional comments about big data challenges, please use the space provided below:

Thank you for taking the time to complete this questionnaire!

Your contribution is very important to understand the role and impacts of big data analytics in the Ethiopian agriculture.

Additional comments

If you have comments, please use the space provided below:

Should you have any question, please contact: tokkiti@gmail.com

Appendix B

Sample Source codes on ANN

```
x=[];
y=[];
i=x';
o=y';
net1= newff(i,o,4);
[net1,tr1]= train(net1,i,o);
net2= newff(i,o,5);
[net2,tr2]= train(net2,i,o);
net2= newff(i,o,6);
[net3,tr3]= train(net3,i,o);
net3= newff(i,o,6);
[net3,tr3]= train(net3,i,o);
net4= newff(i,o,7);
[net4,tr4]= train(net4,i,o);
net5= newff(i,o,8);
[net5,tr5]= train(net5,i,o);
net6= newff(i,o,9);
[net6,tr6]= train(net6,i,o);
net7= newff(i,o,10);
[net7,tr7]= train(net7,i,o);
net8= newff(i,o,11);
[net8,tr8]= train(net8,i,o);
net9= newff(i,o,12);
[net9,tr9]= train(net9,i,o);
net10= newff(i,o,13);
[net12,tr12]= train(net12,i,o);
[net10,tr10]= train(net10,i,o);
net11= newff(i,o,14);
[net11,tr11]= train(net11,i,o);
net12= newff(i,o,15);
[net12,tr12]= train(net12,i,o);
// selected neurons using best MSE
net63= newff(i,o,5);
net63.trainfcn='trainoss';
[net63,tr63]= train(net63,i,o);
net64= newff(i,o,6);
net64.trainfcn='trainoss';
[net64,tr64]= train(net64,i,o);
net65= newff(i,o,11);
net65.trainfcn='trainoss';
[net65,tr65]= train(net65,i,o);
net66= newff(i,o,12);
net66.trainfcn='trainoss';
[net66,tr66]= train(net66,i,o);
net67= newff(i,o,15);
net67.trainfcn='trainoss';
[net67,tr67]= train(net67,i,o);
view (net67);
simpleclassOutputs = sim(net67,i); //to create confusion matrix
plotconfusion(o,simpleclassOutputs);
```