Medical Diagnosis System Using Machine Learning

Submitted By

Dhaval Raval

14MCEN16

तमसो मा ज्योतिर्गमय
NIRMA
UNIVERSITY
INSTITUTE OF TECHNOLOGY
NAAC ACCREDITED ‘A’ GRADE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2016

Medical Diagnosis System Using Machine Learning

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering(Networking Technologies)

Submitted By

Dhaval Raval

14mcen16

Guided By

Prof.Dvijesh Bhatt

Co-Guided By

Prof.Malaram K Kumhar

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2016

Certificate

This is to certify that the major project entitled **”Medical Diagnosis System Using Machine Learning”** submitted by **Dhaval Raval (Roll No: 14MCEN16)**, towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Engineering of Nirma University, Ahmedabad, is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this major project, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Prof. Dvijesh Bhatt
Guide & Assistant Professor,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Prof. Malaram Kumhar
Co-Guide & Assistant Professor,
CSE Department
Institute of Technology,
Nirma University, Ahmedabad

Dr. Gaurang Raval
Associate Professor,
Coordinator M.Tech - NT,
Institute of Technology,
Nirma University, Ahmedabad,

Dr. Sanjay Garg
Professor and Head,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. P. N. Tekwani
Director,
Institute of Technology,
Nirma University, Ahmedabad

Statement of Originality

I, **Dhaval Raval**, 14MCEN16 **14MCEN16**, give undertaking that the Major Project entitled "**Medical Diagnosis System Using Machine Learning**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made.It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

Signature of Student
Date:
Place:

Endorsed by
Prof. Dvijesh Bhatt
(Signature of Guide)

Acknowledgements

I would like to take this opportunity to thank the people who helped me in my project work.First of all, I would like to thank **Prof. Dvijesh Bhatt** for giving me an opportunity to work under his guidance and guiding me in every step of project.It gives me immense pleasure in expressing thanks and profound gratitude to **Prof. Dvijesh Bhatt**, Assistant Professor, Computer Science Department, Institute of Technology, Nirma University, Ahmedabad for his valuable guidance and continual encouragement throughout this work. The appreciation and continual support he has imparted has been a great motivation to me in reaching a higher goal. His guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

I would also like to thank **Prof. Malaram K Kumhar**, Assistant Professor, Computer Science Department, Institute of Technology, Nirma University, Ahmedabad for his valuable guidance and continual encouragement throughout this work.

It gives me an immense pleasure to thank **Dr. Sanjay Garg**, Hon'ble Head of Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his kind support and providing basic infrastructure and healthy research environment.

A special thank you is expressed wholeheartedly to **Dr.P.N.Tekwani**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation he has extended throughout course of this work.

I would also thank the Institution, all faculty members of Computer Engineering Department, Nirma University, Ahmedabad for their special attention and suggestions towards the project work.

- Dhaval Raval
14MCEN16

Abstract

Disease prediction is one of the critical task while designing medical software. Artificial intelligence and Neural network techniques have already been developed to solve this type of medical care problem. Recently, Machine Learning techniques have been successfully utilized in a different applications including to assist in medical diagnosis.It is very effortless and on time process for patients to analyze disease based on clinical and laboratory symptoms and give the more efficient result of particular disease.In this report I have described objectives of ”Swine Flu” disease, necessity of it,how algorithm solves existing problems and technical aspects of this project.

Abbreviations

SVM	Support Vector Machine.
NB	Naive Bayes
ANN	Artificial Neural Network.
CART	Classification and Regression Tree.
RF	Random Forest.
KNN	k-nearest neighbour.
BP	Backpropagation.
FFNC	Feed-Forward Neural Network Construction.

–

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

In computer science, artificial intelligence is used generate more imaginative machine. Learning system is primary requirement for the imaginative system. There are numbers of researchers presently concede that without learning system, machine cant produce effective outputs. Thus, Machine Learning is dominant branches of artificial intelligence. Machine learning algorithms are used to analysis data again and again to produce most effective results. Presently machine learning provides essential of machine for imaginative data scrutinize. Currently medical clinics are very well furnished with fully automatic and those machines are generating huge amount of data, then those data are collected and shared with information systems. Machine learning techniques can be used for the analysis of medical data and it is helpful in medical diagnosis for sensing different specialized diagnostic problems. Using Machine learning, system would take the patient data like symptoms, laboratory data and some of the important attributes as an input and generate the accurate diagnosis report. Then based on the accuracy of the result, machine will decide the training and trained dataset for the future reference. In current scenario, doctor has to collect all the record of the patient and based on that he will give medicine to patients. With this scenario, so much time is wasted due to several reasons which some time produced disaster in human life. When using machine learning classification algorithms, for particular diseases, we can improve the accuracy, speed, reliability and performance of the diagnostic on the current system. Machine learning is capable of offering automatic learning techniques to excerpt common patterns from realistic data and then make sophisticated and accurate decisions, based on the different learning behaviors. But major problem is with medical data because data have huge number of dimensionality and medical application finds the problem frequently change, the human-generated, rule-based heuristics intractable. In this paper, we tried to resolve the issue of current system. Thus, we proposed the new approach which can predict the swine flu diseases. Main purpose of this system to assist in medical diagnosis of Swine Flu. It is very effortless and on time process for patients to analyze disease based on clinic and laboratory symptoms and data to give the more accurate result of Swine Flu disease. Also it will help in early stage detection of diseases.

1.2 Objectives

As per public documents, Swine flu is the disease of applicability and wide spread of swine flu in the Country

- The H1N1 virus killed many people in last several years since 2009. In 2009, the count was 981.In 2010, it was 1763. The rate decreased up to 75 people in 2011 but again increased to 405,218 and 699 in year 2012, 2013 and 2014, respectively.
- According to health ministry of India by 30 march 2015, there are 31,974 reported cases and 1,895 people had lost their lives to the disease.
- In Gujarat, 6495 Cases were reported and 428 Deaths were occurred due to Swine flu by the date 30 March 2015.
- This disease is highly infectious and can be life threatening to humans. Chances of being epidemic in India are very high.
- The high virulence of Swine flu makes it eligible candidate for development of newer methods of detection for patient and public safety.

1.3 Purposes

Main purpose of this system to assist in medical diagnosis. Main purpose of system are below:

- It is very effortless and on time process for patients to analyze disease based on clinical and laboratory symptoms and give the more accurate result of **Swine Flu** disease.
- Early stage detection process

Chapter 2

Requirement Analysis

Requirement analysis explore project to find out potential resource requirements during development and deployment phase. Here I am listing out resources that are used during development phase.

2.1 Hardware Requirements

- Processor: Intel(R) Core(TM) i5 3rd Generation
- Hard disk Space: 30 GB
- RAM: 2 GB DDR3

2.2 Software Requirements

- OS: Windows 7 or above
- Language: R Programing

2.3 Data Requirements

- **Data set**

  We have required data set of Swine Flu for the simulations. Its data collected from the Civil Hospital Ahmadabad.

  There are two types of Data :

  1. **Actual Data** : It is Authentic data and secure data.

2. **Synthetic Data** : Synthetic data are generated to meet specific needs or certain conditions that may not be found in the original, real data. This can be useful when designing any type of system because the synthetic data are used as a simulation or as a theoretical value, situation, etc.

   - Synthetic data generated based on the authentic data using Mean Square Error(MSE). We have generated based on the equivalent the MSE of synthetic data and actual data.
     MSE = 1/n $\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$

- **Data Collections**

  The following details are collected from the suspected Swine flu patients:

  - Name: It is defined the patient's name.
  - Age: It is defined the age of the patient's in data set.
  - City: It is defined the city/village of the patient.
  - State: It is defined the state of the patient.
  - Clinical Symptoms: It is defined the Clinical symptoms like Fever, Fever Duration, Stuffy Nose, Sore Throat, Headache, Respiratory Distress.
  - Laboratory Indicators: It is defined the Laboratory symptoms like N/L Ratio and WBC Count.
  - Diagnostic result with Swine Flu Positive cases and Swine Flu Negative cases.

2.4 Swine Flu Detection time

- **Minor Points**

  - Sample collections from patients
  - Sample sending to respective laboratory

- **Major Points**
  - Sample Differentiation
  - Analytical Media Preparation
  - Analysis Time
  - Data Interpretation
  - Report Generation

Chapter 3

Literature Survey

Disease prediction is most important for medical system to make the best possible medical care decisions. Incorrect decisions are likely to purpose suspensions in medical treatment or even loss of life. A number of disease prediction models are used in medical diagnosis system which are using data mining and machine leaning techniques like

- Bayesian
- Decision tree
- Regression model
- Neural Network
- Single best model
- Ensemble model

In normal medical diagnosis system, it predicates the disease based on the patients symptoms and laboratory data before analyzing the disease.[1]This prediction techniques give the good performance and with less accuracy using medical dataset.

Neural Network is improving the observation capacity of information systems over the training of a limited number of neural networks nodes and collecting their results.Proposed system is enhancing the performance and training of neural networks for the classification with using cross-validation tool for the optimizing the network parameters and architecture.[1]

Using the artificial neural network (ANN) technique for disease prediction are compared below. Limitations of this scenario are as mentioned below: [1]

1. System is only using one data set for validation which does not predictable enough to generate outcomes.

2. System is only exploring the common predictable performance of their models without considering the F-score and precision as measures.

3. Most studies do not provide statistical test results to demonstrate the level of significance of their experimental results.

4. Most studies related to ensemble classifier do not compare the performance difference between individual classifiers and an ensemble classifier consisted of individual classifiers.[1]

Already some researched in automated imaginative systems for medical applications is an essential and impressive. The classification of automated decision support system is a feasible to collaborating physician rapidly and accurate diagnose patients.[2] Automated systems are useful in giving fast and accurate results. It is also helpful in reducing cost and time. It uses patient database to give enhanced results. [2]

The Fuzzy Min-Max network for medical diagnosis system shows how the Classification, Regression Tree (CART) and Random Forest (RF) models are integrated to make a hybrid intelligent system.[3]

Rule Extracting is the pros of the CART and it is in a tree based structure. It is not more flexible to perform same accuracy on medical data samples. Absences of capability of predictions is the pros of the FMM.[3]

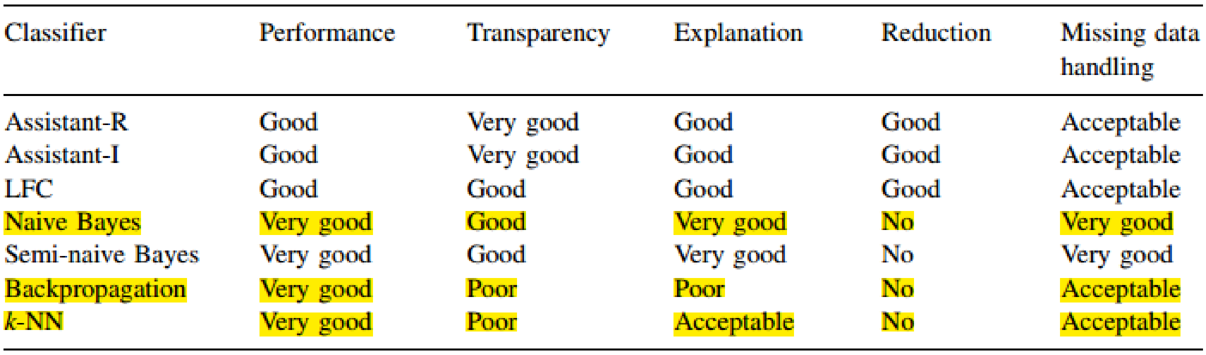
In medical system, accuracy of the DSS is more decisive. The prospective method is not gain the high precision, sensitivity, and specificity rates, but still to contribute description for its prognosis in the structure of a decision tree. [3]

Using **Fuzzy Hierarchical Approach to Medical Diagnosis we can improve the results by following ways::**[4]

- Complexity in diagnosis process.

- Simple fuzzy logic which does not provide hierarchical structure.

- Uncertainty occurred by different diagnosis system.



Classifier	Performance	Transparency	Explanation	Reduction	Missing data handling
Assistant-R	Good	Very good	Good	Good	Acceptable
Assistant-I	Good	Very good	Good	Good	Acceptable
LFC	Good	Good	Good	Good	Acceptable
Naive Bayes	Very good	Good	Very good	No	Very good
Semi-naive Bayes	Very good	Good	Very good	No	Very good
Backpropagation	Very good	Poor	Poor	No	Acceptable
k-NN	Very good	Poor	Acceptable	No	Acceptable

Figure 3.1: The Various Algorithm for Medical Diagnosis.[5]

- Problem occurred in other diagnosis system where grammatical labels comprehend to actual code in a period a numbers of values sensible process can be solved by this approach.

Using Limitation of the profession and aspect for medical diagnosis system using machine learning we can improve the result as it:[5]

- Provides an analysis of the automated data scrutiny.[6]

- Significance the naive Bayesian, neural network, decision trees.

- Specific requirement for machine learning systems[7]

    1. Good Performance

    2. Dealing with missing data

    3. Dealing with noisy data

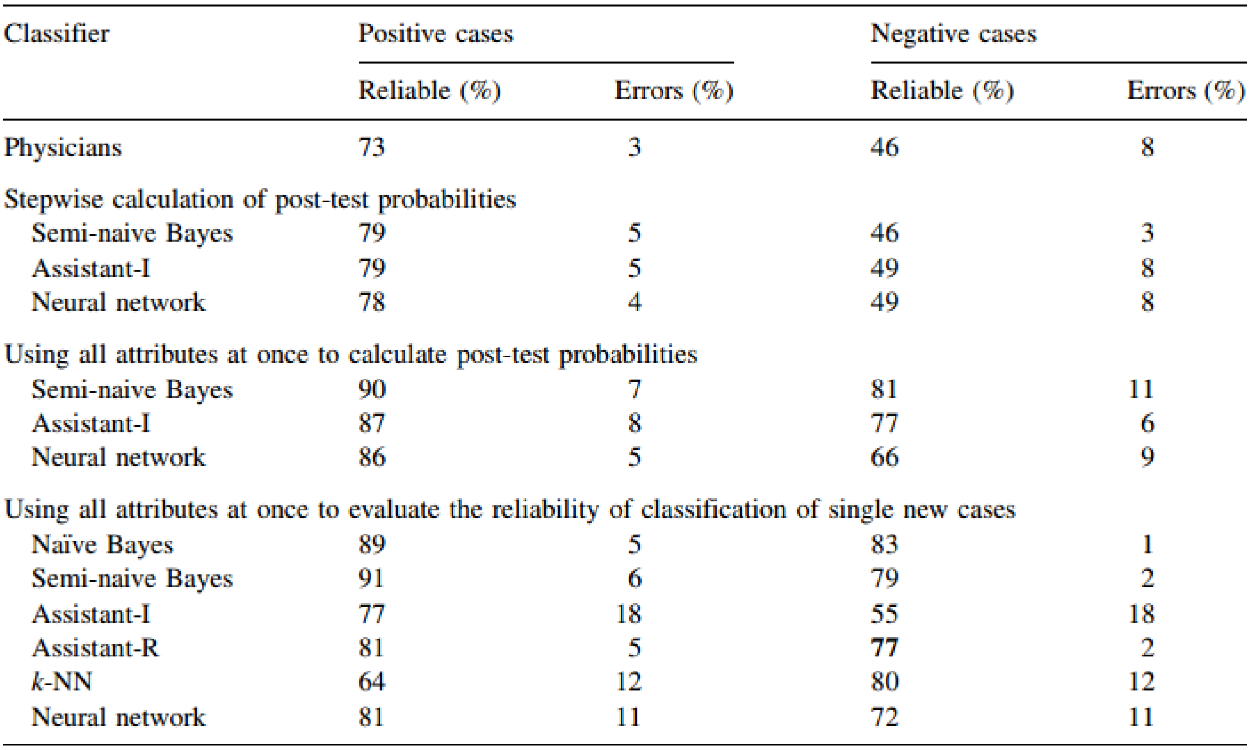
    4. Transparency of diagnostic knowledge

    5. Reduction of the number of tests

Predictive Models for Dengue Outbreak Using Multiple Rule-base Classifiers : Predictive models can be used for dengue outbreak detection. It is working as shown in the figure 3.3. Predictive model uses different rule based classifiers for detection. Classifiers such as

- Decision Tree

- Rough Set Classifier



Classifier	Positive cases		Negative cases	
	Reliable (%)	Errors (%)	Reliable (%)	Errors (%)
Physicians	73	3	46	8
Stepwise calculation of post-test probabilities				
Semi-naive Bayes	79	5	46	3
Assistant-I	79	5	49	8
Neural network	78	4	49	8
Using all attributes at once to calculate post-test probabilities				
Semi-naive Bayes	90	7	81	11
Assistant-I	87	8	77	6
Neural network	86	5	66	9
Using all attributes at once to evaluate the reliability of classification of single new cases				
Naïve Bayes	89	5	83	1
Semi-naive Bayes	91	6	79	2
Assistant-I	77	18	55	18
Assistant-R	81	5	**77**	2
k-NN	64	12	80	12
Neural network	81	11	72	11

Figure 3.2: Results of various classifiers for Heart-Disease Diagnosis.[5]

- Naive Bayes
- Associative Classifier are used for dengue detection.[8]

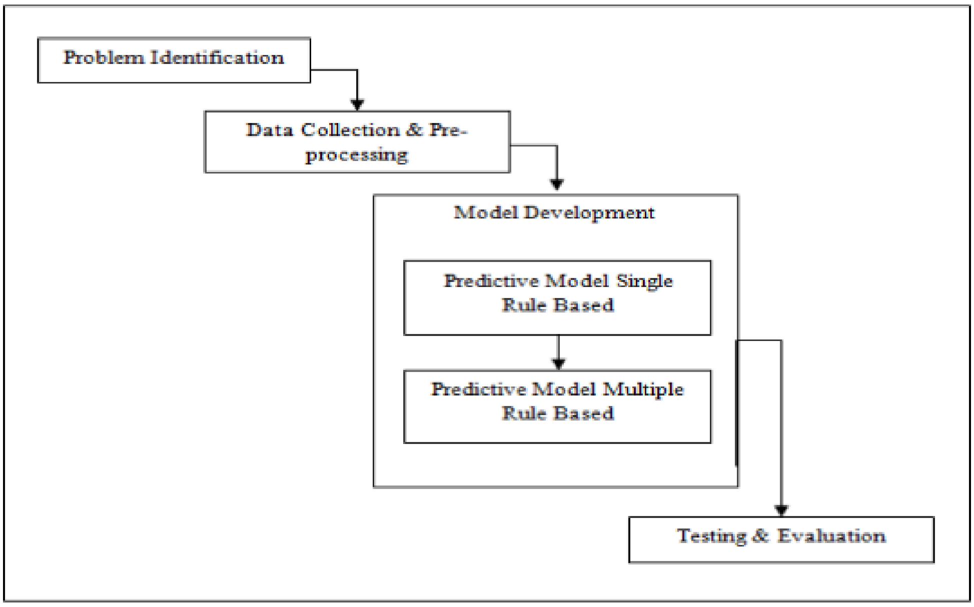

Problem Identification
Data Collection & Pre-processing
Model Development
Predictive Model Single Rule Based
Predictive Model Multiple Rule Based
Testing & Evaluation

Figure 3.3: Research Frame Work [8]

Naive Bayes Classifier, as per figure 3.4, is used previously for prediction of Swine Flu Disease.[9]

Swine flu transmission generally occurs due to sneeze or cough droplets of people.The droplets which contains virus are inhaled by other people and thus the transmission occurs. [9][10]

In second method which uses neonatal screening of dried blood spots and protein microarray to monitor the trends of the 2009 influenza A (H1N1) virus.[11]

Electrocardiogram (ECG) and auscultator blood pressure signals are used to examine two real world classification problems. The outcome of the computation of performance matrices such as accuracy, sensitivity, specificity and the area under receiver operating characteristic curve suggests that with original dataset logistic regression models are good but for noisy dataset, ensemble machine learning models are more appropriate. [12]

In one more approach in which researchers used the data mining along with non-wearable sensor hardware is also proposed to give the difference between on and off medication states.[13]

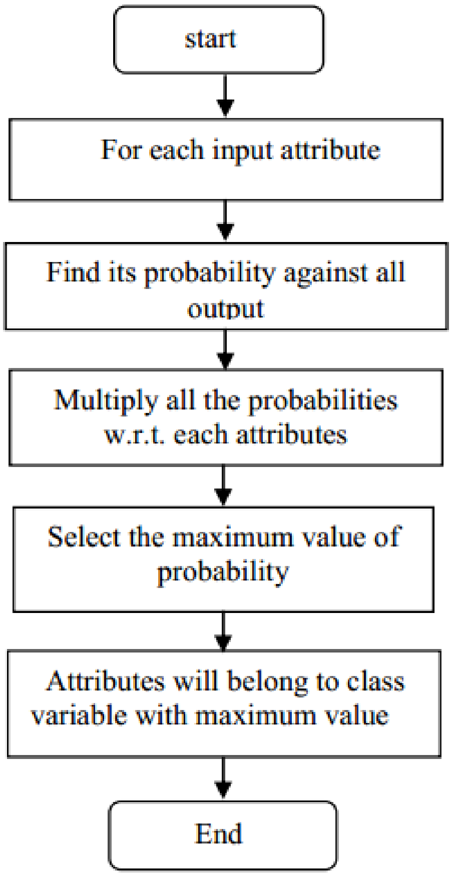

start
For each input attribute
Find its probability against all output
Multiply all the probabilities w.r.t. each attributes
Select the maximum value of probability
Attributes will belong to class variable with maximum value
End

Figure 3.4: Flow Chart of Naive Bayes Algorithm [9]

3.1 Existing Technologies

Various data mining techniques used in medical diagnosis system like :

- Data mining can be considered as Knowledge Discovery from Data (KDD) with three steps:[14]

    1. Data Pre-Processing
    2. Data Modeling
    3. Data Post-Processing

- Classification algorithms such as

    - Decision Tree
    - Decision Rule
    - Logistic Regression
    - Neural Network
    - Naive Bayes
    - Support Vector Machine(SVM)
    - k-nearest neighbour

- Intelligent classifiers for predictions, disease diagnostic, and screening tool of diabetes, breast cancer.

- In heart disease used the neural-fuzzy model.

- C4.5 Rule-PANE develop rules with more observation capability- A Case studies on diabetes,hepatitis,and breast cancer.[15]

Chapter 4

Proposed System

4.1 Existing data mining and Machine learning techniques which are used in medical field :

Techniques	Description
Support Vector Machine	SVM method works based on particular disease and gives the accurate result also. When it was used for another disease, it was not produced the accurate result because for modification in current algorithm is required for particular disease.
Nave Bayes	Probability based checked and found the disease. Using the class conditional probability. Which disease of probability high than it bias on that disease and give the result of the highly probable disease.
Decision Tree	Complexity increases. Time consuming process.
Random Forest	Forest is made using more than one decision tree. Splitting criteria of decision tree is random attribute selection. Random vectors are sampled independently and with equal distribution among all the trees. The class selection process includes voting of every tree and majority voting class is returned.
Clustering Method	Simultaneously change cluster based on symptoms. Its not give the accurate result for the number of diseases.
Logistic Regression	Recursive Procedure. Time consuming process.
Backpropagation	Predefined hidden units. Time complexity increase

Table 4.1: Techniques are used in medical data analysis

4.2 Research Design

The research will be conducted through two main phases. The following subsections will describe each phase briefly.

4.2.1 Phase 1: Dataset Processing

The processing of dataset was carried out on the collected datasets to better refine them to the requirement of the study. Many stages are involved in processing, some of this are: feature extraction,normalization, dataset division, and attribute weighting. These are very necessary in ensuring that the classifier can understand the dataset and properly classify them into the reference classes. The output of this phase is directly passed on to Phase 2 in evaluation the learning techniques.

4.2.2 Phase 2: Evaluation of Individual Machine Learning Techniques

Evaluation of learning techniques is required in this research to measure the performance achieved by a learning algorithm. To do this, a test set consisting of dataset with known labels is used. Each of the technique is trained with a training set, applied to the test set, and then measured the performance of by comparing the predicted labels with the true labels (that were not available to the training algorithm).Therefore, it is important to evaluate the classifiers by training and testing with the dataset obtained from Phase 1 using the following performance metrics; precision, recall, f1-score, and accuracy. The formula used is shown in Table 4.2.

Performance Measure		Description
Percentage Classification	Accuracy	Accuracy is the overall correctness of the model and is calculated as the sum of correct classifications divided by the total number of classifications $\frac{TN + TP}{TN + TP + FN + FP} \quad (4.1)$
	Precision	Precision is a measure of the accuracy provided that a specific class has been predicted $\frac{TP}{TP + FP} \quad (4.2)$
	Recall	Measuring the frequency of the correctly detected patterns as normal by the classifier. $\frac{TP}{TP + FN} \quad (4.3)$
	F-Measure	F1 score (also F-score or F-measure) is a measure of a tests accuracy. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. $2\frac{Precision.Recall}{Precision + Recall} \quad (4.4)$

Performance Measure		Description
Error Rate	False Positive Rate (FPR)	The average of normal patterns wrongly classified as malicious patterns. $$\frac{FP}{TN+FP} \tag{4.5}$$
	False Negative Rate (FNR)	The average of malicious patterns mistakenly classified as normal patterns. $$\frac{FN}{FP+FN} \tag{4.6}$$

Table 4.2: Formula Used to Calculate the Performance

	Actual Class(Observation)	
Expected Class(expectation)	TP(True Positive)Correct Result	FP(False Positive) Unexpected Result
	FN (False Negative) Missing Result	TN (True Negative) Correct Absence of Result

Table 4.3: Classification Context

4.3 Proposed Solution

- Here we are proposing the algorithm using the neural network with feed-forward network. Backpropagation algorithm is used for learning procedure and for training the multilayer feed-forward network.

- It can be utilized for purpose like medical diagnosis, pattern classification, image processing, character recognition etc.

- But Traditional approach of this algorithm is need to be determined the number of units in the hidden layer before training is started in the neural network.

- To overcome this difficulty many algorithms, that construct a network dynamically, had been proposed. Out of them, the well-known constructive algorithms are dynamic node creation (DNC), feed-forward neural network construction (FNNC) algorithm and the cascade correlation (CC) algorithm.

- We have used FNNC. In figure 4.1, we are displaying our algorithm flow. In first step, we are creating and initializing neural network with one hidden layer and attributes of dataset and give random weight to all the feature and hidden layer

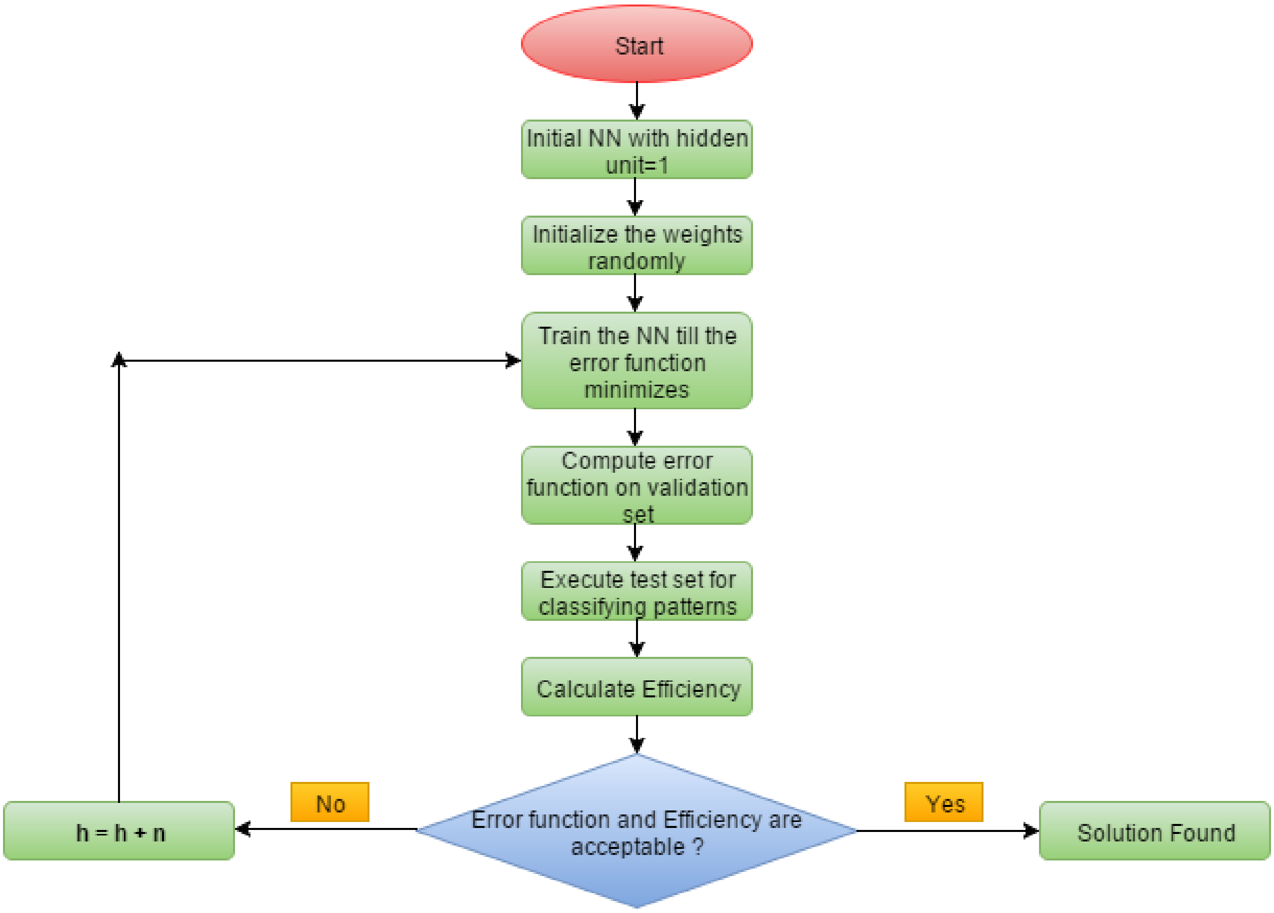

Start
Initial NN with hidden unit=1
Initialize the weights randomly
Train the NN till the error function minimizes
Compute error function on validation set
Execute test set for classifying patterns
Calculate Efficiency
Error function and Efficiency are acceptable ?
No
Yes
h = h + n
Solution Found

Figure 4.1: Flowchart of Proposed Algorithm

nodes links. Using the backpropagation on first hidden layer, we try to minimize the error function. Error function is used as mention below.

Error Function : E(n) = 1/2 $\sum$ k$\in$P ${e_k}^2$ (n)

Where e_k = desired output from k^{th} neuron - actual output from k^{th} neuron

- Once after minimizing the error function, we train the data and then calculate the error function on validate set and try to execute the test to generate the classify patterns with efficiency.

- If calculated efficiency is acceptable means we get accurate result then we stop the execution and find the predicted value else we will add n hidden layers and calculate weights again and initialize it until we get accurate result.

4.4 Comparison between Various Algorithm

Neural Network	**SVM**	**Nave Bayes**	**KNN**
Can be applied to many problems, as long as there is some data	Diagonal separation line	Missing value	Arbitrary decision boundaries
If there is a pattern, then neural networks should quickly work it out, even if the data is noisy	Appropriate for high dimensional data	Accuracy degraded by correlated attribute	Prediction based on local data
There are many free parameters, such as the number of hidden nodes, the learning rate, minimal error,which may greatly influence the final result.	Parametric	Required to determine initial probability	Required similarity measurement, parametric

Table 4.4: Comparison of various Algorithm

Chapter 5

Implementation and Result

5.1 Introduction

The process of dataset processing, and dataset division.We have gathered the dataset of swine flu.Important attributes in our dataset like

1. Fever
2. Fever Duration
3. Stuffy Nose
4. Sore Throat
5. Headache
6. Respiratory Distress
7. N/L Ratio
8. WBC Count

After data processing, the dataset is divided into three sets for training and testing purpose and to investigate the accuracy of result. Two steps of data division are used, the first step is to divide the data into three different groups, and then choose different percentage of training and testing for each group, for first group 50% training and remaining 50% testing, second group 70% training and 30% testing, and the last group 30% training and 70% are testing.Furthermore, a cross-validation of 10% is used to estimate predictive performance of the selected attribute set is used.This chapter addresses the problem of selecting the best machine learning technique for medical diagnosis disease prediction that causes degradation in prediction performance. The main objective of this chapter is to train and test the individual machine learning techniques (SVM, Naive Bayes, KNN, and Neural Network) with the same dataset, implements on the learning techniques, and comparison between the different-different techniques.

Tables 5.1 - 5.5 shows the accuracy, precision and recall of the SVM, Nave Bayes and KNN algorithms were trained and tested across the three sets of dataset and the resulting output of this process. Corresponding charts of the result obtained are shown in Figures 5.1 - 5.5.

Set	Actual Data	Synthetic Data
A	78.00%	71.00%
B	81.00%	84.00%
C	64.00%	68.00%

Table 5.1: Accuracy of Naive Bayes in Varying Dataset

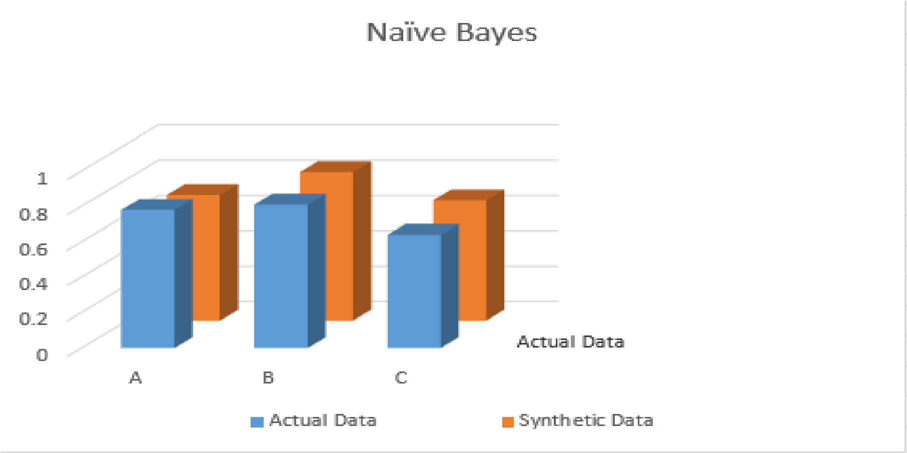


Figure 5.1: Plot of Accuracy Across Varying Dataset

Set	Actual Data	Synthetic Data
A	60.29%	64.00%
B	71.00%	72.00%
C	28.00%	32.00%

Table 5.2: Accuracy of SVM in Varying Dataset

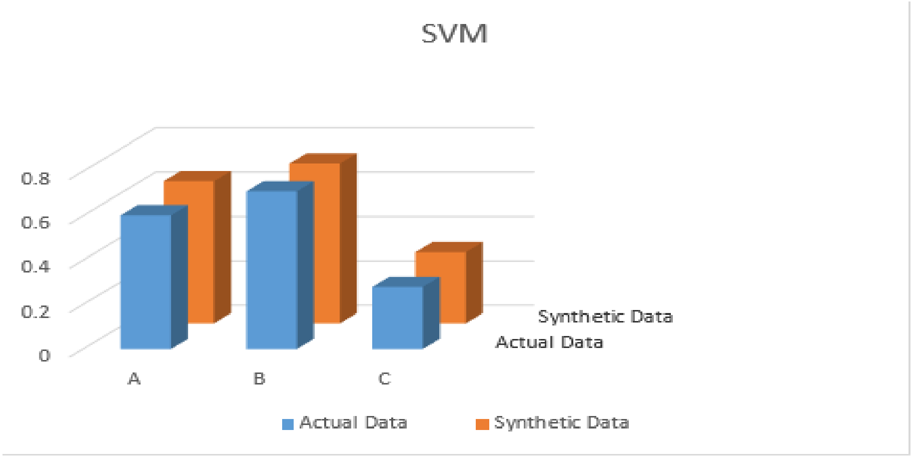


Figure 5.2: Plot of Accuracy Across Varying Dataset

Set	**Actual Data**	**Synthetic Data**
A	66.00%	69.74%
B	71.59%	73.32%
C	22.29%	33.84%

Table 5.3: Accuracy of KNN in Varying Dataset

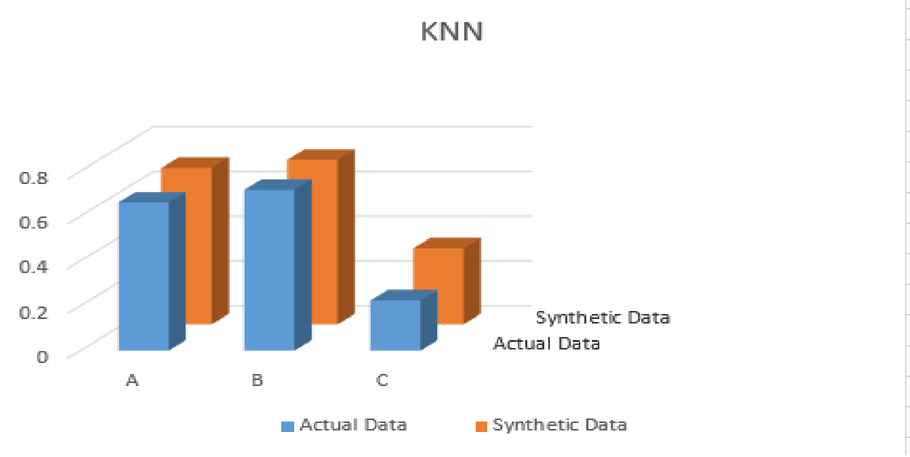


Figure 5.3: Plot of Accuracy Across Varying Dataset

Set	**SVM**	**NB**	**KNN**
A	61.62%	79.11%	64.71%
B	69.38%	78.57%	72
C	27.31%	62.38%	24

Table 5.4: Precision of Individual Technique in Varying Dataset

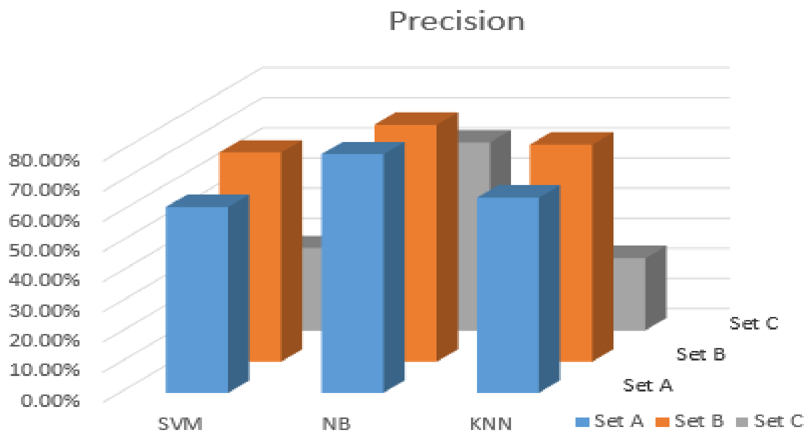


Figure 5.4: Plot of Precision Across Varying Dataset

Set	SVM	NB	KNN
A	61.12%	78.52%	63.67%
B	68.29%	78.80%	71.88
C	27.02%	62.15%	24.31

Table 5.5: Recall of Individual Technique in Varying Dataset

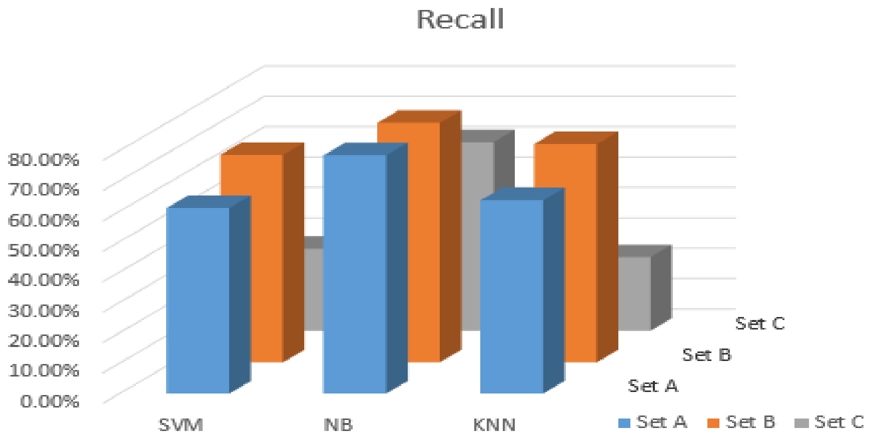


Figure 5.5: Plot of Recall Across Varying Dataset

In Neural Network processing of dataset then based on the script training and testing the data and get the best performance output for proposed algorithm.Show Table 5.6 - 5.9 Here, We have tested using the threshold 0.1 - 0.4.Corresponding charts of the result obtained are shown in Figures 5.6 - 5.9.Based on the output continuously changes in output retrain the data and get the different-different output between threshold value 0.1 - 0.4.

Hidden Neurons	Threshold Value	Iterations	Performance
1	0.1	0	0.2979
2	0.1	0	0.4227
3	0.1	0	0.3694
4	0.1	0	0.3409
5	0.1	0	0.3356
6	0.1	0	0.2512
7	0.1	0	0.3052
8	0.1	0	0.2683
9	0.1	0	0.3733
10	0.1	0	0.4038
15	0.1	0	0.3872
20	0.1	0	0.3440
25	0.1	0	0.3747
50	0.1	0	0.4290
70	0.1	0	0.2972
100	0.1	0	0.3233
150	0.1	0	0.3360
200	0.1	0	0.2960

Table 5.6: Performance of the Threshold value = 0.1

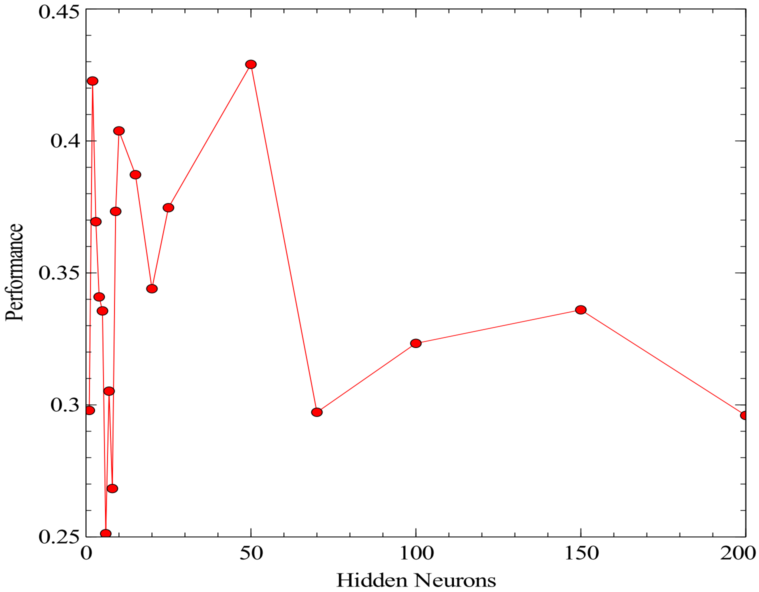


Figure 5.6: Threshold Value-0.1

As Shown in figure 5.10, there are 255 neurons and threshold value is 0.58.Because Threshold value between 0.1 - 0.4 continuously changes in output. Because Changing in threshold gives divergence in output.At the threshold 0.58 get convergence in result. After the processing is complete there are 278 neurons needed to satisfy the threshold value.so there are 23 new neurons added to retrain the system..shows in Figures 5.11 - 5.13 give

Hidden Neurons	Threshold Value	Iterations	Performance
1	0.2	0	0.3564
2	0.2	0	0.3378
3	0.2	0	0.3587
4	0.2	0	0.2631
5	0.2	0	0.2864
6	0.2	0	0.2626
7	0.2	0	0.2708
8	0.2	0	0.3446
9	0.2	0	0.3152
10	0.2	0	0.2745
15	0.2	0	0.3031
20	0.2	0	0.3331
25	0.2	0	0.3090
50	0.2	0	0.3515
70	0.2	0	0.3274
100	0.2	0	0.2638
150	0.2	0	0.3857
200	0.2	0	0.3761

Table 5.7: Performance of the Threshold value = 0.2

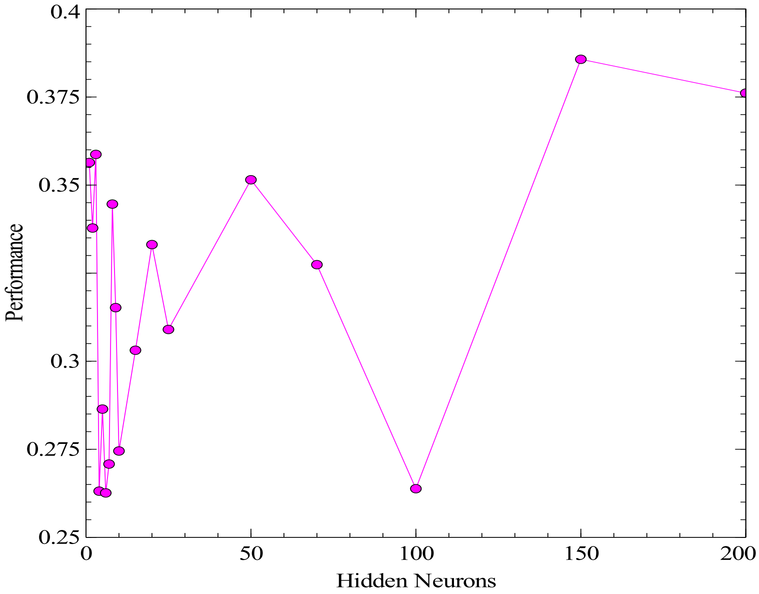


Figure 5.7: Threshold Value-0.2

the performance at the threshold point, generate the confusion matrix at the threshold point and generate the best error histogram chart at the threshold point respectively.

Hidden Neurons	**Threshold Value**	**Iterations**	**Performance**
1	0.3	0	0.3510
2	0.3	0	0.3675
3	0.3	2	0.3256
3	0.3	0	0.3404
4	0.3	0	0.3013
5	0.3	0	0.4442
6	0.3	0	0.3676
7	0.3	0	0.3495
8	0.3	0	0.3998
9	0.3	2	0.3057
9	0.3	0	0.3687
10	0.3	2	0.3173
10	0.3	0	0.3543
15	0.3	0	0.4444
20	0.3	1	0.3596
20	0.3	1	0.3516
20	0.3	0	0.3918
20	0.3	0	0.3791
20	0.3	0	0.3176
50	0.3	0	0.3088
70	0.3	1	0.4203
70	0.3	1	0.4110
100	0.3	0	0.3048
150	0.3	1	0.3695
150	0.3	0	0.3541
200	0.3	0	0.4234

Table 5.8: Performance of the Threshold value = 0.3

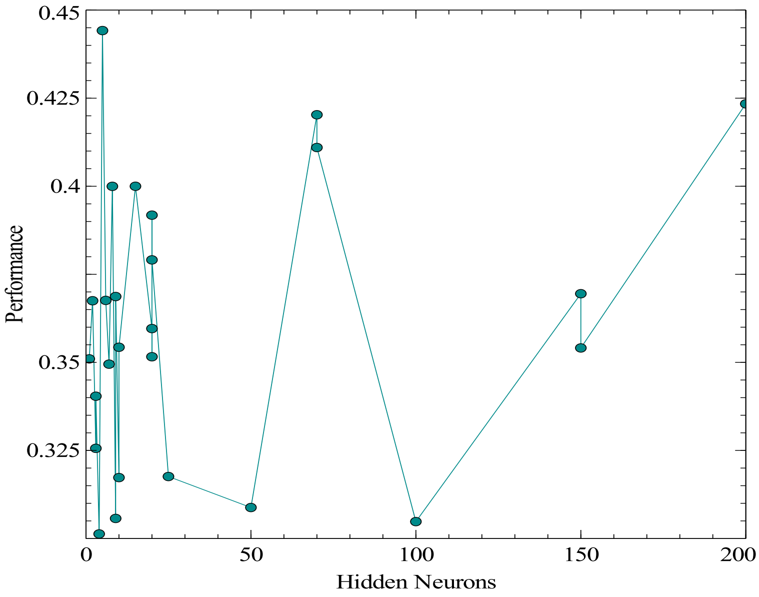


Figure 5.8: Threshold Value-0.3

Hidden Neurons	**Threshold Value**	**Iterations**	**Performance**
1	0.4	19	0.4397
2	0.4	16	0.4362
2	0.4	18	0.4616
3	0.4	4	0.4059
3	0.4	0	0.4026
4	0.4	4	0.4264
4	0.4	13	0.4224
4	0.4	8	0.4993
5	0.4	4	0.4191
5	0.4	10	0.4295
5	0.4	11	0.4379
6	0.4	4	0.4031
6	0.4	1	0.4555
7	0.4	3	0.4538
7	0.4	10	0.4715
8	0.4	3	0.4731
8	0.4	1	0.4219
8	0.4	19	0.4142
9	0.4	7	0.4309
9	0.4	0	0.4164
10	0.4	1	0.4482
10	0.4	1	0.4622
15	0.4	1	0.4257
15	0.4	0	0.4712
20	0.4	2	0.4566
20	0.4	0	0.4306
25	0.4	0	0.4769
50	0.4	1	0.4188
50	0.4	0	0.4123
70	0.4	10	0.4549
70	0.4	1	0.4695
70	0.4	4	0.4432
100	0.4	2	0.4010
100	0.4	36	0.4676
150	0.4	4	0.4557
150	0.4	2	0.4282
200	0.4	0	0.4456

Table 5.9: Performance of the Threshold value = 0.4

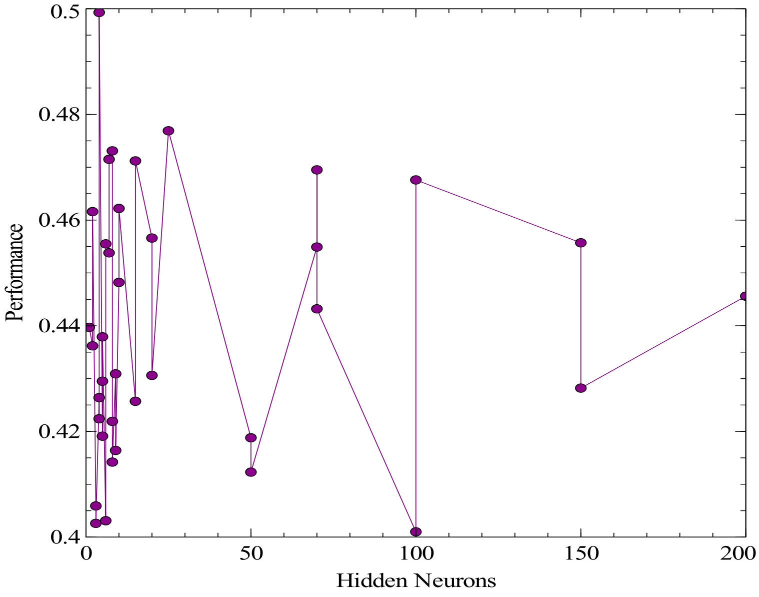
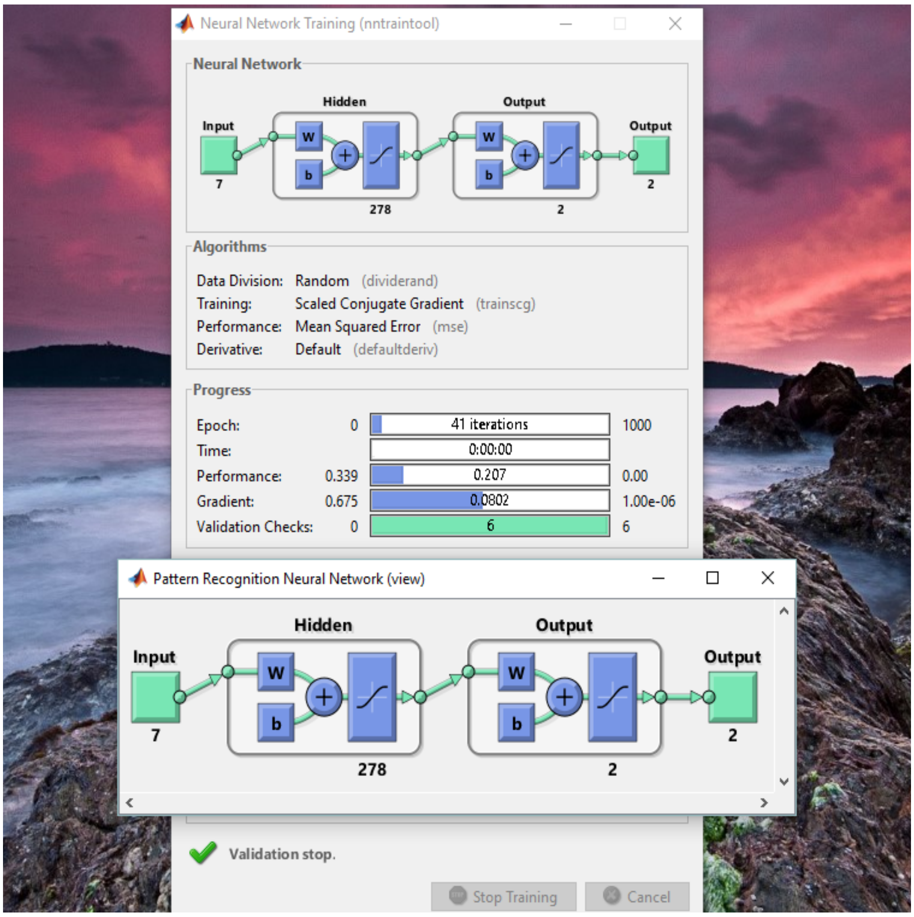


Figure 5.9: Threshold Value-0.4



Algorithms

Data Division: Random (dividerand)
Training: Scaled Conjugate Gradient (trainscg)
Performance: Mean Squared Error (mse)
Derivative: Default (defaultderiv)

Progress

Epoch:	0	41 iterations	1000
Time:		0:00:00	
Performance:	0.339	0.207	0.00
Gradient:	0.675	0.0802	1.00e-06
Validation Checks:	0	6	6

✓ Validation stop.

Figure 5.10: Simulation for the threshold value = 0.58

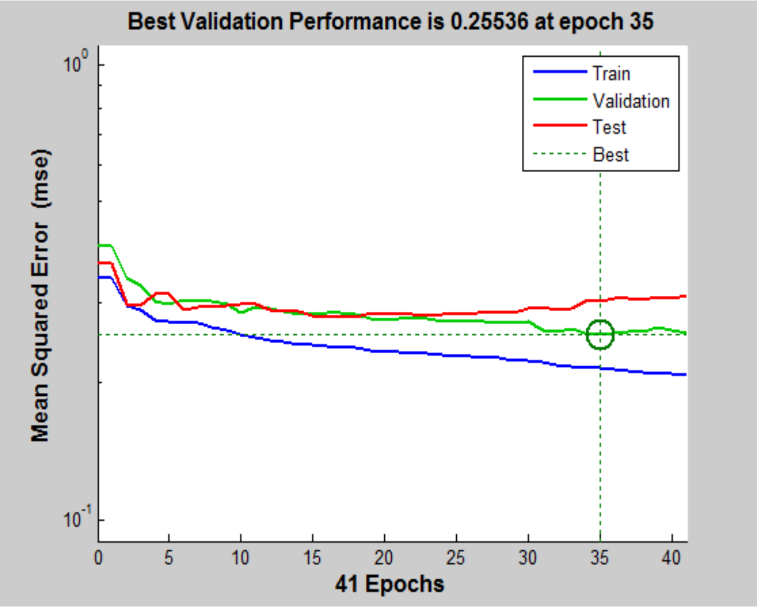


Figure 5.11: Performance of the threshold value = 0.58

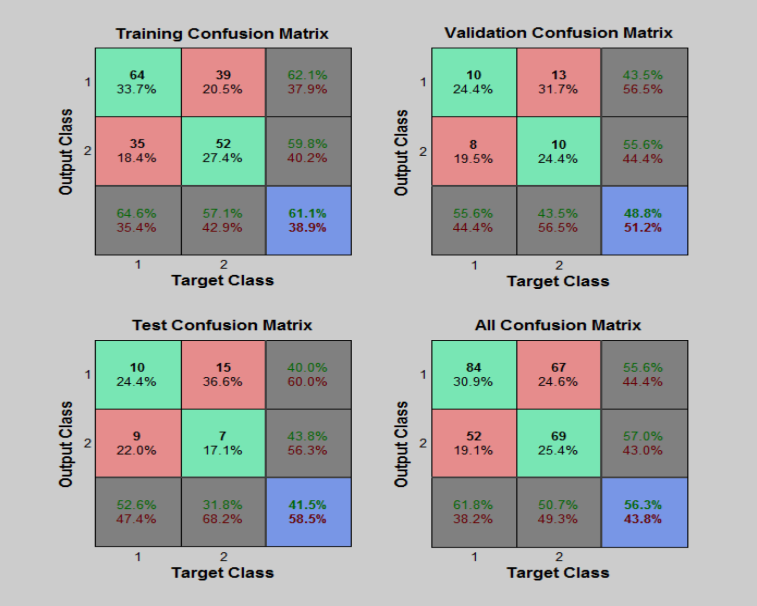


Figure 5.12: Confusion Matrix of the threshold value = 0.58

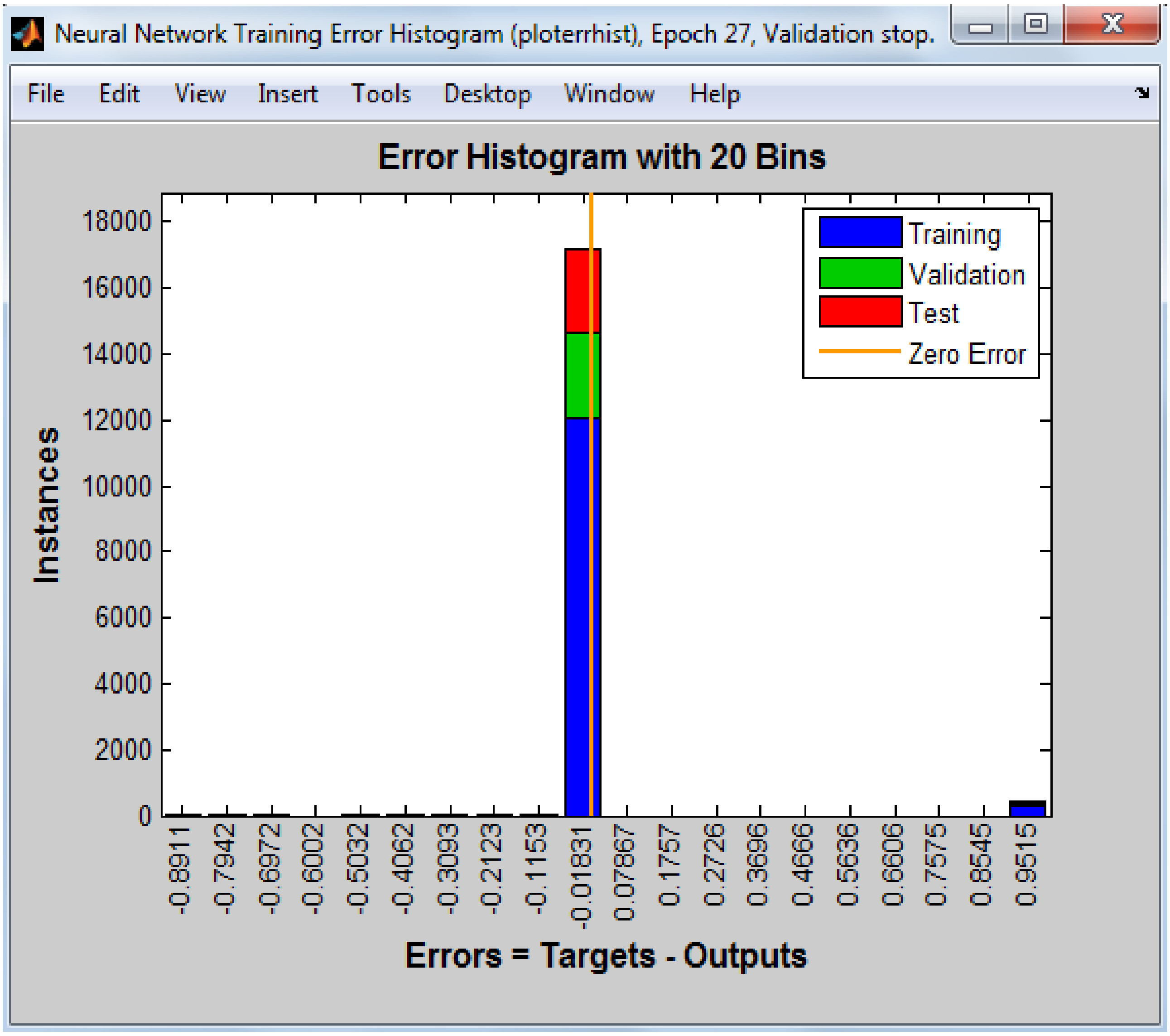


Figure 5.13: Error Histogram of the threshold value = 0.58

5.2 Comparative Study

A comparative study between the SVM,Naive Bayes,KNN and Neural Network are discussed in this section. Table 5.10 - 5.12 shows the results of the accuracy,precision and recall with three set. Also, Figure 5.14 - 5.16 shows the plot of the best performed algorithms in SVM, Naive Bayes, KNN and Neural Network methods, respectively. Table 5.10 - 5.12 shows that NN algorithm performs better in accuracy than the Naive Bayes, SVM and KNN, in Table 5.10 the precision of NB is higher than the SVM,KNN and NN, in Table 5.11 - 5.12 the precision of NN is higher than the SVM, NB, KNN.

Metrics	**SVM**	**NB**	**KNN**	**Propose Method**
Accuracy	60.29	73.00%	66.00%	78.8%
Precision	61.62	78.11%	64.71%	71.72
Recall	61.12	78.52%	63.67%	68.84

Table 5.10: Comparison between each algorithm for SetA

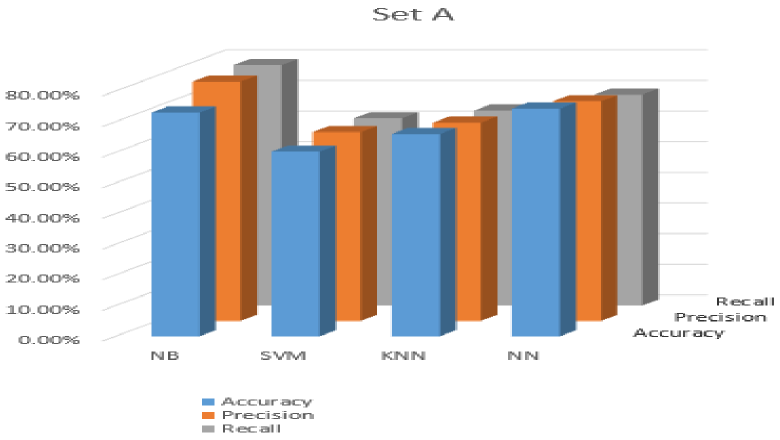


Figure 5.14: Plot of Comparison between each algorithm for SetA

Metrics	**SVM**	**NB**	**KNN**	**Propose Method**
Accuracy	71.00	80.00%	71.59%	82.29%
Precision	69.38	78.57%	72.00%	81.78
Recall	68.29	78.80%	71.88%	80.11

Table 5.11: Comparison between each algorithm for SetB

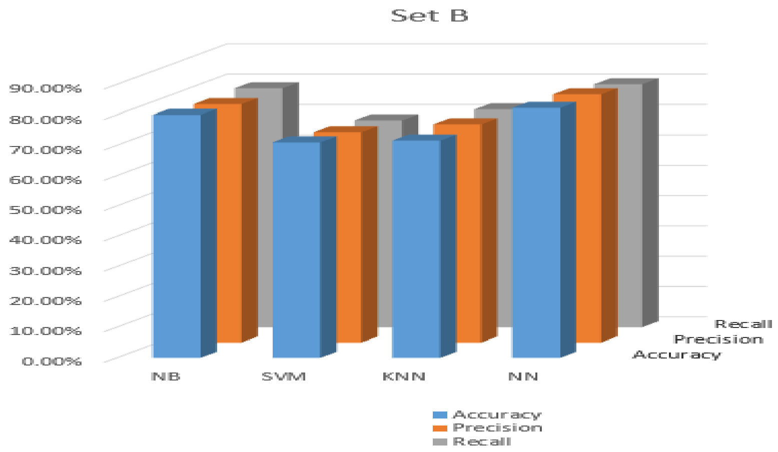


Figure 5.15: Plot of Comparison between each algorithm for SetB

Metrics	SVM	NB	KNN	Propose Method
Accuracy	28.00%	62.58%	22.29%	68.81%
Precision	27.31%	62.38%	24.00%	66.32
Recall	27.02%	62.15%	24.31%	64.28

Table 5.12: Comparison between each algorithm for SetC

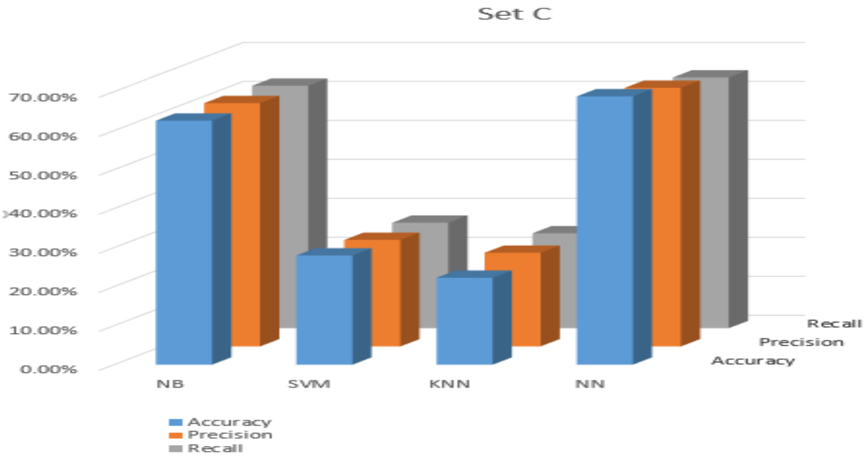


Figure 5.16: Plot of Comparison between each algorithm for SetC

Chapter 6

Conclusion

Phase 1 focuses on dataset gathering and preprocessing. The objective is to process data for use in Part 2.The dataset is validated first after gathering, then normalized, features extraction and finally dataset division. Selected Features for this,to ensure an optimum result from the machine learning techniques.

Phase 2 focuses on proposed system and implementation of training and validating model using machine learning techniques. A defined performance metrics is used as a measurement of accuracy, precision, recall, and f-measure. The objective of this phase is to test the performance of individual learning techniques in the pool of varying dataset as divided and select the most performed of all the techniques.

Phase 3 focuses on the proposed algorithm and implementation based on the proposed algorithm, and comparative study between SVM, Naive Bayes, KNN and Neural Network,also get the result for particular method.Design the Neural Network and if the size of the dataset is increased, the NN will most likely perform better than the other algorithm.

Bibliography

[1] C.-H. Weng, T. C.-K. Huang, and R.-P. Han, "Disease prediction with different types of neural network classifiers," *Telematics and Informatics*, vol. 33, no. 2, pp. 277–292, 2016.

[2] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, 2011.

[3] C. L. ManjeevanSeera, "A hybrid intelligent system for medical data classification," *Expert Systems with applications41 (5)*, pp. 2239–2249, 2014.

[4] S. Zahan, C. Michael, and S. Nikolakeas, "A fuzzy hierarchical approach to medical diagnosis," in *Fuzzy Systems, 1997., Proceedings of the Sixth IEEE International Conference on*, vol. 1, pp. 319–324, IEEE, 1997.

[5] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in medicine*, vol. 23, no. 1, pp. 89–109, 2001.

[6] R. Isola, R. Carvalho, M. Iyer, and A. K. Tripathy, "Automated differential diagnosis in medical systems using neural networks, knn and som," in *Developments in E-systems Engineering (DeSE), 2011*, pp. 62–67, IEEE, 2011.

[7] V. S. H. Rao and M. N. Kumar, "A new intelligence-based approach for computer-aided diagnosis of dengue fever," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 16, no. 1, pp. 112–118, 2012.

[8] A. A. Bakar, Z. Kefli, S. Abdullah, and M. Sahani, "Predictive models for dengue outbreak using multiple rulebase classifiers," in *Electrical Engineering and Informatics (ICEEI), 2011 International Conference on*, pp. 1–6, IEEE, 2011.

[9] A. Borkar and P. Deshmukh, "Naive bayes classifier for prediction of swine flu disease," vol. 5, pp. 120–123, 2015.

[10] B. Thakkar, M. Hasan, M. Desai, *et al.*, "Health care decision support system for swine flu prediction using naive bayes classifier," in *Advances in Recent Technologies in Communication and Computing (ARTCom), 2010 International Conference on*, pp. 101–105, IEEE, 2010.

[11] E. de Bruin, J. Loeber, A. Meijer, G. M. Castillo, M. G. Cepeda, M. R. Torres-Sepúlveda, G. Borrajo, M. Caggana, Y. Giguere, M. Meyer, *et al.*, "Evolution of an influenza pandemic in 13 countries from 5 continents monitored by protein microarray from neonatal screening bloodspots," *Journal of Clinical Virology*, vol. 61, no. 1, pp. 74–80, 2014.

[12] M. Seera, C. P. Lim, W. S. Liew, E. Lim, and C. K. Loo, "Classification of electrocardiogram and auscultatory blood pressure signals using machine learning models," *Expert Systems with Applications*, vol. 42, no. 7, pp. 3643–3652, 2015.

[13] C. S. Tucker, I. Behoora, H. B. Nembhard, M. Lewis, N. W. Sterling, and X. Huang, "Machine learning classification of medication adherence in patients with movement disorders using non-wearable sensors," *Computers in Biology and Medicine*, vol. 66, pp. 120–134, 2015.

[14] L. Hu, G. Hong, J. Ma, X. Wang, and H. Chen, "An efficient machine learning approach for diagnosis of paraquat-poisoned patients," *Computers in biology and medicine*, vol. 59, pp. 116–124, 2015.

[15] Z.-H. Zhou and Y. Jiang, "Medical diagnosis with c4. 5 rule preceded by artificial neural network ensemble," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 7, no. 1, pp. 37–42, 2003.