DESIGN OF AN EFFICIENT CLASSIFIER FOR BIOINFORMATICS APPLICATION

A Dissertation Submitted in partial fulfillment of the requirement

for the award of degree of

MASTER OF ENGINEERING

in

WIRELESS COMMUNICATION

Submitted By

DEEPIKA GARG

Roll. No: 801563003

Under Supervision of

Dr. Amit Mishra

Assistant Professor, ECED

ELECTRONICS AND COMMUNICATION ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA - 147004 (PUNJAB)

JUNE, 2017

DECLARATION

I, **Deepika Garg** hereby declare that the work presented in this thesis entitled **"DESIGN OF AN EFFICIENT CLASSIFIER FOR BIOINFORMATICS APPLICATION"** for the award of degree of Master of Engineering in Wireless Communication Engineering, submitted to Electronics and Communication Department (ECED), Thapar University, Patiala is an authentic record of my own work carried out under the supervision of **Dr. Amit Mishra** Assistant Professor, ECED, Thapar University, Patiala).The matter presented in this has not been submitted either in part of full to any other university or institute for the award of any other degree.

Date : July 17, 2017

Deepika Garg
801563003

It is certified that the above statement made by the student is correct to the best of my knowledge and belief.

Dr. Amit Mishra
Assistant Professor
ECED
Thapar University, Patiala

ACKNOWLEDGEMENT

Simply stated, there is no way I could have ever finished this thesis on my own.

I address my sincere thanks to Almighty God for giving me the inner strength to complete my dissertation and he was always there protecting and saving me.

I would like to thank my supervisor, **Dr. Amit Mishra**, Assistant Professor, Electronics & Communication Engineering Department (ECED), Thapar University, Patiala, for his responsible scientific coordination, guidance in research methodology & for providing stimulating discussions based on the related work.

I express my deep gratitude for all his suggestions and direction he provided for exploring the new avenues of research and immense concern throughout the project work.

I would also like to thank **Dr. Alpana Aggarwal**, Head (ECED), **Dr. Hemdutt Joshi**, P.G Coordinator and **Dr. Ashutosh Kumar Singh**, Programme Coordinator, M.E. (WC) for providing adequate environment for carrying out the work

Last but not least, I would like to extend my gratitude to my parents for their continuous inspiration and support. I would also like to thank all those persons who helped me directly or indirectly during the development of this work.

….Deepika Garg

801563003

ABSTRACT

In today's integrated world, solutions for problems are inter disciplinary by nature. Soft computing proves to be potent for obtaining solutions accurately, and quickly. Moreover, the combination or hybrid of one of more methodology has resulted into the new class of system called Ensemble Models or Hybrid models. This has resulted in creating a classifier with better efficiency used for the design of intelligent systems.

The study employed various machine learning techniques for the prediction of diseases. Particularly, the study applies the several data mining techniques: Decision tree, Neural Networks, Support Vector Machine, Linear Regression, Linear Discriminant Analysis, Naïve Bayes and k-nearest neighbor.

The study deals with the focus on improving the accuracy of classification of machine learning algorithms. Undoubtedly, Support Vector Machine has provided the better results as compare to the other techniques in classification. In this thesis, various parameters of SVM have been exploited in order to get the best results.

The proposed research converges on the hybrid of heterogeneous classifiers for disease prediction. The performance of a classifier is judged by two parameters namely Classification accuracy and Simulation Time.

Another effort has been put up to optimize the classification technique such as SVM using Genetic algorithm in order to get the best fit value. Many techniques have been applied on various diseases dataset. One such technique employed SVM which considered ovarian cancer features for the optimization process.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ABC	:	Artificial Bee Colony
AI	:	Artificial Intelligence
ANN	:	Artificial Neural Network
BN	:	Bayesian Networks
BrNdT	:	Bayesian regularized Neural network decision Tree
BRNN	:	Bayesian Regularized Neural Network
BUPA	:	British United Provident Association
CpG	:	-C-phosphate-G-
CGI	:	Vertebrate CpG islands
DE	:	Differential Evolution
DNA	:	Deoxyribonucleic Acid
DT	:	Decision Tree
EISPACK	:	Eigen System Package
FN	:	False Negative
FP	:	False Positive
FPR	:	False Positive rate
GA	:	Genetic Algorithm
GDA	:	Generalized Discriminant Analysis
HPM	:	Hybrid Prediction Model
ILPD	:	Indian Liver Patient Dataset
k-NN	:	k-Nearest Neighbor
LDA	:	Linear Discriminant Analysis
LINPACK	:	Linear system Package
LS-SVM	:	Least Square Support Vector Machine
MATLAB	:	Matrix Laboratory
MLP	:	Multilayer Perceptron
MRMR	:	Maximum Relevance and Maximum Redundancy
NB	:	Naïve Bayes
NN	:	Neural Networks
PV	:	Plurality Voting
RBF	:	Radial Basis Function
RF	:	Random Forest

ROC	:	Relative Operating Characteristics
RSA	:	Recurrence Surface Approximation
TN	:	True Negative
TP	:	True Positive
TPR	:	True Positive Rate
SVM	:	Support Vector Machine
UCI	:	University of California, Irvine
WHO	:	World Health Organization

CHAPTER - I
INTRODUCTION

1.1 Preamble

Over few decades, healthcare domain has been employing information technology due to its massive data; for the purposes like information storing such as details of patients, and many more. Not only, the task of data collection and its storage but analysis also has been extensively improved, with the help of these machine learning and data mining techniques. Data mining, supposed to work well with such voluminous data. The challenging task is to use data mining in foreseeing the consequences of the particular disease. Data mining classification technique are used for data instances for predicting the group membership.

Artificial Intelligence (AI) is a domain of computer science involved with designing of intelligent computer systems. AI model and implement the theories of intelligence for developing techniques for intelligent problem solving. Artificial Intelligence has developed various tools for the problem solution. Most general methods are shown in Fig.1.1.

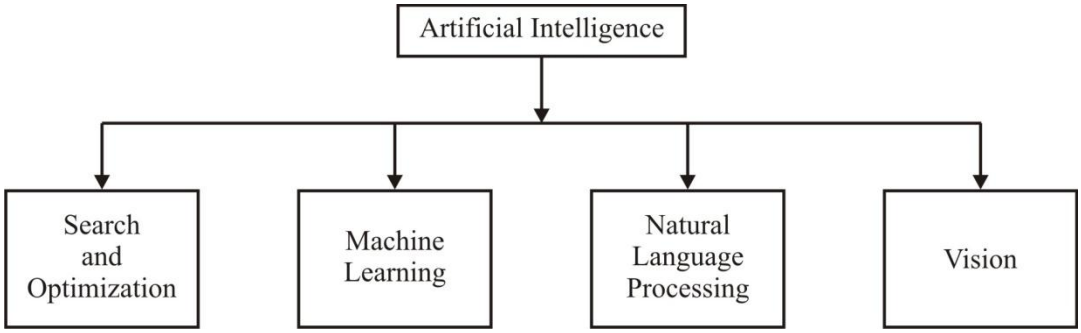


Figure 1.1: Common tools of Artificial Intelligence.

Soft computing is referred to as computational intelligence also. It varies from hard computing (conventional) in the terms of uncertainty, tolerance to imprecision and partial truth. Hard computing methods are based on mathematical approaches, therefore requires high degree of accuracy and precision. In most engineering problems, however input parameters can't be predicted with high accuracy, leading to better estimates of parameters used to solve problems.

The intrinsic characteristics of soft computing techniques have been drawn from biological systems. Soft computing exploits the uncertainty, tolerance for imprecision and partial truth in order to obtain the robust, tractable and low cost solution.

The major components of soft learning are as follows:

- Fuzzy logic
- Evolutionary Computation
- Machine Learning
- Probabilistic Reasoning

Machine learning is the subfield of artificial intelligence which is used in creating the algorithms that can learn and make predictions on the data. It is related with the computational analysis.

1.2 Statistical Data

According to the statistics, India is one among the top 3 countries having high rate of diabetic population. India has more type-2 diabetic patients. The World Health Organization (WHO) has estimated that deaths due to diabetes will be doubled between 2016 and 2030 and presented in Table.1.1.

Table 1.1: Diabetes Statistics (Source: International Diabetes Federation)

Country	2030 [million]
China	129.7
India	101.2
USA	29.6
Brazil	19.6
Bangladesh	16.8

As per the Cancer Statistics 2017, there is an estimation of 23.6M fresh cases of cancer by 2030. The worldwide most common cancers are lung cancer, female breast, bowel and prostate cancer and cancer rates are given in Table 1.2.

Table 1.2: Global Cancer Rates (Source: World Cancer Research Foundation)

Country	Age-Standardized Rate per 100,000 (World)
Denmark	338.1
France	324.6
Australia	323.0
Belgium	321.1
Norway	318.3
United States of America	318.0

As per the latest statistics of India, about 10L cases of liver cirrhosis are diagnosed every year. As per the WHO, Liver disease is ranked as 10^{th} most frequent cause of death in India. Around the world, Liver Cirrhosis is considered as 14^{th} popular cause of death. By2020, it would be 12^{th} leading cause.

1.3 Terms and Concepts

As there is still ongoing development in the field of artificial intelligence, some terms may have been misinterpreted. For avoiding this misinterpretation, several terms used in this dissertation are explained briefly as under.

Sample : It is a set of the feature data and also possesses a label which helps in deciding the category in which it falls under.

Classification : It is a process of predicting the label of the sample.

Over fitting : It is the biasing in training set towards the samples.

Training set : It is a form of supervised learning in which model is trained to get a desired output.

Testing set : It is used to test, whether the model is trained well as per the classification rule.

1.4 Types of learning methods in Machine learning

In machine learning, learning methods are classified into three types based upon the type of the learning signal and feedback available as shown in Fig.1.2.

- Supervised
- Unsupervised
- Reinforced

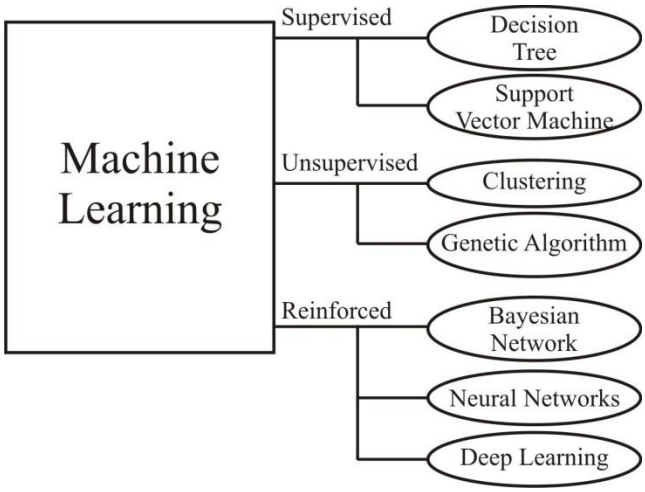


Figure 1.2: Types of Machine Learning

1.4.1 Supervised Learning

The system is made learned to achieve the desired outputs. The basic goal is to make the model learn a general rule which aids in mapping inputs to outputs.

Examples include Support vector machines, Linear Discriminant analysis, Naïve Bayes, k-NN and Decision tree.

1.4.2 Unsupervised Learning

Unsupervised Learning doesn't provide any labeling to the algorithm. It is task of deducing a function that describes the hidden structure from the unlabeled data.

Examples include k- means clustering, neural networks and hierarchical clustering.

1.4.3 Reinforced Learning

In reinforced learning, environment is provided such as to attain a specific goal. The input/output pair which is correct is never provided. It uses samples for making the performance optimize and function approximation for dealing with the large environments.

1.5 Schemes for classification in Machine learning

Classification is one of the significant decision making tool in medical sciences. Researchers have developed several classification schemes. The most commonly schemes used are Neural Networks, Bayesian regularized neural network, k nearest neighbor, radial basis function, decision tree, support vector machine and Fuzzy C-means. However, there is not a single methodology which has highest performance for every dataset/disease. When one classifier has good performance for given dataset, then it may not have that much good performance for other datasets also. Classification is a kind of supervise leaning in which the training set of the correctly identified observations are available. In classifiers, there are two types of problems, one is binary and the other is multi class.

The work is related with the binary classification only. Some of the classification techniques are discussed as under:

1.5.1 Neural Networks (NN)

An Artificial Neural Network (ANN) is described as a system comprising of an extensive number of interconnected neural connecting elements called artificial neurons. These interconnected processing elements have the potent to learn and acquire knowledge and made it accessible for use. For making the NN to acquire knowledge, various leaning methods are present. According to the learning mechanisms, there are various classes of NN such as Single Layer Feed forward Network, Multilayer Feed forward Neural Network and Recurrent Networks.

- ✓ **Features of NN**
  - i) NN's are robust by nature and capable of tolerating fault. Thus, NN can recall complete patterns from partial or noisy patterns.
  - ii) NN's can process data in a distributed manner at high speed.
  - iii) NN's possess capability to predict the new outcomes based on past trends.
  - iv) NN's can be trained with familiar data and tested upon unknown data.
- ✓ **Advantages of NN**
  - i) Perform better due to its complex structure.
  - ii) Not sensitive to outliers in the data and hence perform better.
- ✓ **Disadvantages of NN**
  - i) Lack of clarity due to its complex network.
  - ii) Does not converge due to large number of variables.

1.5.2 Bayesian Regularized Neural Networks (BRNN)

Bayesian Regularization is a mathematical technique which converts a non-linear regression into a statistical well posed problem. The Bayesian regularized neural network is robust in nature (Burden and Winkler 2008). The Bayesian scheme for neural networks is developed on the probabilistic interpretation of network parameters. Bayesian approach includes a probability distribution of network weights. Bayesian approach resolves the over fitting problem. The complex models are also penalized in Bayesian approach [1]. BRNN is the linear combination of ANN and Bayesian methods to determine the optimal regularization parameters. It involves imposing specified prior distributions on the parameters of model. In Eq. (1.1), E_w term is used to anticipate a better generalization and smooth mapping [2].

$$F = \beta E_D(D|w, M) + \alpha E_w(w|M) \tag{1.1}$$

where E_D is sum of squared estimation errors, M is ANN Architecture and $E_w(w|M)$ denotes sum of squares of architecture weights. α and β are regularization parameters also called objective function parameters. The term αE_w represents the weight decay, where α is the weight decay coefficient. To decrease the tendency of over fitting, w should have smaller values.

Eq. (1.1) involves tradeoff between goodness of fit and model complexity. When $\alpha \gg \beta$ then it produces a smoother network response at the expense of goodness of fit. The algorithm is well explained in Fig. 1.3.

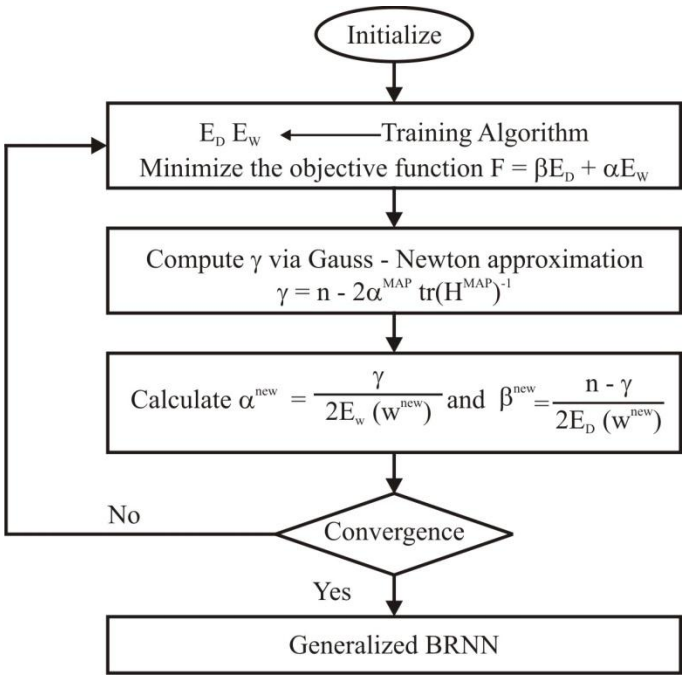


Figure 1.3: Data flow diagram of BRNN Algorithm.

1.5.3 Decision Trees (DT)

Decision Tree is a well-known classification method also known as Classification Tree. A decision tree comprises of a set of internal and leaf nodes. The internal nodes are linked with a splitting criterion, consisting of a splitting attribute and splitting predicates defined on that splitting attribute. The leaf nodes are labeled using a single class label. Decision tree has various advantages which makes its performance outstanding from several other algorithms. Decision tree don't require any input parameter, and its construction technique is relatively fast. Moreover the results are also interpretable [3]. Decision Trees also provide the scalability features for large datasets. Decision tree is capable of classifying both numerical as well as categorical data.

- ✓ **Advantages of Decision Tree**:
  - i) Easier to understand and implement.
  - ii) Not sensitive to usual values of the data, resulting in better performance.
  - iii) Requires less data preparation

1.5.4 Support Vector Machine (SVM)

Traditional classifiers i.e. ANNs are the good classifiers but for training an ANN, a large number of training sets are required. This is not feasible in every real application. SVMs are categorized as a supervised learning model. SVMs work well with small datasets. But SVMs are also capable of managing distributed data in high dimensional datasets. SVMs provide better performance than ANN due to its global solution rather than local minima.

SVM proposed by Vapnik, creates optimal plane that separates the linear variable data by classifying into two classes. The classification is made such that the margin is maximized. The boundary is defined by the nearest data points and is also known as support vectors. The SVM representation is given in Fig.1.4.

Assume that, for a given sample training data, $A = \{x_i, y_i\}_1^N$ in which each input vector $x_i \in R^d$ belongs to class defined by {+1,-1}. The y_i can either take value +1 or -1 depending on which x_i belongs and x_i is a real valued d-dimensional vector.
The classification function of SVM is in the form of Eq. 1.2

$$f(x) = w, \emptyset\,(x) + b \tag{1.2}$$

Where $\{w\}_{i=1}^N$ and b are representing the coefficients and $\emptyset\,(x)_{i=1}^N$ is representing data in feature space. The above said are calculated by minimizing the risk function.

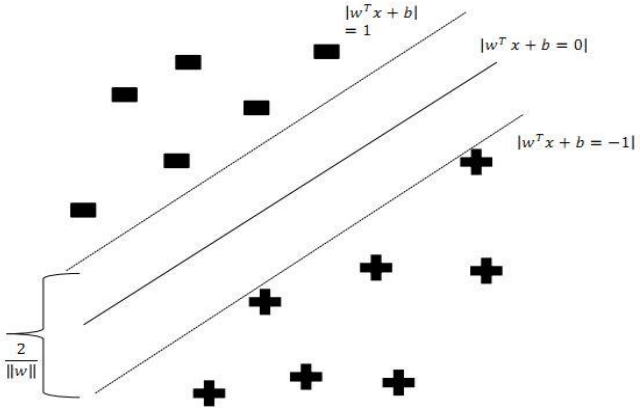


Figure 1.4: SVM Representation

Mathematically, the linear classifier is represented as $y = sgn(w^T x + b)$ (1.3)
The Euclidean distance taken from an instance x_i to the hyper plane $(w^T x + b)$ is represented in Eq. 1.4

$$\frac{|\, w^T x_i + b|}{\|w\|} \tag{1.4}$$

If $|w^T x_i + b| \geq 1$ is confined for all instances, then the minimum distance is $\|w\|$ to the hyperplane. Hence SVM maximizes $\|w\|^{-1}$

The optimization problem can be solved by SVM.

$$(w^*, b^*) = \underset{w,b,\varepsilon_i}{\arg min}\ \frac{\|w\|^2}{2} + C \sum_{i=1}^{m} \varepsilon_i \tag{1.5}$$

such that $$y_i\,(w^T x_i + b) \geq 1 - \varepsilon_i\ (\forall i = 1, \ldots, m) \tag{1.6}$$

$$\varepsilon_i \geq 0 \ (\forall i = 1, \ldots\ldots, m) \tag{1.7}$$

where C is a parameter and denotes the penalty parameter of error term and ε_i is a slack variable.

When data is non-linear, linear classifier can't make the classification good. In that case, conventional method is to map the data points on the higher dimensional feature space, in which non-separable data becomes linearly separable.

The kernel function comes into picture when the data is non-linear in nature. The kernel function is represented as in Eq. 1.8

$$K(x_i, x_j) = \varphi\ (x_i)^T\ \varphi(x_j) \tag{1.8}$$

There are four kind of basic kernels: linear, polynomial, radial basis function and sigmoid. All the basic functions with their basic equation are given in Table 1.3.

Table 1.3: Representation of various kernels of SVM

Kernel	Representation
Linear	$K(x_i, x_j) = {x_i}^T x_j$
Polynomial	$K(x_i, x_j) = (x_i, x_j)^d$
RBF	$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$
Sigmoid	$K(x_i, x_j) = \tan h(k(x_i x_j) + \varphi$

In Table 1, d is a positive integer value and depicts the degree of kernel. Here σ is a real positive value.

1.5.5 Naïve Bayes

Naïve Bayes is based on the Bayes' rule of conditional probability. It makes use of all the attributes and consequently, makes the contribution in decision making, as the attributes are equally important as well as independent of one another.

$$P(H, E) = \frac{P(E_1|H) \cdot P(E_2|H) \ldots\ldots\ldots P(E_n|H)}{P(E)} \tag{1.9}$$

In Eq. (1.9), the probability of the event H is denoted by P (H) and P (H, E) depicts probability of the event H that is conditional on event E. The outcome of the event is denoted by H. E represents combination of all the attribute values. [4]

1.5.6 Linear Discriminant Analysis (LDA)

LDA is a supervised technique used in machine learning for the reduction of dimensionality. It is a generalization of Fisher's linear discriminant. It is used for determining a linear

combination of features which aids in the separation of two or more events. The basic purpose of LDA is the projection of the large feature space onto a smaller subspace. Over fitting is reduced by using the dimensionality reduction, as it minimizes the error in parameter dimension. For a classification task, dimensionality reduction helps in reducing the computational cost.

1.5.7 k-nearest Neighbor (k-NN)

K-nearest neighbor is one of the simplest classification and regression techniques, which is non-parametric by nature. k-NN is also known as instance-based learning. As the object classification is made by the majority vote of its neighbors. The object is assigned to the most common class among its k-nearest neighbors, where k is typically small positive integer. The selection of k is based upon the data. The greater the value of k, less is the classification effected by noise, but it also makes the decision boundaries less distinct.

1.6 Regression

Regression model is one of the oldest prediction models. In Regression models, a specific link structure is created between the inputs and targets. For the prediction of the cases, mathematical equations are used using input variables. There are two forms of regressions: Linear and Logistic regression. In linear regression modeling approach, simple combination of input variables is used to predict the target. Linear regression doesn't handle the missing value cases. Logistic regression is similar to linear regression. Logistic regression uses the link function transformation for target variable.

Linear Regression performs prediction for dependent variable by determining relationship between independent and dependent variables [6]. The graphical relationship between variables can be easily represented by regression model. Mathematical representation of Regression model can be described as in Eq. (1.10)

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i \quad (1.10)$$

where i ranges from $1 \cdots n$. T represents transpose of x which is used to determine the inner product between x_i and β.

1.7 Optimization

Optimization is the procedure of creating something better. Optimization comprises in trying diversity on an initial concept and making the use of that information in order to improve the

idea. Optimization is the technique of regulating the inputs to, or traits of a device. It is a mathematical process, or experiment to come up with the minimum or maximum output.

The input comprises of variables; the function or method is called the cost function, the fitness function or the objective function; and the output is called the cost or fitness. Sometimes the cost needs to be minimized, making the optimization minimized. In some cases, maximizing the function provides the better results.

A practical example called life consists of random events and some decisions made. Quantum theory proposes that an infinite number of dimensions exist, corresponding to the decision made. Life is absolutely nonlinear and clutter plays a significant role. Little disruption in the initial condition might produce different and unpredictable solutions. Optimization is the fundamental tool required in intellectual toolbox and can be represented as in Fig 1.5.

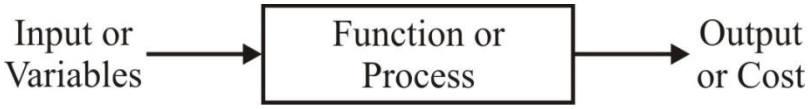


Figure 1.5: Block diagram of optimization.

1.7.1 Categories of Optimization

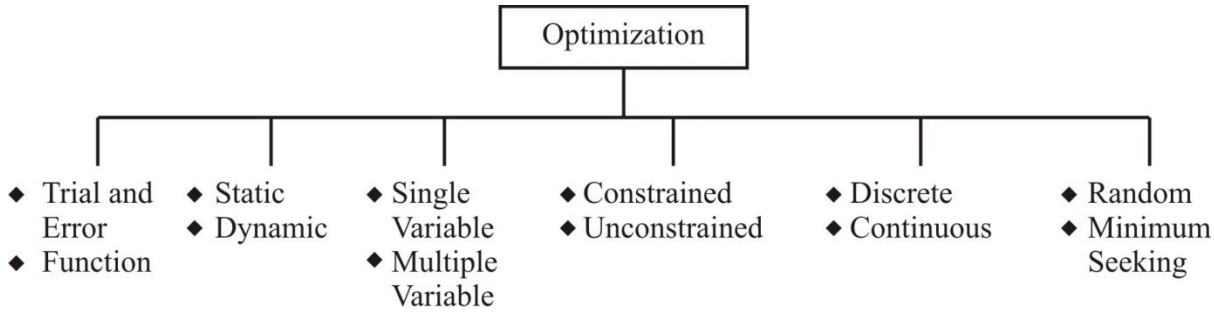


Figure 1.6: Categories of Optimization Algorithm

The categories of optimization can be well seen in Fig. 1.6 and discussed here under.

- ✓ Trial & Error and Function Optimization

  Trial and Error Optimization implies simply, the adjustment of variable which can influence the output without the exact knowledge of the mechanism which results in the output. Experimentalists go for this approach. In comparison to this, the objective function in function optimization is also described by a mathematical formula. Different mathematical usage of the function results in the optimal solution. Theoreticians prefer this theoretical approach.

- ✓ Static and Dynamic Optimization

  Static implies that the output is independent of time whereas dynamic suggests that the output is a function of time. It is difficult to solve the static problem for the best solution, but with the addition of the dimension (time), exaggerate the challenge of cracking the dynamic problem.

- ✓ Single and Multiple Variable

  If only one variable is present, then it is called one-dimensional optimization. If more than one variable is there, it is known as multi-dimensional optimization. With the increase in the dimensions, optimization becomes challenging.

- ✓ Constrained and Unconstrained Optimization

  Unconstrained Optimization accounts the variables to acquire any value. Variables generally have constraints. Constrained Optimization consolidates variable equalities and inequalities upon the cost function. A constrained variable can be transformed to an unconstrained variable with the change in variables.

- ✓ Discrete and Continuous Optimization

  Continuous variables possess infinite number of possible values whereas discrete contain a finite number of possible values. Discrete variable Optimization also called Combinational Optimization.

- ✓ Random and Minimum Seeking

  Traditional optimization algorithms, based on calculus methods, attempt to minimize the cost by beginning from an initial set of variable values. The minimum seekers tend to be fast, but get struck easily in local minima. Whereas, random methods are slower and utilize the probabilistic calculations for determining the variable sets. Random variables determine the global minimum that is better than the local minima.

1.7.2 Differential Evolution Algorithm (DE)

Traditionally, several deterministic gradient based techniques were used for the optimization process. For most of the methods, the objective function required should satisfy the continuity, differentiability and smoothness criteria. The ANN model approximates the non-linear relationship, in which the smoothness criteria may not be satisfied by the objective function. Hence, for the optimization of the input space of the model, gradient based methods are not efficient. Therefore, there becomes a need of exploring alternative optimization techniques which don't have the stringent conditions towards the objective function.

Recently, for the optimization of N-dimensional input space, DE algorithm is made into use for large search spaces. Moreover, DE such as Genetic Algorithm involves survival of the fittest.

The characteristics of DE include:

i) DE requires only scalar values.

ii) DEs can well handle nonlinear objective functions.

iii) DEs don't require conditions such as smoothness, continuity and differentiability as required in the case of traditional techniques.

iv) DEs perform global search.

v) DEs can also handle noisy objective functions.

1.7.3 Genetic Algorithm

Genetic algorithm (GA) is an optimization algorithm and is based upon the process of natural genetics and natural selection. The aim of GA is to minimize or maximize the objective function.

GA has been applied in various fields such as biology, computer science, neural networks, image, pattern recognition and physical sciences. In GA, huge inputs are applied and look for optimal solution.

GA varies from most traditional optimization algorithms in number of ways.

i) GA converts design space into genetic space. GA deals with the coding of variables. Even though the function is continuous, but coding discretizes search space which is the additional benefit of working with the coding of variable space.

ii) Traditional optimization methods use the single point approach whereas GA makes the use of population of points on the single point of time.

✓ Important aspects of GA:

- i) Objective function is defined.
- ii) Genetic representation is defined and implemented.
- iii) Genetic operator is defined and implemented.

✓ Biological Background

Living organisms are made up of cells. In each cell, set of chromosomes are present. A chromosome is made up of genes on the blocks of DNA. Each gene encodes some particular pattern also called traits. Each gene has its position called locus in

chromosome search space. The complete set of the genetic material is known as genome. Genotype is specified set of genes in the genome. The genotype deciding parameters are the organism's phenotype i.e. development after birth and its physical and mental characteristics.

Offspring are created with the recombination, in which a whole new chromosome is generated from genes of the parents. The newly created offspring can be mutated. Mutation refers to the modification in the elements of DNA. While copying genes from parents, the errors are created which leads to the mutation. The success of organism defines the fitness of an organism.

While solving some problem, the goal is to look for the best solution among all. The space of all feasible solutions is termed as search space. Each solution is marked by the value of the fitness for the problem. Solution refers to the minimum or maximum (or extrema) in search space. GA is based on Darwinian Theory i.e. survival of the fittest.

GA algorithm begins with the set of solutions that are represented by chromosomes, also called populations. Solution of one population is taken to form an offspring (new population) with a hope of getting better new population. Solutions are selected based upon their fitness for the generation of offspring. The process is repeated until best solution is obtained.

- ✓ Encoding

  There are numerous ways for representing an individual gene. Usually, Binary encoding is used in GA. Binary encoding provides possible chromosomes for small number of alleles also. For solving minimization or maximization problems, unknown variables are first coded into string structures. Binary encoded strings have 1s and 0s. For converting any integer to binary string, divide the integer by 2.

- ✓ Fitness Function

  Fitness function represented by $F(X)$ is calculated from objective function and is used in successive genetic operations.

  Fitness function is represented by following:

  For maximization problem: $F(X) = f(X)$ (1.11)

  For minimization problem: $F(X) = f(X)$ if $f(X) \neq 0$ (1.12)

$$F(X) = \frac{1}{1+f(X)} \quad \text{if} \quad f(X) = 0 \tag{1.13}$$

The value of the fitness function of the string is called the string's fitness.

GA starts with the population of random strings. To derive the fitness value, each string is being evaluated. The population is operated by following operators:

a) Reproduction

b) Cross-over

c) Mutate

Offspring is further evaluated and then tested for termination. The population is operated iteratively by the above said operators until termination criteria are met. A cycle of operations with the subsequent procedure of evaluation is called as a generation.

a) Reproduction/Selection Operator

The first operator used on population is the reproduction. Chromosomes are selected for cross over, and to become parents from the population and then produce offspring. The necessary idea is that the average strings are taken from current population and subsequently in a probabilistic manner; multiple copies are introduced in mating pool. Several methods of choosing chromosomes for cross-over are

- ✓ Roulette-wheel selection
- ✓ Boltzmann selection
- ✓ Tournament selection
- ✓ Rank Selection
- ✓ Steady-state selection
- ✓ Roulette-wheel selection

It is the selection in which a string is decided from mating pool with the possibility related to the fitness. F_i represents the fitness value for the particular string. The aggregate of all the probabilities for selection of each string for mating pool is always equal to one. The probability of selection of i^{th} string is:

$$p_i = \frac{F_i}{\sum_{j=1}^{n} F_j} \tag{1.14}$$

where n refers to the population size.

The average fitness is described by the Eq 1.15

$$\bar{F} = \sum_{j=1}^{n} \frac{F_j}{n} \quad (1.15)$$

The string possessing the large fitness value will represent wider range in cumulative probability with a higher probability of being replicate in the mating pool. The Roulette-wheel is easier to implement, but noisy in nature.

Roulette- wheel has the following disadvantages specified by Whitley (1989)

(i) The absence of selection pressure results in stagnation of search.

(ii) Premature convergence of search which leads to the quick narrow down of the search.

✓ Boltzmann Selection

Stimulated annealing is used for function minimization or maximization. The method reproduces the technique of cooling of the molten metal slowly which is simulated by parameter like temperature controlling that was introduced by concept of probability distribution given by Boltzmann.

✓ Tournament Selection

For the selection of individuals from population, GA uses the strategy and introduces the selected individuals to the mating pool for the generation of offspring. Selection strategy in GA motivates the collection of better individuals for the mating pool.

Whitley had given the two issues which occur in the evolution technique of genetic search.

i) Population Diversity

Population Diversity refers to the exploitation of the genes from the selected individuals from the population.

ii) Selective Pressure

It is the degree up to which better individuals are preferred.

The population diversity and selective pressure decides the convergence rate of GA.

The tournament selection supports selective pressure by making the tournament competition among number of individuals. The individual with the highest fitness is considered to be the best individual. In mating pool, winners or the tournament competitors are added. The tournament competition is iterated till the mating pool is filled for the generation of new offspring.

- ✓ Rank Selection

  In rank selection, first of all population is being ranked and by taking the chromosomes, and apprehend fitness from above rankings. The best will have the fitness as N and reduces subsequently and worst will have the fitness as 1. This result in slow convergence as best chromosome doesn't have that much difference as others.

- ✓ Steady-state Selection

  In steady state selection, larger part of chromosome should withstand till next generation. With every generation, better individuals are retained while others are eliminated.

b) Cross Over

Reproduction produces clones of good sequences, but not the new ones. Cross-over operator is applied in order to produce a better sequence and for searching the parameter space. Cross over is the recombination operator that occurs in three steps. Primarily the reproduction operator chooses a pair of two individual sequences randomly for mating, then single site cross over is done. There are different types of crossover such as Single-site cross over, two-point cross over and multi-point cross over.

- ✓ Cross Over Rate

  The cross over rate is denoted usually by P_c. i.e. probability of cross-over. It can be calculated by using the ratio of the cross-over operation leads to the search of new strings.

c) Mutation

After performing the cross-over, the strings are exposed to mutation. Mutation of a bit refers to the flipping the bits and transforming 0 to 1 and vice versa. A random number is chosen between 0-1. If random number chosen is smaller than mutation probability, them the outcome is likely to be true, alternatively, it will be false. If the outcome is true at any bit, then that bit is altered, otherwise remains unchanged. The bits of strings are exclusively muted of the probability of mutation.

Mutation leads to introduction of new genetic structures. Mutation also causes alteration in the search space which leads to restoration of lost information to population.

Mutation rate describes the probability of mutation that is used for deriving the number of bits that need to be muted. It also helps in preserving diversity in the population. Typically the population size of 30-200, have mutation rate varied from 0.001 to 0.5.

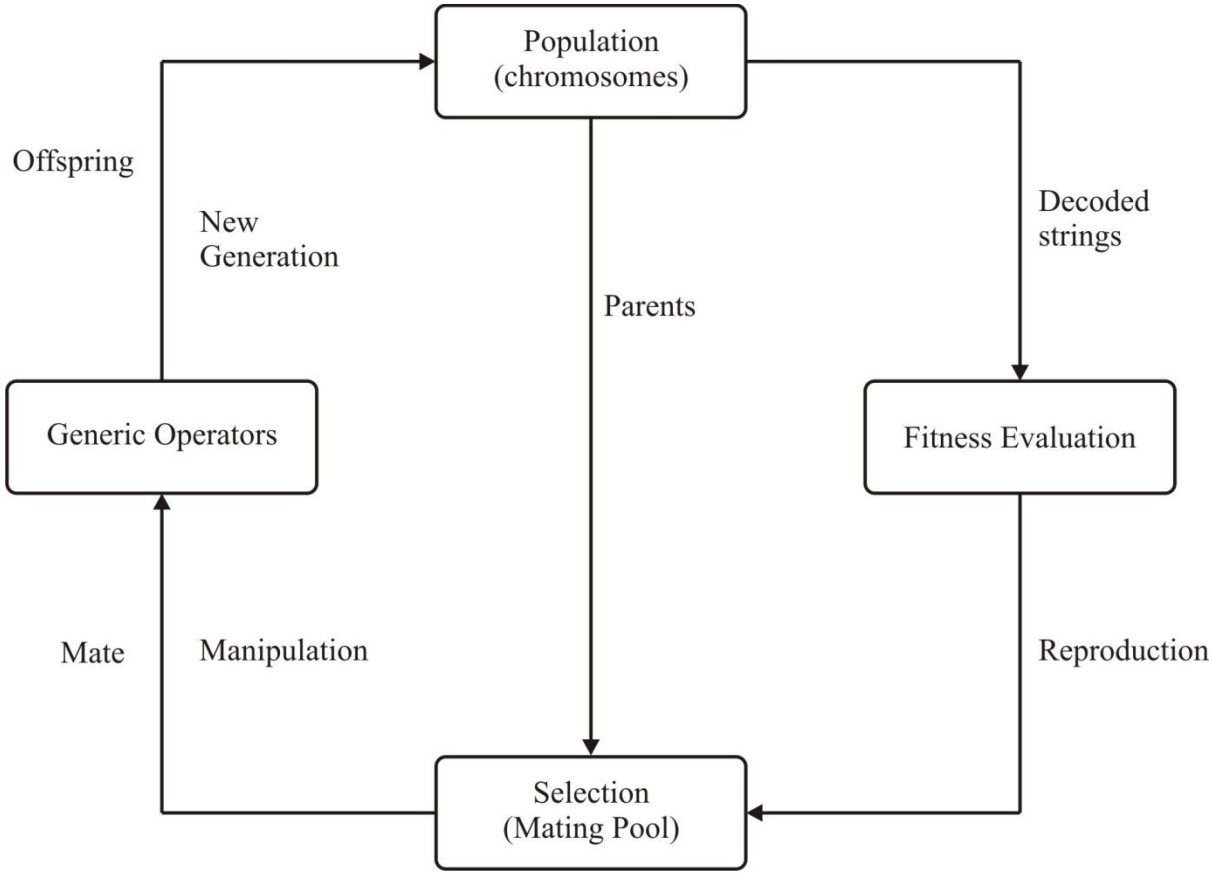


Figure 1.7: GA Cycle

The GA cycle shown in Fig. 1.7 can be explained as below:

Step 1 : Perform the following steps:

(i) Encode problem variable as chromosomes which represent a binary string of fixed length.

(ii) Choose population size, N.

(iii) Define the fitness function to evaluate the probability of chromosome of being selected as a parent chromosome.

Step 2 : Generate population of chromosomes of size N randomly.

Step 3 : Test each chromosome with the fitness function in the population.

Step 4 : Perform following steps until the best fitness value is obtained.

(i) Selection of the pair of chromosomes as a parent chromosome with highest fitness value from the population.

(ii) Apply genetic operators to the selected parent chromosome for producing a pair of offspring chromosomes.

(iii) Allow the offspring and their parents chromosomes to generate the new population.

(iv) Replace current chromosome population with new population.

(v) Calculate fitness value for each chromosome of new population.

Step 5 : Output optimal solutions as the fittest chromosomes for the given problem.

1.7.4 GA as a Soft Computing Tool

GA is based on natural evolution. In this a population of strings is used with initial random parameters. Number of generations is reproduced with operators which represent the extensive elements of evolution likely fitness based selection, competition, recombination and mutation. GA is random by nature. The evolutionary process results in filtering the individuals in population which are closer to fulfilling the objective function for optimization problem. GA contains all the features of soft computing. It promises robust and imprecision tolerant.

✓ Advantages of GA

i) Easy to understand

ii) It Supports multi objective optimization.

iii) It is easily distributed

iv) It is inherently parallel

v) The exploitation of previous or alternate solutions are easy.

vi) It is modular in nature.

vii) It can be used in hybrid applications.

✓ Problems faced while using GA:

i) Selection of population size, mutation rate, mutation operators and cross-over.

ii) Termination criterion.

iii) Scalability and Performance.

1.8 Aim of the Study

My research is concentrated on optimizing and developing a hybrid model of a classifier and verifies its performance analysis on diseases databases such as diabetes, cancer, and liver. In particular research deals with creating an ensemble model for the prediction of disease. The optimization of the SVM is done using GA in order to get the best results.

The specific goals of the research are:

i) Simulate algorithms of various existing classifiers in literature.

ii) Compare and analyze the classification accuracies of various classifiers on different standard datasets.

iii) Compute other performance parameters such as sensitivity, specificity, ROC curve, and simulation time.

iv) Develop a hybrid algorithm for SVM classifier optimized with GA.

v) Develop an efficient hybrid classifier model and compare its performance parameters with existing classifiers.

1.9 Motivation of the Dissertation

The research focuses on predicting disease survivability. The reasons for the motivation behind this research are as:

i) The serious effects of the diseases.

ii) Potential of data mining techniques.

iii) Further understanding of the nature of disease.

iv) Find best machine learning technique using various datasets.

1.10 Organization of the Dissertation

This dissertation is organized in the form of chapters as described below:

Chapter 2: Literature Review, A study has been done about the existing techniques. Detail of various classifiers in different applications is discussed separately.

Chapter 3: Methodology, Steps related to propose work in creating an ensemble are discussed. The details of the approach applied in each step are also provided. The optimization of SVM is also discussed using Genetic Algorithm.

Chapter 4: Results and Discussions, Simulated results of various performance parameters such as sensitivity, specificity, accuracy, and simulation time using various classifiers are discussed in this chapter and also comparison is made with proposed technique. The best fit and mean fit values of the optimized classifier is also discussed.

Chapter 5: Conclusion and Future Scope, The dissertation work is concluded in this chapter and future scope is also given based upon the observation.

CHAPTER - II
LITERATURE REVIEW

2.1 Introduction

This chapter overview the work and methods presented by various researchers in the domain of machine learning. It also includes the classification methods using Support Vector Machine (SVM), Decision Tree (DT), Artificial Neural Network (ANN), Bayesian Regularized Neural Network (BRNN), Naïve Bayes (NB), k-Nearest neighbour (k-NN) and Linear Discriminant Analysis (LDA). The gaps are identified and objectives of dissertation are drawn, along with a brief overview of the proposed methodology.

2.2 Diabetes

Diabetes Mellitus Statistics

According to the World Health Organization (WHO) reports, in every 10 seconds, one person is dying in the world. In every 10 seconds, two new diabetic cases have been found. By 2025, 7 million new diabetic cases will be detected. In context of diabetes, India is considered as the capital of the world. One in every five is Indian diabetic patient. India has 35 million Diabetic patients[7].

Extensive work has been done on Pima Indian diabetes database. The most exhaustive among them is done by Michie et al. (Michie, Spiegelhalter, and Taylor) by considering 22 different algorithms for classification. Accuracy rates were calculated by 12-fold cross validation [8] which is shown in Table 2.1.

Table 2.1: Accuracy of classification methods on Pima Indian Diabetes Data.

Method	Accuracy (%)	Method	Accuracy (%)
AC^2	72.4	Kohonen	72.2
ALLOC80	69.9	Logdisc	77.7
Backprop	75.2	LVQ	72.8
Baytree	72.9	Naivebay	73.8
C4.5	73	NewID	71.1
Cal5	75	Quaddisc	73.8
CART	74.5	RBF	75.7
CASTLE	74.2	SMART	76.8

Method	Accuracy (%)	Method	Accuracy (%)
CN^2	71.1	IndCART	72.9
DIPOL92	77.6	Itrule	75.5
Discrim	77.5	k-NN	67.6

From Table 2.1, it can be analyzed that range of accuracy varies between 67.6% and 77.7%.

Janani *et al.* [9] has evaluated the performance parameters such as Accuracy, Error rate and execution time for PIMA Indian Diabetes dataset. The assessment has been made by using the Decision Table, Naïve Bayes, J48, FT and Multilayer Perceptron. MLP has performed well than other algorithms.

M.S.Barle *et al*. [10] has developed the cascaded model for the classification of PIMA Indian Diabetes Dataset. K-nearest neighbor method is used to extract hidden patterns in dataset. K-means algorithm is integrated with artificial neural network and logistic regression to compute classification accuracy. K-means is also combined with SVM for classification accuracy. K-means with LR has performed best among the methods.

Patil *et al*. [11] proposed a Hybrid prediction model (HPM) by using K means clustering algorithm for pattern extraction. The data instances having zero values are eliminated Decision tree is constructed by using C4.5 classification algorithm. The accuracy achieved in this model is 92.38% for Pima Indian Diabetes dataset.

Polat *et al*. [12] presented an approach based on principal component analysis and adaptive neuro fuzzy interference system for the diagnosis of diabetes disease with accuracy 89.47% .

Humar *et al*. [13] has proposed a hybrid algorithm of artificial neural network and fuzzy neural network and obtained the accuracy value 84.24% by using k-fold cross validation.

Polat *et al*. [14] proposed a cascade learning system using Generalized Discriminant Analysis (GDA) and Least Square Support Vector Machine (LS-SVM) having accuracy of 82.05%.

Carpenter *et al.* [15] developed the ARTMAP-IC algorithm and obtained accuracy as 81%. The training and testing data was randomly chosen as 576 and 192 respectively. The accuracy of logistic regression and KNN were 77% both.

Bioch *et al*. [16] has discarded the patients with glucose and BMIs as zero and the data left was 752. The training and testing data were 500 and 252 respectively. The accuracy of standard neural network was 75.4% and with Bayesian approach, the accuracy was 79.5%.

Smith *et al*. [17] has calculated the accuracy as 76% by using ADAP algorithms. The training and testing data was chosen randomly as 576 and 192 samples respectively.

2.3 Cancer

2.3.1 CpG islands

DNA is a genetic form of gene expression regulation. About 60% of genes in the genome have DNA expansion promoter areas in which cytosine are enriched to guanine substitutions, and these regions are called CpG islands. DNA methylation is a biochemical transition of eukaryotic DNA which resides at the fifth (C5) position of cytosine residue in a 5'-CG-3'known as 5'-methyl-cytosine or CpG Dinucleotide [18]. In vertebrates, the unmethylated CpGs are usually discovered in the CpG Island. Consequently, CpG islands are generally unmethylated. Conventionally, it has been thought that CpG island methylation is responsible for turning off promoters. Recent analysis has demonstrated that CpG methylation is correlated with the stimulating of some genes. In mammals, CpG islands (CGIs) are the core promoter elements. The statistics dispense, in mammals, 60-90% are methylated CpG dinucleotide.

When a gene illustrates higher methylation than that of a reference, the gene is categorized as hypermethylated. Contrarily, a gene illustrating low methylation than that of a reference is called hypomethylated. Cancer cells display large, hypomethylated areas comprising about one-third of TSSs and constant hypermethylation of genetic bodies along with intergenic regions [19]. Hypomethylation results in unconstrained expression and has the potential to stimulate tumor genes called oncogenes. However DNA methylation is a crucial mechanism for tumors to establish drug resistance [20] and activate oncogenes [19].

2.3.2 Malignant Tumor

Cancer has been described as a heterogeneous disease comprising of different subtypes. The early diagnosis and prognosis of tumor/cancer have become the requisite in cancer research, as it can improve survival. Machine learning techniques prove useful by the development of predictive models which detect key features from complex datasets, and leads to accurate decision making[21].

Several studies have been reported based on different techniques which could make the early diagnosis and prognosis of cancer possible.

Hiba Asri *et al*.[3] made a comparison among different machine algorithms such as Decision Tree (C4.5), Naïve Bayes (NB), Support vector machine (SVM) and k nearest neighbors (k-NN) using original Wisconsin Breast Cancer data. The objective is to determine the exactness in classifying data in terms of sensitivity, specificity, accuracy and precision. The results show that SVM provides the highest accuracy that is 97.3% with the lowest error rate. The experiments are conducted in WEKA data mining tool.

Eunhye *et al*. [22] has summarized the current databases such as The cancer Genome Altas (TCGA), Surveillance, Epidemiology and End Results (SEER), Embase and Gene Expression Omnibus (GEO), and used for identification of biomarkers for the breast cancer. It is advantageous for seeking relevant strategies for the diagnosis and treatment of breast cancer.

Raed Ali *et al*. [23] has made the comparison among three Machine learning techniques such as support vector machine (SVM), Bayesian Networks (BN) and Random Forest (RF). For evaluating the performance, Wisconsin Breast Cancer data has been used. The parameters used for comparison are Accuracy, Precision, Recall and Area of ROC. SVM has shown the accuracy, specificity and precision as highest whereas RF has the highest probability of correct classification.

Megha *et al*. [24] has proposed a hybrid approach for diagnosing the breast cancer using ML techniques. Maximum relevance and minimum redundancy (MRMR) is used for feature selection. Parameters such as Accuracy, mean absolute error, Root mean squared error and Kappa statistics such as sensitivity and specificity is used to find the performance of classifiers namely Naive Bayes, SVM, Function Tree and End Meta. SVM with MRMR provides the best accuracy of 99%.

Htet *et al*. [25] has proposed an island based model which overcomes the disadvantages of ANN classifier. The Wisconsin Diagnostic and Prognostic Breast Cancer datasets have been used. In this Differential Evolution (DE) is used to predict the optimal value or near optimal value for parameters of ANN. Island differential evolution neural network has worked well in terms of accuracy, reliability and efficiency. Two different migration topologies are analyzed for the better accuracy and less training time.

Christopher *et al*. [26] discussed that CGI methylation status can be predicted based on the quality of the biological information. It defines a profile based approach which identifies CGIs displaying a differential degree of methylation or lack of methylation across tissues and cell types.

Animesh *et al*. [27] has worked on calculating the smallest subset of features that can provide highly accurate classification of breast cancer. It concluded that Naïve Bayes is the best classifier with maximum accuracy of 97.3 % with only five dominant features and time complexity is 0.1020 millisecond.

Kumar *et.al.* [28] performed a cross validation on the dataset by randomly dividing into training and testing datasets. Naïve Bayes algorithm provided the accuracy of 94.5% and SVM also provided the same accuracy.

Shweta *et al*. [29] developed a probabilistic breast cancer prediction system by using Naïve Bayes classifier with accuracy of 93%.

Gouda *et al.* [30] presented a comparison between various classifiers namely decision tree, Naïve Bayes, Multi-layer perception, sequential minimal optimization and instance based KNN on three datasets of breast cancer such as Wisconsin Breast Cancer, Wisconsin Prognosis Breast Cancer, Wisconsin Diagnosis Breast Cancer by using 10fold cross validation. Naïve Bayes provides the highest accuracy of 92.97%.

Das *et al.* [31] described a computational pattern recognition method which is used to detect the methylation landscape of human brain. They proposed a program called HDFINDER which provides accuracy of 86% for all 22 human autosomes.

Jeetha *et al.* [32] has proposed a method for classification of ovarian dataset using Artificial Neural networks and Genetic Algorithm. The best results are found in neural network with the accuracy of 98%. The proposed model has been find more suitable for kind and unkind ovarian tambours.

2.4 Liver

AlpanaJijja *et al.* [33] has made the comparison between feed forward back propagation (FFBPNN) and cascade correlation feed forward network (CCFFN) on the basis of parameters such as sensitivity, specificity and accuracy. Levenberg- Marquardt algorithm and

Resilient Back propagation algorithm has been used for training the BUPA liver dataset. Cascade correlation network performs better than the feed forward algorithm.

Shrivastava *et al.* [34] has developed multilayer perceptron for classifying liver or non-liver patients based on the liver dataset given. The model has achieved accuracy of 77.77% in the case of 2 hidden layer. A robust model with the learning rate 0.7 has been developed.

Sindhuja *et al*. [35] has discussed the various data mining techniques such as Naïve Bayes, C4.5, Decision Tree, Back propagation Neural network, and Support vector machine on Liver database for computation of parameters such as accuracy, speed, cost and performance. It has been found that C4.5 provides the better result from all the algorithms.

Pakhale *et al*. [36] has summarized the various data mining techniques used for classification on the liver dataset. The survey has been made on classification using Naïve Bayes, Decision tree, artificial neural network, and Support vector machine. The results have been compared using the accuracy and time parameters.

2.5 Optimization

Phan *et al.* [37] has proposed the combination of GA with SVM using feature weighting. The GA is designed using a direction based crossover operator. The comparison has been made with Grid Search. The experiment has been performed on 11 datasets collected from UCI.

Liao *et al.* [38] has presented a Nested real valued genetic algorithm (NRGA) for the optimization of the parameters of SVM and increase the speed of parameter optimization. The experiments have been performed on facial images used for gender classification. It has significantly reduced the computation cost related to the population size of C.

Andy *et al.* [39] has optimized the parameters of SVM using GA. Feature selection has been used for the optimization. The experiment was performed on 6 different datasets using different classification techniques such as k-NN, decision tree and linear discriminant analysis. The best results are found in the case of GA-SVM.

Zhou *et al.* [40] has discussed the optimization of SVM parameters proposing a new genetic operator called improved genetic operator (IO-GA). The paper has also explained the problem encountered like premature convergence using standard genetic algorithm (SGA) and how it limits the accuracy of SVM. The accuracy rate using IO-GA has been increased to 5%, despite of the classification (binary and multi) used.

2.6 Gaps in Study

Several approaches have been proposed in the literature, which have done extensive work on the machine learning and its application in bioinformatics field. But it has been found that very less work has been done in the selection of features of biomedical data from the accuracy point of view. As the medical data is voluminous in nature, so it becomes prime need of considering that features/attributes only that play a significant role in the decision making. All the attributes given are not independent of each other. Some of the attributes are correlated, and hence only increase the computation time as well as complexity. Moreover, this also affects the classification accuracy.

In few of the databases, some values are missing. This becomes an issue in computing the classification accuracy and other parameters. The missing values made the task more challenging to compute the actual accuracy without any assumption. Based on this, the accuracy provided in the literature can't be considered certain.

2.7 Objectives of the Dissertation

On the basis of literature survey of existing methodologies, systems and identified gaps, following are the objectives of this work:

- ✓ To develop an efficient hybrid classifier with the best accuracy and less time complexity
- ✓ To optimize the classification technique (i.e. SVM) using Genetic Algorithm.

2.8 Chapter Summary

In this chapter, various techniques for diseases survivability have been studied. Various classification techniques such as SVM, DT, ANN, BRNN, k-NN, LM and LDA and performance.

Metric such as sensitivity, specificity, ROC, accuracy and time complexity are studied. Literature review for diabetes, cancer and liver diseases is also done. At the end, objectives are drawn based upon the gaps identified in the survey. The existing schemes don't provide that much higher accuracy on various datasets simultaneously. Therefore, it is the need of research to design a time efficient algorithm, which not only offers high accuracy but also outperform other equivalent classifiers based on its training speed and computational complexity. Motivated by this research gap, in this work, an effort has been initiated to

design a Bayesian regularized neural network decision Tree (BrNdT) ensemble mode classifier.

Literatures survey has also been done on the optimization of classification techniques using the evolutionary algorithms. The existing schemes don't consider the relevance of the feature selection in the calculation of performance parameters such as accuracy, sensitivity, specificity and time complexity. The work presented here makes the readers aware the importance of the feature selection before the computation of performance metrics.

CHAPTER - III
ENSEMBLE MODEL

3.1 Introduction

In classifier combination, diverse ensembles have better accuracy than non-diverse ensembles. Classifier combination promises better accuracy than that of an individual classifier. When majority voting combination is used, the combination of independent classifiers will enhance the performance upon the single classifier. Once the ensemble is made together, various combination methods may be used to infer the final class label for an object by using outputs of individual classifier.

Four combination methods have been used: Decision Tree, Support vector machine, Bayesian regularized neural networks and linear regression. Experiments have been carried out on different datasets with different number of classifiers. On each of these, accuracy and time complexity has calculated.

Ensemble methods got the popularity due to its robustness and higher accuracy relative to non-ensemble methods. Inspite of depending on a single classifier, ensemble methods incorporate various classifiers.

The common approach for constituting an ensemble classifier is via a training step. Individual classifiers are trained on different training sets and combined through ensemble algorithm by a voting scheme.

3.2 Materials and methods

3.2.1 Ensemble Classifier

In data mining, each classifier performs prediction based on its supervised learning. Ensemble methods were first invented 10 years back in the machine learning field. The ensemble classifiers weigh several individual classifiers and then fuse them together in order to obtain the results which transcend every individual classifier. The ensemble classifier has better prediction accuracy and classification performance than individual classifiers [41]. The ensemble framework has been shown in Fig. 3.1.

The proposed research converges on the classifiers such as Decision tree, Bayesian regulatory neural network, and linear regression to form ensemble classifier.

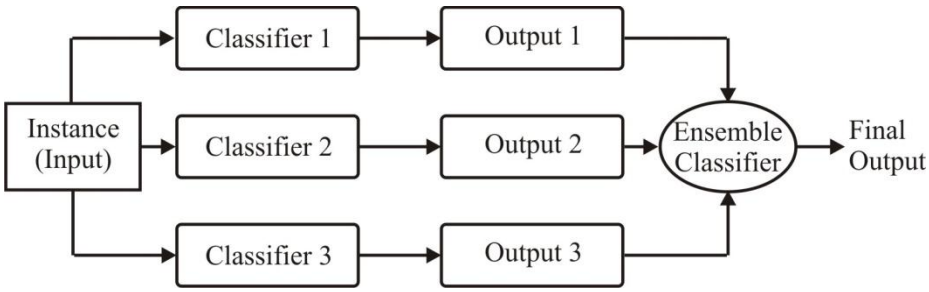


Figure 3.1: Ensemble Framework.

3.2.2 Majority Voting

The technique of classifying the unlabeled instances, based on the high frequency vote (or the highest number of votes) is coined as Majority voting or Plurality Voting (PV). Three classifiers namely decision tree, Bayesian regularized neural network and linear regression combine to form majority voting scheme which is used to predict the class of unlabeled diabetes instances.

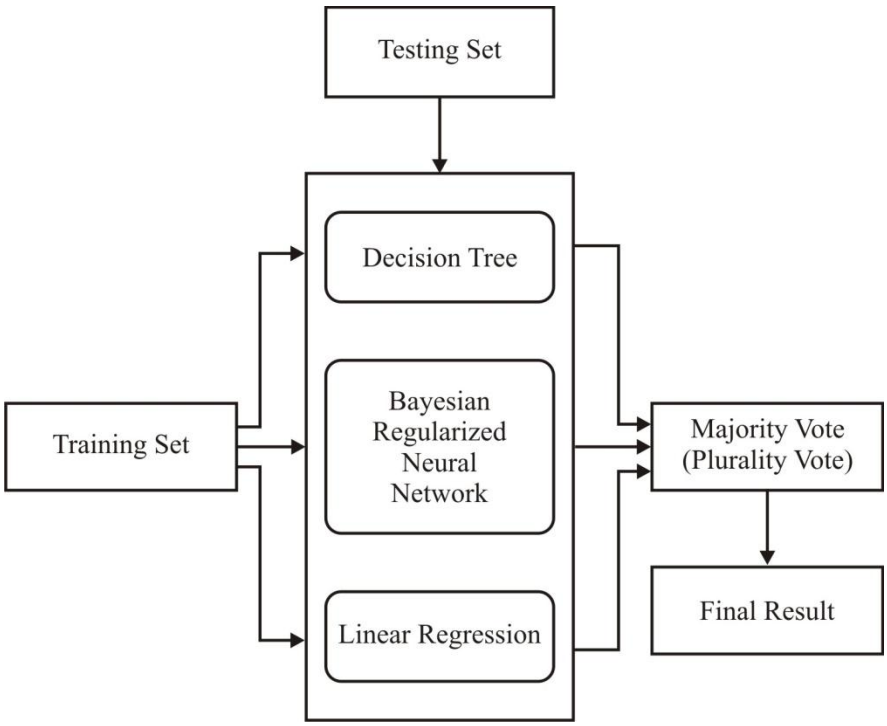


Figure 3.2: Structure of Majority voting Ensemble

Figure 3.2 represents the structure of Majority Voting Ensemble. Mathematically, it can be represented as in Eq. (3.1) [42].

$$\text{class}(x) = \underset{c_i \in \text{dom}(y)}{\arg\max} \left(\sum_k g(y_k(x), c_i)\right) \tag{3.1}$$

Where the classification of the the k'th classifier is denoted by y_k (x) and g(y,c) is an indication function defined as

$$g(y, c) = \begin{cases} 1 & y = c \\ 0 & y \neq c \end{cases} \tag{3.2}$$

A crisp classification can be obtained in the case of probabilistic classifier by following formula

$$y_k(x) = \underset{c_i \in dom(y)}{\arg max} \; \hat{P}_{M_k}(y = c_i|x) \tag{3.3}$$

in which classifier is represented by M_k and the probability of class c is represented by $\hat{\mathbf{P}}_{\mathbf{M_k}}(\mathbf{y} = \mathbf{c_i}|\,\mathbf{x})$ for an instance x. The decision boundaries in the ensemble are shown in Fig. 3.3.

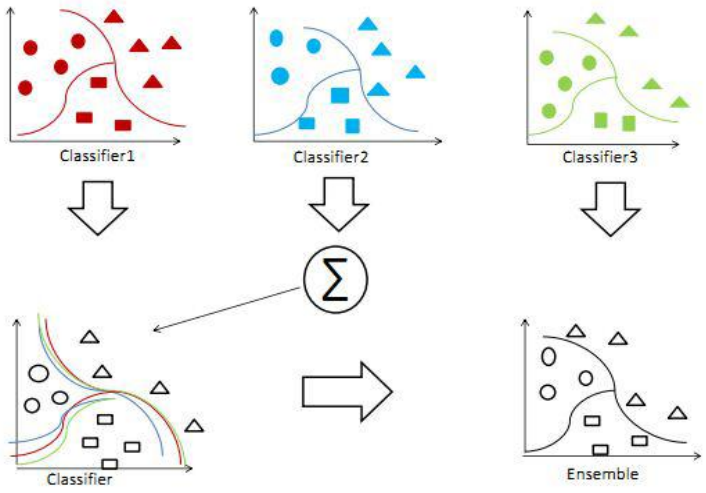


Figure 3.3: Ensemble Model using Majority Voting.

3.2.3 Algorithm for majority voting ensemble and probability of correct labeling

Require:

D : Training dataset

L : Learning algorithm

W : Labels of training dataset

N : Number of Learning algorithm used

d : Output of Classifier

Do n= 1:N

Initialize L_n for every D_n .

Compare W_n with d_n generated from L_n, and update vote.

Evaluate the plurality for output from ensemble.

End

3.2.4 Advantages of Ensemble model

✓ Accuracy: The ensemble model provides the better accuracy than individual classifiers. As it combines the results of multiple weak models, to produce a better result.

- ✓ Scalability and Flexibility: As large databases have become the benchmark in almost every field, it becomes cumbersome to deal with the large datasets. The ensemble methods can scale to large data sets and flexible to binary as well as multivariate classification.
- ✓ Computational Cost: The computational cost of constructing the ensemble makes it applicable in real time applications.

The objective is to raise the accuracy of the classifier and reduce the time complexity. In this, a new algorithm Decision Bayesian Regression Tree (BrNdT) is proposed and this hybrid model has not only the better accuracy but also reduces the time complexity than the existing models with conventional validation only.

3.3 Data Acquisition

The basic purpose is to obtain the data from different repositories and make it into suitable format for analysis. Each dataset contains the feature space that will distinguish between healthy and sick individuals. Every dataset have different attributes and data types. The data is split into training and testing set. The partitioning leads to reduction in computation time.

To determine the performance of the algorithm, experiments have been done on the following datasets. This comprises of 3 real world classification problems gathered from the UCI repository. The characteristics of the datasets are described in the Table 1.

For each dataset, following steps have been performed.

- ✓ Create a stratified arbitrary division between training and testing sets. The range of training and testing sets are also given in Table 1.
- ✓ Exercise the defined algorithm by training the set of instances.
- ✓ Validate the algorithm on testing set of instances.
- ✓ Compute the accuracy and time elapsed for every dataset.

3.4 Selection of Data Base

In this work, standard datasets are selected to verify robustness of the proposed scheme. The datasets namely Pima Indian Diabetes, CpG island, Wisconsin Original Breast Cancer, Wisconsin Diagnostic Breast Cancer, Wisconsin Prognostic Breast Cancer, Ovarian Cancer, ILPD and BUPA represent different real world classification problems available at UCI repository. The detail of datasets used is shown in Table 3.1. The datasets contain the feature space that distinguishes healthy and sick individuals into two classes. Each dataset has different various sets of attributes and data types. For every evaluation, constant proportion of training examples is randomly selected and verified on the testing set.

3.5 UCI Repository

UCI Machine Learning Repository was created by David Aha at UC Irvine in 1987. It contains more than 300 databases as an open access. This site has been cited over 1000 times. This repository is basically used by the researchers related to machine learning and computer science community. Most of the datasets have been collected from the UCI repository itself.

Table 3.1: Dataset employed in the experiment.

Dataset	Train	Test	Number of Attributes	Number of Instances / No. of observations	Positive Instances	Negative Instances	No. of missing instances
Pima Indian Diabetes	460	308	8	768	268	500	NA
CpG Island	297	199	38	495	84	411	NA
Original Breast Cancer	419	280	10	699	241	458	16
Diagnostic Breast Cancer	341	228	30	569	35	212	NA
Prognostic Breast Cancer	119	79	32	198	47	151	4
Ovarian Cancer	130	86	4000	216	95	121	NA
ILPD	350	233	10	583	167	416	NA
BUPA	207	138	6	345	145	200	NA

Note: The classes of the dataset used are binary i.e. 2.

3.5.1 Pima Indian Diabetes

Pima Indian Diabetes data can be obtained from UCI repository of machine learning. (https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes)

The database is created in National Institute of Diabetes and Digestive and Kidney Diseases. Several constraints were adopted on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The various attributes of Pima Indian Diabetes data set and its range is given in table 3.2.

Table 3.2: Description of attributes of Pima Indian diabetes dataset with the range

Attributes	**Min/Max**
Number of times pregnant	0/17
Concentration of plasma glucose in an oral glucose tolerance test in 2hours	0/199
Diastolic blood Pressure (mm Hg)	0/122
Thickness of triceps skin folds (mm	0/99
Serum insulin in 2hour (mu U/ml)	0/846
BMI (Body Mass Index) (weight in kg/ (height in m^2)	0/67.1

Attributes	Min/Max
Diabetes Pedigree Function	0.078/2.42
Age (years)	21/81

3.5.2 CpG Island

Cpg Island dataset methylation defines the status of human chromosomes. The dataset is referred from BioMed Central Bioinformatics [43]. This dataset contains both normalized as well as non-normalized value. But the normalized values have been considered for this work.

The dataset has 38 attributes which belong to four distinct categories namely

i) CGI specific attributes (including G+C content, CpG cluster p-value, Observed / Expected ratio),

ii) Repetitive sequences (eg. Number and type of repetitive elements)

iii) Evolutionary Conservation (e.g. PhaseCon content)

iv) Physiochemical and Structural Properties of DNA (includes twist, tilt, roll, rise, shift, slide).

The attributes of dataset are listed below: The minimum and maximum values of the attributes are shown in Table 3.3

Table 3.3: Description of Attributes of CpG Island Dataset

Sr. No	Attributes	Sr. No	Attributes
1	CpG cluster value	20	AG content
2	Observed/Expected value	21	CT content
3	Average distance between CpGs	22	CC content
4	Standard Deviation	23	GG content
5	G+C Content	24	GA content
6	Repetitive Element Content	25	TC content
7	Repetitive elements	26	Bending
8	PhastCon Content	27	Curvature
9	PhastCon Elements	28	Stacking energy
10	CG content	29	Turn
11	GC content	30	Twist
12	AA content	31	Cleavage
13	TT content	32	Bases per turn
14	TA content	33	Twist Constraint
15	AT content	34	Tilt constraint
16	CA content	35	Roll constraint
17	TG content	36	Shift constraint
18	AC content	37	Slide constraint
19	GT content	38	Rise constraint

Note: As the above table contains only normalized values, so every attribute has the range of 0/1.

3.5.3 Wisconsin Original Breast Cancer

The dataset was collected from University of Wisconsin. Wisconsin Original Breast Cancer contains 699 clinical cases with 9 attributes having range from 1 to 10. This contains 458 benign and 241 malignant cases and listed in Table 3.4.The last attribute of the dataset have 16 instances missing. The elimination of instances will affect the accuracy. The evaluation has been made without replacing the missing values.

Table 3.4: Description of Wisconsin Original Breast Cancer Dataset with the range.

Attributes	Domain	Attributes	Domain
Sample Code number	Id number	Single Epithelial Cell Size	1-10
Clump Thickness	1-10	Bare Nuclei	1-10
Uniformity of Cell Size	1-10	Bland Chromatin	1-10
Uniformity of Cell Shape	1-10	Normal Nuclei	1-10
Marginal Adhesion	1-10	Mitoses	1-10

3.5.4 Wisconsin Diagnostic Breast Cancer

Wisconsin Diagnostic Breast Cancer Data is made available at following link by William H. Wolberg University of Wisconsin Hospital.

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

It has two classes with 569 instances i.e. 357 benign and 212 malignant. It has 10 variables consisting of cellular characteristics computed from digitized image of breast fine needle aspirate. The standard error, mean and worst of these attributes were computed for every image, which results in 30 features. The real valued attributes collected from link https://bigml.com/user/czuriaga/gallery/dataset/51794b29ce5680176800031e and listed in Table 3.5.

Table 3.5: Description of attributes for Wisconsin Diagnostic Breast Cancer dataset.

Attributes	Std. error Min/Max	Mean Min/Max	Worst Case Min/Max
Radius (mean of distances from center to points on the perimeter)	0.11/2.87	6.98/28.11	7.93/36.04
Texture (standard deviation of gray-scale values)	0.36/4.89	9.71/39.28	12.02/49.54
Perimeter	0.76/21.98	43.79/188.50	50.41/251.20
Area	6.80/542.20	143.50/2501	185.20/4254
Smoothness (local variation in radius lengths)	0.00/0.03	0.05/0.16	0.07/0.22
Compactness ($perimeter^2$ / area - 1.0)	0.00/0.14	0.02/0/35	0.03/1.06
Concavity (severity of concave portions of the contour)	0.00/0.40	0.00/0.43	0.00/1.25

Attributes	Std. error Min/Max	Mean Min/Max	Worst Case Min/Max
Concave points (number of concave portions of the contour)	0.00/0.05	0.00/0.20	0.00/0.29
Symmetry	0.01/0.08	0.11/0.30	0.16/0.66
Fractal dimension ("coastline approximation"-1)	0.00/0/03	0.05/0.10	0.06/0.21

3.5.5 Wisconsin Prognostic Breast Cancer Dataset

The Wisconsin Prognostic Breast Cancer Dataset was given by Dr. Wolberg. The cellular characteristics computed from digitized image of breast fine needle aspirate. The standard error, mean and worst of these attributes were computed for every image, which results in 30 features. The other two features are Tumor Size i.e. Diameter of excised tumor (in cm) and Lymph Node status. Recurrence Surface Approximation (RSA) is a programming model which is linear by nature and used to predict "Time to Recur" using recurrent and non-recurrent cases both. The outcome can be recurring or non-recurring. The description of the attributes is given as under in Table 3.6

Table 3.6: Description of Attributes for Wisconsin Prognostic Breast Cancer Dataset.

Attributes	Std. error Min/Max	Mean Min/Max	Worst Case Min/Max
Radius (mean of distances from center to points on the perimeter)	0.19/1.82	10.95/27.22	12.84/35.13
Texture (standard deviation of gray-scale values)	0.36/3.503	10.38/39.28	16.67/49.54
Perimeter	1.153/13.28	71.9/1821	85.1/232.2
Area	13.99/316	361.6/2250	508.1/3903
Smoothness (local variation in radius lengths)	0.002/0.03	0.075/0.144	0.08/0.22
Compactness ($perimeter^2$ / area - 1.0)	0.007/0.14	0.05/0/311	0.05/1.06
Concavity (severity of concave portions of the contour)	0.01/0.14	0.024/0.43	0.02/1.17
Concave points (number of concave portions of the contour)	0.005/0.03	0.02/0.20	0.03/0.29
Symmetry	0.01/0.06	0.130/0.30	0.16/0.66
Fractal dimension ("coastline approximation" - 1)	0.00/0.012	0.05/0.10	0.06/0.21
Tumor Size	0/27		
Lymph Node status	0/1		

3.5.6 Ovarian Cancer

The Ovarian data has been based on Serum proteomic pattern diagnostics and taken from http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp

The profile patterns are created using protein mass spectrometry. Ovarian Cancer contains a record of 216 patients with 121 as ovarian cancer patients and 95 as normal patients. It contains 4000 features.

3.5.7 BUPA

BUPA dataset is also collected from UCI repository. It contains 345 instances with 6 attributes. The selector is the target attribute having binary class 1 or 2. The attributes and its range is given in Table 3.7

Table 3.7 Description of BUPA attributes with the range

Attribute	Min/Max
mcv (Mean corpuscular volume)	65/103
alkphos (Alkaline Phosphotase)	23/138
Sgpt (Alamine Aminotransferase)	4/155
Sgot (Aspartate Aminotransferase)	5/82
Gammagt (Gamma-glutamyl transpeptidase)	5/297
Drinks (Number of alcoholic beverages drunk per day)	0/20

3.5.8 ILPD (Indian Liver Patient Dataset)

Indian Liver Patient Dataset was collected from UCI repository which has further recorded the data from the state Andhra Pradesh of India. The data set has 441 male patients and 142 female patients in records. The record contains 416 patients suffering from liver disease and 167 normal individuals. Selector is the class label given to distinguish liver patients or normal person. The following are the attributes of ILPD with their respective range given in Table 3.8

Table 3.8: Description of ILPD attributes with the range.

Attribute	Min/Max
Age of person	4/90
Gender	M/F
TB (Total Bilirubin)	0.4/75
DB (Direct Bilirubin)	0.1/19.7
Alkphos (Alkaline Phosphotase)	63/2110
Sgpt (Alamine Aminotransferase)	10/2000
Sgot (Aspartate Aminotransferase)	10/4929
TP (Total Proteins)	2.7/9.6
ALB (Albumin)	0.9/5.5
A/G (Ratio Albumin and Globulin Ratio)	0.3/2.8

3.6 Normalization Procedure

Normalization speeds up the training process, as it scales the data in the same range for each input feature. It is beneficial for modeling applications with different scales of input. For the efficient results, data entries are made normalized in the interval of [0, 1]. Different techniques employ different rules such as min, max, sum and product rule. The normalization performed in this work is given in Eq.(3.4).

$$Normalized\ value = \frac{value - minimum}{maximum - minimum} \tag{3.4}$$

3.7 Feature Selection

Data pre-processing includes data cleaning, integration, transformation and feature selection (reduction). There are few attributes in the dataset that are redundant as the have same information as contained in the other attributes. The more the attributes, the more time it takes for the computation process and also affect the accuracy as well. The goal of feature selection is to find out the least number of attributes so that the resultant distribution of data classes is approximate to the original distribution. The removal of irrelevant data is made possible through data reduction process. This not only decreases the simulation time but also increases the classification accuracy.

Feature selection has been an important area of research for decades in the field of data mining and machine learning. A typical process of feature selection includes four basic steps viz.

- ✓ Generation of subset
- ✓ Estimation of subset
- ✓ Stopping condition
- ✓ Confirmation of result.

3.8 Summary

In this chapter, ensemble model has been discussed using majority voting. The algorithm of the ensemble model has also been presented. The standard datasets collected from various sites mainly from UCI repository are described. UCI is a database containing several biomedical datasets from the various sources under one umbrella. The datasets such as PIMA Indian Diabetes, CpG Island, Original Wisconsin Breast Cancer, Diagnostic Wisconsin Breast Cancer, Prognostic Wisconsin Breast Cancer, Ovarian Cancer, BUPA and ILPD with their respective attributes are mentioned. The normalization of few datasets is also performed considering their computational constraints.

4.1 Simulation Software

The proposed model is tested on four standard datasets to establish its robustness. The training phase simulation of BrNdT model is completed with 60 % patterns selected for training and 40% patterns selected for testing. The existing algorithms such as BRNN, SVM, DT and Linear Regression are also simulated with same standard datasets and their performance metrics are recorded for comparison using R as simulation software.

'R' is an integrated tool developed at Bell Laboratories. 'R' version 3.4.0 has been used as a computing platform. 'R' is mainly used for the statistical data analysis and also provides packages which can be easily installed as per the requirement.

4.2 Performance Parameters

The proposed model framework has been depicted in Fig. 4.1. The performance analysis of each model is measured in terms of accuracy and time complexity.

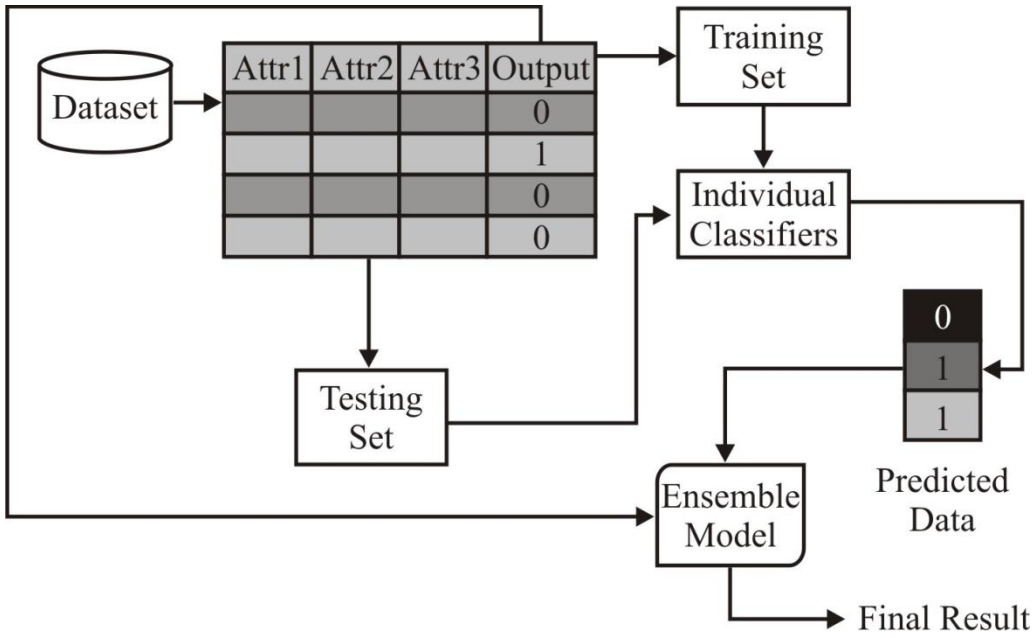


Figure 4.1: Proposed Ensemble framework for classifier.

Classification accuracy is one of the important performance parameters of any classification scheme, since it aids in predicting the correct class and can lead to accurate diagnosis of the patients. The speed (Simulation time) of the classifier also plays an important role. A

classifier with smaller accuracy might be preferred over the higher one if shows significant smaller time complexity.

Other performance parameters such as sensitivity and specificity of the classifier are evaluated from confusion matrix. Confusion matrix consists of two categories of class namely actual and predicted class. The actual class is determined through angiographic method whereas predicted class is simulated through algorithms.

4.2.1 Confusion Matrix

In confusion matrix, four parameters of actual class called True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are evaluated and listed in Table 4.1.

Table 4.1: Confusion Matrix.

Predicted Class	**Actual Class**	
	C1	**C2**
C1	TP	FP
C2	FN	TN

The parameter TP represents the count of instances correctly classified as C1 and TN is the count of instances correctly classified as C2. Similarly, FN is the count of instances falsely classified as C2 and FP represents the count of instances falsely classified as C1.

Based on the parameters used in Confusion Matrix, the performance metric such as sensitivity, specificity, and accuracy are evaluated as given in Table 10 and Table 11.

Accuracy is the ratio of total number of all correctly classified to the overall number of cases of model and given in Eq. (4.1).

$$\textbf{Accuracy} = \frac{TN+TP}{TN+TP+FN+FP} = \textbf{1- Error} \tag{4.1}$$

Sensitivity is the ability to detect disease in the diseased individual population. It is the ratio of true positives correctly classified and given in Eq.(4.2).

$$\textbf{Sensitivity} = \frac{TP}{TP+FN} \tag{4.2}$$

Specificity is the ability of correctly ruling out the disease in population of disease free individuals. It is the ratio of true negatives correctly identified and depicted in Eq. (4.3).

$$\textbf{Specificity} = \frac{TN}{TN+FP} \tag{4.3}$$

The time elapsed is described in Eq.(4.4) for computing the simulation time.

$$\textbf{Time elapsed} = Processing\ Time - Start\ Time \tag{4.4}$$

Table 4.2: Confusion Matrix for Pima Indian Diabetes.

Predicted Class	Actual Class	
	C1	C2
C1	54	21
C2	194	39

(a) BRNN

Predicted Class	Actual Class	
	C1	C2
C1	61	38
C2	177	32

(b) DT

Predicted Class	Actual Class	
	C1	C2
C1	54	22
C2	193	39

(c) LM

Predicted Class	Actual Class	
	C1	C2
C1	58	21
C2	194	35

(d) SVM

Predicted Class	Actual Class	
	C1	C2
C1	65	22
C2	193	28

(e) BrNdT

Table 4.3: Confusion Matrix for CpG Island.

Predicted Class	Actual Class	
	C1	C2
C1	21	9
C2	120	49

(a) BRNN

Predicted Class	Actual Class	
	C1	C2
C1	14	4
C2	125	56

(b) DT

Predicted Class	Actual Class	
	C1	C2
C1	5	1
C2	125	65

(c) LM

Predicted Class	Actual Class	
	C1	C2
C1	0	0
C2	129	70

(d) SVM

Predicted Class	Actual Class	
	C1	C2
C1	14	1
C2	128	56

(e) BrNdT

Table 4.4: Confusion Matrix for Original Breast Cancer.

Predicted Class	Actual Class	
	C1	C2
C1	64	5
C2	210	1

(a) BRNN

Predicted Class	Actual Class	
	C1	C2
C1	57	2
C2	213	8

(b) DT

Predicted Class	Actual Class	
	C1	C2
C1	61	1
C2	214	4

(c) LM

Predicted Class	Actual Class	
	C1	C2
C1	64	6
C2	209	1

(d) SVM

Predicted Class	Actual Class	
	C1	C2
C1	62	1
C2	214	3

(e) BrNdT

Table 4.5: Confusion Matrix for Diagnostic Breast Cancer.

Predicted Class	Actual Class	
	C1	C2
C1	55	8
C2	165	0

(a) BRNN

Predicted Class	Actual Class	
	C1	C2
C1	53	16
C2	157	2

(b) DT

Predicted Class	Actual Class	
	C1	C2
C1	53	5
C2	168	2

(c) LM

Predicted Class	Actual Class	
	C1	C2
C1	54	5
C2	168	1

(d) SVM

Predicted Class	Actual Class	
	C1	C2
C1	54	4
C2	169	1

(e) BrNdT

Table 4.6: Confusion Matrix for Prognostic Breast Cancer.

Predicted Class	Actual Class	
	C1	C2
C1	2	2
C2	59	17

(a) BRNN

Predicted Class	Actual Class	
	C1	C2
C1	4	11
C2	50	15

(b) DT

Predicted Class	Actual Class	
	C1	C2
C1	3	0
C2	61	16

(c) LM

Predicted Class	Actual Class	
	C1	C2
C1	0	0
C2	61	19

(d) SVM

Predicted Class	Actual Class	
	C1	C2
C1	2	0
C2	61	17

(e) BrNdT

Table 4.7: Confusion Matrix for BUPA.

Predicted Class	Actual Class	
	C1	C2
C1	87	52
C2	0	0

(a) BRNN

Predicted Class	Actual Class	
	C1	C2
C1	61	29
C2	23	26

(b) DT

Predicted Class	Actual Class	
	C1	C2
C1	90	49
C2	0	0

(c) LM

Predicted Class	Actual Class	
	C1	C2
C1	95	44
C2	0	0

(d) SVM

Predicted Class	Actual Class	
	C1	C2
C1	2	0
C2	61	17

(e) BrNdT

Table 4.8: Confusion Matrix for ILPD.

Predicted Class	Actual Class	
	C1	C2
C1	2	0
C2	164	68

(a) BRNN

Predicted Class	Actual Class	
	C1	C2
C1	26	27
C2	137	44

(b) DT

Predicted Class	Actual Class	
	C1	C2
C1	0	0
C2	164	70

(c) LM

Predicted Class	Actual Class	
	C1	C2
C1	0	0
C2	164	70

(d) SVM

Predicted Class	Actual Class	
	C1	C2
C1	3	0
C2	164	67

(e) BrNdT

Table 4.9: Sensitivity (%) of various classifiers with different dataset.

	BRNN	DT	LM	SVM	BrNdT
PIMA Indian Diabetes	58.1	65.0	58.1	62.4	**69.9**
CpG Island	30.0	20.0	**71.0**	0.0	20.0
Original Breast Cancer	**98.5**	87.7	93.8	**98.5**	95.4
Diagnostic Breast Cancer	**100.0**	96.4	96.4	98.2	98.2
Prognostic Breast Cancer	10.5	**21.1**	15.8	0.0	10.5
BUPA	100.0	70.1	100.0	100.0	**100.0**
ILPD	37.1	0.0	29.0	0.0	**43.0**

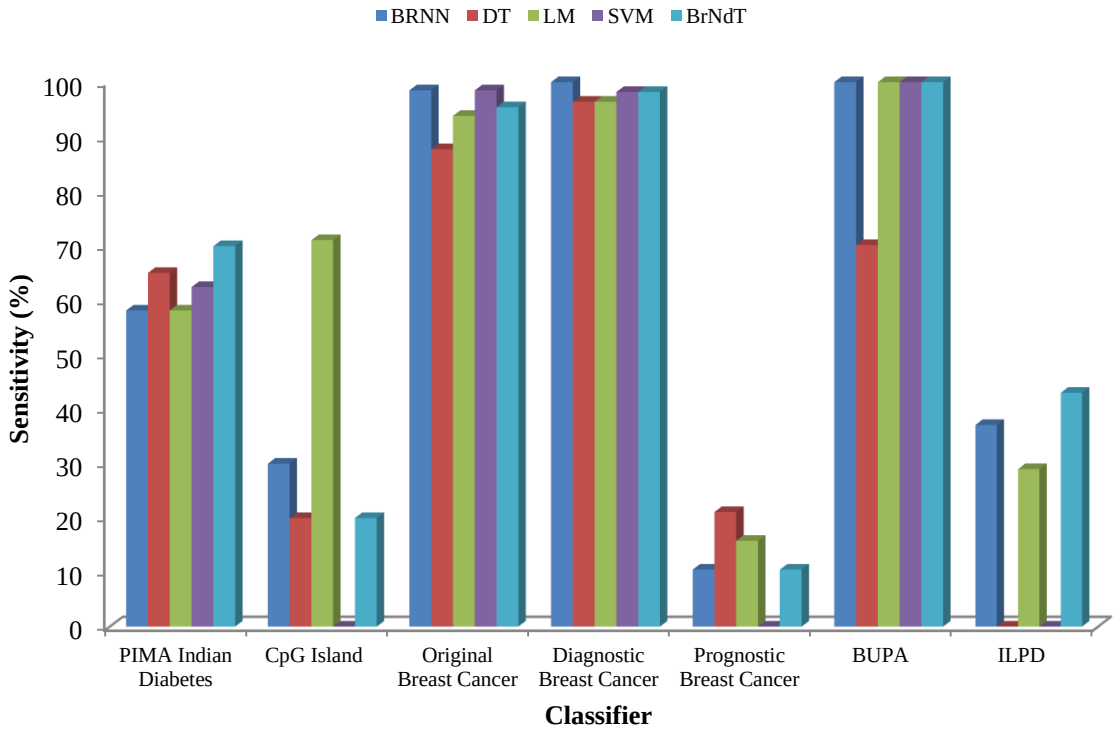


Figure 4.2: Sensitivity of various classifiers with different dataset.

Table 4.10: Specificity (%) of various classifiers with different dataset.

	BRNN	DT	LM	SVM	BrNdT
PIMA Indian Diabetes	**90.2**	82.3	89.8	90.2	89.8
CpG Island	93.0	96.9	99.2	**100.0**	99.2
Original Breast Cancer	97.7	99.1	99.5	97.2	**99.5**
Diagnostic Breast Cancer	95.4	90.8	97.7	97.1	**97.7**
Prognostic Breast Cancer	96.7	82.0	100.0	100.0	**100.0**
BUPA	0.0	**44.2**	0.0	0.0	0.0
ILPD	83.5	100.0	100.0	100.0	**100.0**

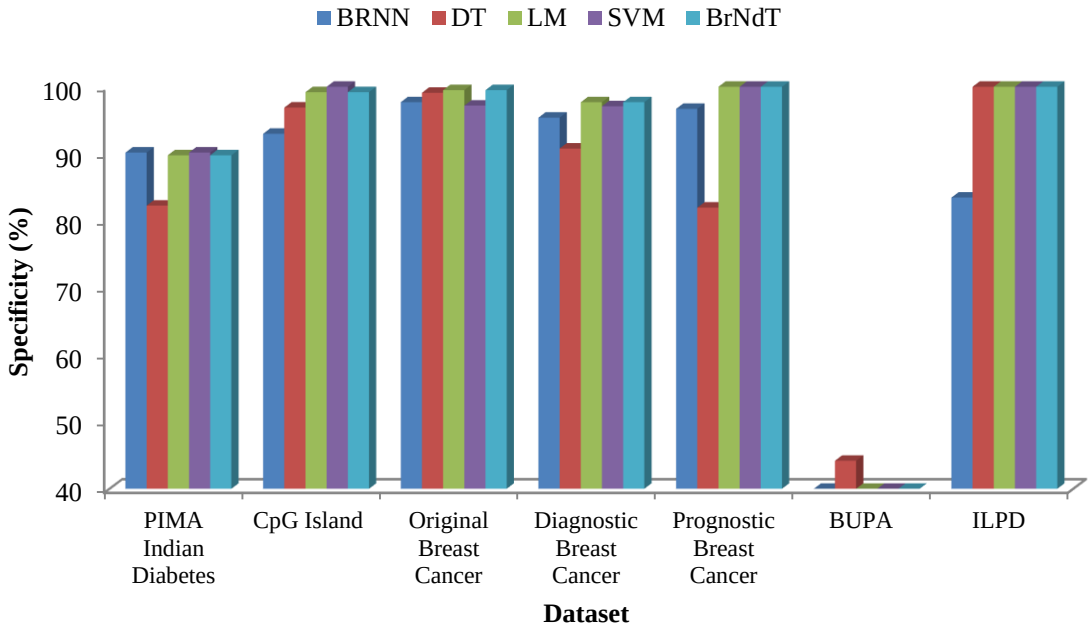


Figure 4.3: Specificity of various classifiers with different dataset.

The confusion matrix listed in Table 4.2 - Table 4.8 has resulted in calculation of metrics such as specificity and sensitivity shown in Fig. 4.2 and Fig. 4.3. The classifiers like BRNN, DT, LM, SVM and the proposed model have behaved varied on different standard datasets taken into consideration. The proposed hybrid model doesn't ensure the best performance in the case of sensitivity and specificity. But also, it has not deteriorated the performance in the context of specificity and sensitivity. The proposed hybrid model, BrNdT has able to maintain the specificity as well as sensitivity comparable to the standard classifiers.

Table 4.11 Accuracies (%) of various classifiers with different dataset

	BRNN	DT	LM	SVM	BrNdT
PIMA Indian Diabetes	80.52	77.27	80.19	81.82	**83.77**
CpG Island	70.85	69.85	66.33	64.82	**71.36**
Original Breast Cancer	97.86	96.43	98.21	97.50	**98.57**
Diagnostic Breast Cancer	96.49	92.11	96.93	97.37	**97.81**
Prognostic Breast Cancer	76.25	67.50	77.50	76.25	**78.75**
BUPA	66.19	60.43	64.03	68.35	**69.78**
ILPD	70.94	69.66	70.09	70.09	**71.37**

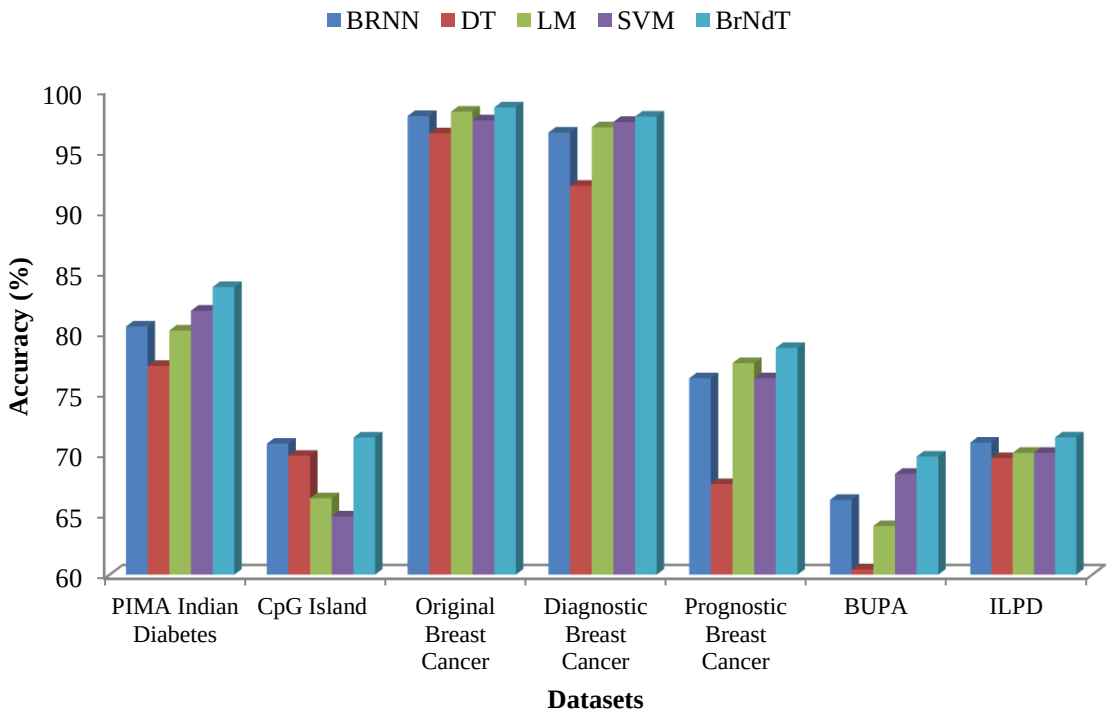


Figure 4.4 Accuracies of various classifiers with different dataset.

Table 4.12: Simulation Time of algorithm(sec) of various classifiers with different dataset.

	BRNN	DT	LM	SVM	BrNdT
PIMA Indian Diabetes	9.45	4.92	3.62	3.76	**3.05**
CpG Island	4.35	4.60	4.33	6.45	**3.60**
Original Breast Cancer	7.08	5.66	4.44	4.37	**2.00**
Diagnostic Breast Cancer	6.42	4.31	5.34	4.86	**3.58**
Prognostic Breast Cancer	4.98	6.38	4.76	4.56	**4.00**
BUPA	7.81	6.49	6.52	5.54	**4.43**
ILPD	7.16	7.83	5.12	4.52	**4.11**

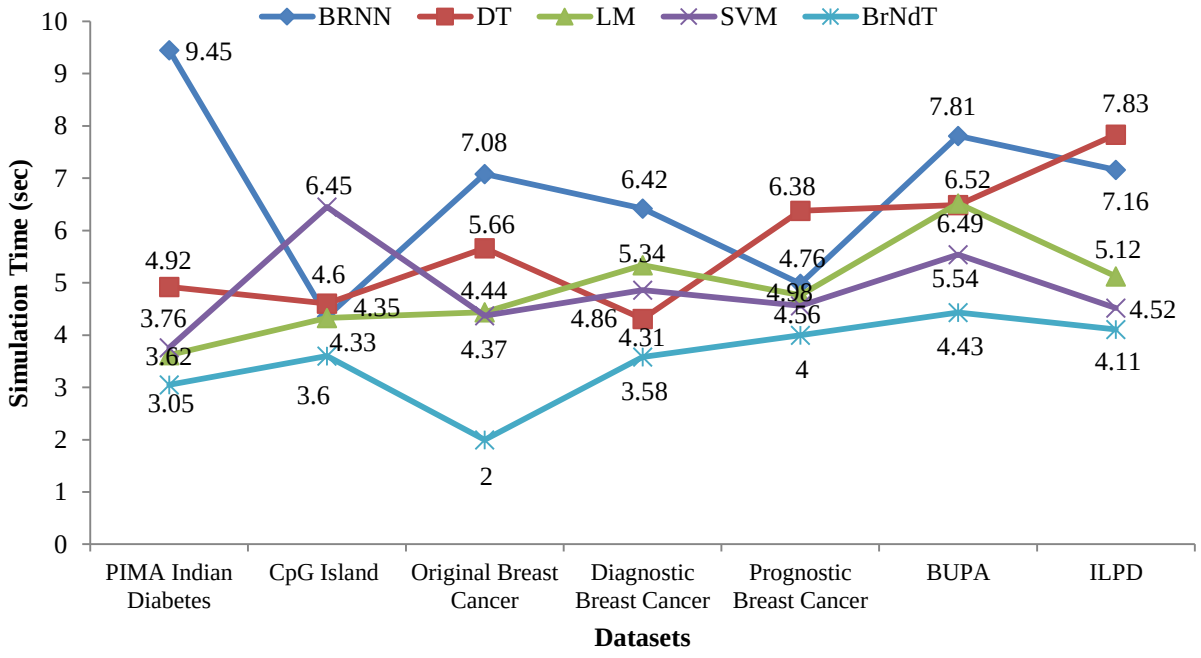


Figure 4.5: Simulation time of various classifiers with different dataset.

The statistical measures such as accuracy and simulation time are computed and listed in Table 4.11 and 4.12. From the analysis done on the above said datasets, it has been found that SVM has better accuracy than Decision Tree, Linear Regression and Bayesian regularized Neural Networks. Now Ensemble classifier is designed by using Decision Tree, Linear Regression and Bayesian Regularized Neural Networks and the performance measures of Ensemble Classifier are compared with that of the SVM. It has been found that Ensemble Classifier performs best than SVM and can be seen in Fig.4.4 and Fig. 4.5.

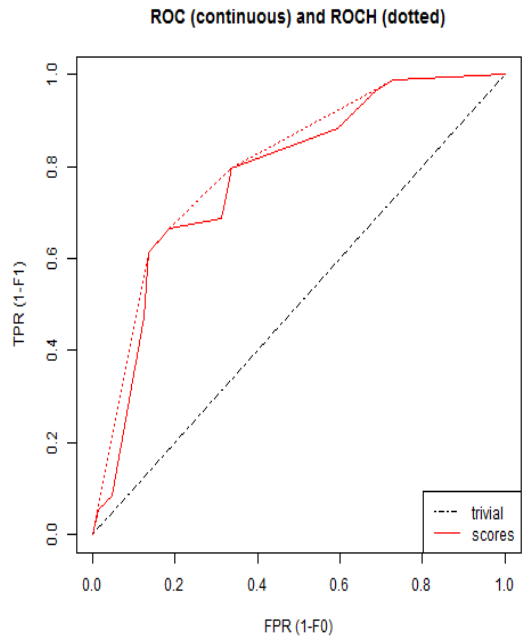
BrNdT has increased the accuracy as well as reduces time complexity to a great extent. So overall it can be analyzed that Ensemble classifiers perform better than individual classifiers in the binary classification.

4.2.2 ROC Curve

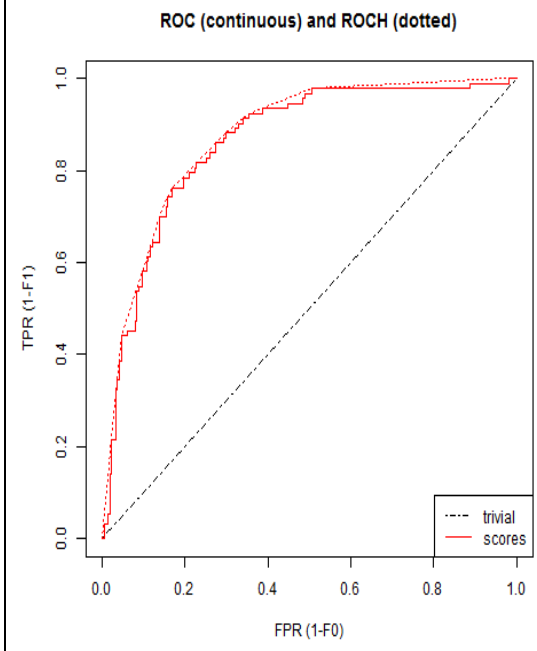
Consider a set of instances, marked as p or n, and ranked by using some scoring function. Let a threshold value σ, the instances scoring below σ is predicted as 'n' and instances scoring higher than σ, are taken as 'p'. For a given threshold value, Total Positive is number of positive instances which are classified correctly and False positive is the number of negative instances, that are misclassified [20].

ROC curve is a graph between True Positive Rate (TPR) versus False Positive Rate (FPR). Each pair of FPR and TPR is corresponding to a point on ROC space. The ROC curve on each dataset is given in Fig 4.6 - Fig. 4.12.

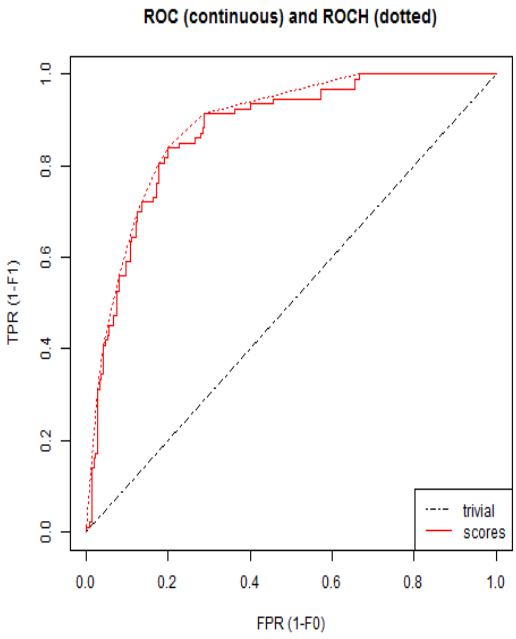
A trivial ranking is used to map the each real value of training set with class label 'p' as 1, and class label 'n' as 0.



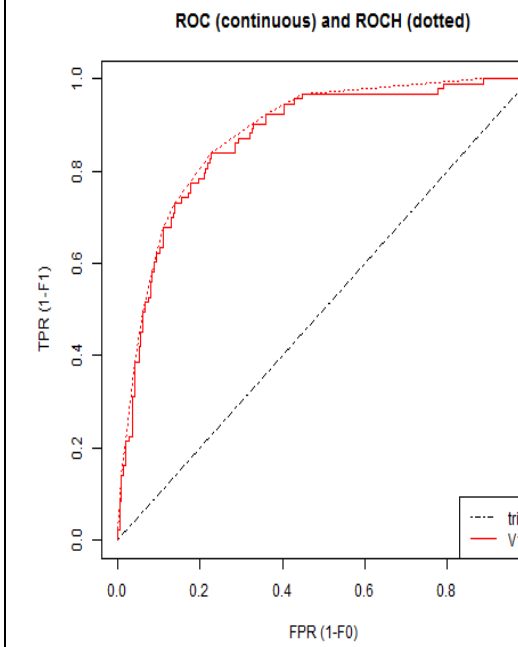
ROC curve of Decision Tree for Pima Indian Diabetes



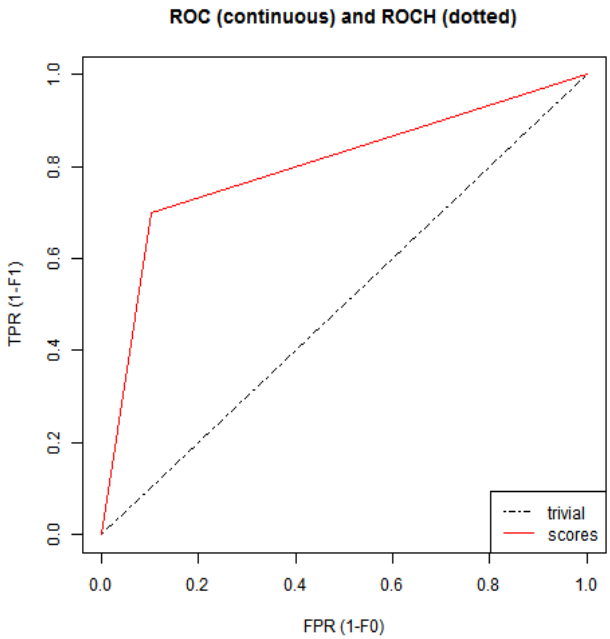
ROC curve of linear regression for Pima Indian Diabetes



ROC curve of BRNN for Pima Indian Diabetes

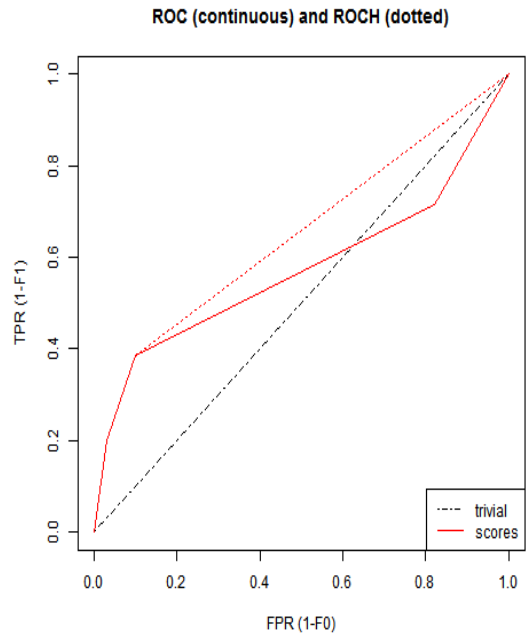


ROC curve of SVM for Pima Indian Diabetes

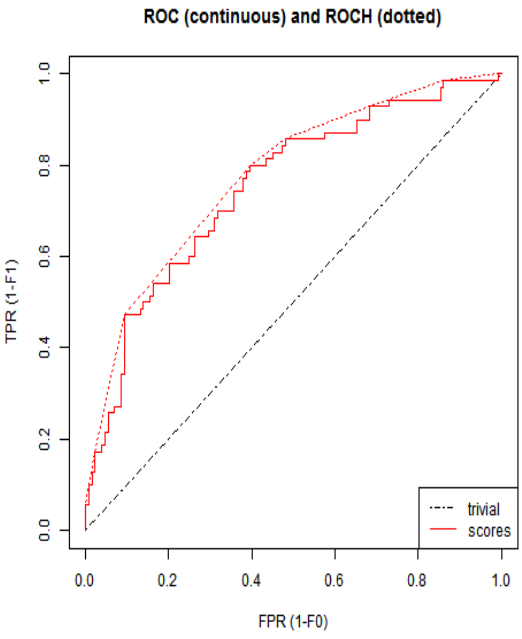


ROC curve of BrNdT for Pima Indian Diabetes

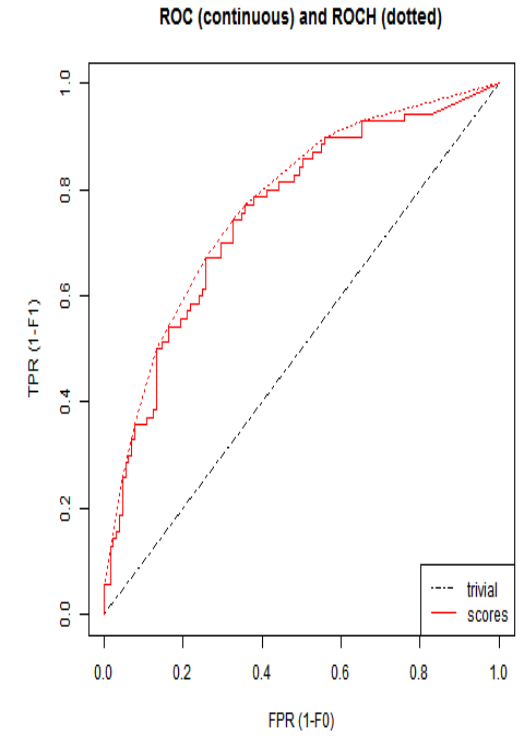
Figure 4.6: ROC curves for PIMA Indian Diabetes dataset.



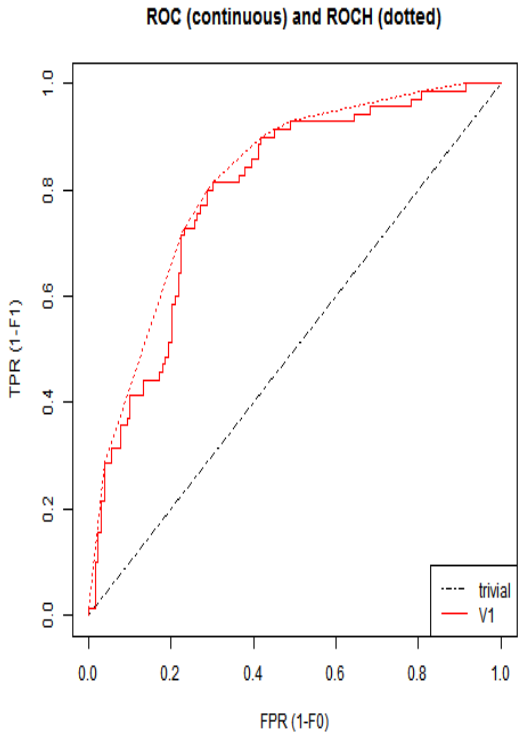
ROC curve of Decision Tree for CpG Island



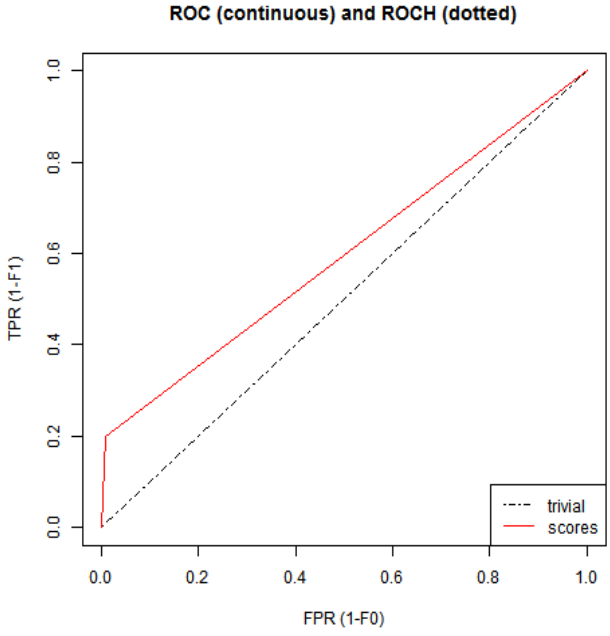
ROC curve of linear regression for CpG Island



ROC curve of BRNN for CpG Island

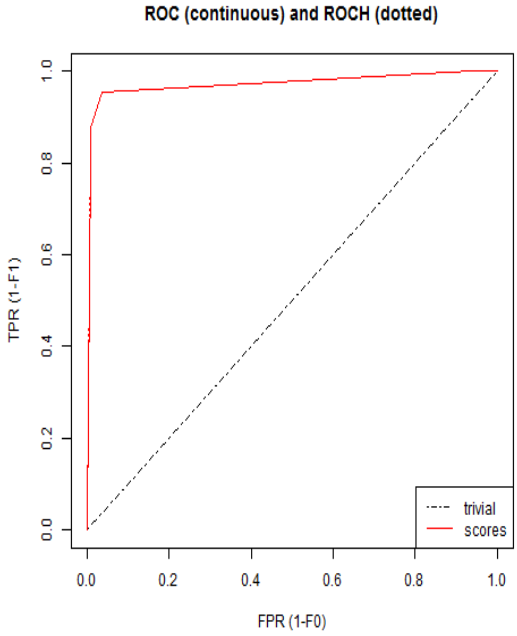


ROC curve of SVM for CpG Island

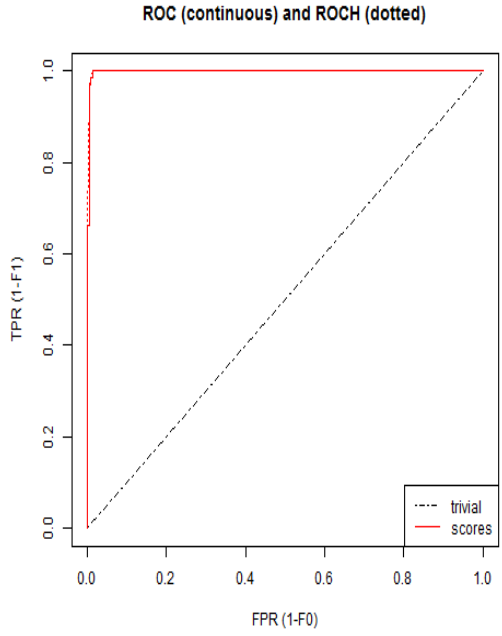


ROC curve of BrNdT for CpG Island.

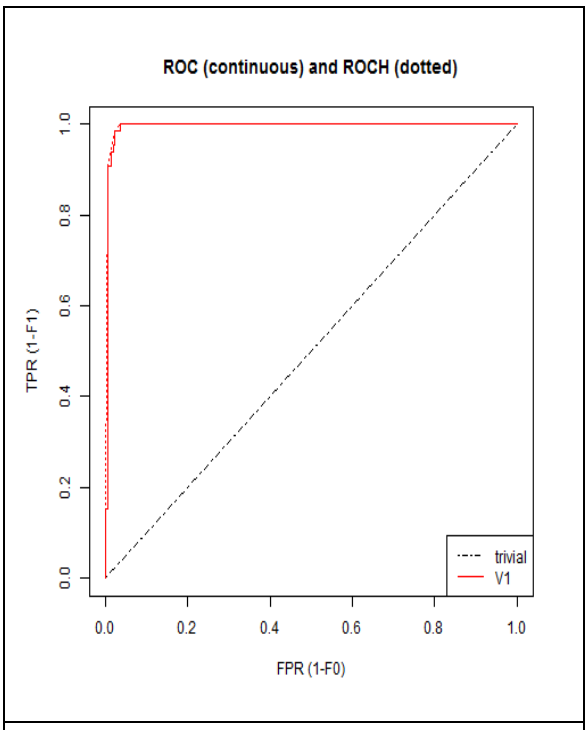
Figure 4.7: ROC curves for CpG Island Dataset.



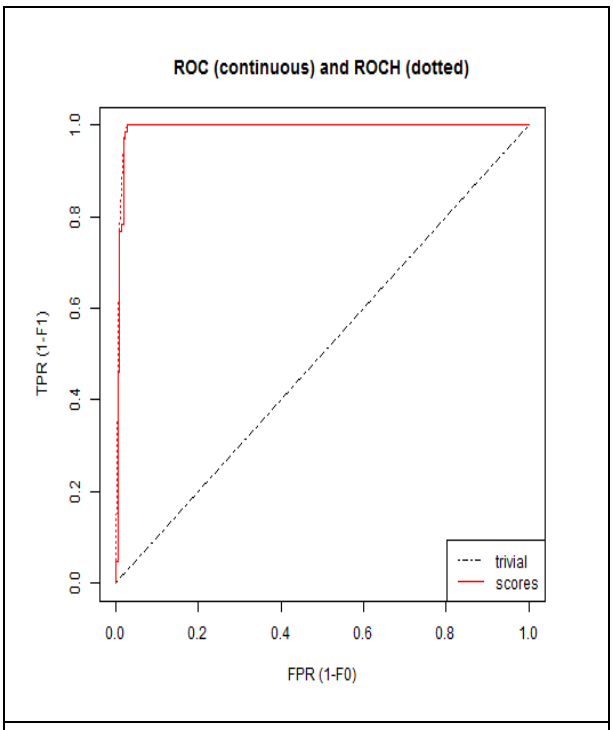
ROC curve of Decision Tree for Original Breast Cancer



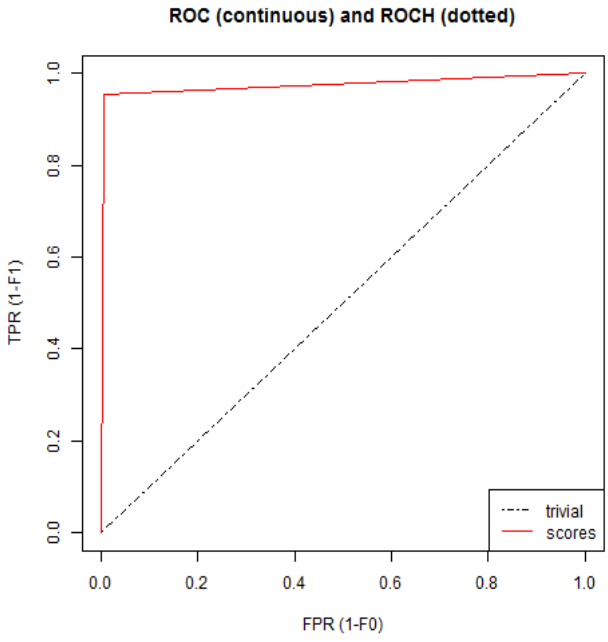
ROC curve of linear regression for Original Breast Cancer



ROC curve of BRNN for Original Breast Cancer

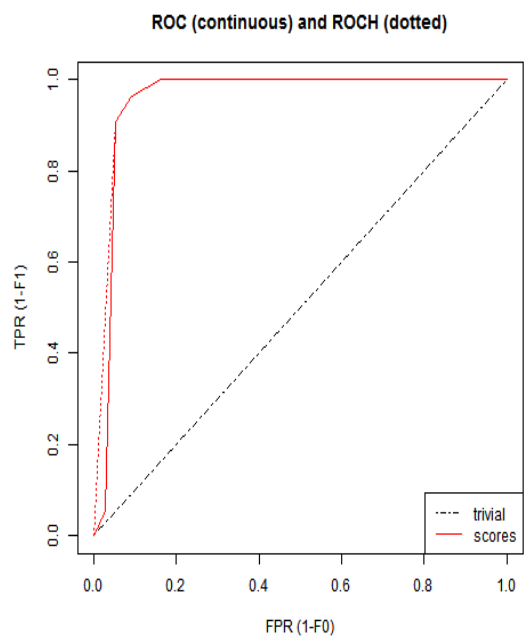


ROC curve of SVM for Original Breast Cancer

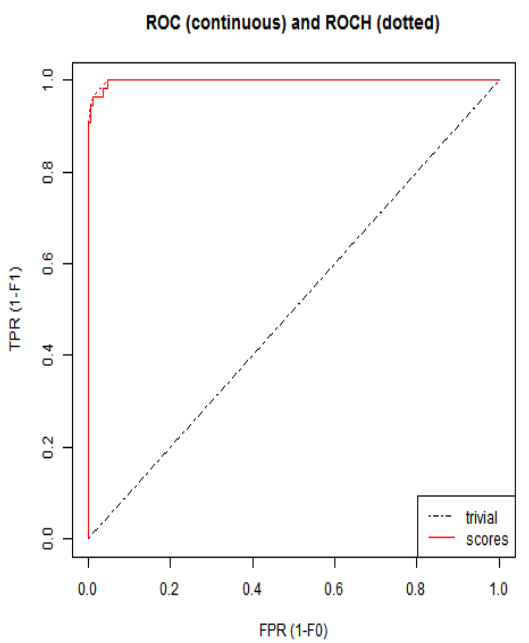


ROC curve of BrNdT for Original Breast Cancer

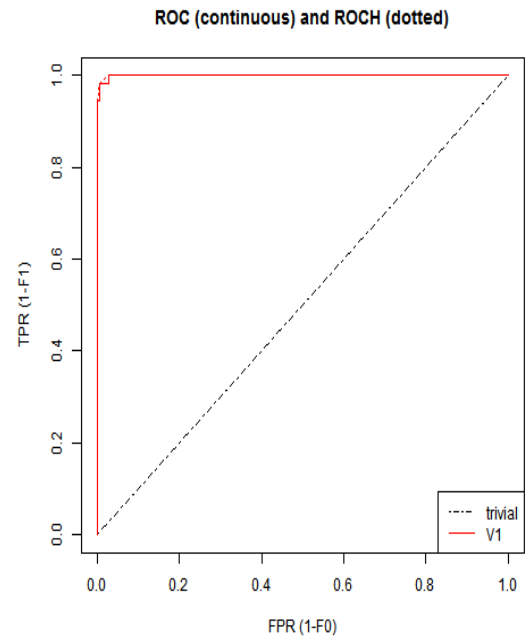
Figure 4.8: ROC curves for Original Breast Cancer.



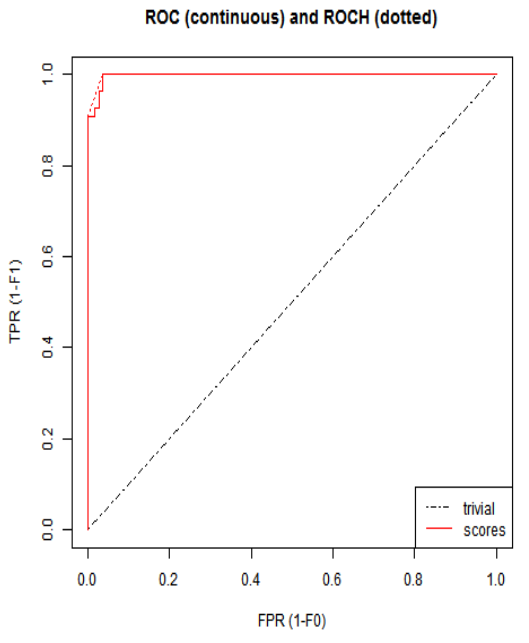
ROC curve of Decision Tree for Diagnostic Breast Cancer



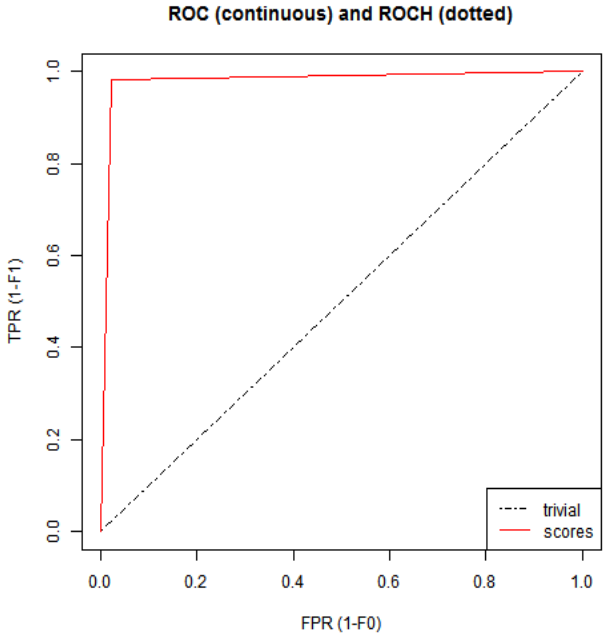
ROC curve of linear regression for Diagnostic Breast Cancer



ROC curve of BRNN for Diagnostic Breast Cancer

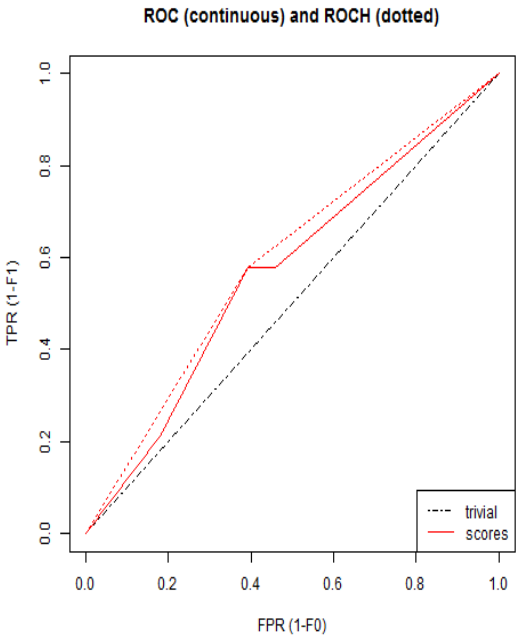


ROC curve of SVM for Diagnostic Breast Cancer

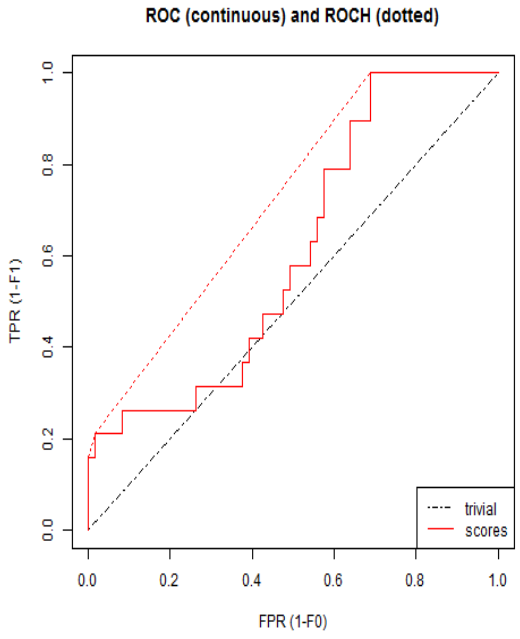


ROC curve of BrNdT for Diagnostic Breast Cancer

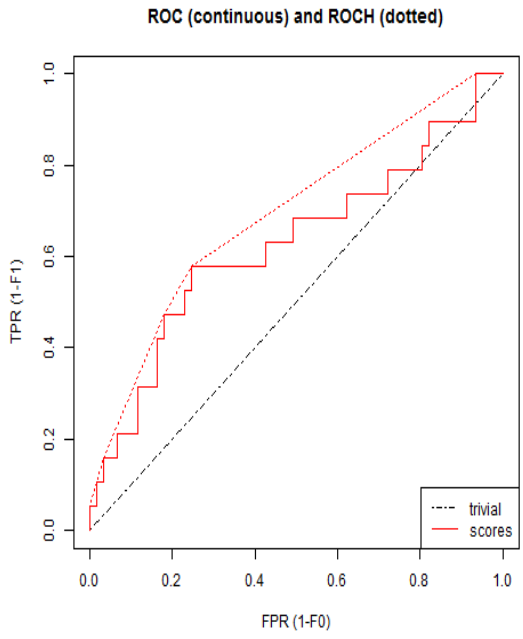
Figure 4.9: ROC curves for Diagnostic Breast Cancer.



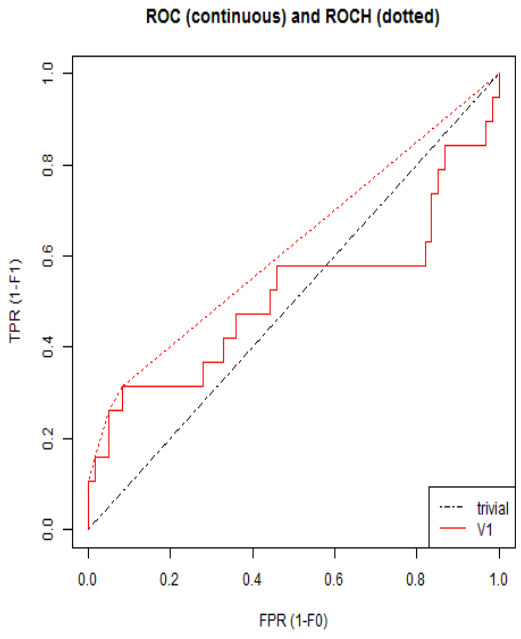
ROC curve of Decision Tree for Prognostic Breast Cancer



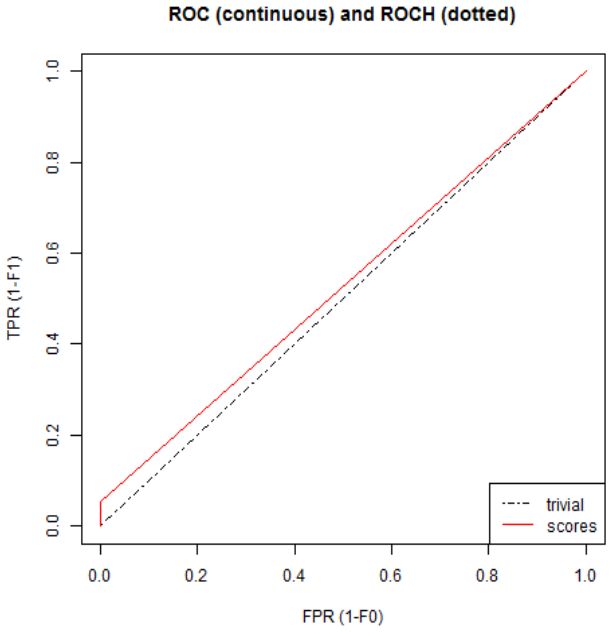
ROC curve of linear regression for Prognostic Breast Cancer



ROC curve of BRNN for Prognostic Breast Cancer

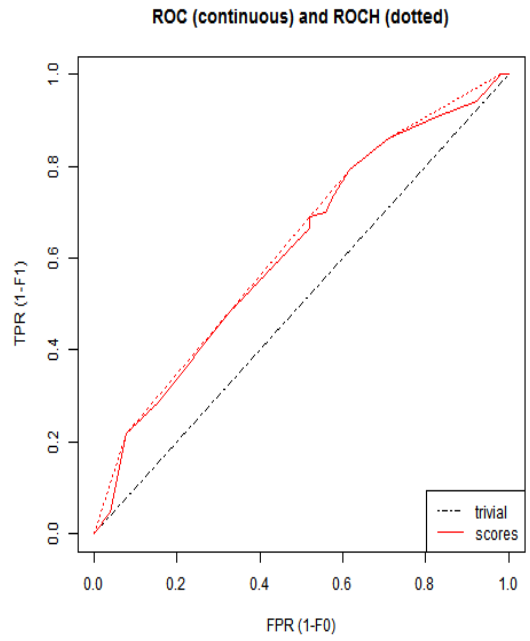


ROC curve of SVM for Prognostic Breast Cancer

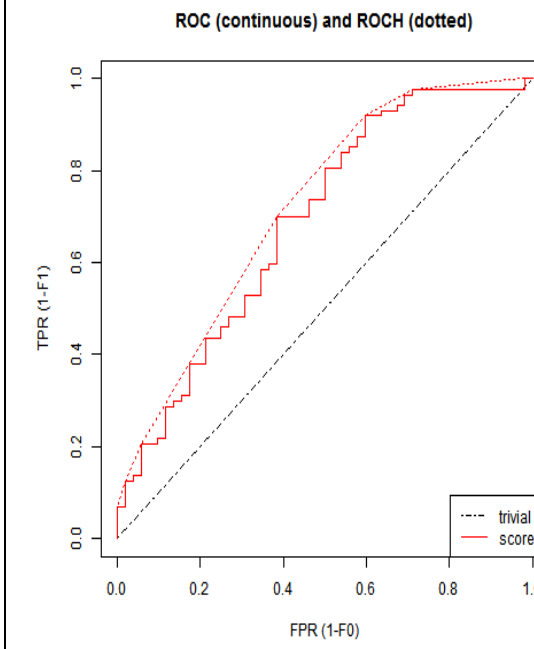


ROC curve of BrNdT for Prognostic Breast Cancer

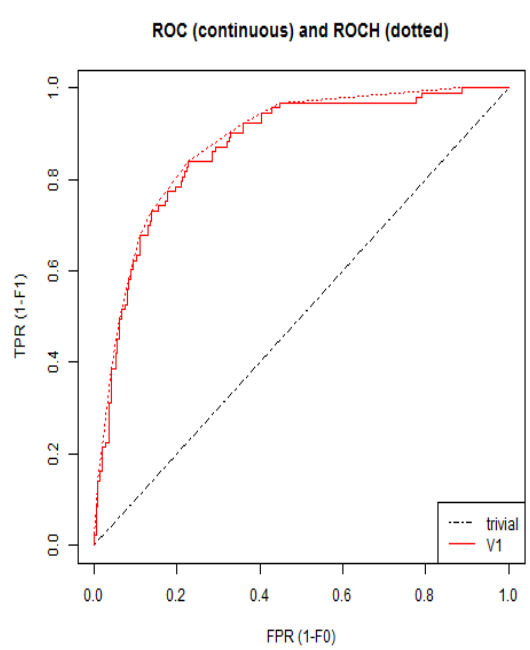
Figure 4.10: ROC curves for Prognostic Breast Cancer.



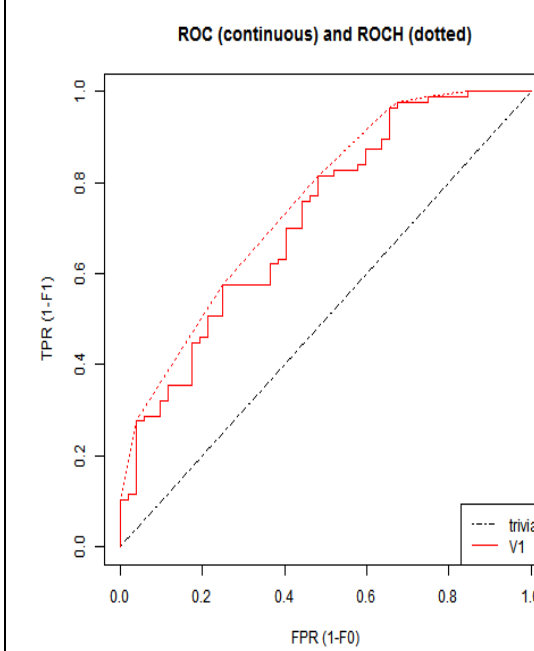
ROC curve of Decision Tree for BUPA



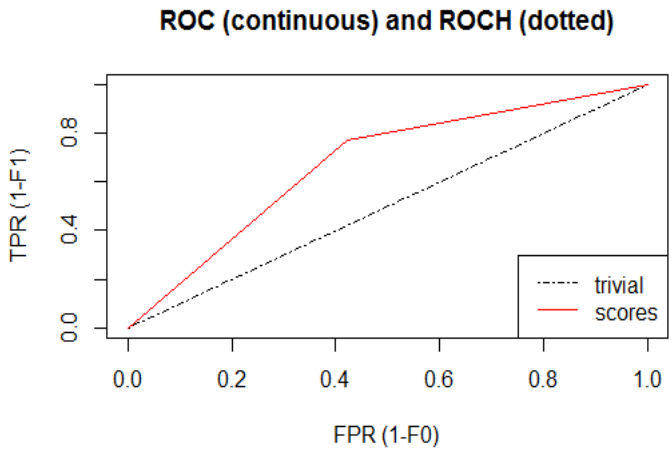
ROC curve of linear regression for BUPA



ROC curve of BRNN for BUPA

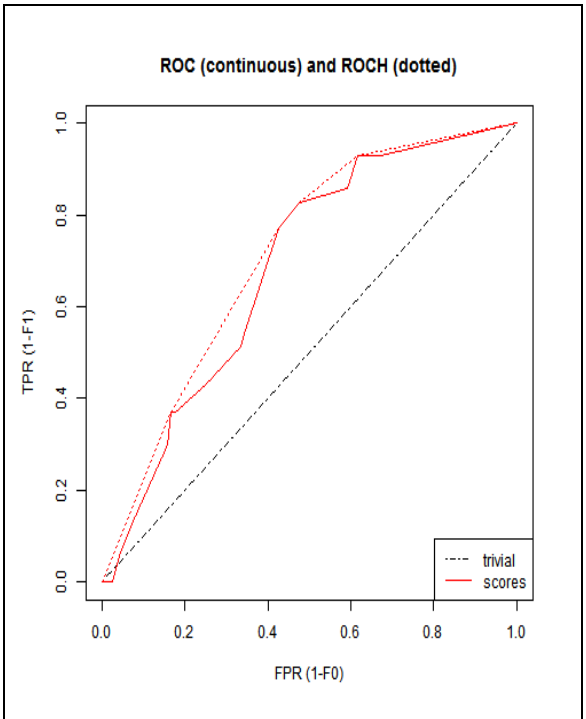


ROC curve of SVM for BUPA

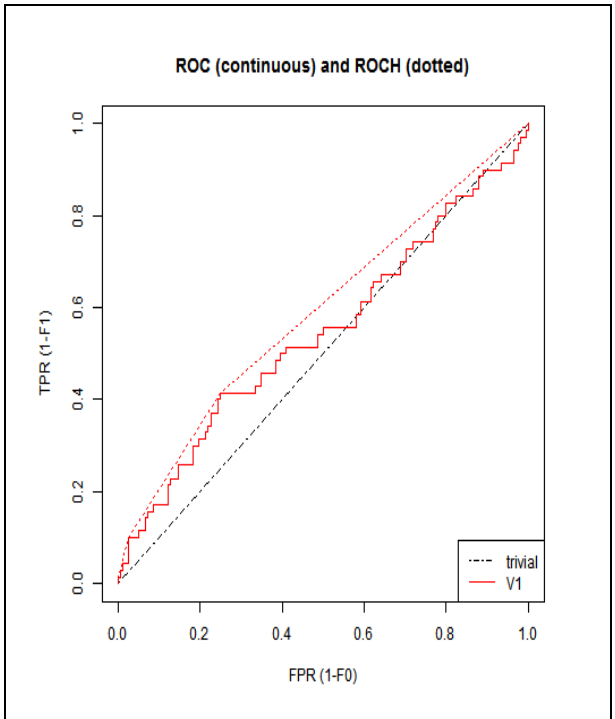


ROC curve of BrNdT for BUPA

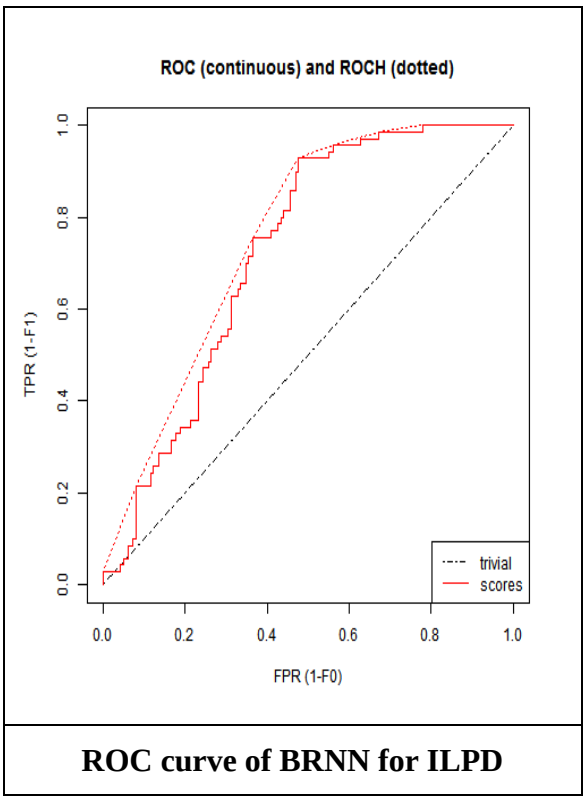
Figure 4.11: ROC curves for BUPA.



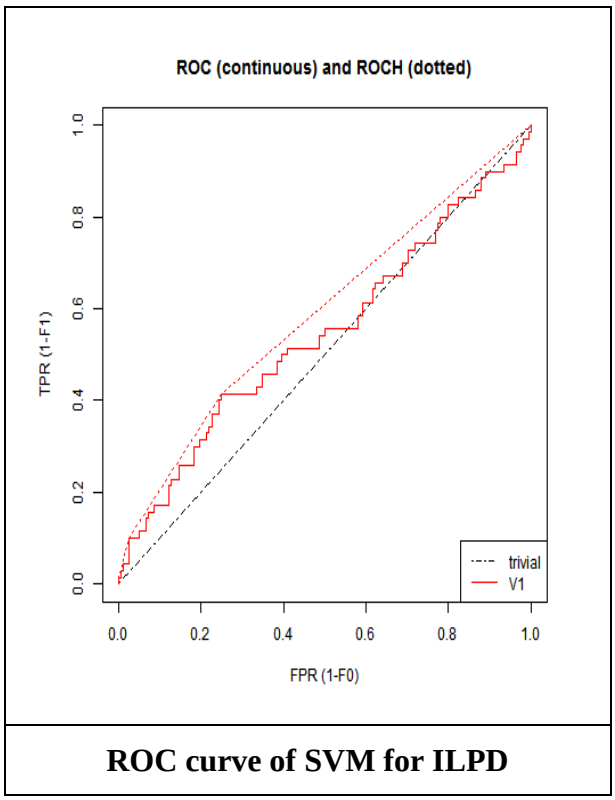
ROC curve of Decision Tree for ILPD



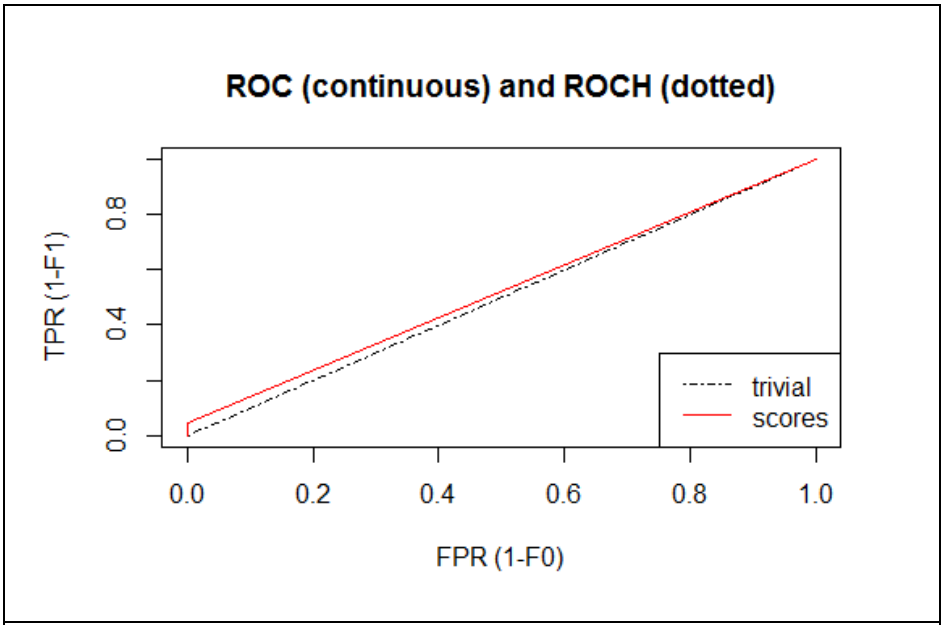
ROC curve of linear regression for ILPD



ROC curve of BRNN for ILPD



ROC curve of SVM for ILPD



ROC curve of BrNdT for ILPD

Figure 4.12: ROC curves for ILPD.

4.3 Optimization

The accuracies of the various classifiers such as SVM, NB, LDA, k-NN, DT and ANN are computed on various datasets using MATLAB also and listed in Table 4.13. The optimization of SVM parameters such as C (soft margin constant) and gamma (kernel parameter) is made with feature selection.

4.3.1 Simulation Software

The name MATLAB stands for Matrix Laboratory. MATLAB was written originally to provide easy access to matrix software developed by the LINPACK (linear system package) and EISPACK (Eigen System package) projects. MATLAB version 8.1.0.604 (R2013a) software has been used on 32 bit operating system for the computations.

MATLAB is a high performance language for technical computing. It integrated computation, visualization and programming environment. Furthermore, MATLAB is a modern programming language environment; it has sophisticated data structure, contains built-in-editing and debugging tools and supports object oriented programming. These factors make MATLAB excellent tool for teaching and research.

4.3.2 Analysis

It has been studied that the different classifiers performed varied on different datasets. It has been found that from the above datasets, only one of the dataset i.e. ovarian cancer had large number of attributes/features in the range of 1000, while the others used dataset in this work are having very less number of features. In that case, the second one, having the most number of features is the CpG island (38 features).

The mapping of data into higher dimensional space based on the concept of kernel cause computational problems. The SVM can be made more robust by the appropriate selection of regularization parameters like kernel parameter (gamma) and soft margin constant (C). The goal is to reduce the misclassification penalty.

4.3.3 Selection of the Parameters

Population size	:	100
Cross over function	:	0.95
Mutation function	:	0.05
Generations	:	100
Stall Generation Limit	:	10
No. of variables	:	20 (features) +1 (C) + 1 (gamma) = 22

Table 4.13: Accuracies (%) of various classifiers with different datasets using MATLAB.

	SVM	**NB**	**LDA**	**k-NN**	**DT**	**ANN**
Ovarian Cancer	99.06	82.87	99.30	90.65	**100.00**	93.75
Diabetes	73.96	73.44	76.82	69.01	**92.58**	72.17
Bupa	69.19	58.55	63.77	63.37	**90.14**	65.38
ILPD	57.38	59.18	61.75	63.23	**91.94**	78.16
CpG	82.19	81.81	84.65	83.00	**97.98**	81.08
Original Breast Cancer	96.84	95.85	96.42	96.28	**97.71**	81.90
Diagnostic Breast Cancer	96.48	94.20	96.83	94.71	**98.95**	30.58
Prognostic Breast Cancer	65.30	63.63	75.75	67.35	**90.40**	76.67

In the case of ovarian cancer dataset, many features are irrelevant in the detection of cancer. Only few features played that significant role in the cancer detection, which would be determined in the next step.

- Feature extraction is simply the process of extraction of relevant/significant features. Before the application of classification algorithm, significant features are selected by using genetic algorithm. By the means of feature selection, selection of the most informative feature leads to a robust model.
- The SVM parameters such as C and gamma are also optimized in order to keep the accuracy as higher as possible.

The optimization with the selected features is shown in Fig. 4.13 and Fig. 4.14. The obtained best fit and mean fit are shown in Fig. 4.15.

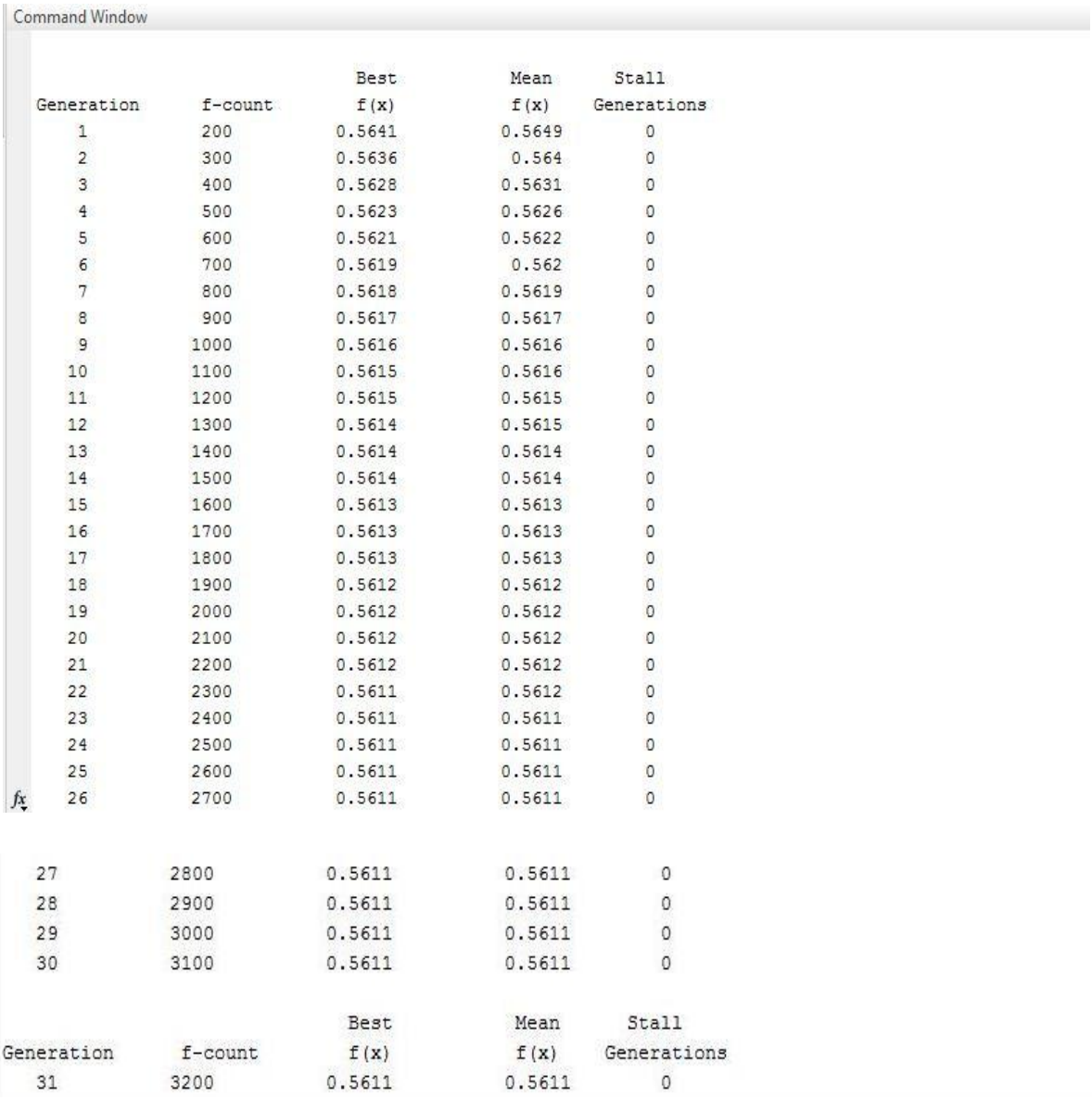


Figure 4.13: Clipping of the generations produced

```
The Optimized Value of C  is 0.4891
The Optimized Value of Gamma is 0.4465
The Selective Features Number is 2148
The Selective Features Number is 676
The Selective Features Number is 2782
The Selective Features Number is 2279
The Selective Features Number is 2391
The Selective Features Number is 2299
The Selective Features Number is 2689
The Selective Features Number is 2071
The Selective Features Number is 2196
The Selective Features Number is 2152
The Selective Features Number is 3130
The Selective Features Number is 3103
The Selective Features Number is 2118
The Selective Features Number is 2901
The Selective Features Number is 2100
The Selective Features Number is 703
The Selective Features Number is 214
The Selective Features Number is 1719
The Selective Features Number is 2122
The Selective Features Number is 2383
fx >>
```

Figure 4.14: Clipping of the selective features.

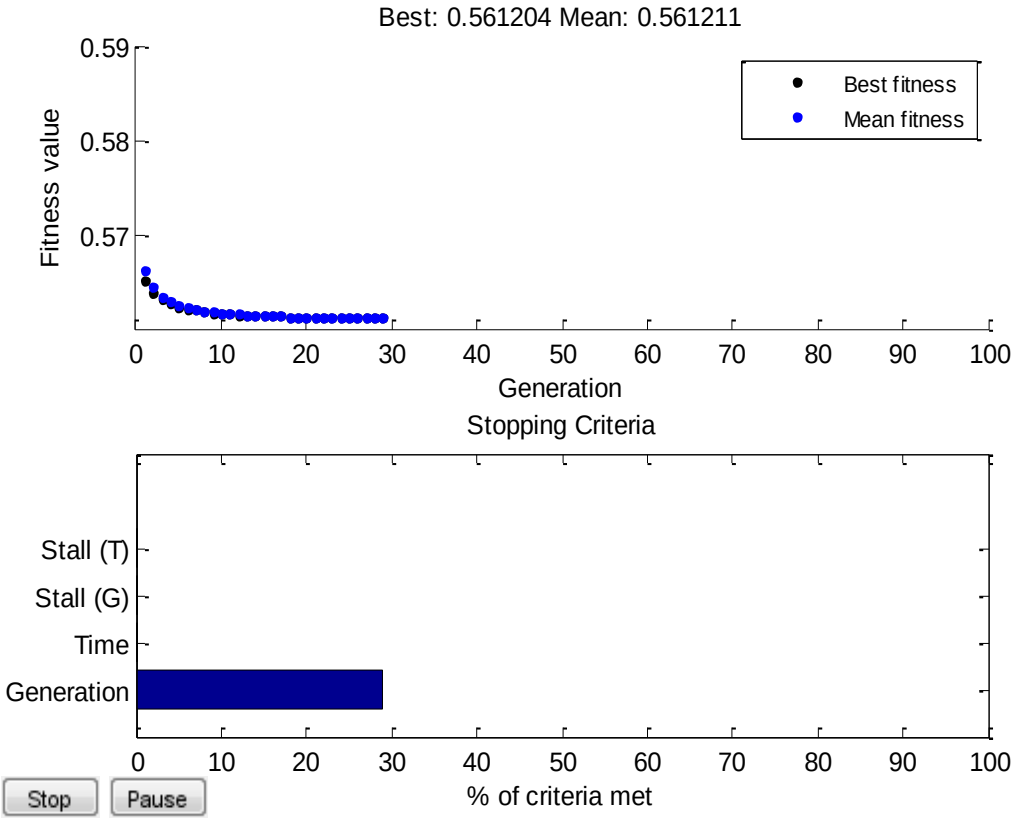


Figure 4.15: Best fit and mean fit values with stall generation

4.4 Problems encountered

- ✓ Missing Instances

  In some of the datasets, few values are missing. In such cases, many algorithms may be used to calculate those missing values [44]. Mode value of the attributes is simplest and easy to implement method to generate these missing values because mode is the most frequent occurrence of attributes. So we can assume that these missing values would be as equal as mode.

- ✓ Non-availability of normalized datasets.

  A common issue encountered with most of the databases is the absence of normalized values. Classification techniques give the maximum performance when its input arguments (features) are in the normalized form.

- ✓ Assumption of range.

  The range is not defined in the source sites of some databases from where the datasets has been obtained. Min/Max is extracted in order to find the range of continuous values.

- ✓ Whenever LDA is applied on the datasets with small number of features, it works well. But as LDA is applied on the datasets having large number of features such as CpG Island and Ovarian Cancer dataset, the error "The pooled covariance matrix of training must be positive definite" has been found.

  Solution: Before the application of LDA, PCA and other dimensionality reduction techniques can be applied for the reduction of the dimensionality of dataset.

- ✓ SVM works well on many of the databases, but on few databases, it causes the following error.

  Error: Maximum number of iteration exceeds.

  Solution: As some parameters are set by default in every system, so in order to avoid the above said error, increase the maximum number of iterations.

- ✓ Independency of datasets on SVM parameters.

  In some databases, the features are not separated from each other and dependent on the parameters such as C and gamma.

  In order to eliminate this error, random values of C and gamma are used. The SVM classifiers give the same accuracy on every attempt, no matter what is the value of C and gamma. The accuracy becomes independent of C and gamma.

- ✓ When GA-SVM is applied on smaller datasets, repetitive features are selected. For instance, with CpG Island having 38 attributes, most number of relevant features is 20. But those 20 features should be unique and distinct. The repetitive features are of no use.

CHAPTER - V
CONCLUSION AND FUTURE SCOPE

Biomedical data is quite difficult to handle due to its uncertainty as well as non-availability. Various efforts have been done by the researchers in order to collect all the possible datasets for the research. The classifier performance can't be made generalize for all types of dataset. As selection of the most suitable algorithm is dependent on various parameters such as the size of data samples, type of data samples, time limitations and type of prediction outcomes. Considering all the above parameters, still an effort is made to approach the generalization as much as possible. An ensemble based Bayesian regularized neural network decision Tree (BrNdT) model has been proposed for binary classification. SVM is the promising classification approach used in bioinformatics applications. As SVM is considered to be one of the most efficient machine learning algorithm for classification. The statistics of many experiments has figured out it to be accurate and best efficient model. But the experiments performed in comparison with the proposed model (BrNdT) have overcome this statistics. The proposed model is tested on different standard datasets to validate its robustness. The simulation results show that the proposed scheme not only offers a better accuracy but also reduces the simulation time of the model significantly compared to other existing methods.

Another effort that has been done in this research is the optimization of one of the classification technique i.e. support vector machine using Genetic algorithm. The main idea behind the optimization is the elimination of the irrelevant features which results in decrease of the computational time as well as the storage. It may also help to keep the classification accuracy as high as possible. But this is not certain in every case.

FUTURE SCOPE

Considering the time constraints, the above discussed research work was feasible. But it can be extended to more number of classifiers available with new kind of datasets. The selected features can be used further for predicting performance metrics such as accuracy and simulation time. The optimization can also be done using other optimization techniques such as particle swarm optimization, ant colony optimization, ABC optimization, and many more. The classifiers other than SVM can also be made optimized in order to obtain the best results possible. The relation between the classifiers and the type of the dataset can also be studied.

As this work is centered around the binary classification only. But real life multi-category class problems are more challenging and complex to solve. The work can be extended to design a multi-category class classifier model in future.

[1] Kayri Murat (2016). Predictive Abilities of Bayesian Regularization and Levenberg-Marquardt algorithms in Artificial neural networks: A comparative empirical study on social data", *Mathematical and Computational Applications Journal*, 21 (2), 1-11.

[2] Hayrettin Okut. Bayesian Regularized Neural Networks for Small n Big p Data, *Artificial Neural Networks- Models and Applications*. Intech, 2016.

[3] Asria, Hiba., Mousannifb, Hajar., Moatassimec, Hassan Al., Noel Thomas (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, *Procedia Computer Science* 83, 1064-1069.

[4] Pirooznia Mehdi et al. (2008). A comparative study of different machine learning methods on microarray gene expression data, *BMC Genomics*, 9, 1-13.

[5] Weiss S.M., Kulikowski (1991). Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning and expert systems, Morgan Kaufmann Publishers Inc, USA.

[6] Mehra. A, (2003). Statistical sampling and regression: simple linear regression, *PreMBA analytical methods*, Columbia business school and Columbia University.

[7] Devi Nirmala, alias Appavu and Swathi (2013). An amalgam KNN to predict Diabetes Mellitus, *International Conference in Emerging Trends in Computing, Communication and Nanotechnology*, pp. 691-695.

[8] Michie D,. Spiegelhalter D.J, & Taylor (1994*).* Machine learning, Neural and Statistical classification, Ellis Horwood, USA.

[9] Janani S., Ramya Chitra D (2016). Examining Classification Techniques in Data Mining for PIMA Indian Diabetes Dataset, *International Journal for Scientific Research & Development*, 4, 629-633.

[10] Barale M.S., Shirke D.T (2016). Cascaded Modeling for PIMA Indian Diabetes Data, *International Journal of Computer Applications*, 139(11), 1-4.

[11] Patil B.M, Joshi R.C, Toshniwal Durga (2010). Hybrid Prediction model for Type-2 Diabetic patient, *Expert Systems with Applications*, 10, 8102-8108.

[12] Polat Kemal, Gunes Salih (2006). An expert system approach based on principal component analysis and adaptive neuro-fuzzy interference system to diagnosis of diabetes disease, *Digital Signal Processing*, 17, 702-710.

[13] Kahramanli Humar, Allahverdi Novruz (2008). Design of a hybrid system for the diabetes and heart Diseases, *Expert Systems with Applications*, 35, 82-89.

[14] Polat Kemal, Gunes Salih, Arslan Ahmet (2008). A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support machine, *Expert Systems with Applications*, 34, 482-487.

[15] Carpenter GA, Markuzon N (1998).ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases, *Neural Networks*, 11, 323-336.

[16] Bioch J.C, Meer, Potharst Rob (1996). Classification using Bayesian neural Nets, *International Conference on neural networks*, 1488-1493.

[17] Smith Jack W, Everhart J.E, Dickson W.C, Knowler W.C and Johannes R.S (1988). Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus, *Proceedings of symposium on computer applications and medical cares*, 261-265.

[18] Ali Isse, Mohamoud Hussein Sheikh ali (2011). An identification and prediction methods for feature-subsets of CpG islands methylation based on human peripheral blood leukocytes of chromosome, *International Conference of the IEEE EMBS Boston*, 3233-3236.

[19] Timp, W & Feinberg (2013). Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host, *Nat Rev Cancer*, 497-510.

[20] Glasspool, R, Teodoridis, J. M, & Brown (2006). Epigenetics as a mechanism driving polygenic clinical drug resistance, *British Journal of Cancer*, 1087-1092.

[21] Konstantina, Themis, Michalis (2014). Machine learning applications in cancer prognosis and prediction", *Computational and Structural Biotechnology Journal*, 13, 8-17.

[22] Lee Eunhye, Moon Aree (2016). Identification of Biomarkers for Breast Cancer Using Databases, *Journal of Cancer Prevention*, 21 (4), 235-242.

[23] Bazazeh Dana and Shubair Raed (2017). Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis", *International Conference on Electronic Devices, Systems and Application*, 2159-2055.

[24] Rathi Megha, Pareek Vikas (2016). Hybrid Approach to predict Breast Cancer using Machine Learning Technique, *International Journal of Computer Science Engineering*, 5(3), 125-136.

[25] Thein Htet Thazin Tike and Tun Khin Mo Mo (2015). An approach for breast cancer diagnosis classification using neural network, *Advanced Computing: An International Journal*, 6 (1), 1-11.

[26] Christopher, Oscar, Igor, Val Coral del (2009). Profile analysis and prediction of tissue-specific CpG island methylation classes, *BioMed Central Bioinformatics*, 10, 1-16.

[27] Hazra Animesh, Mandal Subrata Kumar (2016). Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble algorithms", *International Journal of Computer Applications*, 145, 39-45.

[28] Kumar G. Ravi, Ramachandra G.A. (2016). An Efficient Prediction of Breast Cancer Data using data mining techniques, *International Journal of Innovations in Engineering and Technology*, 2,.139-144.

[29] Kharya Shweta, Agrawal Shikha, Soni Sunita (2014).Naïve Bayes Classifiers: A probabilistic detection model for breast cancer", *International Journal of Computer Applications,* 92, 26-31.

[30] Salama Gouda I, Abdelhalim M.B and Zeid Magdy Abd-elghany (2012). Breast cancer diagnosis on three different datasets using multi-classifiers, *International Journal of Computer and Information Technology*, 1, 36-43.

[31] Das, Nevenka, Zhenyu (2005). Computational prediction of methylation status in human genomic sequences, *Proceedings of the National Academy of Sciences of the United States of America,* 103, 10713-10716.

[32] Jeetha B.Rosiline, Malathi M (2013). Diagnosis of Ovarian cancer using artificial neural network, *International Journal of Computer Trends and Technology*, 4(10), 3601-3606.

[33] AlpanaJijja, Rai Dinesh, Mathur Priyanka (2017). Comparative Analysis of Feedforward Back propagation and Cascade Correlation Algorithm on BUPA liver disorder, *International Journal of Engineering and Technology*, 8(6), 2912- 2917.

[34] Shrivastava Pooja, Kesharwani Yukti (2016). AN efficient classifier using multilayer perceptron for classification of liver patient, International Journal of Research in Applied Science and Engineering, 4(6), 646-648.

[35] Sindhuja D, Priyadarsini R Jemina (2016). A survey on classification techniques in data mining for analyzing liver disease disorder, *International Journal of Computer Science and Mobile Computing*, 5(5), 483-488.

[36] Pakhale Harsha, Xaxa Deepak Kumar (2016). A survey on diagnosis of Liver disease classification, *International Journal of Engineering and Technology*, 2(3), 132-138.

[37] Phan Anh Viet, Nguyen Minh Le, Bui Lam Thu (2017). Feature weighting and SVM parameters optimization based on genetic algorithms for classification problems, *Springer*, 46 (2), 455-469.

[38] Liao Pin, Zhang Xin, and Li Kunlun (2015). Parameter Optimization for support vector machine based on nested Genetic Algorithm, *Journal of Automation and Control Engineering*, 3(6), 507-511.

[39] Andy, Fernando Michael, Halim Kristanto, and Sanjaya Gradiyanto (2013). Optimization Features using GA-SVM approach, International Journal of Science and Research, 4(9), 193-197.

[40] Zhou Jing, Maruatona Omaru O,and Wang Wei (2011). Parameter Optimization for Support vector machine classifier with IO-GA, *Complexity and Data Mining (IWCDM,* 117-120.

[41] Polikar Robi (2006). Ensemble based systems in Decision making", *IEEE Circuits System Magazine*, 6, 21-45.

[42] Rokar Lior (2010). Ensemble based Classifiers, Artificial Intelligence Review, 33, 1-39.

[43] Christopher, Oscar, Igor, Val Coral del (2009).Profile analysis and prediction of tissue-specific CpG island methylation classes, *BioMed Central Bioinformatics*, 10(1), 116-132.

[44] Pelckmans K, Brabanter J De, Suykens J.A.K, Moor B.De (2005). Handling missing values in support vector machine classifiers, *Neural networks, Elsevier*, 18(5-6), 684-692.

LIST OF PUBLICATION

"Bayesian regularized Neural network decision Tree ensemble model for Genomic Data classification", *Applied Artificial Intelligence*, Communicated, 2017. (*Science Citation Indexed*).