A-Semi Automatic Decision Support System for Predicting Heart Failure Using Advanced Classification Techniques

By

Muhammad Subhan Khan

NUST201362650MMCS25113F

A thesis submitted to the faculty of Information Security Department, Military College of Signals, National University of Sciences and Technology, Rawalpindi in partial fulfilment of the requirements for the degree of MS in Computer Software Engineering

August 2017

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS Thesis written by NS **Muhammad Subhan Khan** Registration No.**NUST201362650MMCS25113F**, of **Military College of Signals** has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: _____

Name of Supervisor_____

Date: _____

Signature (HoD):_____

Date: _____

Signature (Dean/Principal):_____

Date: _____

Declaration

I certify that this research work titled "*A-Semi Automatic Decision Support System for PredictingHeart Failure Using Advanced Classification Techniques"*is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Dedication

Dedicated to my exceptional parents, my kindhearted uncle, my encouraging wife and adored siblingswhose tremendous support and cooperation led me to this wonderful accomplishment.

Acknowledgements

I am thankful to my Creator Allah Subhana-Watala to have guided me throughout this work at every step and for every new thought which You setup in my mind to improve it. Indeed I could have done nothing without Your priceless help and guidance. Whosoever helped me throughout the course of my thesis, whether my parents or any other individual was Your will, so indeed none be worthy of praise but You. Each and every grain of knowledge that I have is from you, I myself is not even capable to write a single word without Your Will.

I am profusely thankful to my beloved parents Mr. & Mrs. Muhammad Naeem Khan who raised me when I was not capable of walking and continued to support me throughout in every department of my life. I am also thankful to my uncle Mr. Muhammad Naveed Khan who throughout my career and studies stayed very supporting and helping. I amthankful to my encouraging wife Tabinda Subhanand my adored siblingsMuhammad Furqan Khan and Hadia Khan for always motivating me in my life and achieving more than I am capable of.

I would also like to express special thanks to my supervisor Dr. Hammad Afzal for his help throughout my thesis and also for Advance Databases course which he has taught me. I can safely say that I haven't learned any other engineering subject in such depth than the ones which he has taught. His support for me throughout my thesis was incomparable.

I would also like to thank Dr.Naima Iltaf for being on my thesis guidance and evaluation committee member.I would also like to specially thank Mr. Athar Mohsin for being a helping instructor and guiding me during my stay here at Military College of Signalsand I would also like to express my special thanks to Ms. Ayesha Naseer for her help.

Finally, I would like to express my gratitude to all the fellows and family members who have rendered valuable assistance to my study.

Abstract

Heart failure is considered as one of the major reasons of death worldwide. The amount of deaths from heart failure surpasses the amount of deaths resulting from any other causes. This point makes heart disease as one of the most dangerous disease resulting in human deaths worldwide. Latest studies have concentrated on the usage of machine learning and data mining techniques to build predictive models that are capable to forecast the occurrence of heart failure. The presence of medical data leads to the need of smart data mining tools in order to extract valuable knowledge. Scientists have been using several statistical analysis and data mining techniques to increase the disease diagnosis accuracy in medicinal healthcare. Numerous researchers have presented various data mining techniques for heart disease diagnosis. Using a single data mining technique shows an acceptable level of accuracy for disease prediction. In recent times, more investigation is carried out in the direction of hybrid models which demonstrate incredible enhancement in cardiac vascular disease prediction accuracy. The purpose of the suggested research is to predict the heart disease in a patient more accurately. Therefore, we have suggested a Decision Support System'sframework to help the medical practitioners, doctors and decision makersto collect and understand information and construct a ground work for effective decisionmaking. The suggested framework will play a constructively significant role in medical field and hence likely to increase the quality of medication.The suggested model will overcome the conventionalperformance limitations by applying an ensemble of three diverse classifiers (SVM, LR and NB). Efficiency of the suggested ensemble technique is examined by comparison of results withnumerous renowned classifiers as well as ensemble methods. The experimental assessment expresses thatthe suggested framework dealt with all sorts of attributes and accomplished greaterprediction accuracy. For analysis, the data sets (Statlog and SPECTF) are collected from UCI data repository [14] and [25].

Key Words:*Naïve Bayes, Support Vector Machine, Linear Regression, Machine Learning,Ensemble, Split Validation, Heart Disease Prediction, Statlog, SPECTF*

Table of Contents

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

Machine learning in medical field is a procedure of identifying unseen patterns and knowledge from medical data; analyses the information and use the information to predict disease in a patient. [31].The fundamental purpose of data mining and machine learning is to mine unseen knowledge from medical data and organize it into understandable information [17]. Data mining techniques offers a variety of predictive models that can be utilized to predict and classify medical diseases. Once discovering knowledge from the medical data is completed, the learning stage begins; where a predictive model is constructed. This learning process develops the idea of machine learning and can be defined as “the complex computation process of automatic pattern recognition and intelligent decision makingbased on training sample data” [32]. Furthermore, the machine learning classifiers are characterized as supervised and unsupervised classifiers conditional to the usefulness of data. In supervised learning, training data must consists of a label class and a model is created on it. Few examples are Naïve Bayes (NB), Support Vector Machine and (SVM) etc. In unsupervised classifiers, sample data set does not contain any such thing as class label e.g. K-mean clustering and Self-Organization Map (SOM).

Analysis of medical data is considered as a complex job that needs to be performed accurately and proficiently. Inaccurate results can lead to disastrous situations where the quality of medication is compromised. According to the study held at World Health Organization, every year around 10 million people expire due to heart disease. Apart from that, heart attacks are the main reason for death in many developed countries like USA as around 5 million patients in

America are suffering from heart diseases. Such large numberof deaths make heart disease as one of the most hazardous disease in the world [1].

Cardiac Vascular or heart disease is the leading reason of demise for both males and females, with more than fifty percent of the expirieshappening in males. One out of every four persons is suffering with and passes away from heart disease, and in the United States of America alone, around0.6 million Americans goes in the mouth of death annually [40].

Heart attack or cardiac arrest is a serious ailment and it's one of the primary sources of death globally [1]. Throughout the world, there are more than 25 million adults who are suffering from cardiac vascular diseases. A huge percentage of heart patients die within the first year of disease, therefore timely and accurate prediction of heart disease in earlier stages is very essential [2]. Heart failure is also the key reason for numerous hospitalizations among the patients who are more than 65 years old, in such age heart disease is considered very fatal [3].

There are several symptoms associated to cardio vascular diseases. Some patients have severe chest ache while some feel exhaustion. Patients feel uneasiness, heaviness, and immensity in their arm or beneath the breastbone. Sweating, queasiness or faintness are also considered some of usual symptoms of heart disease. However around 50 percent of the heart patients do not feel any significant change in their health till a cardiac arrest happen. According to another study held by World Health Organization, unusual behavior and persistent disability is the main outcomes of heart disease [1].

Timely Diagnosis of heart disease has a vital influence on the safety of patient, as it can progress towards an efficient and well organized treatment before any severe condition occurs. On the other hand, diagnosis of heart disease in advanced phases can seriously affect the health of

patient and can lead to devastating situations. In alter stages the cost of medication is also increased with the risks of its treatment [2].

It has been observed that a large number of research studies has been conducted in medical field and in particular for heart disease. But there is no final verdict on which study is supposed to be the best among them. So there is a continuous need and room of improvement in models and frameworks that previously exist.This ambiguity about which model symbolizes an ideal way out has been handled in this study by presenting a newensemble,which is capable to utilize the capabilities of numerous classifiers.Due to the availability of large amount of data for heart disease, there is a requirement of an intelligent decision support system for predicting heart disease to extract meaningful information.This study aims the same to provide better and accurate result for heart disease data sets.

Machine learning in medical field is a struggle towards making error free decision for medical practitioners thus increasing patient's health security. Disastrous conditions can occur if the syndrome is not detected in a timely and correct manner. Furthermore, an effective dealing with the ailment can decrease medical and medication blunders as well as they can also reduce the cost of treatment at the later stages of disease. Hence an accurate, efficient and reliable decision support system is the need of hour in the field of medical diagnosis [22].

So keeping these circumstances in mind, data mining has a key role in the domain of heart disease diagnosis. Classification techniques can be used to identify the unseen patterns from existing medical records to categorize healthy and heart disease patients. [11].Considering the

treatment of patients, the prediction of heart disease is very important. This research study reveals a framework for heart disease prediction which uses different data mining practices like Support Vector Machine (SVM), Naïve Bayes and Linear Regression .Data normalization is performed for specific attributes in the prepressing stage and also some specific attributes are selected to achieve better results. The possibility of having a heart disease is calculated using various attributes like Age, Chest Pain Type, Sex, Serum Cholesterol, Maximum heart rate and slope etc.

The rest of the document is ordered as follows: Chapter 2 narrates to literature review. Chapter 3 relates the Data Set(s) Description. Chapter 4 represents the Proposed Framework. In Chapter 5,we discussed the Suggested Technique. While,Chapter 6 is related Results and Findings when we applied the framework on different heart data sets. And finally, in Chapter 7 the document concludes with Conclusion and future Work.

CHAPTER 2: LITERATURE REVIEW

In the past, researchers have applied various data mining approaches for improved and accurate prediction of heart disease to assist the healthcare experts. In this section we will discuss some of these studies from the past. A lot of these techniques and approaches are based on hybrid ensemble approaches which is a trend in machine learning for medical sciences.

Shadab et al. [4] presented a web based application for the prediction of heart disease. The suggested algorithm is constructed on Naïve Bayes that reveals the unseen patterns in the specified data set. The benefit of using Naïve Bayes is that it uses very few records of data set for training purposes and ignore related attributes in calculating the probability of occurrence of a disease in a patent.

Chen et al. [5] presented an efficient and accurate system, named as 'Heart Disease Prediction System (HDPS)' for the diagnosis of cardiac diseases. Machine Learning and Statistics are the key methods that are utilized in the suggested framework. Learning Vector Quantization (LVQ) is used for training purposes of data set while ROC curve is used to verify the accuracy of the final outcomes. The suggested framework shows 80% accuracy in the concluding results. From critique point of view, the study uses LVQ approach but it does not include any particulars regarding the implementation. Similarly, the study also lacks the association of this research with the related previous studies to evidently recognize its advantages over the other frameworks.

Pandey et al. [6] used J48 classification algorithm to create a model heart disease prediction. They used the holdout validation technique to divide the data set into training and testing data sets. Final results exhibited that the accuracy of 75.73 % was achieved by pruning the J48

decision tree classification algorithm. They performed experiments using the UCI datasets available at [8]

Bashir et al. [7] suggested an ensemble base classifier which used a combination of three classifiers namely Support Vector Machine, Naïve Bayes and Decision tree for the prediction of cardiac disease. Every classifier was trained separately and then the prediction was made by merging the results from all classifiers depending on the majority voting. The accuracy of this technique was recorded as 81.82 %.

Uppin et al. [8] predicted heart disease existence by executing C4.5 decision tree algorithm. Attributes were filtered in this approach and seven out of thirteen attributes were used for classification purposes. The overall accuracy of this classifier was listed as 85.96 %.Sotelsek-Margalefet al. [19] suggested an expert system titled as MIDAS for medicinalanalysis from patient's histories. The system is centred on dataextraction and machine learning methodologies which practisesformerly diagnosed patient's record for forthcomingdiagnosis. The Natural Language Processing (NLP) methodshave been applied for purpose of classification of textual data. The MIDAS is regarded as theforemost expert system in the pastutilized for septicblood disease analysis. The data set is acquired from the Cincinnati Children's Hospital and contains proper characterizedclasses. Weka tool is utilized in amalgamation with C4.5 decision tree algorithm and k-Nearest-Neighbour machine learning classifier. The results indicated a higher accuracy headed towardsmedicinal disease diagnosis. The valuation of suggested framework is completed on the foundation of ICD-code estimation accuracy.

Mahmood et al. [9] suggested a novel pruning approach with the intention to increase the accuracy of heart disease prediction. An amalgamation of pre and post pruning was utilized for the pruning purpose of C4.5 decision tree classifier. The conclusions indicated that the new approach considerably decreased the tree size and accomplished the accuracy of 76.51%.

Aljaaf et al [10], proposed a multi-level model for evaluating the risk of heart failure in cardiac vascular patients. They used C4.5 classifier and the output class was classified in to five risk categorize. In addition to this, they improved the early diagnosis of cardiac failures by including three major risk elements in the heart disease dataset. The final result showed that the predictive model's accuracy is calculated as 86.53 %.

Chaurasia et al [29], calculated the accuracy in predicting heart disease by using the most basic forms of decision tree classifiers. ID3, CARD and DT decision trees are used to obtain results. 10 fold cross validation technique is used for the purpose of evaluation. 83.49 %, 82.5%and 72.9% are the accuracy results that are achieved by CARD, DT and ID3 respectively.

Shouman et al [30] with the intent to increase the accuracy in decision trees suggested a model that concentrated to integrate K- means clustering algorithm with it. Enhancement in the results are achieved by performing the inlier process with two clusters. The accuracy achieved by this suggested technique is specified as 83.9%.

Ghumbre et al [34] suggested a technique which concentrated on applying Support Vector Machine and radial basis function approach that operate independently for the prediction of

cardiac diseases. Chita et al [35] made a comparison of single machine learning classifiers with convolutional neural network. They used ensemble based classifiers and significant accurate results were obtained from the leant model. Das et al [36], another previous study, which is based on ensemble classifiers. Neural network's ensemble is used and a greater accuracy is obtained when matched with the performance of independent/single classifiers. Ubeyli et al [37] applied several machine learning classifier for the purpose of disease prediction and accessed the Support Vector Machine attained the highest accurateness.

Bashir et al [18] suggested a diverse ensemble based framework for the diagnosis of various diseases. They used multiple data sets regarding to heart disease , among them two are also used in this study as well i.e. Statlog and SPECTF [14] [25]. The used a multilayer ensemble architecture where in first layer five well known classifiers (Support vector machine , Naïve Bayes , Kth nearest neighbor , Quadratic Discriminant Analysis and Linear Regression) were utilized and in the second layer of ensembles two more classifiers (Artificial Neural Network and Random Forest) were utilized. For SPECTF Data set the accuracy is recorded as 83.03 and for Statlog data set the accuracy is calculated as 87.93.

Bashir et al [22] also suggested an ensemble based technique which was fundamentally constructed to eradicate the challenges of conventional machine learning techniques. Five machine learning classifiers were utilized for the purpose of training the predictive model for classification. Instance based learners, NaïveBayes, Decision tree based on Gini index, Decision tree based Information Gain and Support Vector Machine. Multiple data sets for heart disease prediction were used in this study. Among them two data sets are also part of this study i.e.

SPECTF and Statlog. The accuracy for SPECTF was recorded as 72.73 and for Statlog it was 87.37

The following figure shows some results for data mining approaches used in the past studies for finding the accuracy of heart disease prediction. These approaches are taken from the recent past and all of them have shown a considerable amount of accuracy while prediction of heart disease. Also we have observed that most of these techniques have considered ensemble based techniques for constructing their learning model. Our main intention is to create such a framework that is viable for predicting heart disease in cardiac vascular patients with an improved accuracy (from the accuracy obtained from previously constructed models and frameworks).

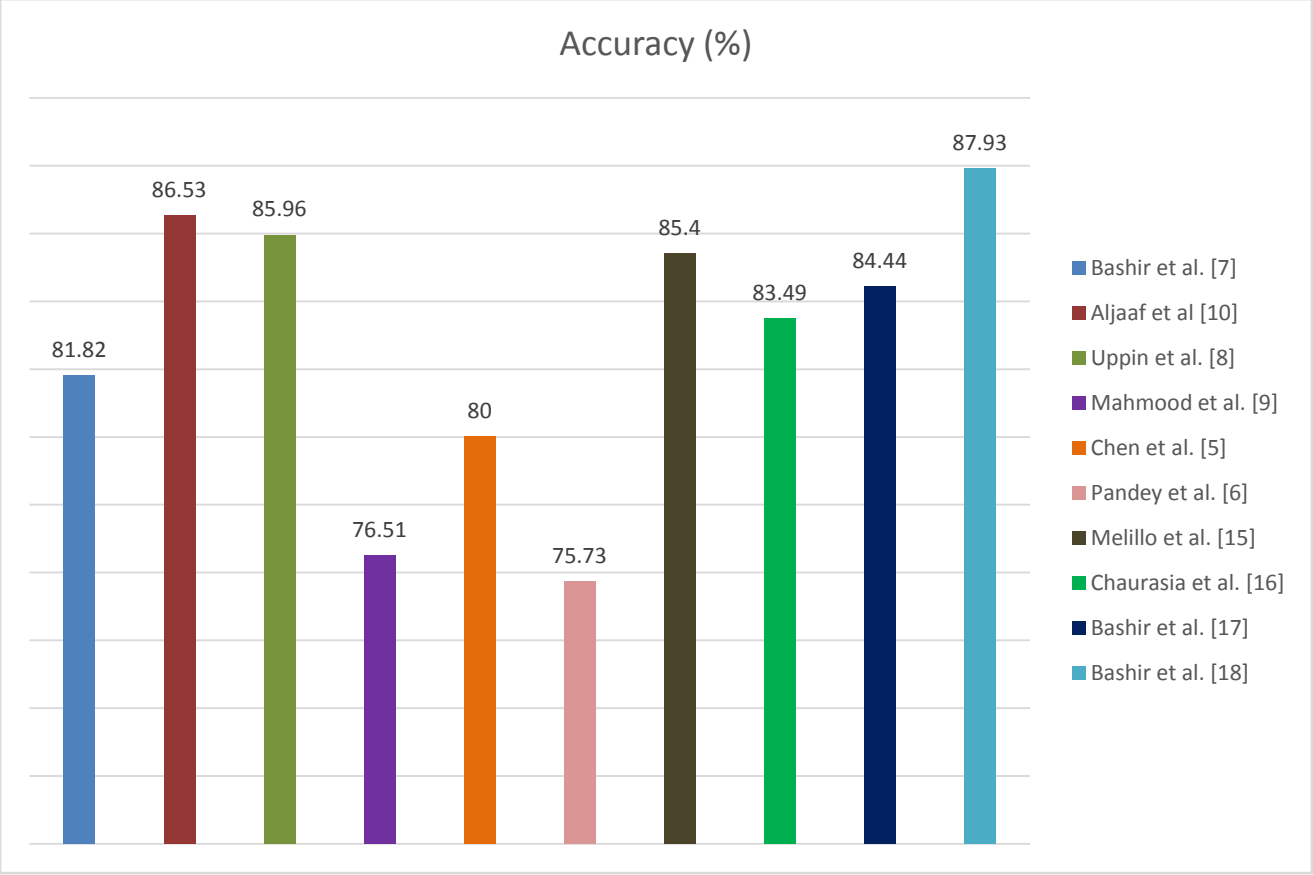

Figure 1: Prediction Accuracy Results of Previous Studies

Considering the achievements of various data mining techniques and ensemble technique in particular, it is very advantageous to use this approach for efficient and accurate prediction of heart disease. So we have presented an ensemble based framework which is built on majority voting mechanism. In addition various pre-processing techniques like data normalization is applied on selected attributes. Also, on the basis of attribute weights, specific attributes are selected from the data set to run ensemble based classifiers on it. Outlier detection is also used to detect the outlying records from the data set.The proposed framework showed significant improvement in the accuracy of heart disease prediction as compared to the previous studies

CHAPTER 3: DATA SET(S) DESCRIPTION

Two data sets have been used for this research study and are utilized to evaluate our suggested framework. The data sets are named as Statlog and SPECTF (Single Proton Emission Computed Tomograohy-Continuos Features).Both data set are quite well known and several past studies have included them to evaluate their techniques. The Statlog data set is retrieved from UCI online data sets repository at [14]. This data set is related to medical field and is specific to heart disease patient's data. The second data set SPECTF is also retrieved from UCI data repository [25]. The following table will tell about the high level details of the two data sets.

Data Set	Number Of Records	Number Of Features
Statlog	270	6 real +1 ordered + 3 binary + 3 nominal + 1 binary class = **14**
SPECTF	267	44 continuous + 1 binary class = **45**

Table 1 : Data Sets Statistics [22]

3.1 Statlog Data Set:

To begin with, statlog heart disease data from the famous UCI data set library [14] has been used for the purpose of training and testing the suggested framework. The data set consist of 75 attributes but 13 attribute (plus one label class) have been extracted from a larger data set which help in derive the presence or absence of heart disease more accurately. 6 attributes are identified as real, one attribute is ordered and 3 are nominal attributes. Table 1 shows some description about the attributes in the Statlog heart disease data set [14]. The data contains 270 instances, in

which 150 represent data of healthy patients and the remaining 120 represent patient with presence of heart disease. The data does not contain any missing value.

Table 2: Statlog Heart Disease Data Set Description

Attribute Name	Description
Age	Age of patient effected with heart disease (years)
Sex	Gender of heart patient (Male or Female)
Chest Pain Type	Ranges from 1 to 4 for several chest pain types like- typical angina, atypical angina, non-angina pain and asymptomatic.
Resting Blood Pressure	Heart Patient's Blood Pressure measure in mmHg
Serum Cholesterol	Serum Cholesterol level of cardiac vascular patient measured in mg/dl
Fasting Blood Sugar	Fasting blood sugar greater than 120 mg/dl **(**specified as Yes/No**)**
Resting Electrocardiographic Results	Values specified as 0, 1 or 2
Maximum Heart Rate	Heart Patient's Extreme heart beating rate , specified as an integer ranges from 71 – 202
Exercise Induced	Exercise induced Angina (indicated as Yes/No),
Old Peak	ST depression prompted by exercise related to relaxation.
Slope	Slope of the peak exercise ST segment (varies from 1 -3)
Number Of major Vessels	Quantity of major vessels painted by fluoroscopy(varies from 0 - 3)
Thal	3 = normal; 6 = fixed defect; 7 = reversible

In the past few well known research studies have used statlog data set for the purpose of evaluating machine learning based frameworks. There is still room for improvement is accuracy of results previously obtained. Therefore we have also utilized statlog data set in this study as well.The following figure shows some results for data mining approaches used in the past studies for finding the accuracy of heart disease prediction (Statlog data set).Some of the important and recent results obtain by using data mining techniques on statlog heart disease data set is shown in Figure 2

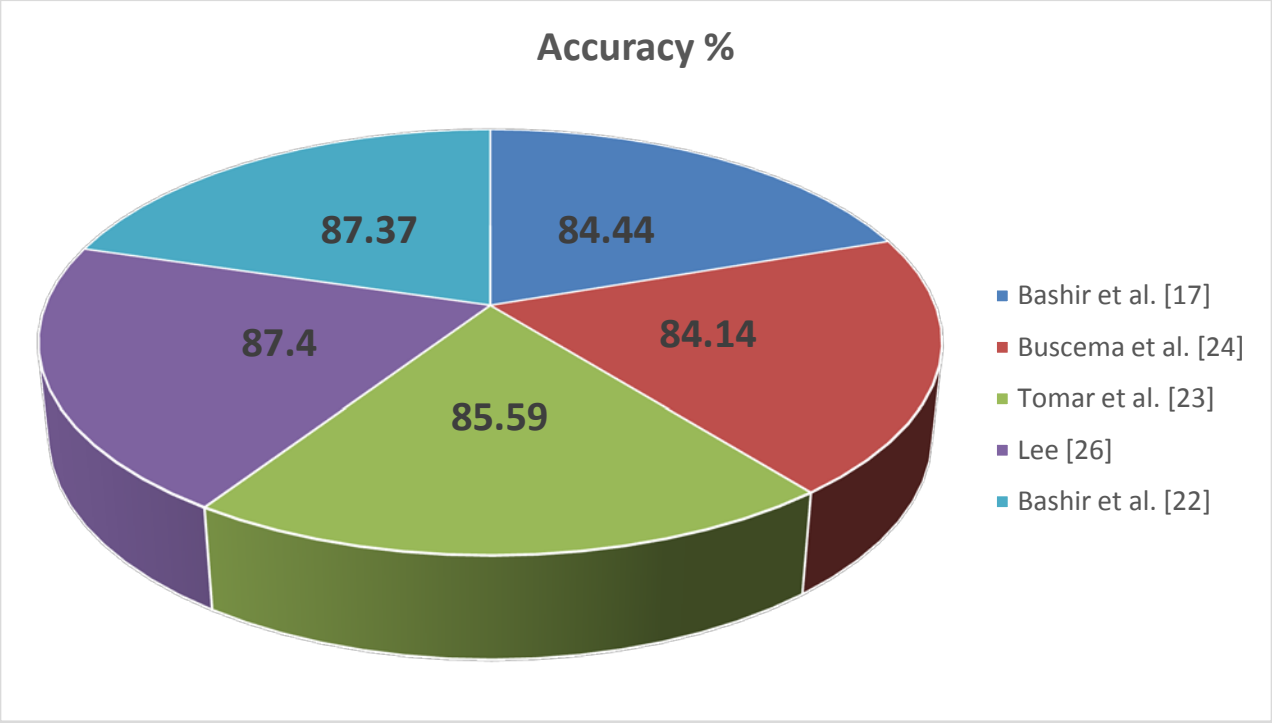

Figure 2: Prediction Accuracy Results of Previous Studies (Statlog Data Set)

3.2 SPECTF Data Set:

In addition to statlog data set another well-known heart set i.e. SPECTF [25] is also used for the purpose of predicting heart disease in the patients using the suggested framework. SPECTF data set consist of 267 records 45 attributes. All attributes except the label class are continues in nature [25]. 212 patients are classified as healthy whereas the remaining 55 are categorized as unhealthypatients. Some of the important and recent results obtain by using data mining techniques on SPECTF heart disease data set is shown in Figure 3.This data-set explains analyzing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Every patient is categorized into binary class i.e. healthy and unhealthy. Table 3 shows some description about the attributes in the SPECTF heart disease data set [25].

Table 3: SPECTF Heart Disease Data Set Description [25]

Feature	Range	Feature	Range	Feature	Range
F1R	{23-83}	F8S	{8-77}	F16R	{39-80}
F1S	{21-81}	F9R	{3-80}	F16S	{9-84}
F2R	{38-82}	F9S	{11-83}	F17R	{28-77}
F2S	{34-84}	F10R	{35-86}	F17S	{16-80}
F3R	{21-82}	F10S	{17-81}	F18R	{18-76}
F3S	{25-89}	F11R	{36-86}	F18S	{13-77}
F4R	{21-85}	F11S	{40-86}	F19R	{33-80}
F4S	{20-83}	F12R	{27-84}	F19S	{18-79}
F5R	{12-77}	F12S	{32-85}	F20R	{11-83}

F5S	{22-78}	F13R	{12-79}	F20S	{6-81}
F6R	{11-81}	F13S	{6-80}	F21R	{5-82}
F6S	{12-82}	F14R	{8-80}	F21S	{8-83}
F7R	{32-79}	F14S	{17-86}	F22R	{11-82}
F7S	{28-80}	F15R	{13-79}	F22S	{4-73}
F8R	{23-77}	F15S	{7-78}	DIAGNOSIS	{0-1}

SPECTF is also famous among the previous research studies on heart disease data sets. Both individual data mining classifiers and hybrid ensemble approaches are used here as well for the evaluation. Ensemble techniques hold an edge over individual classifiers for this data set as well.The following figure shows some results for data mining approaches used in the past studies for finding the accuracy of heart disease prediction (SPECTF data set). It can be seen that there is no final verdict on the best technique and there is supposed to be a sufficient room for improvement in prediction accuracy of heart disease presence or absence in a patient.

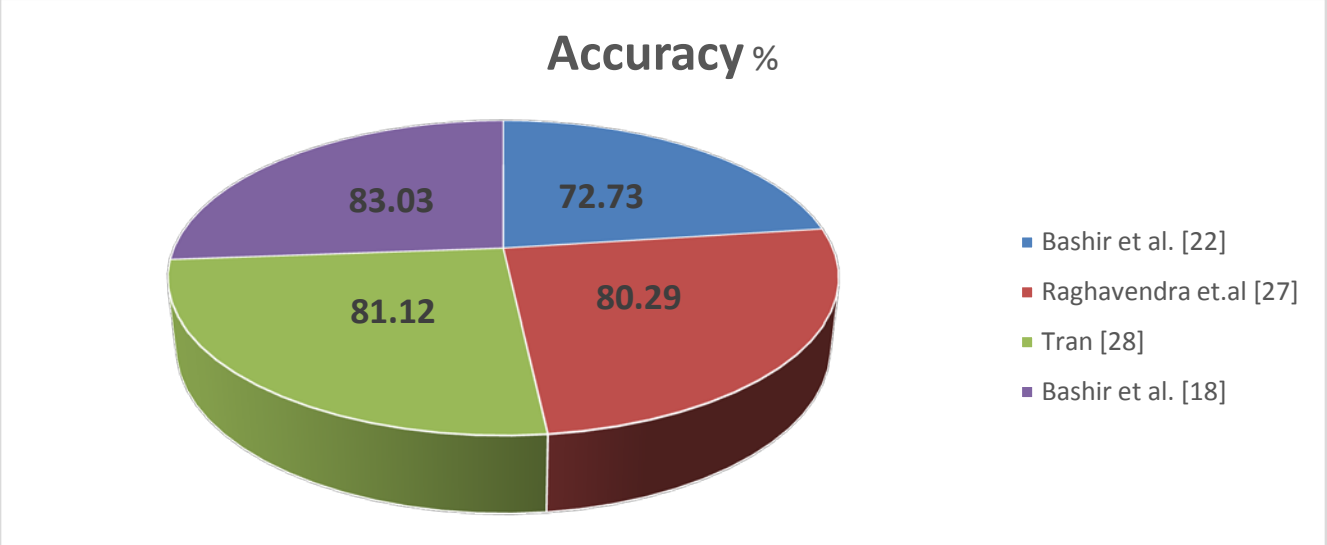

Figure 3: Prediction Accuracy Results of Previous Studies (SPECTF Data Set)

CHAPTER 4: PROPOSED FRAMEWORK

The suggested framework is based on a new amalgamation of three classifiers i.e. SVM, Naïve Bayes and Linear Regression. These suggested set of classifiers are then joined by majority voting mechanism [12]. Split Validation is used to divide the data set into training and testing data sets. Data normalization is done by using z-transformation method on a subset of attributes for both training and testing data sets. Chi Square Technique is then used to calculate the weights of attributes and top twelve attributes are taken into further consideration. For training data set, outlier detection technique is also used to remove the outliers from the data. This section contains the description of various classifiers and pre-processing technique that is used by the suggested framework

4.1 Classifiers:

4.1.1 Support Vector Machine:

The typical support vector machine classifier is a non-probabilistic binary classifier which uses a set of input data and maps it into two probable output classes. The fundamental theory of Support Vector Machine classifier is constructed on Statistical learning principle. The classifier produces a combination of hyper planes which are utilized in high dimensional space for the purpose of regression and classification. Primarily this classifier was designed for classifying binary data but later it was extended for handling multiclass data sets as well. [13].

The classifier i.e. SVM produces a hyper plane or a variety of such planes that can be utilized in a large number of regression and classification problems. A variety of kernel functions e.g. Gaussian and polynomial functions are applied for the purpose of data mapping .The below

mentioned rule is used for the purpose of classification in Support Vector Machine classifier. [13]

$$\mathbf{Sgn(f(x, y, z))}$$

$$\mathbf{f(x, y, z) = <y.\ x> + z}$$

In the above equation (y,z) signifies a problem that is complex in nature and x represents the classifier.

4.1.2 Naïve Bayes:

The famous Naïve Bayes algorithm emphases on the principle that the feature itself determine the existence or absence of disease in a cardiac vascular patient. It considers that the features are not dependent on one another [19]. For the purpose of training Naïve Bayes algorithm's model of probability, a supervised learning technique can be applied [20]. The classification results of Naïve Bayes classifier are considered reasonably good despite of the reason that the assessments of its probability does not have a greater worth [21]. The Bayesian formula mentioned below is applied to compute the class of disease from a give data set.

$$\mathbf{P(Ck|X) = P(Ck) \times P(X|Ck)/ P\ (X) \ldots\ldots\ (a)}$$

Here X is considered as the record which is going to be classified and Ck is its particular class.P(Ck|X) symbolizes the probability of X associated to class Ck.

Due to the scarceness of data the direct calculation of P(Ck|X) is not feasible. Hence, P(X|Ck) will be further estimated as :

$$\mathbf{P(X) = \pi\ (j=1 \rightarrow\ n)\ \ P(Xj|C)\ \ \ldots\ldots\ (b)}$$

Here **Xj**represent the jth element in X. Now from equation (a) and (b) we concluded the equation as.

$$\mathbf{P(Ck|X) = P(Ck) \times \{\pi\ (j=1 \rightarrow\ n)\ \ P(Xj\ |Ck)/ P\ (X)\}}$$

4.1.3 Linear Regression:

Linear Regression is a classifier that is used for forecasting in the field of mathematics. It is a statistical technique which analyses the power of relationship among one dependent attribute (In our case it's the class attribute) and various other independent attributes. Just like the way classification is applied to analyses categorical classes, Regression is utilized for the predication of continues values.

Machine learning, or more explicitly the science of predictive modeling is mainlyassociated with reducing the inaccuracy of a model or predicting the most correctresults, at the cost of understandability. As such, linear regression was established in the arena of statistics and is considered as a model for accepting the associationamong input and output statistical variables, however has been adopted by machine learning. In Linear regression, we intent to calculate a continuous variable **y** related with a known input vector **x**. Here **y** and **x** are dependent and independent variables respectively. Some other notations are as follows: [33]

y = symbolizes the continuous model of the dependent variable.

t = represents distinct noisy interpretations of the dependent variable.

Regression specifies that y is considered a function of x. The accurate working of this function is ruled by an indefinite parameter vector known as w and represented as: [33]

$$\mathbf{y = y\ (x,\ w)}$$

If y is linear in w, then the regression is considered as linear. The other way around, it can be represented as: [33]

$$y = w_t * \Phi * (x)$$

Here **Φ(x)** is approximately considered a possible nonlinear function of x and generally linear regression is represented by the following equation: [33]

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

4.2 Data AcquirementandPreprocessing Component:

The fundamentalneed of data attainment and pre-processing component is to acquire data fromvarious online heart disease data set repositories and afterwardspolishit to transfer into a structure that is appropriate for laterexploration. Every data set has diverseset of characteristics and data formats. So that's why it's very important that to execute an ensemble based classifier, the data must be refined for obtaining best results. The pre-processing stageincludesseveral steps that are executed on every data set successively. It includes outlier detection, Normalization by z-score method and chi square technique to calculate individual weights of attributes.

4.2.1 Outlier Detection (Distances):

An outlier is said to be an instance that is numericallyoutlying from the remaining setofrecords of the sample data set. An outlying instance is the one that seems to straydistinctlyand evidently from other records of the sample data set. Outliers are frequently (not all the time) symbolic of measurement fault and such instances (outliers) should be removed from the specified data set so that they won't contribute in creating a learning model and hence final prediction of label class.

The suggested ensemble technique utilized an outlier detection mechanism which implements outlier search based on the outlier detection methodology suggested by Ramaswamy et al.[38]. In thatstudy, a modelfor distance-based outliers was suggested that is fundamentallyconstructed on the distance calculated from a defined point to its *k-th*nearbyneighbour. The ranking of each point is dependent on its distance to its *k-th*closestneighbour and the uppermost*n* points in this ordering are referred as outliers. Here the valuesof *n* and *k*iscalculated by the *total sum of neighbours* and *amount of outlier's*variablesrespectively. The outlier detection techniquecreate a new Booleanfeaturecalled 'outlier' in the specifieddata set. If the value of this additional feature is calculated as true, than this record of sample data set will be considered as an outlier. Similarly if the value of the added feature is false, than this record is not an outlier is the sample data set [38].

4.2.2 Z-Score Normalization:

Normalization is a well-known pre-processing methodusually applied to re-scale feature values to encapsulate in a definite range. Data normalization is considered very effective and vital when researchers are dealing with features of various scales.Normally data sets have different set of attributes/features which are scaled on different ranges. Applying machine learning approaches on them may not provide the efficient results.Few data mining practicesapply the Euclidean distance for normalization so that all the featuresmust have the similar scale for a rational comparison amongst them. Likewise, normalization is a procedureapplied to smooth the playing arena when considering at features that extensivelydiffer in size as a consequence of the units nominated for illustration. [39]

Z-score Normalization is a well-known technique under the banner of normalizationand is also named as statistical normalization. The ambition of statistical normalization is to transform a feature value into Normal distribution through mean value; which is equals to 0 and variance which is specified as 1. The numericalequation of statistical normalization is specified as Z = (X-u) /s .Z-Score Normalization is also referred as Standard Normal distribution, N (0, 1). Though, the scale of the standard Normal distribution is not sandwiched among0 and 1howeveraround ranges from (-) 3 to (+) 3 (in fact ∞ to ∞ however by applying -3 to +3 you actuallytargeted 99.9% of your information). [39]. The equation of z-score normalization is shown below: [39]

$$d' = \frac{d - mean(P)}{std(p)}$$

Here mean (P) represent sum of all the features of P and std (P) represents the standard deviation of P [39]. The document will be followed by another pre-processing technique i.e. Normalization. Z-Score Normalization will be explained as a type of normalization which is used in the framework in pre-processing stage.

4.2.3 Feature/Attribute Selection Techniques

In this section we will discuss the famous feature extraction techniques (other than the one we have used) which we can use to extract mostvaluable features from the specified data sets [14] and [25]. Feature are selected on the basis of their individual weights that are calculated by chi square technique. The Chi square technique is explained in the following heading.

4.2.3.1 Chi Square-Calculating Individual Attributes Weights:

The Weight by Chi Squared statisticalapproachdetermines the weight of specific data set's features in accordance to the class feature by applying the chi-squared statistic. The greater the weight of afeature, the more important and vital it will befor the machine learning process. As chi-squared statistic could only be considered for nominal class labels, thus this approachcan only be useful for data sets with nominal class label.

The chi-square weight calculation approach is a nonparametric statistical approachutilized to decide if a scattering of perceived frequencies varies from the theoreticallyanticipated frequencies. The equation for the calculation of the chi-square statistic is determinedby:

$$X2 = Sigma\ [(P\text{-}A)\ 2/\ A]$$

Here *X2* isreferredas chi-square statistic, *P* is the perceived frequency and *A*is the anticipated frequency. Normally the chi-squared statistic sums up the inconsistenciesamongst the expectedamount of times every outcome happens (supposing that the model is correct) and the observed amount of times every outcome happens, by adding the squares of the inconsistencies, normalized by the anticipated numbers, over all the classes [41]. This Feature Selection technique is used for the suggested ensemble based framework.

4.2.3.2 Feature Extraction by principal component analysis (PCA)

The principal component analysis (PCA) method undertakes that most noticeable and valuable attribute in the dataset is one which has the greatest spread and variance. This concept is built on the point that the dimension with the greatest variance signifies the dimension which has the greatest amount of entropy and therefore represents the maximum amount of data. Eigen vector signifies A and B coordinates for a specified data set. The minutest Eigen vectors will purely symbolize noise units, however the largest Eigen vectors frequently relate to the principal

components that describe the information. Dimensionality decline through PCA is then fulfilled merely by representing the data onto the major Eigen vectors of its covariance matrix. Hence, we attain a true K-dimensional plane of the unique L-dimensional data, where $K \leq L$. The Singular Value Decomposition (SVD) is a method to accomplish PCA analysis and is specified by [42]

$$[U, V, W] = VWD (A)$$

$$\text{Thus, } A = UVWT$$

4.2.3.3 Feature extraction by forward selection and backward elimination

The feature short listing procedure includes lessening of features by choosing merely those attributes which bestow toward ultimate diagnosis of disease and setting rest as excluded. These excluded attributes will not be utilized for succeeding components and examination. There are numerous stages included in the method of attribute selection and recognition. The generation process and selection method are two of the most vital stages. The generation method includes generation of subgroup of attributes whereas selection process will assess these attributes on the foundation of diverse assessment standards. The generation process can effect in blank set, subgroup founded on arbitrarily nominated features set or a set founded on entire features set. Forward selection technique is utilized for blank set which repeatedly enhances the attributes in feature set and back-ward elimination method is utilized for all attributes set which repeatedly eradicates the unrelated features from feature set. The relevancy of features is calculated based on wrapper methodologies. The foremost emphasis of wrapper methodologies is classification accuracy. The approximation accuracy of every feature set is calculated that is contender of addition or elimination from the information. The feature selection procedure lasts till pre-specified amount of attributes are attained or some edge conditions are reached. [43]

4.2.3.4 F-score feature selection

F-score attribute choosing technique is used to choose the most suitable and applicable attributes from datasets. F-score technique can differentiate among two classes having real values. For a specified dataset, F-score of a specific attribute is calculated by given mathematical formula: [44]

$$F(i) = \frac{(X_i'^{(+)} - X_i')^2 + (X_i'^{(-)} - X_i')^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(X_{k,i}^{(+)} - X_i'^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(X_{k,i}^{(-)} - X_i'^{(-)})^2}$$

The discernment among positive and negative sets is displayed in numerator while the denominator symbolizes one within each of the two sets. An edge value is utilized to choose the suitable attributes from feature set. If the F-score of a specified attribute is larger than edge value then the attribute is added to chosen feature sample, else it is excluded from feature sample. The edge value is achieved by examining the average of F-scores of all attributes.

CHAPTER 5: SUGGESTED TECHNIQUE

The suggested technique is fundamentally built on ensemble of three well known classifiers i.e. Support Vector Machine, Naïve Bayes and Linear Regression. In the recent times, ensemble based classifiers and frameworks are considered very efficient and accurate in the field of medical data mining. The results predicted by ensembletechniques are significantly better than the performance of individual classifiers [17].

So keeping this in mind, the suggested technique has utilized this concept of ensemble classier and contributed in improving accuracy of heart prediction. The suggested technique also involves data acquirement and gathering from a famous online data set repository i.e. UCI machine learning repository and can be accessedat [14] and [25], pre-processing (for both training and testing data sets) and training of classifier based on majority voting scheme. The training model is then applied on the test data for predicting the presence of heart disease with improved accuracy in cardiac vascular patients.

The intention is to suggest a technique, which can serve in the medical field by increasing the accuracy of prediction in heart disease prediction. The large amount of data present in medical field creates a need for this as timely and accurate prediction is very vital in the medication of patient. Also the room for improvement in recent studies make it more vital to work towards improving the performance of frameworks in medical disease prediction field.Thedetailed suggested technique in shown belowin figure 4. Each component is diagramed to explain the flow of the technique we have used to predict heart disease with greater accuracy. Modules are divided as training and testing modules on the diagram to explain the techniques used in both

modules. Apart from that the pre-processing component is also displayed to show the techniques used in the preprocessing stage.

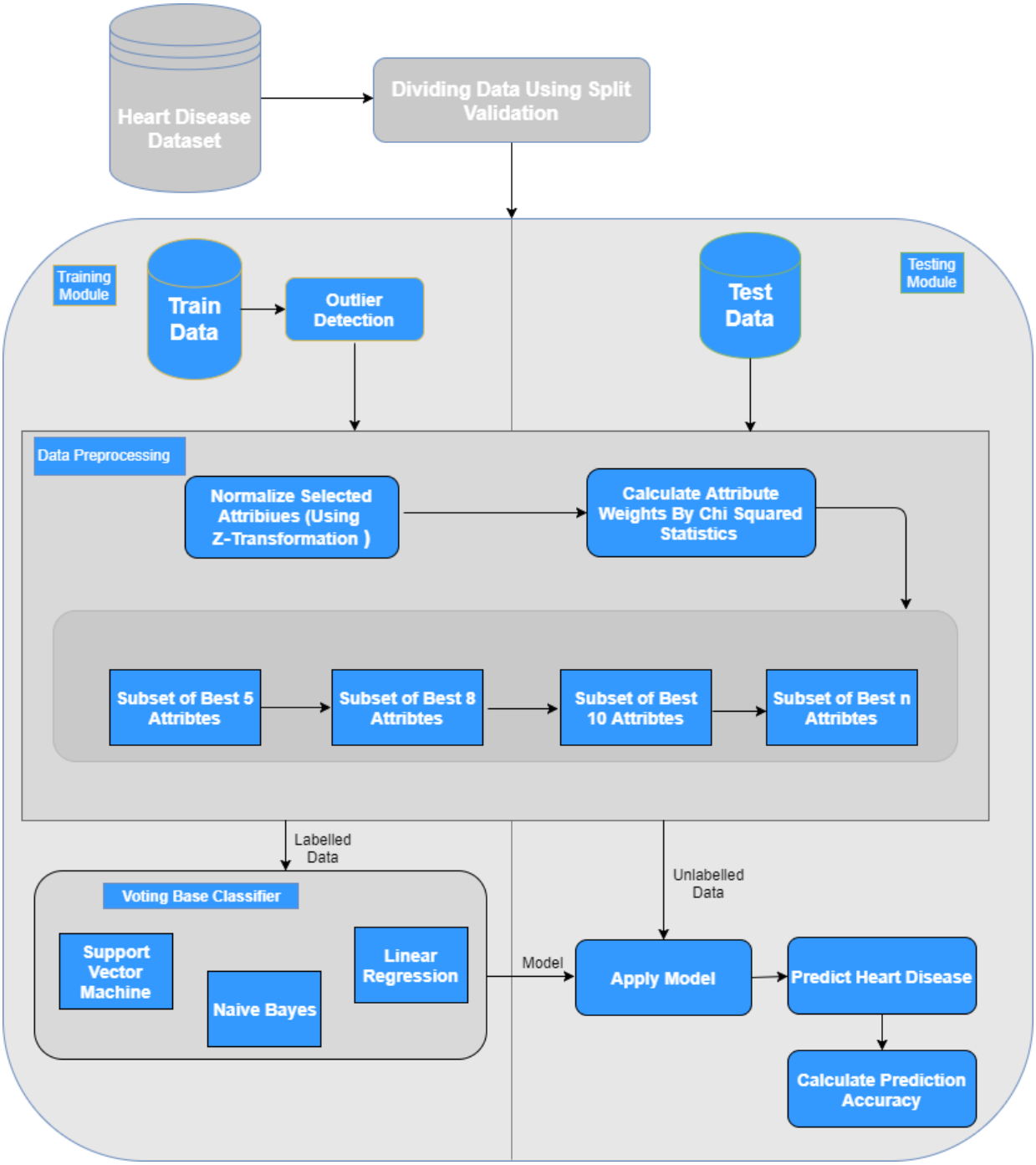

Figure 4: Architecture of Suggested Ensemble Technique

Here we are going to elaborate the step by step formation of our suggested technique .Initially the heart disease data sets (Statlog and SPECTF) are gathered from UCI data repository [14] and [25]. The acquired data contains a total of 270 and 267 records respectively, describing the presence or absence of heart disease. The detailed description of data sets can be acquired from chapter 3 of this document. Following steps are included in the pre-processing phase for both data sets

- Split Validation technique is used to divide the acquired data set into training and testing data set. This training data set will serve to generate a classifier model which will be used by testing data to predict and evaluate results.For Statlog data set 188 records are considered for training the classifier whereas the remaining instances are used to test the model obtained from classifier based on the majority voting scheme. Similarly for SPECTF 182 records are used for the purpose of training of ensemble and the remaining are used for testing the validity of the generated model.

- For both training and testing data set, z-score normalization is performed on the subset of attributes, especially rescaling the attributes with values spreading over larger range. It's basically a need to set your entire data on the similar scale: if the scales for variousattributes are majorly dissimilar, than this can have a knock-on consequence on your capacity to learn (conditional towhich procedures you're following to do it). More details about Z-Score normalization can be found in Chapter 4 of this document.

- Pre-processing also includes selection of best attributes (for both training and testing data sets) on the basis of their individual weights.Chi-Square, which determine the individual weights of the attributes provide the basis attribute selection. Best attributes/features are ordered on the basis of their weights calculated by Chi-Square and hence selected for further execution.More details about Chi-Square technique can be found in Chapter 4 of this document.

-

After the above steps, the pre-processed data serves as an input for the ensemble based classifier which is a combination of three well known classifiers. The collaborative architecture utilized SVM, Naïve Bayes and Linear Regression classifiers to categorize the test data into patient with heart disease and person with no heart disease. The concluding results is predicted on the basis of majority voting scheme (These sort of ensemble techniques are very famous in the recent times). It states that in order to be categorized as a particular class, at least two of the selected classifiers should vote for this class. It means that the instance will be classified as true if and only ifat least two classifiers out of three, categorize it as true. Similarly labeled class will be valued as false if two classifiers categorize it as false.So the test instance is allotted to the class which has maximum votes (two classifiers predicting similar result). The predicted results are then compared with the actual result and hence the performance accuracy of the framework is calculated.

The next section of this document will contain such result and findings which have evaluated the performance accuracy of our suggested technique. These findings are obtained by applying the suggested technique on two data sets (Statlog and SPECTF) accessed from [14] and [25].

CHAPTER 6: RESULTS AND FINDINGS

In this study, we have trained three machine learning classifiers by using heart disease (statlog) data set to efficiently predict presence and absence of heart disease in a heart patient with higher accuracy. We have also applied the suggested ensemble framework on another heart disease data set to ensure the validity of the suggested ensemble framework (SPECTF). The results will proceed after we will discuss the results of statlog data set.

The accuracy of individual classifiers is calculated initially and then it is compared with the accuracy of suggested ensemble technique. We have seen improved results of accuracy when suggested ensemble technique is executed. **Specificity** and**Sensitivity** are also calculated which identifies the quantity of patients that are accurately classified as healthy and ratio of patients classified rightly as sick respectively. Numerically these two measures are calculated as: [7].

Specificity = True Negatives / True Negatives + False Positives

Sensitivity = True Positives / True Positives + False Negatives

Accuracy calculates the ratio of true estimations for predicting the actual class label from the test data while using the classifiers used in the suggested framework. Numerically it is calculated as: [7]

Accuracy = True Positives + False Negative / TP + FP + TN + FN

A confusion matrix of heart disease data set (statlog) along with its specificity, sensitivity and accuracy is shown in the table below. Here we can see that a large number of healthy patients i.e. 45 are correctly specified as healthy (True positives). Similarly 29 sick patients are correctly predicted as sick (True negatives). The specificity and sensitivity is calculated as approximately **83%**and **96 %.**

Table 4: Confusion Matrix of Suggested Ensemble Framework for Statlog Data

Label	Healthy Patient	Sick Patient	Specificity	Sensitivity	Accuracy
Healthy Patient	45	6	82.9	95.75	90.24
Sick Patient	2	29			

The table below will show the comparison of individual classifiers with our suggested ensemble technique.Here it is can be observed that the suggested ensemble technique have improved results of accuracy. Naïve Bayes, Support Vector Machine and Linear Regression showed 82.9%, 85.3% and 86.5% accurate results respectively. And the suggested ensemble technique has a better and improved accuracy of **90.24%**. Individual classifiers lag behind from the suggested ensemble technique in predicting accuracy for heart disease patients for statlog data set. This trend is also observed in the recent times as well, where the hybrid approaches like ensemble technique outweigh individual data classifiers.

Table 5: Comparison of Individual classifiers with Suggested Technique for statlog data

Classifiers	Accuracy
Naïve Bayes	82.9
SVM	85.3
Linear Regression	86.5
Suggested Ensemble Technique	**90.24**

A comparison of our proposed framework for predicting heart disease in statlog data set with different published studies is shown in the table below:

Table 6: Comparison of Suggested Ensemble Technique with famous techniques for Statlog Dataset.

Technique/Study	Accuracy	Specificity	Sensitivity
Bashir et al. [17]	84.44	86	86
Buscema et al. [24]	84.14	N/A	N/A
Tomar et al. [23]	85.59	89	85.71
Lee [26]	87.4	N/A	N/A
Bashir et al. [22]	87.37	87.27	87.50
Suggested Ensemble Technique	**90.24**	**83**	**95.75**

Here we can see that the suggested ensemble technique showed an accuracy improvement in predicting cardiac diseases in its patients.

We have also applied the suggested framework on another famous heart disease data set known as SPECTF. This Heart is also a famous heart disease data set and is accessible on UCI at [25]. The data set consist of 267 records and 45 attributes (44 continuous, 1 binary).

The table below will show the comparison of individual classifiers with our suggested ensemble technique.Here it is can be observed that the suggested ensemble technique have improved results of accuracy. Naïve Bayes, Support Vector Machine and Linear Regression showed approximately 83%, 83% and 82% as prediction accuracy respectively. And the suggested ensemble technique has a better and improved accuracy of **84.52%.**So as seen in the recent times, the latest trends showed that ensemble techniques have greater prediction accuracy as compared to the individual machine learning classifiers. Here for SPECTF we have seen same sort of a trend.

Table 7 : Comparison of Individual classifiers with Suggested Technique for SPECTF data

Classifiers	**Accuracy**
Naïve Bayes	83.3
SVM	83.3
Linear Regression	82.3
Suggested Ensemble Technique	**84.52**

A comparison of our proposed framework for predicting heart disease in SPECTF data set with different published studies is shown in the table below. It can be seen that our new suggested ensemble technique have improved and accurate results as compared to some famous studies for the same data set.

Table 8: Comparison of Suggested Ensemble Technique with famous techniques for SPECTF Dataset.

Technique/Study	Accuracy	Specificity	Sensitivity
Bashir et al. [22]	72.73	73.33	72.67
Raghavendra et al. [27]	80.29	N/A	N/A
Tran. [28]	81.12	N/A	N/A
Bashir et al. [18]	83.03	99	7.45
Suggested Ensemble Technique	**84.52**	**88**	**94.29**

CHAPTER 7: CONCLUSION AND FUTURE WORK

Accuracy of results is considered an important factor in the medical arena as it linkedwith the life of a patient. Machine learning in the medical fieldstudies the formerpractices of researchers and investigates them to recognize the over-all trends and possibleanswers to the present problematic conditions. So we intended to use machine learning techniques and to achieve better and accurate results for heart disease prediction.

The purpose of the suggested research work is to predict the heart disease in a cardiac vascular patient with an improved accuracy. For this purpose, three well-known machine learning classifiers i.e. Support Vector Machine, Naïve Bayes and Linear Regression are applied using an ensemble technique to analyze the heart disease data set. We observed an improved and accurate prediction of heart disease with our suggested ensemble framework.

The suggested ensemble framework is fundamentally based on threemain components. The first component is data acquirement and data pre-processingwhich includesacquiring data from various online data sources and pre-processthem so that the data is a feasible input for the ensemble and a training model can be efficiently generated. Data Pre-processing include outlier detection, data normalization - which is done by using z-transformation method and Chi Square Technique is then used to calculate the weights of attributes and top attributes are taken into further consideration.

In the second component, ensemble based classifier’s training of data is then executed on the training records; and then the generated model isutilized to forecast class labels that are not knownin test data set records. The prediction and valuationis the third component of the proposed ensemble framework.The evaluation of our suggested ensemble based framework is executed on two different heart disease datasets obtained from [14] and [25].The examination of resultsspecifies that suggestedensemble technique has attained higher accuracy of heart disease prediction for both heart disease datasets.

As there is a lot of medical data accessible on online repositories to extract meaningful information from it, this technique can be extended to predict other diseases like cancer, hepatitis and diabetes etc. This will in turn enhance the chances of improved medication in a proper cost effective way for the patients, can save them from ailment and save a large number of human lives.AlsoIn future, aweb application can be built which is accessible for public and medical practitioners over internet and can execute any query regarding medical data efficiently and in a user friendly way.

REFERENCES

[1]. World Health Organization (WHO), “The top 10 causes of death, Factsheet", Available at: http://www.who.int/mediacentre/factsheets/fs310/en/.

[2]. European Society of Cardiology, “Heart failure: Preventing disease and death worldwide”, Available at: http://www.escardio.org/communities/HFA/Documents/whfa-whitepaper.pdf

[3]. VL. Roger, “The heart failure epidemic”, International Journal of Environmental Research and Public Health, 7(4), 1807-1830; 2010.

[4].Pattekari S.A., Parveen, A., Prediction System for Heart Disease Using Naive Bayes, International Journal of Advanced Computer and Mathematical Sciences, ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294

[5]. Chen, A.H., Huang, S.Y., Hong, P.S., Cheng, C.H., Lin, E.J., HDPS: Heart Disease Prediction System, Computing in Cardiology, (2011)

[6.] A.K. Pandey, P. Pandey, K.L. Jaiswal, and A.K. Sen, “A heart disease prediction model using decision tree”, IOSR Journal of Computer Engineering (IOSR-JCE), vol. 12, Issue 6, PP 83-86, Aug. 2013.

[7] S. Bashir, U. Qamar, and M.Y. Javed, “An ensemble based decision support framework for intelligent heart disease diagnosis”, International Conference on Information Society (i-Society 2014), London, IEEE, 2014.

[8].S.K. Uppin, and M.A. Anusuya, “Expert system design to predict heart and diabetes diseases”, International Journal of Scientific Engineering and Technology, vol. 3, no.8, pp: 1054-1059, 2014.

[9]. A. M. Mahmood and M. R. Kuppa, “Early detection of clinical parameters in heart disease by improved decision tree algorithm”, 2010 Second Vaagdevi International Conference on Information Technology for Real World Problems, Warangal, IEEE, 2010.

[10]. A. J. Aljaaf, D. Al-Jumeily, A. J. Hussain, T. Dawson, P Fergus and M. Al-Jumaily “Predicting the Likelihood of Heart Failure with a Multi Level Risk Assessment Using Decision Tree”,2015.

[11]. Palaniappan, S., Awang, R.: Intelligent Heart Disease Prediction System Using Data Mining Techniques. 978-1-4244-1968-5/08/ ©IEEE (2008)

[12]. Shouman , M., Turner, T., Stocker, R.: Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients. In: International Journal of Information and Education Technology, Vol. 2, No. 3 (2012).

[13]. Wang, S., Mathew, A., Chen, Y., Xi, L., Ma, L., Lee, J.: Empirical Analysis of Support Vector Machine Ensemble Classifiers.In: Expert Systems with applications, (2009)

[14]. Statlog heart disease data set Accessed From:https://archive.ics.uci.edu/ml/datasets/Statlog+(Heart)

[15]. P. Melillo, N.D. Luca, M. Bracale and L, Pecchia, "Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability", IEEE Journal of Biomedical and Health Informatics, vol. 17, issue 3, 2013.

[16]. V. Chaurasia and S. Pal, "Early prediction of heart diseases using data mining techniques" Caribbean Journal of Science and Technology, vol.1, 208-217, 2013.

[17]. S. Bashir, U. Qamar, F.H. Khan and L. Naseem, "HMV: A medical decision support framework using multi-layerclassifiers for disease prediction" Journal of Computational Science 13, 10–25, 2016

[18]. S. Bashir, U. Qamar and F.H. Khan, "IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework" Journal of Biomedical Informatics 59,185–200,2016

[19]. Sotelsek-Margalef, A., Villena-Román, J.,: MIDAS: An Information-Extraction Approach to Medical Text Classification. In:Procesamientodellenguaje Natural, pp. 97-104 (2008).

[20]. Zhang, H., The optimality of Naïve Bayes. American association of artificial intelligence, (2004)

[21]. Manning, D., Raghavan, P., Schutze,H.: Introduction to Information retrieval, Cambridge university, (2008)

[22]. S. Bashir, U. Qamar and F.H. Khan, "A MULTICRITERIA WEIGHTED VOTE-BASED CLASSIFIER ENSEMBLE FOR HEART DISEASE PREDICTION" Computational Intelligence,2015

[23]. D. Tomar and S. Agarwal, "Feature selection based least square twin support vector machine for diagnosis of heart disease," International Journal of Bio-Science and Bio-Technology, vol. 6, no. 2, pp. 69–82, 2014

[24]. M. Buscema, M. Breda, and W. Lodwick, "Training with Input Selection and Testing (TWIST) algorithm: a significant advance in pattern recognition performance of machine learning," Journal of Intelligent Learning Systems and Applications, vol. 5, no. 1, pp. 29–38, 2013

[25]. SPECTF heart Disease data set Accessed from :https://archive.ics.uci.edu/ml/datasets/SPECTF+Heart

[26]. S.-H. Lee, “Feature selection based on the center of gravity of BSWFMs using NEWFM,” Engineering Applications of Artificial Intelligence, vol. 45, pp. 482–487, 2015

[27].Raghavendra, S., and M. Indiramma. "Classification and Prediction Model using Hybrid Technique for Medical Datasets." *analysis* 127.5 (2015).

[28].Tran Thanh, Dat. "Dropout-based Support Vectors Regularization." (2017).

[29]. V. Chaurasia and S. Pal, “Early prediction of heart diseases using data mining techniques” Caribbean Journal of Science and Technology, vol.1, 208-217, 2013.

[30]. M. Shouman, T. Turner and R. Stocker, “Integrating decision tree and kmeansclustering with different initial centroid selection methods in the diagnosis of heart disease patients”, Proceedings of the InternationalConference on Data Mining, 2012.

[31]. C. Fernández-Llatas, J.M. García-Gómez (Eds.), Data Mining in ClinicalMedicine, Humana Press, 2015.

[32]. S. Dua, X. Du, Data Mining and Machine Learning in Cyber Security, CRC Press, 2011.

[33]. Linear Regression by J. Elder Available at: http://www.eecs.yorku.ca/course_archive/2011-12/F/4404-5327 / lectures/ 03%20Linear%20Regression.pdf

[34].GHUMBRE, S., C. PATIL, and A. GHATOL. 2011. Heart disease diagnosis using support vector machine. *In* International Conference on Computer Science and Information Technology (ICCSIT'), Pattaya, Thailand

[35].CHITRA, R., and D. V. SEENIVASAGAM. 2013. Heart disease prediction system using supervised learning classifier. International Journal of Software Engineering and Soft Computing,

[36] DAS, R., I. TURKOGLU, and A. SENGUR. 2009. Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications

[37]. E.D. Übeyli, Implementing automated diagnostic systems for breast cancerdetection, Expert Syst. Appl. 33 (4) (2007)

[38] Ramaswamy, Sridhar, Rajeev Rastogi, and Kyuseok Shim. "Efficient algorithms for mining outliers from large data sets." *ACM Sigmod Record*. Vol. 29. No. 2. ACM, 2000.

[39]. Saranya, C., and G. Manikandan. "A study on normalization techniques for privacy preserving data mining." *International Journal of Engineering and Technology (IJET)* 5.3 (2013): 2701-2704.

[40]. Centers for Disease Control and Prevention (CDC), Deaths: leading causes for 2008, National Vital Statistics Reports, Vol. 60, No. 6, June 6, 2012.

[41].McHugh, Mary L. "The chi-square test of independence." *Biochemiamedica: Biochemiamedica* 23.2 (2013): 143-149.

[42] C.Fernández-Llatas, J.M. García-Gómez (Eds.), Data Mining in ClinicalMedicine, Humana Press, 2015.

[43] An introduction to feature extraction, in: I. Guyon, A. Elisseeff (Eds.), FeatureExtraction, Springer, Berlin/Heidelberg, 2006, pp. 1–25

[44] Combining SVMs with various feature selection strategies, in: Y.W. Chen, C.J.Lin (Eds.), Feature Extraction, Springer, Berlin/Heidelberg, 2006, pp. 315–324.