Medical Diagnosis Using Data Mining

Submitted in partial fulfillment of the requirements
of the degree of

Bachelors of Engineering

by

Asad Siddiqui (62)
Huzaifa Vakil (68)
Sohel Tharani (67)
Zain Momin (33)

Guide:
Dr. Anupam Choudhary

Computer Engineering Department
Rizvi College of Engineering

University of Mumbai

2018-2019

CERTIFICATE

This is to certify that the project entitled **"Medical Diagnosis Using Data Mining"** is a bonafide work of **"Asad Siddiqui(62), Huzaifa Vakil(68), Sohel Tharani(67), Zain Momin(33)"** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **"Bachelor of Engineering"** in **"Computer Engineering"**.

(Dr. Anupam Choudhary)
Guide

(Prof. Shiburaj Pappu)
Head of Department

(Dr. Varsha Shah)
Principal

Project Report Approval for B.E.

This Project report entitled ***Medical Diagnosis Using Data Mining*** by ***Asad Siddiqui, Huzaifa Vakil, Sohel Tharani, Zain Momin*** is approved for the degree of Bachelor of Computer Engineering.

Examiners

1.---

2.---

Guide

Date: 1.---

Place: 2.---

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: ---------------------------------------
(Asad Siddiqui, Roll No. 62)

Date: ---------------------------------------
(Huzaifa Vakil, Roll No. 68)

Date: ---------------------------------------
(Sohel Tharani, Roll No. 67)

Date: ---------------------------------------
(Zain Momin, Roll No. 33)

Abstract

Data mining is the method of discovering patterns in huge datasets. These patterns are then used to make decisions and also to find future trends. Data mining techniques have been widely used in industries such as finance, retail and telecommunication. However, data mining techniques have not been as widely used in clinical decision support systems for prediction and diagnosis of various diseases. There are huge amounts of medical data that has just been stored hospital and clinic databases that aren't being used. Thus, the reason we have proposed this system is to find a way to create trends between stored data and to use these trends to help medical professions make a diagnosis in an efficient, quick and accurate manner.

Neural Networks are one amongst many data mining analytical tools that can be utilized to form predictions for medical data. Neural Networks are modeled as a circuit of neurons. The connections of the neurons are modeled as weights. These weights can have a positive or negative value; a positive value shows an excitatory connection whereas a negative value shows an inhibitory connection. An activity known as linear combination is performed where, inputs are weighted and then added. Then in the last step an activation function is applied which regulates magnitude of output. Back propagation algorithm is a method used to train the Neural Network. It does this training by computing certain error value at output and distributing it backwards through networks layers.

Genetic algorithms are a search heuristic that has been motivated by Charles Darwin's theory of natural evolution. The algorithm functions in a way such that only fittest individuals are selected for reproduction. The algorithm repeatedly modifies a population of individual solutions. At every step, the genetic algorithm selects individuals at random from the current population to be parents and uses them to produce the children for the next generation. Over sequential generations, the population "evolves" toward an optimal solution. This is a form of optimization so as to choose the weights that satisfy certain conditions. Thus, Genetic algorithms are used for optimization.

One of the most important applications of systems is in diagnosis of heart diseases because it is one of the leading causes of deaths all over the world. Almost all systems that predict heart diseases or diabetes use clinical dataset having parameters and inputs from complex tests conducted in labs. The proposed system predicts heart diseases based on risk factors like age, family history, diabetes, hypertension, high cholesterol, tobacco, smoking, alcohol intake, obesity or physical inactivity, etc.

The proposed system presents a method for prediction of heart disease and diabetic patients using major risk factors. This system incorporates two of the most successful data mining tools, neural networks and genetic algorithms. The proposed system will be implemented as a web-based application, where user will give answers to the predefined questions. The system will retrieve the data from stored database collection and compares the user values with trained data set using Multilayer perceptron Neural Network. Back propagation algorithm will be used to train the network using the weights optimized by Genetic algorithm.

Keywords: Data Mining, Neural Network, Genetic Algorithm, Back Propagation Algorithm.

Index

List of Figures

List of Tables

Chapter 1

Introduction

The extraction of patterns that are hidden and relationships from databases that are huge in size, Data mining joins statistical analysis, machine learning and databases. In several areas of medical services, that includes effectiveness of surgical procedures, medical tests, all of the medication and the relations that are present in clinical and diagnosis data, in all of these, data mining is applies.

Medical diagnosis is a process that is very complex and requires many years of experience in that expertise. Heart attack is one of the main causes of death throughout the world. Hence there are huge databases pertaining to the causes of heart attack, symptoms of it and hence it gives a huge opportunity for data science and machine learning to hop on to this opportunity and make predictions and inferences hence to arrive on a conclusion that can help further people suffering from heart diseases or those people that are showing symptoms pertaining to heart disease. Most of the times symptoms also include other side effects such as discomfort in the shoulders, back pain, shortness of breath, nausea, light headedness.

We have proposed the classification of heart patients in which our system will make use of machine learning and find significant patterns in heart patients to help them in further treatment. We have applied intervals with equal binning all over the data we have made use of the neuro genetic network using back propagation algorithms. Pre-Processing of the data is done to make the data efficient and useful to make predictions and inferences.

Our idea was inspired by Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg that made use of Neural Network and Genetic Algorithm. The study is improved by us by using hybrid system in data mining. The objective of our Proposed system is to build an intelligent system that is focused on heart diseases using previous databases. 13 attributes were proposed but in the end it was deemed necessary to include only those factors that actually contributed to the heart disease conclusions incorporating optimal model construction time.

Chapter 2

Literature Survey

Data mining techniques have been widely used in clinical decision support systems for prediction and diagnosis of various diseases with good accuracy. These techniques have been very effective in designing clinical support systems because of their ability to discover hidden patterns and relationships in medical data. One of the most important applications of such systems is in diagnosis of heart diseases because it is one of the leading causes of deaths all over the world. Almost all systems that predict heart diseases use clinical dataset having parameters and inputs from complex tests conducted in labs.

2.1 Paper 1

"Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors"

In Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg paper on "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors"; they explained how data can be used for prediction of heart diseases, based on blood pressure, smoking habit, cholesterol and blood pressure levels, diabetes. Life style risk factors which include eating habits, physical inactivity, smoking, alcohol intake, obesity are also associated with the major heart disease risk factors and heart disease [1,2].There are studies showing that reducing these risk factors for heart disease can actually help in preventing heart diseases [3]. Researchers have used these prediction algorithms in adapted form of simplified score sheets that allow patients to calculate the risk of heart diseases [4]. The Framingham Risk Score (FRS) is a popular risk prediction criterion which is used in algorithms for heart disease prediction [5]. Researchers are using data mining techniques for the diagnosis of many diseases such as heart disease, stroke and cancer and many data mining techniques have been used in the diagnosis of heart disease with good accuracy. They have been applying different data mining techniques such as naïve bayes, neural network, decision tree, bagging, kernel density, and support vector machine[6]-[8] for prediction and diagnosis of heart diseases. The system [9] which uses neural based learning classifier for classifying data mining tasks showed that neural based learning classifier system performs equivalently to supervise learning classifier. This system requires a Global Optimization Toolbox and the Neural Network Toolbox got implementing the algorithm [10]. ANN is initialized with the 'configure' function, with each weigh between -1.0 to +1.0. These weights are then passed to the genetic algorithm which uses the mean square error as the fitness function. The interconnecting weights and thresholds of the trained neural network are passed to the genetic algorithm. The fitness function used is mean square error (mse) which is calculated as below:

$$mse = \sum_k (O_k - T_k)^2 / n \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(1)}$$

Accuracy can be calculated as:

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN) \qquad \ldots\ldots\ldots\ldots\ldots\ldots.. \text{(2)}$$

Where TP, FP, TN & FN denotes True Positives, False Positives, True Negative & False Negatives respectively.

2.2 Paper 2

"Heart Disease Prediction System Using Weight Optimized Neural Network"

T.Manju, K.Priya, R.Chitra stated in "Heart Disease Prediction System Using Weight Optimized Neural Network" that to extract hidden patterns and relationships from large databases, Data mining merges statistical analysis, machine learning and database technology. In several areas of medical services, including prediction of effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data, Data Mining techniques have been applied. Medical diagnosis is a very composite process, entailing precise patient data, a philosophical understanding of the medical literature and many years of clinical experience [11]. This paper also explains that how the neural network is trained with Heart Disease Data Set by using Feed Forward Neural Network Model and Back Propagation Learning Algorithm with parameter as weight.

2.3 Paper 3

"Review of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques"

In the paper "Review of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques", R. Chitra and V. Seenivasagam stated that there are many data mining algorithms that can predict heart disease but accuracy and time taken may vary in different algorithms. So by creating a system with less accuracy would create misunderstandings about the heart disease. And the system would fail to predict the disease. The Table 1. In this paper shows the accuracy and time taken by using respected algorithms.

The Algorithm Used	**Accuracy**	**Time Taken**
Naïve Bayes	52.33%	609ms
Decision List	52%	719ms
K-NN	45.67%	1000ms

Table 2.1: Performance Study of Data Mining Algorithms

In this paper, the system presented by Latha Parthiban and R. Subramanian [12] shows that using a hybrid system like Neuro-Fuzzy has good training performance and classification accuracies.

Also, Olatubosun Olabode et al. [13] classifies the Cerebrovascular disease by using artificial neural network with back propagation error method. The Multi-layer perceptrons artificial neural networks with back-propagation error method were feed-forward nets with one or more layers of nodes between the input and output nodes. These additional layers contain hidden units or nodes that were not directly connected to both the input and output nodes. The neural network was trained using back propagation algorithm with sigmoid function on one hidden layer with the 16 input attributes. This type of system would give an accuracy of more than 85%.

Chapter 3

Problem Statement

A prediction system helps people who do not have sufficient expertise to evaluate whatever the patient has heart disease or not based on the data that is feeded to the system. It provides patient and doctor with information to help them decide what will be the further course of action. The proposed work is different from existing traditional based doctor's consultation systems since the existing only considers the doctors recommending the medicines based on the knowledge of the said doctor. Doctor can't recommend things that are not in his/her knowledge domain. The proposed system uses combination of machine learning and data mining. Machine learning model is used for finding similarity between patient's data which would help the system for prediction and data mining is used for getting to said prediction by making patterns in the feeded data. Then it uses prediction to tell the doctor or the patient that based on the history, whatever the patient has heart disease or not using the famous genetic algorithm. Thus, the use of both methods can help to manage data conceived prediction problem and traditional medical approach in our proposed project.

Chapter 4

Report on Present Investigation

4.1 Theory:

Neural network learning is a type of supervised learning technique where the network is learned using known output.

4.1.1 General Structure of MLFFN:

A neural network consists of layers of interconnected artificial neurons, as shown in Figure 1. A neuron in a neural network is sometimes called a “node” or unit; all these terms mean the same thing and are interchangeable. A multilayer feed forward neural network consists of a layer of input units, one or more layers of hidden units, and one output layer of units. A neural network that has no hidden units is called a Perceptron. However, a Perceptron can only represent linear functions, so it isn’t powerful enough for the kinds of applications to be solved [14].

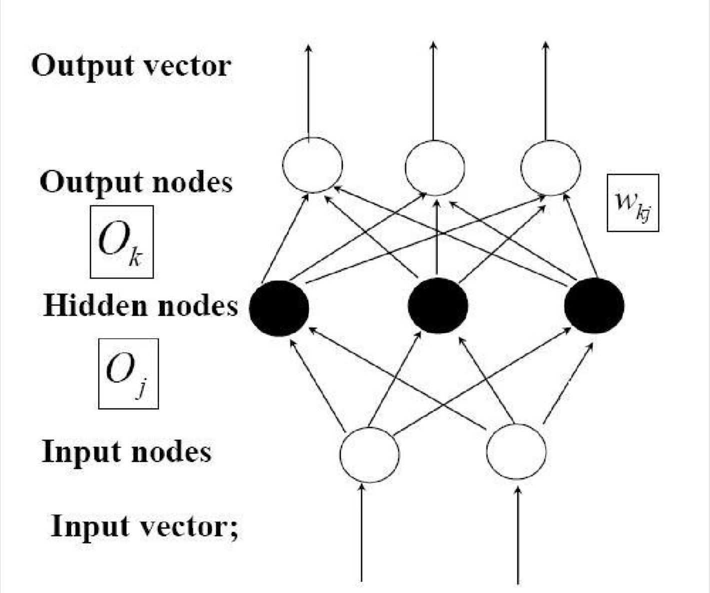


Fig 4.1: Typical structure of a Multilayer Feedforward Neural Network

This structure is called multilayer because it has a layer of processing units (i.e., the hidden units) in addition to the output units. These networks are called feed forward because the output from one layer of neurons feeds forward into the next layer of neurons. There are never any backward connections, and connections never skip a layer. Typically, the layers are fully connected, meaning that all units at one layer are connected with all units at the next layer. So, this means that all input units are connected to all the units in the layer of hidden units, and all the units in the hidden layer are connected to all the output units.

4.1.2 MLFFN Training and Propagation

In this work the neural network is trained with Heart Diseases dataset by using feed forward neural network model and backpropagation learning algorithm with the parameter as a weight.

The Back-Propagation Training Algorithm

Step 1: Initialization

Set all the weights and threshold levels of the network to random numbers uniformly distributed inside a small range:

$$\left(-\frac{2.4}{F_i}, +\frac{2.4}{F_i}\right)$$

where *Fi* is the total number of inputs of neuron *i* in the network. The weight initialization is done on a neuron-by- neuron basis.

Step 2: Activation

Activate the back-propagation neural network by applying inputs $x_1(p), x_2(p), x_n(p)$ and desired outputs $y_{d,1}(p), y_{d,2}(p), ..., y_{d,n}(p)$.

(a) Calculate the actual outputs of the neurons in the hidden layer:

$$y_j(p) = sigmoid\left[\sum_{i=1}^{n} x_i(p).w_{ij} - \theta_j\right]$$

where *n* is the number of inputs of neuron *j* in the hidden layer, and *sigmoid* is the *sigmoid* activation function.

(*b*) Calculate the actual outputs of the neurons in the output layer:

$$y_k(p) = sigmoid\left[\sum_{j=1}^{m} x_{jk}(p).w_{jk}(p) - \theta_k\right]$$

where m is the number of inputs of neuron k in the output layer.

Step 3: Weight training

Update the weights in the back-propagation network propagating backward the errors associated with output neurons.

(a) Calculate the error gradient for the neurons in the output layer:

$$\delta_k(p) = y_k(p).[1 - y_k(p)].e_k(p)$$

$$e_k(p) = y_{d,k}(p) - y_k(p)$$

Calculate the weight corrections:

$$\Delta w_{jk}(p) = \alpha.y_j(p).\delta_k(p)$$

Update the weights at the output neurons:

$$w_{jk}(p+1) = w_{jk}(p) + \Delta w_{jk}(p)$$

Step 4: Iteration

Increase iteration p by one, go back to *Step 2* and repeat the process until the selected error criterion is satisfied.

Learning

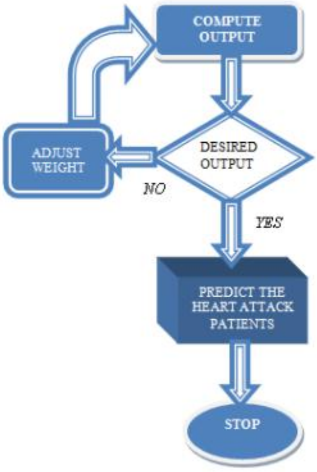


Fig 4.2: Learning of Weight Adjust

The algorithm used to train the network is the Backpropagation Algorithm [15]. The general idea with the backpropagation algorithm is to use gradient descent to update the weights so as to minimize the squared error between the network output values and the target output values. Then, each weight is adjusted, using gradient descent, according to its contribution to the error. It won't go into the actual derivations here, it can find that in your text and in other sources. This process occurs iteratively for each layer of the network, starting with the last set of weights, and working back towards the input layer hence the name backpropagation, with the output calculation, the weight update calculation depends on the type of problem we're trying to solve.

4.1.3. Genetic algorithm

Genetic Algorithm (GA) is an optimization technique inspired by natural selection and natural genetics. Its main advantage is that GA only uses the fitness function but not gradient or other attached information in the optimizing process. The GA's fitness function is used to estimate the individuals' optimization degree by optimizing computation. Those individuals who have much higher fitness will have more opportunities to be duplicated to the next generation [16].

Genetic Algorithm are Selection, Cross over, Mutation, Accepting.

a) Select two parents' chromosomes from a population according to their fitness.

b) Crossover with a crossover probability crossover of a parent to form new offspring is the exact copy of parents.

c) Mutation with a mutation probability mutates new offspring at each locus.

d) Accepting place new offspring in the new population.

4.1.4. Genetic Algorithm and Neural network

Genetic algorithm is an optimization algorithm that mimics the principles of natural genetics. It finds acceptably good solutions to problems acceptably quickly. Genetic algorithm have been used in, to reduce the actual data size to get the optimal subset of attributed sufficient for heart disease prediction. In many applications, knowledge that describes desired system behavior is contained in datasets.

When datasets contain knowledge about the system to be designed, a neural network promises a solution because it can train itself from the datasets. Neural networks are adaptive models for data analysis particularly suitable for handling nonlinear functions. By combining the optimization technique of genetic algorithm with the learning power of neural network, a model with better predictive accuracy can be derived.

4.2. Proposed Methodology:

4.2.1. Working principle

In a proposed system using data mining intelligent technique, in this system predict more accurately the presence of heart disease with reduced number of attributes. Heart Attack predicting system have the 13 risk factors, in this risk factors take 6 main factors and give the weights randomly high value, it will adjust the weight for the use of reduce error in the automated system for the prediction. Weight Optimized dataset will be processing on the neural network multi layer feed forward network.

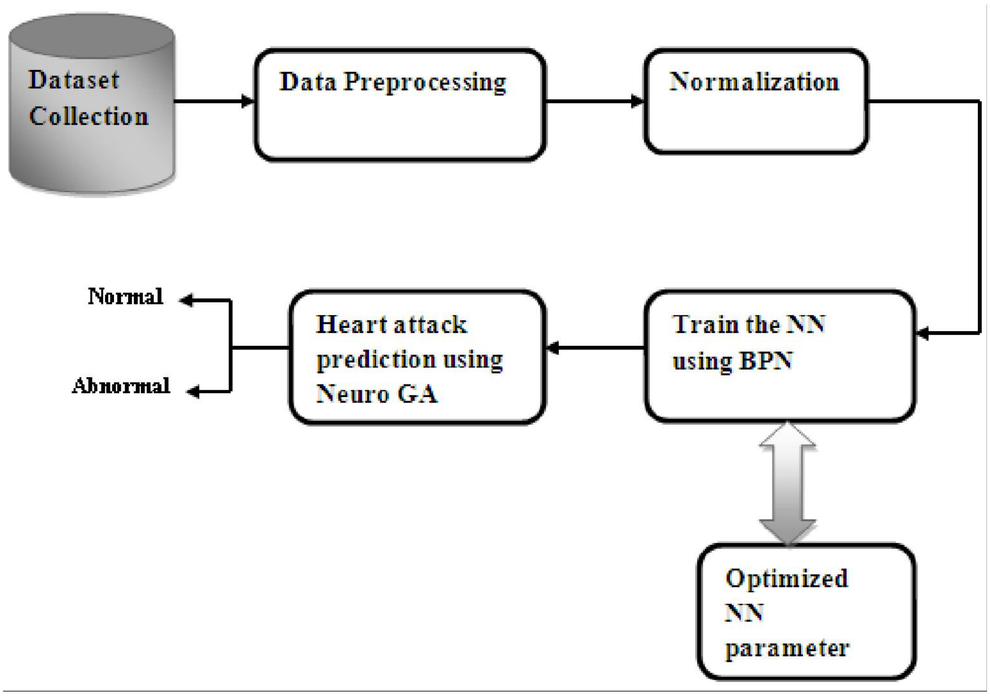


Fig 4.3: Proposed System for the Heart Attack Prediction

In the first step the heart disease consist of dataset which is collected from the medical institute. The dataset consists of patient information, patient history, Gene diagnosis disease database which contains the symptoms of Heart disease. Then it performs the preprocessing technique. So that noisy data, duplicate records, missing data, inconsistent data are removed. Then perform the normalization, in this work it perform the min-max normalization. So that negative values are removed. Then train the NN by using the BPN. Then predict the heart disease using the neuro genetic algorithm. Then diagnosis the disease whether the patient suffer from HD or not.

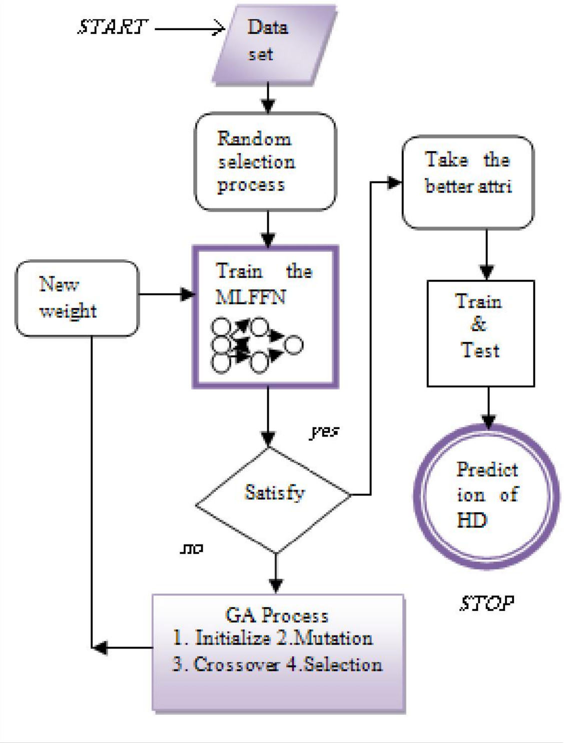


Fig 4.4: Flow chart-Multi Layer Feed Forward Neural Network Genetic Approach

The process of neural network with genetic algorithm presented below,

1. Initialize the process of predicting the heart attack.

2. Extract the risk factors of patient’s details.

3. Selection process start with assign the weight randomly to each attributes.

4.Each attributes given in the hidden layer of neural network and it will start process of training.

5.In this training adjust the value and choosing the absolute value.

6.Satisfy the training process yes it will go to the feature subset selection.

7. Do the testing function of exact prediction the heart attack patients.

8. Otherwise not satisfy the condition, in the approach will be genetic algorithm do the conditions and adjust the weight and return the step3.

9. Finally the result of patients, predict the heart attack via affect patients or not affect the patient.

Sample geneo type for weight initialization is:

$$\{w_{11}, w_{12}, \ldots., w_{1n}, w_{21}, w_{22}, \ldots., w_{2n}, w_{n1}, w_{n2}, \ldots, w_{nh}\}$$

Weights from input to hidden layer is $\{w_{11}, w_{12}, \ldots.., w_{nh}\}$

Weight from hidden to output layer is $\{w_{11}, w_{12}, \ldots.., w_{nm}\}$

The disease data ‘x’ of thr connection GA loop is

$$\{w_{11}, w_{12}, \ldots., w_{1n}, w_{21}, w_{22}, \ldots., w_{2n}, w_{n1}, w_{n2}, \ldots, w_{nm}, b_1, b_2\}$$

Where b1 is bias 1 and b2 is bias2.After the initialization stage, the genetic reproduction, crossover, mutation are applied to the output.

4.2.2. Proposed Algorithm

1. In the first step the heart disease consists of dataset which is collected from the medical institute and Initialize the process.
2. Extract the risk factors of patient’s details which consists of patient information, patient history, Gene diagnosis disease database which contains the symptoms of Heart disease.
3. Selection process start with assigning the weight randomly to each attribute.
4. Each attribute is given in the hidden layer of neural network and it will start process of training.
5. Compare the actual and desired output.
6. Calculate the error in each neuron.
7. Propagates this error by using the Backpropagation algorithm.
8. If the training process is satisfy, then it will go to the exact heart attack prediction of patients.
9. Otherwise it will be go to genetic algorithm adjust the weight and return to the step 3.
10. Finally, the result of patients will be predicting the heart attack via affect patients or not affect the patient.

4.2.3 Research Methodology

Attribute value to be taken into project for disease

Heart Disease Dataset Attributes

Sr. No.	**Risk Factors**	**Values**
1.	Age (Years)	20-34 (0) 35-50 (1) 51-60 (2) >60 (3)
2.	Sex	Male (0), Female (1)
3.	Blood Cholesterol	Below 200 mg/dL – Low (1) 200-239 mg/dL - Normal (2) 240 mg/dL – High (3)
4.	Blood Pressure	Below 120 mmHg – Low (1) 120 to 139 mmHg – Normal (2) Above 139 mmHg – High (3)
5.	BP Treatment	Yes (1) No (2)
6.	Hereditary	Yes (1) No (2)
7.	Smoking	Yes (1) No (2)

8.	Alcohol Intake	Yes (1) No (2)
9.	Physical Activity	Low (1) Normal (2) High (3)
10.	Diabetes	Yes (1) No (2)
11.	Diet	Poor (1) Normal (2) Good (3)
12.	Obesity	Yes (1) No (2)
13.	Stress	Yes (1) No (2)
Output	Heart Disease	Present (0) Not Present (1)

Table 4.1 Heart Disease Dataset Attributes

4.3 Feasibility Study

The very first phase in any system developing life cycle is preliminary investigation. The feasibility study is a major part of this phase. A measure of how beneficial or practical the development of any information system would be to the organization is the feasibility study. The feasibility of the development software can be studied in terms of the following aspects:

1.Operational Feasibility

2. Technical Feasibility

3. Economical Feasibility

4. Motivational Feasibility

5. Legal Feasibility

4.3.1 Operational Feasibility

The dataset will be maintained and dataset will be under manual human maintenance to assure clear records. Hence operational feasibility is assured.

4.3.2 Technical Feasibility

- Intel Compactible Processor
- At least 512MB RAM
- A mouse or other pointing device
- At least 250 MB free hard disk space

4.3.3 Economical Feasibility

Once the hardware and software requirements get fulfilled, there is no need for the user of our system to spend for any additional overhead. For the user, the project will be economically feasible in the following aspects:

- This project will reduce a lot of paper work. Hence the cost will be reduced.
- Our project will reduce the time that is wasted in laboratory tests.
- Our project will reduce the money that is wasted in laboratory tests.

4.3.4 Legal Feasibility

The licensed copy of the required software is quite cheap and easy to get. So, from legal point of view the proposed system is legally feasible.

4.4 System Development Life Cycle

The System Development Life Cycle is the process of developing information systems through investigation, analysis, design, implementation, and maintenance. The System Development

Life Cycle (SDLC) is also known as Information Systems Development or Application Development.

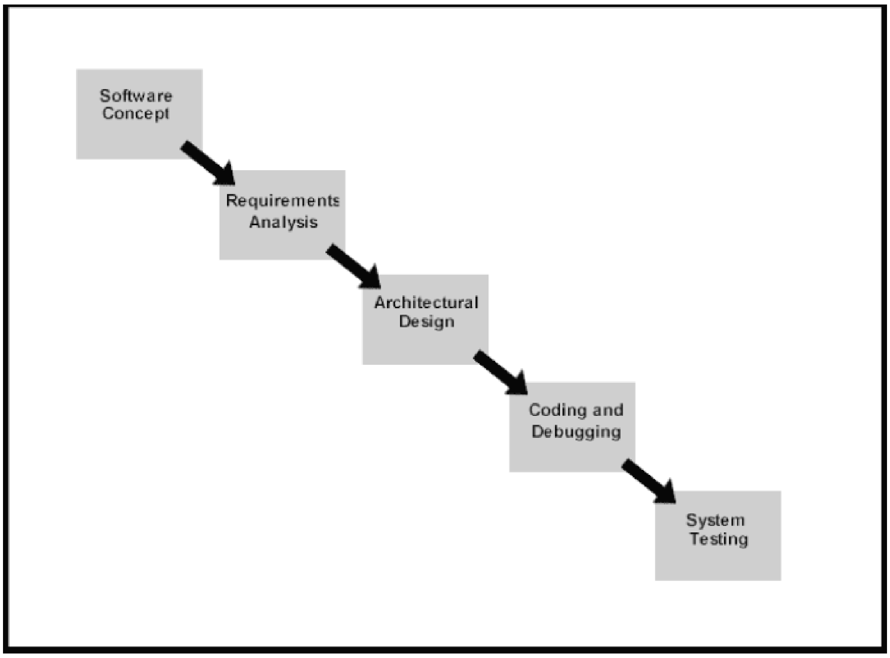


Fig 4.5: System Development Life Cycle

Below are the steps involved in the System Development Life Cycle. Each phase within the overall cycle may be made up of several steps.

4.4.1 Software Concept

The first step is to identify a need for the new system. This will include determining whether a business problem or opportunity exists, conducting a feasibility study to determine if the proposed solution is cost effective, and developing a project plan.

This process may involve end users who come up with an idea for improving their work. Ideally, the process occurs in tandem with a review of the organization's strategic plan to ensure that IT is being used to help the organization achieve its strategic objectives. Management may need to approve concept ideas before any money is budgeted for its development.

4.4.2 Requirements Analysis

Requirements analysis is the process of analyzing the information needs of the end users, the organizational environment, and any system presently being used, developing the functional requirements of a system that can meet the needs of the users. Also, the requirements should be recorded in a document, email, user interface storyboard, executable prototype, or some other form. The requirements documentation should be referred to throughout the rest of the system development process to ensure the developing project aligns with user needs and requirements. Professionals must involve end users in this process to ensure that the new system will function adequately and meets their needs and expectations.

4.4.3 Architectural Design

After the requirements have been determined, the necessary specifications for the hardware, software, people, and data resources, and the information products that will satisfy the functional requirements of the proposed system

can be determined. The design will serve as a blueprint for the system and helps detect problems before these errors or problems are built into the final system. Professionals create the system design, but must review their work with the users to ensure the design meets users' needs.

4.4.4 Coding and Debugging

Coding and debugging is the act of creating the final system. This step is done by software developer.

4.4.5 System Testing

The system must be tested to evaluate its actual functionality in relation to expected or intended functionality. Some other issues to consider during this stage would be converting old data into the new system and training employees to use the new system. End users will be key in determining whether the developed system meets the intended requirements, and the extent to which the system is actually used.

4.4.6 Maintenance

Inevitably the system will need maintenance. Software will definitely undergo change once it is delivered to the customer. There are many reasons for the change. Change could happen because of some unexpected input values into the system. In addition, the changes in the system could directly affect the software operations. The software should be developed to accommodate changes that could happen during the post implementation period.

There are various software process models like:-

- Prototyping Model
- RAD Model
- The Spiral Model
- The Waterfall Model
- The Iterative Model

Of all these process models we've used the Iterative model (The Linear Sequential Model) for the development of our project.

4.5 The Iterative model

The waterfall model derives its name due to the cascading effect from one phase to the other as is illustrated in Figure1.1. In this model each phase well defined starting and ending point, with identifiable deliveries to the next phase.

This model is sometimes referred to as the linear sequential model or the software life cycle.

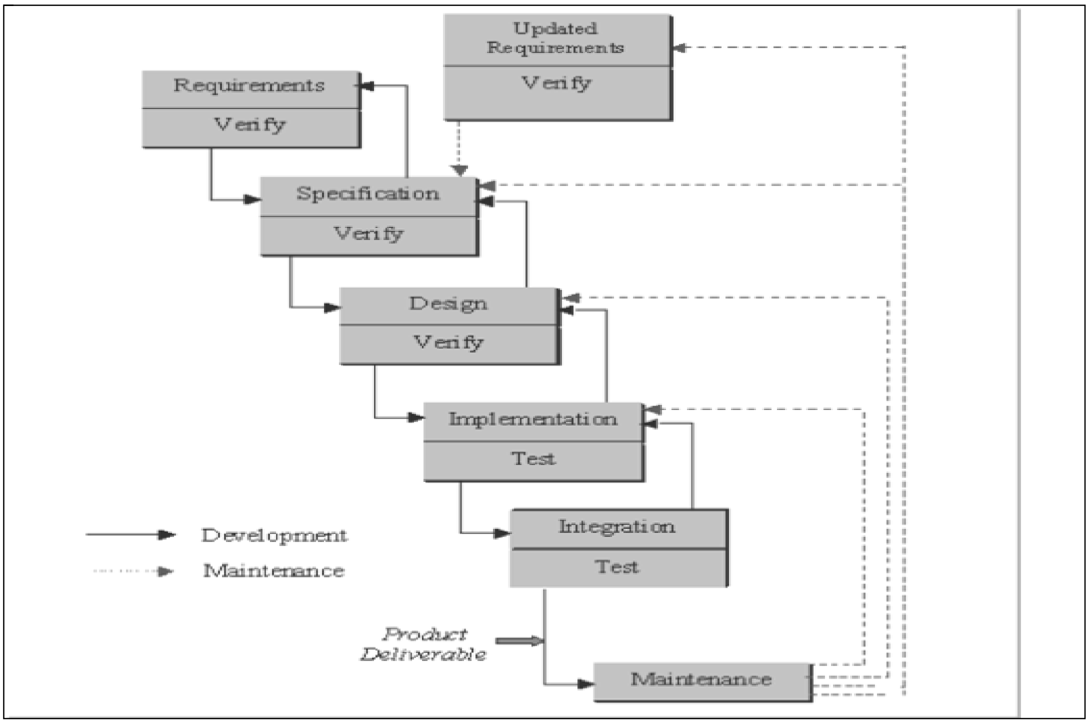


Fig 4.6: Software life cycle

The model consists of six distinct stages, namely:

1. In the **requirements analysis** phase

(a) The problem is specified along with the desired service objectives (goals)

(b) The constraints are identified

2. In the **specification phase** the system specification is produced from the detailed definitions of (a) and (b) above. This document should clearly define the product function.

3. In the system and software **design phase**, the system specifications are translated into a software representation. The software engineer at this stage is concerned with:

- Data structure
- Software architecture
- Algorithmic detail
- Interface representations

The hardware requirements are also determined at this stage along with a picture of the overall system architecture. By the end of this stage should the software engineer should be able to identify the relationship between the hardware, software and the associated interfaces. Any faults in the specification should ideally not be passed 'down stream.

4. In the **implementation and testing** phase stage the designs are translated into the software domain

- Detailed documentation from the design phase can significantly reduce the coding effort.
- Testing at this stage focuses on making sure that any errors are identified and that the software meets its required specification.

5. In the **integration and system testing** phase all the program units are integrated and tested to ensure that the complete system meets the software requirements. After this stage the software is delivered to the customer [**Deliverable – The software product is delivered to the client for acceptance testing.**]

6. The **maintenance phase** the usually the longest stage of the software. In this phase the software is updated to:

- Meet the changing customer needs

- Adapted to accommodate changes in the external environment

- Correct errors and oversights previously undetected in the testing phases

- Enhancing the efficiency of the software

Observe that feedback loops allow for corrections to be incorporated into the model. For example, a problem/update in the design phase requires a 'revisit' to the specifications phase. When changes are made at any phase, the relevant documentation should be updated to reflect that change.

Advantages of the Iterative Model:

- Testing is inherent to every phase of the Iterative model
- It is an enforced disciplined approach
- It is documentation driven, that is, documentation is produced at every stage

Disadvantages of the Iterative Model:

The waterfall model is the oldest and the most widely used paradigm. However, many projects rarely follow its sequential flow. This is due to the inherent problems associated with its rigid format. Namely:

- It only incorporates iteration indirectly, thus changes may cause considerable confusion as the project progresses.
- As The client usually only has a vague idea of exactly what is required from the software product, this IM has difficulty accommodating the natural uncertainty that exists at the beginning of the project.
- The customer only sees a working version of the product after it has been coded. This may result in disaster any undetected problems are precipitated to this stage.

4.6 UML Diagrams

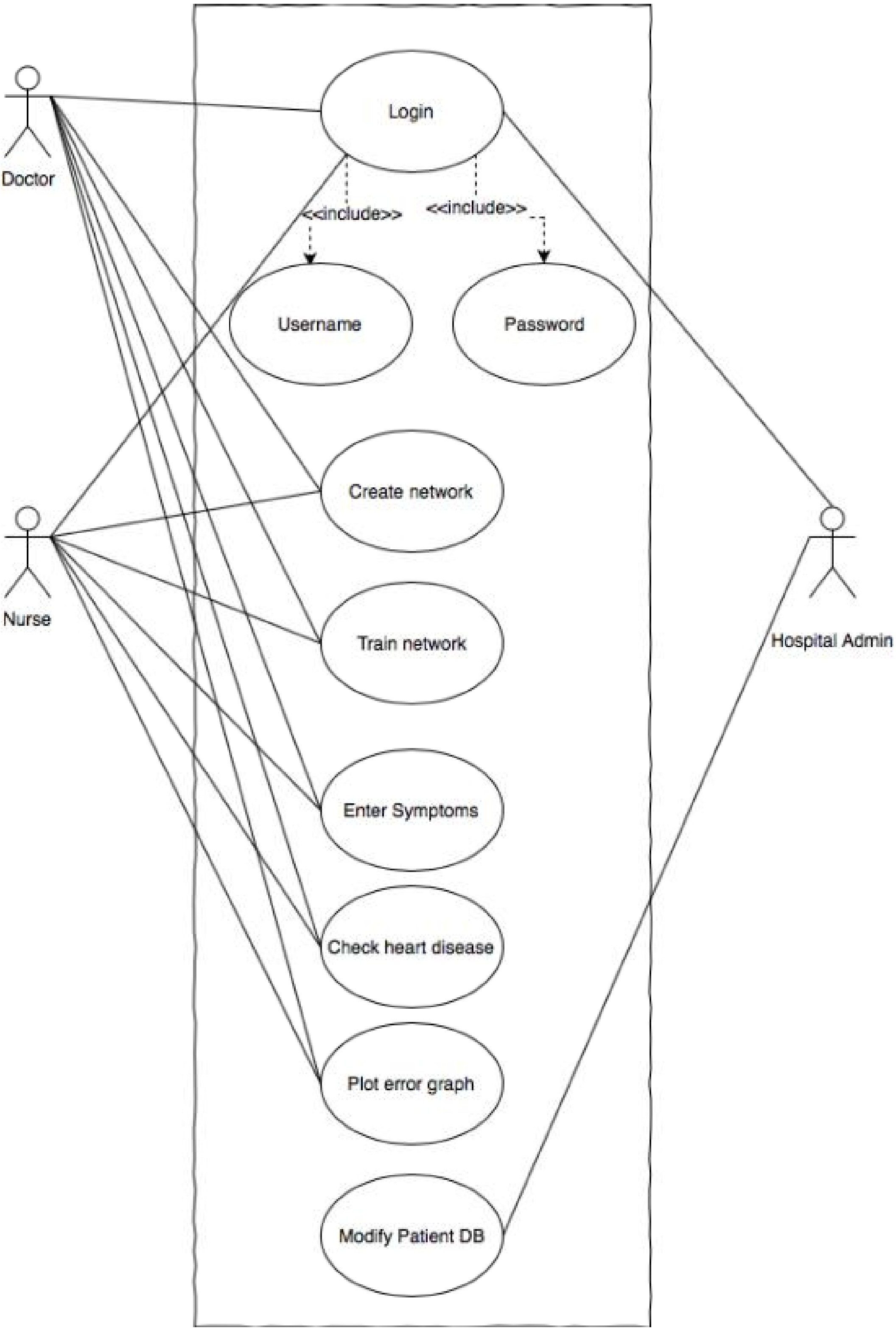


Fig. 4.7: Use Case Diagram

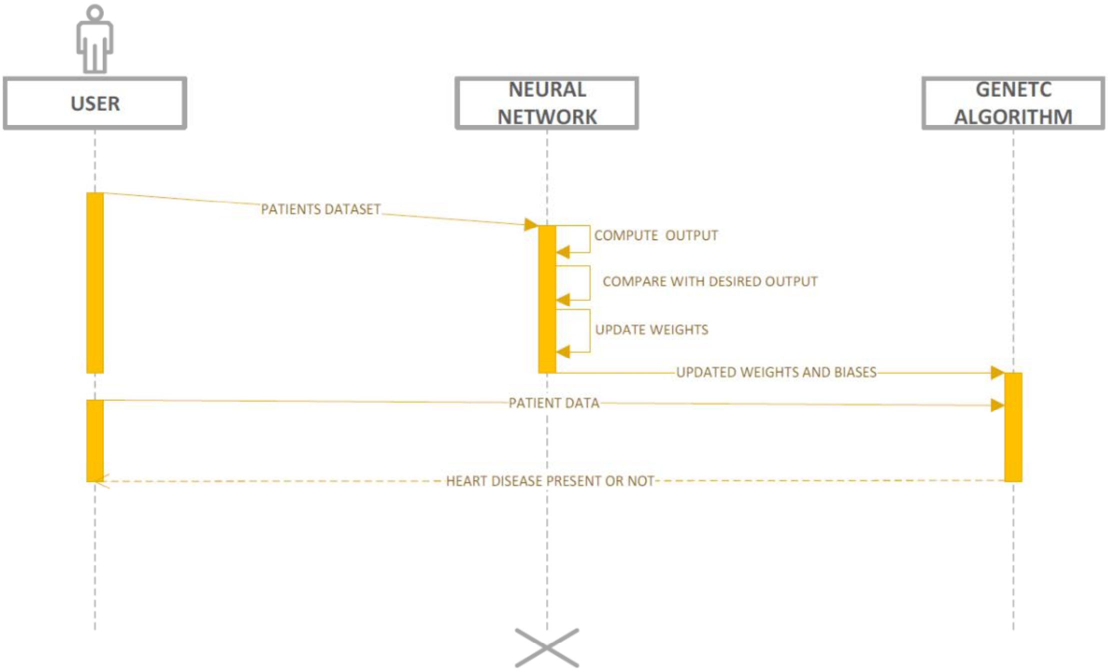


Fig. 4.8: Sequence Diagram

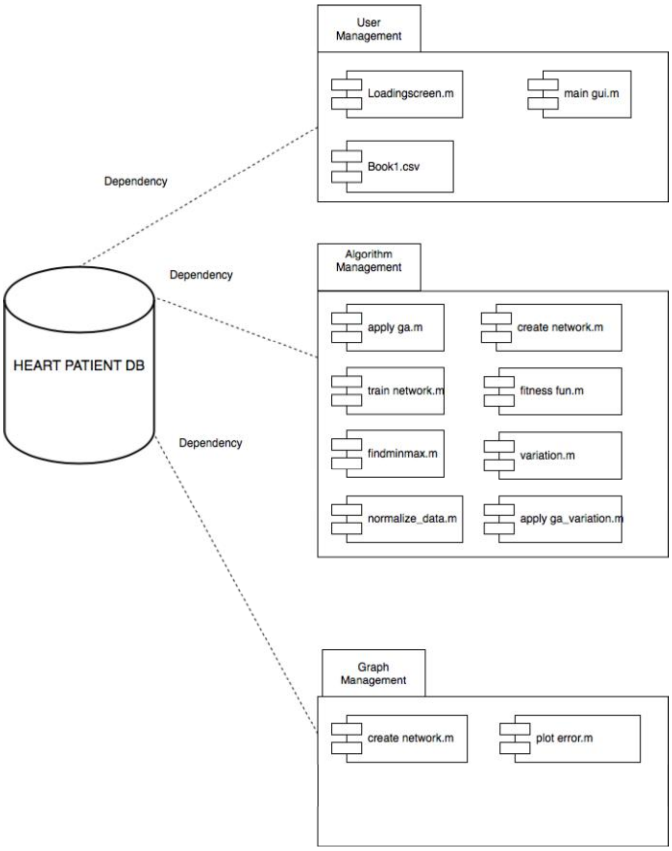


Fig. 4.9: Component Diagram

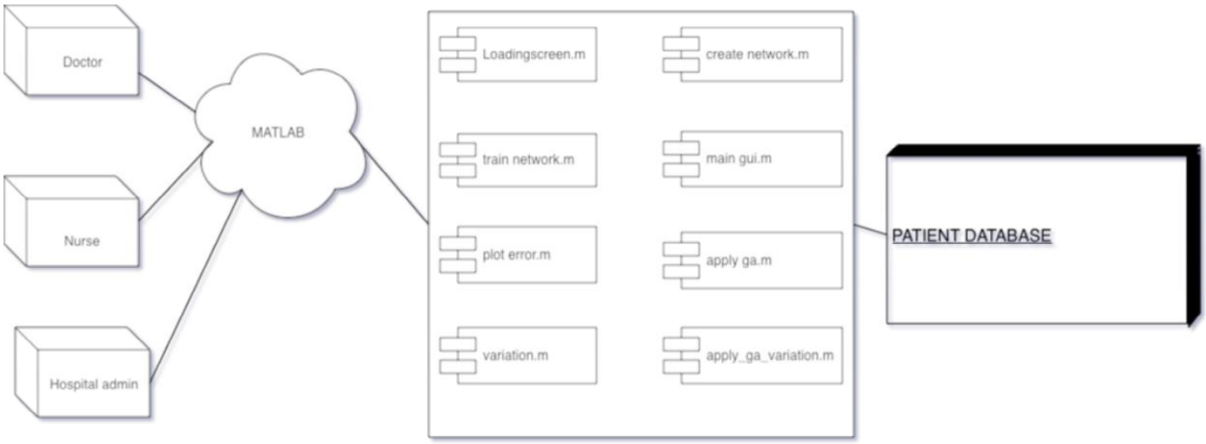


Fig. 4.10: Deployment Diagram

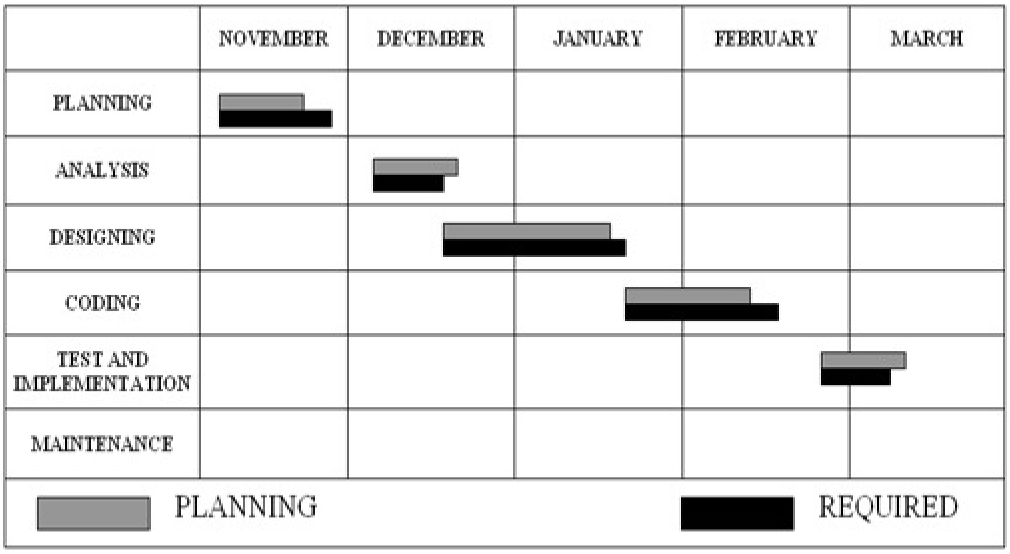
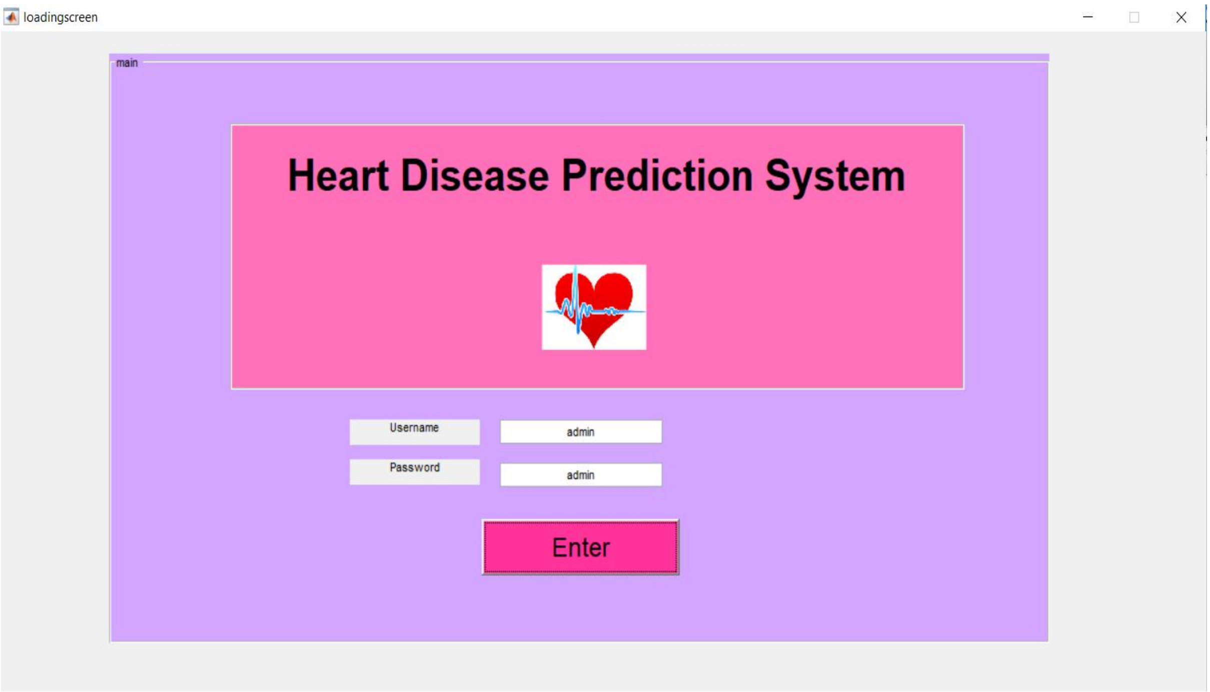


Fig. 4.11: Gantt Chart

4.7 Snapshots

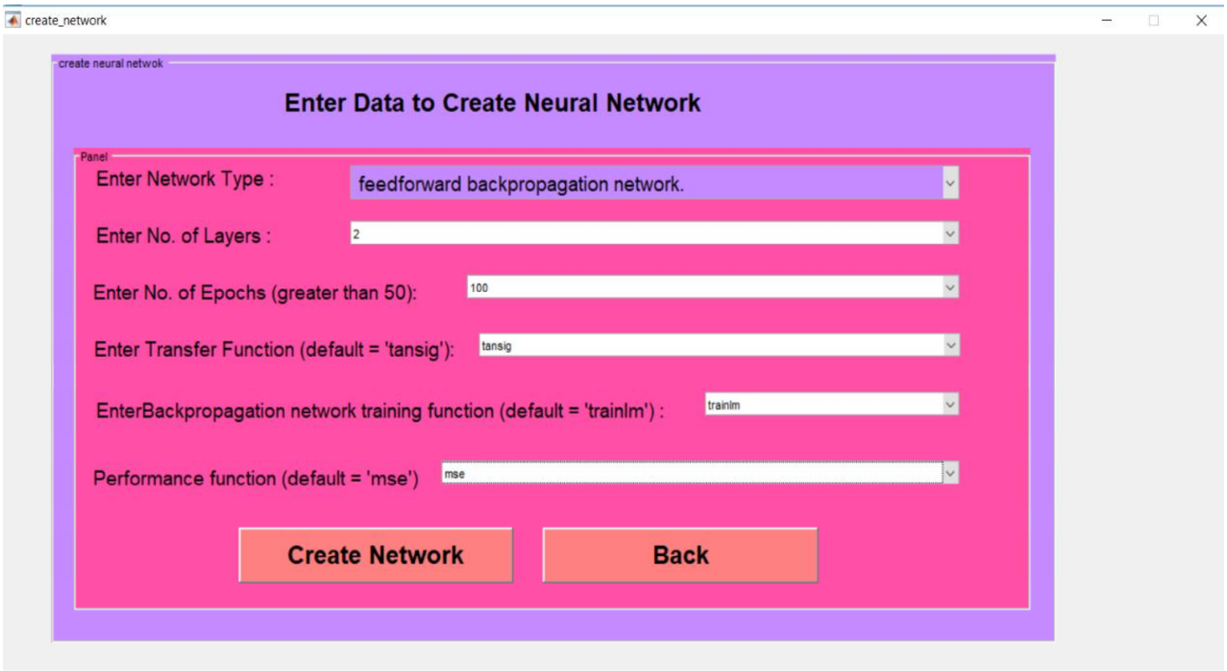


Heart Disease Prediction System

Username: admin

Password: admin

Fig. 4.13: Loading Screen (Home Page)



Enter Data to Create Neural Network

Enter Network Type : feedforward backpropagation network.

Enter No. of Layers : 2

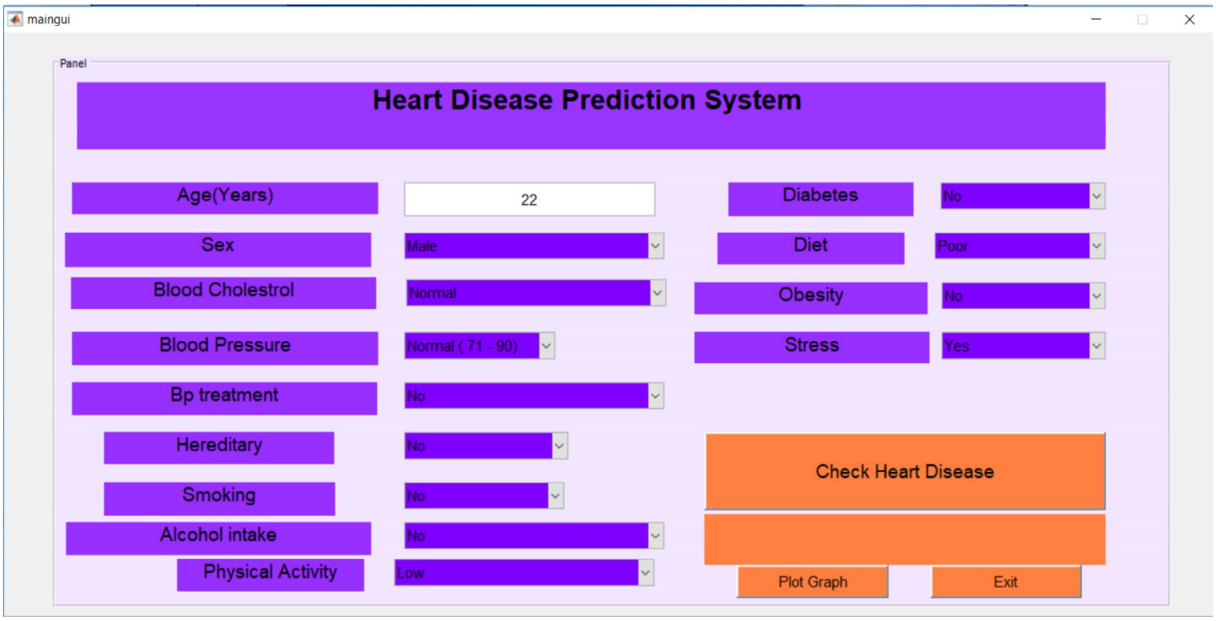
Enter No. of Epochs (greater than 50): 100

Enter Transfer Function (default = 'tansig'): tansig

EnterBackpropagation network training function (default = 'trainlm') : trainlm

Performance function (default = 'mse') mse

Fig. 4.14: Creation of Neural Network



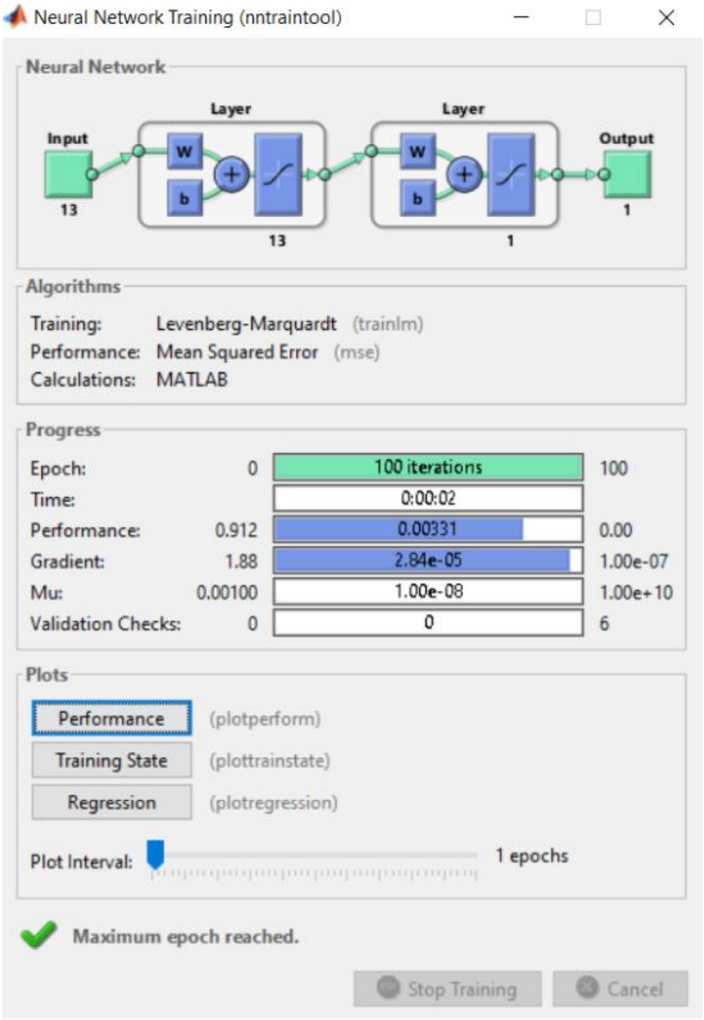
maingui

Panel

Heart Disease Prediction System

Age(Years)	22	Diabetes	No
Sex	Male	Diet	Poor
Blood Cholestrol	Normal	Obesity	No
Blood Pressure	Normal (71 - 90)	Stress	Yes
Bp treatment	No		
Hereditary	No		
Smoking	No		
Alcohol intake	No		
Physical Activity	Low		

Fig. 4.15: Main GUI (Accepts Details of Patient)



Neural Network Training (nntraintool)

Neural Network

Input
13
Layer
W
b
13
Layer
W
b
1
Output
1

Algorithms

Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Calculations: MATLAB

Progress

Epoch:	0	100 iterations	100
Time:		0:00:02	
Performance:	0.912	0.00331	0.00
Gradient:	1.88	2.84e-05	1.00e-07
Mu:	0.00100	1.00e-08	1.00e+10
Validation Checks:	0	0	6

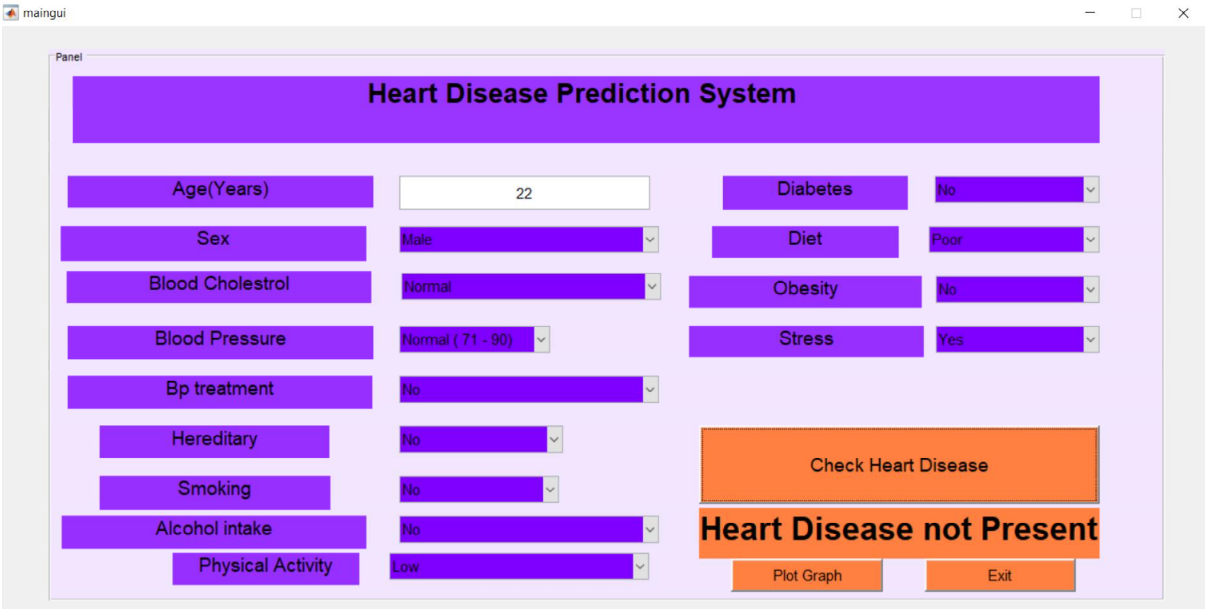
Plots

Performance (plotperform)
Training State (plottrainstate)
Regression (plotregression)

Plot Interval: 1 epochs

✓ Maximum epoch reached.

Fig. 4.16: Network Training

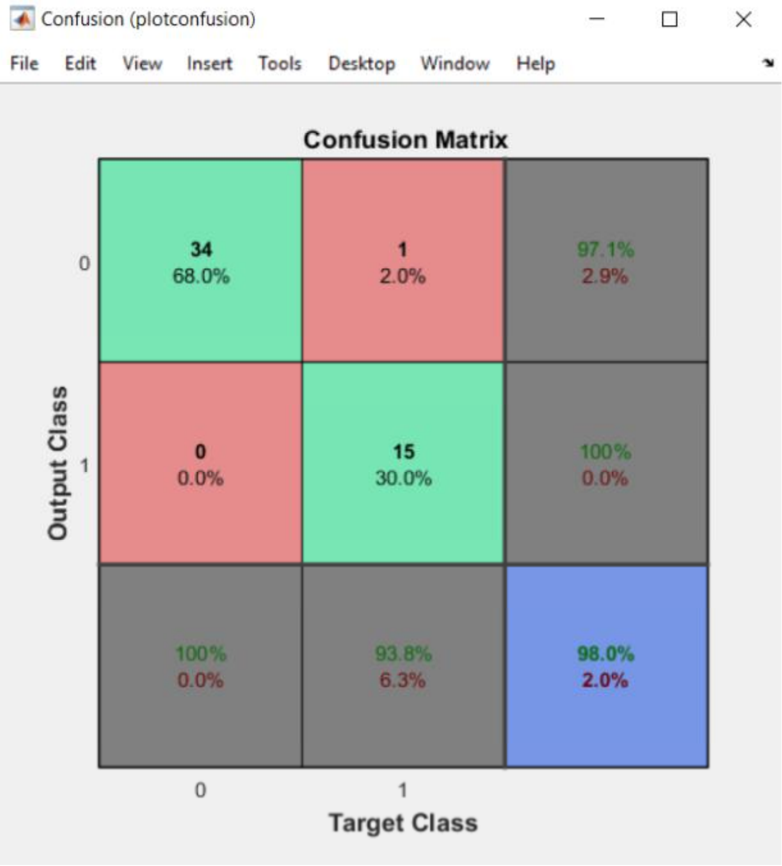


Heart Disease Prediction System

Age(Years): 22
Sex: Male
Blood Cholestrol: Normal
Blood Pressure: Normal (71 - 90)
Bp treatment: No
Hereditary: No
Smoking: No
Alcohol intake: No
Physical Activity: Low
Diabetes: No
Diet: Poor
Obesity: No
Stress: Yes

Heart Disease not Present

Fig. 4.17: Output (shows Heart Disease Present or Not Present)



Confusion Matrix

Output Class \ Target Class	0	1	
0	34 68.0%	1 2.0%	97.1% 2.9%
1	0 0.0%	15 30.0%	100% 0.0%
	100% 0.0%	93.8% 6.3%	98.0% 2.0%

Fig. 4.18: Confusion Matrix (Shows Accuracy)

4.8 MATLAB

MATLAB (matrix laboratory) is a multi-paradigm numerical computing environment and proprietary programming language developed by MathWorks. MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, C#, Java, Fortran and Python.

Although MATLAB is intended primarily for numerical computing, an optional toolbox uses the MuPAD symbolic engine, allowing access to symbolic computing abilities. An additional package, Simulink, adds graphical multi-domain simulation and model-based design for dynamic and embedded systems.

As of 2018, MATLAB has more than 3 million users worldwide. MATLAB users come from various backgrounds of engineering, science, and economics.

4.9 Deep Learning Toolbox

Deep Learning Toolbox™ (formerly Neural Network Toolbox™) provides a framework for designing and implementing deep neural networks with algorithms, pretrained models, and apps. You can use convolutional neural networks (ConvNets, CNNs) and long short-term memory (LSTM) networks to perform classification and regression on image, time-series, and text data. Apps and plots help you visualize activations, edit network architectures, and monitor training progress.

For small training sets, you can perform transfer learning with pretrained deep network models (including SqueezeNet, Inception-v3, ResNet-101, GoogLeNet, and VGG-19) and models imported from TensorFlow™-Keras and Caffe.

To speed up training on large datasets, you can distribute computations and data across multicore processors and GPUs on the desktop (with Parallel Computing Toolbox™), or scale up to clusters and clouds, including Amazon EC2® P2, P3, and G3 GPU instances (with MATLAB Distributed Computing Server™).

Chapter 5

Results and Discussion

5.1 Results

In this section, the experimental results of the heart attack disease system for prediction are explained. Here, the evaluation of the sensitivity, specificity and accuracy occurs by comparing the performance of the proposed system with neural network-based system.

In the proposed system, we find the risk factors of heart attack patients and obtained results are evaluated with namely sensitivity, specificity, and accuracy. After detecting a positive disease, Sensitivity evaluates the diagnostic test correctly. By eliminating a given condition, Accuracy measures correctly figured out diagnostic test. In order to find these metrics, we first compute some of the terms like, True positive (TP), True negative (TN), false positive (FP) and false negative (FN).

Sensitivity = TP/(TP+FN) …………………………………………………………..(a)

Specificity = TN/(TN+FP) ………………………………………………………….. (b)

Accuracy = (TN+TP)/(TN+TP+FN+FP)…………………………………………………...(c)

The confusion matrix shows the number of samples which have been classified into the two correctly/falsely classes of C1 and C2. The entries of this matrix are used to explain the performance measures [7]. In the confusion matrix, the correctly classified number of samples of class C1 falls under true positive (TP); false negative (FN) is the number of the samples of class C1 which have been falsely classified as C2; and false positive (FP) is the number of the samples of class C2 which have been falsely classified as C1.

	Actual Class		
Predicted Class	C1	True Positive (TP)	False Positive (FP)
	C2	False Negative (FN)	True Negative (TN)

Table 5.1: Confusion Matrix

The genetic optimized NN is trained and tested using sample of 50 patient data. Utilizing accuracy, sensitivity and specificity, the performance of the system is compared with the neural network-based system. In the True Positive value 34 and True Negative is 15. Then False positive 1 and False Negative 0. The accuracy is given by the 95.02%.

Based on the implementation result Multi-Layer Feed Forward Neural Network and Genetic Network evaluate the best accurate performance.

BPN is widely used in the learning algorithm in Neural Network for the many applications. However, BP learning depends on weights in the MLFFNN. Due to this, GA has been used to obtain the optimal parameter value and weight for the BP learning. So that the performance of GA is increased better than the MLFFNN.

5.2 Future Scope

We have implemented the neuro genetic algorithm using a small dataset. But, imagine if this was a system employed by very large and huge datasets giving us more precise and derivative conclusions.

In today’s world, everything is being analyzed by data and using that data to derive conclusions, inference and predictions from the given data has evolved many industries and is not stopping in the near future.

If our project is implemented in large-industry datasets, it can be diversified into looking for the symptoms of the disease or the causation of that specific pattern making use of various machine learning algorithms. Our project may become the base of diversification, and to facilitate this acceleration of development, it is necessary to have the basics of machine learning model.

Our project is focused on the prediction of one disease. It can be used to have predictions of different diseases other than heart disease which our project proposes. Our project can assure that different diseases prediction is possible in today’s world as, in the old days, data was considered to be a black box where it was said to be not accurate but the statistics and the huge influx of data says otherwise.

Chapter 6

Conclusions

The proposed heart system has taken 50 patients suffering from heart diseases for machine learning and intelligent systems, weighted based on frequency in the datasets. The usage of multi-layer feed forward neural network optimized with genetic algorithm should be adjusted by adjusting the variable and given the better improved results were compared with other neural-based systems to get a measure of accuracy, sensitivity and specificity. This work also demonstrates about GA-NN prediction by improving rate at which hidden neurons are optimized . By using this, predictions and conclusions become more clear and accurate.

Appendix

The Back-Propagation Training Algorithm

- Initialization

  $\left(-\frac{2.4}{F_i}, +\frac{2.4}{F_i}\right)$

- Activation
  - Actual Outputs in Hidden Layer

    $$y_j(p) = sigmoid\left[\sum_{i=1}^{n} x_i(p).w_{ij} - \theta_j\right]$$

  - Actual Outputs in Output Layer

    $$y_k(p) = sigmoid\left[\sum_{j=1}^{m} x_{jk}(p).w_{jk}(p) - \theta_k\right]$$

- Weight Training
  - Error Gradient

    $\delta_k(p) = y_k(p).[1 - y_k(p)].e_k(p)$

    $e_k(p) = y_{d,k}(p) - y_k(p)$

  - Weight Corrections

    $\Delta w_{jk}(p) = \alpha.y_j(p).\delta_k(p)$

  - Update Weights

    $w_{jk}(p+1) = w_{jk}(p) + \Delta w_{jk}(p)$

- Iteration

Mean Square Error

$mse = \sum_k (O_k - T_k)^2/n$

Accuracy

Accuracy = (TP + TN) / (TP + FP + TN + FN)

References

[1] Mozaffarian D, Wilson PW, Kannel WB, Beyond established and novel risk factors: lifestyle risk factors for cardiovascular disease. Circulation 117: 3031–3038, 2008.

[2] Poirier P, Healthy lifestyle: even if you are doing everything right, extra weight carries an excess risk of acute coronary events. Circulation 117: 3057–3059, 2008.

[3] Wood D, De Backer, Prevention of coronary heart disease in clinical practice: recommendations of the Second Joint Task Force of European and other Societies on Coronary Prevention. Atherosclerosis 140: 199–270, 1998.

[4] Anderson KM, Odell PM, Wilson PWF, Kannel WB. Cardiovascular disease risk profiles. Am Heart J., 121: 293–298, 1991.

[5] Kannel WB, An investigation of coronary heart disease in families. The Framingham offspring study. Am J Epidemiol 110: 281–290, 1979.

[6] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks," International Journal on Computer Science and Engineering (IJCSE), vol. 2, no. 2, pp. 250-255, 2010.

[7] H. Yan, et al., "Development of a decision support system for heart disease diagnosis using multilayer perceptron," in Proc. of the 2003 International Symposium on, vol. 5, pp. V-709- V-712.

[8] M. C. Tu, D. Shin, and D. Shin, "Effective Diagnosis of Heart Disease through Bagging Approach," Biomedical Engineering and Informatics, IEEE, 2009.

[9] Hai H.Dam, Hussain A.Abbass and Xin Yao, “Neural – Based Learning Classifier Systems”, IEEE Transactions on Knowledge and Data Engineering, Vol.20, No.1, pp.26-39, 2008.

[10] J. Guo, and W. J. Sun, Theory of Neural Network and its Implementation with MatLab, Electronic Industry Press, Beijing, 2005.

[11] P.K.Anooj, “Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules”, Journal of King Saud University –Computer and Information Sciences. xxx,xxx-xxx, (2011).

[12] Latha Parthiban and R. Subramanian, “Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm”, International Journal of Biological and Life Science, Vol. 15, pp. 157 - 160, 2007.

[13] Olatubosun Olabode and Bola Titilayo Olabode, “Cerebrovascular Accident Attack Classification Using Multilayer Feed Forward Artificial Neural Network with Back Propagation Error”, Journal of Computer Science, Vol. 8, No. 1, pp.18 - 25, 2012.

[14] Lynne E. Parke, “Notes on Multilayer, Feedforward Neural Networks”CS425/528: Machine Learning, (2010).

[15] Asha Gowda Karegowda, A.S. Manjunath, M.A. Jayaram, Application Of Genetic Algorithm Optimized Neural Network Connection Weights For Medical Diagnosis Of Pima Indians Diabetes”, International Journal on Soft Computing (IJSC), Vol.2, No.2,pp. 15-22, (2011).

[16] www.myreaders.info/html/artificialintelligencehtml.

Paper Publications

A. **Medical Diagnosis Using Data Mining**

Published In JETIR (www.JETIR.org) ISSN UGC Approved & 5.87 Impact Factor Published in Volume 5 Issue 10 , October-2018

MEDICAL DIAGONOSIS USING DATA-MINING

Zain Momin, Asad Siddiqui, Huzaifa Vakil, Sohel Tharani

Department of Computer Engineering,

Rizvi College Of Engineering, Bandra, Mumbai.

Abstract:

The amount of data that is collected from the healthcare industry is enormous and hence to go through that data for the improvement of the diagnosis in the future is next to impossible .In summary, the amount of data that is collected is not mined properly. Even though it will be very advantageous for us and the doctors that are giving the diagnosis to plough through the data for simple and effective diagnosis. Our research, hence, focuses on this niche concept of exploiting the collection of data for diseases like diabetes, hepatitis and heart diseases and to devise a decision tree that will make intelligent decisions medically to help the physicians. Various such decision tree algorithms are C4.5 algorithm, ID3 algorithm and CART algorithm. These types of algorithms are used to compare the effectiveness and correction rate among them.

1. Introduction:

There are at least millions of people that are diagnosed and treated in the healthcare industry on a monthly basis. To go through the diagnosis and the treatment of every one of them is inhumane. But in this world of technology, we don’t need to do all the work if a set system using specific algorithms can do our tedious work. The amount of information that can be churned into important diagnosis and treatment is unimaginable. The process becomes easier for both the physicians and the doctors concerned to treat the disease and give the proper diagnosis. A support system solely based on the decisions that were made and which can analyze the different factors such as relationships between the past patient and the current patient, all the diseases in the population, history of the family, and test results, would prove very useful and would be revolutionary in the field of medicine. This concept, called the Decision Support System (DSS) is very broad because of many approaches and the domain which has no end as the patients in the healthcare industry are never-ending based on which decisions are made. It can be summarized as the computerized system that helps make decisions. A DSS application consists of many subsystems .However, the development of a system like this one is a very daunting and tedious task. Many factors are attributed in making such a system but inadequate information has been identified as a major challenge. Hence it, has become our responsibility in such a technical age to take care of such problems like inadequate information and make powerful medical decision support system (MDDS) to support the exploding demand and the increasingly difficult and excruciating diagnosis decision process.

The medical diagnosis is a very difficult process as there are various factors that come into the picture. Every patient has their own story to tell and their own background. To suffice such information and provide diagnosis using specific algorithms is very difficult and that is why we use the concepts of soft computing methods such as decision tree classifiers have shown great potential to be applied in the development of MDSS of various diseases, majorly aimed at diseases like heart diseases, diabetes and many other. The aim is identification of the most important risk factors based on the classification rules that need to be extracted.

2. Overview of related work:

According to the World Health Organization (WHO) fact sheet on diabetes, an estimated 3.4 million deaths are caused due to high blood sugar. Also heart disease is a leading cause of death in world, WHO claims 3.8 million and 3.4 million deaths in males and females, respectively.

Up to now, many studies have reported that they have focused on medical diagnosis. Data Mining Techniques Used in Diagnosis System Classification technique is the most frequently used data mining tasks with a majority of the implementation of Bayesian classifiers, neural networks, and Association Rule. The data mining techniques that have been applied to medical data include Apriori and FPGrowth, decision tree algorithms like ID3, C4.5, C5, and CART, and, Naïve Bayesian, combination of K-means, Self Organizing Map (SOM) and Naïve Bayes. Different approaches have been applied on these studies to achieve high accuracies i.e. 77% or higher using different datasets. Some examples are given below:

D. Senthil Kumar, G. Sathyadeviand S. Sivanesh's results concluded the idea of medical diagnosis and decision tree algorithm for effective classification and also found 83.184% accuracy with cart algorithm which is greater than ID3.

Robert Detrano's experimental results showed correct classification accuracy of approximately 77.00% with a logistic-regression-derived discriminant function.

Vector Machines gets support from Ischemic Heart Disease (IHD) which have high accuracy and serve as excellent classifiers and predictors. Nonlinear proximal Support Vector Machines (PSVM) uses Classifiers when it is tree based.

Polat and Gunes designed an expert system to diagnose the diabetes disease based on principal component analysis. To diagnose the diabetes, Polat et al. developed a cascade learning system.

The Centers for Disease Control and prevention shows gestational diabetes which was presented by National Center for Chronic Disease Prevention and Health Promotion.

The new framework known as duo-mining tool was presented by Jaya Rama Krishniah et al., which is used for diagnosing diabetes. Many classification algorithms like KNN, SVM, and decision tree was applied by Jaya Rama Krishniah for type-2 diabetes. SVM algorithm has highest accuracy among all the algorithm with value of 96.39%.

Aljarullah et al., proposed J48 algorithm to diagnose type-2 diabetes which is used for constructing a decision tree. The accuracy of the model is 78.68%.

Adidela et al., presented the type of diabetes by using Fuzzy ID3 method. The author uses the system for predicting the disease from data set as it initially clusters the data and applies the classification algorithms on clustered data. The author presented a combination of classification method where they developed EM algorithm for clustering and fuzzy ID3 algorithm to attain decision tree for each cluster.

G. Parthiban et al, applied Naïve Bayes method to diagnose heart related problems which are occurring in diabetic patients.

3. About the datasets

The aim of this study is evaluation and development of a Clinical Decision Support System for the treatment of patients with Diabetes and Heart Disease. According to survey and World Health Organization (WHO), every year, the leading of deaths is Heart Disease.

Elsevier, in the journal of American College of Cardiology, the death rate due to cardiovascular diseases (CVD) declined by a significant 41 % in US between 1990 and 2016, whereas in Indi it rose by 34 % from 155.7 to 209.1 deaths per one lakh population in the same period. In India, the leading cause of mortality is Cardiovascular Diseases (CVDs). In India, the heart ailments caused more than 2.1 million deaths in all ages that is more than a quarter of all deaths.

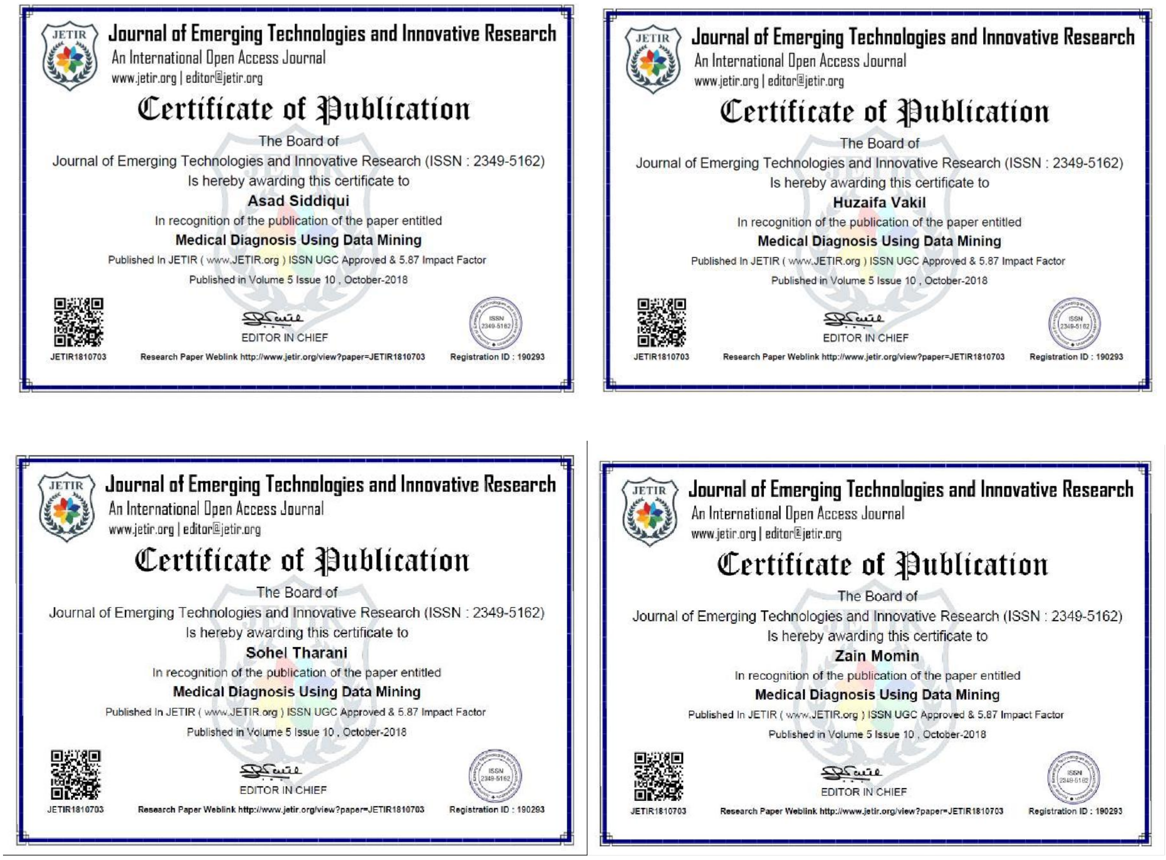
Diabetes is also called as diabetes mellitus. It is a disease that result in too much sugar in blood (high blood glucose). Diabetes mellitus is a metabolic disorders in which there are high blood sugar levels. The number of people with diabetes has risen from 108 million in 1980 to 422 million in 2014. The global prevalence of diabetes among adults over 18 years of age has risen from 4.7% in 1980 to 8.5% in 2014. There are currently 246 million diabetic people all over the world. According to the International Diabetes Federation, the number of diabetic patients would rise to 380 million by 2025. Due to the rapid change in lifestyles, the deaths due to diabetes had increased by 50% between 2005 and 2015. Now in India, the seventh most common cause of death is diabetes.

All these datasets used in this study are taken from UCI KDD Archive. Also this study used Cleveland Clinic Foundation dataset known as "Cleveland Clinic Foundation Heart Disease Dataset"

4. Conclusion

Decision tree algorithms are some of the most efficient and powerful methods of classification. The data sets we have used will test the efficiency and correction rate of the algorithms. As expected through our intuition and prior knowledge of the algorithms, CART algorithm showed the best performance of all the algorithms. The outcomes we have reached will give great advances to doctors making decisions on a patient with heart disease or diabetes. We have surveyed the data through decision tree algorithms ID3, CART and C4.5 and we have concluded that CART algorithm has the best rate of success with regards to performance of rules and accuracy. The results of all this processed data is stored in a database so doctors can access it with ease and to allow them to look at the data and form similarities with the diagnoses. To further improve this decision support, the medications that patients are consuming should also be added to the data set to make the results even more accurate.

Certificates



Journal of Emerging Technologies and Innovative Research
An International Open Access Journal
www.jetir.org | editor@jetir.org

Certificate of Publication

The Board of
Journal of Emerging Technologies and Innovative Research (ISSN : 2349-5162)
Is hereby awarding this certificate to
Asad Siddiqui
In recognition of the publication of the paper entitled
Medical Diagnosis Using Data Mining
Published In JETIR (www.JETIR.org) ISSN UGC Approved & 5.87 Impact Factor
Published in Volume 5 Issue 10 , October-2018

EDITOR IN CHIEF

ISSN 2349-5162

JETIR1810703 Research Paper Weblink http://www.jetir.org/view?paper=JETIR1810703 Registration ID : 190293

Journal of Emerging Technologies and Innovative Research
An International Open Access Journal
www.jetir.org | editor@jetir.org

Certificate of Publication

The Board of
Journal of Emerging Technologies and Innovative Research (ISSN : 2349-5162)
Is hereby awarding this certificate to
Huzaifa Vakil
In recognition of the publication of the paper entitled
Medical Diagnosis Using Data Mining
Published In JETIR (www.JETIR.org) ISSN UGC Approved & 5.87 Impact Factor
Published in Volume 5 Issue 10 , October-2018

EDITOR IN CHIEF

ISSN 2349-5162

JETIR1810703 Research Paper Weblink http://www.jetir.org/view?paper=JETIR1810703 Registration ID : 190293

Journal of Emerging Technologies and Innovative Research
An International Open Access Journal
www.jetir.org | editor@jetir.org

Certificate of Publication

The Board of
Journal of Emerging Technologies and Innovative Research (ISSN : 2349-5162)
Is hereby awarding this certificate to
Sohel Tharani
In recognition of the publication of the paper entitled
Medical Diagnosis Using Data Mining
Published In JETIR (www.JETIR.org) ISSN UGC Approved & 5.87 Impact Factor
Published in Volume 5 Issue 10 , October-2018

EDITOR IN CHIEF

ISSN 2349-5162

JETIR1810703 Research Paper Weblink http://www.jetir.org/view?paper=JETIR1810703 Registration ID : 190293

Journal of Emerging Technologies and Innovative Research
An International Open Access Journal
www.jetir.org | editor@jetir.org

Certificate of Publication

The Board of
Journal of Emerging Technologies and Innovative Research (ISSN : 2349-5162)
Is hereby awarding this certificate to
Zain Momin
In recognition of the publication of the paper entitled
Medical Diagnosis Using Data Mining
Published In JETIR (www.JETIR.org) ISSN UGC Approved & 5.87 Impact Factor
Published in Volume 5 Issue 10 , October-2018

EDITOR IN CHIEF

ISSN 2349-5162

JETIR1810703 Research Paper Weblink http://www.jetir.org/view?paper=JETIR1810703 Registration ID : 190293

Fig. 6.1 Certificates

B. Heart Disease Prediction (Medical Diagnosis) using Data Mining

Paper Submitted in International Conference on Global Technology Initiatives (ICGTI), Rizvi College of Engineering

Paper Publish In-Progress

Heart Disease Prediction (Medical Diagnosis) using Data Mining

Asad Siddiqui[a], Huzaifa Vakil[b], Sohel Tharani[c], Zain Momin[d]

Mentor: Dr. Anupam Choudhary

[a,b,c,d]Rizvi College of Engineering

[a]asad-siddiqui@hotmail.com, [b]vakilhuzaifa@gmail.com, [c]sohel.tharani786@gmail.com, [d]zmomin29@gmail.com

Abstract: Medical domain application development is a rapidly expanding area of research. These applications have been very useful in designing clinical support systems because of the ability to find patterns in medical data. Important application of such system is the diagnosis of diseases in the heart, which has a high mortality rate all around the world. One of the data mining techniques used for medical data are neural networks. Model selection for a neural network, which automatically obtains knowledge from the patient's clinical data, uses selection of optimal number of hidden nodes, relevant input variables and optimal connection weights. In this paper, the use of Multi-Layer Feed Forward Neural Network that incorporates Genetic Algorithm and Back Propagation network (BPN) for heart attack projection is shown. GA initializes and optimizes the connection weights of neural network. Risk factors such as age, family history, diabetes, hypertension, high cholesterol, tobacco smoking, alcohol intake, and physical activity are utilized by system. Patients with heart disease have a lot of these negative risk factors thus allowing system to come to a conclusion. This system based on such factors will allow a doctor to find an early diagnosis of heart disease before complex and expensive and time-consuming traditional methods. This system was created in MATLAB and has up to 95.02% accurate results.

Key Words: Multi-Layer Feed Forward Neural Network, Genetic Algorithm, Back Propagation Network.

1. INTRODUCTION

The extraction of patterns that are hidden and relationships from databases that are huge in size, Data mining joins statistical analysis, machine learning and databases. In several areas of medical services, that includes effectiveness of surgical procedures, medical tests, all of the medication and the relations that are present in clinical and diagnosis data, in all of these, data mining is applies.

Medical diagnosis is a process that is very complex and requires many years of experience in that expertise. Heart attack is one of the main causes of death throughout the world. Hence there are huge databases pertaining to the causes of heart attack, symptoms of it and hence it gives a huge opportunity for data science and machine learning to hop on to this opportunity and make predictions and inferences hence to arrive on a conclusion that can help further people suffering from heart diseases or those people that are showing symptoms pertaining to heart disease. Most of the times symptoms also include other side effects such as discomfort in the shoulders, back pain, shortness of breath, nausea, light headedness.

We have proposed the classification of heart patients in which our system will make use of machine learning and find significant patters in heart patients to help them in further treatment. We have applied intervals with equal binning all over the data we have made use of the neuro genetic network using back propagation algorithms. Pre-Processing of the data is done to make the data efficient and useful to make predictions and inferences.

Our idea was inspired by Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg that made use of Neural Network and Genetic Algorithm. The study is improved by us by using hybrid system in data mining. The objective of our Proposed system is to build an intelligent system that is focused on heart diseases using previous databases. 13 attributes were proposed but in the end it was deemed necessary to include only those factors that actually contributed to the heart disease conclusions incorporating optimal model construction time.

2. DATA MINING TECHNQUES

Data Mining techniques are used to analyze, explore and extract medical data using high level instructions in order to find random and not known patterns. Data mining techniques are used by the researchers so that they can diagnose many fatal diseases such as diabetes, stroke, heart disease and cancer and these techniques are proven to give good accuracy for heart disease. Data mining techniques such as naïve bayes, neural network, bagging, kernel density, decision tree and support vector machine for prediction and diagnosis of heart disease are used by researchers. Some of the systems have shown that neural based learning classifier to classify data mining tasks performs equivalently to supervised learning classifier. IEHPS intelligent and effective heart attack prediction system was built with the help of data mining and neural networks and it expressed that in order to mine the frequent patterns, extracting significant patterns for heart disease prediction using K-means clustering and used MAFIA algorithm can be used. Polatet al., developed system using hybrid fuzzy and k-nearest neighbour approach for the prediction of heart disease, which had 87% accuracy in diagnosis [1]. System's using hybrid fuzzy and k-nearest neighbour method for the prediction of heart disease which was able to provide 87% accuracy in diagnosis was developed by Polatet al. In another system where neural network was implemented showed 94.02% accuracy in diagnosis of heart disease. Using genetic algorithm and CANFIS, Latha and Subramanian were able to propose an intelligent heart disease prediction system which has very low mean square error. Keeping in mind the various techniques discussed, this paper proposes a system using neural network and genetic algorithm in order to predict the risk of heart disease. To optimize neural network weights genetic algorithm is used. We have used such hybrid techniques on risk factors for prediction of heart disease. The main objective of this system is to use in clinical decision support as well as to use as a risk indicator for people to keep track of their health in the future.

3. DISEASE DATASET

Databases in hospitals and clinics have collected large amounts of data about patients which are being virtually unused, this information is being used in our system. Details of patients have been taken. The data set has been created based on 13 attributes/risk factors. Thus, we use this patient database to predict heart disease. Dataset was obtained from UCI[1]. The data of 50 people was collected from surveys done by American Heart Association [2].

Sr. No.	Risk Factors	Values
14.	Age (Years)	20-34 (0) 35-50 (1) 51-60 (2) >60 (3)
15.	Sex	Male (0), Female (1)
16.	Blood Cholesterol	Below 200 mg/dL – Low (1) 200-239 mg/dL - Normal (2) 240 mg/dL – High (3)
17.	Blood Pressure	Below 120 mmHg – Low (1) 120 to 139 mmHg – Normal (2) Above 139 mmHg – High (3)
18.	BP Treatment	Yes (1) No (2)
19.	Hereditary	Yes (1) No (2)
20.	Smoking	Yes (1)

		No (2)
21.	Alcohol Intake	Yes (1) No (2)
22.	Physical Activity	Low (1) Normal (2) High (3)
23.	Diabetes	Yes (1) No (2)
24.	Diet	Poor (1) Normal (2) Good (3)
25.	Obesity	Yes (1) No (2)
26.	Stress	Yes (1) No (2)
Output	Heart Disease	Present (0) Not Present (1)

Table 6.1: Research Paper Dataset

4. RESULT AND ANALYSIS

In this section, the experimental results of the heart attack disease system for prediction are explained. Here, the evaluation of the sensitivity, specificity and accuracy occurs by comparing the performance of the proposed system with neural network-based system.

In the proposed system, we find the risk factors of heart attack patients and obtained results are evaluated with namely sensitivity, specificity, and accuracy. After detecting a positive disease, Sensitivity evaluates the diagnostic test correctly. By eliminating a given condition, Accuracy measures correctly figured out diagnostic test. In order to find these metrics, we first compute some of the terms like, True positive (TP), True negative (TN), false positive (FP) and false negative (FN).

Sensitivity = TP/(TP+FN) (a)

Specificity = TN/(TN+FP) (b)

Accuracy = (TN+TP)/(TN+TP+FN+FP) (c)

The confusion matrix shows the number of samples which have been classified into the two correctly/falsely classes of C1 and C2. The entries of this matrix are used to explain the performance measures [7]. In the confusion matrix, the correctly classified number of samples of class C1 falls under true positive (TP); false negative (FN) is the number of the samples of class C1 which have been falsely classified as C2; and false positive (FP) is the number of the samples of class C2 which have been falsely classified as C1.

Predicted Class	**Actual Class**		
	C1	True Positive (TP)	False Positive (FP)
	C2	False Negative (FN)	True Negative (TN)

Table 6.2: Research Paper Confusion Matrix

The genetic optimized NN is trained and tested using sample of 50 patient data. Utilizing accuracy, sensitivity and specificity, the performance of the system is compared with the neural network-based system. In the True Positive value 34 and True Negative is 15. Then False positive 1 and False Negative 0. The accuracy is given by the 95.02%.

Based on the implementation result Multi-Layer Feed Forward Neural Network and Genetic Network evaluate the best accurate performance.

BPN is widely used in the learning algorithm in Neural Network for the many applications. However, BP learning depends on weights in the MLFFNN. Due to this, GA has been used to obtain the optimal parameter value and weight for the BP learning. So that the performance of GA is increased better than the MLFFNN.

5. CONCLUSION

The proposed heart system has taken 50 patients suffering from heart diseases for machine learning and intelligent systems, weighted based on frequency in the datasets. The usage of multi-layer feed forward neural network optimized with genetic algorithm should be adjusted by adjusting the variable and given the better improved results were compared with other neural-based systems to get a measure of accuracy, sensitivity and specificity. This work also demonstrates about GA-NN prediction by improving rate at which hidden neurons are optimized . By using this, predictions and conclusions become more clear and accurate.

Acknowledgements

I am profoundly grateful to Prof. Anupam Choudhary for his expert guidance and continuous encouragement throughout to see that this project rights its target.

I would like to express deepest appreciation towards Dr. Varsha Shah, Principal RCOE, Mumbai and Prof. Shiburaj Pappu, HOD Computer Department whose invaluable guidance supported me in this project.

At last I must express my sincere heartfelt gratitude to all the staff members of Computer Engineering Department who helped us directly or indirectly during this course of work.

Asad Siddiqui (62)
Huzaifa Vakil (68)
Sohel Tharani (67)
Zain Momin (33)