HPCC: An Ensembled Framework for the Prediction of the onset of Diabetes

Thesis submitted in partial fulfilment of the requirements for the award of degree of

Master of Engineering

In

Computer Science & Engineering

Submitted By

Harnoor Kaur

(801532020)

Under the supervision of:

Dr. Shalini Batra

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

July 2017

Certificate

I hereby certify that the matter which is being presented in the seminar report titled, "**HPCC: An ensemble Framework for the Prediction of the onset of Diabetes**", in partial fulfilment 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 a survey carried out by me, under the supervision of ***Dr. Shalini Batra*** and refers others researcher's work which is duly listed in the reference section.

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

Harnoor Kaur

801532020

ME (CSE)

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

Dr. Shalini Batra

Associate Professor

Department of CSE

Thapar University

Patiala

Acknowledgement

This research work would be incomplete without acknowledging the people who supported and guided me for the successful completion of this task.

First of all I wish to acknowledge the benevolence of God who gave me courage and strength to face the challenges and to overcome the obstacles that occurred while working on this task.

This research work is the outcome of the excellent guidance and positive attitude of my guide

Dr. Shalini Batra, Associate Professor, Computer Science and Engineering Department, Thapar University. I am very thankful to them for their valuable guidance, constant encouragement, keen interest, invaluable co-operation, generous attitude and above all his blessings. They have been a source of inspiration and motivation for me.

I am grateful to **Dr. Maninder Singh**, Head, Computer Science and Engineering Department, Thapar University and **Dr. Ashutosh Mishra**, PG Coordinator, for the motivation and inspiration for the completion of this thesis.

I am thankful to my parents and my friends who encouraged and supported me with their whole-hearted co-operation for completing this thesis.

Harnoor Kaur

(801532020)

Abstract

Diabetes being one of the chronic diseases worldwide needs proper diagnosis and treatment since this is on the spread and is on the way of becoming the main cause of many other medical disorders. The advent of such a disease must be nipped in the bud if a person is found to be prone to it. Such an experiment has been done in this study which tells about the onset of diabetes in the females of Pima Indian origin of Arizona. Diagnosing Diabetes is one of the problems that require high level of accurate analysis and prediction. Data scientists have attempted several data analytics methods in order to improvise the examination of data sets. Previously, various data mining techniques have been implemented in the healthcare systems, however, the hybridization in addition to single technique in the identification of the disease shows promising outcomes, and can be useful in further investigating its treatment and can help in reducing the cost if the treatment. Traditional techniques which are used for clinical decision support systems are grounded on a single classifier or combination of various classifiers which are used for the diagnosis of the disease and its prediction. Recently much heed has been paid to improve the performance of disease prediction with the use of ensemble-based methods. Using ensemble methods in decision support systems assist in analyzing theses type of diseases more effectively. To improve the performance of weak classifiers boosting and bagging techniques can be used. These techniques are based on combining the outputs and functionality of the various classifiers used. A weighted majority vote or a simple majority vote which has been used in this study are the most common rules for the implementation of bagging and boosting. In this paper, we compare the performance of bagging and boosting with our hybrid approach called Hierarchical and Progressive Combination of Classifiers (HPCC) through the study of the famous Pima Indians Diabetes Dataset and the best classifier is chosen on the basis of the accuracy achieved.

Keywords—Data Mining, Heart Disease, Classification, Bagging, Boosting

Table of Contents

List of Figures

List of Tables

List of abbreviations

DM	Diabetes Mellitus
IDF	International Diabetes Federation
WHO	World Health Organization
IDDM	Insulin-Dependent Diabetes Mellitus
NIDDM	Non–Insulin Dependent Diabetes Mellitus
FPG	Fasting Plasma Glucose
OGTT	Oral Glucose Tolerance Test
NDIC	The National Diabetes Information Clearing House
KDD	Knowledge Discovery in Databases
AI	Artificial Intelligence
ML	Machine Learning
SVM	Support Vector Machine
UCI	University of California
KNN	K-Nearest Neighbors
CART	Classification and Regression Trees
LDA	Linear Discriminant Analysis
MWSVM	Morlet Wavelet SVM classifier
FNN	Fuzzy Neural Network
ANN	Artificial Neural Network
GRNN	General Regression Neural Network
BMI	Body Mass Index
MV	Majority Voting
PV	Plurality Voting
GLM	Generalized Linear Model
IG	Information Gain
PLS	Partial Least Squares

OWNN	Optimal Weighted Neural Network
HPCC	Hierarchical and Progressive Combination of Classifiers
GBM	Generalized Boosted Regression Models
GAMBoost	Generalized Additive Model on the basis of likelihood Boosting
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives

CHAPTER 1

INTRODUCTION

Diabetes mellitus (DM), popularly known as diabetes, is a very chronic disease and comes in the category of intensely increasing metabolic diseases of the world. It is a disease in which the body does not produce or use insulin properly, the hormone that ‘‘unlocks” the cells of the body, allowing glucose to enter and fuel them. It is linked with an abnormal hike in the level of glucose (hyperglycemia) in the blood of a person, occurring either due to an insufficient production of insulin by the gland which is responsible for this i.e. pancreas (classified as Type 1 diabetes) or the failure of cells in an effective response to the insulin produced by gland pancreas (categorized as Type 2 diabetes)[1].It is a non-communicable disease and it is being very closely observed by the International Diabetes Federation (IDF) and the World Health Organization, It is a major health problem in both industrial and developing countries, and its number is rising. (WHO) because cases of diabetes are increasing continuously worldwide at an alarming rate [2].

Considering diabetes death rate, there are approximately 12 deaths due to diabetes in every 1 lac. It can be assumed that this is a quite small proportion but still this fraction is only the deaths of people due to diabetes. Actually, diabetes mellitus is a major cause of many other diseases like cardiovascular strokes and heart disorders which are at the top of the long-lasting non-communicable diseases having high mortality rates [1]. It may also cause the damage of cells in the body and organs such as blood vessels, kidneys, nerves, eyes and heart may get severely damaged and impaired for lifetime.

Diabetes used to be most prevalent in adults and was once called “adult-onset” diabetes [1]. It is also alarming that the age of patients suffering from it is tending to decrease at a high rate. Moreover, females are more likely than males to get into the grab of diabetes and the percentage of obese patients is higher than the non-obese [2]. While its causes are not completely understood by now, scientists have a belief that both environmental triggers and genetic factors are involved in its inception.Going by the reports of the International Diabetes Federation, expenditures

for the health care of the diabetes patients are expected to be $490 billion for the year 2030, which accounts for 11.6% of the total health care expenditures on all the other diseases in the world [1]. Moreover, diabetes is possibly an independent factor of risk contributing to lot of microvascular complications. Its patients are more prone to microvascular impairment thereby exposing them to cardio vascular diseases 40% more than the non-diabetic individuals. These microvascular impairments and subsequent cardiovascular diseases ultimately lead the patient to nephropathy, retinopathy and neuropathy. Studies have shown that the age of the people with diabetes may get cut short by as much as 15 years [1].

Computational intelligence has always strived to play a very important part in the areas of decision making and health diagnosis. Medical diagnosis processes can be well characterized by the use of intelligent computational classification tasks [2]. Data mining is the process of analyzing huge amount of data and characterizing it into useful information. It plays an important role for disease diagnosis and prediction in medical domain. In particular, *classification algorithms* are quite helpful in categorizing the data important for the process of decision making by medical experts. Further to improve the accuracy of the classifier various data pre-processing methods and ensemble approaches have been proposed. The objective of this study is to promote and encourage good health of people. People who are found to be prone to diabetes should consult a specialist as soon as possible for the formal diagnosis of the symptoms to prevent themselves from being the victim of serious diabetes. It also considerably reduces the accompanied cost by sidestepping unwanted and expensive medical tests to be carried out.

1.1 History

For about 2,000 years, diabetes had been known to be a distressing and fatal disease. In the first century A.D., a physician from Greece named Aretaeus, termed the disparaging nature of the disease, as "diabetes". In the 17^{th} century later a physician from London named Dr. Thomas Willis, examined his patients for diabetes by checking and testing that if the urine sample had a sweet taste he used to diagnose them with "honeyed" diabetes which were technically known as diabetes mellitus.

But this way of analyzing and monitoring blood sugars had gone essentially untouched or unchanged until the advent of the 20th century.

Before the discovery of the drug named insulin which is really necessary for the control and easy digestion of glucose, nothing could be done for the patients who were suffering from this disease except they could be given only natural remedies for its cure which could lower the sugar level in their body. Diets with very low calories persisted but their bodies were left weak and because of this they would always remain starving the whole day. But then in the year 1921, doctors from Toronto started treating patients who were on their death-beds because of diabetes with the drug called insulin which was able to lower the sugar levels in their bodies drastically to normal levels. And since then insulin has found its ways into the medical market for the cure and successful treatment of diabetes mellitus.

It has been approximately two thousand years that Aretaeus remarked diabetes as "the mysterious sickness". It was a long and tiring process which lead to its discovery, as generations of scientists and physicians have added their combined strength and knowledge in finding a cure for this disease. It was from this wealth of knowledge and the hard work put in that the discovery of insulin emerged to be a remarkable discovery in a small lab of Canada. And since that invention, medical discoveries and innovations have made life easier for people suffering from diabetes.

At the advent of 21st century, many researchers of diabetes continued to pave the road of progress towards a cure against it. Today, no one can tell that what shape the road will take in the near future; maybe yet another histrionic discovery like insulin be waiting around the corner, or it may even be possible that the researchers will have to stay content with the slow evolution of progress [3].

1.2 Types of diabetes mellitus

There are two main categories of diabetes: First one being the IDDM also called the Type I diabetes and second one being the NIDDM also known as the Type II diabetes. One more type of diabetes called the gestational diabetes or the diabetes during pregnancy have also been added to the list. This condition of sugar-level increase in the body can also be divided on the basis of insulin dependency as the categories can be “non-insulin-dependent diabetes mellitus” and “insulin-dependent

diabetes mellitus" [4]. But this categorization is of no use now as the patients now are divided on the basis of treatment they receive and not according to the pathogenesis [5].

1.2.1 IDDM or Type I Diabetes

This type of diabetes in only found in 5—10% of diabetes cases; but still its prevalence is on an increase across the world with its short-term and long-term repercussions. This type depicts the beta-cell destruction which is carried out in the gland called pancreas that eventually leads to the occurrence of the condition called diabetes mellitus during which "insulin is the prerequisite for the survival of the human". And this treatment of diabetes became mandatory to prevent the advent of the condition called ketoacidosis, which ultimately leads to coma and then death. Management and treatment of Type I diabetes best works in the environment where the patient is been taken care of by a diverse health team as it needs continuous supervision with respect to many aspects, which include administration of insulin, monitoring of blood glucose, undertaking the meals according to meal plans, and the most important being the screening which is required for complications related to diabetes [6].

1.2.2 NIDDM or Type II Diabetes

This is the most susceptible form of diabetes. People suffering from this kind of diabetes are at a much greater risk of being prone to cardiovascular disorders such as stroke and heart attack in case the diagnosis of diabetes is handled carelessly and the treatment is not taken seriously. They are also at a high risk of foot and leg amputation which occurs due to the damage of blood vessels and nerves of the body, loss of sight and renal failure which require transplantation or dialysis on a regular basis for its treatment [7]. Before the patient is diagnosed with type II diabetes, he or she is almost firstly a prey to "prediabetes" which is a condition where the levels of blood glucose go beyond the normal levels but are not high enough to be diagnosed as diabetes. Recent surveys and researches have shown that a diabetic person can also sometimes encounter life-long damages to the body which includes

breakdown of the circulatory system and severe damage to the proper functioning of heart [8]. In the case of Type II diabetes, either the patient's body is unable to produce insulin in enough amount which can control the blood glucose levels as insulin is necessary to extract energy from glucose or the insulin so produced by gland pancreas is not acceptable by the cells of the body. After the consumption of food, body breaks all the consumed sugar and starch into glucose which acts the energy fuel for the cells of the body. With rise in the levels of glucose which starts to build up in the blood instead of being used by the cells, the condition of diabetes and its complications encountered by the person's body.

1.2.3 Diabetes in Pregnancy (Gestational Diabetes)

Gestational diabetes also known as Pregnancy diabetes in layman language is the diabetes which is diagnosed in a pregnant women. Ladies who are generally overweight, have obesity in their genes, medical history of diabetes in their family or have suffered from pregnancy diabetes earlier are more prone to this kind of medical condition. If gestational diabetes is left undiagnosed and not treated on time, chances are that the child birth may face some complications. It makes both the baby and the mother prone to Type II diabetes for their entire lives and the child may suffer from diabetes at a very early age itself [9].

1.3 Diagnosis

The diagnosis of diabetes mellitus is rather easy as and when the patient starts showing the symptoms of hyperglycaemia and his body has a blood glucose value of 200 mg/dL that is 11.1 mmol/L or greater than this.

Tests used for the basic diagnosis of diabetes have been explained as follows:

- A fasting plasma glucose (FPG) test is used to measure the blood glucose in a person when he/she has not eaten anything in the last 8 hours. This test is helpful in the detection of diabetes and prediabetes so that their cure may be speeded up.

- An oral glucose tolerance test (OGTT) is used to measure blood glucose of a person who has been fasting for the last 8 hours. But before the test is conducted the patient is made to drink a glucose-containing beverage just 2 hours before his blood sample is taken for the analysis. This test is used to diagnose and detect both prediabetes and diabetes. The FPG test is the most favorable test for the diagnosis of diabetes because of its very good convenience and a merely low cost. However, it may be possible that this test may miss some diabetes or prediabetes that are identified only with the OGTT test. The FPG test thus is the most dependable when it is conducted in the morning as the patient had the earlier night's dinner and nothing else. Researches have shown that the OGTT test is more sensitive than the FPG test for diagnosing and revealing both prediabetes and diabetes, but the problem with it is that it is less convenient to be administered.
- A random plasma glucose test which is also known as the casual plasma glucose test, measures blood glucose level without any concern to the time when the person who is being tested had eaten anything for the last time. This test along with an invigilation of other symptoms has been used to diagnose and detect diabetes, but they are not successful for the detection of prediabetes.

  Test results which indicate that the particular person has diabetes should be confirmed with a second test which is to be conducted on a different day [10]. The current WHO diagnostic criteria for diabetes should be maintained – fasting plasma glucose ≥ 7.0mmol/l (126mg/dl) or 2–h plasma glucose ≥ 11.1mmol/l (200mg/dl)

1.4 Adverse effects of Diabetes Mellitus

Besides many physical side-effects of diabetes it also stands as a firm reason behind the many generalized side-effects on a person's mind and soul which have been explained as follows:

1.4.1 Quality of Life

Diabetes is a disease with adverse after effects. Almost every diabetic person seriously feels that this disease strongly affects their lives, and most of them

feel burdened by the various demands of this disease, a practice that could be called "diabetes *overwhelmus*," since lot of people feel stunned by the continuous and tiring burden of this disease and its management along with its treatment. These social and emotional burdens may be exaggerated by the physical disorders of hyperglycaemia or hypoglycaemia and by the long-lasting physical distress of the /complications related to diabetes. When it comes to quality of life for men with diabetes, sexual problems are too common. The National Diabetes Information Clearing House (NDIC) states that erectile dysfunction can range from 20% to as high as 75% of men with diabetes [11].

1.4.2 Physiological Effects

It is widely accepted that diabetes can have an adverse effect on a patient's quality of life, which can strongly affect a person's obligation to diabetes management and treatment. The increasing awareness of the significance of quality of life has made the researchers put up a wide range of queries regarding their inter-relationships. Some physiological factors such as personality type, coping style, health-related beliefs and social support can make dynamic effect on the quality of life of the patient. The effects can either be direct or indirect, considering the harmful impact of this disease and its demands. Actually, the psychosocial factors so discussed become the most powerful predictors and analysts of the quality of life, often overshadowing the effects of important factors related to the disease [12].

1.4.3 Depression

The emotional state of people suffering from diabetes is important to be taken good care of and thus is also fundamental to the assessment of the overall health of the person, predominantly for the people with life-long complications such as diabetes mellitus. Such people need both psychological and emotional support for surviving with diabetes or because of the conditions external to this situation.

When it comes to diagnosis, the side effects because of the medicines and treatments, the development of a complication or dealing with the regular

responsibility of managing and treating diabetes can make them mar their emotional well-being. In some of the cases this can lead to anxiety, depression, phobias and eating disorders. The occurrence of depression is about twice as high in diabetic people as compared to the population in general [13]. People have been diagnosed with a long-lasting health problem which is physical in nature such as diabetics are four times more prone to be diagnosed with depression than people not suffering from it (www.diabetes.co.uk/diabetes-and-depression). Depression can have a serious impact on a person's well-being and their ability and motivation to self-manage their condition. Depression is the most common psychiatric disorder witnessed in the diabetes community. It may develop because of stress but also may result from the metabolic effects of diabetes on the brain [14]. Some studies have suggested that women with diabetes may be more likely to suffer from depression compared with their male counterparts [15].

1.5 Data Mining

Data mining is the process of automatically discovering and deriving useful information in from data which is saved in large data repositories. Data mining techniques help in analyzing large databases so as to find unique and useful patterns that might otherwise not be known to the researchers working in that field. They also provide abilities to predict the outcome also known as the target value of a future observation. Further, they can also be used in the addressing some vital challenges in biology such as multiple sequence alignment, prediction of protein structure and the modeling of pathways in Biochemistry. Data mining is the integral part of knowledge discovery in databases (KDD), which is the overall process of converting raw data into useful information [16].

The following data mining tasks are associated in this work in the diagnosis of diabetes:

- **Predictive tasks**: The objective of these tasks is to predict the value of a particular attribute (target or dependent variable) based on the values of other attributes (explanatory or independent variables). These tasks refer to the building of a model for the retrieval of the value of the target variable as a

function of the explanatory variables being used as attributes. There are two different types of tasks under the category of predictive modelling: classification (for discrete type target variables), and regression (for continuous type target variables) [17].

- **Descriptive tasks:** Here, the main goal of the task is to derive patterns (clusters, correlations, trajectories, anomalies and trends) that collectively or separately recapitulate the fundamental inter-relationships in the data. Since descriptive tasks are often fact-finding by nature, thus techniques of post processing are used to explain and validate the results that are relevant [17].
- **Cluster analysis:** This is the technique which is used to find groups of observations that closely related to each other so that the observations that reside in the same cluster are more similar to each other qualitatively than the observations that belong to different clusters [17].
- **Anomaly detection:** This part helps in identify those observations whose characteristics are considerably different from rest of the data in the dataset. Such observations are known as outliers or anomalies. The basic aim of this procedure is to determine the actual anomalies and avoid the incorrect labelling of the normal objects as anomalous [16].

1.5.1 Knowledge Discovery and Data Mining

Data mining is an integral part of Knowledge Discovery from Database (KDD) which is the overall process of converting raw data into useful information as shown in figure 1.1. The term Knowledge Discovery in Databases, or KDD in short, refers to the elaborated process of extracting knowledge from the raw data, and puts an emphasis on the "high-level" application of particular methods of data mining used for such purpose. It is of great interest to researchers who follow a career in machine learning, databases, pattern recognition, statistics, and knowledge acquisition for expert systems along with the use of artificial intelligence and data visualization. The unifying goal of this process is knowledge extraction from data in the context of extremely large databases which cannot be maneuvered manually. It is done by the use of data mining methods (algorithms) to

extract (identify) useful facts and figures in the form of knowledge, using a database along with the required preprocessing, subsampling, and transformations of the data. This process consists of series of steps of transformation, starting from the data preprocessing to post-processing of data mining results so achieved. The overall procedure of discovering and understanding varying patterns from the data involves the iterative execution of the steps as follows:

1. Development of an understanding about the
    - the domain of the application.
    - prior knowledge related to the application.
    - the goals which are to be achieved by the end-user
2. Creation of the target data set which refers to the selection of a data set, or focus on the variable subset on which the discovery is to be performed for the data mining results.
3. Data cleaning and preprocessing.
    - Elimination of outliers or noise.
    - Collection of essential information to account or model for noise.
    - Approaches to handle the missing data fields.
4. Data projection and reduction.
    - Finding features which are actually useful for data representation depending upon the goal of the task.
    - Use of dimensionality reduction to reduce the number of attributes which are to be taken into consideration.
5. Choosing the approach for data mining.
    - To decide if the goal of the KDD process is regression, classification, or clustering, etc.
6. Choice of the algorithm(s) for data mining.
    - Selection of the method(s) which are to be used for the search of patterns in the dataset available.
    - Determining which parameters and models may be the most suited here.

7. Data mining.
    - Search for the patterns which are of high interest in a particular form or representation such as in the form of classification trees or rules, clustering, regression, and so on.
8. Inferring the patterns so mined.
9. Fusing the knowledge discovered.

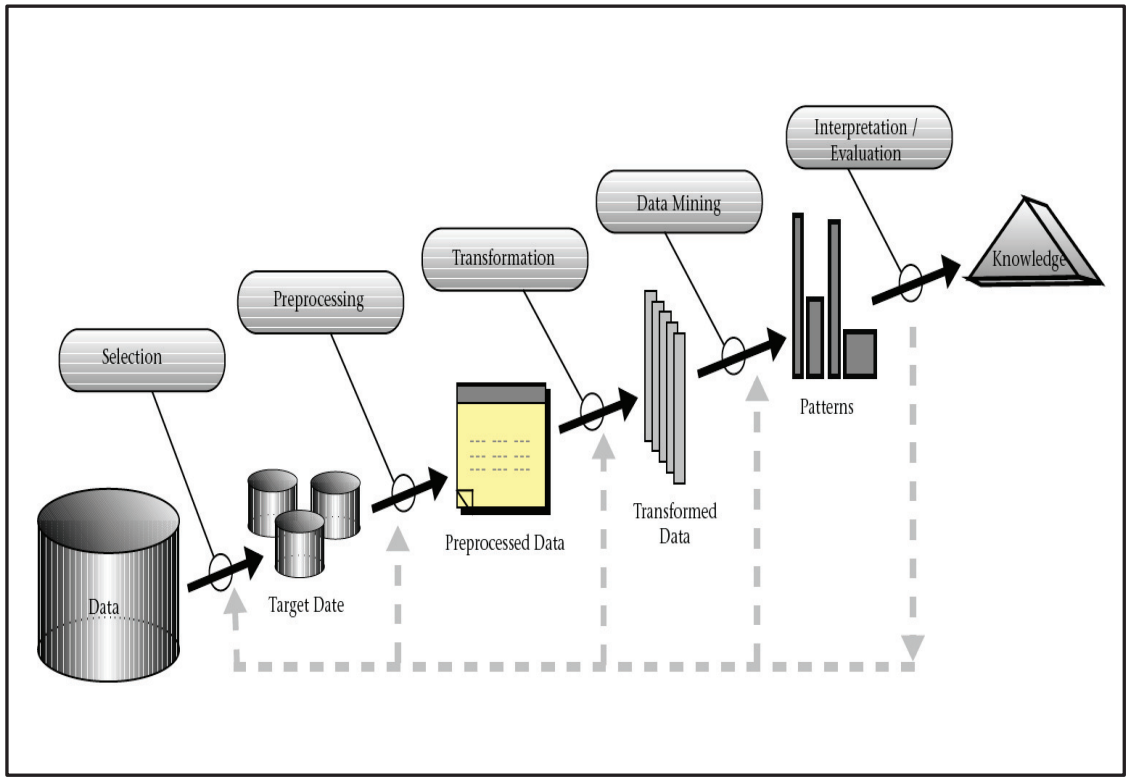

Fig 1.1: Process of Knowledge Discovery from Database

The data which is to be given as the input can be stored in a number of formats (such as file systems, spreadsheets and relational tables) and they may be kept in a repository which is centralized or be circulated across any number of sites may they be remote or not. The basic aim of preprocessing is the transformation of the raw data into a suitable format for the analysis to follow. The steps involved in data preprocessing have been explained above vividly. Because of the numerous ways available for the data to be collected and stored, data preprocessing appears to be the most arduous task and the most time-consuming step in the overall process of knowledge discovery.

The dimensionality of a dataset is said to be the number of attributes that the instances in the data used have. Dataset with a fewer number of dimensions are likely to be different and better than high-dimensional or moderate data in the terms of quality of data. Thus, the complications related to the examination of high-dimensional data are sometimes known as the curse of dimensionality. Thus, a significant inspiration in data preprocessing is dimensionality reduction. When working on the sequential data, it is important to take into account temporal automated correlation; i.e., if two dimensions are close in time, then the values of those dimensions are often very similar and thus one of them can be ignored without affecting the result to any extent. The term dimensionality reduction is often earmarked for those techniques that reduce the dimensionality of a data set by the creation of new attributes that tend to be a combination of the old attributes. The dimensionality reduction by the selection of newer attributes that form subset of the old attribute is also sometimes referred as feature subset selection or feature selection. The advantages of dimensionality reduction are as follows:

(i) Mining algorithms perform better if the attributes in the dataset used are lesser.

(ii) Dimensionality reduction can lead to the elimination of unrelated features and the noise present.

(iii) Can lead to a model which is more understandable and involves lesser number of attributes.

(iv) May allow the easy visualization of data.

(v) Even if this process is unable to reduce the data to two or three dimensions, data can still be visualized by looking at the pairs or triplets of the attributes so left, and the number of such combinations of attributes is greatly reduced.

(vi) The amount of memory and time required by the data mining algorithm is reduced drastically with the reduction in dimensions of the dataset.

"Closing the Loop" is the phrase mostly used to refer to the practice of integrating the results of data mining into the performance of the decision support systems available. For example, taking the case of business applications, the perceptions

made on the basis of the data mining results can be combined with the tools for campaign management so that marketing promotions can be conducted and tested effectively. Such kind of assimilation requires a post-processing step that assures the researcher or the user that only useful and valid results are amalgamated into the decision support system. An important example of post-processing is visualization which allows the researchers to explore and analyze the data and the data mining results from a variety of viewpoints can be so achieved. Hypothesis testing methods or statistical measures can also be applied during post processing to eliminate false results of data mining.

1.6 Thesis Outline

Chapter two provides the literature review on the applications of data mining where in research papers discuss novel techniques for achieving better performance of the classifiers.

Chapter three presents the Problem statement in brief with the objectives of the study.

Chapter four gives the methodology used in various data mining techniques. The hybrid is made by the combination of certain data models when they are arranged in the increasing order of their accuracy.

Chapter five discusses the results and the performance analysis is being done for determining the best classifier used in this study on the basis of accuracy.

Chapter six concludes the thesis with a scope for future enhancement.

CHAPTER 2

LITERATURE REVIEW

This chapter summarizes all the relevant work that has been done in this field of diabetes prediction. Some research work from the heart disease prediction has also been used in this study so as to make use of the complete information available relevant to the work done in this study.

Today, data mining, a process of finding previously unknown patterns and trends in data to build predictive models, has grown so fast that it has been used in many applications. It is an approach to data analysis and knowledge discovery which originated from work of statistics and machine learning as an interdisciplinary field. Intelligence is the capability of learning, understanding and finding solutions for problems in a specific domain and Artificial Intelligence is a branch of Computer Science which deals with the study of "intelligent agents" who are trained to respond according to the environment and their perception about it thereby increasing their chances of success at some predefined goal. Machine Intelligence can be used for medical data mining as it extracts biomedical and health care knowledge for clinical decision making and generates scientific hypothesis from large medical data. KDD, which includes data mining techniques, has become a popular research tool for healthcare researchers. The need for data mining arises from the data which seems to increase rapidly every single day for the majority of domains related to information processing, and the need to find a way to mine and get knowledge from databases. Its applications also can benefit healthcare providers such as hospitals, clinics, physicians, and patients by identifying effective treatments and best practices.

Health is a common theme in most cultures. Among definitions still used, probably the oldest is that health is the absence of disease. Health is not mainly about issues of' doctors, social services and hospitals. It is an issue of social justice. The widely accepted definition of health given by the WHO (1948) is 'Health is a state of complete physical, mental and social wellbeing and not merely an absence of disease or infirmity'.

Modern medicine has evolved tools and techniques which may be used in various combinations for the assessment of physical health. They include self-assessment of overall health, inquiry into symptoms of ill health and risk factors, inquiry into the use of medical services, standardized questionnaires for cardiovascular diseases and clinical examination.

In recent years, data mining has been used widely in the areas of science and engineering, bioinformatics, genetics and medicine. It is a collection of algorithmic ways to extract informative patterns from raw data. It plays an important role in tackling the data overload in medical informatics.

With the development of information technology, extensive medical data is available. Medical data classification plays a crucial role in many medical applications. It is the process of transforming descriptions of medical diagnoses and procedures to universal codes. Diagnosis codes are used to track diseases and other health conditions, even chronic diseases such as diabetes mellitus and heart disease. Medical classification is widely used in hospitals for the statistical analysis of diseases and therapies. It addresses the problems of diagnosis, analysis and teaching purposes in medicine. Medical data has made a great progress over the past decades in the development and use of classification algorithms. In healthcare, medical data can be transformed into aggregations, to calculate average values per patient and compare with other values, to group data into clusters of similar data *etc*. However, few challenges include data mining methodology that are user interaction, performance and scalability. The relationships of the disorders and the effects of symptoms that are unexpectedly seen in the patients can be evaluated by the medical practitioners via these techniques very easily and thus medical data mining is taking a roll of the path of success. Knowledge of the risk factors related with diabetes help the health care professionals to identify the patients who are at a high risk of having diabetes. Data mining techniques and Statistical analysis help the healthcare professionals in the diagnosis of this disease.

Gandhi *et al.* (2014) proposed a novel technique for the diagnosis of diabetes for the very common dataset of Pima Indians0. The SVM classifier has been used to predict the patients who are prone to or are on the onset of daibetes. Feature selection is

performed using F-score and k-mean clustering methods to obtain optimal set of features. F-score gives better performance of classification than other feature selection methods like relief and relief filtering methods. For data normalization Z-Normalization has been used. Accuracy, sensitivity and specificity have been used for calculating the performance of the model named SVM. Embedding feature selection and data normalization improves the performance of SVM classifier. High accuracy of proposed technique can be considered as a good candidate for disease diagnosis. It has achieved an accuracy of 98%, Sensitivity of 97.77%, Specificity of 97.79% [18].

Priyadarshini *et al.* (2014) experimented and used the concept of modified extreme learning machine to identify the patients of being diabetic or non-diabetic basing on some previously given data which in turn helps the medical people to identify whether someone is affected by diabetes or not. It also describes and compares the application of two popular machine learning methods: Back propagation neural network and modified Extreme learning machine which are used as binary classifiers to address the diabetes prediction problem. These two approaches are applied on same type of multi class classification datasets and the work tries to generate some comparative inferences from training and testing results [19].

Ibrahim *et al.* (2013) presented the innovative and unique hybrid model by discovering Agglomerative Hierarchical Clustering along with the classifier named Decision Tree to be applied on the diabetes dataset of Pima Indians. The experiments so conducted were used to compare the performance in the terms of accuracy of the classifier which is decision tree against the performance of the same classifier when improved with the use of Hierarchical Clustering. Results have shown that the hybrid model so developed has attained the highest accuracy with 80.8% as and when compared to 76.9% of the model which is the standard for this very study. This is an encouraging result for the acceptance of this hierarchical clustering in a classifier which is rule-based in nature [20].

Aslam *et al.* (2013) used genetic programming model for the classification of diabetes. Genetic Programming has been used in this study to generate new attributes or features from the combinations of existing diabetes attributes or features, without any prior knowledge that which of the probability distribution has been used. The proposed technique has been divided into three stages: first one being the feature selection which is to be performed with the use of F-score selection, t-test, Kullback–Leibler divergence test, GP and Kolmogorov–Smirnov test. The results so achieved after the feature selection are then used to generate a systematic list of original features of the dataset where features have been arranged in decreasing order of importance such that the feature with highest importance comes at the first spot and vice versa. The performance of features generated through GP for classification purposes is tested with the use of support vector machine (SVM) and k-nearest neighbor (knn) classifiers. The results so obtained are then compared with the results of other methods used and it is realized that the proposed method demonstrates the best performance over the other methods available [21].

Rajesh *et al.* (2013) point out that Gestational Diabetes Mellitus refers to any irregular carbohydrate which starts or is surfaced during pregnancy for the first time in a female. It also includes the possibility of the presence of some unknown glucose intolerance that had preceded the state of pregnancy. This condition is found to occur more in the urban areas as compared to the rural areas. Global screening for 50-g oral glucose test is recommended to be done when the lady is at the 24-28th week of her gestation period. Women which appear to have 1-Hour glucose level greater than 140 mg/dl are referred to undergo the diagnostic Oral Glucose Tolerance Test also known as OGTT in common. In this study those women were enrolled who were in the 24th and 28th week of their gestation period and were made to attend an antenatal care (ANC) clinic at a tertiary care hospital in Rohtak, Haryana. After taking a consent from the women who were going to participate in this study, they were made to undertake a standard 2-h 75 g oral glucose tolerance test. A performa consisting of general information on social and economic status, demographic characteristics, parity, family history of diabetes, hypertension and education level was filled up by the women in this survey. Results showed that a total of 607 women were the

participants of this study and 43 were diagnosed with GDM that is 7.1% of all the women participants [22].

Anuja Kumari *et al.* (2013) have illustrated the application of SVM classifier which is a supervised method of machine learning for the diagnosis of diabetes in the patients of the dataset Pima Indians from the UCI repository. They have made use of the simple conceptions of kernel function selection and SVM and the experiments have been conducted on Matlab. The level of effectiveness of the SVM model has been calculated on the basis of the number of correct and incorrect classifications. The data set was evaluated using 10 fold cross validation error rate, the error rate focuses True Positive (Sensitivity), True Negative, False Positive (Specificity), False Negative and Accuracy. The diagnostic performance of the developed model was evaluated using ROC curve also known as the Receiver Operating Characteristic curve and it was plotted between true positive rate and false positive rate which describe the degree of how positively the disease is been predicted. Finally the results were shown that the performance parameters such as the classification sensitivity, accuracy and specificity of the models named SVM and RBF have been found out to be 80%, 78% and 76.5% respectively [23].

Aishwarya *et al.* (2013) have depicted that machine learning is an effective technique which can be used for the purposes of both classification and prediction just on the basis of recursive learning. It allows training and test classification system, with Artificial Intelligence. Automatic learning has fetched a greater amount of interest in medical domain due to less amount of time for detection and less interaction with patient, saving time for patients care. It has provided the greatest amount of support for predicting whether the patient has disease named diabetes with the correct use of both training and testing datasets. Amongst the most promising and flourishing techniques which are used in machine learning lies Support Vector Machine which stand for SVM used for classification purposes [24].

Zolfaghari *et al.* (2012) proposed a framework- an effective biological machine learning algorithm for the diagnosis of diabetes in female patients. In this study they

tried to predict the presence of diabetes based on ensemble of Support Vector Machine and Back propagation Neural Network. The predictive accuracy was 88.04% and it was very promising with regard to the other classification systems in the literature for this problem. [25].

Karthikeyani *et al.* (2012) have illustrated the classification from the supervised data mining category of algorithms based on diabetes disease dataset in which different classification algorithms like C4.5 decision tree, Support Vector Machine, Regression and Classification Trees, KNN and Prototype Neural Network classification have been used to analyze the Pima Indian Diabetes dataset with 9 attributes and 768 instances. Here they have compared the performance of the data evaluated using 10 fold Cross Validation error rate, precision value and time for computation. They have achieved the results using Tanagra tool. A classification rate of 86%, 85%, 78% and 74% of accuracy was obtained for C4.5, CART, K-NN and SVM algorithms respectively [26].

Rajesh *et al.* (2012) aimed for the mining and discovery of relationships in diabetes data so that an efficient level of classification can be carried on the Pima Indians dataset. They have used C4.5 algorithm as well as known decision tree induction learning technique which are used to predict the class label as the target of the dataset. The final outcome consisted of patterns used to find out whether the patient who is being examined is suffering from diabetes or not. An accuracy of 91% was obtained with the use of C4.5 algorithm [27].

Sarojini Balakrishnan *et al.* (2011) have proposed a system to improve the diagnostic accuracy of diabetic disease by selecting informative features of Pima Indians Diabetes dataset. They propose a hybrid prediction model that combines two different functionalities of data mining clustering and classification with F-score selection approach to identify the optimal feature subset of the Pima Indians Diabetes dataset. The process of feature selection using the F-score method and clustering using k-means select the most optimal subsets of features from the medical datasets thus to enhance the performance of the SVM classifier. In their

study accuracy for the dataset namely diabetes dataset comes out to be 98.9427%, for cancer dataset it comes to be 99% and for heart disease dataset it comes out to be 100%. They identify the significant features of the medical datasets. Hence, the results prove that when the optimal subset of features is derived for the above mentioned datasets the accuracy of the classifier improves manifolds [28].

Çalişir *et al.* (2011) proposed an automatic diagnosis system for the prediction of diabetes on the basis of Linear Discriminant Analysis (LDA) which has been combined with the popular Morlet Wavelet Support Vector Machine Classifier and their ensemble LDA–MWSVM has been introduced. The structure of the system so introduced has been divided into primarily three stages: The feature reduction and the selection of optimal features subset has been performed with the use of Linear Discriminant Analysis (LDA) technique and the process of classification has been concluded with the use of Morlet Wavelet SVM classifier. The features of both healthy and diabetic patients so obtained in the very first stage are provided as inputs to the MWSVM classifier in the succeeding stage. Then, in the third stage, the performance of this ensemble so made is evaluated on the basis of sensitivity, accuracy, specificity. The accuracy for the classification done by this system was found out to be 89.74% [29].

Kavitha, *et al.* (2010) discussed that knowledge discovery in databases and data mining is an interdisciplinary area focusing on the methodologies for extracting useful knowledge from data. They stated the necessity of effective identification of information, contextual data, obvious and valuable for decision making from a large collection of data has been on a steady increase recently. This is an interactive and iterative process encompassing several subtasks and decisions and is known as Knowledge Discovery from Data. This paper has depicted the use of KDD in any field where prediction and classification is to be done for the purpose of data mining using a data set. The central process of knowledge discovery is the transformation of data into knowledge for decision making, known as data mining [30].

Porter *et al.* (2009) showed the benefits of using DM in the healthcare domain. It made a theoretical contribution, as it exhibited a formal presentation of the DM process, while integrating several concepts from other disciplines. The results that were shown in this study can help decision makers in determining a health policy related to diabetes. It presented a model for identifying diabetes patients from large medical datasets. This approach clusters the similar kind of diabetes patients into different groups of subpopulations. Data transformation and discretization techniques have been used to improve the quality of data which is been worked upon. The clustering algorithms helped to gain higher accuracy as and when compared to the models already available for the classification of the females of Pima Indians heritage on the basis of the onset of diabetes and its symptoms [31].

Jianchao Han *et al*. (2008) followed the process of data mining so as to determine whether the person from the Pima Indians dataset is suffering from diabetes or not. This work focuses on data pre-processing, including attribute identification and selection, outlier removal, data normalization and numerical discretization, visual data analysis, hidden relationships discovery, and a diabetes prediction model construction. RapidMiner was used for attribute modification so that better understanding of the attributes can be gained which can further be used to build a predictive model to determine whether the patient is suffering from diabetes mellitus or not. Two main options considered were: the ID3 Algorithm and the Decision Tree [32].

Kahramanli *et al.* (2008) have worked for the reliability of results in a classification mechanism where crisp and fuzzy values are to be used, and in this regard an innovative method has been developed in this study. The hybrid approach has ensembled two kinds of neural network that are fuzzy neural network (FNN) and artificial neural network (ANN) into one. The proposed method was investigated on two real-time disease datasets which were Cleveland heart disease dataset and the Pima Indians diabetes dataset. Sensitivity, accuracy and specificity were taken as the evaluation parameters to calculate the performance of the ensemble so formed as these three are the most important. The accuracies so obtained for these datasets

were achieved by *k*-fold cross-validation and their values come out to be 86.8% and 84.24% for Cleveland heart disease dataset and Pima Indians diabetes dataset respectively [33].

Kayaer *et al.* (2003) gauged the performance of classifier general regression neural network (GRNN) on the diabetes disease dataset. In the process of classification of patients into diabetic or non-diabetic the accuracy of GRNN is comparable to the accuracy of the classification done by which has been used as the model for this purpose in the reference work. The accuracy achieved by GRNN is the promising with the advantage that the structure of GRNN is rather simpler than the other structures except ARTMAP-IC and the value of accuracy is found out to be 80.21% which is comparable to the accuracy of ARTMAP-IC having value 81% [34].

Carpenter *et al.* (1998) discussed the simulations examination of the accuracy of prediction on four health related medical datasets which are breast cancer, Pima Indian diabetes, gall bladder removal and heart disease dataset. ARTMAP-IC as proposed in this study resulted in the performance which was at par or rather better than the performance of multi-surface pattern separation, logistic regression, the ADAP perceptron, K nearest neighbour (KNN), CLASSIT, C4 and instance-based (IBL). ARTMAP was based on the dynamics that were stable, fast and highly scalable. A voting technique was used to improve the prediction accuracy by training the classifier system for a several number of times on various orderings of the input dataset. Distributed representations, instance counting and voting were put together to form estimates on the basis of confidence for evaluating the predictors by the way of classification accuracy [35].

After going through various studies conducted in the field of Diabetes research it has been analyzed that in majority of the research papers focus has been on use of single classifier and we feel that accuracy of the results can be improved by using more than one classifier on the given data sets. Hence coming chapters discuss ensemble approach which has been applied in this work.

CHAPTER 3

PROBLEM STATEMENT

Discovery of new information in terms of patterns or rules from large amounts of data is basis of machine learning. Disease prediction plays an important role in data mining. Diagnosis of a disease requires the performance of a number of tests on the patient. However, use of data mining techniques, can improve the diagnosis process considerably. Diabetes data mining is important because it allows doctors to see which features or attributes are more important for diagnosis such as age, weight, etc. This will help the doctors diagnose diabetes more efficiently. There are various data mining techniques in use in healthcare industry to improve the performance of the various classification techniques and enable the choice of the best among them. The research presented in this thesis is intended to address the challenge of improving the prediction model to predict the onset of diabetic patients and providing timely response in predicting the disease. Briefly the important research functions are therefore stated as

- How various data mining techniques can be used in health care industry and to identify their performance in prediction?
- How does a classification techniques help in developing the prediction model so as to predict accurately the risk of heart disease among diabetic patients?

3.1 Objectives of the Research

Application of data mining in analyzing the medical data is a good method for investigating the existing relationships between variables. Nowadays, data stored in medical databases are growing in an increasingly rapid rate. It has been widely recognized that medical data analysis can lead to an enhancement of health care. The primary objective of the research work is the effective development of prediction model using various classification techniques to predict the onset of diabetes and evaluate its performance in the prediction of the target value. The following are the objectives laid out for this work:

- To study various classification techniques used in machine learning.

- To study the use of various techniques which are to be employed for the purpose of aggregating the results of more than one classifier.
- To propose an ensemble framework for prediction of onset of diabetes.
- To evaluate the performance of proposed ensemble framework with other classifiers using various evaluation parameters such as accuracy, sensitivity and specificity.

CHAPTER 4

PROPOSED APPROACH: HPCC

4.1 Machine Learning

Machine learning is a kind of artificial intelligence (AI) layout that fulfils the technicalities of computers by making them able enough to learn without the computer being explicitly programmed. It is the study of computer algorithms that improve automatically through experience and has been central to AI research since the field's inception. Machine learning is the science of making our computers to act like humans without getting them explicitly programmed. Many researchers and scientists think of machine learning as the best way to make progress on the path of achieving human-level AI. It lays the focus on the creation of computer programs which can improve and change themselves when encountered with new data. Its goals are to learn complicated patterns and to make intelligent decisions based on input data automatically. To deal with a problem in a computer, one first plan an appropriately competent algorithm that deals with the problem and then designs and implements that algorithm in software or hardware. One cannot solve the problem without implement and design an algorithm for that problem. When we are unable to solve a problem manually then Machine Learning extend what can we do with a computer, and how we play with the programmed algorithm.

Machine learning as a logical train inspects the computational premise of learning; consequently, it is basic regardless of the possibility that we are just keen on how people and creatures learn. Machine learning is organized around three primary research foci (Michalski *et al.* 1997) [37] that are:

- Task-oriented studies – The development and analysis of learning systems to improve performance in a pre-determined set of tasks (also known as engineering approach).
- Cognitive simulation – The investigation and computer simulation of human learning processes

- Theoretical analysis – The theoretical exploration of the space of possible learning methods and algorithms independent of application domain.

4.1.1 Types of Machine Learning

(i) Supervised learning: This algorithm breaks down the information for preparation and inherently leads to the production of a derived capacity, which can finally be used for the mapping of new cases or instances to be encountered. It classifies the data mainly on the basis of features of the instances and the classes that are available as options. Supervised learning as regression (for persistent yields) and order (for discrete yields) is an important constituent of Machine Learning. For example, you have input (x) and a yield (Y) and you utilize a calculation to take in the processing capacity called yield with the assistance of this information.

$$Y = f(X) \tag{1}$$

The point lies in the exact realization of the registering capacity so well that when you utilize new information (x) you are able to foresee yield factors (Y) for that information. Supervised learning manages to learn a capacity from accessible information. A supervised learning algorithm breaks down the information for preparation and leads to the production of a construed work, which can finally be used for the mapping of new illustrations. Here are some examples are shown below.

- Classification of the e-mails as important or spam
- Labeling of the web pages on the basis of their content
- Voice and Speech recognition.

There are a number of supervised learning algorithms, for example, neural networks (NN), Naive Bayes classifiers and Support Vector Machines (SVMs) etc.

Supervised learning problem can be grouped into classification and regression problem.

a) **Classification**: The goal of the classification algorithm is to predict the target class: yes or no. For predicting two target value or class we use binary classification, *i.e.* to predict student profile status fail or

pass. When we have to predict for more than two target data class we use the multiple classifications, *i.e.* considering all the details of the students to estimate which students will earn more points.

b) **Regression**: The goal of regression algorithm is to predict continuous or discrete values. Once in a while, the foreseeing quality can be utilized to locate the straight connection between the attributes. Basic regression algorithm such as linear, polynomial, *etc.* is used in machine learning problems. Some famous regression algorithm of supervised learning are follows-

Linear Regression: It is used to gauge genuine esteems (cost of apartments and houses, the quantity of calls, deals on aggregate etc.) in a way that the opinion of a persistent variable(s) is sustained. In this case, we set up a connection amongst autonomous factors by the fit of the best line found amongst many linear solutions available. This line which fits the best is called the relapse line expressed in a straight form as:

$$Y = a * X + b \quad (2)$$

(ii) Unsupervised learning: The primary aim of this kind of learning is to design and model the basic structure of the data used and to derive knowledge about the distribution of data in order to learn more about it. Unsupervised learning is a type of machine learning algorithm that draws references from a dataset with input data without labeled responses. It is different from the supervised learning or the reinforcement learning in a way the in this case the learner is trained on the set of data with unlabeled instances. Unsupervised learning technique is further categorized into association and clustering problems:

a) **Clustering:** Clustering is the technique of the arrangement of instances into subsets or sub-populations (also called bunches) so that the instances which are similar to each other or are comparable in some sense belong to the same cluster or group. A clustering issue is a place you need to find the innate groupings in the information, for example, gathering clients by obtaining conduct. It is an approach under the unsupervised learning

technique and is an investigation process for the retrieval of factual information which is to be used in many fields. Following are some real world examples of clustering:

- In the field of astronomy, with the help of the auto class system a new kind of star was discovered, based upon the clustering of astro-physical measurements.
- Clustering can be applied in the area of e-commerce where it is common to cluster users into groups on the basis of their web-surfing behavior and purchasing activities. By the clustering results so obtained the merchant can send personalized and customized advertisements to the persons concerned.

b) Association: This type of learning is the place where you need to decide the portrayal of expensive segments of your information, for example, individuals that purchase X additionally tend to purchase Y. Some prominent cases of unsupervised learning calculations are:

- K-means for clustering problems.
- Association rule mining using Apriori algorithm.

(iii) Semi-supervised learning: To defeat the drawback of supervised learning algorithms that they can't make utilization of unlabeled information, semi-supervised learning (SSL) has been proposed to use both marked and unlabeled information [17]. Common approaches to semi-supervised learning include-

- Generative Models
- SVMs
- Graph-Based Algorithms

4.2 Dataset Used

The objective of this popular dataset is to carry out the diagnosis of the onset of diabetes in Pima Indians. Based on personal data of the patient, such as age and the outcomes of various medical checkups being carried out, e.g., Body mass index (BMI), blood pressure etc., it is decided whether or not a particular Pima Indian individual is on the risk of having diabetes. Pima Indian Diabetes Dataset can be freely downloaded from the available machine learning database in the UCI

repository [36]. The Pimas are the group of people who live in Native America which is now in an area having central and southern Arizona. Patients considered in this data set are females with minimum 21 years of age residing in Phoenix, Arizona. This comes under the category of a binary class problem having class value 1 to be construed as "positive for diabetes" and class value 0 to be considered as "negative for diabetes". There are 500 instances of class 1 and 268 instances of class 0. The objective of this study is to examine the connection between the diagnosis results and a given list of parameters that provide the medical attributes of the patient. The dataset consists of 768 instances and 9 variables having no missing values. The attributes with the class variable have been extracted in the basis of minimal quality tests done on the Pima females and the results so found have been taken as attributes and their values have been taken as attribute values. A total of 8 attributes are to be treated as the input variables for any model which is being used for classification purposes. These attributes and their characteristics are shown with their description in Table 4.1.

Table 4.1: Attributes of the Dataset and their description.

No.	Variable	Description	Value
1.	times_pregnant	Number of times the female got pregnant	Numeric
2.	glucose_conc	Plasma glucose concentration of 2 hours in a test of oral glucose tolerance.	Numeric
3.	diastolic_bp	Diastolic blood pressure (mm/ Hg)	Numeric
4.	Triceps	Triceps skin thickness (mm)	Numeric
5.	serum	2-Hour serum insulin concentration (mu U/ml)	Numeric
6.	BMI	Body mass index (weight in kg/(height in m)^2)	Numeric
7.	diabetes_func	Diabetes pedigree Function	Numeric
8	age	Age (years)	Numeric
9.	class	Class variable	Numeric(0 or 1)

4.3 Proposed Technique

The proposed technique takes into account the comparison of bag free models which are Decision Tree, Support Vector Machine, Generalized Linear model along-with the bagged approach with GLM and SVM which are then compared with the Boosted models which include GAMBoost and GBM. The final comparison is done with the Hybrid approach which inculcates the qualities of seven models into itself namely Neural Network, Support vector machine, Naïve Bayes, Decision tree, PLS, GLM, OWNN and makes an ensemble model which has the accuracy greater than all the models specified till now. The proposed framework of this study is depicted in Figure 4.1.

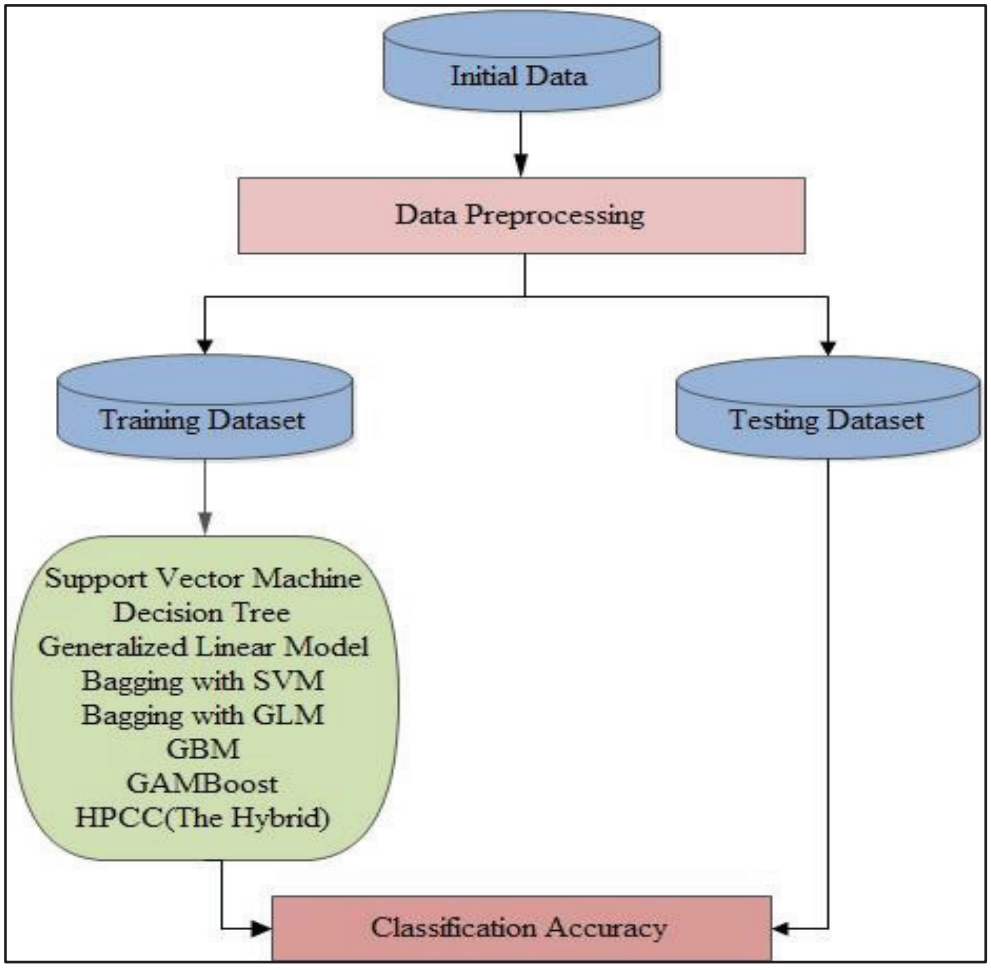

Fig 4.1: Proposed Framework

The ensembling is done on the basis of Majority Voting since the class variable of this dataset has only 2 values (0 and 1). Majority voting is the technique of combining the results of more than one classifier on the basis of highest frequency

vote. If at least two out of three classifiers label an unlabeled instance as 0 then the final output would be 0 and vice versa. This approach is also known by the name of plurality voting (PV). The framework of Majority Voting shown in Fig. 4.2 depicts the advantage of combining the results of the classifiers namely Neural Network (NN), Support Vector Machine (SVM) and Decision Tree.

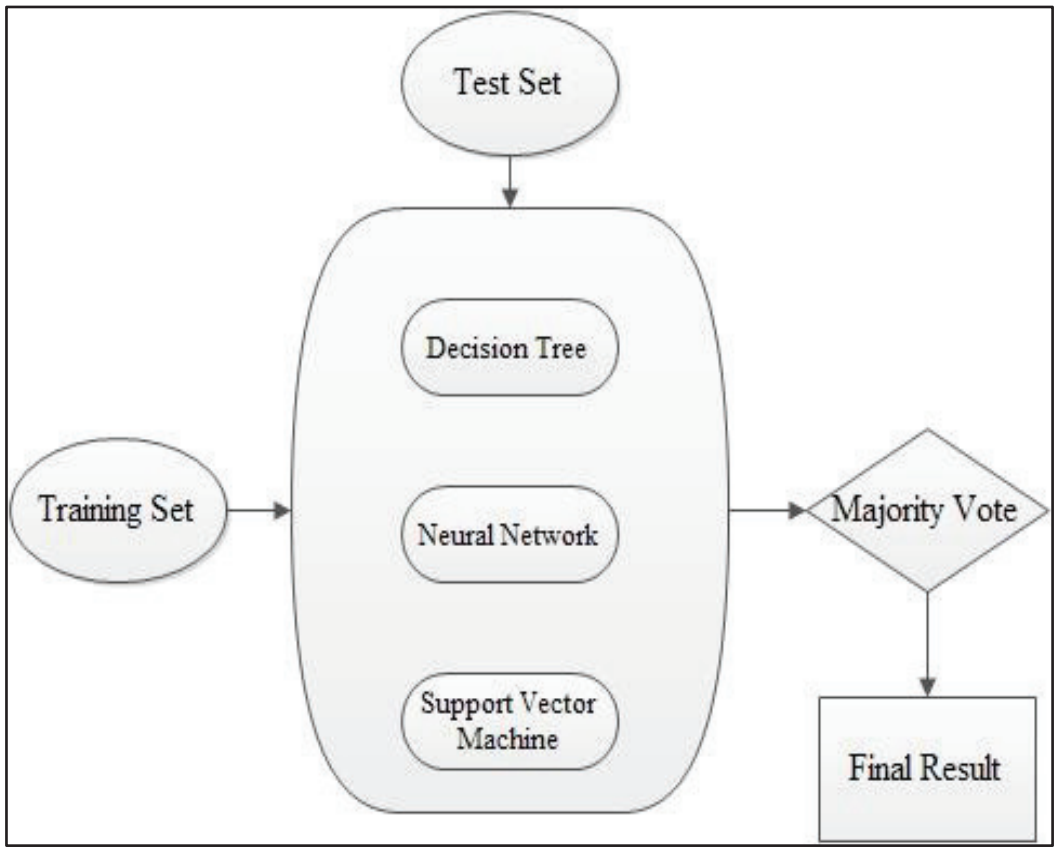

Fig 4.2: Majority Voting

The Models used in the study have been explained as follows:

4.3.1 Generalized Linear Model:

Normally linear models are centralized for the implementation of statistics through them. They form the core and integral part of the knowledge which is expected from any ordinary statistician. They provide the basis for a wide range of statistical approaches. The model takes the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots\ldots\ldots\ldots\ldots + \beta_p x_p + \varepsilon \quad (3)$$

where ε is normally distributed. The first part is used to generalize the y part of the equation; the second is the ε part which normalizes the equation on the basis of

normal distribution; and the third is the *x* part which acts as the input for the execution of linear model. Linear model is unable to handle the responses which are non-normal in nature like proportions and counts. There evolves the generalized linear model to rectify this short-coming which represents all the response types like binary, categorical *etc*.

In Generalized linear model (GLM) each outcome Y of the dependent variables is presumed to have been generated from a particular distribution of the exponential distributions family, which includes binomial, normal, gamma and Poisson distributions and many other. The mean μ of the distribution which is used has dependence on the independent variables i.e. $X = \{ x_1, x_2, x_3 \ldots x_n\}$ and the relation so formed is depicted by the equation as follows:

$$E(Y) = \mu = g^{-1}(X\beta) \quad (4)$$

here E(Y) is the probable value that Y can take; (Xβ) being the linear predictor, g is used as the link function and β is the linear combination of unknown parameters. The starting point for generalized linear models is the familiar general linear model (GLM), the most widely taught and used method of data analysis in psychology and the behavioral science today. The GLM is comprised of both analysis of variance, multiple regression and analysis of covariance. Multiple regression and analysis of variance allow researchers to study the relationships between one or more independent variables and a single continuous dependent variable. In multiple regression, the independent variables may be continuous or categorical; in analysis of variance, the independent variables are categorical, so that it can be considered a special case of multiple regression. The form of relationship between each independent variable and the dependent variable can be linear and curvilinear. These relationship can be general or conditional, potentially involving interactions between two or more independent variables.

4.3.2 Support Vector Machine:

SVM is a model under the category of supervised machine learning which is used mainly for the purpose of binary classification. A classification predictor is generated for each test set as input and corresponding output is produced which takes values of the two classes available thus creating a non-probabilistic binary classifier [38]. It is

a representation of the illustrations mapped in such a way that the illustrations of distinct categories are bifurcated by a gap which is clear and is as wide as possible. New illustrations are then plotted into the same space and expected to fit into a category on the basis that on which side of the predesigned gap they fall. This technique hence constructs a linear maximum margin hyperplane in a space with infinite dimensionality, which can further be used for regression, classification and other tasks. It is defined by a weight vector which is denoted by 'w' and bias been represented by 'b' which is the distance of hyperplane from the center. The non-linear separation of dataset is carried out by the use of a kernel function.

The classification rule which is used by the SVM classifier is depicted as follows:

$$Sgn(f(x,w,b)) \quad (5)$$

$$f(x,w,b) = \langle w.x \rangle + b \quad (6)$$

where f(w,b) presents maximum margin hyperplane for the complex problem and x denotes the example to be classified. Each base classifier which is being used in the generation of the ensemble is trained on the training dataset so as to make them worthy to be used for the prediction of diabetes. The feature space and the predicted class or target labels of the instances of dataset are to the knowledge of each classifier that has been trained, and which ultimately becomes capable of predicting sick and healthy persons from the dataset. The linear margin hyper-plane so found is maximum because a good bifurcation is attained by the hyperplane having the greatest distance to the closest training point belonging to any class (also known as the functional margin), as generally, the greater the margin, the lesser is the error of the classifier [38]. The primary benefit of SVM is its maximum classification accuracy. It is utilized for pattern recognition and is fundamentally designed for the two-class classification issue. It performs outstanding with the perfect margin for good separation and is really effective in spaces with high dimensionality. Let's take the scenario, where we are dealing with three hyper-planes namely A, B and C and all these hyper planes are partitioning the classes of the dataset very well. Now in order to find out the most appropriate hyper-plane the right hyper-plane right thumb rule is used: "Select that hyper-plane as the most suited one which separates the two

classes of the dataset better". And in this case, hyper-plane named "B" is doing this job perfectly. This has been shown in figure 4.3.

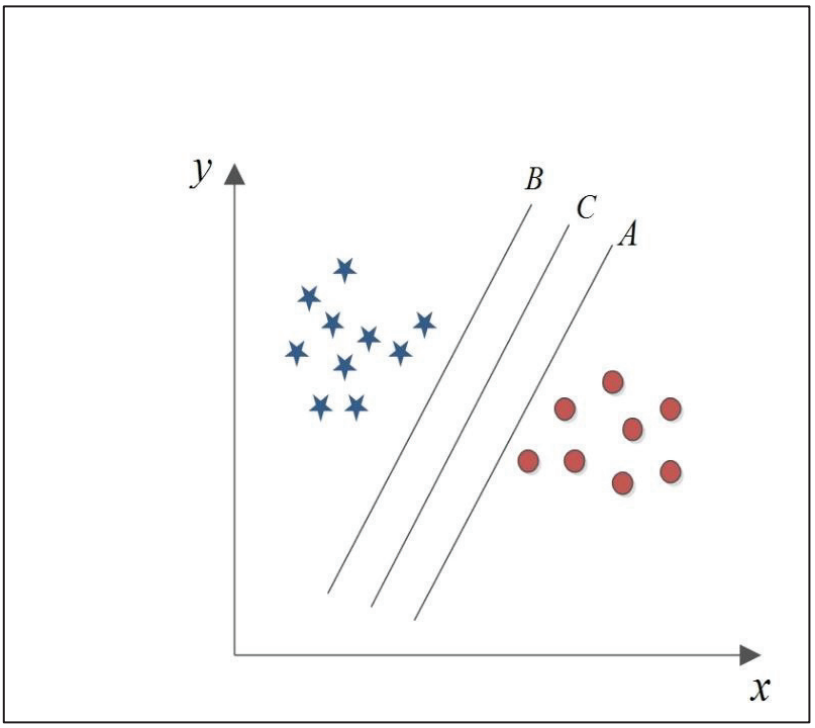

Fig 4.3: Hyperplanes

It is not good for large datasets because required training time is high. It does not perform well on noisy datasets i.e. target class is overlapping.

4.3.3 Decision Tree

Decision tree learning utilizes a tree as a predictive model that generates and converts perceptions about a thing (spoken to in the branches) into decisions about the objective esteem (spoken to in the clears out) of the thing which is being analyzed. It is one of the predictive modeling approaches utilized as a part of insights, information mining and Machine Learning. Decision Tree algorithm is a very easy technique that is used to make a decision by dividing the inputs into smaller decisions. It is used to predict the target class of the instance of the dataset which is being used just on the basis of much fewer variables given to it as inputs and with the help of a structured decision tree. Like other models, this one also includes mathematics but the mathematics used here is not of a very complex level. Given below is an example of a much simpler decision tree for the purpose of understanding the basic concept in figure 4.4.

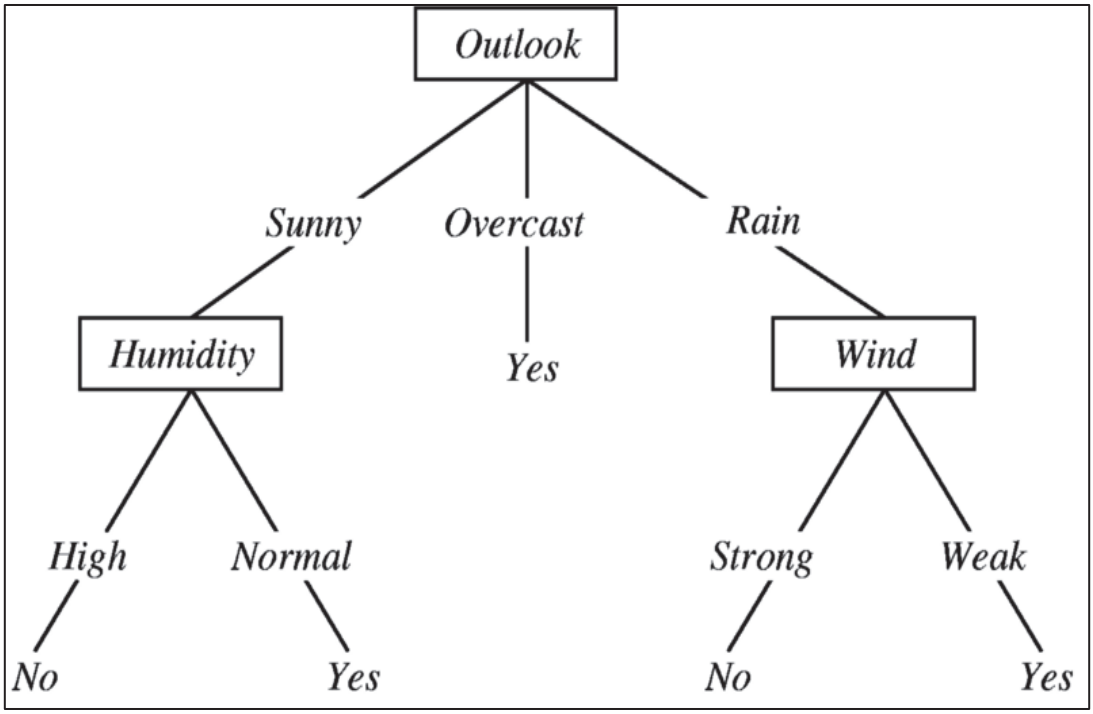

Fig 4.4: Decision Tree

The goal here is to make a decision on whether the person should play golf on a particular day or not. This is done by making an informed decision on the basis of temperature, wind, humidity and also that the weather is cloudy, sunny or rainy. For example if it is Rainy and the wind is weak then the decision for playing golf will have the outcome 'yes' provided the weather is not humid or mild.

The tree is divided into three kinds of nodes namely internal nodes or the leaf nodes where the condition is the internal node and what decision comes out as the output is the leaf node [39]. The values that the leaf nodes possess are the predefined classes of the dataset. Here in this example the decisions yes or no and the leaves and the factors windy, sunny etc. are the internal nodes. The root node is the very first nodes and is that attribute from the instance which has been selected as the base to build the decision tree upon. The branches of the tree work as the possible values that the particular internal node or the root node may possess. The most powerful features are the features which get selected earlier during the process of making of the tree and they take the place as the root node or the nodes at the upper levels. The unimportant features either occupy the lower levels or do not find a place in the final decision tree. These trees are built in the top-down fashion and work upon the recursive

divide-and-conquer strategy. The making of the tree goes on until a termination criteria is achieved. The process starts by the demarcation of a root node from the input features which has the closest relationship with the output variable. The further nodes are selected by the calculation of the Information Gain (IG) and the formula to calculate this is as follows:

$$IG\ (parent,child) = Entropy(parent) - [p(c_1) * Entropy(c_1) + p\ (c_2) * Entropy\ (c_2)] \quad (7)$$

here $Entropy(c_j) = (-p(c_j) * \log(p(c_j)))$ and $p(c_j)$ is a probability of the child node j. The node having the greatest value of IG is selected as a parent node for the generation to follow. This process is iterated till it gets the leaf node and the tree is completed. There exist a number of algorithms of decision tree which include C4.5, ID3 and CART. Every technique uses a different measure for the selection of the best split so as to find out the most suitable fit to construct the tree [40]. The following is an example of a rather complex decision tree on one of the classification problems in figure 4.5 where the income comes out to be high, medium or low. The high class has been further divided on the basis of age which has two categories on is the person who has an age less than 31 years i.e. (<=31years) and the one who has an age between 31 to 40 years i.e. (31-40 years). The medium class has been further categorized on the basis of whether the person is a student or not. If he is a student then no more process is required, his income will obviously fall in the category of medium. But if he is not a student then the decision has to be made on the basis of his age. If his age is <=30 then the income will not be medium else it will be medium for the case if the age is 31-40. And if his age is >40 then the decision will be on the basis if he has fair CR or excellent CR as in the former case the income will be medium else not medium in the latter case. The low class has been first categorized on the basis of CR where fair CR means low income and in the case of Excellent CR the decision lies on the basis of age. The person with excellent CR and age >40 will not have low income but the person with the age 31-40 will have low income. This decision tree has been depicted in figure 4.5

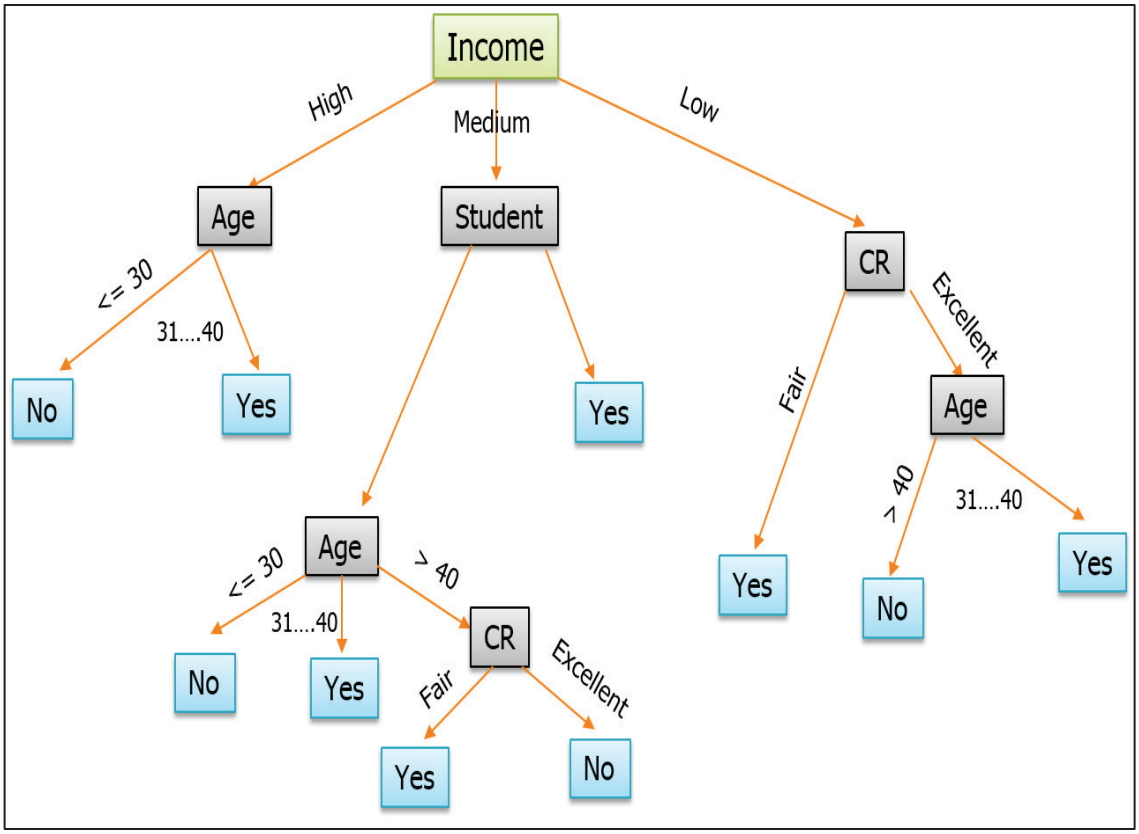

Fig 4.5: Decision Tree

4.3.4 Bagging

Some problems of pattern recognition are too complex to be solved by the use of a single classifier. Such kind of a situation occurs when the distribution of data is too complex or the other condition is when the data used is of high dimensionality i.e. it has a large number of attributes. In such scenarios, the collective output of the aggregated classifiers gives better results instead of a single classifier and it can be used so as to improve the accuracy of the model so formed. Although, it has been still not found out that what exactly the classifiers must consist of in order to generate the most effective ensemble but still, the approach of bagging is in huge practice as it increases the accuracy of the base classifier manifolds.

Bagging which stands for the process popularly known as Bootstrap Aggregation, combines the outcome of the normal or base classifiers while treating each distinct model equally with same weight to predict the final outcome. To improve the prediction results, each base classifier is trained using a randomly drawn sample set also called a bootstrap sample drawn from the original training set with replacement

[38]. It was created by Leo Breiman [41], used to combat the effects of over fitting and lessen the value of variance of the classification model [42]. Using an unstable learning algorithm would be to use bagging (e. g., neural networks or decision trees), result in largely different classifiers when small changes in the learning set [43]. It is a framework which assumes that the dataset has M instances and the procedure begins with the random generation of a training dataset. Then, a model which gives a class prediction is generated. This process is repeated several times where every time training dataset is generated with data replacement. The final output is achieved by majority voting of all the model predictions. There are two reasons for using bagging. The first is that the use of bagging seems to enhance accuracy when random features are used. The second is that bagging can be used to give ongoing estimates of the generalization error (PE∗) of the combined ensemble of trees, as well as estimates for the strength and correlation. It basically aims to increase the accuracy of the predictor by the generation of number of versions of a single predictor and then using these varying versions to create an aggregated predictor. The dataset which is to be used as the training dataset is drawn randomly from the main dataset by the implementation of replacements of instances present in the dataset. Then these training datasets which are also known as bags are used to train the classifier separately. The output so received is combined on the basis of majority voting which has been explained earlier. Majority voting is used in case the problem is of classification type. Average or weighted voting is used in case the dataset is of regression type. The single output so received acts as the final output which can be used for comparison purposes on the basis of accuracy, sensitivity and also specificity. This single output can be changed by varying the combination techniques for combining the outputs of the training datasets used. Thus we can say that bagging performs better than a single classifier because it combines the qualities of the single classifiers into one. It is easily implementable and does not require a lot of parameters to be tuned and this model performs better even if the data being used is noisy [41, 44]. The bagging framework has been precisely explained in figure 4.6.

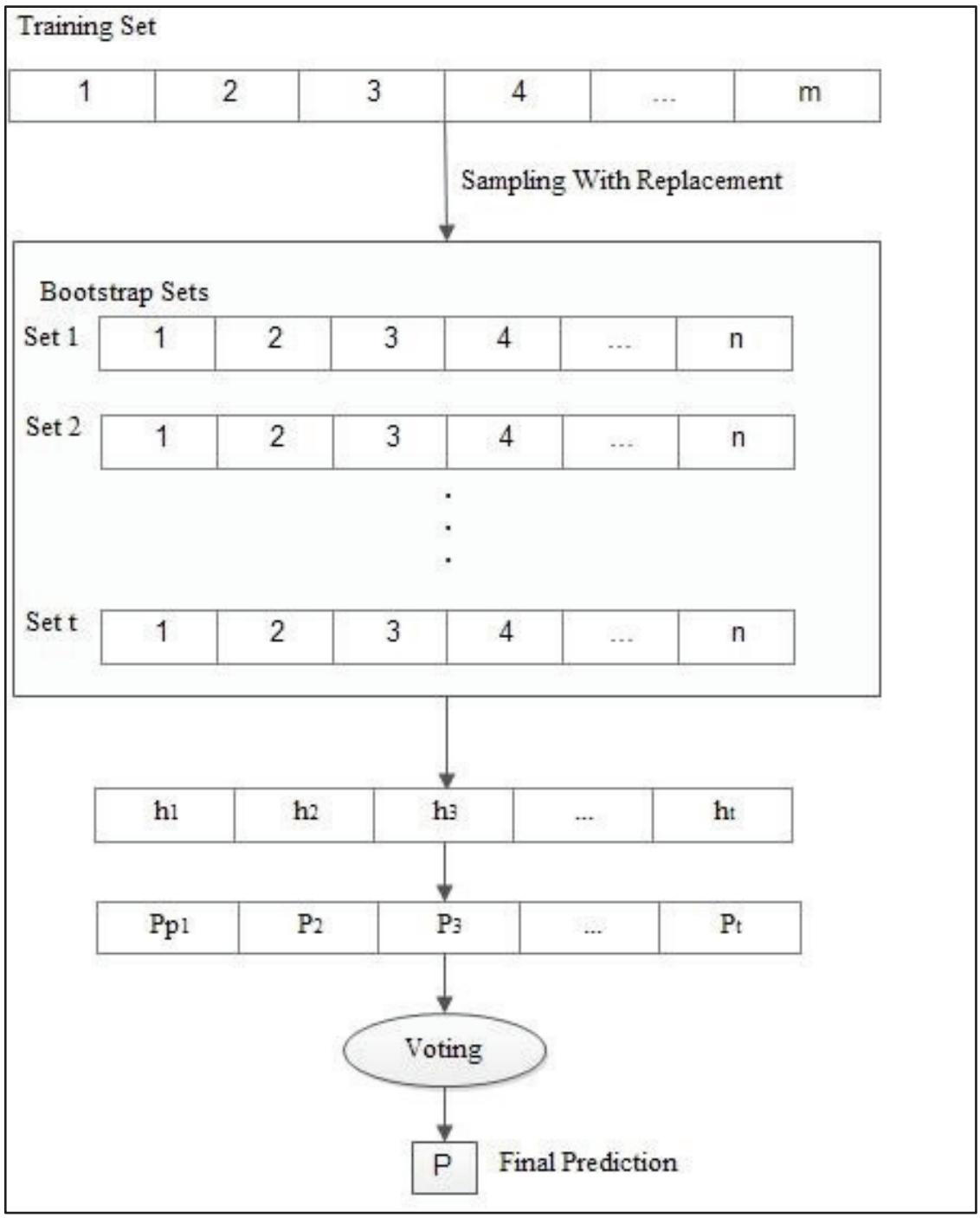

Fig 4.6: Bagging Framework

In this study bagging has been combined with the models namely SVM and GLM to increase the accuracy of the models already calculated under the bag free models category.

The algorithm for the process of bagging is: Taking the input as the dataset B having b training tuples and m being the number of models to be used for the bagging to

proceed, it gives the output in the form of an ensembled model which takes into account the classification model used in the process of bagging.

Algorithm: Bagging.

Input:

- B, a dataset consisting of b number of training tuples;
- m, the number of models which are to be incorporated in the ensemble;
- a classification learning scheme (neural network, decision tree algorithm, linear model, partial least squares etc.).

Output: The ensemble—a compound or aggregated model, D*.

Method to be followed:

for i=1 to m do // create m number of models:

create a bootstrap sample or bag, Bi , by the sampling the instances of B with replacement;

using Bi and the learning scheme to derive a model, Di ;

end for

4.3.5 Boosting:

Boosting was introduced by Schapire [45] to decrease error of a weak classifier by the process of iterative construction of another classifier being trained on the misclassifications of the weak classifier [45]. It is an iterative process of generating a strong classifier which contains series of weighted classifiers complementing one another. These base classifiers are trained on different subsets with the subsets being drawn deterministically from the original dataset.

The contrast between more than one methodologies is that in bagging the integrally built base model is left to risk, while in boosting we attempt to create reciprocal base models by learning resulting models, considering the missteps of past models. The methodology begins by learning the respectable starting point model on the whole learning set with similarly weighted cases. For the following base models, we need them to effectively anticipate the cases that have not been accurately anticipated by past base models. Subsequently, we increment the weights of these illustrations (or reduction the weights of the effectively anticipated cases) and take in another base model. We quit adapting new base models when some stopping standard is satisfied.

The emphasis which has to be laid on every instance is determined by a weight that is allocated to each instance in the training dataset chosen at every step [1]. Boosting [46] has a whole family of equal family members, such as winning, utilizing voting in favor of coalitions to join the forecasts of a base model learned by a solitary learning algorithm. The contrast in between two methodologies is that the built-in base models are dropped on the occasion of completing, while we try to model the supplementary model by learning further models, keeping in mind the mistakes of the previous model. With the same learning examples, learning a respectable starting display on the whole showing set begins the procedure. For the following base model, we need gauge the illustrations which have not been appropriately anticipated by past base models. Accordingly, we increment the heaviness of these cases (or shed pounds of exact prescient illustrations) and take in another base model. When some stop criteria are satisfied, we stop learning new base models. The main aim of boosting is to convert a weak classifier into a strong classifier. Figure 4 shows the procedure of boosting. In this work, for Boosting GBM (Generalized Boosted Regression Model) and GAMBoost (Generalized linear and additive models by likelihood based boosting) have been used.

- GBM: Fits generalized boosted regression models.
- GAMBoost: It is used to fit a generalized additive model on the basis of likelihood boosting. It is mainly suited for models with very large number of predictors having non-linear influence. It also provides smooth functional estimates of covariate influence functions combined with confidence bands and approximate degrees of freedom.

To evaluate the performance of all the models used, accuracy of each model in the terms of classification needs to be considered. This is done on the basis of the calculation of confusion matrix which helps out to find both the classification rate and the misclassification rate. After the model is generated, it is tested on the test dataset and on the basis of TP, FP, TN and FN *i.e.* true positives, false positives, true negative and false negatives respectively the accuracy is determined using the following formula:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}) \quad (8)$$

The boosting framework is shown in figure 4.7.

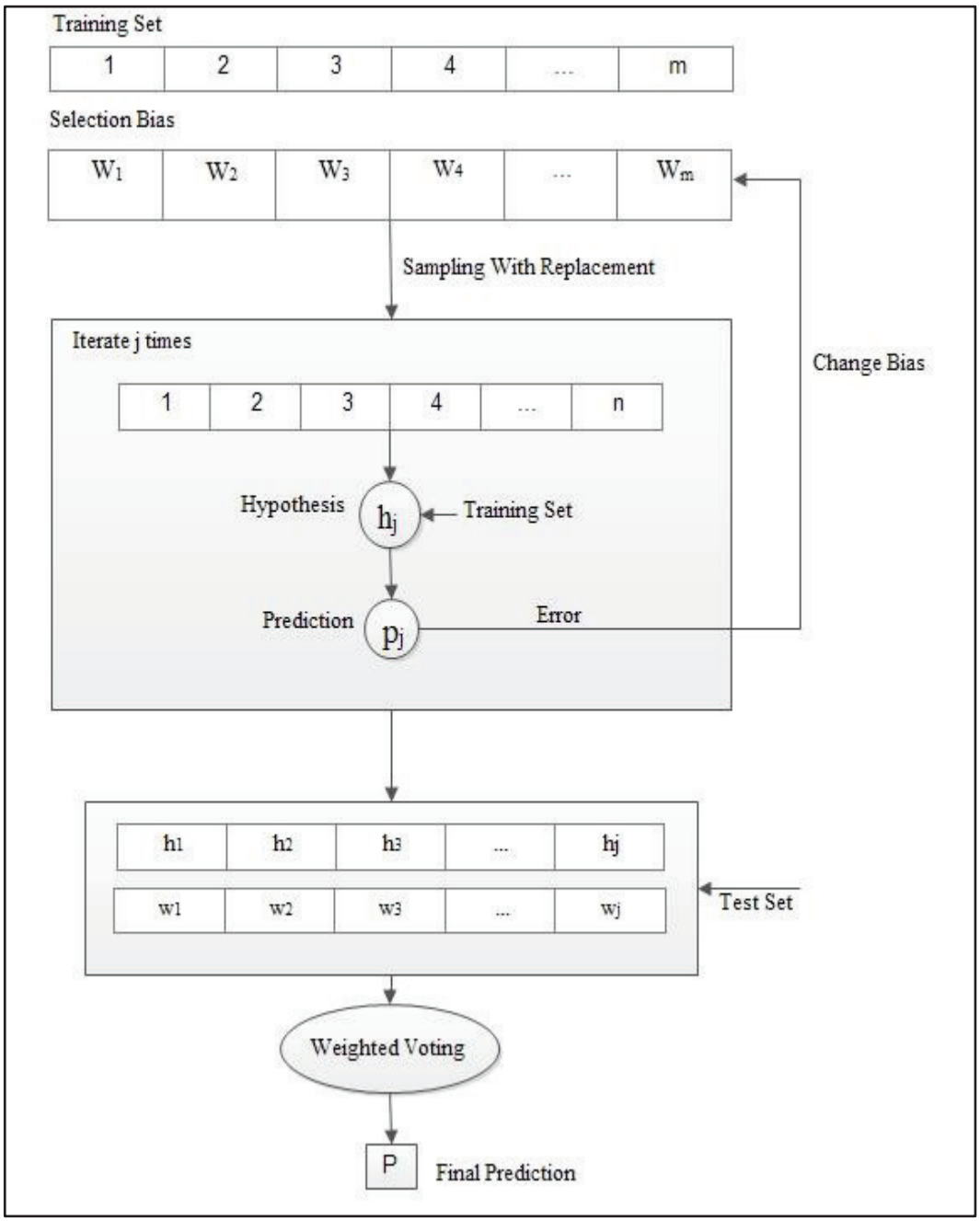

Fig 4.7: Boosting Framework

4.3.6 Hybrid approach

The proposed approach take into consideration 7 classifiers namely Support vector machine, Naïve Bayes, Decision tree, Partial Least Squares (PLS), Neural Network, Generalized Linear Model (GLM) and Optimal Weighted Neural Network (OWNN),

trained on the standard dataset and their accuracy is calculated by testing each model on the corresponding test dataset. The models so generated are then arranged according to their accuracy in ascending order. The models with the least accuracy are combined for the first level of the ensemble to give the Combination Result 1 which has the accuracy better than the models so used on the basis of Majority Voting. Then two new models are combined with Combination Result 1 to get Combination Result 2 again on the basis of Majority Voting and Combination Result 2 has accuracy greater than both the models used in the second level and Combination Result 1. Similarly the hybrid model is generated by combining Combination Result 2 and two new models. The output of the hybrid is tested against the test dataset and the accuracy is found to be greater than the accuracy of all the models used for its construction and the Combination Results.

The hybrid thus is called Hierarchical and Progressive Combination of Classifiers (HPCC). The layout of the ensemble is shown in the Fig.4.8.

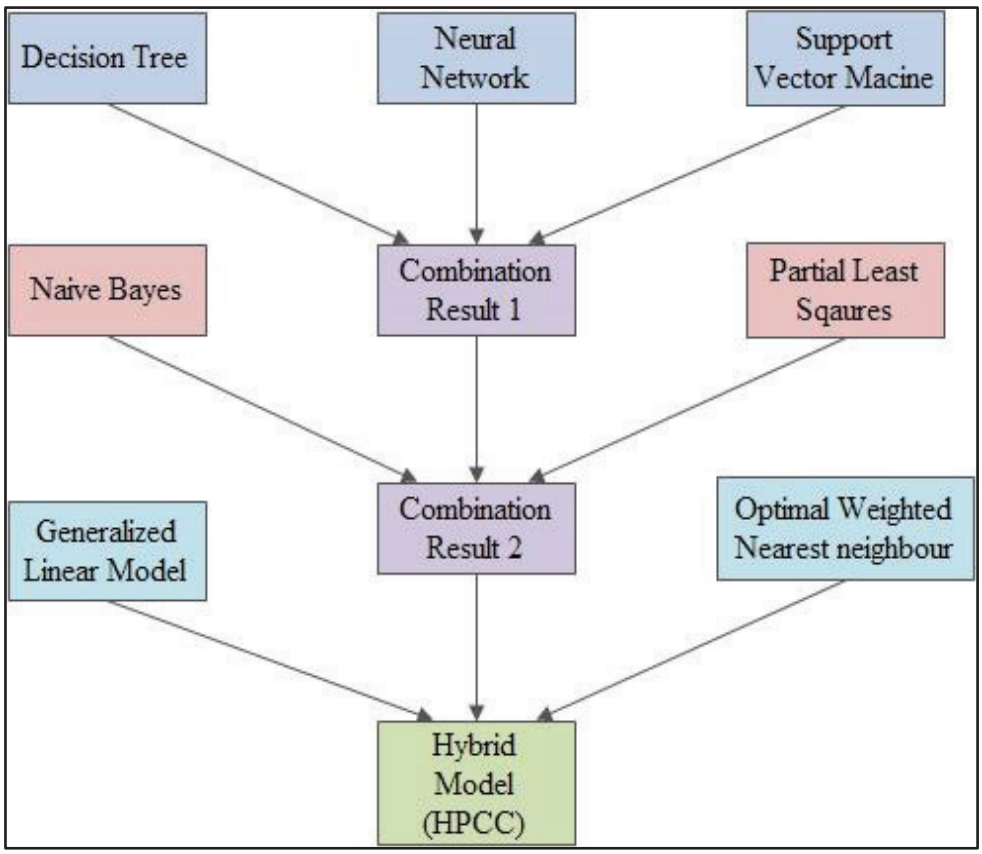

Fig 4.8: The hybrid model HPCC

The Models used in the ensemble have been explained as follows

For Level I-

a) **Decision Tree**: Discussed in Sub-Section C under Section 4.3.

b) **Support Vector Machine**: Discussed in Sub-Section B under Section 4.3.

c) **Neural Network**: This model was developed by imitating the idea behind the functionality of the nervous system of a human body where the neurons act as the nodes for the input of any nervous activity of the body. This system is used to calculate the function output when a large number of variables are given as an input to it. It can be used for both numerical and categorical kind of data. First comes number of hidden layers that are defined as the sum of number of output and input variables and this value is divided by two. For example in this dataset there are eight input variables and one output variable. Therefore the number of hidden layers for this dataset will be equal to four. The values of these layers is calculated from the value of the sigmoid function used which is given by the following formula

$$f(x) = \frac{1}{1+e^{x}} \tag{9}$$

where x is sum of product of numeric weights and input values.

Then, the value for the output layers is calculated the same way just taking the hidden layers as the input. The result is obtained by considering the highest value of the output layers. The structure for the algorithm of neural network has been shown in Fig. 4.9.

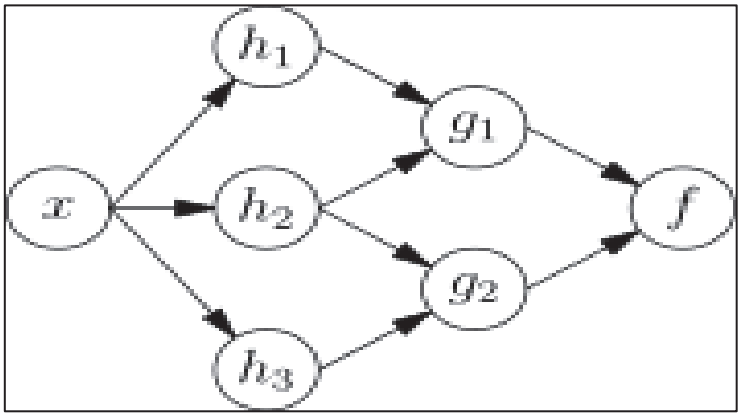

Fig 4.9: Neural Network

This figure shows the decomposition of function ***f*** on the basis of dependencies between the variables used which are indicated by the use of arrows. These process can be inferred in two distinct ways.

The first interpretation is known as the functional view where the input variable ***x*** is converted into a 3D vector named ***h***. ***h*** is then converted into a 2D vector named ***g*** and ***g*** is later converted into 1D functional value ***f***. This interpretation of neural network is frequently used in the context of the problems requiring optimization.

The second interpretation is known as the probabilistic view where the random variable **F=*f*(G)** is dependent on the random variable **G=*g*(H)** which in fact is dependent on **H=*h*(X)** which finally is dependent upon the random variable **X**. This interpretation is frequently used in the problems based on graphical models.

For Level II-

d) Naïve Bayes: It is an algorithm which is grounded on the simple Bayes' Theorem. Only the training dataset is required along with the features of the given class and there is no need of the complete covariance matrix as these two things are independent of each other [47]. It is based on the hypothesis that tells us that a particular person is suffering from a particular disease or not is completely independent of the feature space.

Let $class_a$ be the diabetes group with *a* risk factor that is the target class '*a*' and S be the set of input variables that are possessed by the instances of the dataset used in the model, assuming that all the variables are independent of each other. To predict the target class of diabetes risk factor of the person being examined, the model works as follows:

$$P(class_a|S) = \frac{P(S|class_a) \times P(class_a)}{P(S)} \quad (10)$$

where $P(class_a|S)$ is the posterior probability of that instance of the training dataset from the set of variables S that will occur to be $class_a$. $P(S|class_a)$ is the likelihood of a training data set of $class_a$ and variable S

where S is equal to S_1 union S_2 union ... union S_M. $P(class_a)$ is a probability of diabetes risk group a.

The above model can be written as formula in the following way:

$$P(class_a|S) = \frac{P(S_1|class_a)\times P(S_2|class_a)\times .....P(S_M|class_a)\times P(class_a)}{P_s} \quad (11)$$

e) **Partial Least Squares:** Partial least squares regression also known as PLS regression stands for the statistical method which resembles in some sorts to principal components regression which is also called PCR regression in which we find hyperplanes having the maximum variance between the independent variables and response, whereas in PLS we find a linear regression model by the projection of the observable variables and the predicted variables into a new space.

PLS is generally used to find out vital relationships between the two matrices (*A* and *B*). PLS model will try and find the multidimensionality in the direction of space of *A* that will be able to explain the maximum multidimensional variance in the direction of space of *B*.

f) **The Majority voting result of Level I which uses Decision tree, Support Vector Machine and Neural Network.**

For Level III-

g) **Generalized Linear Model**: Discussed in Sub-Section A under Section 4.3.

h) **Optimal Weighted Neural Network**: This is almost similar to the neural network with some modifications in it. A weighted ensemble is generated by using the set of independently trained statistical models that are often known by the name of base learners. The final predicted output of the ensemble is the linear aggregation of the outputs predicted by the individual models

$$Y_P(x) = \sum_{i=1}^{P} w_i y_i\,(x) \quad (12)$$

where P is the number of individual models, y_i is the prediction output of the i-th member, and w_i is a decreasing function of the prediction error of the i-th member over the whole training set. Thus, each ensemble member is weighted according to its individual performance.

i) **The Majority voting result of Level II which uses Partial Least Squares, Naïve Bayes and Majority Result of Level I.**

CHAPTER 5

RESULTS

For the analysis and the measurement of the performance of the classifiers used, *accuracy, sensitivity, and specificity* are used. They are used because these three measures are more useful in the medical field than any other criteria of measurement. For calculation of accuracy, specificity and sensitivity a confusion matrix is needed which has been explained further.

In a confusion matrix:

Actual class is the class to which the instance belongs in the original dataset.

Predicted class is the class to which the instance is classified by the algorithm used.

The confusion matrix has been shown in Table 5.1.

Table 5.1: Confusion Matrix

		ACTUAL CLASS	
		0	1
PREDICTED CLASS	0	True Positive(TP)	False Positive(FP)
	1	True Negative(TN)	False Negative(FN)

TP (True Positive) refers to the number of samples or instances which actually belong to class 0 and also have been correctly classified to class 0 itself.

TN (True Negative) refers to the number of samples or instances which actually belong to class 1 and also have been correctly classified to class 1 itself.

FN (False Negative) refers to the number of samples or instances which actually belong to class 0 but have been wrongly classified to class 1.

FP (False Positive) refers to the number of samples or instances which actually belong to class 1 but have been wrongly classified to class 0.

In this study, the hybrid approach has shown the maximum accuracy and has proved to be the best in the performance with regard to the prediction of the onset of

diabetes in Pima Indians of Arizona. This study has made the comparison amongst three bag free models which are Support Vector Machine (SVM), Decision Tree, Generalized Linear Model (GLM), two bagging models which are Bagging with GLM and Bagging with SVM, two boosting models GAMBoost and GBM and the Hierarchical and Progressive Combination of Classifiers HPCC which is the hybrid ensemble of 7 basic classifiers. The comparison has been made on the basis of the accuracy, sensitivity and specificity and the accuracy of each model or the ensemble so formed and it has been shown in the Table 5.2.According to the Confusion Matrix stated above, Accuracy is to be calculated as follows:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}) \quad (13)$$

Accuracy is the measure of how well the classifier is working in predicting the class/target value of the instance in the test dataset as compared to its actual value. Higher the accuracy, better the model is and in this case the hybrid has shown the highest accuracy. Accuracy is actually the weighted arithmetic mean of both Precision and Inverse Precision which are weighted by the Bias present and it can also said to be the weighted arithmetic mean of Recall and Inverse Recall which are weighted by amount of Prevalence present.

Table 5.2: Classifiers and their Accuracy

CLASSIFIER	ACCURACY
Support Vector Machine	74.32%
Decision Tree	76.19%
Generalized Linear Model	79.22%
Bagging with SVM	81.09%
Bagging with GLM	82.39%
GBM	82.16%
GAMBoost	82.43%
HPCC	83.34%

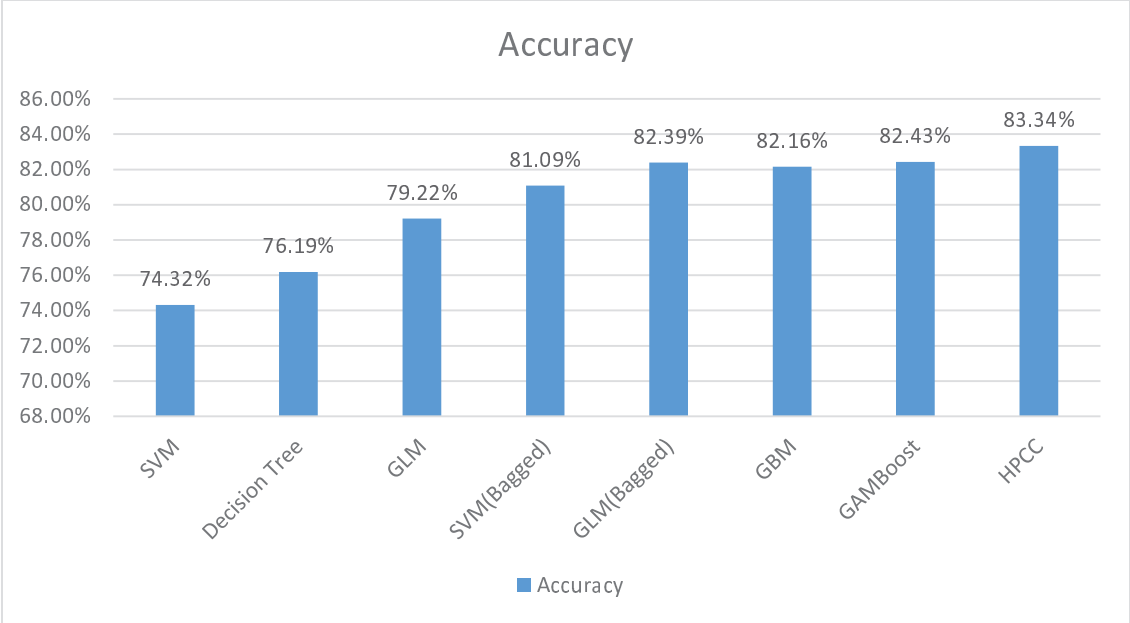

Fig 5.1: Graph of Accuracy

According to the Confusion Matrix stated above, Sensitivity is to be calculated as follows:

Sensitivity=TP/(TP+FN) (14)

Sensitivity here refers to the number of TP i.e. true positives when divided by the total number of instances that in fact are the instances belonging to the positive class which refers to the sum of false negatives and true positives, where false negatives are instances which actually belong to the negatives class but have been predicted to belong to the positive class by the classifier used.

Table 5.3: Classifiers and their Sensitivity

CLASSIFIER	SENSITIVITY
Support Vector Machine	0.769
Decision Tree	0.782
Generalized Linear Model	0.813
Bagging with SVM	0.832
Bagging with GLM	0.846
GBM	0.856

GAMBoost	0.853
HPCC	0.876

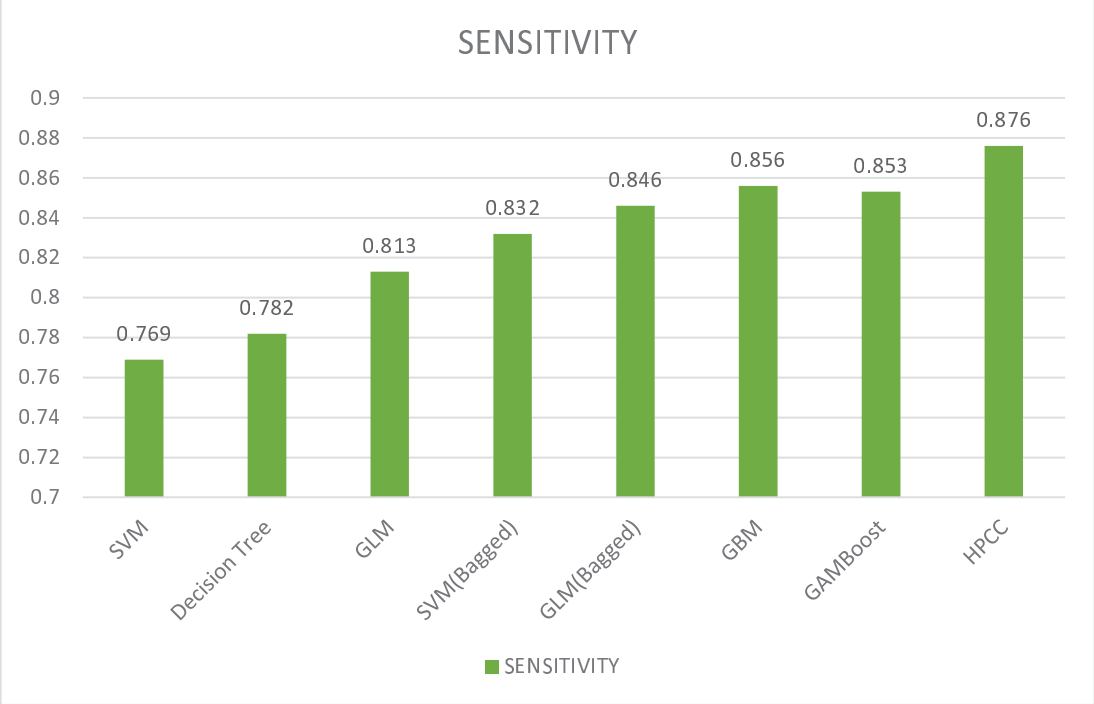

Fig 5.2: Graph of Sensitivity

Specificity is calculated by the formula:

Specificity=TN/(TN+FP) (15)

Specificity here refers to the number of TN i.e. true negatives when divided by the total number of instances that in fact are the instances belonging to the negative class which refers to the sum of false positives and true negatives, where false positives are instances which actually belong to the positive class but have been predicted to belong to the negative class by the classifier used.

Table 5.4: Classifiers and their Specificity

CLASSIFIER	SPECIFICITY
Support Vector Machine	0.732
Decision Tree	0.752

Generalized Linear Model	0.779
Bagging with SVM	0.808
Bagging with GLM	0.809
GBM	0.815
GAMBoost	0.810
HPCC	0.829

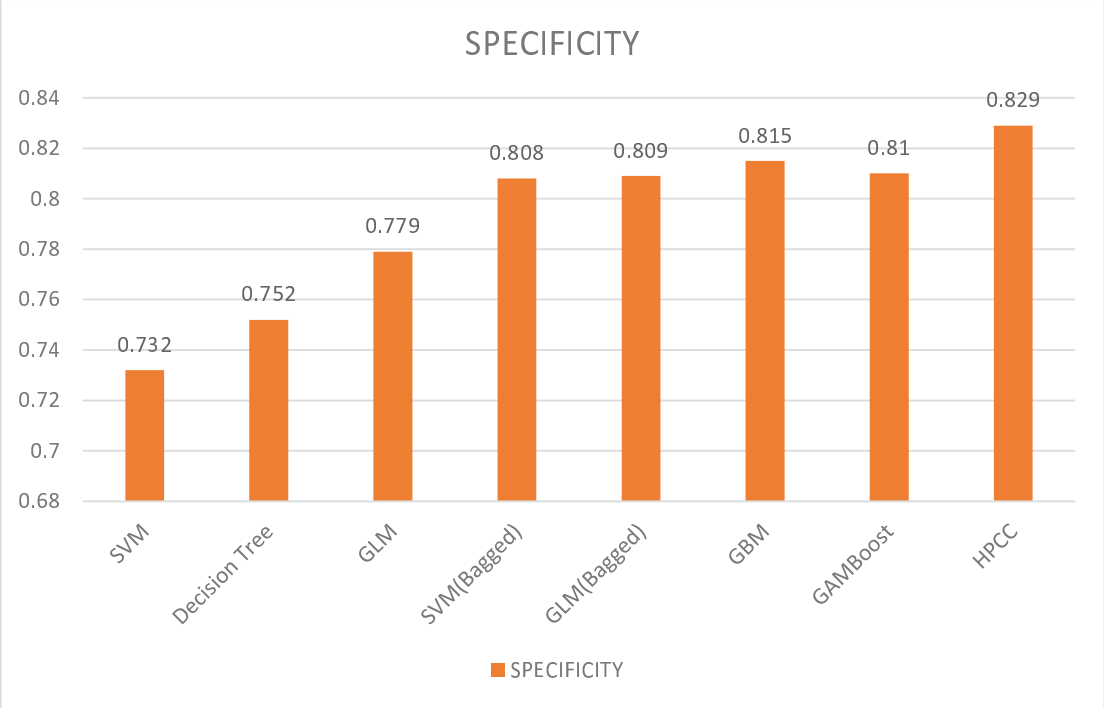

Fig 5.3: Graph of Specificity

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

This study can be concluded as the introduction of a novel approach which generates an ensemble of seven models having accuracy greater than all the other models it has been compared to. It has been shown that, by the use of the hybrid technique HPCC, it is possible to predict diabetes vulnerability in patients and that too with much reasonable accuracy than the other models used for comparison purposes. The proposed model has rendered high scalability and accountability and it is a far more robust technique than the earlier approaches used for the achievement of same kind of objective may it be in finding the vulnerability of heart disease or diabetes. Classifiers of such type are helpful in the early and correct detection of the onset of diabetes in the patients who are at a risk of it. On the basis of this detection the person concerned can be warned before-hand to change his/her lifestyle which will in turn prevent the patients from being affected by heart disorders like stroke and attack. These improvements will help lower the mortality rates and reduction in the costs of medical aid as well as health care of the state.

6.1 Future Scope

Future scope for the potential research is also discussed in a detailed way that certainly would pave a way for researchers in the future.

- This methodology can be further extended to predict and detect many more types of ailments such as thyroid, gall bladder malfunctioning, pregnancy complications etc.
- It can also be used to predict the onset of disorders which result from a particular disease like visual impairment in the case of persons suffering from diabetes or high chances of getting a stroke or an attack in case the person suffers from heart related disorders.

- This technique can further be expanded and enhanced, by the use of stacking as one of the ensembling techniques in order to increase the accuracy of the system so formed.
- Each individual model can be improved in its working by the use of feature selection which removes the unnecessary features thereby reducing the time of computation and increasing the accuracy of the system.

References

[1]. Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K., "Performance analysis of data mining classification techniques to predict diabetes", Procedia Computer Science, vol. 82, pp. 115-121, 2016.

[2]. Nai-arun, N., & Moungmai, R., "Comparison of classifiers for the risk of diabetes prediction", Procedia Computer Science, vol. 69, pp. 132-142, 2015. [3]. SATLEY M, The History of Diabetes, *Diabetes Health, vol*. 86, no. 1, pp.83-87, 2008.

[4]. World Health Organization, Diabetes Mellitus: Report of a WHO Study Group, *Geneva. World Health Org*, 1985

[5]. World Health Organization, Definition, diagnosis and classification of diabetes mellitus and its complications: report of a WHO consultation, *Definition, diagnosis and classification of diabetes mellitus and its complications: report of a WHO consultation* pp. 59-59, 1999.

[6]. Daneman D., "Type 1 diabetes", *The Lancet*, vol. 367, no. 9513, pp.847-858, 2006.

[7]. Hao, K., Di Narzo, A. F., Ho, L., Luo, W., Li, S., Chen, R., & Pasinetti, G. M., "Shared genetic etiology underlying Alzheimer's disease and type 2 diabetes", *Molecular aspects of medicine*, vol. 43, pp. 66-76, 2015.

[8]. DePaula, A. L., Macedo, A. L. V., Rassi, N., Machado, C. A., Schraibman, V., Silva, L. Q., & Halpern, A., "Laparoscopic treatment of type 2 diabetes mellitus for patients with a body mass index less than 35", *Surgical endoscopy*, vol. 22, no. 3, pp. 706-716, 2008.

Chicago [9]. Flegal, K. M., Ezzati, T. M., Harris, M. I., Haynes, M. G., Juarez, R. Z., Knowler, W. C., & Stern, M. P., "Prevalence of diabetes in Mexican Americans, Cubans, and Puerto Ricans from the Hispanic health and nutrition examination survey", *Diabetes care*, vol. 14, no. 7, pp. 628-638, 1991.

[10]. Merz, C. N. B., Buse, J. B., Tuncer, D., & Twillman, G. B., "Physician attitudes and practices and patient awareness of the cardiovascular complications of diabetes" *Journal of the American College of Cardiology*, vol. 40, no. 10, pp. 1877-1881, 2002.

[11]. "Diabetes" [Online] Available: "http://.diabetes.about.com/od/" [Accessed on 30th may, 2017].

[12]. Rubin, R. R., & Peyrot, M., "Quality of life and diabetes", *Diabetes/metabolism research and reviews*, vol. 15, no. 3, pp. 205-218, 1999.

[13]. Lin, E. H., Katon, W., Von Korff, M., Rutter, C., Simon, G. E., Oliver, M., & Young, B., "Relationship of depression and diabetes self-care, medication adherence, and preventive care", *Diabetes care*, vol. 27, no. 9, pp. 2154-2160, 2004.

[14]. Kovacs, M., Mukerji, P., Drash, A., & Iyengar, S., "Biomedical and psychiatric risk factors for retinopathy among children with IDDM", *Diabetes Care*, vol. 18, no. 12, pp. 1592-1599, 1995.

[15]. Wilkinson, G., Borsey, D. Q., Leslie, P., Newton, R. W., Lind, C., & Ballinger, C. B., "Psychiatric morbidity and social problems in patients with insulin-dependent diabetes mellitus", *The British Journal of Psychiatry*, vol. 153, no. 1, pp. 38-43, 1998

[16]. Patil MS., "Intelligent and Effective Heart Attack Risk Prediction from Heart Disease Warehouses Using Data Mining and Neural Networks", 2011

[17]. Hand DJ, Mannila H, Smyth P., "Principles of data mining", *MIT press*, 2001.

[18]. Gandhi, K. K., & Prajapati, N. B., "Diabetes prediction using feature selection and classification", International Journal of Advance Engineering and Research Development, 2014.

[19]. Priyadarshini, R., Dash, N., & Mishra, R., " A Novel approach to predict diabetes mellitus using modified Extreme learning machine", Electronics and Communication Systems (ICECS), 2014 International Conference, pp. 1-5, 2014.

[20]. Ibrahim, N. H., Mustapha, A., Rosli, R., & Helmee, N. H., "A hybrid model of hierarchical clustering and decision tree for rule-based classification of diabetic patients", International Journal of Engineering and Technology (IJET), vol. 5, no. 5, pp. 3986-91, 2013.

[21]. Aslam, M. W., Zhu, Z., & Nandi, A. K., "Feature generation using genetic programming with comparative partner selection for diabetes classification", Expert Systems with Applications, vol. 40, no. 13, pp. 5402-5412, 2013.

[22]. Rajput, R., Yadav, Y., Nanda, S., & Rajput, M., "Prevalence of gestational diabetes mellitus & associated risk factors at a tertiary care hospital in Haryana", *The Indian journal of medical research*, vol. 137, no.4, pp. 728, 2013.

[23]. Kumari, V. A., & Chitra, R., "Classification of diabetes disease using support vector machine", *International Journal of Engineering Research and Applications*, vol. 3, no. 2, pp. 1797-1801, 2013.

[24]. Aishwarya, R., & Gayathri, P., "A Method for Classification Using Machine Learning Technique for Diabetes", 2013.

[25]. Zolfaghari, R., "Diagnosis of diabetes in female population of pima indian heritage with ensemble of bp neural network and svm", *Int. J. Comput. Eng. Manag*, vol. 15, pp. 2230-7893, 2012.

[26]. Karthikeyani, V., Begum, I. P., Tajudin, K., & Begam, I. S., "Comparative of data mining classification algorithm (CDMCA) in diabetes disease prediction", *International Journal of Computer Applications*, vol. 60, no. 12, 2012.

[27]. Rajesh, K., & Sangeetha, V. "Application of data mining methods and techniques for diabetes diagnosis", *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 2, no. 3, 2012.

[28]. Balakrishnan, S., Narayanaswamy, R., & Paramasivam, I., "An empirical study on the performance of integrated hybrid prediction model on the medical datasets", *International Journal of Computer Applications*, vol. 29, no. 5, pp. 1-6, 2011.

Chicago [29]. Çalişir D, Doğantekin E., "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier", *Expert Systems with Applications*, vol. 38, no. 7, pp. 8311-5, 2011.

[30]. Shouman, M., Turner, T., & Stocker, R., "Using data mining techniques in heart disease diagnosis and treatment", *Electronics, Communications and Computers (JEC-ECC)-2012 Japan-Egypt Conference* pp. 173-177, 2012.

[31]. Porter, T., & Green, B., "Identifying diabetic patients: a data mining approach", *AMCIS 2009 Proceedings*, 2009.

[32]. Han, J., Rodriguez, J. C., & Beheshti, M., "Diabetes data analysis and prediction model discovery using rapidminer", *Future Generation Communication and Networking, 2008. FGCN'08*, vol. 3, pp. 96-99, 2008

[33]. Kahramanli H, Allahverdi N., "Design of a hybrid system for the diabetes and heart diseases", *Expert systems with application,* vol. 35, no. 1, pp. 82-90, 2008.

[34]. Kayaer, K., & Yıldırım, T., "Medical diagnosis on Pima Indian diabetes using general regression neural networks", *Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP)*, pp. 181-184, 2003.

[35]. Carpenter, G. A., & Markuzon, N., "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases", *Neural Networks*, vol. 11, no. 2, pp. 323-336, 2012.

[36]. "Dataset" [Online] Avaialbale: "https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes" [Accessed on 3rd Feb, 2017].

[37]. Michalski, R. S., & Kaufman, K. A., "Data mining and knowledge discovery: A review of issues and a multistrategy approach", 1997.

[38]. Bashir, S., Qamar, U., & Khan, F. H., "BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting", Australasian physical & engineering sciences in medicine, vol. 38, no. 2, pp. 305-323, 2015.

[39]. Nai-arun, N., & Moungmai, R., "Comparison of classifiers for the risk of diabetes prediction", Procedia Computer Science, vol. 69, pp. 132-142, 2015.

[40]. Ali, J., Khan, R., Ahmad, N., & Maqsood, I., "Random forests and decision trees", *IJCSI International Journal of Computer Science Issues*, vol. 9, no. 5, pp. 272-278, 2012.

[41]. Breiman, L., "Bagging predictors", Machine learning, vol. 24, no. 2, pp. 123-140, 1996.

[42]. Nai-Arun, N., & Sittidech, P., "Ensemble Learning Model for Diabetes Classification", Advanced Materials Research, vol. 931, pp. 1427-1431, 2014.

[43]. Dietterich, T. G., "Ensemble methods in machine learning", *Multiple classifier systems*, vol. 1857, pp. 1-15, 2000.

[44]. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J., "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann, 2016.

[45]. Schapire, R. E., "The boosting approach to machine learning: An overview", *Nonlinear estimation and classification*, pp. 149-171, 2003.

[46]. Freund, Y., & Schapire, R. E., "Experiments with a new boosting algorithm", *icml*, vol. 96, pp. 148-156, 1996.

[47]. Palaniappan S, Awang R, "Intelligent heart disease prediction system using data mining techniques", *International conference on computer system and applications(AICCSA)*, pp. 108–115, 2008.

Research Publication

H. Kaur and S. Batra, "HPCC: An Ensembled Framework for the Prediction of the onset of Diabetes", *International Conference on Signal Processing, Computing and Control (ISPCC 2017)*.

[ACCEPTED]

VIDEO URL

HPCC: An Ensembled Framework for the Prediction of the Onset of Diabetes-
https://www.youtube.com/watch?v=TUAm0BM_i8I

Harnoor

ORIGINALITY REPORT

%18	%11	%10	%9
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.cs.bham.ac.uk Internet Source	%1
2	pearson.com.cn Internet Source	%1
3	Bashir, Saba, Usman Qamar, and Farhan Hassan Khan. "BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting", Australasian Physical & Engineering Sciences in Medicine, 2015. Publication	%1
4	Submitted to King Mongkut's University of Technology Thonburi Student Paper	%1
5	www.slideshare.net Internet Source	<%1
6	research.ijcaonline.org Internet Source	<%1
7	Harald Romsdorfer. "Speech prosody control	

	using weighted neural network ensembles", 2009 IEEE International Workshop on Machine Learning for Signal Processing, 09/2009 Publication	<%1
8	Calisir, D.. "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier", Expert Systems With Applications, 201107 Publication	<%1
9	Lewis, Nathan E., Neema Jamshidi, Ines Thiele, and Bernhard Ø. Palsson. "Metabolic Systems Biology", Encyclopedia of Complexity and Systems Science, 2009. Publication	<%1
10	en.wikipedia.org Internet Source	<%1
11	citeseerx.ist.psu.edu Internet Source	<%1
12	spikelab.jbpierce.org Internet Source	<%1
13	Submitted to Asian Institute of Technology Student Paper	<%1
14	www.bioinformatics.ege.edu.tr Internet Source	<%1
15	pythia.inf.brad.ac.uk Internet Source	<%1

16	Submitted to Victoria University Student Paper	<%1
17	ijact.in Internet Source	<%1
18	Parthiban, G. and Srivatsa, S. K.. "Comparing Naive Bayes and Decision Tree Techniques for Predicting the Risk of Diabetic Retinopathy", International Journal of Applied Engineering Research, 2015. Publication	<%1
19	Xiao Fang. "Are You Becoming a Diabetic? A Data Mining Approach", 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 08/2009 Publication	<%1
20	Submitted to CSU, Dominguez Hills Student Paper	<%1
21	Submitted to University of Maryland, University College Student Paper	<%1
22	Submitted to SASTRA University Student Paper	<%1
23	www.archive.org Internet Source	<%1
	Submitted to BITS, Pilani-Dubai	

24	Student Paper	<%1
25	www.biharanjuman.org Internet Source	<%1
26	Submitted to EDMC Student Paper	<%1
27	Submitted to Institute of Graduate Studies, UiTM Student Paper	<%1
28	www.marshall.edu Internet Source	<%1
29	I Dooley. "Changes in intraocular pressure and anterior segment morphometry after uneventful phacoemulsification cataract surgery", Eye, 02/19/2010 Publication	<%1
30	docslide.us Internet Source	<%1
31	www.i-scholar.in Internet Source	<%1
32	ira.lib.polyu.edu.hk Internet Source	<%1
33	Bashir, Saba, Usman Qamar, and Farhan Hassan Khan. "A Multicriteria Weighted Vote-Based Classifier Ensemble for Heart Disease	<%1

Prediction : A Novel Ensemble for Heart Disease Prediction", Computational Intelligence, 2015.
Publication

34	Ahmed, Ibrahim M., Marco Alfonse, Mostafa Aref, and Abdel-Badeeh M. Salem. "Reasoning Techniques for Diabetics Expert Systems", Procedia Computer Science, 2015. Publication	<%1
35	Submitted to Higher Education Commission Pakistan Student Paper	<%1
36	Submitted to Study Group Australia Student Paper	<%1
37	Peña-García, Antonio, Rocío de Oña, Pedro Antonio García, and Juan de Oña. "Personal factors influencing the visual reaction time of pedestrians to detect turn indicators in the presence of Daytime Running Lamps", Ergonomics, 2016. Publication	<%1
38	www.razorrobotics.com Internet Source	<%1
39	Amato, Umberto, Anestis Antoniadis, and Italia De Feis. "Additive model selection", Statistical Methods & Applications, 2016. Publication	<%1

40	Bashir, Saba, Usman Qamar, and M. Younus Javed. "An ensemble based decision support framework for intelligent heart disease diagnosis", International Conference on Information Society (i-Society 2014), 2014. Publication	<%1
41	Submitted to Universiti Sains Malaysia Student Paper	<%1
42	www.airccse.org Internet Source	<%1
43	Illhoi Yoo. "Data Mining in Healthcare and Biomedicine: A Survey of the Literature", Journal of Medical Systems, 05/03/2011 Publication	<%1
44	Submitted to Universiti Teknologi Malaysia Student Paper	<%1
45	Submitted to National College of Ireland Student Paper	<%1
46	Submitted to West Cheshire College Student Paper	<%1
47	Submitted to University of Nottingham Student Paper	<%1
48	yourhealth-check.com Internet Source	<%1

49	Chetty, Naganna, Kunwar Singh Vaisla, and Nagamma Patil. "An Improved Method for Disease Prediction Using Fuzzy Approach", 2015 Second International Conference on Advances in Computing and Communication Engineering, 2015. Publication	<%1
50	www.pitara.com Internet Source	<%1
51	ijceronline.com Internet Source	<%1
52	stanford.wellsphere.com Internet Source	<%1
53	www.inderscience.com Internet Source	<%1
54	Li, Jing-Song, Hai-Yan Yu, and Xiao-Guang Zhang. "Data Mining in Hospital Information System", New Fundamental Technologies in Data Mining, 2011. Publication	<%1
55	Alireza Kajabadi. "Data mining cardiovascular risk factors", 2009 International Conference on Application of Information and Communication Technologies, 10/2009 Publication	<%1

56	fumblog.um.ac.ir Internet Source	<%1
57	Muntean, Maria, Honoriu Vălean, Adrian Tulbure, Ioan Ileană, Manuella Kadar, and George Caruntu. "", Advanced Topics in Optoelectronics Microelectronics and Nanotechnologies V, 2010. Publication	<%1
58	Submitted to Cranfield University Student Paper	<%1
59	Submitted to Corinthian Colleges Student Paper	<%1
60	Submitted to VIT University Student Paper	<%1
61	evidenceframework.org Internet Source	<%1
62	www.authorstream.com Internet Source	<%1
63	Submitted to University of Greenwich Student Paper	<%1
64	howtopreventheartdisease.blogspot.com Internet Source	<%1
65	dmiftp.uqtr.ca Internet Source	<%1

66	Abdel-Aal, R.E.. "Improved classification of medical data using abductive network committees trained on different feature subsets", Computer Methods and Programs in Biomedicine, 200511 Publication	<%1
67	Jianchao Han. "Discovering Decision Tree Based Diabetes Prediction Model", Communications in Computer and Information Science, 2009 Publication	<%1
68	ijettcs.org Internet Source	<%1
69	Submitted to University of South Australia Student Paper	<%1
70	Zanin, M., D. Papo, P.A. Sousa, E. Menasalvas, A. Nicchi, E. Kubik, and S. Boccaletti. "Combining complex networks and data mining: Why and how", Physics Reports, 2016. Publication	<%1
71	www.type2diabetesadvice.co.uk Internet Source	<%1
72	Du, Pufeng, and Chao Xu. "Predicting multisite protein subcellular locations: progress and challenges", Expert Review of Proteomics, 2013.	<%1

Publication

73	Altana, C., A. Muoio, F. Schillaci, G A P. Cirrone, G. Lanzalone, S. Tudisco, F. Brandi, G. Cristoforetti, P. Koester, L. Fulgentini, L. Labate, and L. A. Gizzi. "Thomson parabola spectrometer: A powerful tool for on-line plasma analysis", 2015 4th International Conference on Advancements in Nuclear Instrumentation Measurement Methods and their Applications (ANIMMA), 2015. Publication	<%1
74	renoir.villanova.edu Internet Source	<%1
75	Antonio J. Rivera. "A study on the medium-term forecasting using exogenous variable selection of the extra-virgin olive oil with soft computing methods", Applied Intelligence, 04/06/2011 Publication	<%1
76	Banu, R. Karthiya, and R. Ravanan. "Analysis of e-learning in data mining — A dreamed vision for empowering rural students in India", 2011 International Conference on Recent Trends in Information Technology (ICRTIT), 2011. Publication	<%1
77	Submitted to University of Wales Swansea	

	Student Paper	<%1
78	Submitted to University of Sheffield Student Paper	<%1
79	Bioinspired Smell and Taste Sensors, 2015. Publication	<%1
80	Submitted to The Nelson Mandela Africa Institution of Science and Technology Student Paper	<%1
81	Submitted to Mugla University Student Paper	<%1
82	hera.ugr.es Internet Source	<%1
83	crm.ittoolbox.com Internet Source	<%1
84	Submitted to John F. Kennedy Memorial High School Student Paper	<%1
85	Submitted to King's College Student Paper	<%1
86	www.slidesearch.org Internet Source	<%1
87	cran.espol.edu.ec Internet Source	<%1

88	Weizhong Yan. "Application of Random Forest to Aircraft Engine Fault Diagnosis", The Proceedings of the Multiconference on "Computational Engineering in Systems Applications", 10/2006 Publication	<%1
89	www.sciencedirect.com Internet Source	<%1
90	Kahramanli, H.. "Design of a hybrid system for the diabetes and heart diseases", Expert Systems With Applications, 200807 Publication	<%1
91	Carbonell, Jaime G., Ryszard S. Michalski, and Tom M. Mitchell. "An Overview of Machine Learning", Machine Learning, 1983. Publication	<%1
92	www.daneshbod.com Internet Source	<%1
93	www.uptodate.com Internet Source	<%1
94	pdfs.semanticscholar.org Internet Source	<%1
95	www.jove.com Internet Source	<%1
96	www.nsclass.ca	

Internet Source

<%1

97	Kandhasamy, J. Pradeep, and S. Balamurali. "Performance Analysis of Classifier Models to Predict Diabetes Mellitus", Procedia Computer Science, 2015. Publication	<%1
98	Lecture Notes in Computer Science, 2012. Publication	<%1
99	www.ijeat.org Internet Source	<%1
100	CIRP Encyclopedia of Production Engineering, 2014. Publication	<%1

EXCLUDE QUOTES	OFF	EXCLUDE MATCHES	< 8 WORDS
EXCLUDE BIBLIOGRAPHY	ON		