Multiclass Diagnosis Model for Heart Disease using PSO based SVM

A Thesis Report

submitted in partial fulfillment of the requirements

for the award of degree

of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Prerna Dembla

(801432021)

Under the supervision of:

Ms. Tarunpreet Bhatia

Lecturer

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

July 2016

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "***Multiclass Diagnosis Model for Heart Disease using PSO based SVM***", in partial fulfillment of the requirements for the award of degree of Master of Engineering in ***Computer Science and Engineering*** submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of ***Ms. Tarunpreet Bhatia*** and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Prerna Dembla
801432021
ME (CSE)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Ms. Tarunpreet Bhatia
Lecturer
Computer Science and
Engineering Department
Thapar University

Countersigned by

Dr. Maninder Singh
Head
Computer Science and Engineering Department
Thapar University
Patiala

Dr. S. S. Bhatia
Dean (Academic Affairs)
Thapar University
Patiala

ACKNOWLEDGEMENT

I would like to express my most sincere appreciation and deep sense of gratitude and indebtedness to my guide **Ms. Tarunpreet Bhatia,** Lecturer, Computer Science and Engineering Department, Thapar University, Patiala for providing me continuous guidance throughout the research. It has been possible to proceed in the correct direction and successfully complete the thesis with her valuable suggestions. I would like to thank her for encouraging my research and for being actively involved in my work.

I am also thankful to Dr. Maninder Singh, Head of Department, CSED and Dr. Ashutosh Mishra, P.G. Coordinator, for the motivation and inspiration that triggered me for the thesis work.

I am also very thankful to the entire faculty and staff members for the direct-indirect help, cooperation and support.

Prerna.

Prerna Dembla

(801432021)

ABSTRACT

One of the most significant domains of Machine learning is in the Healthcare Industry which helps the medical professionals in the automation of medical diagnosis process and in the development of a disease prediction system that is highly powerful in reducing the patient mortality rate. The count of people dying every year from heart disease is increasing drastically. The multiclass model for diagnosing heart disease has been proposed using PSO based SVM. It classifies the heart disease into 5 classes namely healthy, low-risk, medium-risk, high-risk and danger. The severity of disease increases from healthy to danger. Principal Component Analysis has been used as a dimensionality reduction step to choose the subset of attributes that best reflects the original heart dataset. Support vector machine is a promising supervised method that classifies data by functional hyperplane which separates two classes from each other. The accuracy of SVM was enhanced by global stochastic optimization technique called Particle swarm optimization. The proposed algorithm is implemented for both 2-class and 5-class problems. The performance of multiclass model has been estimated on the basis of accuracy, recall, precision and F-measure. The results indicate that the attained classification accuracy is very promising as compared to the other existing algorithms. The proposed approach can be successfully used for determining the severity level of heart disease.

Keywords: Support vector machine, Heart disease, Particle swarm optimization, Classification, Principal component analysis, Medical diagnosis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANN Artificial Neural Network

AUC Area under curve

BP Back Propagation

ELM Extreme Learning Machine

GA Genetic Algorithm

KNN K-Nearest Neighbor

LVQ Learning Vector Quantization

MLP Multilayer Perceptron

PCA Principal Component Analysis

PNN Probabilistic Neural Network

PSO Particle Swarm Optimization

RBF Radial Basis Function

SA Simulated Annealing

SMO Sequential Minimal Optimization

SMOTE Synthetic Minority Oversampling Technique

SVM Support Vector Machine

WAC Weighted Associative classifier

CHAPTER 1

INTRODUCTION

1.1 Introduction to heart disease

The Heart is one of the most important organs in the human body. It is the center of the circulatory system. The heart functions as a pump that propels blood to different parts of the human body through a network of blood vessels, supplying a constant supply of oxygen as well as other vital nutritional components. Without the proper working of the heart, multiple other organs would stop working. If the heart ever stops functioning and ceases to pump blood, the body will shut down and within very less time a person will expire.

According to the association's 2015 Heart Disease and Stroke Statistics Update [1], cardiovascular disease is the leading worldwide cause of death, accounting for 17.3 million deaths per year and by 2030 it is estimated to increase to more than 23.6 million. In 2011, nearly about 787,000 people from the U.S. died from cardiovascular diseases, heart disease and stroke. It has been claimed that heart disease causes more death than all types of cancer combined. The count of people dying every year from cardiovascular disease is increasing drastically. If heart disease is identified and diagnosed precisely at an early stage and proper subsequent treatment is provided then considerable life can be saved and death rate can be reduced.

Machine learning presents various algorithms for analysis of medical data. It helps in diagnosis and prediction of healthcare problems untimely. Patient data is gathered with the help of data collection equipment and stored in a computer system in the form of medical records for treatment. Machine learning algorithms help in the diagnosis process of a new patient by analyzing the data pattern of the patient admitted in the past. It examines the disease, symptoms faced and the adequate treatment provided to the patient and uses that information for a newly admitted patient. Machine learning automatically learns through experience and performance of algorithm gets improved with each experience. The diagnosis system involves dividing the dataset into various classes depending on the values of the attribute measured for each instance of a patient. Machine learning has attained notable results and can be successfully used in the healthcare industry. In this thesis, a multiclass

model for diagnosing heart disease has been proposed using PSO based SVM. The proposed approach develops a heart disease diagnosis system that determines the criticality level of a newly presenting patient. It uses support vector machine classifier to classify heart disease dataset into multiple classes.

1.2 Motivation of the thesis

The diagnosis process is a decision making process in which decisions are made by the medical experts with the help of their knowledge and the experience they get with the treatment of patients suffering from same problem and symptoms. Disease diagnosis is a complicated process and may lead to wrong assumption as some factors are associated with many organs. Hence, there is a need to automate the medical diagnosis process and develop a diagnosis system to determine the stroke level of heart disease with higher precision and without causing any delay in the proper subsequent action. The number of tests is conducted on the patient for the diagnosis of a disease. With the use of machine learning technique for predicting the disease, the number of tests can be reduced which saves time and provides quality services at a reasonable cost.

1.3 Organization of the thesis

The work is presented in thesis as:

Chapter 2 gives an overview of machine learning and the various classification techniques along with their advantages and disadvantages. Chapter 3 describes the literature review on machine learning techniques. The various approaches proposed by researchers are studied to help medical experts in predicting various diseases. Chapter 4 describes the problem statement and objective of this work. Chapter 5 illustrates the proposed algorithm and their implementation. Chapter 6 gives the result of proposed algorithm and comparison of our method with different approaches. Chapter 7 concludes the work done and proposes the future enhancement of work.

CHAPTER 2

BACKGROUND

2.1 Introduction to Machine learning

Machine learning is a domain of artificial intelligence involving the construction of algorithms that automatically learns through experience and performance of algorithm gets improved with each experience. Algorithm operates by detecting some pattern in input data and building a model based on input data to make precise predictions for new data.

2.2 Machine learning Techniques

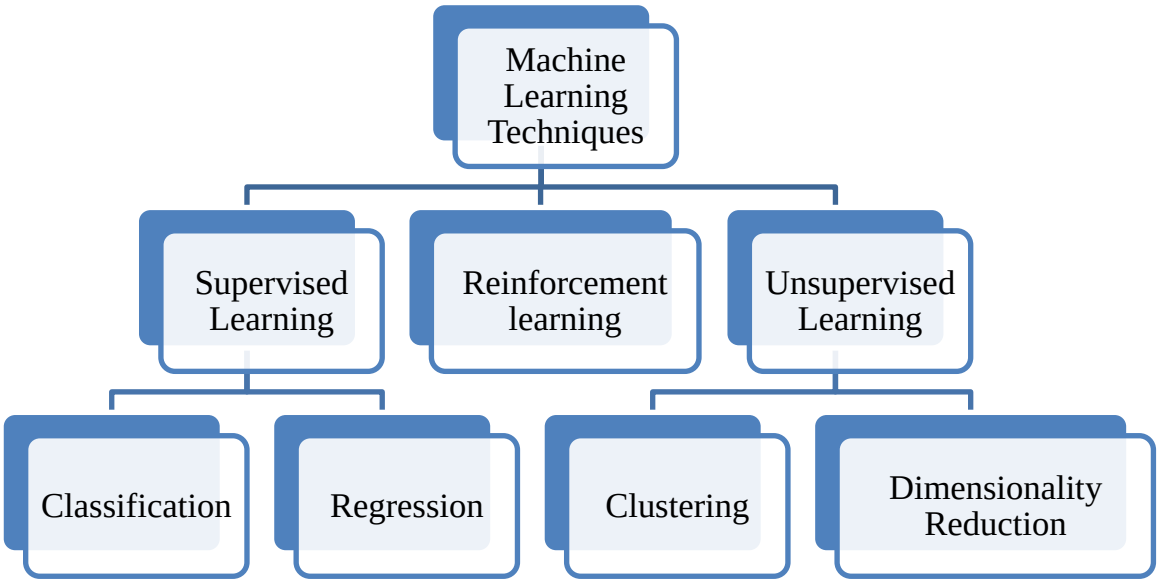

Figure 2.1: Machine learning techniques

2.2.1 Supervised Learning

In case of supervised machine learning, algorithm induces a mapping function from given labeled training dataset to map new input data to its desired output [2]. Labeled training dataset comprises of examples, which is a pair of input data and its output value. Classification problems are the most common method of supervised learning. For example, suppose we have a record of previous heart patients, including blood

pressure, sex, age, cholesterol, weight, etc. within a year and want to predict whether a new patient is prone to a heart attack or not.

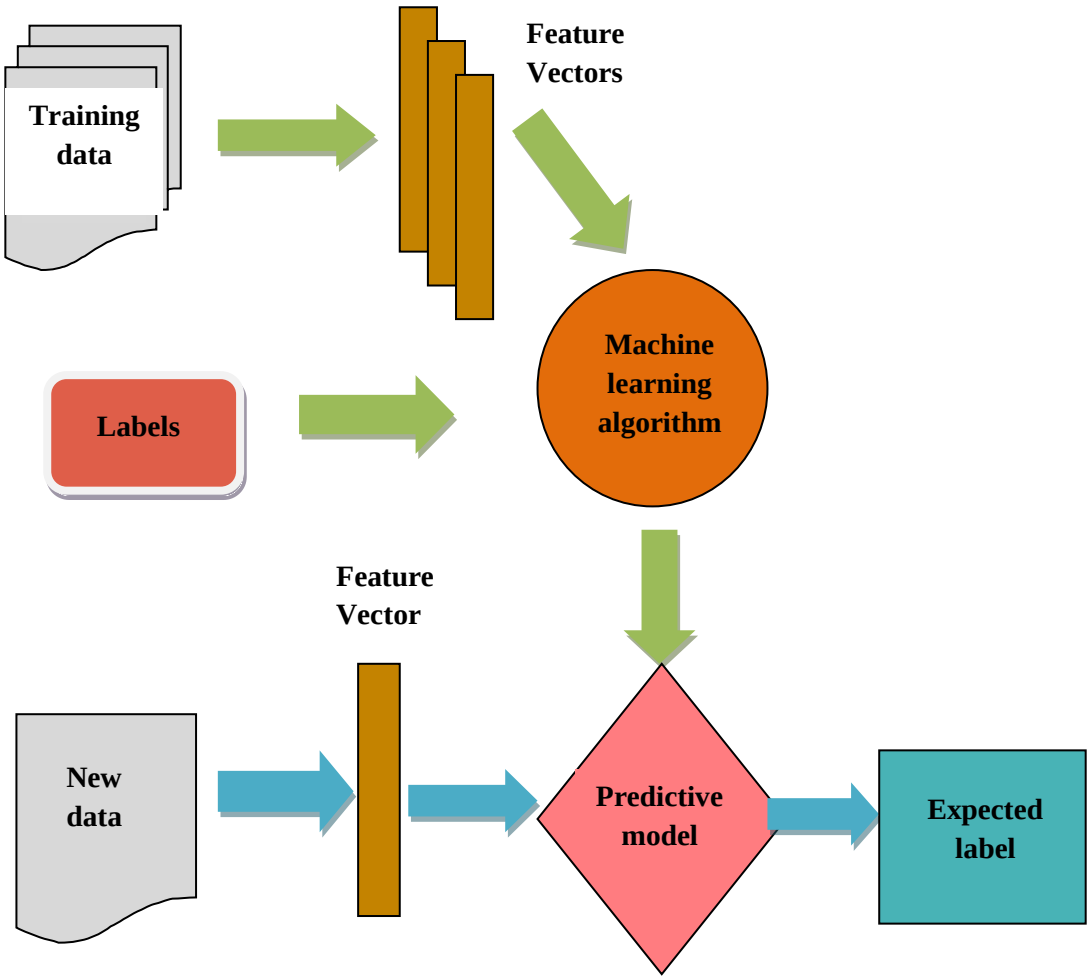

Figure 2.2: Supervised learning

The problems solved by supervised learning are:

- **Classification**

  Classification is a process of classifying the data into their respective classes. It mainly includes two phases. The first phase is the training step and building classifier in which a classifier is trained to analyze the given data records and the class with which they are associated. The second phase is the testing step in which model classifies the test dataset on the basis of pattern analyzed in the first phase [5]. Different classifiers can be applied to the same dataset and best classifier can be selected by comparing the performance metrics like accuracy, specificity, robustness, speed, precision and recall.

- **Regression**

  Regression is a process of predicting the output variable on the basis of the training set. The output variable is in the form of continuous value. It can be symbolized in the form of regression tree where the data gets divided with each branch and results in final predicted value. The prediction of price in stock market is an example of regression.

2.2.2 Unsupervised Learning

In case of unsupervised machine learning, algorithm infers a mapping function to find hidden patterns and correlation between them from unlabelled input dataset [3]. Input dataset comprises of examples, each example is an input data with no explicit output value. For example, we have to discover close-knit group of friends in facebook.

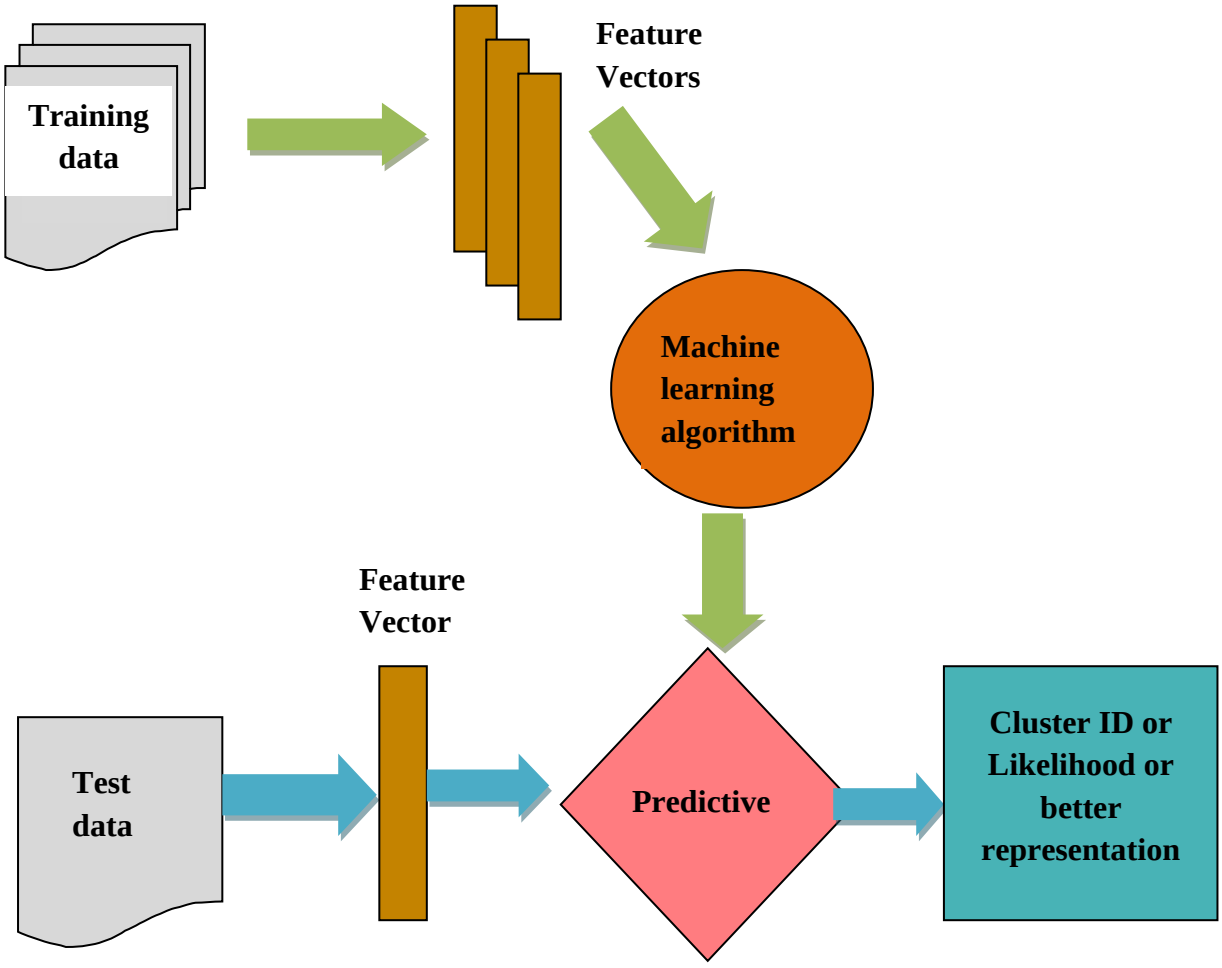

Figure 2.3: Unsupervised learning

The problems solved by unsupervised learning are:

- **Clustering**

  Cluster analysis is the most common unsupervised learning method. The goal is to find similarities in the given training data by metrics like Euclidean distance and group them in a cluster. Algorithm operates by assigning example to one of the cluster depending upon its similarity with the cluster. It works well when there is an adequate data.

- **Dimensionality reduction**

  Dimensions of data can be reduced either by choosing the most significant features (Feature selection) or transforming features to a small set of features (Feature extraction). Every feature in the given dataset contributes differently. Moreover, high dimension of data results in more computation cost.

  Advantages

  1. Visualization of the data becomes easy.
  2. The computation cost and space to store the data get reduced.
  3. The performance of model built after dimensionality reduction step generally gets improved.

2.2.3 Reinforcement learning

Reinforcement learning enables the machine to perform some action in an environment in such a way that the machine gets rewarded for each right action to signify success and get punished for each wrong action [4]. Hence, the aim of this learning is to make such decisions that maximize rewards. The machine can make decision on the basis of learning in the past that which action at what scenario helps in contributing to success.

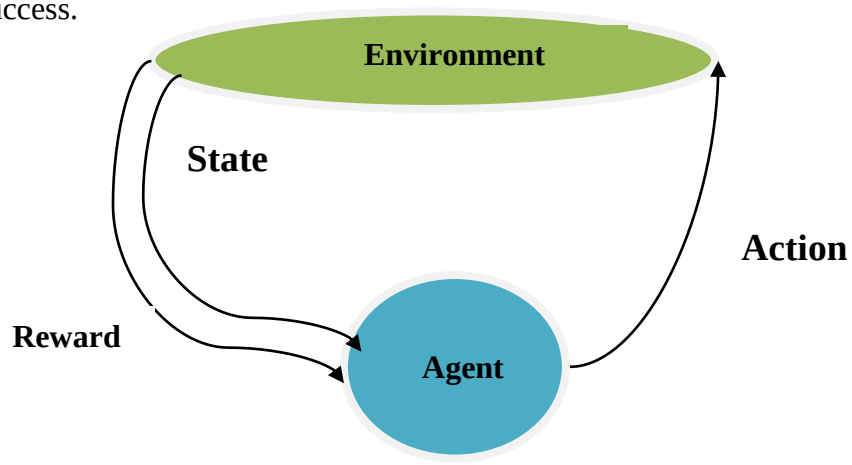

Figure 2.4: Reinforcement learning

2.3 Classification algorithms

2.3.1 Naïve Bayesian

It is a statistical and supervised method of classification based on Baye's theorem. The approach is named as naïve as it assumes that all the variables are independent of each other and contribute independently to the probability. It works well for very large datasets. It is very simple to use and give better results than highly complicated classification algorithm. It takes a set of input from the training set and applies bayes theorem to compute the probability for each class. Predicted class is the one having the highest probability [6] [7].

Bayes Rule:

Bayes theorem gives a method of calculating posterior probability P(c|d) from P(c), P(d) and P(d|c) by the following equation:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

where:

P(c|d) is the posterior probability of class c given some training data d.

P(c) is the prior probability of class c.

P(d) is the prior probability of training data d.

P(d|c) is the likelihood of training data d given class c.

Advantages

1. It is very simple to implement and fast method to predict the class in test data set.
2. It works well in case of multi-class problem.
3. It gives good results if the input variables are categorical.
4. It outperforms to other complex methods like logistic regression.
5. To estimate the parameters, small amount of training data is required

Disadvantages

1. If variable has some category which is present only in testing set but was not present in the training set, then model was not able to predict this category. Smoothing technique such as Laplace estimation is the solution for this. Sometimes, it is considered as a bad estimator.

2. It assumes that variables are completely independent to each other but it is not always practical.

2.3.2 K-nearest neighbor

K-nearest neighbor (KNN) is one of the simplest used algorithms for classification. It is also called as a lazy learning algorithm as it takes into account complete training set and does not perform any generalization on training set [8]. It considers K-nearest neighbors for classification of given instance 'X' [9]. The distance between the attributes of X ($x_1, x_2, \ldots x_n$) and its neighbors ($y_1, y_2, \ldots y_n$) is measured by Euclidean distance using the following formula:

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots\ldots (x_n - y_n)^2}$$

The neighbors of X vote to predict the class of X and X is alloted to the class having majority votes by its neighbors.

Algorithm to find the class of instance X

1. Given a new instance 'X' and count of nearest neighbors 'k' which is to be considered.
2. Find k training set instances which are nearest to instance 'X' using Euclidean distance.
3. Find the majority of votes by considering each neighbor vote.
4. Instance 'X' is assigned to the majority winning class.

Advantages

1. Robust to noisy training data.
2. It gives good results for large training data also.
3. It is very simple algorithm to implement.
4. It is highly adaptive as it uses only local information.

Disadvantages

1. As the number of attributes increases, accuracy of algorithm decreases.
2. It requires more memory for large size training dataset which makes the prediction process slow.
3. Computationally intensive recall.

2.3.3 Multilayer Perceptron

Artificial Neural Network (ANN) is the technique used for processing information and is based on the working of Brain. The most common neural network model is Multilayer perceptron. This is supervised learning algorithm which means that it acquires knowledge through learning [10]. M.Minsky and S.Papert introduced concept of Multilayer perceptron in 1969. The single layer perceptron is basically used for solving problems that are linearly separable. For non linearly separable problems, more layers are added to the neural network. These are also known as Feed Forward neural networks because the input signals propagate layer by layer in the system. These are mostly used for classification, pattern recognition, approximation and prediction [11].

The architecture of Multilayer Perceptron (MLP) consists of 3 layers:

1. Input Layer: This is the first layer. It accepts the input vector or pattern from the user. It standardized the value of each variable in the range of -1 to 1 and passes the result to first hidden layer.
2. Hidden Layers: There can be more than one hidden layers depending upon the complexity of the problem. This layer extracts more meaningful information from input. The value of input neuron is multiplied with the weight associated with it. The combined value is produced by adding together the resulting values. This combined value is then fed to the last layer i.e. output layer.
3. Output Layer: It receives the values from the last hidden layer, weighs them and generates the target value.

Back Propagation (BP) Algorithm

The back-propagation algorithm is mainly used to training neural networks. The intelligence of the networks is based on the valves of weights. Using back propagation we can adjust the values of these weights. The neural network learns in small iterative steps. The system is initialized with random weights and known input vector is fed to the system. The obtained output is compared with desired output and error value is calculated. This value is propagated back to the system and weights are adjusted to reduce the error signals. These steps keep on repeating until the output reaches some predetermined threshold value. This procedure is called as Training phase.

Advantages

1. They are used for solving non linear models.
2. They can learn how to do task using training sets.
3. They are used in generalization problems. They can classify unknown or incomplete problems sharing the characteristics with known and complete problems.
4. They are highly fault tolerant. The whole system does not halt even after some neuron or interconnections fails.

Disadvantages

1. It can easily traps into the local minima instead of settling into global minimum of energy.
2. It is highly sensitive to feature scaling.

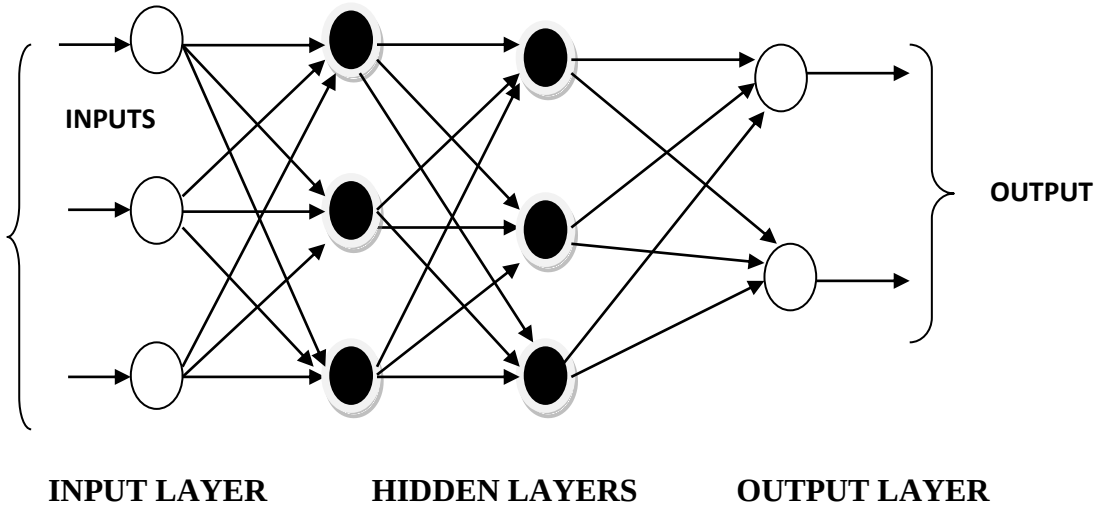

Figure 2.5: Multilayer perceptron

2.3.4 Decision tree

Decision tree are supervised method used for the prediction of categorical as well as numerical value [12].They represents the data instances along with their class label in the form of a tree. A set of rules can be deduced from the tree which can be used to classify the unknown data record to its output value. A test on an attribute is performed on the internal node. The result of the test is depicted by the branch of tree and class label are present at the leaf node [13].

Advantages

1. It can be used for the prediction of categorical as well as numerical value.
2. Decision trees are very easy to understand as data can be visualized.

3. Multi-output problems can be solved by decision trees.
4. It is very easy to deal with irrelevant attributes with the help of information gain.

Disadvantages

1. Decision trees are prone to overfitting.
2. If the dataset is imbalanced then biased trees are generated. The dataset should be balanced before giving it as an input to decision tree.
3. Problems like multiplexer or XOR are not generalized well by decision tree.
4. A completely different tree can be generated with minute deviation in the data. Ensemble can solve this problem.

2.3.5 Support Vector Machine

Support Vector Machine is a supervised method of classification invented by Vladimir Vapnik and Chervonenkis in 1963 and proposed as a kernel based learning method for classification of non linear data in 1993 [14]. It is able to predict classes from both linear and non linear data and works well on reasonably sized datasets only. The original training data is transformed into higher dimension using non linear mapping and search for the linear optimal separating hyperplane within higher dimension. Hyperplane is the decision surface that separates the data from two classes in such a manner that data of one class are on one side of the hyperplane and of other class are on other side. Hyperplane can be found using support vectors and margins.There can be more than one hyperplane possible that classify the data. The hyperplane that denotes the largest margin between the decision boundaries of two classes is the best one [15].

Suppose *f* is a function that maps input to output for SVM classification then,

$$f: I \rightarrow O$$

Let the dataset be given as {X, Y} where

$$X = \{x_i \mid 1 \leq i \leq n\}, \text{ a set of n training tuples}$$

$$Y = \{y_i \mid 1 \leq i \leq n\}, \text{ associated class label}$$

Each y_i belongs to either +1 or -1, that corresponds to two classes of dataset.

$$y_i \in \{+1, -1\}$$

Assume that the pattern characterized by the subset $y_i = +1$ and $y_i = -1$ are linearly separable.

The decision surface can be represented in the form of a hyperplane as

$$f(x) = 0$$

$$W.X + b = 0$$

where W represents weight vector and b represents the bias value.

Hence, any point from one class lies above the separating hyperplane satisfies,

$$W.X + b > 0$$

Similarly any point from another class lies below the separating hyperplane satisfies,

$$W.X + b < 0$$

P1 and P2 are the two planes:

$$P1:\ x_i.w + b = +1$$

$$P2:\ x_i.w + b = -1$$

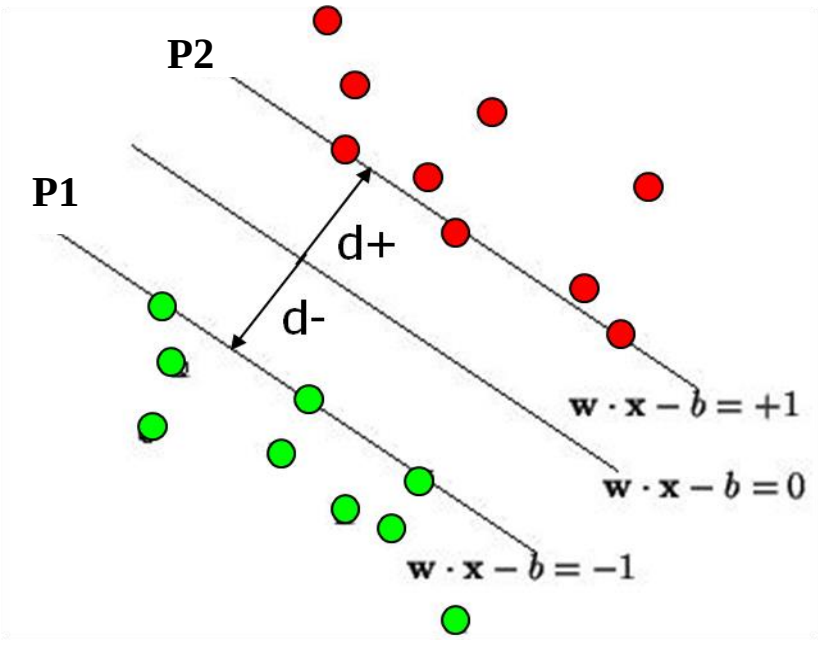

Figure 2.6: Decision boundary and margins of SVM

The equations of the planes P1 and P2 can be combined to form

$$Y_i(W.X + b) \geq 1 \text{ , for all i}$$

Which can be solved by lagrange multipliers $(\alpha_i \geq 0 (i = 1,2, \ldots . n))$

The decision vector can be represented in the form of following equation

$$y(x) = sgn\left[\sum_{i=1}^{n} \alpha_i y_i \langle x_{i,} x\rangle + b\right]$$

In case of non linear problems, SVM with kernel functions are used by which data get transformed to a higher dimensional space [16]. Selection of different kernel functions produces different results. The decision function can be represented as:

$$y(x) = sgn\left[\sum_{i=1}^{n} \alpha_i y_i K\langle x_{i,} x\rangle + b\right]$$

where $K\langle x_{i,} x\rangle = \langle \emptyset(x_i), \emptyset(x)\rangle$ is kernel function and $\emptyset(x)$ is mapping to high dimension space.

The kernel functions of SVM are:

1. Linear: $K\langle x_{i,} x\rangle = x^T x_j$
2. Polynomial: $K\langle x_{i,} x\rangle = (\gamma x^T x_j + r)^d, \gamma > 0$
3. Radial basis function: $K\langle x_{i,} x\rangle = \exp\left(-\gamma \|X_i - X_j\|^2\right), \gamma > 0$

Advantages

1. They are not vulnerable to overfitting because of presence of regularization parameter.
2. SVMs always output a unique solution as there is no local minima concerned.
3. By selecting appropriate values of C and r, they provide good generalization.
4. They give good results in case of complex decision boundaries also.

Disadvantages

1. SVMs can be very slow as they are computationally expensive and take more time in training.

2. Proper selection of kernel is needed for good results.
3. For large scale tasks, SVM have more memory requirements and high algorithmic complexity.

CHAPTER 3

LITERATURE REVIEW

Many machine learning approaches are proposed by researchers to help medical experts in the implementation of highly reliable diagnosis system for the classification of disease with techniques like artificial neural network, learning vector quantization, decision tree, neural networks, naïve bayesian, k-nearest neighbor etc. This section summarizes some works of machine learning in the health care industry that found in literature survey.

Zhang et al. [17] proposed an efficient coronary heart disease prediction system using Support Vector Machine. In this, Principal Component Analysis (PCA) was used to extract the important features and different kernel functions were utilized as a classifier. The highest classification accuracy is achieved with Radial Basis Function (RBF). To find the optimal parameters values, Grid search method was employed and optimal values were found to be c=1 and g=0.0909. The highest classification accuracy reached is 88.6364%. It was used for prediction of two classes .

Naib et al. [18] suggested classification system of primary tumors using multiclass classifier with Random Forest. In this, Synthetic Minority Oversampling Technique (SMOTE) is used as a preprocessing step using to remove biasness towards majority classes by adding the instances of minority classes. The experimental study was implemented in weka tool. The dataset comprises of total 22 classes of tumor. The classification is performed with different machine learning algorithms. The result shows that random forest with 10 random trees outperforms with the accuracy of 85.7% and ROC area of 0.997.

Jabbar et al. [19] proposed a novel approach which combines K-nearest neighbor with genetic algorithm to predict two classes of heart dataset which are healthy and sick. The approach has been evaluated using the cross validation on 6 medical datasets from UCI dataset repository and heart disease AP. In this, genetic algorithm is used to select the high ranked features which are more significant to others. The accuracy of classification reached is 81.4% when 5 nearest neighbors are considered. The disadvantage of genetic algorithm is that it performs many evaluations to evaluate fitness value and is not a good way to find local optima.

Santhanam et al. [20] suggested a Heart disease prediction method. In this study, PCA1, PCA2, PCA3, PCA4 methods which are produced by applying various operations on principal components extracted from PCA and a new method called exponentiated estimate of the coefficient are considered for feature selection. The performance is evaluated using Feed Forward Neural Networks and regression classifier. The maximum accuracy reached of regression classifier is 92% with PCA1. The proposed system was for classification of two classes.

Lin et al. [21] proposed an SA-SVM method in which Simulating annealing (SA) approach is used, which searches for continuous decision variable to find optimal feature subset and parameter values. The proposed method is implemented on a number of datasets taken from UCI repository and results are compared to the grid search method and many other classification methods. The implementation of the method on heart disease dataset for prediction of absence or presence of disease reveals that SA-SVM achieves an accuracy of 93.33% which is better than Grid search method that achieved an accuracy of 81.37%.

Ismail et al. [22] presented a classification approach called GA-SVM for lymph disease diagnosis in which genetic algorithm (GA) is used to reduce the number of features of the dataset from 18 features to 6 features. The experiments were performed with 10-fold cross validation. Different kernel functions were employed and for each function, performance was evaluated by measures like accuracy, sensitivity, area under curve (AUC), F-measure. The result indicates that GA-linear classifier achieved best results of 83.1% accuracy with 82.6% sensitivity, 82.7% F-measure and 84.9% AUC.

A system for prediction of two classes of heart disease by using Learning Vector Quantization (LVQ) neural network algorithm is proposed by Sonawane et al. [23]. LVQ is a supervised version of quantization algorithm based on competitive learning. In this system, the input to the neural network is Cleveland heart disease database having 13 attributes of disease. To reduce error, the hidden layer neurons can be varied. The output layer indicates the absence or presence of heart disease. The system performance is evaluated with different number of neurons in hidden layer and iterations. In this system, 85.55% of maximum accuracy is achieved for 18 neurons and 100 iterations.

Ismaeel et al. [24] suggest a new method called Extreme Learning Machine (ELMs) for prediction of heart disease. ELM is a fast learning algorithm and can obtain outstanding generalization results. This system does not separate the data into training and testing steps. The error is measured with varying number of neurons. It has been observed that the error value gets declined with the increase in the number of neurons. This approach uses all the data in the prediction of 5 classes representing stroke level from 0 to 4 and achieved about 80% accuracy.

Bhatia et al. [25] proposed a system based on SVM and integer-coded genetic algorithm. A Genetic algorithm is used as a feature extraction step by extracting top N best features and accuracy is taken as a fitness measure. The experiment was performed with different number of features, different kernel function, and its parameter values. The result depicts that RBF kernel function outperforms with an accuracy of 90.57% for two class problem and 72.55% for five class problem.

Soni et al. [26] designed a GUI based interface to predict the absence or presence of heart disease using Weighted Associative classifier (WAC) in which each attribute is assigned with different weights based on their capability of prediction. The results illustrated that WAC gives better results than existing Associative classifier with an accuracy of 81.51% for two class problem and 57.75% for five class problem. The reason for less accuracy achieved in the case of multi-class problem is that dataset is highly imbalanced.

Wiharto et al. [27] presented automatic diagnosis system for predicting 5 classes of heart disease which are healthy, sick low, sick medium, sick high and serious. The experiment was conducted with different multi-class algorithms of SVM, Naïve Bayesian, Multilayer Perceptron and Adaboost. The performance metrics used are recall, precision and F-measure. Binary tree multiclass approach of SVM achieved maximum accuracy of 61.86% as compared to other approaches. Due to the imbalanced dataset, classes' having less number of instances does not show good results.

Sunila et al. [28] proposed an improved Multilayer perceptron algorithm which works on multiple subsets of training set. The majority probability rule is used to combine the results from different subsets. The experiment is implemented with 10-fold cross validation on Cleveland, Switzerland and Hungarian datasets. The result shows that

proposed approach is better than MLP algorithm and has attained an accuracy of 82.8%.

Vadicherla et al. [29] suggested a sequential minimal optimization (SMO) technique of SVM for heart disease diagnosis system. The system is proposed for two classes. SMO helps in training of SVM by finding the optimal values of multipliers required during training phase. The result reveals that SMO shows good results even on large dataset and performance time is also improved.

Bascil et al. [30] presented a comparative analysis of methods used in the hepatitis disease diagnosis. The dataset comprises of 155 instances and 19 features. The system is applicable for classification of two classes that are die and live. The dataset is taken from UCI data repository. In this study, probabilistic neural network (PNN) was proposed using 10 fold cross validation technique. The LDA-ANFIS structure [31] obtained the best results followed by FS-FUZZY-AIRS [32]. The PNN approach can be used effectively in the prediction of hepatitis disease.

CHAPTER 4

PROBLEM STATEMENT

4.1 Problem Statement

Heart disease is the leading worldwide reason of death. The count of people dying every year from heart disease is increasing drastically. There is a need to detect and diagnose the disease at an early stage so that considerable life can be saved. The diagnosis process is a complicated process in which doctors make a decision with the help of their knowledge and the clinical data available which may lead to a wrong assumption due to the complex association among various factors. The work presented in this thesis is intended to automate the medical diagnosis process and develop a prediction system to detect the heart disease using machine learning with higher accuracy.

4.2 Gaps in Study

On the basis of literature survey, there are following gaps in the study of diagnosis process of heart disease:

1. Most of the work has been done on a two class problem of heart disease diagnosis with various machine learning techniques. In this study, work is done on the multiclass problem. The stroke level of heart disease is also considered. Stroke 0 represents the absence of heart disease. Stroke 1, 2, 3 and 4 correspond to the presence of disease with each higher stroke level depicting the severity of heart disease.
2. Decision trees are prone to overfitting of data and may not be able to generalize well due to the presence of noise in the training data. This problem can be solved by SVM. SVM are less prone to overfitting because of the presence of regularization parameter. SVM always output a unique solution as there are no local minima concerned.
3. Performance of SVM classifier model is optimized by finding the best values of SVM parameters that maximize accuracy using Particle Swarm Optimization (PSO) technique.

4.3 Aims and Objectives

The aims and objectives of our work are:

1. To develop a heart disease prediction system that is highly precise, efficient and useful in early diagnosis which lessens the patient mortality rate.
2. To predict the severity level of heart disease by considering 5 classes for classification so that proper subsequent treatment is provided.
3. To implement Support vector machine (SVM) for the accurate classification of heart disease and to apply Particle Swarm Optimization (PSO) technique to find the best values of parameters of SVM.
4. To measure the performance of diagnosis system and to compare our results with other approaches using performance metrics like accuracy, precision and recall.

4.4 Research Methodology

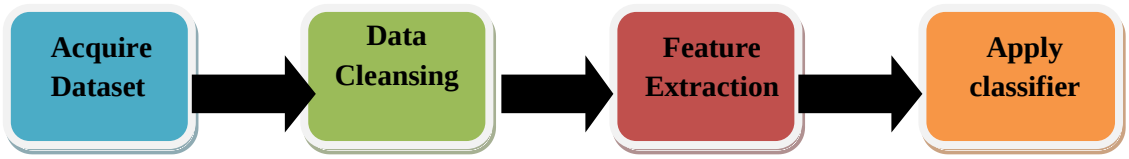

Figure 4.1 Research methodology used

Four phases are included in the proposed medical diagnosis process for early and precise detection of heart disease which are as follows:

Phase 1: In the first phase, the appropriate dataset related to the problem to work upon is selected. Proper selection of dataset is very important and make sure that an adequate number of records are present to address the problem. The dataset from a trustworthy source is collected and get familiarized with the data. The data is analyzed properly by understanding the meaning of each attribute present in the dataset.

Phase 2: In the second phase, clean the data and prepare it to give as input to the modeling tool. Data cleansing includes dealing with missing values, purging of redundant information, removing inconsistencies and errors which make the quality of data better and efficient to find useful patterns from the data. It includes all the steps to make the data complete for further preprocessing. It is a time-consuming step and very important step because the solution is highly affected by the quality of data.

Phase 3: In the third phase, a subset of features is selected which best reflects the original dataset. Each feature in the given dataset contributes differently. Some features are more significant to others while some features are irrelevant and add no useful information to the data which degrades the efficiency of the system. Moreover, high dimension of data results in more computation cost. So, there is a need to reduce the dimensions without affecting the quality of data. Dimensions of data can be reduced either by choosing the most significant features (Feature selection) or transforming features to a small set of features (Feature extraction).

Phase 4: In the fourth phase, a classifier is implemented to classify the data into their respective classes. Classification mainly includes two phases. The first phase is the training step and building classifier in which a classifier is trained to analyze the given data records and the class with which they are associated. It analyzes the pattern in the training set. The second phase is the testing step in which model classifies the test dataset on the basis of pattern analyzed in the first step. Different classifiers can be applied to the same dataset and best classifier can be selected by comparing the performance metrics like accuracy, specificity, robustness, speed, precision and recall.

CHAPTER 5

PROPOSED ALGORITHM AND IMPLEMENTATION

The multiclass model for diagnosing heart disease has been proposed using PSO based SVM that is highly efficient in terms of accuracy, precision and recall. It uses support vector machine classifier to classify heart disease dataset into 5 strokes. Stroke 0 represents the absence of heart disease. Stroke 1, 2, 3 and 4 correspond to the presence of disease with each higher stroke level depicting the severity of heart disease. The PSO is implemented to enhance the accuracy by finding optimal values of SVM parameters.

5.1 Major Components

The proposed algorithm mainly comprises of 4 major components which are as follows:

1. **SMOTE:** Synthetic Minority Oversampling Technique (SMOTE) is a popular algorithm to over-sample the minority class. SMOTE usually operates on continuous variables [33]. It generates synthetic examples of minority class by selecting a neighbor from k nearest neighbors of that minority class example. The difference between the attributes of minority class example and its neighbor is computed which is multiplied by an arbitrary number in the range [0, 1]. Synthetic example is created by the addition of computed value with minority class example considered. It results in the selection of synthetic example of minority class on the line segment between a minority class example and its neighbor [34].
2. **PCA:** Principal Component Analysis (PCA) is used as a preprocessing step to select the subset of features which best reflects the original heart dataset. Each feature in the given dataset contributes differently. Some features are more significant to others. The goal of PCA is to transform a number of correlated variables of a dataset to a new set of a small number of variables which are linear combinations of original variables called Principal Components [35]. The original dataset is replaced by its principal components after the application of PCA.
3. **PSO**: Particle Swarm Optimization (PSO) is a global optimization method used to find the optimal values of the error penalty parameter C and kernel

parameter of SVM which plays a vital role in improving the accuracy of SVM [36]. It searches for the best solution of parameters by executing a number of iterations.

4. **SVM:** Support Vector Machine (SVM) is used to predict the stroke level of heart disease. SVM solves the multiclass problem of 5 classes by dividing it into a series of two-class problems.

5.2 Working of proposed algorithm

The working of our proposed algorithm is shown in Figure 5.10. It comprises of 7 steps listed below to predict the severity level of heart disease.

5.2.1 Loading Heart Disease Dataset: The Cleveland Heart Dataset is used that is taken from UCI Machine Learning Dataset Repository which was contributed by Detrano [37]. The dataset comprises of 303 instances and 14 attributes of disease as shown in Figure 5.1. The dataset is divided into 5 classes represented by attribute 'num' which represents 5 different stroke level of disease. Increasing stroke level depicts the more severity of heart disease and immediate adequate treatment is recommended. The dataset comprises of 6 missing values which is neglected for this study. The total number of instances considered is 297.

The 5 classes of heart disease are as follows:

Class 0: Healthy

Class 1: Low-risk

Class 2: Medium- risk

Class 3: High-risk

Class 4: Danger

1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
2	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
3	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
4	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
5	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
6	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
7	56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
8	62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
9	57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
10	63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
11	53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
12	57	1	4	140	192	0	0	148	0	0.4	2	0	6	0

Figure 5.1: The Cleveland heart dataset

5.2.2 Balancing dataset using SMOTE: The Heart disease dataset is imbalanced as each class label is not approximately equally distributed. The number of instances of class 0 is 160, class 1 is 54, class 2 is 35, class 3 is 35 and class 4 is only 13 which is depicted in Figure 5.2. SMOTE algorithm is applied to the imbalanced dataset for reducing the biases among the majority classes. After the implementation of SMOTE, the number of instances of class 0 is 160, class 1 is 162, class 2 is 157, class 3 is 157 and class 4 is 156 which are approximately balanced as depicted in Figure 5.3.

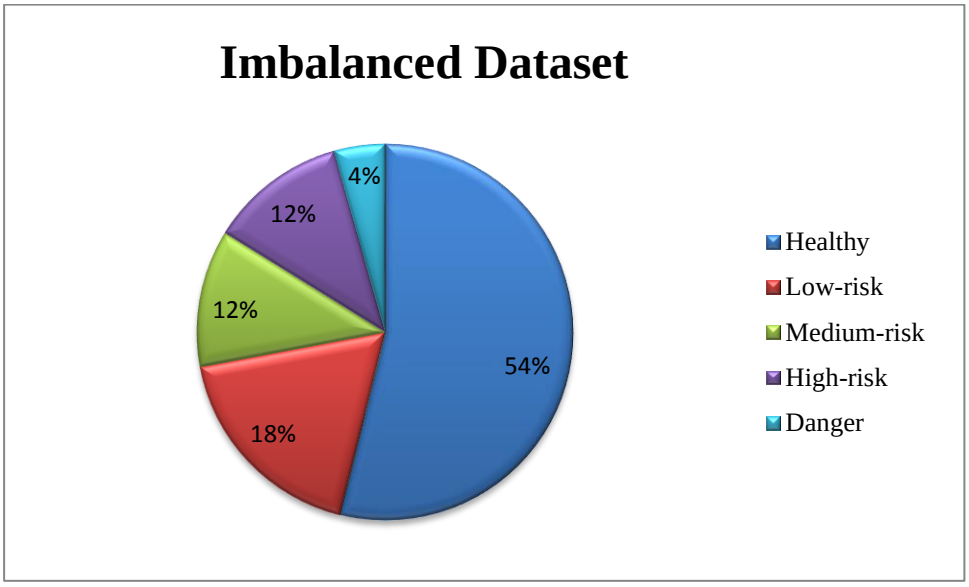

Figure 5.2: Imbalanced dataset

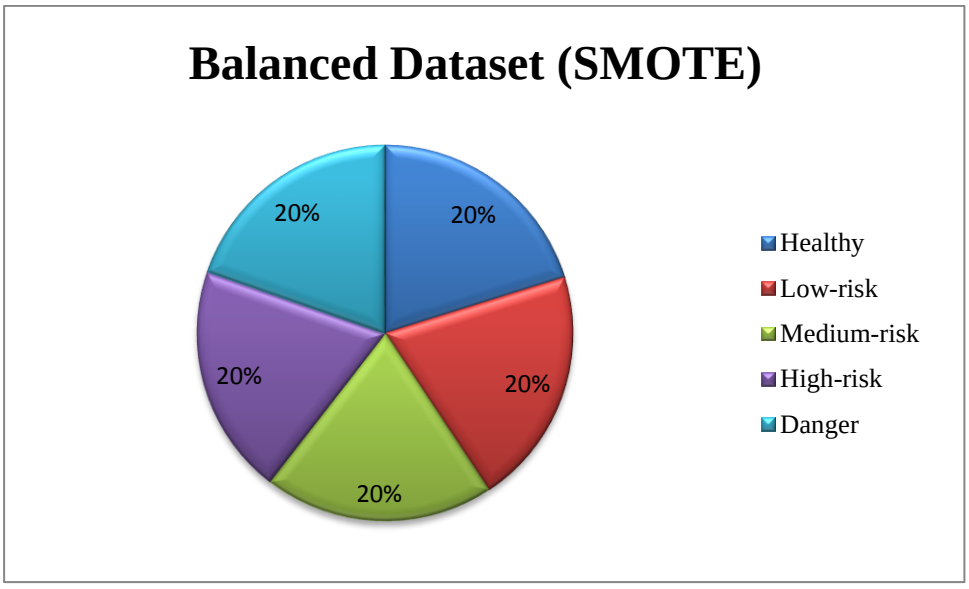

Figure 5.3: Balanced dataset

It generates synthetic examples of minority class by setting parameters nearestNeighbors and percentage depending upon the amount of oversampling required for a particular minority class. The value of nearest neighbors of the minority class sample considered is 5. The pseudo code for SMOTE algorithm is shown in Figure 5.4.

Algorithm: SMOTE (m, p, k, n)
Input: m- Number of minority class examples; p%- Percentage of SMOTE required; k- Number of nearest neighbors; n-Number of minority classes.
Output: Synthetic examples for each minority class.

```
1.  for  i ← 1 to n do      // For each minority class i
2.          if  p < 100 then
3.                  Randomize the m minority class samples
4.                  m = (p/100) × m
5.                  p = 100
6.          endif
7.          p = (int)(p/100)
8.          num_atrib : Number of attributes present in the dataset
9.          MinArr[ ][ ] : minority class instances array
10.         count=0         // stores the count of synthetic samples created
11.         SynthArr[ ][ ]: synthetic instances array
12.         For j←1 to m do//For each minority class example j in minority
        class i
13.                 NeighArr =Indices of k- nearest neighbors for j
14.                 while( p != 0)
15.                         rand=Select a number in the range [1,k]
16.                         for k ← 1 to num_atrib do
17.                           dif = MinArr [NeighArr [rand]][k] – MinArr [j][k]
18.                           interval = select a random no in the range  [0,1]
19.                           SynthArr [count][k] = MinArr[j][k] + interval * dif
20.                         Endfor
21.                         count=count +1
22.                         p = p – 1
23.                 endwhile
24.                 return  Synthetic examples
25.         endfor
26. Endfor
```

Figure 5.4: Pseudo code for SMOTE

5.2.3 Normalizing Dataset: The Cleveland Heart Dataset is dealing with the attributes having different units and scales. For example, the resting blood pressure (trestbps) is in mm Hg and ranges from 94 to 200 while the sex being 0 or 1, and the cholesterol is in mg/dl ranges from 126 to 564. To have the fair comparison, all parameters should have the same scale. Normalization makes the data scalable into a small specific numeric range. The dataset after normalization of values is shown in Figure 5.5.

If $x = (x_{1,} x_2, \ldots . x_k)$ are the data points, f(x) will normalize the dataset using the following equation

$$f(x_i\ ) = \frac{x_i - mean(x)}{var(x)}$$

where

x_i= Data point i where $1 \leq i \leq$ k

$mean(x)$= The average of all the data instances

$var(x)$ = The sample deviation of all the data instances

1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
2	0.865507	0.48665	-2.93127	0.715365	-0.32309	3.760461	0.973392	0.313984	-0.96297	0.725934	2.355669	-1.11433	0.293914
3	-2.3781	0.48665	-0.51932	-0.20725	0.016619	-0.26552	-1.25954	2.011625	-0.96297	1.769579	2.355669	-1.11433	-1.43974
4	-1.87909	-2.05171	-1.7253	-0.20725	-0.9026	-0.26552	0.973392	1.323392	-0.96297	-0.0568	-1.39851	-1.11433	-1.43974
5	-0.00777	0.48665	-1.7253	-0.82232	-0.26314	-0.26552	-1.25954	1.598685	-0.96297	-0.57862	-1.39851	-1.11433	-1.43974
6	0.116982	-2.05171	0.686656	-0.82232	2.094852	-0.26552	-1.25954	0.910452	1.222898	-0.75256	-1.39851	-1.11433	-1.43974
7	0.116982	0.48665	0.686656	0.407827	-1.1424	-0.26552	-1.25954	0.222219	-0.96297	-0.92651	0.478581	-1.11433	0.293914
8	-0.00777	-2.05171	-1.7253	0.407827	0.895871	-0.26552	0.973392	0.45163	-0.96297	-0.14377	0.478581	-1.11433	-1.43974
9	-1.50482	0.48665	-1.7253	-0.82232	0.276398	-0.26552	-1.25954	1.369274	-0.96297	-1.27439	-1.39851	-1.11433	0.871798
10	-0.50679	0.48665	-0.51932	2.376068	-1.00251	3.760461	-1.25954	0.86457	-0.96297	-0.83953	-1.39851	-1.11433	0.871798
11	0.116982	0.48665	-0.51932	1.022903	-1.62199	-0.26552	-1.25954	1.415156	-0.96297	0.11714	-1.39851	-1.11433	-1.43974
12	-0.25728	0.48665	0.686656	0.407827	-0.20319	-0.26552	-1.25954	0.772806	-0.96297	-0.23074	-1.39851	-1.11433	-1.43974

Figure 5.5: Normalized dataset

5.2.4 Selecting feature subset using PCA: Principal Component Analysis is used as a dimensionality reduction step to select the subset of features which best reflects the original heart dataset. The PCA is a statistical method which transforms a number of correlated variables of a dataset to a new set of a small number of variables which are linear combinations of original variables called Principal Components. The original dataset is replaced by its principal components after the implementation of PCA. The procedure for PCA is shown in Figure 5.7.

The contribution of each principal component is represented byR_p.

$$R_p = \frac{\lambda_p}{\sum_{i=1}^{n} \lambda_i}$$

where

R_p= The contribution of principal component p .

λ_p= The variance accounted for by component p.

n = The number of principal components.

The cumulative contribution rate is given by R_c as below:

$$R_c = \frac{\sum_{i=1}^{c} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

where

R_c = The cumulative contribution of principal component c .

λ_i = The variance accounted for by component i.

n = The number of principal components.

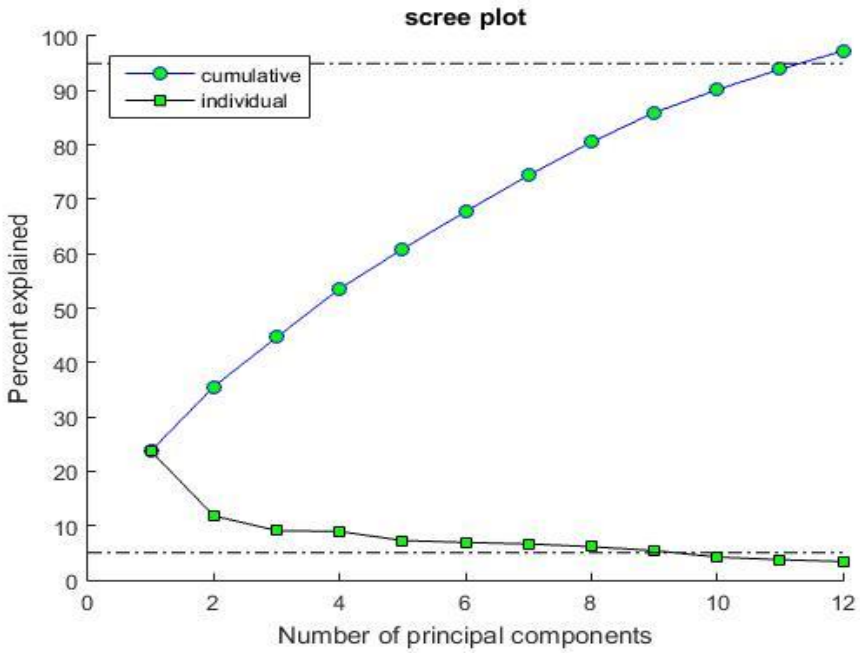

Figure 5.6: Scree Plot of principal components

As shown in Figure 5.6, the contribution of each principal component is depicted by the green squares on the line and cumulative contribution rate is depicted by the green circles on the line. Principal components are ordered according to their variance with

the largest variation at the first position and the contribution decreases with each succeeding principal component. Large variance reflects the more information of the data. Each succeeding principal component is orthogonal to its preceding principal component. With the increase in principal components, the cumulative contribution rate increases.

Algorithm for PCA

Input: The input data matrix X of size $N \times d$ where N is the number of instances and d is the number of dimensions.

Output: Principal Components

Procedure:
1. Calculate and subtract the mean in every dimension d of the dataset to centralize the data.

2. Construct the covariance matrix Cov of d*d as:

$$Cov = \frac{1}{N}\sum_{p=1}^{N}(x_p - \mu)\,(x_p - \mu)^T$$

where $\{x_p,\ p = 1,2, \ldots . N\}$ is given N input data records with mean μ.

3. Calculate the eigen values $(\lambda_1, \lambda_2, \ldots \lambda_d)$ and $(e_1, e_2, \ldots e_d)$ eigen vectors from the covariance matrix Cov such that

$$\lambda \times e = Cov \times e$$

4. Choose the m eigen vectors corresponding to m largest eigen values where $m \leq d$.

5. Compute the $d \times m$ dimensional matrix W from the above selected m eigen vectors where eigen vectors are represented by columns.

6. The original dataset X is transformed via W onto m-dimensional new subspace Y.

$$y = W^T \times x$$

where x is a $d \times 1$dimensional vector representing one data record and y is transformed $m \times 1$ dimensional vector representing data record in the new subspace Y.

Figure 5.7: Procedure for PCA

5.2.5 Selecting training and testing set: The 10-fold cross validation method is used to select the training and testing set. The dataset is divided into 10 equal parts out of

which 9 parts are used for training and the left over part is used for testing. This procedure is repeated 10 times in such a way that testing is performed on each part.

5.2.6 Finding the best parameters of SVM using PSO: The PSO is a global optimization technique inspired by the intelligence and movement of fish schooling or bird flocking. The error penalty parameter C and kernel parameter of SVM has a crucial role in improving the accuracy of SVM. If C is large, there is less final training error but if it exceeded beyond a certain limit, the generalization ability of the classifier may lose and encounter overfitting. If C is too small, we may come across underfitting. PSO is used to find optimal values of SVM parameters. It uses a group of swarm partles, each of which corresponds to a point in n-dimensional space and searches for the best solution by executing a number of iterations [38]. At each iteration, these particles get attracted towards the prime solution attained by them thus far ($lbest$) and the prime solution attained by any particles in entire space ($gbest$).

$$lbest_i = pos_i(x^*) \text{ s.t. } fitness\big(pos_i(x^*)\big) = \max_{x=1,\dots,n}[\,fitness\big(pos_i(x)\big)]$$

$$gbest = pos_{i^*}(x^*) \text{ s.t. } fitness\big(pos_{i^*}(x^*)\big) = \max_{\substack{i=1,\dots,P \\ x=1,\dots,n}}[\,fitness\big(pos_i(x)\big)]$$

where i denotes the particle number, pos denotes the particle position, x denotes the iteration index and n denotes the current iteration counter.

The particle modifies its position and velocity by the following equations

$$v_i(t+1) = wv_i(t) + c_1\, rand_1\big(lbest_i(t) - pos_i(t)\big) + c_2\, rand_2\big(gbest(t) - pos_i(t)\big)$$

$$pos_i(t+1) = pos_i(t) + v_i(t+1)$$

where v is the particle velocity, w is the initial weight used to control the local exploitation and global exploration abilities of swarm, $rand_1$ and $rand_2$ are two random numbers in the range [0,1] c_1 and c_2 are two positive constants. The pseudo code for PSO algorithm is shown in Figure 5.8.

5.2.7 Training SVM classifier: Support Vector Machine is implemented on training data with the parameters returned by the PSO. For multiclass classifier, our proposed method uses one against one method by constructing k (k-1)/2 classifiers for each possible pair of classes for k class problem, each classifier is trained on data to separate the samples of two classes from each other. To get multiclass classification, SVM applies each of the k(k-1)/2 classifier and predict the class label by each one of

them. The final class is the one getting the maximum votes by the classifiers. The radial basis kernel function is used to implement the data.

Algorithm: PSO(p, m)
Input: p- Number of particles; m- Maximum iteration allowed
Output: $gbest$- the position of a particle having global best value.

1. for particle ← 1 to p do // *For each particle p*
2. Initialize particle with random position $pos_{particle}$ and velocity $v_{particle}$ vector.
3. $lbest_{particle} = pos_{particle}$
4. endfor
5. do
6. for particle ← 1 to p do
7. $fitness_{particle}$= Fitness value of particle
8. if $fitness_{particle} > fitness(lbest)$
9. $lbest_{particle} = pos_{particle}$
10. endif
11. endfor
12. i← index of the particle having best f$itness(lbest_{particle})$ value.
13. $gbest = lbest_i$
14. for particle ← 1 to p do
15. Update $v_{particle}$ using equation
16. Update $pos_{particle}$ using equation
17. endfor
18. While (max iteration or minimum error reached)
19. return $gbest$

Figure 5.8: Pseudo code for PSO

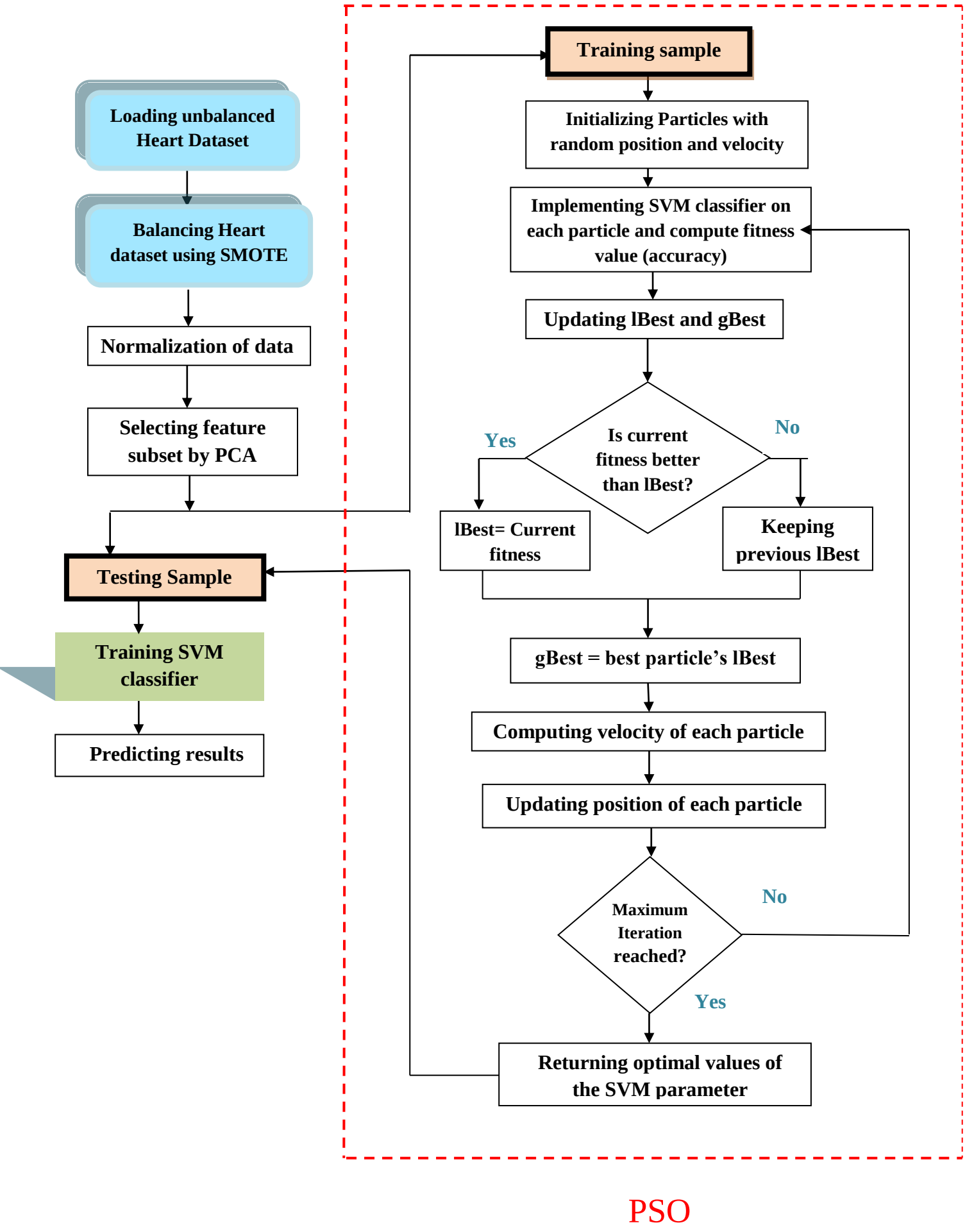

Figure 5.9: Flow chart of proposed algorithm

CHAPTER 6

SIMULATION RESULTS AND DISCUSSIONS

The experimental study is implemented by using MATLAB simulator. The Cleveland Heart dataset is used that consists of 14 attributes and 303 instances. Out of 303 instances, 6 values are missing which are removed for this study. The dataset is imbalanced and is being divided into five unequally distributed classes which are healthy, low risk, medium risk, high risk and danger. The SMOTE algorithm is applied to balance the dataset. The SVM classifier with radial basis function is implemented using K-fold cross validation. The performance of proposed method is improved by implemented PSO technique.

6.1 Performance metrics

The performance of the system is evaluated in terms of accuracy, precision and recall using the below parameters.

TN_i= True negative for class i

TP_i= True positive for class i

FP_i= Fasle positive for class i

FN_i= Flase negative for class i

N= Number of classes

M= Macro-averaging

1. **Accuracy**: It is the percentage of correctly predicted samples to the total number of samples.

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{N}$$

2. **Precision**: It is the fraction of retrieved samples that are predicted correctly.

$$\text{Precision}_M = \frac{\sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i}}{N}$$

3. **Recall**: It is the percentage of correctly predicted samples that are retrieved.

$$\text{Recall}_M = \frac{\sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i}}{N}$$

4. **F-score**: It is the weighted average of precision and recall.

$$\text{Fscore}_{\text{M}} = \frac{(\beta^2 + 1)\text{Precision}_{\text{M}}\text{Recall}_{\text{M}}}{\beta^2\ \text{Precision}_{\text{M}} + \ \text{Recall}_{\text{M}}}$$

6.2 Simulation results

The results of our proposed method are:

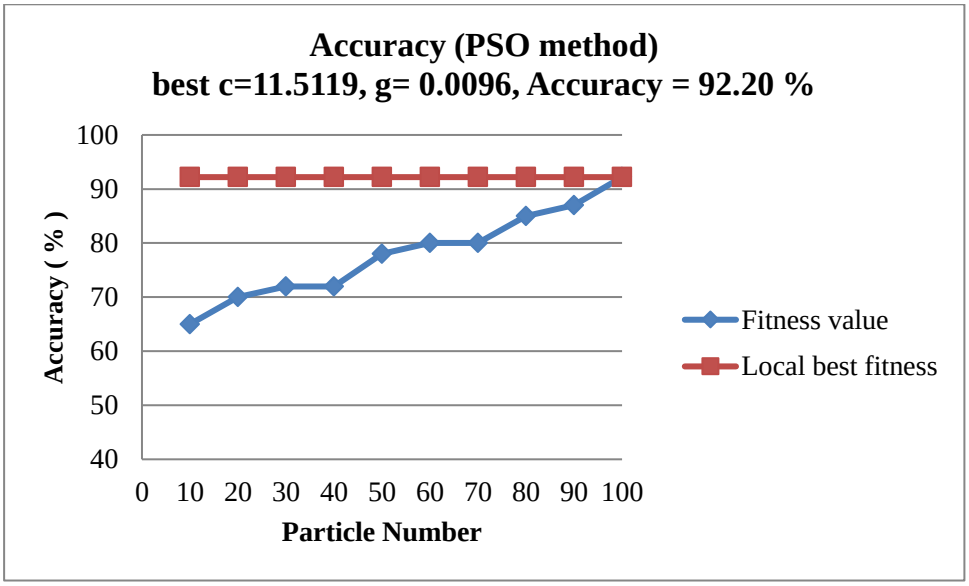

Figure 6.1: Values of SVM parameters by PSO in first fold

The PSO technique finds the best solution of SVM parameters by executing a number of iterations. At each iteration, it keeps track of the prime solution attained by each particle thus far (lbest) and the prime solution attained by any particle in entire space (gbest). The optimal values of c and g are found to be 11.5119 and 0.0096 respectively by executing 100 iterations for first fold of dataset. In Figure 6.1, Blue curve represents the accuracy achieved by a particle and red line depicts the best local accuracy achieved by a particle**.** Table 6.1 represents the initial parameters of algorithm. Table 6.2 represents the value of parameters c and g achieved in each fold.

Table 6.1: PSO parameters

Parameter	Value
Population size	100
Number of iterations	100
c1	1.5
c2	1.7
K	0.6
wV	1
Cmin	0.001
Cmax	300
Gmin	0.001
Gmax	300

Table 6.2: Value of c and g for 10-fold

Number of Fold	Value of c	Value of g	Accuracy(%)
1	11.5119	0.0096	92.20
2	4.1231	0.0149	87.01
3	100	0.0451	84.523
4	100	0.0202	87.50
5	10.2490	0.0122	86.25
6	12.0711	0.0178	78.31
7	89.2445	0.0246	86.075
8	12.4205	0.0422	85.36
9	100	0.0239	90.66
10	56.1783	0.0194	90.90

The class labels of heart dataset can be visualized in Figure 6.2. Different colors are assigned to each class. Multidimensional scaling [39] is used to plot the classes in such a way that the between class distance is conserved as much as possible. The class labels along with decision boundaries are shown in Figure 6.3.

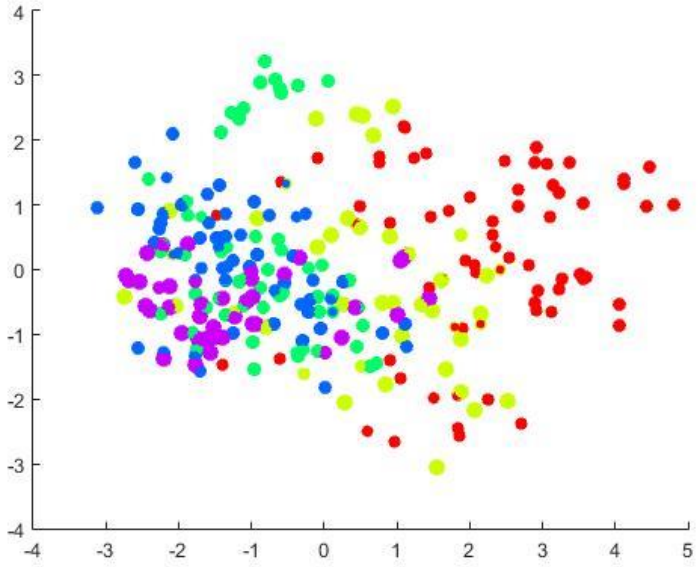

Figure 6.2: Visualization of data

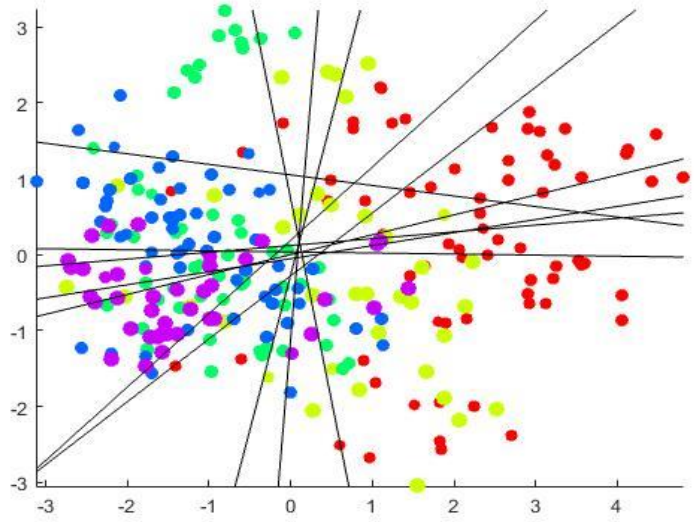

Figure 6.3: Decision boundaries of SVM

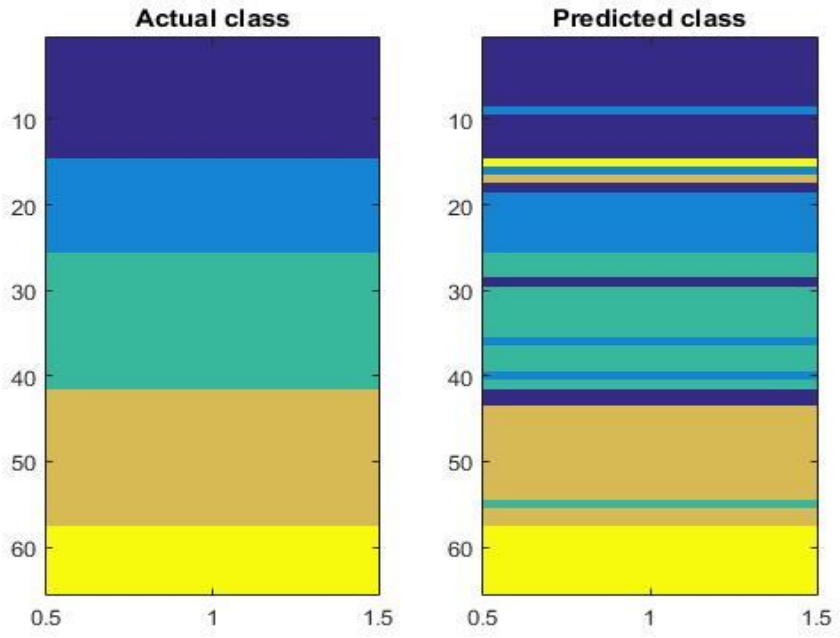

Figure 6.4: Comparison of actual and predicted class of proposed algorithm

The results of proposed algorithm can be visualized in Figure 6.4. Five colors in rectangular box above represent classes 0 to 5 from bottom to top. Left and right rectangular box represent actual and predicted class respectively. The error in the final results is depicted by the change in the color in right rectangular box.

The precision, recall for five classes and overall accuracy is shown in Figure 6.5.

		Truth data						
		Class 1	Class 2	Class 3	Class 4	Class 5	Classification overall	Producer Accuracy (Precision)
Classifier results	Class 1	144	13	5	2	1	165	87.273%
	Class 2	9	131	10	9	2	161	81.366%
	Class 3	3	9	133	10	3	158	84.177%
	Class 4	4	6	7	132	5	154	85.714%
	Class 5	0	3	2	4	145	154	94.156%
	Truth overall	160	162	157	157	156	792	
	User Accuracy (Recall)	90%	80.864%	84.713%	84.076%	92.949%		

Overall accuracy (OA): 86.49%

Figure 6.5: Confusion matrix for 5-class problem

The overall accuracy of proposed model is 86.49%

The F-score of Class 0= 88.615%

Class 1= 81.114%

Class 2= 84.444%

Class 3= 84.888%

Class 4= 93.548%

6.3 Comparison of results

Table 6.3 summarizes the comparative analysis of proposes algorithm with different existing approaches. The result shows that proposed method achieved highest accuracy than other approaches for classification of 2-class problem as well as 5-class problem. This shows that our proposed algorithm can be successfully used for the diagnosis of heart disease.

Table 6.3: Comparison of classification accuracy

Techniques	**Accuracy(2-class)**	**Accuracy(5-class)**
PCA1-regression [21]	92.0%	_
Grid search SVM [22]	81.37%	_
SA-SVM [22]	93.33%	_
GA and 5-nearest neighbor [20]	81.4%	_
LVQ [24]	85.55%	_
SVM-integer coded GA [26]	90.57%	72.55%
Weighted associative classifier [27]	81.51%	57.75%
Binary tree-SVM [28]	_	61.86%
Extreme learning Machine [25]	_	80%
Proposed Method	95.29%	86.49%

The comparative analysis of classification accuracy of proposed algorithm is done with existing algorithms. In 2-class problem, proposed method predicts the absence or presence of heart disease and achieves highest accuracy of 95.29% as shown in Figure 6.6. In 5-class problem, it classifies the heart dataset into stroke levels and achieves an accuracy of 86.25% which is promising than other methods as shown in Figure 6.7

6.3.1 Comparison of 2-class problem

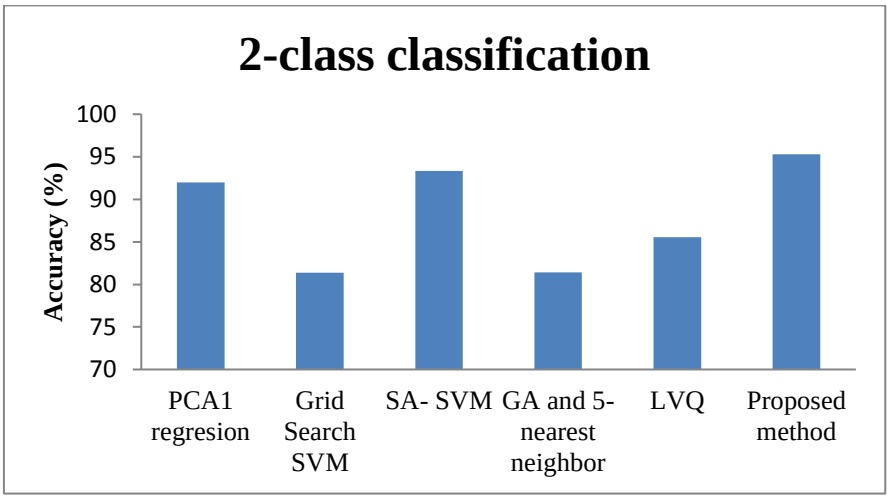

Figure 6.6: Classification accuracy comparison for 2-class problem

6.3.2 Comparison of 5-class problem

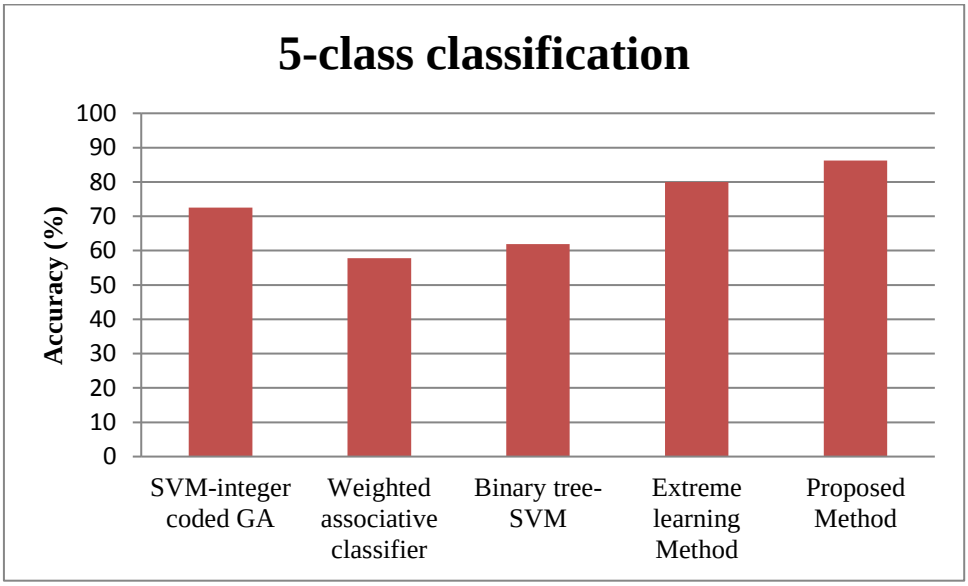

Figure 6.7: Classification accuracy comparison for 5-class problem

CHAPTER 7

CONCLUSION AND FUTURE SCOPE

7.1 Conclusion

The multiclass model for diagnosing heart disease has been proposed using PSO based SVM. Our proposed algorithm classifies the data into five classes namely healthy, low-risk, medium-risk, high-risk and danger. Each class depicts the severity level of heart disease in the increasing order. The subset of features that best reflects the original dataset are extracted by means of PCA. The support vector machine has been used to implement the diagnosis system and parameters of SVM are optimized using Particle swarm optimization technique. The experimental result of proposed algorithm shows that RBF kernel gives the best result. The accuracy achieved for 2-class problem is 95.29% and for 5-class problem is 86.49% which is better than other existing approaches. The results obtained prove that the proposed algorithm can be successfully used for the determination of stroke level of disease.

7.2 Future scope

Future work involves optimization of SVM parameters with other methods such as scatter search method, Cuckoo search etc and comparing results with our proposed algorithm. The multiclass problem in our work is solved by one against one approach. The performance of system will be evaluated with other multi class algorithms of SVM like one against all and error correcting output code. The optimization techniques for other kernel function of SVM will be searched so that they can also be effectively used for the heart disease diagnosis.

REFERENCES

[1] S. Goenka et al. ,"Preventing cardiovascular disease in India-translating evidence to action,'' *Current science*, vol. 97, no. 3, pp. 367– 377, 2009.

[2] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," *Proceedings of the 23rd international conference on Machine learning*, pp. 161–168, Pittsburgh, ACM 2006.

[3] Z. Ghahramani, "Unsupervised Learning BT - Advanced Lectures on Machine Learning," *Advanced lectures on machine learning*, vol. 3176, no. 5, pp. 72–112, Springer 2004.

[4] L.P. Kaelbling, M.L. Littman and A.W. Moore, "Reinforcement Learning: A Survey," *Journal of artificial intelligence research*, vol. 4, pp. 237– 285, 1996.

[5] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," *Computing, Communications and Networking Technologies 2013 Fourth International Conference (ICCCNT)*, pp. 1–7, IEEE, 2013.

[6] D.S. Medhekar, M.P. Bote, and S. D. Deshmukh, "Heart Disease Prediction System using Naive Bayes," *International Journal of Enhanced Research in Science Technology and Engineering*, vol. 2, no. 3, Elsevier 2013.

[7] D. Lowd and P. Domingos, "Naive Bayes models for probability estimation," *Proceedings of the 22nd international conference on Machine learning*, pp. 529–536, ACM 2005.

[8] A. Kataria and M. D. Singh, "A Review of Data Classification Using K-Nearest Neighbour Algorithm," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 6, pp. 354–360, 2013.

[9] T. Cover and P. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions*, vol. 13, no. 1, pp. 21–27, IEEE 1967.

[10] L. M. Silva, J. Marques de Sa, and L. a. Alexandre, "Data classification with multilayer perceptrons using a generalized error function," *Neural Networks*, vol. 21, no. 9, pp. 1302–1310, 2008.

[11] S. K. Pal and S. Mitra, "Multilayer Perceptron, Fuzzy Sets, and Classification," *Neural Networks, IEEE Transactions*, vol. 3, no. 5, pp. 683–697, 1992.

[12] W. Du and Z. Zhan, "Building decision tree classifier on private data," *Proceedings of the IEEE international conference on Privacy, security and data mining*, vol. 14, pp. 1–8, Australia 2002.

[13] J.R. Quinlan, "Induction of Decision Trees," *Expert System*, vol. 1, no. 1, pp. 81–106, 2007.

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[15] S. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks*, vol. 12, no. 6, pp. 783–789, 1999.

[16] V. D. Sanchez A, "Advanced support vector machines and kernel methods," *Neurocomputing*, vol. 55, no. 1, pp. 5–20, Elsevier 2003.

[17] Y. Zhang et al. , "Studies on application of Support Vector Machine in diagnose of coronary heart disease," *Electromagnetic Field Problems and Applications 2012 Sixth International Conference (ICEF)*, Dalian, IEEE 2012.

[18] M. Naib, "Predicting Primary Tumors using Multiclass Classifier Approach of Data Mining," *International Journal of Computer Applications* ,vol. 96, no. 8, pp. 9–13, 2014.

[19] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," *Procedia Technoogy.*, vol. 10, pp. 85–94, Elsevier 2013

[20] T. Santhanam and E. P. Ephzibah, "Heart Disease Classification Using PCA and Feed Forward Neural Networks," *Mining Intelligence and Knowledge Exploration*, pp. 90–99, Switzerland, Springer 2013.

[21] S.-W. Lin, et al. , "Parameter determination of support vector machine and feature selection using simulated annealing approach," *Applied soft computing*, vol. 8, no. 4, pp. 1505–1512, 2008.

[22] H. I. Elshazly, A. M. Elkorany, and A. E. Hassanien, "Lymph diseases diagnosis approach based on support vector machines with different kernel functions," *Computer Engineering & Systems 9th International Conference (ICCES)*, Cairo, pp. 198–203, 2014.

[23] J. S. Sonawane and D.Patil, "Prediction of Heart Disease Using Learning Vector Quantization Algorithm*," In IT in Business, Industry and Government Conference on IEEE (CSIBIG)*, Indore, pp. 1–5, 2014.

[24] S. Ismaeel, A. Miri, and D. Chourishi, "Using the Extreme Learning Machine Technique for Heart Disease Diagnosis," *In Humanitarian Technology Conference Canada International(IHTC2015)* , no. 1, pp. 1–3, Canada , IEEE 2015.

[25] S. Bhatia, P. Prakash, and G. N. Pillai, "SVM Based Decision Support System for Heart Disease Classification with Integer-Coded Genetic Algorithm to Select Critical Features," In *Proceedings of the World Congress on Engineering and Computer Science (WCECS)* ,pp. 22-24, 2008.

[26] J. Soni, U. Ansari, and D. Sharma, "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers," *International Journal on Computer Science and Engineering*, vol. 3, no. 6, pp. 2385–2392, 2011.

[27] Wiharto, H. Kusnanto, and Herianto, "Performance Analysis of Multiclass Support Vector Machine Classification for Diagnosis of Coronary Heart Diseases," *International Journal on Computational Science & Applications*, vol. 5, no. 5, pp. 27–37, 2015.

[28] P. Panday and N. Godara, "Decision Support System for Cardiovascular Heart Disease Diagnosis using Improved Multilayer Perceptron," *International Journal of Computer Applications* ,vol. 45, no. 8, pp. 12–20, 2012.

[29] E. Technologies, D. Vadicherla, and S. Sonawane, "Decision Support System for Heart Disease Based on Sequential Minimal Optimization in Support," *International Journal of Engineering Sciences and Emerging Technologies*, vol. 4, no. 2, pp. 19–26, 2013.

[30] M. S. Bascil and H. Oztekin, “A study on hepatitis disease diagnosis using probabilistic neural network,” *Journal of medical systems*, vol. 36, no. 3, pp. 1603–1606, 2012.

[31] E. Dogantekin, A. Dogantekin, and D. Avci, “Automatic hepatitis diagnosis system based on Linear Discriminant Analysis and Adaptive Network based on Fuzzy Inference System,” *Expert Systems with Appications*, vol. 36, no. 8, pp. 11282–11286, Elsevier 2009.

[32] K. Polat and S. Güneş, “Medical decision support system based on artificial immune recognition immune system (AIRS), fuzzy weighted pre-processing and feature selection,” *Expert Systems with Appications*, vol. 33, no. 2, pp. 484–490, Elsevier 2007.

[33] N. Chawla and K. Bowyer, “SMOTE: Synthetic Minority Over-sampling Technique Nitesh,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[34] W. Juanjuan et al. , “Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding,”, *8th International Conference on Signal Processing*, Vol. 3, Beijing , IEEE 2006.

[35] A. Ilin and T. Raiko, “Practical approaches to principal component analysis in the presence of missing values,” *Journal of Machine Learning Research*, vol. 11, pp. 1957–2000, 2010.

[36] J. Kennedy, “*Encyclopedia of machine learning*, vol. 46, no. 1, pp. 760–766, US, springer 2011.

[37] “Cleveland dataset.” [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data. [Accessed: Jun. 7, 2016].

[38] D. P. Rini, S. M. Shamsuddin, and S. S. Yuhaniz, “Particle Swarm Optimization: Technique, System and Challenges,” *International Journal of Computer Applications*, vol. 14, no. 1, pp. 19–27, 2011.

[39] A. Buja et al. , “Data Visualization with Multidimensional Scaling,” *Journal of Computational and Graphical Statistics*, vol. 17, no. 2, pp. 444–472, 2008.

LIST OF PUBLICATIONS

- P. Dembla and T. Bhatia, "PCA-SVM-PSO Decision Support System for Enhanced Prediction of Heart Disease [Communicated]
- P. Dembla and T. Bhatia, "Multiclass Diagnosis Model for Heart Disease using PSO based SVM[Communicated]

VIDEO LINK

https://www.youtube.com/channel/UCbyzUiAGN4PAAxVewM6-3ig