

A
PROJECT REPORT
ON
“Medical Diagnosis: A Neural Network Approach”

Submitted in partial fulfillment of the requirement
for the award of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE & ENGINEERING

BY
NAVEEN CHANDRA YADAV
PRAFULL GAJBHIYE

UNDER THE SUPERVISION OF:
MR. VAIBHAV KANT SINGH
ASSISTANT PROFESSOR
(Computer science & Engineering)



Department of Computer Science & Engineering
Institute of Technology
Guru Ghasidas Vishwavidyalaya, Bilaspur(C.G.) India
Session: 2012-13

CERTIFICATE

This is to certify that the Project entitled **“Medical Diagnosis: A Neural Network Approach”** embodied in the dissertation has been carried out to my satisfaction by **Naveen Chandra Yadav** and **Prafull Gajbhiye** at the department of Computer Science & Engineering.

I declare that this project is carried out under my supervision and guidance towards the partial fulfilment of the requirement for the award of Bachelor of Technology degree in Computer Science & Engineering.

The work is original as it has not been earlier submitted either in part or full for any purpose before.

Approved By:

Dr. Manish Shrivastava

Head Of Department

Computer Science & Engineering,

IT.GGV, Bilaspur (C.G.)

Guided By:

Mr. Vaibhavkant Singh

Assistant Professor

Computer Science & Engineering

IT.GGV, Bilaspur (C.G.)

DECLARATION

We hereby declare that the work presented in this dissertation entitled **“Medical Diagnosis: A Neural Network Approach”** has been done by us and this dissertation embodies our own work.

Naveen Chandra Yadav

Prafull Gajbhiye

Dept. of Comp. Sci. & Engg.

IT.GGV, Bilaspur,(C.G)

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people who made it possible and whose constant guidance and encouragement crown all efforts with success. This acknowledgement transcends the reality of formality when we would like to express deep gratitude and respect to all those people behind the screen who guided, inspired and helped us for the completion of our project work.

We are very thankful to our project coordinator **Mr. Vaibhavkant Singh** and Head of department of computer science & engineering **Dr. Manish Shrivastava** for giving us an opportunity to do the project and giving us continuous help and valuable suggestions to guide us in the successful completion of the project work. We are also thankful to **Dr. Shailendra Kumar**, Director I.T. for allowing us to work on this project.

We also thank all our friends who have directly or indirectly helped us in completing the task successfully.

ABSTRACT

Breast cancer is one of the fatal disorders causing death in people and second most common reason of cancer death in women. Instead of its fatal nature, extent of damage and cancer spreading can be reduce with good possibility if diagnosed in early stage .In this project report an artificial neural network inspired from biological neural network is used for pattern classification of benign and malignant cells. A feed forward neural network model with back propagation learning algorithm for training the neural network using breast cancer database is simulated with all the variable network constraints to make it efficient, robust and fault tolerated pattern classifier.

Contents

1. Introduction.....	1
1.1 What is Cancer?.....	1
1.1.1 What is Breast Cancer	
1.1.2 Breast Cancer Statistics	
1.2 Goals & Objectives: Malignant or Benign?.....	5
1.3 Classification.....	6
1.4 Why Neural Network.....	7
1.5 Method: Artificial Neural Network Approach.....	8
1.5.1 Introduction to Artificial Neural Network Approach	
1.5.2 Applications in Medical Stream	
2 Related Works.....	14
2.1 Existing Systems.....	14
2.2 The Origins of the Data.....	15
3. System Overview.....	16
3.1 Requirements.....	16
3.2 System Specification.....	16
3.3 Feasibility study.....	17
4. Theoretical Foundations: The Engineering Model.....	19
4.1 Introduction.....	19
4.2 Details of Proposed Models.....	19
4.2.1 Deployed Algorithms	
4.3 Training the Neural Network.....	23
4.4 Deployed WDBC database.....	24
5. System Analysis: System Model and System Architecture.....	26
5.1 SDLC paradigm.....	26
5.2 UML Diagrams/Specifications.....	27
5.2.1 Structural Model - Class Diagram,	
5.2.1 Behavioral Model – Use Case Diagram, Activity Diagram and Sequence Diagram	
6. Implementations.....	33
6.1 Software development platform.....	33
7. 7. Results & Conclusion.....	36
7.1 System Performance (Results).....	36
7.2 Conclusions.....	39
8. Future Scope.....	40

References

Appendices

☐ Terms

List of Tables

Understanding Significance of each of 9 input units of data.....	20
Neural Network Performance analysis statics.....	36
Statics table stating observation on each of 9 logical input.....	38

List of Figures

Anatomy of Breast	2
Classification of breast cancer sample.....	5
A Simple Neural Network.....	8
Decision Regions Created by a multilayer perceptron.....	10
A Backpropagation network.....	12
Algorithm Workflow.....	22
Logical ANN.....	22
Separation of data for various experiments to performed on neural network.....	25
UML Diagram Series	
Class Diagram.....	28
Use Case Diagram.....	29
Activity Diagram User level.....	30
Activity Diagram Admin level.....	31
Sequence Diagram.....	32
Application Snapshots 1.....	34
Application Snapshots 1.....	35
Overfitting Effect between training and validation data.....	37

Chapter 1

1. Introduction

Artificial Neural Networks (ANNS) are programs built to model the brain's neural-syntax structure. With their remarkable ability to learn the meaning of complicated data, neural networks can be used to detect patterns that are too complex for a human or another computer program to notice. The more experience a neural network has, the better the network can learn to think and analyze scenarios. There are numerous advantages in using ANNS, such as their ability to use adaptive learning provide projections of new situations. Neural networks are self-organized, so they are not dependent on the knowledge of the programmer. ANNS have a wide variety of applications, and with all of their potential, it is not surprising they are used in the medical field.

1.1 What is Cancer?

Cancer occurs as a result of mutations, or abnormal changes, in the genes responsible for regulating the growth of cells and keeping them healthy. The genes are in each cell's nucleus, which acts as the "control room" of each cell. Normally, the cells in our bodies replace themselves through an orderly process of cell growth: healthy new cells take over as old ones die out. But over time, mutations can "turn on" certain genes and "turn off" others in a cell. That changed cell gains the ability to keep dividing without control or order, producing more cells just like it and forming a tumor.

A tumor can be benign (not dangerous to health) or malignant (has the potential to be dangerous). Benign tumors are not considered cancerous: their cells are close to normal in appearance, they grow slowly, and they do not invade nearby tissues or spread to other parts of the body. Malignant tumors are cancerous. Left unchecked, malignant cells eventually can spread beyond the original tumor to other parts of the body.

1.1.1 What is breast cancer?

The term “breast cancer” refers to a malignant tumor that has developed from cells in the breast. Usually breast cancer either begins in the cells of the lobules, which are the milk-producing glands, or the ducts, the passages that drain milk from the lobules to the nipple. Less commonly, breast cancer can begin in the stromal tissues, which include the fatty and fibrous connective tissues of the breast.

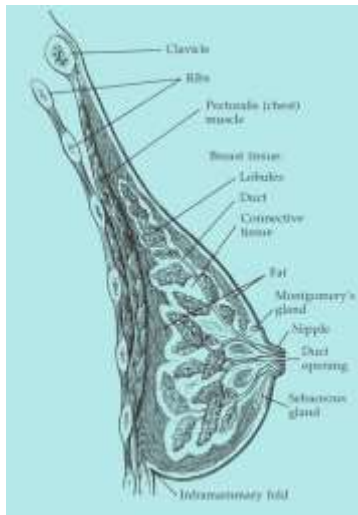


Fig.1.1 Depicting Anatomy of Breast

In Short **Breast cancer is an uncontrolled growth of breast cells.**

The breast is made up mainly of lobules (milk-producing glands in women), ducts (tiny tubes that carry the milk from the lobules to the nipple), and stroma (fatty tissue and connective tissue surrounding the ducts and lobules, blood vessels, and lymphatic vessels)

Over time, cancer cells can invade nearby healthy breast tissue and make their way into the underarm lymph nodes, small organs that filter out foreign substances in the body. If cancer cells get into the lymph nodes, they then have a pathway into other parts of the body. The breast cancer's stage refers to how far the cancer cells have spread beyond the original tumor.

Breast cancer occurs mainly in women, but men can get it, too. Many people do not realize that men have breast tissue and that they can develop breast cancer.

Like all cells of the body, a man's breast duct cells can undergo cancerous changes. But breast cancer is less common in men because their breast duct cells are less developed than those of women and because they normally have lower levels of female hormones that affect the growth of breast cells.

1.1.2 Breast Cancer Statistic:

- Breast cancer is the top cancer in women worldwide and is increasing particularly in developing countries where the majority of cases are diagnosed in late stages.
- In US about 1 in 8 women (just under 12%) will develop invasive breast cancer over the course of her lifetime.
- About 85% of breast cancers occur in women who have no family history of breast cancer. These occur due to genetic mutations that happen as a result of the aging process and life in general, rather than inherited mutations.

Breast cancer is one of the leading causes of death among women. However, there is clear evidence that early diagnosis and subsequent treatment can significantly improve the chance of survival for patients with breast cancer.

Diagnosis is the identification of the nature and cause of anything or a diagnosis can be regarded as an attempt to classification of an individual's condition into separate and distinct categories that allow medical decisions about treatment and prognosis to be made. Subsequently, a diagnostic opinion is often described in terms of a disease or other condition, but in the case of a wrong diagnosis, the individual's actual disease or condition is not the same as the individual's diagnosis.

A medical diagnosis is concerned with identifying the disease which is most likely to cause a given set of clinical findings. A clinician uses several sources of data and puts the pieces of the puzzle together to make a diagnostic impression. The initial diagnostic impression can be a broad term describing a category of diseases instead of a specific disease or condition.

Mammography_μ has become one of the major diagnostic procedures with a proven capability for detecting early-stage, clinically occult breast cancers. However, breast cancer in their early stage is small and frequently their radiographic appearance differs only from that of normal.

tissue or benign abnormalities. Because of this subtlety, the potential for misclassification by radiologist is substantial. Only 10-30% of cases that have suspicious Mammographically suspicious findings and are subjected to biopsy prove to be malignant.

On the other hand, approximately 10-30% of patients with breast cancer misdiagnosed by Mammography (have the cancer missed or not detected on their Mammograms 10-14).

Besides the subtle nature of radiographic lesions associated with breast cancer, many errors in radiological diagnoses can be attributed to human factors such as subjective or varying decision criteria and simple oversight. etc., 15-17 Studies suggest that these errors may occur even with experienced radiologists.18-19. These errors may be reduced by the use of automated detection machines that can locate and classify possible lesions.

Thus Medical usage demands neural networks to achieve accuracy with their diagnosis and reduce malignant false negatives. Neural technology is the attempt to replicate the brain through artificial intelligence.

Motivations

GLOBOCAN 2008

CANCER FACT SHEET

International Agency for Research on Cancer

Incidence
Mortality

Breast Cancer Incidence, Mortality and Prevalence Worldwide in 2008 : Summary

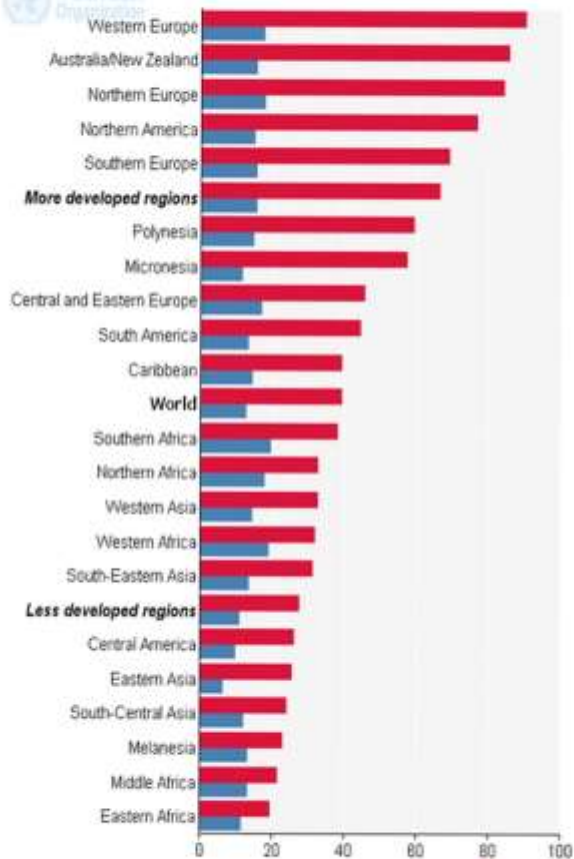
Estimated numbers (thousands)	Cases	Deaths	5-year prev.
World	1384	458	5189
More developed regions	692	189	2808
Less developed regions	691	269	2380
WHO Africa region (AFRO)	68	37	216
WHO Americas region (PAHO)	320	82	1272
WHO East Mediterranean region (EMRO)	61	31	215
WHO Europe region (EURO)	450	139	1770
WHO South-East Asia region (SEARO)	203	93	630
WHO Western Pacific region (WPRO)	279	73	1081
IARC membership (22 countries)	740	214	2865
United States of America	182	40	761
China	169	44	631
India	115	53	315
European Union (EU-27)	332	89	1329

At a glance

Breast cancer is by far the most frequent cancer among women with an estimated 1.38 million new cancer cases diagnosed in 2008 (23% of all cancers), and ranks second overall (10.9% of all cancers). It is now the most common cancer both in developed and developing regions with around 690 000 new cases estimated in each region (population ratio 1:4). Incidence rates vary from 19.3 per 100,000 women in Eastern Africa to 89.7 per 100,000 women in Western Europe, and are high (greater than 80 per 100,000) in developed regions of the world (except Japan) and low (less than 40 per 100,000) in most of the developing regions.

The range of mortality rates is much less (approximately 6-19 per 100,000) because of the more favorable survival of breast cancer in (high-incidence) developed regions. As a result, breast cancer ranks as the fifth cause of death from cancer overall (458 000 deaths), but it is still the most frequent cause of cancer death in women in both developing (269 000 deaths, 12.7% of total) and developed regions, where the estimated 189 000 deaths is almost equal to the estimated number of deaths from lung cancer (188 000 deaths).

International Agency for Research on Cancer



GLOBOCAN 2008 (IARC)

Estimated age-standardised rates (World) per 100,000

Incidence
Mortality

1.2 Goals & Objectives: Malignant or Benign?

The development of this model is aimed to reduce the diagnosis time as well as increasing the accuracy percentage in classifying mass in breast to either benign, or malignant.

The Question: Malignant or Benign?

- Can a breast cancer neural network be optimized to improve the success of diagnosis using Fine Needle Aspirates (FNA)?
- Specifically, can the network be optimized to reduce the number of malignant false negatives while handling original, unformatted data?

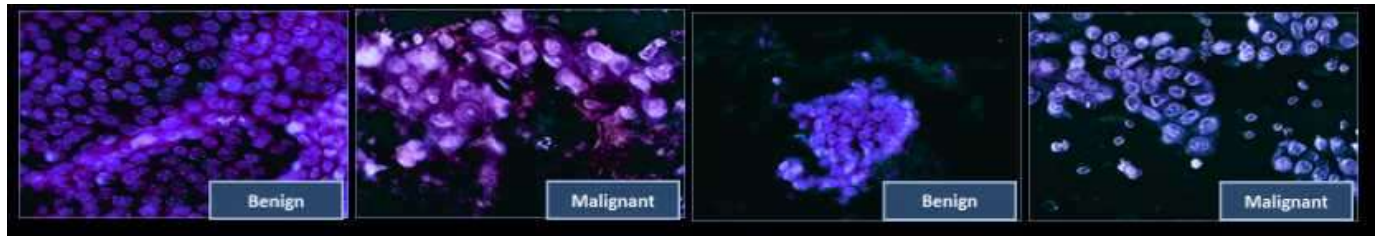


Fig.1.2 Classification of breast cancer sample using FNA

1.3 Classification

Classification is a form of data analysis which can be used to extract model describing important data classes. Such analysis can help provide us with a better understanding of the data at large. It predicts categorical (discrete, unordered) labels.

Data classification is a two-step process

- **Learning:** Training data are analyzed by a classification algorithm. In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or “learning from” a training set established.
- **Classification:** Test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples.

Evaluation of classification according to the following criteria:

Accuracy: The accuracy of a classifier refers to the ability of a given classifier to correctly predict the class label of new or previously unseen data

It can also be estimated using one or more test sets that are independent of the training set.

Pattern Classification/ recognition

A pattern is a qualitative or quantitative description of an object or concept or event. A pattern class is a set of patterns sharing some common properties. Pattern Classification/ recognition refer to the categorization of input data into identifiable classes by recognizing significant features or attributes of the data.

Neural Networks can recognize, classify, convert and learn pattern over a given set of input data.

1.4 Why Neural Network

Artificial neural networks can perform these exciting tasks:

- Recognize something it has never seen before!
- Classify objects into classes, based on generalized training!
- Match patterns to patterns; forward and back: bi-directional association!
- Predict the future, by extracting patterns in the past!

Artificial neural networks detect patterns too complex to be recognized by humans. Neural networks can be explicitly programmed to perform a task by manually creating the topology and then setting the weights of each link and threshold. However, this by-passes one of the unique strengths of neural nets: the ability to program themselves. The most basic method of training a neural network is trial and error. If the network isn't behaving the way it should, change the weighting of a random link by a random amount. If the accuracy of the network declines, undo the change and make a different one. It takes time, but the trial and error method does produce results. The task is to mirror the status of the input row onto the output row

Why Backpropagation Algorithm

On Deploying hit and trail method for training the neural network it takes times. Unfortunately, the number of possible weightings rises exponentially as one adds new neurons, making large general-purpose neural nets impossible to construct using trial and error methods. The back-propagation algorithm compares the result that was obtained with the result that was expected.

It then uses this information to systematically modify the weights throughout the neural network. This training takes only a fraction of the time that trial and error method take. It can also be reliably used to train networks on only a portion of the data, since it makes inferences. The resulting networks are often correctly configured to answer problems that they have never been specifically trained on.

1.5 Method: Artificial Neural Network Approach

1.5.1 Introduction to Neural Network

Where does intelligence emerge? From computational point of view there are two important ways to answer this question. One is based on symbolism, and the other, based on connectionism. The former approach models intelligence using symbols, while the latter using connections associated with weights.

Biological neurons transmit electrochemical signals over neural pathways. Each neuron receives signals from other neurons through special junction called synapse. Some inputs tend to excite the neurons; others tend to inhibit it. When the cumulative effect exceeds a threshold, the neurons fire and send a signal down to other neurons.

Artificial neurons model these simple biological characteristics. Each artificial neuron receives a set of inputs. Each input is multiplied by a weight analogous to a synaptic strength. The Sum of all weighted inputs determines the degree of firing called the activation level.

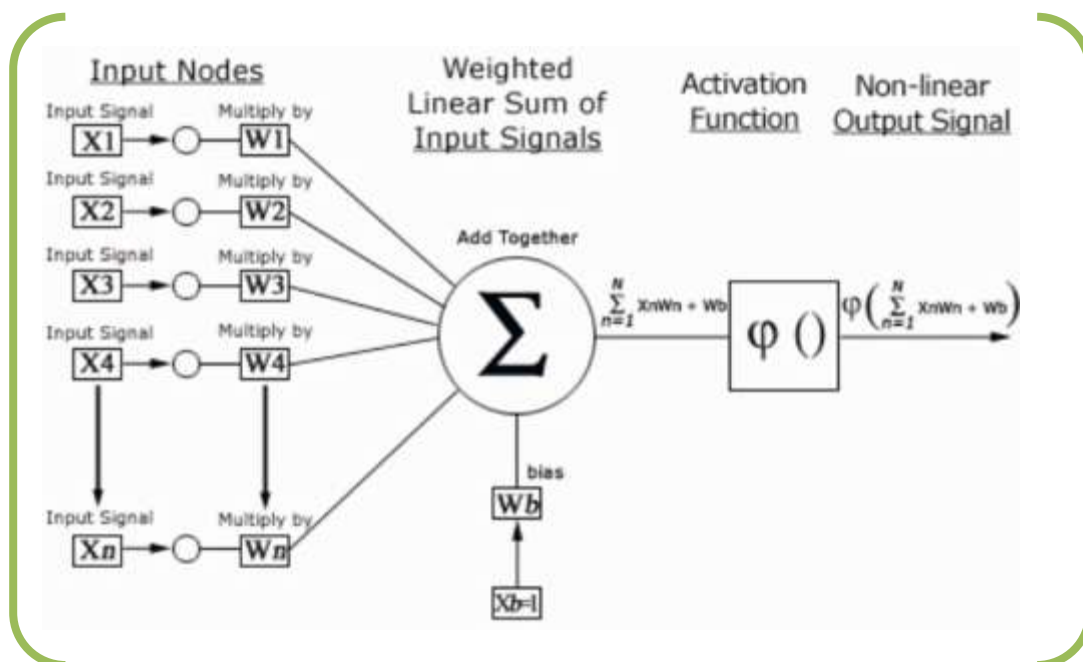


Fig 1.3 depicts An Artificial neural network

Neural Network Models

There are three aspects involved in the construction of a Neural Networks.

- Structure: The architecture and topology of Neural Networks.
- Encoding: The method of changing weights (Training).
- Recall: The method and capacity to retrieve information.

Various Neural Networks models exist and among this Feed Forward Neural Network are considered in this study for the construction of neural network. Because, this model, besides being popular and simple, is easy to implement and appropriate for classification applications.

Classification Model

A neural network classifies a given object presented to it according to the output activation. For binary outputs, 1 corresponds to one class and 0 to the other.

A Single Layer perceptron consists of an input and an output layer. The activation function employed is hard limiting function. An output unit will assume the value 1 if the sum of its weighted input is greater than its threshold.

In terms of classification an object will be classified by unit j into class A if

$$\sum W_{ji} X_i > \theta_j$$

Where W_{ji} is the weight from the unit i to unit j , X_i is the input from unit i , and θ_j is the threshold on the unit j .

A Multilayer perceptron is a feed forward neural network with at least one hidden layer. It can deal with nonlinear classifications problems because it can form more complex regions (rather than just hyper planes). Each node in the first layer (above the input layer) can create hyper planes to create convex decision regions. Each node in the third layer can combine convex regions to form concave regions. The idea is illustrated in the following figure. It is thus possible to form any arbitrary regions with sufficient layers and sufficient hidden units.

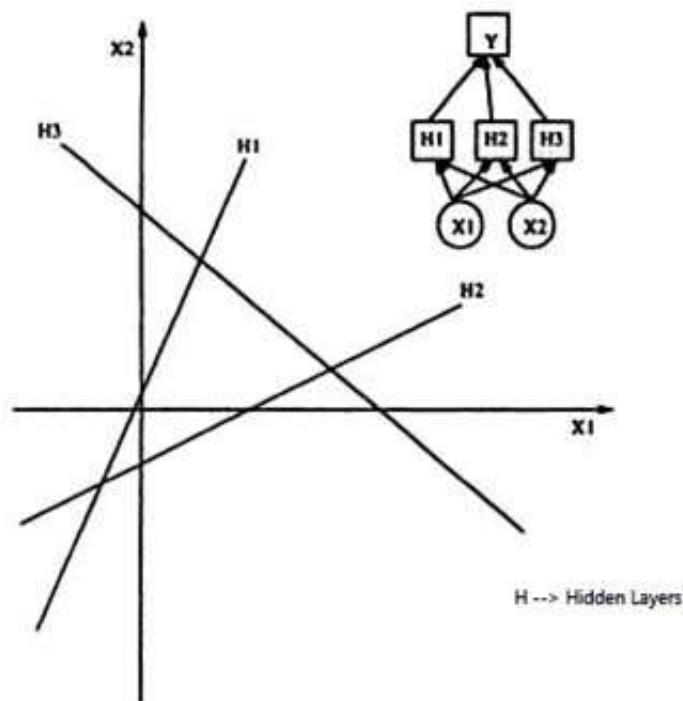


Figure 1.4 Decision Regions Created by a multilayer perceptron

Training Neural network

In the training (or learning) phase, a set of training instance is given. Each training instance is typically described by a feature vector (called an input vector). It may also be associated with a

desired outcome (a concept, a class, etc.) which is encoded as another vector (called a desired output vector) Starting with some random weight setting the neural network is trained to adapt itself to the characteristics of the training instances by changing weight inside the network. In each training cycle, we present an instance to the network. It generates an output vector, which is compared with the desired output vector (if available). In this way, the error for each output unit is calculated and then used to update relevant weights.

In a multilayer network, the error of hidden units are not observed directly but can be estimated with some heuristic. Each weight change is hoped to reduce the error. When all instances are examined the network will start over with the first instance and repeat. Iterations continue until the system performance (in terms of error magnitude) has reached to a satisfactory level.

Advantages of Neural Networks for Classification

- Neural Networks are more robust than decision trees because of the weights.
- The Neural Networks improves its performance by learning. This may continue even after the training set has been applied.
- There is a low error rate and thus a high degree of accuracy once the appropriate training has been performed.
- Neural Networks are more robust than decision trees in noisy environment.

Feed Forward Networks with Backpropagation

The feed forward backpropagation network is a very popular model in neural network. It does not have feedback connections, but the errors are back propagated during training. Backpropagation learning consists of two passes through the different layers of the network: a forward pass and backward pass.

In forward pass, input vector is applied to the sensory nodes of the network and its effect propagates through the network layer by layer. Finally, a set of outputs is produced as the actual response of the network. During the forward pass the synaptic weights of the network are all fixed.

During the backward pass, the synaptic weights are all adjusted in accordance with an error correction rule. The actual response of the network is subtracted from a desired (target) response to produce an error signal. This error signal is then backpropagated through the network, against the direction of synaptic connections.

Backpropagation algorithm can be improved by considering momentum and variable learning rate. Momentum allows a network to respond not only to the local gradient, but also to the recent trends in error surface. Acting like a low pass filter, momentum allows the network to ignore small features in the error surface. Without momentum, a network may get stuck in a shallow local minimum.

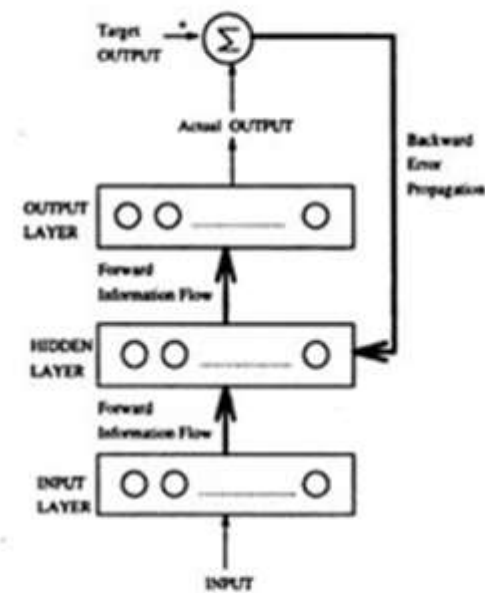


Figure 1.5: The backpropagation network.

In

backpropagation with momentum, the weight change is in a direction that is a combination of the current and previous gradients. This is a modification of gradient descent whose advantages arise chiefly when some training data are very different from the majority of the data. Convergence is sometimes faster if a momentum term is added to the weight update formulas. The performance of algorithm is very sensitive to the proper setting of the learning rate. If the learning rate is set too high, the algorithm may oscillate and become unstable. If the learning rate too small, the algorithm

will take too long to converge. It is not practical to determine the optimal setting for the learning rate before training and in fact the optimal learning rate changes during the training process, as the algorithm moves across the performance surface. Performance of the backpropagation can be improved by allowing the learning rate to change during the training process. An adaptive learning rate will attempt to keep the learning step size as large as possible while keeping the learning process stable. The learning rate is made responsive to the complexity of the local error surface.

1.5.2 Application in Medical Stream

Keeping in view of the significant characteristics of NN and its advantages for the implementation of the classification problem, Neural Network technique is considered for the classification of data related to medical field in this study. Owing to their wide range of applicability and their ability to learn complex and nonlinear relationships including noisy or less precise information Neural Networks are very well suited to solve problems in biomedical engineering. By their nature, Neural Networks are capable of high-speed parallel signal processing in real time. They have an advantage over conventional technologies because they can solve problems that are too complex-problems that do not have an algorithmic solution or for which an algorithmic solution is too complex. Neural Networks are trained by examples instead of rules and are automated. This is one of the major advantages of neural networks over traditional expert systems. When NN is used in medical diagnosis they are not affected by factors such as human fatigue, emotional states and habituation. They are capable of rapid identification, analyses of conditions, and diagnosis in real time. With the spread of Neural Networks in almost all fields of science and engineering, it has found extensive application in biomedical engineering field also. The applications of neural networks in biomedical computing are numerous. Various applications of ANN techniques in medical field like medical expert system, cardiology, neurology, rheumatology, mammography and pulmonology were studied. In this study medical data related to Breast Cancer is considered for classification purpose to identify the disease.

Chapter 2

2. Related Works

2.1 Existing System

In 2007, the University Sains Malaysia and University Malaysia Perlis collaborated to collect data similar to Wisconsin's. However, their data related to diagnosing pre-cancerous stages of breast cancer. This data was utilized to build diagnostic programs. Seven implementations were tested and results proved encouraging; accuracy attained above 75.38% the purely AI based programs did not fare as well as conventionally programmed implementations, but showed promise.

With the success of Malaysia's programs when working with raw data, it is time that the Wisconsin set is revisited. Most of the trials using Wolberg's data were performed in the 1990s, and with the improvements to modern technology, successful neural network diagnostics while working with raw data could be a reality.

2.2 The Origins of Data

In the early 1990's, Researchers at the University of Wisconsin recognized the relevance of improving breast cancer diagnostics. Under the supervision of Dr. William Wolberg, 699 patients had fine needle aspirates (FNAs) of their breast masses. FNAs are tests that stick a needle into a mass to extract cells from the mass for observation. These masses had been biopsied via fine needle aspirates. Doctors then rated nine inputs such as clump thickness, uniformity of cell shape, bland chromatin and mitoses on a scale of one to ten, one being indicative of a benign mass and ten being indicative of a malignant tumor, before they definitively determined the diagnosis of the mass. These results were published on the University of California Irvine's (UCI) Machine Learning Repository for public use.

Chapter 3

3. System Overview

3.1 Requirements: The Big Data

The core of the project is the University of Wisconsin Original Breast Cancer Data Set. An optimized version of that dataset was created to improve the performance of the application during initialization. The contents of the data did not change; however, the way it was represented did only in case of specifying whether it belongs to the benign or malignant class.

3.2 System Specification/Materials Required

Laptop or Computer meeting minimum technical specifications:

CPU: 2.8 GHz Intel Pentium/ Core2Duo

OS: Windows 7 or 8/ Linux

HDD: Free space 4 GB for Program data

Having Data obtained from: University of Wisconsin Breast Cancer Data Set

IDE

Eclipse Juno, DevC++ (Windows), GCC compiler (Linux)

Deployed Programming language

C++, Several benchmarks are performed over Java vs. C++_μ to check which is best among performing operations over operands and various functions and in the end results states C++ is good at handling computations over integral values, floating point values, handling iterative loops and memory management. So we equipped C++.

μ: <http://www.developer.com/java/article.php/3856906/Java-vs-C-The-Performance-Showdown.htm>

3.3 Feasibility Study

Feasibility study is a process to check possibilities of system development. It is a method to check various different requirements and availability of financial & technical resources.

Before starting the process various parameters must be checked like:

- Estimated finance is there or not?
- The man power to operate the system is there or not?
- The man power is trained or not?
- All the above conditions must be satisfied to start the project.

There are three different ways feasibility can be tested

- 1) Economical Feasibility
- 2) Technical Feasibility
- 3) Operational Feasibility.

Economical Feasibility:

In economical feasibility, analysis of the cost of the system is carried out. In addition if he/she wants to see archives of particular equity then he has to refer to all the old newspapers. For research reports he has to buy another magazine. So Instead of buying no of magazines user has to just go online and with a single click he can get whatever information he wants. So our project of online share news passes the test of economical feasibility.

Technical Feasibility:

It is basically used to see existing computer, hardware and software etc., weather it is sufficient or additional equipment's are required?

Minimum System Requirement is such that it can be affordable by of the user who is having computer. All the user requires is compatible browser and some development packages to be installed so our system is fully technical feasible.

Operational Feasibility:

Once the system is designed there must be trained and expert operator. If there are not trained they should given training according to the needs of the system. From the user's perspective our system fully operational feasible as it just requires some knowledge of computer. Operators only need add daily prices of various equities and there are enough validations available so operator does not require any special technical knowledge. So our system also passes the test of operational feasibility.

Chapter 4

4. Theoretical Foundations: The Engineering Model

Introduction

Originally inspired by biological models of mammalian brains, ANN has emerged as a powerful technique for data analysis. Neural Networks consists of compositions of single, nonlinear processing units that are organized in a densely inter connected graph. A set of parameters, called weights, are assigned to each of the edges of the graph. These parameters are adapted through the local interactions of processing units in the network. By repeatedly adjusting these parameters, the neural network is able to construct a representation of a given data set. This adaptation process is known as training. Neural Network is able to solve highly complex problems due to the nonlinear processing capabilities of its neurons. In addition, the inherent modularity of the neural network structure makes it adaptable to a wide range of applications

4.1 Details of Proposed Model

Controlled Variables

The neural network should adhere to the following standards:

- Each trial shall use a different set of randomly selected instances for testing.
- Neural network will use the University of Wisconsin Original Breast Cancer Database.
- All trials will be run on the same computer.
- The number of iterations for simulator is user defined at runtime of the application

Test Variable

The level of diagnostic success achieved by the networks.

Abridged versions of the procedures for this project are detailed below.

Experiment Phase: 1

1. Understand significance of each of 9 inputs: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and marginal adhesion.

Table 1: Understanding Significance of each of 9 input units of data

Input	Description
Clump Thickness	Assesses if cells are mono- or multi-layered.
Uniformity of Cell Size	Evaluates the consistency in size of the cells in the sample.
Uniformity of Cell Shape	Estimates the equality of cell shapes and identifies marginal variances.
Marginal Adhesion	Quantifies how much cells on the outside of the epithelial tend to stick together.
Single Epithelial Cell Size	Relates to cell uniformity, determines if epithelial cells are significantly enlarged.
Bare Nuclei	Calculates the proportion of the number of cells not surrounded by cytoplasm to those that are.
Bland Chromatin	Rates the uniform "texture" of the nucleus in a range from fine to coarse.
Normal Nucleoli	Determines whether the nucleoli are small and barely visible or larger, more visible, and more plentiful.
Mitoses	Describes the level of mitotic (cell reproduction) activity.

2. Architect and optimize a neural network model, each network should adhere to the following control variable standards:

- Reserve some instances for testing. In other words, those instances should not be used for training.
- Each trial shall use a different set of randomly selected instances for testing.
- The network should be optimized to yield best results.

3. Analyze results to determine success and failures of implementations.
4. Determine if neural networks were successful at predicting malignant versus benign.
5. Identify areas for improvement.

Experiment Phase: 2 Deployments of the Algorithms

1. Design a pseudo code for breast cancer neural network, including the following algorithmic components:

- **Artificial/Logical Input Layer.** Convert inputs to binary inputs to simulate the on/off firing of neurons.
- **Sigmoid Function.** A logistic function that removes the linearity from processing.
- **Summation Function.** Matrix math function to propagate neural firings through the network.
- **Step Function.** Incorporate malignancy weightings.

Inconclusive Assessment. Evaluate data elements classes using trained networks.

2. Define a neural network model with artificial input layer.
3. Identify a way to weight malignant false negatives higher.
4. Implement a neural network in C++
5. Implement logic that allows the network to rule masses inconclusive.

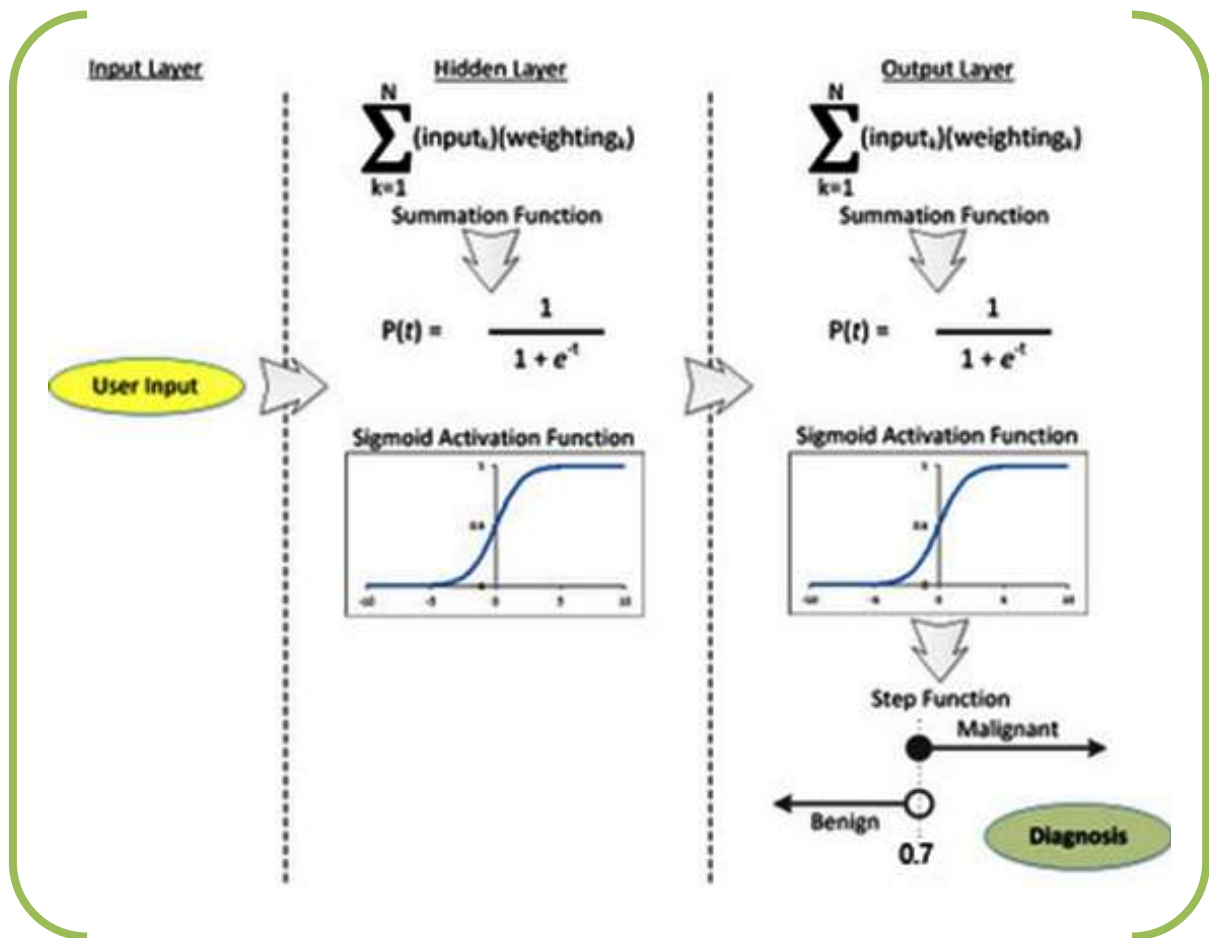


Figure 4.1 The Algorithmic Workflow

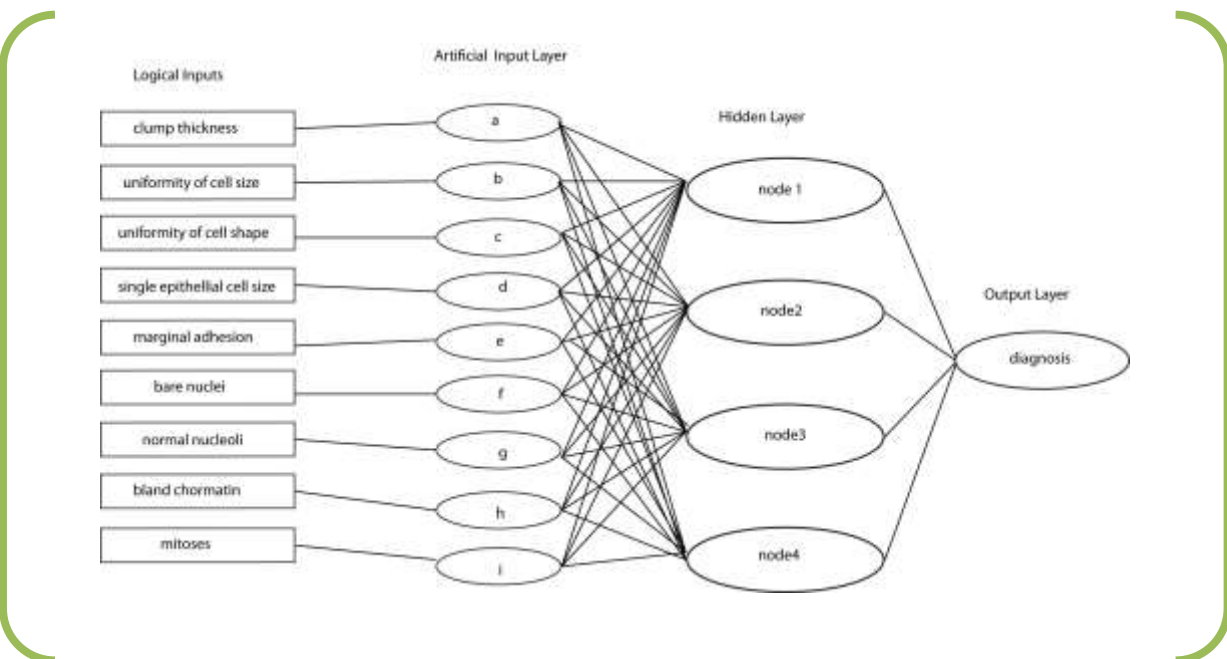


Figure 4.2 Logical ANN model

4.3 Training the Neural Network

In this experiment the neural network is trained with Breast Cancer database by using feed forward neural network model and backpropagation learning algorithm with momentum and variable learning rate. The cancer database consists of 9 attributes. The input layer of the network consists of 9 neurons to represent each attribute as the cancer database consists of 9 attributes. The number of classes is 2, one Benign and another is Malignant.

So one neuron in the output layer is sufficient to represent these two classes. The description of the backpropagation algorithm is specified in the above is used to train the neural network during the training process. Several neural networks are constructed with and without hidden layers i.e., single and multi-layer networks and trained with cancer dataset.

4.4 Deployed WDBC database

Classification of Cancer Dataset

One of the leading causes of death of women is breast cancer. Mammography has been proved to be an effective diagnostic procedure for early detection of breast cancer. An important sign in its detection is the identification of micro calcification of mammograms, especially when they form clusters. In this experiment the medical data related to breast cancer is considered. This database was obtained from the university of Wisconsin hospital, Madison from Dr. William H. Wolberg. This is publicly available dataset in the Internet.

Descriptions of Database:

- Number of instances we had taken 683
- Number of attributes: 10 plus the class attribute
- Attributes 1 through 10 will be used to represent instances
- Each instance has one of 2 possible classes: benign or malignant
- Class distribution: Benign & Malignant

Attribute information:

Attribute Domain

1. ID number	sample_ID
2. Clump thickness	1-10
3. Uniformity of cell size	1-10
4. Uniformity of cell shape	1-10
5. Marginal adhesion	1-10
6. Single epithelial cell size	1-10
7. Bare nuclei	1-10
8. Bland chromatin	1-10
9. Normal nucleoli	1-10
10. Mitosis	1-10
11. Class	(2 for benign, 4 for malignant)

The Data elements are divided into 3 categories which are as follows:

- Training
- Testing
- Validation

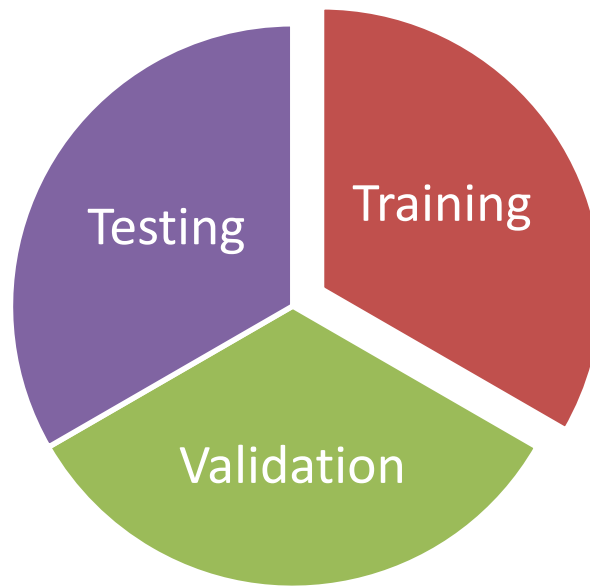


Fig.4.3 depicts the separation of data for various experiments to be performed on neural network

During training the neural network only training data is provided during evaluation the rest of the data elements are given to the network to figure out how much efficient the network is capable to identify the input units to their respective classes.

Chapter 5

5. System Analysis: System Model and System Architecture

5.1 SDLC paradigm

A software can be built using a top-down approach, a bottom-up approach, or a combination of both. The top-down approach starts with the overall design and planning. It is useful in cases where the technology is mature and well known, and where the business problems that must be solved are clear and well understood. The bottom-up approach starts with experiments and prototypes. This is useful in the early stage of business modeling and technology development. It allows an organization to move forward at considerably less expense and to evaluate the benefits of the technology before making significant commitments. In the combined approach, an organization can exploit the planned and strategic nature of the top-down approach while retaining the rapid implementation and opportunistic application of the bottom-up approach.

From the software engineering point of view, the design and construction of software may consist of the following steps: problem analysis, requirements study, planning, Application design, modules integration and testing, and finally deployment of the application.

Large software systems can be developed using two methodologies: the waterfall method or the spiral method. The waterfall method performs a structured and systematic analysis at each step before proceeding to the next, which is like a waterfall, falling from one step to the next. The spiral method involves the rapid generation of increasingly functional systems, with short intervals between successive releases.

For the deployment of our medical diagnosis application we had deployed the waterfall paradigm.

5.2 UML Diagrams/Specifications

- Structural Model - Class Diagram,
- Behavioral Model – Use Case Diagram, Activity Diagram and Sequence Diagram

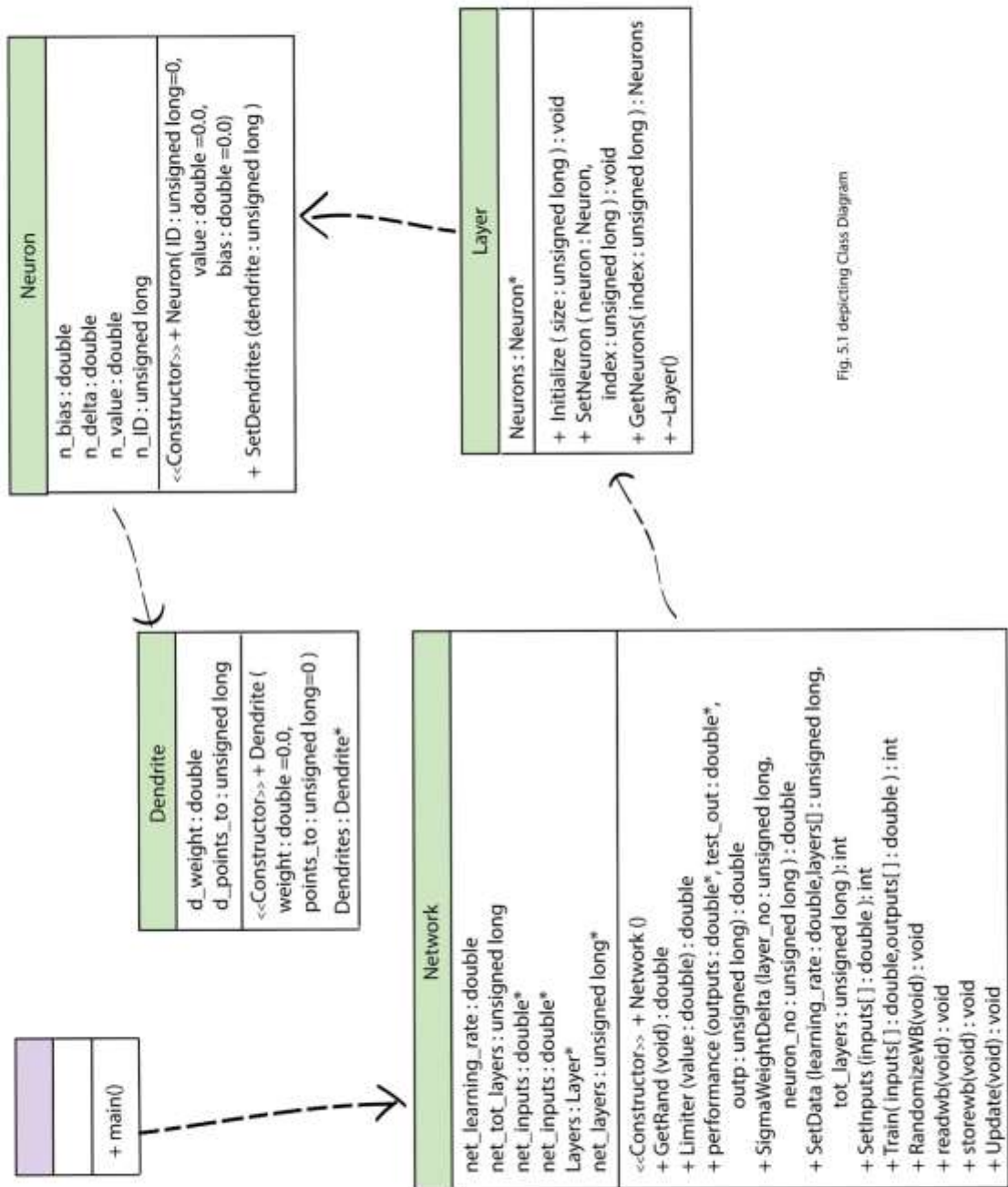


Fig. 5.1 depicting Class Diagram

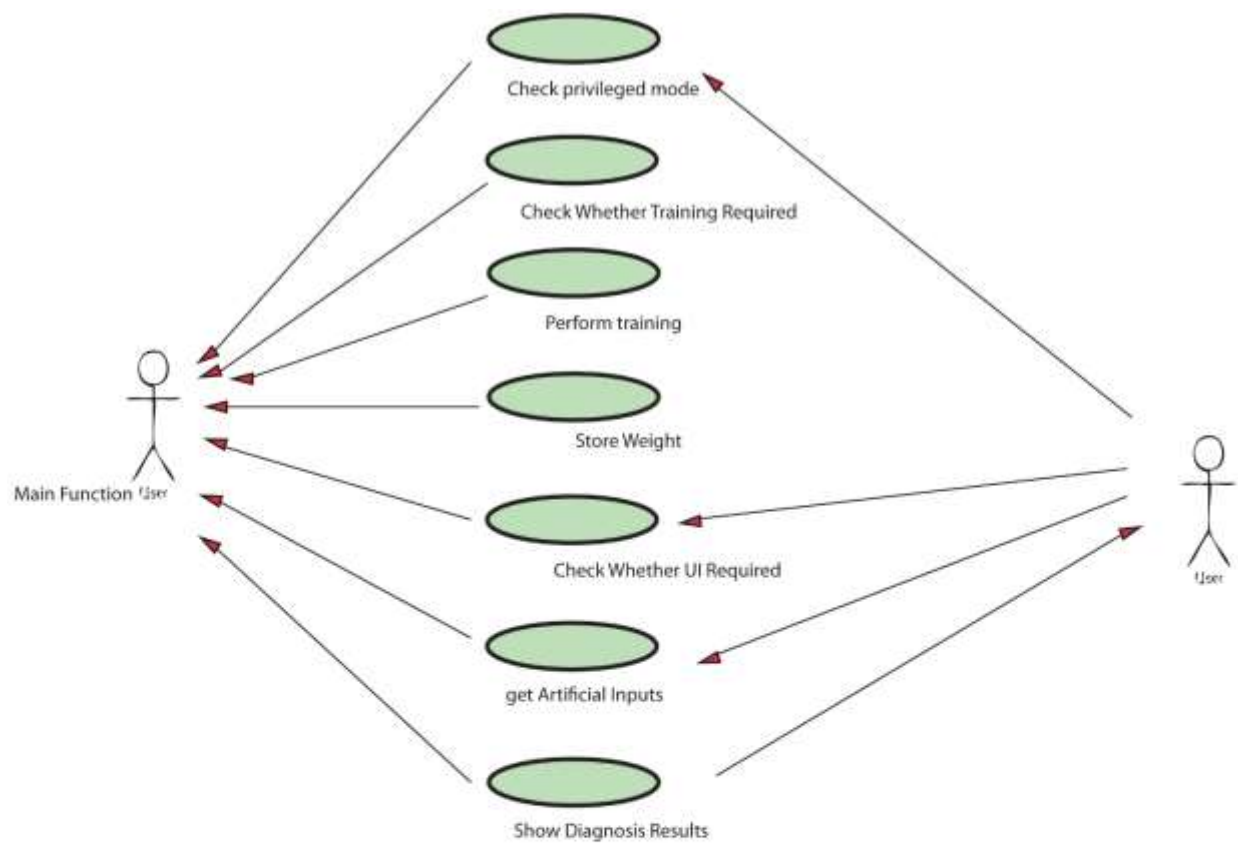


Fig. 5.2 Depicting Use Case Diagram

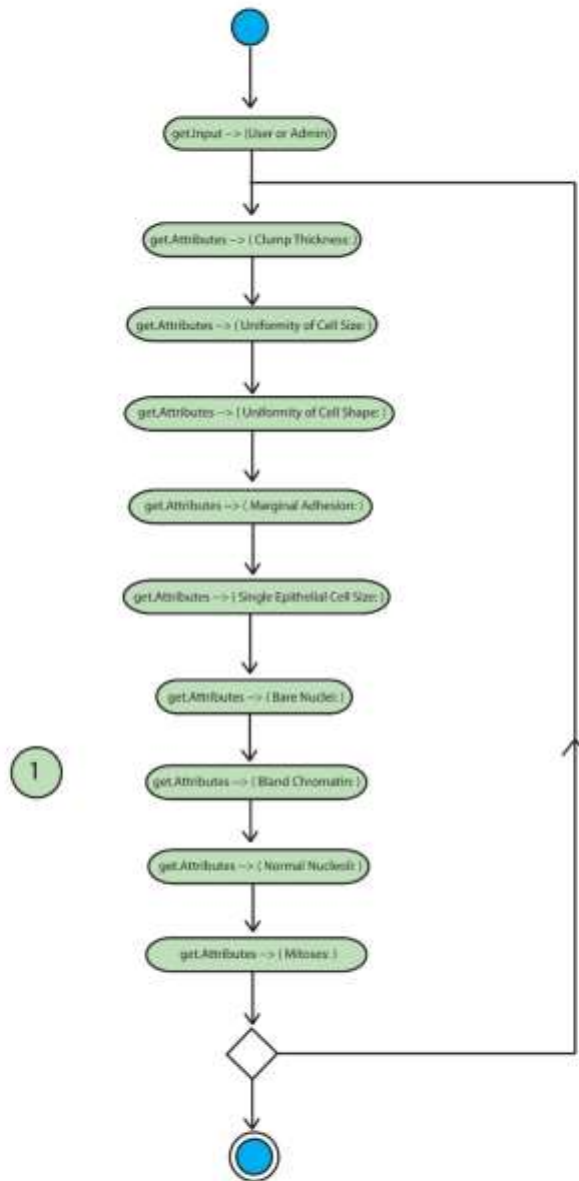


Fig 5.3 depicting Activity diagram at User mode

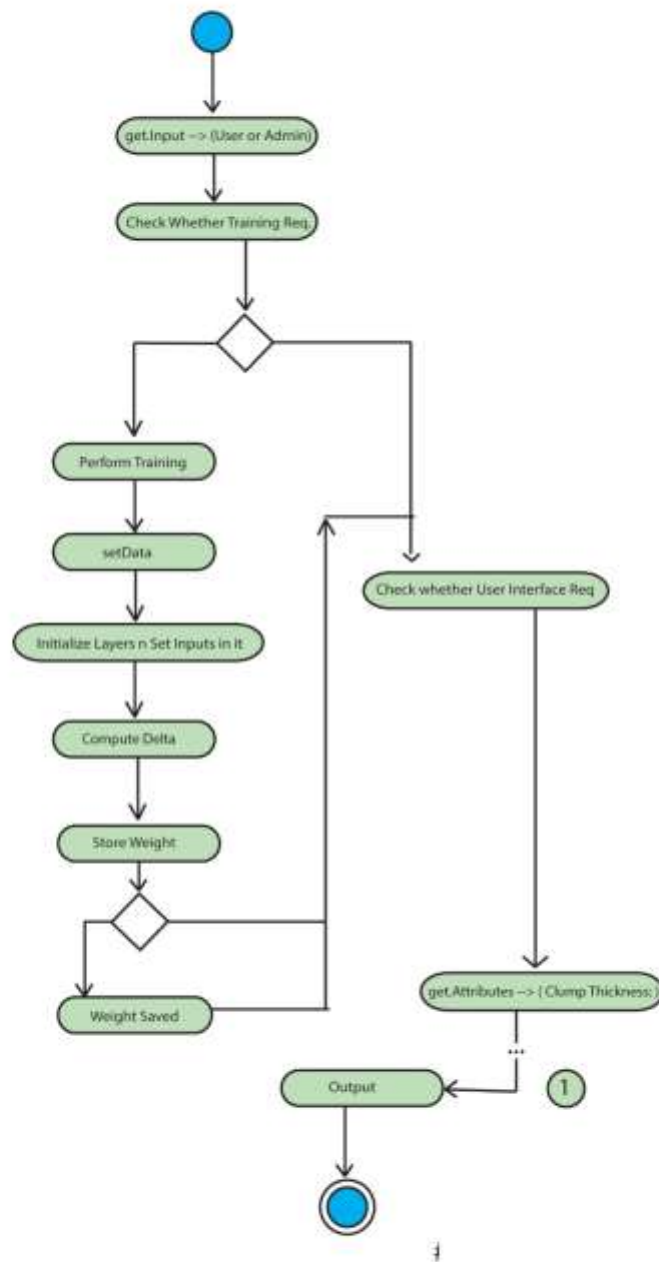
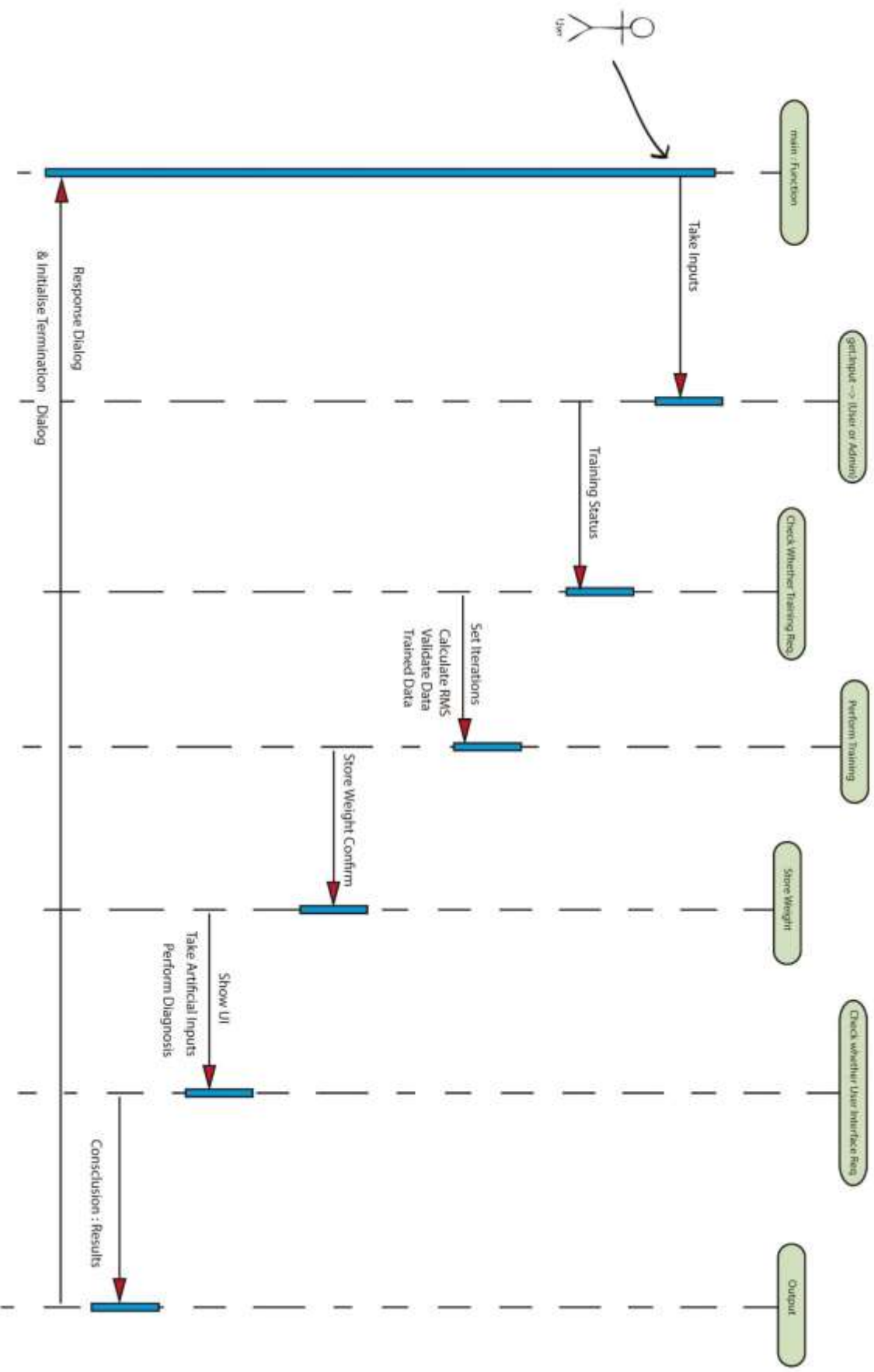


Fig 5.4 depicting Activity diagram at Admin mode

Fig. 5.5 depicting sequence diagram



Chapter 6

6. Implementations

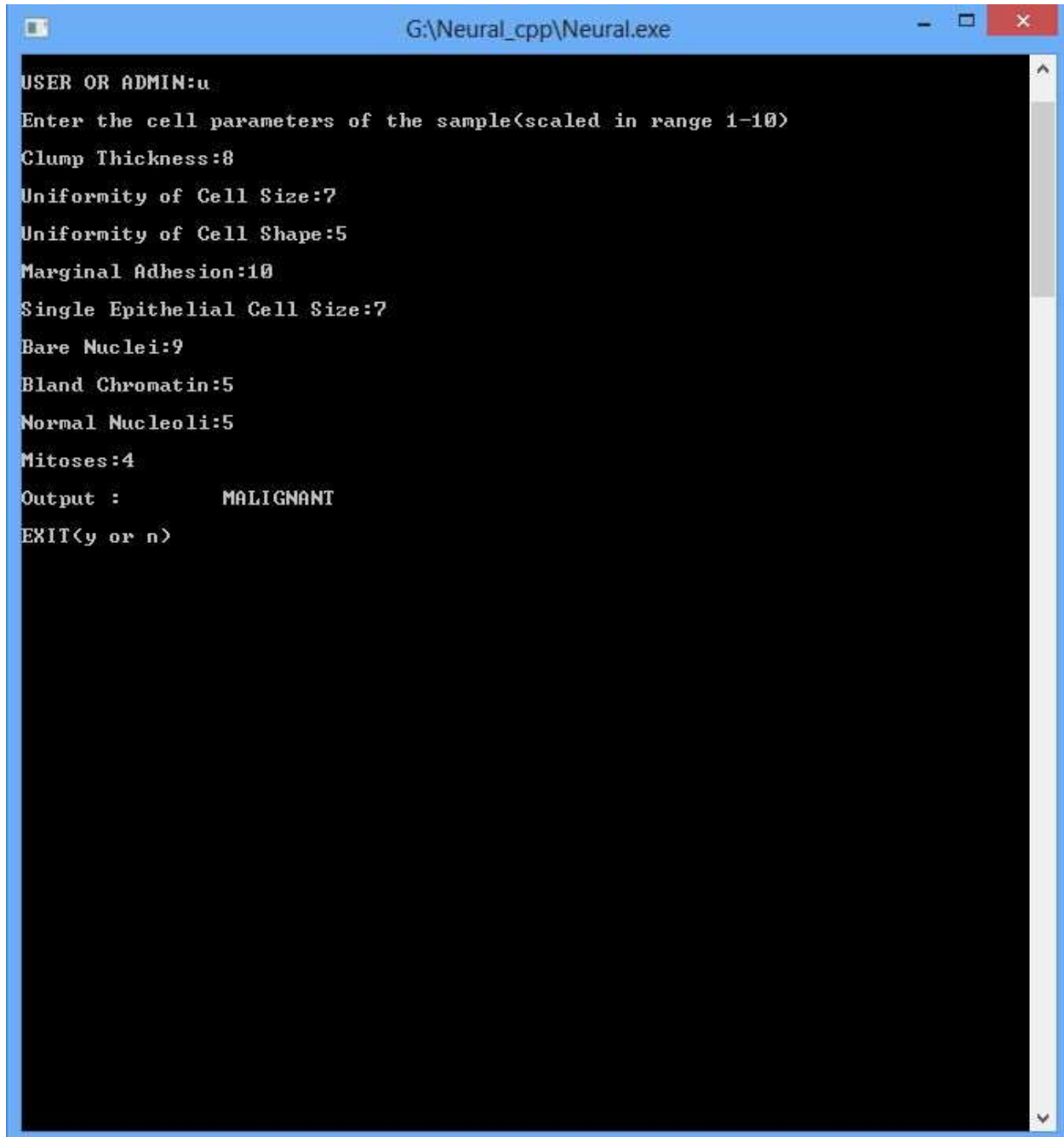
The implementation stage of software development is the process of converting a system specification into an executable system. It always involves processes of software design and programming but, if an incremental approach to development is used, may also involve refinement of the software specification.

6.1 Software development platform

We had employed the windows platform for development of our project application

DevC++ is an excellent IDE to work out best practices on C++ projects.

6.3 Snapshots



```
G:\Neural_cpp\Neural.exe

USER OR ADMIN:u
Enter the cell parameters of the sample(scaled in range 1-10)
Clump Thickness:8
Uniformity of Cell Size:7
Uniformity of Cell Shape:5
Marginal Adhesion:10
Single Epithelial Cell Size:7
Bare Nuclei:9
Bland Chromatin:5
Normal Nucleoli:5
Mitoses:4
Output :      MALIGNANT
EXIT(y or n)
```

Figure 6.1 depicts application snapshot in User mode

```
G:\Neural_cpp\New folder\Neural.exe
Neural Network parameters...
Total Layers:3
Input Layer Node:9
Hidden Layer Node:25
Output Layer Node:1
Learning Rate:0.1

USER OR ADMIN:a

Training required or not<enter y or n>:y

Enter number of training Iterations : 20

Starting Training...
Training : 1
Training : 2
Training : 3
Training : 4
Training : 5
Training : 6
Training : 7
Training : 8
Training : 9
Training : 10
Training : 11
Training : 12
Training : 13
Training : 14
Training : 15
Training : 16
Training : 17
Training : 18
Training : 19
Training : 20
Ending Training.

FALSE POSITIVE on training data:0
FALSSE NEGATIVE on training data:11
Root Mean Square(RMS) error of training data: 0.160376
Accuracy on training data:95.0893

FALSE POSITIVE on validation data:12
FALSSE NEGATIVE on validation data:7
Root Mean Square(RMS) error of validation data: 0.179235
Accuracy on validation data:90.6404
Weight storage required or not(y or n):
```

Figure 6.2 depicts application snapshot in admin mode

Chapter 7

7. Results & Conclusion

Following are the results which are observed during the overall performance of the system runtime

7.1 System Performance (Results)

Neural Network Performance Analysis

Neural Network is bound to various constraints such as number of hidden layers, number of nodes in hidden layer, learning rate these constraints are dealt simultaneously to obtain an efficient topology with small root mean square (RMS) error. Training is done to obtain a low RMS error with maximum upper bound iteration to avoid endless looping. It is observed that more complex the network higher the overfitting. Number of layer increased increases the over fitting. It is found that more complex network requires more training for better performance but if over trained then network start memorizing and performance is good for training data but degrades for validation data. Number of nodes in hidden layer is also varied to found effects on performance and shows negligible changes and it is hard to reach at any conclusion. After playing these network architectural constraints a most accurate architecture is obtained with 1 hidden layer (HL) and 25 hidden layer nodes (HLN). The most important parameter is learning rate. Learning rate has higher influence in comparison to network topology. Lower learning rate result better over fitting effect but higher learning rate result faster convergence.

Table 2 : PERFORMANCE STATISTICS

	Instances Taken	Wrong classifications	Accuracy
Training	224	0	100%
Validation	203	5	97.54%
Testing	256	20	92.19%
Total	683	25	96.34%

	Training	Validation
RMS Error	0.01305	0.09338

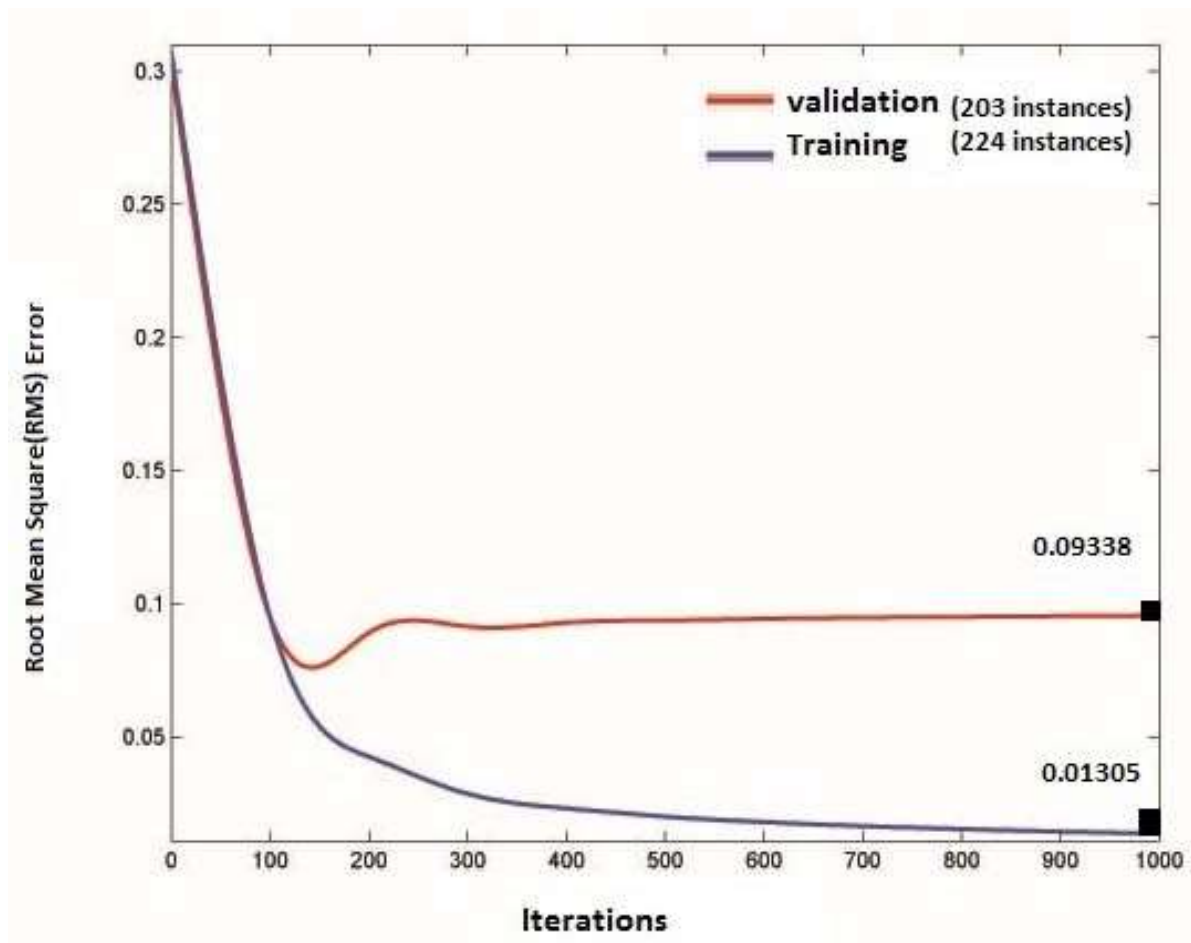


Fig 7.1 Show Overfitting Effect between training and validation data

Overfitting Effect.. μ
(HL=1, HLN=25, LR=0.1)

On completing the training the performance is recoded on training, validation and testing instances.

μ -> Overfitting Effect Refer to Appendix

Table 3 Statics table stating observation on each of 9 logical input

Domain	1	2	3	4	5	6	7	8	9	10	Sum
Clump Thickness	139	50	104	79	128	33	23	44	14	69	683
Uniformity of Cell Size	373	45	52	38	30	25	19	28	6	67	683
Uniformity of Cell Shape	346	58	53	43	32	29	30	27	7	58	683
Marginal Adhesion	393	58	58	33	23	21	13	25	4	55	683
Single Epithelial Cell Size	44	376	71	48	39	40	11	21	2	31	683
Bare Nuclei	402	30	28	19	30	4	8	21	9	132	683
Bare Nuclei	150	160	161	39	34	9	71	28	11	20	683
Normal Nucleoli	432	36	42	18	19	22	16	23	15	60	683
Mitoses	563	35	33	12	6	3	9	8	0	14	683
Sum	2843	850	605	333	346	192	207	233	77	516	

7.2 Conclusions

In this experiment, an artificial neural network was implemented to improve the success for classification and reduce the number of malignant false negatives. Through neural network implementation utilizing novel approaches including heavy malignant weighting, an inconclusive diagnosis option, and an artificial neural input layer, success was achieved. All aspects of the hypothesis were proven correct, specifically:

1. A neural network tuned to heavily weight malignant tendencies improves diagnostic results.
2. Increasing the number of training samples has a positive correlation to the success rate. The more samples collected, the more accurate the network will become.

With the developed artificial network, only 25 samples out of 683 samples were classified as malignant false negatives. Malignant false negatives are the most dangerous misdiagnosis because they are life-threatening. The neural network achieved predictive success of 96.34% with 99.41% sensitivity to malignancy.

Based on the combination of these findings, our application for Breast Cancer may be ready to diagnose actual patients. More global participation is required to confirm the findings and increase the predictive success on blind samples. The impressive results achieved when all data is included in training is a good omen for the potential of the neural network.

Chapter 8

8. Future Work : Changing the global health architecture

This neural network implementation could be leveraged to diagnose other forms of cancer or assist with evaluation of other areas of study as long as information can be clearly captured in numeric terms as inputs. There are already quality data sets for prostate cancer and ovarian cancer.

Looking forward, the cloud service could be marketed to collect more data globally and improve results. As is shown by the correlation between success and data samples, the network needs more samples to improve accuracy and reduce inconclusive diagnosis. The learning capability proved successful; neural networks can be crafted to prove useful for medical diagnostics.

Predicting the Global healthcare solution utilization in the Future :

International Classification of Diseases (ICD)

We'll develop; the International Classification of Diseases (ICD) is the standard diagnostic tool for epidemiology, health management and clinical purposes. This includes the analysis of the general health situation of population groups. It is used to monitor the incidence and prevalence of diseases and other health problems.

It is used to classify diseases and other health problems recorded on many types of health and vital records including death certificates and health records. In addition to enabling the storage and retrieval of diagnostic information for clinical, epidemiological and quality purposes, these records also provide the basis for the compilation of national mortality and morbidity statistics by WHO Member States. It is used for reimbursement and resource allocation decision-making by countries.

Further Enhancement to this Project

As Back Propagation Neural Network (BPNN) are commonly used in medical field the prototype model will be used as a classifier as well. However, BPNN has several issues and weaknesses to be addressed. The other main drawback regarding the use of artificial neural networks is the training problem. While training feedforward neural networks with the gradient-based backpropagation with momentum factor algorithm usually yields good results, there is no guarantee that the solution it reaches is optimum. Indeed, one of the problems with this method is that both convergence speed toward the solution and the possibility of reaching a solution depend on the choice of the learning rate and the proportionality constant of momentum.

Therefore, the architecture will be enhanced with Genetic Algorithm (GA) technology. By combining the GA and BPNN, a faster classifier model can be developed, without downgrading the classification performance.

References

1. Neural Networks in Artificial Intelligence by Limin Fu
2. "Final Year Project Handbook" by Prof. David Vernon
3. WHO for providing the Global report and motivating
4. Sario J (2010). "Breast cancer in the young patient". *The American surgeon* 76 (12): 1397–1401. PMID 21265355.
5. US NIH: Male Breast Cancer.
6. "World Cancer Report". International Agency for Research on Cancer. 2008.
7. American Cancer Society Homepage, (20 July 2005))Citing Internet sources URL: <http://www.cancer.org>
8. Schuermann, Juergen (1996). *Pattern Classification: A Unified View of Statistical and Neural Approaches*. New York: Wiley. ISBN 0-471-13534-8.
9. Jain, Anil.K.; Duin, Robert.P.W.; Mao, Jianchang (2000). "Statistical pattern recognition: a review". *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22** (1): 4–37. doi:10.1109/34.824819.
10. ThinkQuest 2000 Internet Challenge(Bernard Willers, Sep Vrba) team C007395 <http://www.thinkquest.org/>.
11. Haykin, S. S., 1999; "Neural Networks: A Comprehensive Foundation," 2nd Edition, Upper Saddle River, N.J.: Prentice Hall.
12. R. Rojas. Neural Networks: a systematic introduction. Springer-Verlag, 1996
13. Sunghwan Sohn and Cihan H. Dagli. Ensemble of Evolving Neural Networks in classification. *Neural Processing Letters* 19: 191-203, Kulwer Publishers, 2004
14. M. McInerney, and A.P. Dhawan., Use of Genetic Algorithms with Backpropagation in Training of Feedforward Neural Networks, *Proceeding of IEEE International Conference on Neural Network*, 1993,
15. D.W. Ruck., S.K. Rogers., M. Kabrisky., P.S Meibeck., and M.E. Oxley., Comparatives Analysis of Backpropagation & the Extended Kalman Filter for Training Multilayer Perceptrons, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, June 1992, Vol 14, No 6, pg 686-691

16. Dr. William H. Wolberg, (no date), Breast Cancer Wisconsin Dataset (online) (<http://www.radwin.org/michael/projects/learning/about-breast-cancer-wisconsin.html>) (1 July 2005)
17. Application of neural networks in medicine - a review (Kornel Papik¹, Bela Molnar¹, Rainer Schaefer², Zalan Dombovari¹, Zsolt Tulassay¹, Janos Feher¹)
18. Dr. A. Kandaswamy, Applications of Artificial Neural Networks in Bio Medical Engineering. The Institute of Electronics and Telecommunication Engineers, Proceedings of the Zonal Seminar on Neural Networks, Nov 20-21, 1997.
19. George Cybenko. Neural Networks in Computational Science and Engineering. IEEE Computational Science and Engineering, 1996, pp.36-42.

Appendix

Biopsy

A biopsy is done when other tests show that you might have breast cancer. The only way to know for sure is for you to have a biopsy. During this test, cells from the area of concern are removed so they can be studied in the lab. There are several kinds of biopsies. The doctor will use the one best for you.

Types of biopsies

Fine needle aspiration (FNA) biopsy: For this test, a very thin (fine), hollow needle is used to pull out fluid or tissue from the lump. The needle used in an FNA is thinner than the one used for blood tests. Your doctor might use ultrasound to guide the needle into the lump. Medicine may be used to make the skin numb. If the fluid drawn out is clear, the lump is most likely a benign cyst (not cancer). Bloody or cloudy fluid can mean either a cyst or, very rarely, cancer. If the lump is solid, small pieces of tissue are taken out. These will be looked at under a microscope to see if they are cancer.

An FNA biopsy is the easiest type of biopsy to have, but it has some downsides. It can sometimes miss a cancer if the needle is not placed among the cancer cells. And even if cancer cells are found, it is usually not possible to tell whether the cancer is invasive.

Mammograms

A mammogram is an x-ray of the breast. A screening mammogram is used to look for breast disease in women who do not seem to have breast problems. A mammogram can also be used when women have symptoms such as a lump, skin change, or nipple discharge. This is called a diagnostic mammogram. Diagnostic mammograms are also used to follow up abnormal screening mammograms.

Mammograms often don't work as well in younger women, mostly because their breasts are dense and this can hide a tumor. This is also true for pregnant women and women who are breast feeding. Since most breast cancers occur in older women, this is usually not a major problem. But it is a

problem for young women who have a genetic risk factor for breast cancer because they often get breast cancer at a younger age. For this reason, some doctors now suggest MRI along with mammograms for screening these women.

Overfitting

In statistics and machine learning, overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship. Overfitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model which has been overfit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data.

The possibility of overfitting exists because the criterion used for training the model is not the same as the criterion used to judge the efficacy of a model. In particular, a model is typically trained by maximizing its performance on some set of training data. However, its efficacy is determined not by its performance on the training data but by its ability to perform well on unseen data. Overfitting occurs when a model begins to memorize training data rather than learning to generalize from trend. As an extreme example, if the number of parameters is the same as or greater than the number of observations, a simple model or learning process can perfectly predict the training data simply by memorizing the training data in its entirety, but such a model will typically fail drastically when making predictions about new or unseen data, since the simple model has not learned to generalize at all.

Sigmoid Function

The activation levels of nodes can be discrete (e.g 0 and 1) or continuous across range (eg.[0,1]) or unrestricted. This depends on the activation (transfer) function chosen. If it is a hard limiting function, then the activation levels are 0 (or -1). A sigmoid function has activation levels limited to a continuous range of reals [0,1].

$$F(x) = \frac{1}{1+e^{-x}}$$

Step Function

In mathematics, a function on the real numbers is called a step function (or staircase function) if it can be written as a finite linear combination of indicator functions of intervals. Informally speaking, a step function is a piecewise constant function having only finitely many pieces.