DEVELOPMENT OF DATA MINING METHODOLOGIES AND MACHINE LEARNING MODELS TO UNDERSTAND CARDIOVASCULAR DISEASE MECHANISMS

A THESIS
SUBMITTED TO THE DEPARTMENT OF
ELECTRICAL AND COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Burak KOLUKISA
January 2020

DEVELOPMENT OF DATA MINING METHODOLOGIES AND MACHINE LEARNING MODELS TO UNDERSTAND CARDIOVASCULAR DISEASE MECHANISMS

A THESIS
SUBMITTED TO THE DEPARTMENT OF
ELECTRICAL AND COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Burak KOLUKISA
January 2020

SCIENTIFIC ETHICS COMPLIANCE

I hereby declare that all information in this document has been obtained in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Burak KOLUKISA

REGULATORY COMPLIANCE

M.Sc. thesis titled **"DEVELOPMENT OF DATA MINING METHODOLOGIES AND MACHINE LEARNING MODELS TO UNDERSTAND CARDIOVASCULAR DISEASE MECHANISMS"** has been prepared in accordance with the Thesis Writing Guidelines of the Abdullah Gül University, Graduate School of Engineering & Science.

Prepared By

Burak KOLUKISA

Advisor

Assistant Professor
Burcu BAKIR GÜNGÖR

Head of the Electrical and Computer Engineering Graduate Program

Professor V. Çağrı GÜNGÖR

ACCEPTANCE AND APPROVAL

M.Sc. thesis titled **"DEVELOPMENT OF DATA MINING METHODOLOGIES AND MACHINE LEARNING MODELS TO UNDERSTAND CARDIOVASCULAR DISEASE MECHANISMS"** and prepared by Burak Kolukisa has been accepted by the jury in the Electrical and Computer Engineering Graduate Program at Abdullah Gül University, Graduate School of Engineering & Science.

02 / 01 / 2020

JURY:

Assistant Professor Burcu BAKIR GÜNGÖR :……..

Assistant Professor Ufuk NALBANTOĞLU :……..…

Assistant Professor Ahmet SORAN :……..

APPROVAL:

The acceptance of this M.Sc. thesis has been approved by the decision of the Abdullah Gül University, Graduate School of Engineering & Science, Executive Board dated ….. /….. / …….. and numbered ……..……. .

….. /….. / ……..

Graduate School Dean

Professor İrfan ALAN

ABSTRACT

DEVELOPMENT OF DATA MINING METHODOLOGIES AND MACHINE LEARNING MODELS TO UNDERSTAND CARDIOVASCULAR DISEASE MECHANISMS

Burak KOLUKISA
MSc. in Electrical and Computer Engineering
Supervisor: Assistant Professor Burcu BAKIR GÜNGÖR
January 2020

World Health Organization (WHO) reported that in 2016, 31% (17.9 million) of the total deaths in the world were caused by Coronary Artery Disease (CAD) and it is estimated that around 23.6 million people will die from CAD in 2030. In the following years, this disease will cause millions of more deaths and the diagnosis and treatment will cost billions of dollars. CAD, which is a sub-category of Cardiovascular Disease (CVD), is the inability to feed the heart with blood as a result of the accumulation of fatty matter called atheroma on the walls of the arteries. With the development of machine learning and data mining techniques, it became possible to diagnose Cardiovascular Diseases (CVD), especially CADs, at a lower cost via checking some physical and biochemical values. To this end, in this thesis, for CVD diagnosis problem, different computational feature selection (FS) methods, dimension reduction, and different classification algorithms have been evaluated; and a domain knowledge-based FS method, an ensemble FS method and a probabilistic FS method have been proposed. Via experimenting on two publicly available data sets, i.e., UCI Cleveland and Z-Alizadehsani, this thesis aims to generate a robust model for the diagnosis of CVD, at a lower cost. In our experiments, our proposed solution achieved 91.78% accuracy and 93.50% sensitivity on the diagnostic tests.

Keywords: Machine Learning, Ensemble Feature Selection, Domain Knowledge Based Feature Selection, Classification, Cardiovascular Disease Diagnosis

ÖZET

KARDİOVASKÜLER HASTALIK OLUŞUM MEKANİZMALARINI ANLAMAK İÇİN VERİ MADENCİLİĞİ YÖNTEMLERİ VE MAKİNE ÖĞRENMESİ MODELLERİNİN GELİŞTİRİLMESİ

Burak KOLUKISA
Elektrik ve Bilgisayar Mühendisliği Bölümü Yüksek Lisans
Tez Yöneticisi: Dr. Öğr. Üyesi Burcu BAKIR GÜNGÖR
Ocak 2020

Dünya Sağlık Örgütü'nün 2016 yılında yayınladığı bir rapora göre, Koroner Arter Hastalığı (KAH), dünyadaki toplam ölümlerin %31'ine (17,9 milyon), neden olmaktadır. Ayrıca, 2030'da yaklaşık 23,6 milyon insanın KAH'dan dolayı öleceği tahmin edilmektedir. Bu hastalığın önümüzdeki yıllarda, milyonlarca ölüme neden olacağı, tanı ve tedavisinin milyarlarca dolara mal olacağı düşünülmektedir. KAH, arterlerin duvarlarında aterom denilen yağlı madde birikiminin bir sonucu olarak, kalbin kanla yeterince beslenmemesi durumudur. Makine öğrenmesi ve veri madenciliğinin yöntemlerinin gelişmesiyle birlikte, bazı fiziksel ve biyokimyasal değerleri kontrol ederek, Kardiovasküler Hastalığı (KVH) ucuz ve zahmetsiz bir şekilde teşhis etmek mümkündür. Bu bağlamda, bu tezde, KVH teşhisi için farklı hesaplamalı öznitelik seçme (ÖS) yöntemleri, doğrusal ayırt edici analizler ve farklı sınıflandırma algoritmaları değerlendirilmiş; ve bir alan bilgisi temelli ÖS yöntemi, bir topluluk ÖS yöntemi ve bir olasılıksal ÖS yöntemi önerilmiştir. Bu tez çalışması, halka açık iki veri seti olan UCI Cleveland ve Z-Alizadehsani verileri üzerinde deneyler yaparak, KHV'ları daha düşük maliyetle teşhis edebilecek sağlam bir model geliştirmeyi amaçlamaktadır. Önerilen çözüm, yapılan deneylerdeki tanı testlerinde %91.78 doğruluk ve %93.50 duyarlılık ulaşmıştır.

Anahtar kelimeler: Makine Öğrenmesi, Topluluk Öznitelik Seçme, Alan Bilgisi Temelli Öznitelik Seçme, Sınıflandırma, Kardiovasküler Hastalık Tanısı

Acknowledgements

I would like to extend my thanks to my supervisor Asst. Prof. Burcu Bakır Güngör for her supervision and supports. I also want to thank my family and especially my dear wife, for their patience and tolerance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

According to a report published by the World Health Organization (WHO) in 2016, 31% (17.9 million) of the total deaths in the world are caused by ischaemic heart diseases (IHD), which is the former name of coronary artery disease (CAD). Among other diseases, CAD is the leading cause of the deaths in 2016, as shown in the Figure 1.1. It is estimated that around 23.6 million people will die from CAD in 2030 [1]. CAD mainly develops when the major arteries that supply your heart became atherosclerosis. Atherosclerosis is the accumulation of fatty matter called atheroma on the walls of the artery. Atherosclerosis causes narrowing and occlusion on the vessels. The complete occlusion can cause a heart attack. CAD happens over time, and the diagnosis of CAD is difficult until a blockage or a heart attack emerges.

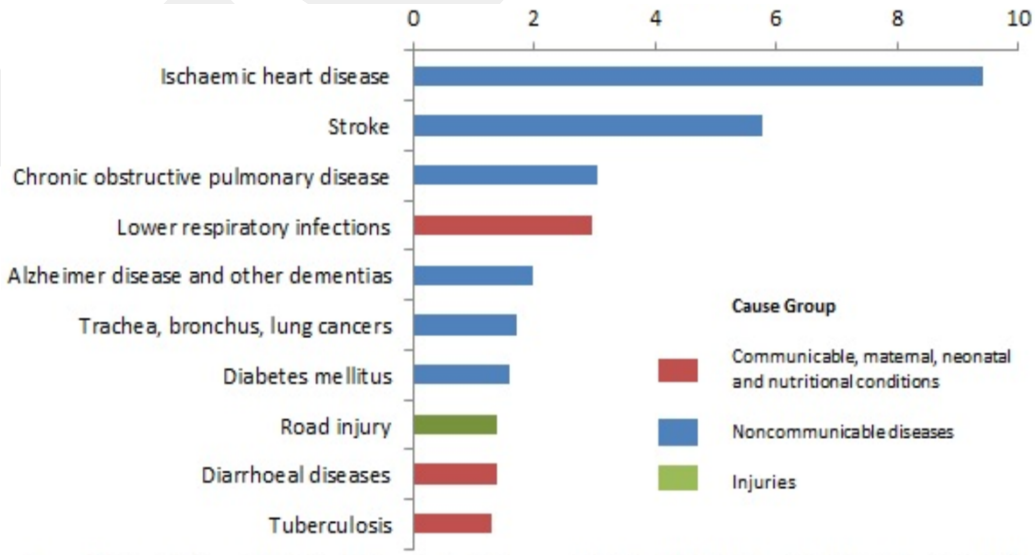

Figure 1.1 The leading causes of death in 2016

Cardiovascular disease (CVD) is the most general term that embraces all kinds of diseases that affect the heart or blood vessels, including CAD and heart diseases [2,3]. The diagnosis of this disease requires specific devices and experts. The most commonly used methods for the diagnosis of CVD are as follows.

Blood Test: Heart diseases can be determined with the laboratory test results. During a heart attack, the heart muscle cells die, and therefore proteins start to circulate in the blood. When the high amount of protein is detected in the blood, it indicates that the person has recently had a heart attack.

Electrocardiogram (ECG): In this method, the patient, is connected to the device via placing its sensor on the chest, wrists, and ankles of the patient. The test shows the rhythm of the heart and how fast it beats. EKG / ECG can help detect a heart attack.

Stress Testing: For this test, the patient need to perform effort or use pills to increase heart rate. Due to the narrowing of the arteries, the heart muscles cannot provide enough blood supply, and this causes shortness of breath and chest pain, which are the symptoms of atherosclerosis and coronary heart disease.

Echocardiography: This test visualizes the heart by using sound waves. It provides information about the shape, size, valves, chambers, and working of the heart. Thus, echocardiography allows to diagnose abnormal contractions of the heart and previously damaged parts.

Coronary Angiography and Cardiac Catheterization: Coronary angiography is an invasive test. The dye is injected into the veins from the arm, neck, or legs. The photos of the blood vessels of the heart are taken, and the system checks for the blockages in the large coronary arteries.

Chest X-Ray: Chest X-ray (radiography) is a non-invasive medical test that allows doctors to diagnose and treat medical conditions in a fast and easy way. -rays are one of the oldest methods of medical imaging. While producing the images, the body is exposed to a small amount of ionizing radiation.

Electron-Beam Computed Tomography (EBCT): EBCT detects calcium accumulation in the walls of the coronary arteries. These calcium deposits are early signs of coronary heart disease.

Cardiac Magnetic Resonance Imaging (MRI): This method generates a 3D image using radio waves within the strong magnetic field generated by magnets. Cardiac MRI clearly distinguishes certain anatomical structures from other structures and detects the differences between healthy and unhealthy tissues.

More than three-quarters of cardiovascular diseases occur in low- and middle-income countries (LMC)'s [4]. The testing phase of CVDs may not be economically feasible in these countries. Therefore, many people in LMC's who suffer from CVD disease die at a young age because of late diagnosis [5]. Since the people in LMC's often cannot benefit from early diagnosis and treatment programs compared to people in high-income countries, they are adversely affected. The diagnosis can be insufficient with inexperienced doctors. In this respect, advanced computer systems and effective methodologies can help to extract reliable information from medical data sets at a lower cost. Nowadays, information technologies are widely used in the medical field. Data mining and machine learning methods could make it possible to diagnose CVDs by examining specific parameters, rather than the outputs of these expensive devices. In this context, researchers have worked on the applications of different data mining and machine learning algorithms on several publicly available CVD data sets, as reviewed in [6]. Although existing studies present valuable insights, there is no generally accepted and standardized machine-learning model for CVD diagnosis. Besides, for the CVD diagnosis problem, none of these studies present

a detailed performance evaluation of different classification methods and feature selection algorithms in terms of accuracy, sensitivity, precision, F-measure, and area under the curve. This thesis aims to fulfill this gap and show that for CVD diagnosis, not only one single performance measure is essential, but also other performance metrics, such as sensitivity, precision, F-measure, and area under the curve need to be evaluated. And to the best of our knowledge, none of the existing studies offer a single model, which works on different CVD data sets.

To address these problems, in this thesis, a single model is developed for two publicly available CVD data sets. Seven different feature selection methods and one proposed feature selection method based on domain knowledge are applied on these two CVD data sets. Seven single classification algorithms, and one ensemble (voting) classification algorithm is used to create a model that can help the diagnosis of CVD and can help medical doctors, especially in LMCs.

The rest of this thesis is organized as follows: Chapter 2 provides an overview of the current machine-learning based CVD diagnosis literature, used methods, and presents a table, which summarizes and compares the results that are obtained in the existing studies. In Chapter 3 introduces CVD data sets, and explains the applied methods. Chapter 4 presents the performance evaluations of our results, which are obtained with different methods. Chapter 5 presents the details of our proposed method. Chapter 6 compares our findings with previous studies and discusses advantages and disadvantages of the proposed method, and finally, Chapter 7 concludes the thesis via emphasizing the outcomes of this study and via stating the future work.

Chapter 2

Literature Review

As of 2019, 149 research articles have been published on machine learning-based CVD diagnosis [6]. Figure 2.1 provides detailed information about the frequency of these studies in different years. The first CVD diagnosis study based on machine learning is published in 1992 [7]. Since 2008, there has been a significant increase in the number of articles, which focus on the applications of machine learning to CVD diagnosis. Although a plethora of studies provide valuable information about CVD symptoms, there is no generalized and accepted model to predict CVD. Besides, all of these existing studies focus on single performance evaluation metric, and they do not provide a detailed performance evaluation in terms of sensitivity, precision, F-measure, and AUC.

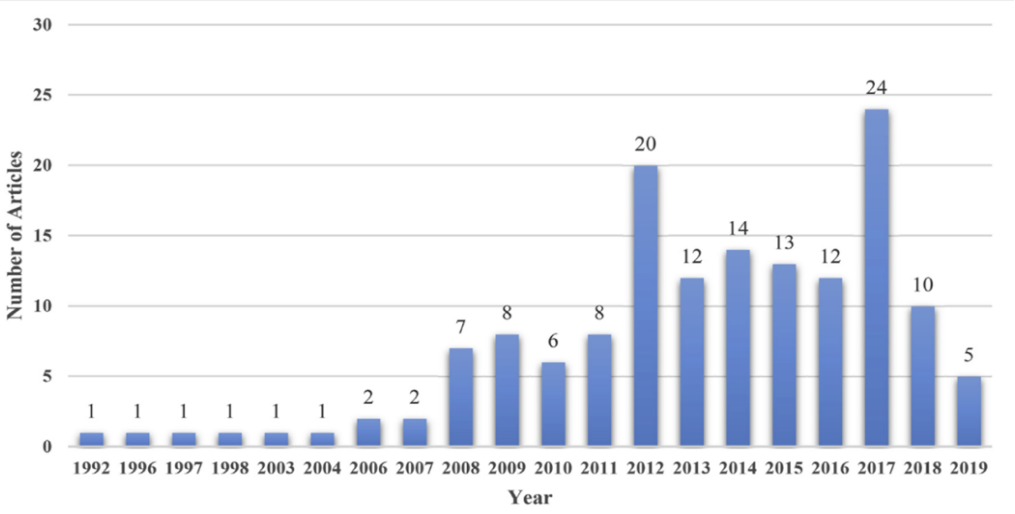

Figure 2.1 Numbers of research articles published on CVD by years

Authors	Year	Data set	Method	SN (%)	SP (%)	FM	AUC	ACC (%)
Akay [7]	1992	RWJUH	NN	84	89	-	-	86
Kemal Polat et al [8]	2007	Cleveland	KNN	92.30	92.30	-	-	87
Sellappan et al [9]	2008	UCI	NB	-	-	-	-	95
Resul Das et al [10]	2009	Cleveland	ANN	80.95	95.91	-	-	89.01
Anbarasi et al [11]	2010	UCI	DT	-	-	-	-	99.2
Shouman et al [12]	2011	Cleveland	DT	77.90	85.20	-	-	84.1
Alizadehsani et al [13]	2012	Alizadehsani	SMO	97.22	79.31	-	-	92.09
Alizadehsani et al [14]	2013	Alizadehsani	Bagging	83.05	74.60	-	-	79.54
Rajalaxmi et al [15]	2014	Cleveland	BABC	-	-	-	-	86.4
El-Biary et al [16]	2015	Cleveland	DT	-	-	-	-	78.54
Verma et al [17]	2016	Cleveland	MLR	-	-	-	-	90.28
Frantisek Babic et al [18]	2017	Alizadehsani	DT	-	-	-	-	86.67
Kolukisa et al [19]	2018	Alizadehsani	SVM	95.8	85.1	0.950	0.904	92.74
Redd et al [22]	2019	UCI	RF	88	95.4	-	-	92.16

SN: Sensitivity, SP: Specificity, FM: F-Measure, AUC: Area Under Curve, ACC: Accuracy

Table 2.1 List of different classification methods used for CVD diagnosis by years

Regardless of the performance-metrics used in these studies, the performance criteria should be evaluated based on the used data set, since data sets have different sample sizes and features. So far, the studies that focus on heart diseases have used 67 different data sets [6]. While the smallest data set consists of 20 samples and 9 features collected from Indian facilities, the largest data set has 24,000 samples and 11 features. A comparison of different classification methods for CVD diagnosis by years is given in Table 2.1.

Akay [7] et al. study was the pioneer study, which presented a CAD diagnosis based on machine learning in 1992. This study analyzes the benefit of using Neural Networks for diagnosing CAD using clinical features and extracting information from diastolic heart sounds. This study is conducted on 63 unhealthy and 37 healthy samples. The data set is collected at the Cardio dynamics Laboratory of Robert Wood Johnson University Hospital (RWJUH), NJ, USA. This study reported that the Neural Network correctly identified CAD with a

performance of 84% sensitivity and 89% specificity. This system also achieved an overall accuracy of 86% for the CAD data set [7].

Polat et al. [8] used a fuzzy weighted method based on k-nearest neighbor algorithm as a pre-processing step and Artificial Immune Recognition Systems (AIRS) as a classifier with 15-fold cross-validation. On the UCI Cleveland data set, the authors reported an accuracy of 87% for diagnosing CAD [8].

Sellappan et al. [9] developed an Intelligent Heart Disease Prediction System (IHDPS), which is implemented on .NET platform, using data mining techniques. The authors obtained the data set from the Cleveland Heart Disease Database [28], which contains 13 features and 909 samples. The authors tested three different classification methods and the best result is obtained with Naive Bayes with an accuracy of 95%.

Das et al. [10] presented an ensemble neural network method as the basis of their proposed system and used SAS (Statistical Analysis System) basic software. The authors reported that the ensemble-based approach created more effective models by combining the prior probabilities of predicted values from multiple primitive models. Although three independent neural network models were used as an ensemble model, the performance result was not improved. The authors achieved 89.01% classification accuracy from the UCI Cleveland heart disease data set.

Anbarasi et al. [11] used a data set containing a total of 13 features and 909 samples from UCI data set. The number of features reduced to six by using the Genetic Algorithm to determine the features with a high impact on Cardiovascular disease. After working with three different classification algorithms, the authors reported that the best result is obtained from the decision trees with 99.2% accuracy.

Shouman et al. [12] investigated different Decision Tree techniques to achieve better results for diagnosing heart diseases. The authors proposed a model that outperforms the J4.8 Decision Tree and Bagging algorithm in heart disease diagnosis. Performance metrics such as sensitivity, specificity, and accuracy results were compared, and the highest accuracy of 84.1% is obtained with the frequency discretion Gain Rate Decision Tree method.

Z-Alizadehsani et al. [13] applied a cost-sensitive CAD diagnosis algorithm named MetaCost on the Z-Alizadehsani data set. Initially, a particular feature selection method was applied to reduce the number of features to 34. Then, three new features (LAD, LCX, RCA) were created using feature extraction methods. Decision Tree, kNN, Naïve Bayes, SVM, and Sequential Minimal Optimization (SMO) algorithms are used in MetaCost. The proposed solution is reported to generate high sensitivity of 97.22%, and accuracy of 92.09%, which makes SMO algorithm better than the other alternatives.

Z-Alizadehsani et al. [14] reported that most people who come to the hospital with chest pain do not have cardiovascular disease, and therefore do not require angiography. This article aims to diagnose CADs with a lower cost and non-invasive method, using data analysis and data mining. For this reason, they investigate the accuracy of electrocardiographic (ECG) and medical test parameters to predict CAD patients in the need of angiography. They achieved 79.54% accuracy on the Left Anterior Descending (LAD) artery with bagging algorithm via Information Gain Feature Selection.

Rajalaxmi et al. [15] designed an algorithm that can remove irrelevant features from the data set and get more accurate performance metrics. A Binary Artificial Bee Colony algorithm is used as the feature selection and Naive Bayesian classifier is used during classification. The authors reported that BABC–Naive Bayesian obtained 84.4% accuracy on the UCI Cleveland data set.

El-Bialy et al. [16] attempted to apply machine-learning analysis to the UCI machine-learning repository, which has four CVD sub-data sets. The authors tested pruned decision tree and fast decision tree. Characteristic features are extracted from the UCI repository using combined UCI data set. As a result, the classification accuracy of the best-selected features on the combined data set is reported as 78.06%, which is higher than the average of all the UCI CVD sub-data sets.

Verma et al. [17] presented a novel hybrid method to diagnose CAD. For dimension reduction, the authors used a correlation-based feature subset (CFS) selection with particle swarm optimization (PSO) method. For classification, the authors used decision tree (C4.5), multi-layer perceptron (MLP), multinomial logistic regression (MLR), and fuzzy unordered rule induction algorithm (FURIA). The authors tested this approach on a data set consisting of 335 samples and 26 features. The authors reported that MLR achieved 88.4% accuracy. Also, the authors tested their approach on the UCI Cleveland heart disease data set. Their proposed hybrid method via MLP achieved an accuracy of 90.28%, outperforming other machine learning techniques.

Frantisek et al. [18] analyzed three publicly available CVD data sets: UCI, South African Heart Disease, and Z-Alizadehsani data sets. The authors performed both predictive and descriptive analyses. In the predictive analysis, Decision Trees, Naive Bayes, Support Vector Machine, and Neural Network classifiers are used to decide whether a person has a heart disease. In the descriptive analysis, the association and decision rules are used to extract steps to support decisions during the diagnosis. The authors achieved 89.93% accuracy with Neural Networks on UCI Cleveland data set, 73.70% accuracy with SVM on South Africa Heart Disease, and 86.67% accuracy on Z-Alizadehsani data set.

In our previous studies, different data sets were analyzed using linear discriminant analysis and a new hybrid feature selection method via classification

techniques. The goal was to reduce the computational cost by reducing the number of features and to generate a model that performs satisfactory results on each data set. Using an ensemble feature selection method with MLP classifier, we achieved 88.11% and 82.50% accuracy values on Z-Alizadehsani and Cleveland data sets, respectively. The best accuracy value of 92.74% is obtained with Fisher Linear Discriminant Analysis (FLDA) via SVM classifier on the Z-Alizadehsani data set [19,20]. Additionally, to our other studies, we proposed a novel Self Optimized and Adaptive Ensemble Machine Learning Algorithm for the diagnosis of CVD, the system automatically selects the most effective machine learning models without any pre-processing and feature selection method. We achieved 88.38% and 83.43% accuracy values on Z-Alizadehsani and Cleveland data sets, respectively [21].

Reddy et al. [22] proposed to combine all five CVD data sets in the UCI machine-learning repository. They filled in the in the missing values by taking the average of the available data and hence obtained a data set including 1190 samples. The authors used three different percentage splits for classification methods. After reducing the size of the features to 8, they achieve consistent performance results, but the results degrade when the feature size is reduced to 6. The authors report that the best result of 92.36% accuracy is obtained with Random Forest Classifier using the raw data set.

Chapter 3

Materials and Methods

3.1 Data sets

In this thesis, two publicly available CVD data sets, i.e., UCI Cleveland and Z-Alizadehsani data sets from the UCI machine-learning repository are analyzed.

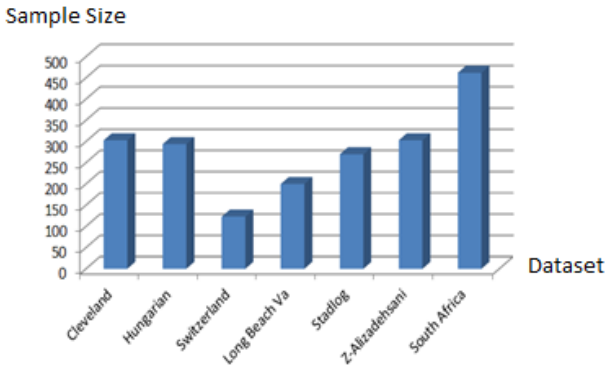

Figure 3.1.1 Publicly available CVD data sets

The characteristics of publicly available CVD data sets are shown in Figure 3.1.1. The UCI Heart Disease Data set contains 4 sub-data sets: UCI Cleveland, VA Long Beach, Hungary, and Switzerland, which were collected in Cleveland, Budapest, Zurich, and Basel in 1998. These data sets contain a total of 76 attributes, but 14 of these attributes are used throughout this thesis, as shown in Table 3.1.2. Although all UCI data sets are still actively used in current studies, the UCI Cleveland data set is the most commonly used one since other data sets have so many missing values. The UCI Cleveland data set was collected by Detrano [23]. The UCI Cleveland data set contains 303 samples, of which only

six samples have missing values, which are ignored in this study, instead of applying data correction. Each sample in this data set is labeled as either (i) healthy people with vessels narrowed less than 50%, or (ii) the ones that have CAD.

Data set	Attribute	CVD	Controls	Total
Z-Alizadehsani	55	216	87	303
UCI Cleveland	14	165	138	303

Table 3.1.1 The distribution of class labels in CVD data sets

The Z-Alizadehsani data set includes 303 data samples and 55 features, which are divided into the following 4 groups, namely “Demographic”, “Symptom and Examination”, “ECG” and “Lab and Eco”. This data set was collected at the Shaeheed Rajaei Cardiovascular, Medical, and Research Center in Tehran. Tables 3.1.2 and 3.1.3 show the details of the features included in Z-Alizadehsani data sets. For two CVD data sets, the number of samples having CVD (cases) and controls are show in Table 3.1.1.

No	Attribute- Description	Value
1	Age	29 - 77
2	Sex	M, F
3	CP (Typical, Atypical, Non-Anginal Pain, Asymptomatic)	1,2,3,4
4	Trestbps (Resting Blood Pressure)	94 - 200
5	Chol (Serum Cholesterol in mg/dl)	126 - 564
6	Fbs (Fasting Blood Sugar > 120)	Yes, No
7	Rectecg (Resting Electrocardiographic)	0,1,2
8	Thalach (Maximum Heart Rate Achieved)	71 - 202
9	Exang (Exercise Induced Angina)	Yes, No
10	Oldpeak (ST Depression Induced by Exercise Relative to Rest)	0 – 6.2
11	Slope (The Slope of The Peak Exercise ST Segment)	1,2,3
12	Ca (Number of Major Vessels Colored by Fluoroscopy)	0,1,2,3
13	Thal (Normal, Fixed Defect, Reversible Defect)	3,6,7
14	Num (Diagnosis of Heart Disease)	Yes, No

Table 3.1.2 The description of the attributes in the UCI Cleveland data set

No	FT	Attribute - Description	Values
1	Demographic	Age	30-86
2		Weight	48-120
3		Length	140-188
4		Sex	M, F
5		BMI (Body Mass Index Kg/m^2)	18-41
6		DM (Diabetes Mellitus)	Yes, no
7		HTN (Hypertension)	Yes, no
8		Current Smoker	Yes, no
9		Ex-Smoker	Yes, no
10		FH (Family History)	Yes, no
11		Obesity (MBI > 25)	Yes, no
12		CRF (Chronic Renal Failure)	Yes, no
13		CVA (Cerebrovascular Accident)	Yes, no
14		Airway Disease	Yes, no
15		Thyroid Disease	Yes, no
16		CHF (Congestive Heart Failure)	Yes, no
17		DLP (Dyslipidemia)	Yes, no
18	Symptom and examination	BP (Blood Pressure mmHg)	90 – 190
19		PR (Pulse Rate ppm)	50-110
20		Edema	Yes, No
21		Weak Peripheral Pulse	Yes, No
22		Lung Rales	Yes, no
23		Systolic Murmur	Yes, no
24		Diastolic Murmur	Yes, no
25		Typical Chest Pain	Yes, no
26		Dyspnea	Yes, no
27		Function Class	1,2,3,4
28		Atypical	Yes, no
29		Nonanginal CP	Yes, no
30		Exertional CP (Exertional Chest Pain)	Yes, no
31		Low Th Ang (Low Threshold Angina)	Yes, no

FT: Feature Type

Table 3.1.3 The description of the attributes in the Z-Alizadehsani data set

No	FT	Attribute - Description	Values
32	ECG	Q Wave	Yes, no
33		ST Elevation	Yes, no
34		ST Depression	Yes, no
35		T inversion	Yes, no
36		LVH (Left Ventricular Hypertrophy)	Yes, no
37		Poor R progression (poor R wave progression)	Yes, no
38		BBB	-
39	Laboratory and echo	FBS (Fasting Blood Sugar in mg/dl)	62–400
40		Cr (Creatine in mg/dl)	0.5–2.2
41		TG (Triglyceride in mg/dl)	37–1050
42		LDL (Low Density Lipoprotein in mg/dl)	18-232
43		HDL (High Density Lipoprotein in mg/dl)	15 -111
44		BUN (Blood Urea Nitrogen in mg/dl)	6–52
45		ESR (Erythrocyte Sedimentation Rate in mm/h)	1–90
46		HB (Hemoglobin in g/dl)	8.9–17.6
47		K (Potassium in mEq/lit)	3.0–6.6
48		Na (Sodium in mEq/lit)	128–156
49		WBC (White Blood Cell in cells/ml)	3700–18,000
50		Lymph (Lymphocyte in %)	7–60
51		Neut (Neutrophil in %)	32–89
52		PLT (Platelet in 1000/ml)	25–742
53		EF-TTE (Ejection Fraction in %)	15–60
54		Region RWMA	0,1,2,3,4
55		VHD (Valvular Heart Disease)	1-4

FT: Feature Type, RWMA: Regional Wall Motion Abnormality

Table 3.1.4 The description of the attributes in the Z-Alizadehsani data set (continued)

3.2 Performance Evaluation Metrics

Accuracy is an essential criterion for performance evaluation results and shows the overall effect of the model. Most of the current studies aim to improve the accuracy of CVD diagnosis, but the accuracy criterion may not be adequate in medical studies. In CVD diagnosis, other performance metrics are also critical. Sensitivity is a metric that indicates how many of the actual CVD patients are correctly labeled by the model as a patient. Precision metric indicates how many of those labeled as CVD by the model are actually CVD. These are important

details to be examined in the medical field. Therefore, to overcome the pitfalls of interpreting a particular performance metric, other measures like sensitivity and precision need to be examined.

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Table 3.2.1 Traditional Confusion Matrix

Traditional Confusion Matrix helps to assess the performance of the model in several aspects, as shown in Table 3.2.1. The definitions of the terms in traditional confusion matrix are given below.

True Positive (TP): Prediction of a sick person as sick
True Negative (TN): Prediction of a healthy person as healthy
False Positive (FP): Prediction of a healthy person as sick
False Negative (FN): Prediction of a sick person as healthy

Accuracy: Accuracy is the ratio of the correctly predicted labels to the total number of predictions made, as shown in Eqs (3.2.1). Focusing on accuracy criteria can work well when there are equal number of samples in each class, if it is a balanced data set. If there is a vast difference in the number of samples between two classes, the model can achieve high performance by predicting the largest sample class. On the other hand, when the sample sizes of two classes are close to each other, the accuracy rate will probably decrease. For some settings, this may not be a real problem, however, while working with fatal diseases, other parameters, e.g., sensitivity, becomes vital. The consequences of not diagnosing a sick person are far more severe than sending a healthy person to the test. For these reasons, it is necessary to consider several performance evaluation metrics for CVD diagnosis.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FN} + \text{FP} + \text{FN}) \quad \textbf{(3.2.1)}$$

Sensitivity: Sensitivity corresponds to the ratio of positive data points that are correctly predicted as positive, to all the positive data, as shown in Eqs (3.2.2). It is also called True Positive Rate (TPR) or Recall. In the context of the CVD data set, it shows how many of the labels are correctly predicted among all people with CVD, which is a very critical performance evaluation result for CVD.

$$\text{Sensitivity, Recall} = \text{TP}/(\text{TP} + \text{FN}) \quad \textbf{(3.2.2)}$$

Specificity: Specificity corresponds to the ratio of positive data points incorrectly predicted as positive, to all negative data, as shown in Eqs (3.2.3). It is also called False Positive Rate (FPR). In the context of CVD data set, it shows how many labels are correctly estimated among all the healthy people.

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \quad \textbf{(3.2.3)}$$

Precision: Precision corresponds to the ratio of correctly labelled sick patients to all sick-labelled patients, as shown in Eqs (3.2.4). In the context of CVD data set, this metric indicates how many of those labelled as CVD are actually CVD.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad \textbf{(3.2.4)}$$

F-Measure: F-Measure is a combination of Precision and Sensitivity metrics under a single parameter, as shown in (3.2.5). F-score, which is known as F1, represents the harmonic mean of these two metrics, which uses the weight of two metrics equally. It shows how sensitive and how robust the classifier is.

$$\text{F} - \text{Measure} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)} \quad \textbf{(3.2.5)}$$

Area Under Curve: A receiver operating characteristic curve (ROC) is a widely used technique to visualize the performance of a binary classifier and analyze the behavior of a diagnostic system. Area Under Curve (AUC) is calculated after placing the TPR and FPR metrics in the coordinate system (x, y). AUC is one of the best ways to summarize the performance of a ROC in a single number.

3.3 Data Mining

Data Mining is the process of finding potentially meaningful data from previously unknown patterns in a large data set, finding out relationships among data, and creating appropriate models for decision support mechanisms that will make accurate predictions. Today, data mining is used in many areas such as marketing, IoT, anomaly detection, banking, medical studies, sports, etc. Especially in recent years, data mining methods have been widely used in healthcare due to the vast amount of the data produced by the health sector.

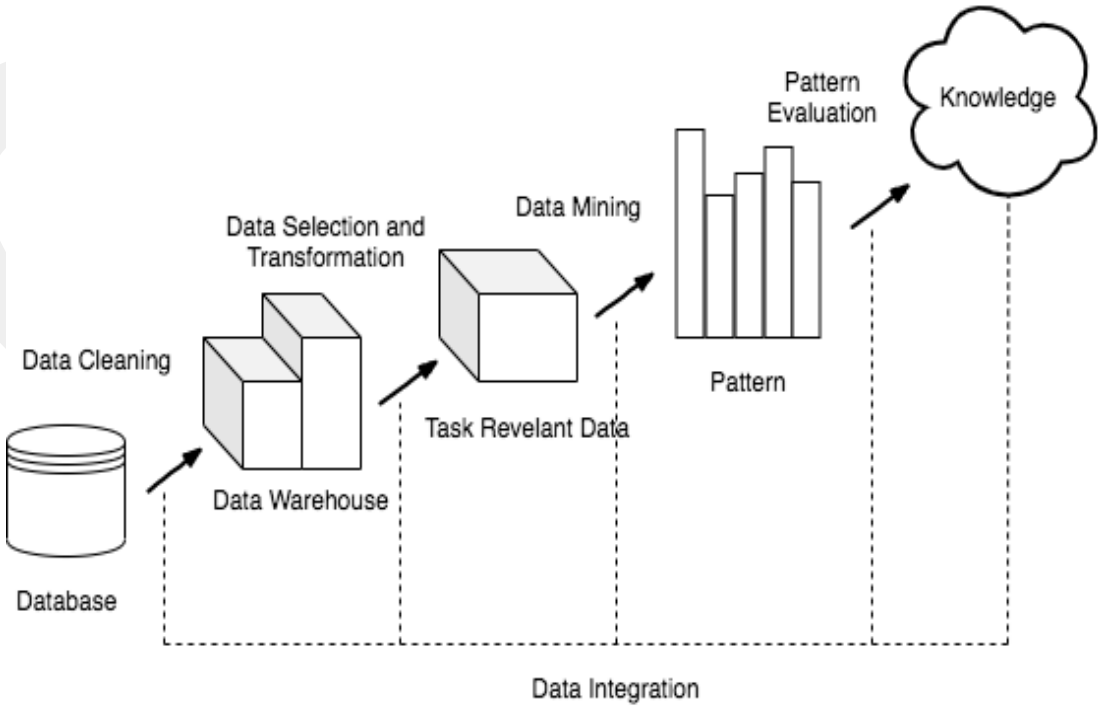

Figure 3.3.1 The Knowledge discovery of databases steps

Data mining is the analysis phase of the "Knowledge discovery of databases" (KDD) process. As shown in Figure 3.2, the KDD steps are as following:

Data Cleaning: Data can be incomplete, noisy, or inconsistent. Data cleaning defined as the removal of noisy and irrelevant data from the collection and filling out missing values.

Data Integration: Data mining often requires data integration, in which heterogeneous data from various sources are combined in a Data Warehouse. Correct integration can help to reduce/prevent redundancy and inconsistencies in the data set. It is critical for the performance evaluation results. Data integration uses several tools or Extract-Load-Transformation (ETL) process to integrate into the target system. This process is repeatable and traceable.

Data Selection: Data selection is the process of deciding the type and source of data to be analyzed and to be collected.

Data Transformation: Data Transformation converts the data to the appropriate form as required by mining procedures, which may be more efficient, and the patterns may be easier to understand.

Data Mining: Data mining process aims to extract potentially useful patterns in a particular representational form that decides the purpose of the model using classification or characterization.

Pattern Evaluation: Pattern Evaluation is the identification of the interesting patterns that represents the knowledge, based on given measures.

Knowledge Representation: This step provides data mining results to the users with a support of visualization and information representation techniques. It generates reports, tables, classification rules, etc.

KDD is an iterative process in which steps can be improved. New data can be integrated and transformed to achieve more relevant results. Preprocessing steps are performed in the steps of data cleaning and data integration [24].

3.4 Machine Learning

Machine Learning (ML) is a method that provides inferences from present data by using mathematical and statistical methods that makes predictions about the unknown. It has become popular as the amount of available data increases and access to these data gets easier. There are five main machine learning categories, as shown in Figure 3.4.1.

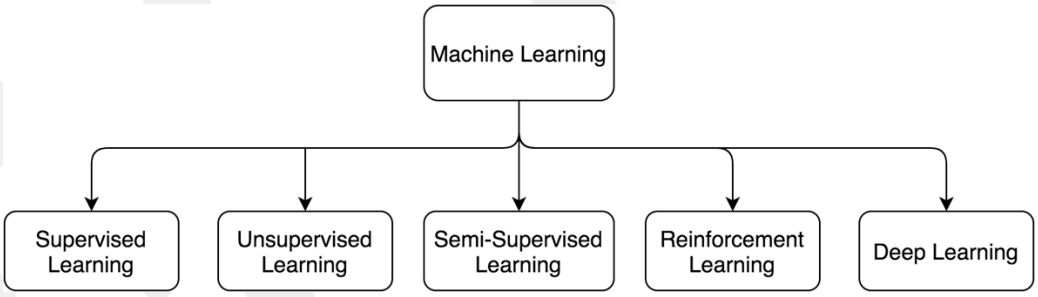

Figure 3.4.1 The machine learning sub-categories

3.4.1 Supervised Learning

Supervised learning generates a function that maps inputs to classes. This function is obtained using various classification and regression methods. Using the generated function, supervised learning models try to estimate which data point belong to which category, by processing new incoming data. Examples of supervised learning algorithms are support vector machine, multi-layer

perceptron, and regression. It may produce excellent results if the system has enough data to train [25].

3.4.2 Unsupervised Learning

Unsupervised learning is used on unlabeled data sets, and when the data set is massive. The goal is to model the hidden patterns in the given input data set, and it tries to determine which samples belong to which class using various unsupervised learning algorithms [25], such as clustering, k-means, and apriori algorithms. In massive data sets, the absence of labels could happen. In such situations, the first step is the use of unsupervised learning, and then supervised learning could be used.

3.4.3 Semi-Supervised Learning

Semi-supervised learning is a technique between supervised (labeled) and unsupervised (unlabeled) learning. Generally, the number of unlabeled data is considerably higher than the amount of labeled data. A small amount of labeled data among many unlabeled data can achieve good improvement on learning.

3.4.4 Reinforcement Learning

Reinforcement learning is a machine learning approach inspired by behaviorism, which is concerned with what actions a subject must take to achieve the highest amount of reward in an environment. It allows methods to automatically determine the ideal behavior in a given context to maximize performance. The main difference between Reinforcement Learning and Supervised/Unsupervised Learning is that; the latter one works with the data set, while the other one is working with trial and error [25].

3.4.5 Deep Learning

Deep Learning is designed by considering how the human brain works. It builds a programmable artificial neural network to make the right decision without the help of people. Unlike the sub-categories of machine learning, they do not

need guidance. Deep learning is a decision-making and learning structure in itself. Deep learning is widely used in the fields of image and speech recognition, computer vision.

3.5 Feature Selection

In the machine learning process, the performance of a model depends on the inputs, and the higher quality inputs are assured to generate higher quality outcomes. The feature selection adjusts the inputs using different methods. Feature selection removes redundant features, which are not related to the target variable, or which have no predictive power. The main objectives of feature selection methods are to eliminate noise, prevent overfitting, enable machine learning algorithms to run faster, reduce the complexity of the model, enable easy interpretation of the model, and improve performance results. Feature selection is grouped in three main categories as: i) The filter-based method, which is independent of the classifier; ii) The wrapper-based method, which interacts with the classifier; iii) The embedded method which combines the advantages of both methods, and performs feature selection and classification concurrently [41,42]. Each method has its advantages and disadvantages. If the data set is not large, then the embedded and wrapper methods should be preferred. If the data set is large, the filter method should be preferred.

Nowadays, new feature selection methods are emerging, and the number of feature selection methods is increasing. Among all these feature selection methods, it becomes difficult to decide which one is suitable for the data sets. However, among these feature selection methods, Chi-square, gain ratio, relief f methods have become popular [43]. In this thesis, chi-square (CS), gain ratio (GR), information gain (GS), relief f (RF), support vector machine (SVM), bee search (BS), conditional mutual information maximization (CMIM), and domain knowledge (DK) based feature selection methods are used. The proposed domain

knowledge-based feature selection method ranks the features based on the medical doctor's expertise, as described in detail in the Proposed Method section. All possible combinations of eight above-mentioned feature selection methods are ensembled in this thesis, as described in detail in the Proposed Method section. Hence, we tried to obtain the best ensemble features selection method, which achieves excellent results on two CVD data sets.

3.5.1 Chi-Square (CS)

Chi-Square (CS) is a well-known statistical hypothesis test, which is a univariate filter that organizes each feature independently by class. A contingency table is created with two selected features, and the observed, expected, and degrees of freedom values are calculated. With these values, the CS value is calculated, as shown in Eqs. (3.5.1.1). The importance of a feature depends on the CS value, and the higher the chi-square value, the greater the importance of the feature.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad \textbf{(3.5.1.1)}$$

c = degrees of freedom
O = observed value(s)
E = expected value(s)
i = combination of the values of two features

3.5.2 Information Gain (IG)

Information Gain (IG) is an entropy-based attribute selection measure, which is commonly used in decision trees. IG is a symmetrical measure and it ranks the features using the entropy criterion by using Eqs. (3.5.2.1, 3.5.2.2, 3.5.2.3). IG has a bias towards the features with more values, even when these features are more informative than the other features [44].

$$H(Y) = -\sum p(y) log_2(p(y)) \quad \textbf{(3.5.2.1)}$$

$$H(Y|X) = \sum p(x) \sum p(y|x) log_2(p(y|x) \quad \textbf{(3.5.2.2)}$$

$$IG(Y|X) = H(Y) - H(Y|X) \quad \textbf{(3.5.2.3)}$$

3.5.3 Gain Ratio (GR)

Gain Ratio (GR) is an adjustment of information gain that decreases its prejudice on high valued features that have no predictive power. GR adds the split information concept, and the GR value is calculated by the information gain divided to the split information value, as shown in Eqs. (3.5.3.1, 3.5.3.2).

$$GR = \frac{IG}{Split\ Info} \quad \textbf{(3.5.3.1)}$$

$$Split\ Info = \sum w\ log_2 w \quad \textbf{(3.5.3.2)}$$

3.5.4 Relief F (RF)

The relief algorithm is a filter-based feature selection approach and it is initially designed for binary classification problems with numerical or discrete features [26,27]. Contrary to some other algorithms, which are unaware of the contextual information, relief algorithm estimates the quality of features according to the relationships between features instead of acting independently. The extension of relief algorithm is relief-f (RF) algorithm, which could handle multiclass problems and it is more robust and able to deal with incomplete and noisy data [28]. RF feature selection selects top-ranking features from the data set by assigning different weights to each feature via comparing to its neighbors. The disadvantage of the relief algorithm is that when it tries to discern redundant features, it does not distinguish features even if they have very low relevance.

3.5.5 SVM Attribute Evaluation (SVM)

The support vector machine (SVM) classifier tries to separate data groups by drawing parallel lines between classes. It is an efficient algorithm, which is widely used in many different domains; and one of these domains is feature selection. SVM attribute evaluation is an embedded feature selection method, which is also known as recursive feature elimination for support vector machines, introduced by Guyon [29] in 2002, and it is firstly used for gene selection for cancer classification. SVM feature selection assigns the scores of the features by using the square of the weights obtained by the SVM classifier and removes the irrelevant features by training an SVM classifier iteratively. For multi-class problems, the selection of attributes is made by using a one vs. all method for each class.

3.5.6 Metaphor Search Methods: Bee Search (BS)

The artificial bee colony (ABC) is a swarm based meta-heuristic algorithm, which is the behavior of honeybees in search of food, these behaviors are modeled to solve optimization problems firstly by Dervis Karaboga [46] in 2005. There are different variations of the artificial bee colony algorithm, and it can be simulated as a neighbor search algorithm in its simplest form. The advantages of the ABC algorithm are easy to implement, flexible, and easy to control parameters [47]. In the Weka program, there is a package called metaphor search methods which contain nine methods, these methods are inspired by a specific animal species [30] and used for feature selection to obtain the best feature.

3.5.7 Conditional Mutual Information Maximization (CMIM)

Conditional Mutual Information Maximization (CMIM) feature selection method first ranks the features according to their conditional entropy and mutual information with the class to predict. Then it allows the addition of a new feature to the selected set of features if and only if the feature carries additional information.

3.6 Dimension Reduction

The goal of dimension reduction is to project the data set to a lower-dimensional area by reducing the variables and obtaining a set of principal variables with good class-separability to avoid over-fitting. Dimension reduction technique is widely used in statistics, machine learning, and information theory. This approach can be divided into two main categories. These are Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

3.6.1 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a commonly used dimension reduction technique, and it finds the feature subspace that optimizes the separation between classes. Linear Discriminant Analysis is a generalization of multiclass Fisher Linear Discriminant Analysis (FLDA). The purpose of LDA is to reduce the dimension, improve computational performance, and reduce overfitting in the models.

3.7 Classification Methods

In this thesis, seven single classifier techniques and one ensemble classifier is used. These seven techniques include the k-nearest neighbor, logistic regression, linear discriminant analysis, naive Bayes, support vector machine, multilayer perceptron, random forest, and one ensemble technique including voting classifiers.

Scikit-learn is a free machine learning software library that runs in the python programming language. It includes the implementations of several regression, clustering, and classification algorithms [45]. In data mining and machine learning studies, the data splitting process, cross-validation step and training-test sets drastically affect the performance results. In the stratified k-fold cross-validation method, the data set is divided into k parts and the classes are

distributed proportionally. While the k-1 part is used for training, and the other one part is used for testing, and this process is repeated k times. The values obtained in each round are summed up, and hence the performance of the model is evaluated. In this thesis, the following classifiers are applied in the python program using the scikit-learn library with the stratified 10-fold cross-validation method on the CVD data set.

3.7.1 k-Nearest Neighbor (kNN)

k-Nearest Neighbor (kNN) algorithm is a supervised machine learning algorithm that can be used in both classification and regression problems. Especially, it is one of the most widely used methods in classification. kNN algorithm can handle both continuous and discrete attributes. The principle behind nearest neighbor methods is to find and estimate the label from a predetermined number of training samples closest to the new point. For the metric distance, the standard Euclidean distance is the most common choice among many other distance definitions. In kNN, the sharpness between the classes began to be soft with the increase of k, the neighborhood number. If the number of classes is 2 in a data set, the k value is not recommended to surpass the square root of the sample size.

3.7.2 Logistic Regression (LR)

Logistic regression (LR) is a statistical machine learning algorithm that tries to define a logarithmic line that best distinguishes outcome variables on extreme ends. LR is the extended version of linear regression, where it allows us to build more complex decision boundaries by putting higher-order polynomials such as stochastic gradient descent. In this way, it is expected to achieve better results on complex data sets.

3.7.3 Linear Discriminant Analysis (LDA)

The generalized version of Fisher's linear discriminant is Linear discriminant analysis (LDA). LDA is a method used in several methods, such as statistics, pattern recognition, and machine learning. The resulting combination may be used as a linear classifier. LDA clearly tries to find the difference between the two or more classes. The aim of LDA is to prevent overfitting and to reduce cost.

3.7.4 Naïve Bayes (NB)

Naïve Bayes (NB) is a classification technique that utilizes both statistical and probabilistic methods. It is easy to build and performs well on large data sets, and also it adapts itself according to the type of data set. Naive Bayes classification method is a set of supervised learning algorithms based on strong hypotheses about the "naive" assumption of conditional independence of common variables in the application of Bayes' theory [33]. The NB classifier assumes there is independence between the conditional expectation variables on the response and the numerical distribution of the mean and standard deviation digital indicators from the training data set.

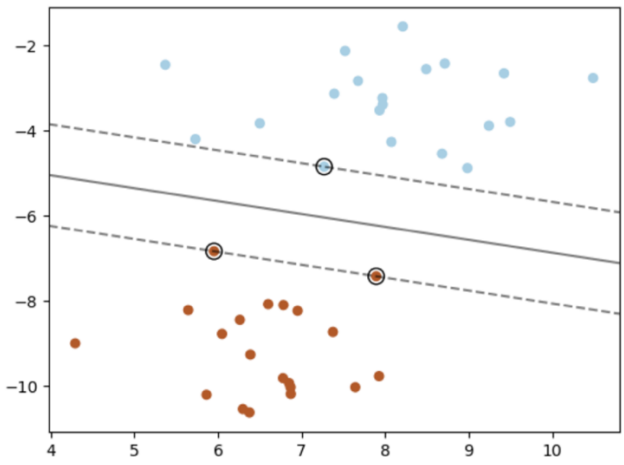

Figure 3.7.4.1 SVM method separates two classes by drawing two parallel lines

3.7.5 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a very efficient supervised learning method used for classification, regression, and outlier detection, which is utilized in many different domains. SVM has a simple method in which it tries to separate two groups by drawing two parallel lines between two classes [34], as shown in the Figure 3.7.4.1. While bringing lines closer, a common boundary line is obtained. This line is used as a decision boundary to separate the classes. SVM is effective in high-dimensional space and when the number of dimensions is greater than the number of samples. It is also efficient in terms of memory.

3.7.6 Multilayer Perceptron (MLP)

MLP is a classical type of Neural Network. It is suitable for classification prediction problems, where inputs are associated with a class. Multilayer perceptron is often applied to supervised learning problems. They train on a set of input-output pairs and learn to model the correlation between those inputs and outputs. There can be one or more non-linear layers, called hidden layers [35], as shown in Figure 3.7.6.1.

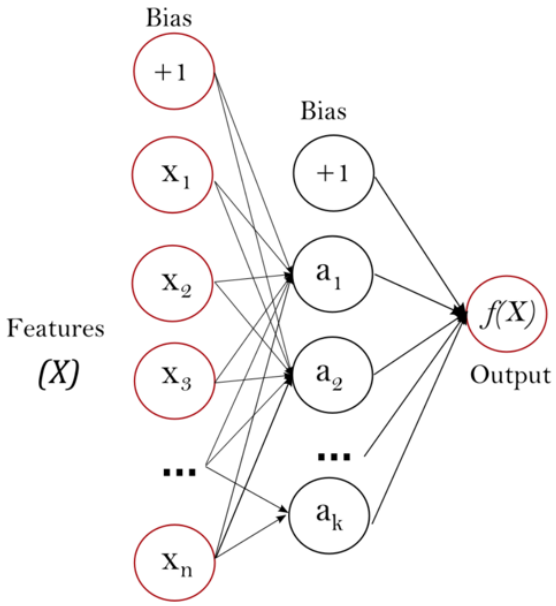

Figure 3.7.6.1 Representation of one hidden layer MLP

3.7.7 Random Forest (RF)

Random Forest is applicable to both regression and classification problems and is one of the most popular machine learning models. One of the problems in decision trees is overfitting, and RF tries to solve this problem of decision trees. Random Forest generates 10s or 100s of decision trees, and each decision tree makes an individual prediction. Via evaluating the predictions of individual decision tress. RF make the final decision as the majority of the predictions.

3.7.8 Ensemble Methods

In machine learning, the purpose of the embedded-based methods is to achieve better performance evaluations than single algorithms, using constituent algorithms alone to improve robustness over a single classifier and overcome overfitting [36,37]. There are different ensemble methods, which are bagging, boosting, voting, and so on. These approaches construct a new model, and then classify data points by taking a weighted average of each classifier's predictions.

Soft voting (Weighted Average Probabilities): This approach could be used when classifiers can estimate the probability of belonging to a class. It achieves the final result by averaging the probabilities obtained by the calculation of individual algorithms.

Hard voting (Majority Class Labels): This classifier generates the class labels via getting the majority of the votes assigned by each individual classifier.

Chapter 4

Proposed Methods

As summarized in Figure 4.1, in this thesis, the proposed model includes several feature selection techniques, dimension reduction, and several classifiers with parameter optimization on 10-fold cross-validation. In order to experiment with the proposed method, UCI Cleveland and Z-Alizadehsani data sets, which are obtained from the UCI machine learning repository are used. The proposed model is realized using Python and Weka.

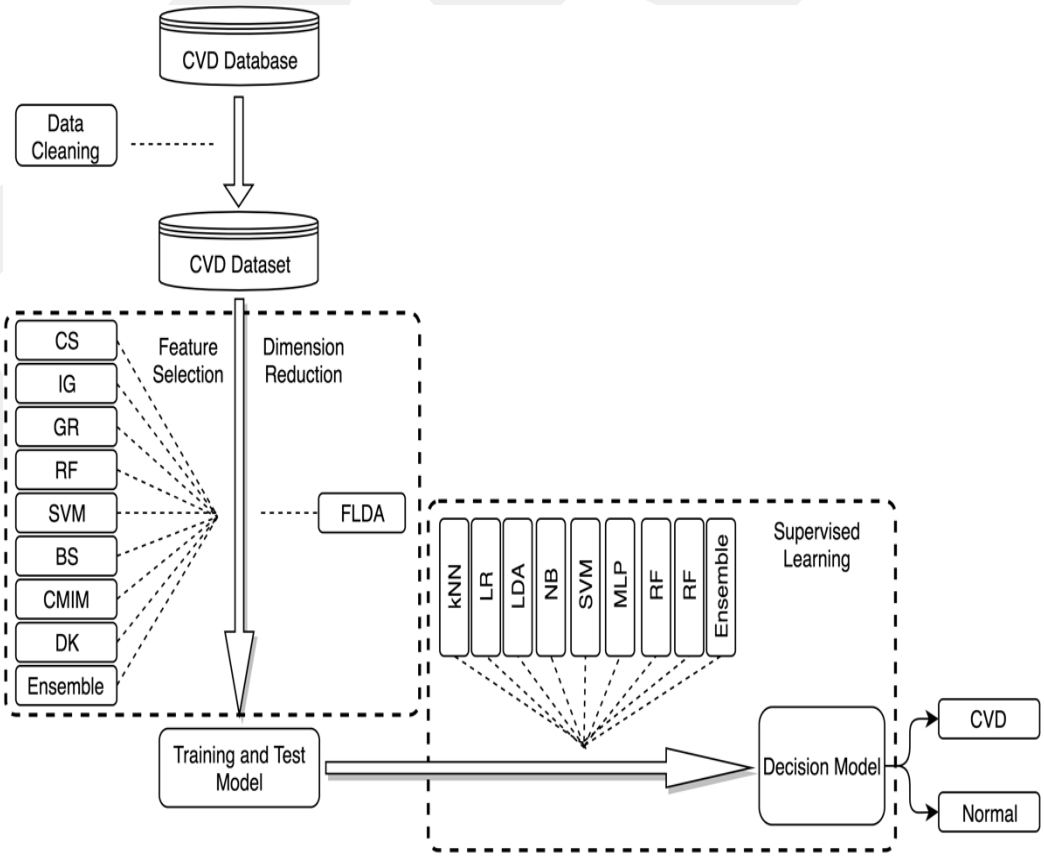

Figure 4.1 Schematic representation of the proposed model

As part of data cleaning, instead of filling six missing samples in the UCI Cleveland data set via synthetic data, these samples are removed from the data set. In addition to seven existing feature selection methods, we also apply our proposed feature selection method, which is based on the domain knowledge and the expertise of the cardiologists. To compensate the differences of feature scores between the proposed feature selection method and other feature selection techniques, the following two methods are conducted. (i) The ensemble feature selection methods are utilized with all sub-sets of eight different feature selection methods to find the best ranking of the features. (ii) The probabilistic score of each feature is obtained with 8 different feature selection methods. These probabilistic scores are further used to determine the selection rate of the features. For the classification task, we experimented seven single classifiers and one ensemble classification algorithm. For KNN, MLP and SVM Classifiers, we perform parameter optimization. For the Z-Alizadehsani data set, top 5-25 features are selected and trained, and for the Cleveland, data set, top 5-12 features are selected and trained. This realization of the proposed model could be better explained step by step as follows:

- Data cleaning: In the UCI Cleveland data set, samples containing missing data are excluded from the data sets.
- The feature scores are obtained with seven different feature selection methods.
- As the eighth feature selection method, we incorporate the domain-knowledge for feature selection task and asked expert cardiologists to rank these features.
- While some of the features achieve high scores in the ranking of the domain-knowledge based feature selection method, the same features achieve low scores in the other seven feature selection methods. To address this inconsistency, two alternative solutions are implemented.

  (i) The ensemble feature selection methods were tested with all sub-sets of eight different feature selection methods to find the best features ranks.

(ii) The probabilistic score of each feature is obtained as a result of 8 different feature selection methods. These probabilistic scores determine the selection rate of the features.

- In this way, in the first alternative method, the best features selected by the methods will be run in classification algorithms. In the second alternative method, it gives a chance to the features that are highlighted as essential in the domain-knowledge feature selection method, but are not identified as critical in other feature selection methods. In this way, the proposed algorithm enables to overcome local maximum.
- Linear discriminant analysis (LDA) is used to projects the data set to a lower-dimensional space by reducing the variables, LDA reduced CVD data set to one dimension.
- On Z-Alizadehsani data set, we experiment different classifiers via starting with the first five features and via increasing the number of features one by one until we include twenty-fifth feature. We repeat the same experiment on the UCI Cleveland data set, but this time until we include twelfth feature.
- The above-mentioned two alternative solutions (sub-set of all ensemble methods and probabilistic methods) are tested on Z-Alizadehsani data set since it has enough number of features (55 features). In the first proposed method, each classifier is tested with 255 different feature combinations. Here 255 refers to the number of all possible sub-sets of 8 different feature selection methods, as calculated in Eqs. (4.1). In the second proposed method, each classifier is tested 50 times to obtain the best performance results. In the UCI Cleveland data set, since it includes 13 features only.
- Seven single classifiers (k-nearest neighbor, logistic regression, linear discriminant analysis, naïve bayes, support vector machine, multilayer perceptron, and random forest) and one ensemble classifier (voting; hard, soft) is included in out experiments. Parameter optimization are conducted for SVM, MLP, kNN classifiers, and other classifiers are used with default parameters.

4.1 Domain Knowledge-Based Feature Selection

According to the medical literature, cardiovascular diseases are diagnosed via referring to the Framingham Heart Study (FHS), which is conducted at the University of Boston in 1948 and supported by the National Institute of Heart Lung and Blood (NHLBI). In this study, 5209 men and women are observed to determine the main factors that cause cardiovascular diseases. These participants went through physical examinations and lifestyle interviews to assess the relationship between cardiovascular diseases and other factors. In 2008, NHLBI generated a risk calculator utilizing several studies including FHS. The calculator determines ten years risk score for the cardiovascular disease using sex, age, total cholesterol, HDL cholesterol, untreated systolic blood pressure (SBP), treated SBP, current smoking status, and diabetes factors [31]. In the last 50 years, 1200 articles have been published in well-known medical journals, which refer to FHS while diagnosing CVD. Therefore, the FHS study is considered as the leading clinical practice, which is fundamental for cardiovascular disease diagnosis [32]. Also, our cardiologist collaborators examined the features of the two publicly available CVD data sets. Throughout the examination, they have determined the essential features according to their medical expertise. Throughout this thesis, we will refer to the features which are selected by cardiologists as "Clinically Important Features (CIF)". When scoring the features in the CVD data set according to the domain knowledge, the features, which are contained in the CIF and FHS are scored high, and the features which are not included in CIF and FHS are scored low. Therefore, while diagnosing CVD using data mining and machine learning, it could be evaluated whether the factors used in the computational model are compatible with cardiovascular medical literature.

4.2 Ensemble Feature Selection Method

Feature selection (FS) methods aims to identify essential features, in other words attributes, in a given data set. While higher scores are assigned to fundamental features, lower scores are assigned to useless features. Feature

selection step is critical to increase the calculation speed and to improve the performance results. However, not every feature selection method has the same performance in each data set. For this reason, it is better to define the beneficial properties by an ensemble feature selection method than a single method. Table 4.2.1 shows a simple illustration of how an ensemble feature selection is calculates scores for each attribute. Assume that three different feature selection algorithms (FS1, FS2, FS3) are performed on the data set. In this ensemble feature selection method, for each attribute, the average of the scores obtained for three different feature selection methods are calculated as a new score. As illustrated in Table 4.2.1, if a single FS method had been applied then attribute 2 and can be considered insignificant, whereas this attribute has a high score when ensemble FS method is applied. Also, the scores of the attribute 5 are low in all three FS methods and the score of the ensemble FS results is also low, indicating that there is no need to hesitate when removing the attribute 5 from the data set.

Attribute	FS1	FS2	FS3	Score of Attribute
Attribute 1	5	4	2	(5+4+2) / 3 = 3.66
Attribute 2	1	5	5	(1+5+5) / 3 = 3.66
Attribute 3	4	1	4	(4+1+4) / 3 = 3.00
Attribute 4	3	3	3	(3+3+3) / 3 = 3.00
Attribute 5	2	2	1	(2+2+1) / 3 = 1.66

FS: Feature Selection

Table 4.2.1 Toy Example of the Ensemble Feature Selection Methodology

While scoring the attributes with ensemble FS method, in order to scan the overall search space (different combinations of 8 feature selection methods), all different sub-sets of different feature selection methods are tested. As shown in Eqs. (4.1), 255 different sub-sets of 8 different feature selection methods are generated and ensembled.

$$\binom{8}{1} + \binom{8}{2} + \binom{8}{3} + \cdots + \binom{8}{8} = 2^8 - 1 \quad \textbf{(4.1)}$$

Features are ranked from the highest score to lowest score according to the obtained ensembles scores. Hence, a total of 255 feature ranking lists have emerged, as shown in Figure 4.2.1. This process is carried out on Z-Alizadehsani and UCI Cleveland data sets separately. As shown in Figure 4.2.2, for each feature-ranking list (out of 255 ranking lists), different numbers of features are tested in our experiments. While the number of tested features ranges from top 5 to 13 for the UCI Cleveland data set, it ranges from top 5 to top 25 for the Z-Alizadehsani data set. We would like to remind that while the UCI Cleveland data set includes 13 features, and the Z-Alizadehsani data set includes 42 additional features (55 features in total). After the top 25 features, the addition of further features did not change the performance significantly for the Z-Alizadehsani data set. Hence, for each feature-ranking list among 255 ranking lists, we realize our classification experiments with a maximum of 13 and 25 features, for UCI Cleveland and Z-Alizadehsani data sets, respectively.

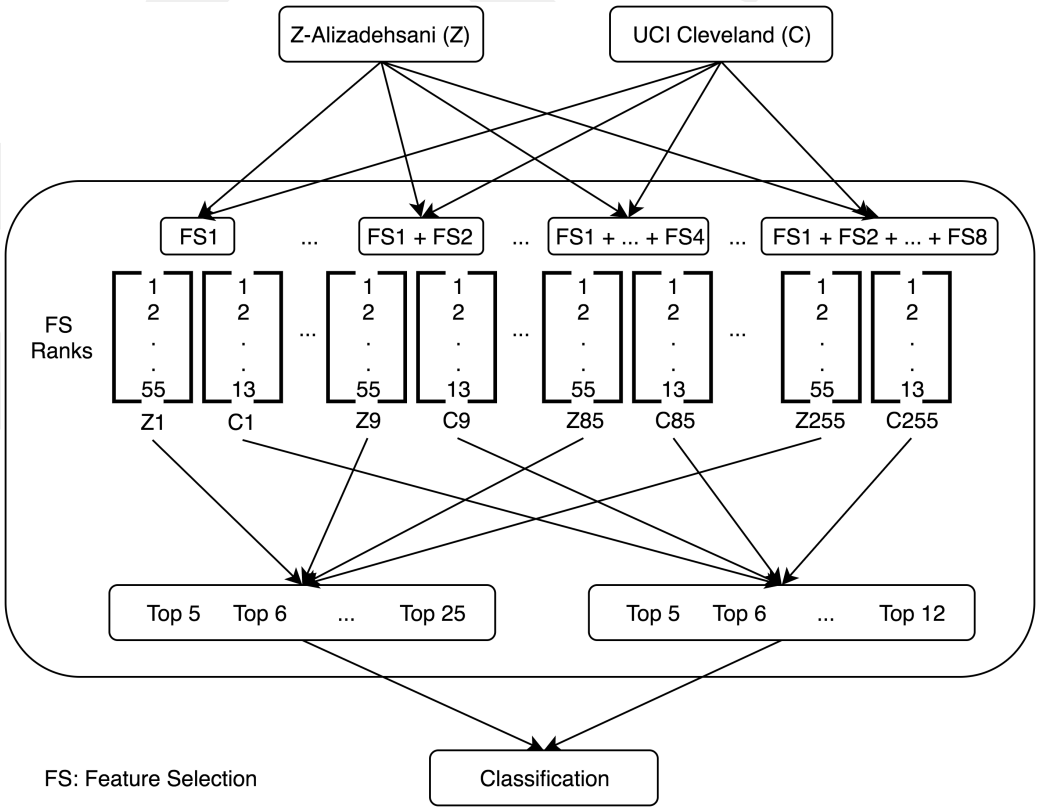

Figure 4.2.1 Schematic representation of the proposed ensemble feature selection method

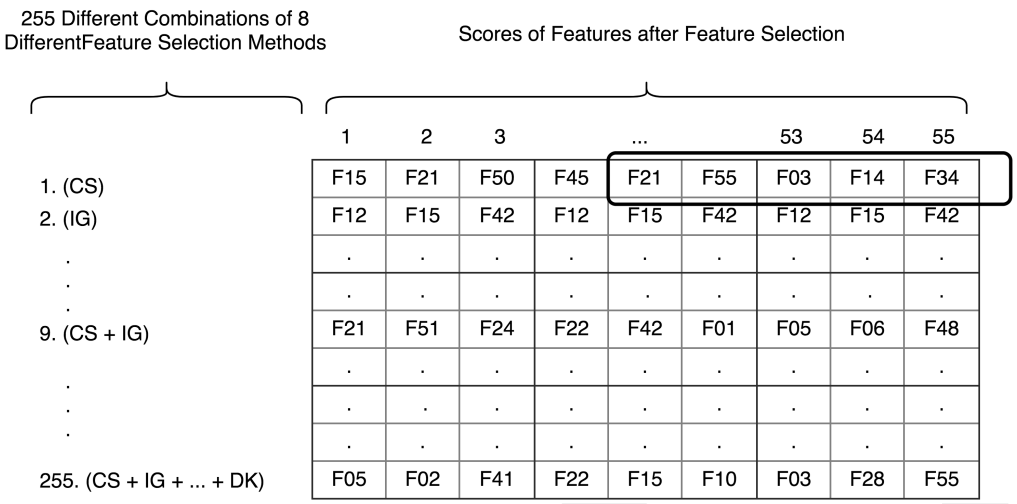

Figure 4.2.2 Visualization of Top 5 Feature Selection Process for in 255 different combinations of subsets of 8 different feature selection methods.

4.3 Probabilistic Feature Selection Method

Probabilistic feature selection is applied on the Z-Alizadehsani data set, to give a chance to the features, which have low scores but could have a positive effect on the performance results. The purpose of this feature selection method is to overcome the problem of getting stuck in the local maximum and to incorporate the features to the classification processes, if the medical specialists consider these features necessary, but computational feature selection methods do not consider important.

Attribute	Score of Attribute	Probabilistic Score
Attribute 1	3.66	3,66 / (3.66+3.66+3.00+1.66) = 0.305
Attribute 2	3.66	3.66 / (3.66+3.66+3.00+1.66) = 0.305
Attribute 3	3.00	3.00 / (3.66+3.66+3.00+1.66) = 0.250
Attribute 4	3.00	3.00 / (3.66+3.66+3.00+1.66) = 0.250
Attribute 5	1.66	1.66 / (3.66+3.66+3.00+1.66) = 0.138

Table 4.3.1 Toy Example of the Proposed Probabilistic Feature Selection Methodology

As explained in Section 4.2, all possible sub-sets of eight FS methods have been generated in the ensemble feature selection method. Among these 8 FS

methods, while 7 FS methods are computational FS methods, the last one refers to the feature selection method based on domain knowledge, as explained in Section 4.1. When the rankings of the features are examined, it is seen that some features obtain high scores from medical experts but have lower scores in other feature selection methods. For example, Externotinal CP, Ex-smoker, and LDL features are critical for CVD diagnosis. However, these features got lower ranks when computational feature selection techniques are applied.

In order to give a chance to lower scored features, the probabilistic ensemble FS method is proposed as following. In this approach, the feature scores are considered as a selection probability for each feature. During the calculation of attribute scores in the ensemble FS method, all of the eight feature selection rankings are averaged together, as illustrated in Table 4.2.1. The probabilistic score is computed by dividing the ensemble score into the overall score. An example of the calculation of the probabilistic score in the probabilistic FS method is shown in Table 4.3.1.

Chapter 5

Performance Evaluation

In this thesis, we conduct the performance evaluations of the proposed method on two publicly available CVD data sets, i.e., Z-Alizadehsani and UCI Cleveland data sets. We applied eight different feature selection techniques, dimension reduction, and we tested seven different classification algorithms for the diagnosis of the CVD using stratified 10-fold cross-validation.

5.1 Feature Selection

In this section, we present the performance evaluations of all feature selection methods. The findings were shown in tables with their ranks and score values. Firstly, we present the feature rankings and scores obtained via seven different feature selection methods. Secondly, we present the feature rankings and scores obtained via ensembling all possible combinations of eight feature selection methods (seven computational feature selection method and one domain knowledge-based feature selection method). Thirdly, we present the feature rankings and scores obtained via probabilistic feature selection approach, which aims to find out the features that have low scores but could have a positive effect on the performance of the classifier.

CS, IG, GR, RF, and SVM feature selection methods rank the features from 55 to 1, and from 13 to 1, for Z-Alizadehsani and UCI Cleveland data sets, respectively. In our proposed method, while the highest-ranked feature (ranked 1st) gets the biggest score (55), the lowest-ranked feature (ranked 55th) gets the smallest score (1). BS, CMIM and DK feature selection methods rank relevant features and the other features are not ranked. In our proposed method, for the

features, which are not ranked, the remaining score is divided by the number of the remaining number of features, and each feature gets the same score.

5.1.1 Chi-Square

Chi-Square (CS) feature selection method is used on two publicly available data sets. For the Z-Alizadehsani data set, CS method scores the relevant features in the order of importance from 55 to 1. CS scores of the attributes for the Z-Alizadehsani data set are shown in Table 5.1.1.1. For the UCI Cleveland data set, CS method scores the relevant features in the order of importance from 13 to 1. CS scores of the attributes for the UCI Cleveland data set are shown in Table 5.1.1.2.

Rank	Attribute	Score	Rank	Attribute	Score
1	Typical Chest Pain	55	**29**	FBS	27
2	Atypical	54	**30**	Weak Peripheral Pulse	26
3	Region RWMA	53	**31**	CR	25
4	HTN	52	**32**	HB	24
5	EF-TTE	51	**33**	BUN	23
6	Nonanginal	50	**34**	LVH	22
7	DM	49	**35**	Edema	21
8	Tinversion	48	**36**	Na	20
9	VHD	47	**37**	PLT	19
10	St Depression	46	**38**	Length	18
11	Age	45	**39**	BMI	17
12	Q Wave	44	**40**	CHF	16
13	BP	43	**41**	LDL	15
14	Diastolic Murmur	42	**42**	Lung rales	14
15	St Elevation	41	**43**	FH	13
16	Lymph	40	**44**	Thyroid Disease	12
17	Dyspnea	39	**45**	ESR	11
18	HDL	38	**46**	Neut	10
19	Poor R Progression	37	**47**	EX-Smoker	9
20	Function Class	36	**48**	Obesity	8
21	TG	35	**49**	Weight	7
22	PR	34	**50**	WBC	6
23	K	33	**51**	LowTH Ang	5
24	Current Smoker	32	**52**	Systolic Murmur	4
25	Airway disease	31	**53**	DLP	3
26	CRF	30	**54**	CVA	2
27	Sex	29	**55**	Exertional CP	1
28	BBB	28			

Table 5.1.1.1 Rankings and the scores of the attributes that are obtained using CS feature selection method for the Z-Alizadehsani data set

Rank	Attribute	Score	Rank	Attribute	Score
1	Thal	13	8	Restecg	6
2	Cp	12	9	Thalach	5
3	Ca	11	10	Chol	4
4	Exang	10	11	Age	3
5	Slope	9	12	Trestbps	2
6	Sex	8	13	FBS	1
7	Oldpeak	7			

Table 5.1.1.2 Rankings and the scores of the attributes that are obtained using CS feature selection method for the UCI Cleveland data set

5.1.2 Information Gain

Information Gain (IG) feature selection method is used on two publicly available data sets. For the Z-Alizadehsani data set, IG method scores the relevant features in the order of importance from 55 to 1. IG scores of the attributes for the Z-Alizadehsani data set are shown in Table 5.1.2.1. For the UCI Cleveland data set, IG method scores the relevant features in the order of importance from 13 to 1. IG scores of the attributes for the UCI Cleveland data set are shown in Table 5.1.2.2.

Rank	Attribute	Score	Rank	Attribute	Score
1	Typical Chest Pain	55	**29**	Lung rales	27
2	Atypical	54	**30**	Thyroid Disease	26
3	Region RWMA	53	**31**	CVA	25
4	Age	52	**32**	Obesity	24
5	EF-TTE	51	**33**	DLP	23
6	HTN	50	**34**	Systolic Murmur	22
7	DM	49	**35**	Weight	21
8	BP	48	**36**	WBC	20
9	Nonanginal	47	**37**	BMI	19
10	Tinversion	46	**38**	Length	18
11	FBS	45	**39**	Current Smoker	17
12	ESR	44	**40**	EX-Smoker	16
13	VHD	43	**41**	Na	15
14	K	42	**42**	FH	14
15	Q Wave	41	**43**	HB	13
16	St Elevation	40	**44**	PR	12
17	St Depression	39	**45**	Edema	11
18	Poor R Progression	38	**46**	Exertional CP	10
19	Diastolic Murmur	37	**47**	TG	9
20	Dyspnea	36	**48**	CR	8
21	CRF	35	**49**	LDL	7
22	Weak Peripheral Pulse	34	**50**	PLT	6
23	Airway disease	33	**51**	Function Class	5
24	BBB	32	**52**	BUN	4
25	LowTH Ang	31	**53**	HDL	3
26	Sex	30	**54**	Lymph	2
27	LVH	29	**55**	Neut	1
28	CHF	28			

Table 5.1.2.1 Rankings and the scores of the attributes that are obtained using IG feature selection method for the Z-Alizadehsani data set

Rank	Attribute	Score	Rank	Attribute	Score
1	Thal	13	8	Age	6
2	Cp	12	9	Sex	5
3	Ca	11	10	Restecg	4
4	Oldpeak	10	11	FBS	3
5	Exang	9	12	Trestbps	2
6	Thalach	8	13	Chol	1
7	Slope	7			

Table 5.1.2.2 Rankings and the scores of the attributes that are obtained using IG feature selection method for the UCI Cleveland data set

5.1.3 Gain Ratio

Gain Ratio (GR) feature selection method is used on two publicly available data sets. For the Z-Alizadehsani data set, GR method scores the relevant features in the order of importance from 55 to 1. GR scores of the attributes for the Z-Alizadehsani data set are shown in Table 5.1.3.1. For the UCI Cleveland data set, GR method scores the relevant features in the order of importance from 13 to 1. GR scores of the attributes for the UCI Cleveland data set are shown in Table 5.1.3.2.

Rank	Attribute	Score	Rank	Attribute	Score
1	Typical Chest Pain	55	**29**	LVH	27
2	Nonanginal	54	**30**	CVA	26
3	Atypical	53	**31**	Sex	25
4	Region RWMA	52	**32**	Obesity	24
5	Q Wave	51	**33**	DLP	23
6	St Elevation	50	**34**	Systolic Murmur	22
7	EF-TTE	49	**35**	EX-Smoker	21
8	Age	48	**36**	FH	20
9	Poor R Progression	47	**37**	Length	19
10	Diastolic Murmur	46	**38**	BUN	18
11	CRF	45	**39**	Current Smoker	17
12	Weak Peripheral Pulse	44	**40**	Weight	16
13	BP	43	**41**	BMI	15
14	HTN	42	**42**	HDL	14
15	DM	41	**43**	Exertional CP	13
16	LowTH Ang	40	**44**	Na	12
17	K	39	**45**	Lymph	11
18	Tinversion	38	**46**	PLT	10
19	CHF	37	**47**	Function Class	9
20	FBS	36	**48**	Neut	8
21	ESR	35	**49**	HB	7
22	Airway disease	34	**50**	TG	6
23	VHD	33	**51**	Edema	5
24	St Depression	32	**52**	WBC	4
25	BBB	31	**53**	LDL	3
26	Dyspnea	30	**54**	CR	2
27	Thyroid Disease	29	**55**	PR	1
28	Lung rales	28			

Table 5.1.3.1 Rankings and the scores of the attributes that are obtained using GR feature selection method for the Z-Alizadehsani data set

Rank	Attribute	Score	Rank	Attribute	Score
1	Ca	13	8	Sex	6
2	Thal	12	9	Age	5
3	Exang	11	10	Restecg	4
4	Thalach	10	11	FBS	3
5	Cp	9	12	Chol	2
6	Oldpeak	8	13	Trestbps	1
7	Slope	7			

Table 5.1.3.2 Rankings and the scores of the attributes that are obtained using GR feature selection method for the UCI Cleveland data set

5.1.4 Relief F

Relief f (RF) feature selection method is used on two publicly available data sets. For the Z-Alizadehsani data set, RF method scores the relevant features in the order of importance from 55 to 1. RF scores of the attributes for the Z-Alizadehsani data set are shown in Table 5.1.4.1. For the UCI Cleveland data set, RF method scores the relevant features in the order of importance from 13 to 1. RF scores of the attributes for the UCI Cleveland data set are shown in Table 5.1.4.2.

Rank	Attribute	Score	Rank	Attribute	Score
1	Typical Chest Pain	55	**29**	ESR	27
2	Atypical	54	**30**	Diastolic Murmur	26
3	HTN	53	**31**	FBS	25
4	DM	52	**32**	BP	24
5	Tinversion	51	**33**	HDL	23
6	Nonanginal	50	**34**	TG	22
7	Age	49	**35**	Thyroid Disease	21
8	Current Smoker	48	**36**	Na	20
9	DLP	47	**37**	FH	19
10	Dyspnea	46	**38**	CVA	18
11	VHD	45	**39**	St Depression	17
12	EF-TTE	44	**40**	CHF	16
13	Edema	43	**41**	Exertional CP	15
14	LVH	42	**42**	LowTH Ang	14
15	Region RWMA	41	**43**	LDL	13
16	Obesity	40	**44**	PLT	12
17	Sex	39	**45**	Lung rales	11
18	Weight	38	**46**	HB	10
19	Length	37	**47**	Weak Peripheral Pulse	9
20	Systolic Murmur	36	**48**	EX-Smoker	8
21	BMI	35	**49**	WBC	7
22	Neut	34	**50**	Poor R Progression	6
23	Function Class	33	**51**	K	5
24	Q Wave	32	**52**	CRF	4
25	BBB	31	**53**	PR	3
26	St Elevation	30	**54**	Airway disease	2
27	BUN	29	**55**	CR	1
28	Lymph	28			

Table 5.1.4.1 Rankings and the scores of the attributes that are obtained using RF feature selection method for the Z-Alizadehsani data set

Rank	Attribute	Score	Rank	Attribute	Score
1	Cp	13	8	Oldpeak	6
2	Thal	12	9	FBS	5
3	Sex	11	10	Thalach	4
4	Ca	10	11	Age	3
5	Slope	9	12	Trestbps	2
6	Exang	8	13	Chol	1
7	Restecg	7			

Table 5.1.4.2 Rankings and the scores of the attributes that are obtained using RF feature selection method for the UCI Cleveland data set

5.1.5 SVM Attribute Evaluation

SVM Attribute Evaluation (SVM) feature selection method is used on two publicly available data sets. For the Z-Alizadehsani data set, SVM method scores the relevant features in the order of importance from 55 to 1. SVM scores of the attributes for the Z-Alizadehsani data set are shown in Table 5.1.5.1. For the UCI Cleveland data set, SVM method scores the relevant features in the order of importance from 13 to 1. RF scores of the attributes for the UCI Cleveland data set are shown in Table 5.1.5.2.

Rank	Attribute	Score	Rank	Attribute	Score
1	Age	55	**29**	LVH	27
2	Region RWMA	54	**30**	Na	26
3	Typical Chest Pain	53	**31**	Poor R Progression	25
4	Tinversion	52	**32**	Airway disease	24
5	TG	51	**33**	Function Class	23
6	PR	50	**34**	FBS	22
7	St Elevation	49	**35**	BBB	21
8	DM	48	**36**	Atypical	20
9	Nonanginal	47	**37**	Weight	19
10	HTN	46	**38**	Obesity	18
11	FH	45	**39**	LowTH Ang	17
12	Lung rales	44	**40**	BP	16
13	Current Smoker	43	**41**	Edema	15
14	HB	42	**42**	CR	14
15	EF-TTE	41	**43**	Diastolic Murmur	13
16	Dyspnea	40	**44**	K	12
17	Q Wave	39	**45**	Lymph	11
18	DLP	38	**46**	WBC	10
19	BUN	37	**47**	Thyroid Disease	9
20	Sex	36	**48**	Neut	8
21	ESR	35	**49**	LDL	7
22	PLT	34	**50**	CVA	6
23	BMI	33	**51**	EX-Smoker	5
24	Length	32	**52**	Exertional CP	4
25	HDL	31	**53**	Weak Peripheral Pulse	3
26	St Depression	30	**54**	CHF	2
27	VHD	29	**55**	CRF	1
28	Systolic Murmur	28			

Table 5.1.5.1 Rankings and the scores of the attributes that are obtained using SVM feature selection method for the Z-Alizadehsani data set

Rank	Attribute	Score	Rank	Attribute	Score
1	Ca	13	8	Slope	6
2	Oldpeak	12	9	Trestbps	5
3	Thalach	11	10	Chol	4
4	Thal	10	11	FBS	3
5	Exang	9	12	Restecg	2
6	Cp	8	13	Age	1
7	Sex	7			

Table 5.1.5.2 Rankings and the scores of the attributes that are obtained using SVM feature selection method for the UCI Cleveland data set

5.1.6 Metaphor Search Methods: Bee Search

For the Z-Alizadehsani data set, BS feature selection method scores the relevant features in the order of importance from 55 to 47, and the remaining features are scored as 23. These results are shown in Table 5.1.6.1. For the Cleveland data set, Bee Search (BS) scores the relevant features in order of importance between 13 and 7, and the remaining features scored as 3 points, and the results are shown in Table 5.1.6.2.

Rank	Attribute	Score	Rank	Attribute	Score
1	Age	55	**29**	HDL	23
2	HTN	54	**30**	St Depression	23
3	Typical Chest Pain	53	**31**	VHD	23
4	Tinversion	52	**32**	Systolic Murmur	23
5	FBS	51	**33**	LVH	23
6	ESR	50	**34**	Na	23
7	K	49	**35**	Poor R Progression	23
8	EF-TTE	48	**36**	Airway disease	23
9	Region RWMA	47	**37**	Function Class	23
10	DM	23	**38**	BBB	23
11	BP	23	**39**	Weight	23
12	Atypical	23	**40**	Obesity	23
13	Nonanginal	23	**41**	LowTH Ang	23
14	Q Wave	23	**42**	Edema	23
15	TG	23	**43**	CR	23
16	PR	23	**44**	Diastolic Murmur	23
17	St Elevation	23	**45**	Lymph	23
18	FH	23	**46**	WBC	23
19	Lung rales	23	**47**	Thyroid Disease	23
20	Current Smoker	23	**48**	Neut	23
21	HB	23	**49**	LDL	23
22	Dyspnea	23	**50**	CVA	23
23	DLP	23	**51**	EX-Smoker	23
24	BUN	23	**52**	Exertional CP	23
25	Sex	23	**53**	Weak Peripheral Pulse	23
26	PLT	23	**54**	CHF	23
27	BMI	23	**55**	CRF	23
28	Length	23			

Table 5.1.6.1 Rankings and the scores of the attributes that are obtained using BS feature selection method for the Z-Alizadehsani data set

Rank	Attribute	Score	Rank	Attribute	Score
1	Cp	13	8	Age	6
2	Restecg	12	9	Sex	5
3	Thalach	11	10	Trestbps	4
4	Exang	10	11	Chol	3
5	Oldpeak	9	12	FBS	2
6	Ca	8	13	Slope	1
7	Thal	7			

Table 5.1.6.2 Rankings and the scores of the attributes that are obtained using BS feature selection method for the UCI Cleveland data set

5.1.7 Conditional Mutual Information Maximization

For the Z-Alizadehsani data set, Conditional Mutual Information Maximization (CMIM) feature selection method scores the relevant features in the order of importance from 55 to 43, and the remaining features are scored as 21. These results are shown in Table 5.1.7.1. For the Cleveland data set, CMIM scores the relevant features in order of importance between 13 and 7, and the remaining features scored as 3 points, and the results are shown in Table 5.1.7.2.

Rank	Attribute	Score	Rank	Attribute	Score
1	Age	55	**29**	St Depression	21
2	DM	54	**30**	VHD	21
3	HTN	53	**31**	Systolic Murmur	21
4	BP	52	**32**	LVH	21
5	Typical Chest Pain	51	**33**	Na	21
6	Atypical	50	**34**	Poor R Progression	21
7	Nonanginal	49	**35**	Airway disease	21
8	Q Wave	48	**36**	Function Class	21
9	Tinversion	47	**37**	FBS	21
10	ESR	46	**38**	BBB	21
11	K	45	**39**	Weight	21
12	EF-TTE	44	**40**	Obesity	21
13	Region RWMA	43	**41**	LowTH Ang	21
14	TG	21	**42**	Edema	21
15	PR	21	**43**	CR	21
16	St Elevation	21	**44**	Diastolic Murmur	21
17	FH	21	**45**	Lymph	21
18	Lung rales	21	**46**	WBC	21
19	Current Smoker	21	**47**	Thyroid Disease	21
20	HB	21	**48**	Neut	21
21	Dyspnea	21	**49**	LDL	21
22	DLP	21	**50**	CVA	21
23	BUN	21	**51**	EX-Smoker	21
24	Sex	21	**52**	Exertional CP	21
25	PLT	21	**53**	Weak Peripheral Pulse	21
26	BMI	21	**54**	CHF	21
27	Length	21	**55**	CRF	21
28	HDL	21			

Table 5.1.7.1 Rankings and the scores of the attributes that are obtained using CMIM feature selection method for the Z-Alizadehsani data set

Rank	Attribute	Score	Rank	Attribute	Score
1	Cp	13	8	Age	6
2	Restecg	12	9	Sex	5
3	Thalach	11	10	Trestbps	4
4	Exang	10	11	Chol	3
5	Oldpeak	9	12	FBS	2
6	Ca	8	13	Slope	1
7	Thal	7			

Table 5.1.7.2 Rankings and the scores of the attributes that are obtained using CMIM feature selection method for the UCI Cleveland data set

5.1.8 Domain Knowledge Based Feature Selection

For the Z-Alizadehsani data set, the cardiologists score the relevant features in the order of importance from 55 to 43, based on FHS and their own expertise. The remaining features are scored as 21, and the results are shown in Table 5.1.8.1. For the Cleveland data set, medical doctors score the relevant features in the order of importance from 13 to 7, and the remaining features are scored as 3 points, the results are shown in Table 5.1.8.2.

Rank	FT*	Attribute	Score	Rank	FT*	Attribute	Score
1	CIF	Typical Chest Pain	55	**29**	-	PR	21
2	CIF	Exertional CP	54	**30**	-	St Elevation	21
3	CIF	Q Wave	53	**31**	-	Lung rales	21
4	CIF	Region RWMA	52	**32**	-	HB	21
5	FHS RF	Age	51	**33**	-	Dyspnea	21
6	FHS RF	Sex	50	**34**	-	DLP	21
7	FHS RF	Weight	49	**35**	-	BUN	21
8	FHS RF	BMI	48	**36**	-	PLT	21
9	FHS RF	Obesity	47	**37**	-	St Depression	21
10	FHS RF	DM	46	**38**	-	VHD	21
11	FHS RF	FBS	45	**39**	-	Systolic Murmur	21
12	FHS RF	HTN	44	**40**	-	LVH	21
13	FHS RF	BP	43	**41**	-	Na	21
14	-	Current Smoker	21	**42**	-	Poor R Progression	21
15	-	EX-Smoker	21	**43**	-	Airway disease	21
16	-	FH	21	**44**	-	Function Class	21
17	-	LDL	21	**45**	-	Length	21
18	-	HDL	21	**46**	-	BBB	21
19	-	Weak Peripheral Pulse	21	**47**	-	LowTH Ang	21
20	-	CHF	21	**48**	-	Edema	21
21	-	CRF	21	**49**	-	CR	21
22	-	Atypical	21	**50**	-	Diastolic Murmur	21
23	-	Nonanginal	21	**51**	-	Lymph	21
24	-	Tinversion	21	**52**	-	WBC	21
25	-	ESR	21	**53**	-	Thyroid Disease	21
26	-	K	21	**54**	-	Neut	21
27	-	EF-TTE	21	**55**	-	CVA	21
28	-	TG	21				

Table 5.1.8.1 Rankings and the scores of the attributes that are obtained using DK feature selection method for the Z-Alizadehsani data set

Rank	FT*	Attribute	Score	Rank	FT*	Attribute	Score
1	CIF	Cp	55	**8**	FHS RF	Age	21
2	CIF	Exang	54	**9**	FHS RF	Sex	21
3	CIF	Oldpeak	53	**10**	-	Restecg	21
4	CIF	Thal	52	**11**	-	Thalach	21
5	FHS RF	Trestbps	51	**12**	-	Slope	21
6	FHS RF	Chol	50	**13**	-	Ca	21
7	FHS RF	Fbs	49	**14**	-	Age	21

Table 5.1.8.2 Rankings and the scores of the attributes that are obtained using DK feature selection method for the UCI Cleveland data set

5.1.9 Ensemble Feature Selection

The score of each attribute was obtained by the ensemble method, which combines eight different feature selection techniques, as described in Section 4.2. The combined score of each attribute is the average of the scores obtained from eight methods. When eight different feature selection techniques are used, the ensemble scores of each attribute are shown in Table 5.1.9.1 and 5.1.9.2, for the UCI Cleveland and Z-Alizadehsani data sets, respectively. We calculated ensemble scores of each attribute when different numbers (1 to 8) of feature selection techniques are applied. Hence, for each data set, 247 (255-8) ensemble scores are calculated for each attribute.

No	Attribute	CS	IG	GR	RF	SVM	BS	CMIM	DK	Score
1	Cp	12	12	9	13	8	13	13	13	7.15
2	Thal	13	13	12	12	10	7	7	10	6.46
3	Exang	10	9	11	8	9	10	10	12	6.07
4	Ca	11	11	13	10	13	8	8	2	5.84
5	Oldpeak	7	10	8	6	12	9	9	11	5.53
6	Thalach	5	8	10	4	11	11	11	2	4.76
7	Restecg	6	4	4	7	2	12	12	2	3.76
8	Sex	8	5	6	11	7	3	3	5	3.69
9	Slope	9	7	7	9	6	3	3	2	3.53
10	Age	3	6	5	3	1	3	3	6	2.30
11	FBS	1	3	3	5	3	3	3	7	2.15
12	Trestbps	2	2	1	2	5	3	3	9	2.07
13	Chol	4	1	2	1	4	3	3	8	2

Table 5.1.9.1 Ensemble scores of each attribute in the UCI Cleveland data set when eight different feature selection techniques are used

No	Attribute	CS	IG	GR	RF	SVM	BS	CMIM	DK	Score
1	Typical Chest Pain	55	55	55	55	53	53	51	55	54
2	Age	45	52	48	49	55	55	55	51	51.25
3	HTN	52	50	42	53	46	54	53	47	49.625
4	Region RWMA	53	53	52	41	54	47	43	52	49.375
5	DM	49	49	41	52	49	23	54	48	45.625
6	Tinversion	48	46	38	51	52	52	47	19	44.125
7	EF-TTE	51	51	49	44	41	48	44	19	43.375
8	Nonanginal	50	47	54	50	47	23	49	19	42.375
9	Q Wave	44	41	51	32	39	23	48	53	41.375
10	Atypical	54	54	53	54	20	23	50	19	40.875
11	BP	43	48	43	24	16	23	52	47	37
12	FBS	27	45	36	25	22	51	21	48	34.375
13	St Elevation	41	40	50	30	49	23	21	19	34.125
14	VHD	47	43	33	45	29	23	21	19	32.5
15	ESR	11	44	35	27	35	50	46	10	32.25
16	Dyspnea	39	36	30	46	40	23	21	19	31.75
17	Sex	29	30	25	39	36	23	21	50	31.625
18	Current Smoker	32	17	17	48	43	23	21	46	30.875
19	K	33	42	39	5	12	49	45	19	30.5
20	St Depression	46	39	32	17	30	23	21	19	28.375
21	Diastolic Murmur	42	37	46	26	13	23	21	19	28.375
22	Poor R Progression	37	38	47	6	25	23	21	19	27
23	BMI	17	19	15	35	33	23	21	49	26.5
24	LVH	22	29	27	42	27	23	21	19	26.25
25	Obesity	8	24	24	40	18	23	21	49	25.875
26	BBB	28	32	31	31	21	23	21	19	25.75
27	FH	13	14	20	19	45	23	21	45	25
28	DLP	3	23	23	47	38	23	21	19	24.625
29	HDL	38	3	14	23	31	23	21	43	24.5
30	Weight	7	21	16	38	19	23	21	49	24.25
31	Airway disease	31	33	34	2	24	23	21	19	23.375
32	Length	18	18	19	37	32	23	21	19	23.375
33	Lung rales	14	27	28	11	44	23	21	19	23.375
34	TG	35	9	6	22	51	23	21	19	23.25
35	Weak Peripheral Pulse	26	34	44	9	3	23	21	19	22.375
36	CRF	30	35	45	4	1	23	21	19	22.25
37	Systolic Murmur	4	22	22	36	28	23	21	19	21.875
38	BUN	23	4	18	29	37	23	21	19	21.75
39	Function Class	36	5	9	33	23	23	21	19	21.125
40	PR	34	12	1	3	50	23	21	19	20.375
41	CHF	16	28	37	16	2	23	21	19	20.25
42	Thyroid Disease	12	26	29	21	9	23	21	19	20
43	HB	24	13	7	10	42	23	21	19	19.875
44	Edema	21	11	5	43	15	23	21	19	19.75
45	Na	20	15	12	20	26	23	21	19	19.5
46	LowTH Ang	5	31	40	14	17	23	2	19	18.875
47	EX-Smoker	9	16	21	8	5	23	21	46	18.625
48	PLT	19	6	10	12	34	23	21	19	18
49	Lymph	40	2	11	28	11	11	21	19	17.875
50	Exertional CP	1	10	13	15	4	23	21	54	17.625
51	CVA	2	25	26	18	6	23	21	19	17.5
52	LDL	15	7	3	13	7	23	21	44	16.625
53	Neut	10	1	8	34	8	23	21	19	15.5
54	CR	25	8	2	1	14	23	21	19	14.125
55	WBC	6	20	4	7	10	23	21	19	13.75

Table 5.1.9.2 Ensemble scores of each attribute in the Z-Alizadehsani data set when eight different feature selection techniques are used

5.1.10 Probabilistic Feature Selection

This approach aims to find out the attributes that have low scores but could have a positive effect on the performance of the classifier and hence, we intend to avoid the local maximum. Following our proposed procedure as described in Section 4.2., we calculated the probabilistic scores for each attribute. For the Z-Alizadehsani data set, the probabilistic scores are shown in Table 5.1.10.1, when eight feature selection methods are used. We calculated probabilistic scores of each attribute when different numbers (1 to 8) of feature selection techniques are applied. Hence, 255 (2^8-1) probabilistic scores are calculated for each attribute.

No	Attribute	Score	P. S.	No	Attribute	Score	P. S.
1	Typical Chest Pain	54	0.035	**29**	HDL	24.5	0.015
2	Age	51.25	0.033	**30**	Weight	24.25	0.015
3	HTN	49.625	0.032	**31**	Airway disease	23.375	0.015
4	Region RWMA	49.375	0.032	**32**	Length	23.375	0.015
5	DM	45.625	0.029	**33**	Lung rales	23.375	0.015
6	Tinversion	44.125	0.028	**34**	TG	23.25	0.015
7	EF-TTE	43.375	0.028	**35**	Weak Peripheral Pulse	22.375	0.014
8	Nonanginal	42.375	0.027	**36**	CRF	22.25	0.014
9	Q Wave	41.375	0.026	**37**	Systolic Murmur	21.875	0.014
10	Atypical	40.875	0.026	**38**	BUN	21.75	0.014
11	BP	37	0.024	**39**	Function Class	21.125	0.013
12	FBS	34.375	0.022	**40**	PR	20.375	0.013
13	St Elevation	34.125	0.022	**41**	CHF	20.25	0.013
14	VHD	32.5	0.021	**42**	Thyroid Disease	20	0.013
15	ESR	32.25	0.021	**43**	HB	19.875	0.012
16	Dyspnea	31.75	0.020	**44**	Edema	19.75	0.012
17	Sex	31.625	0.020	**45**	Na	19.5	0.012
18	Current Smoker	30.875	0.020	**46**	LowTH Ang	18.875	0.012
19	K	30.5	0.019	**47**	EX-Smoker	18.625	0.012
20	St Depression	28.375	0.018	**48**	PLT	18	0.011
21	Diastolic Murmur	28.375	0.018	**49**	Lymph	17.875	0.011
22	Poor R Progression	27	0.017	**50**	Exertional CP	17.625	0.011
23	BMI	26.5	0.017	**51**	CVA	17.5	0.011
24	LVH	26.25	0.017	**52**	LDL	16.625	0.010
25	Obesity	25.875	0.016	**53**	Neut	15.5	0.010
26	BBB	25.75	0.016	**54**	CR	14.125	0.009
27	FH	25	0.016	**55**	WBC	13.75	0.008
28	DLP	24.625	0.016				

P.S: Probabilistic Score

Table 5.1.10.1 Probabilistic scores of each attribute in the Z-Alizadehsani data set when eight different feature selection techniques are used

5.2 Classification Methods

In our study, seven single classifiers (k-nearest neighbor, logistic regression, linear discriminant analysis, naïve bayes, support vector machine, multilayer perceptron, and random forest) and one ensemble classifier (voting; hard, soft) method were used. During our experiments, Python (with the scikit-learn library) programing language and stratified 10-fold cross-validation is used. Parameter estimation was performed using the grid search method in k-NN, MLP, and SVM classifiers. In other classifiers, default parameters were used. Various performance metrics were calculated for two different CVD data sets.

In this section, we present the performance evaluations of all classifiers and all feature selection methods. The findings were shown in tables with their sensitivity and accuracy values in different settings: (i) Using the raw data set, (ii) Using the data set after the FLDA method is applied, (iii) Using the data set after all subsets of the combination on the ensemble feature selection method is applied, (iv) Using the data set after each probabilistic feature selection method is applied. Also, the average results of ensemble feature selection and probabilistic feature selection methods are shown in the tables. The subset combination of the feature selection, which yields the highest performance results on both CVD data sets using the same classifier(s), is marked in bold. The features of the best performance results that are obtained from the probabilistic feature selection method are shown as tables in the appendix.

MLP classifier obtained the best performance result in both CVD data sets with different sub-sets of feature selection combination. LDA classifier generated one of the best performance results using the same sub-set combination of the feature selection method. NB classifier performed one of the lowest performance results in both CVD data sets. For these reasons, the performance metrics of MLP, LDA, and NB classifiers in various settings are shown in the figures according to their best results, average results, median results, worst result and raw results.

5.2.1 k-Nearest Neighbors

k-Nearest Neighbors (kNN) classifier is used in two different data sets, which are Z-Alizadehsani and UCI Cleveland. While running the kNN classifier, the number of neighborhoods (k) ranges from 3 to 11. With the Z-Alizadehsani data set, 84.86% accuracy, 92.87% sensitivity, 85.57% precision, 0.841 F-Measure, and 0.891 AUC is achieved when k is 9. The UCI Cleveland data set gives 84.41% accuracy, 81.37% sensitivity, 85.93% precision, 0.825 F-Measure, and 0.903 AUC when k is 11. Table 5.2.1.1 shows in detail the best results of the kNN classifier sorted with respect to accuracy values in various situations. Among different probabilistic feature selection runs, the results with the top three accuracies are listed in Table 5.2.1.1. When the feature selection methods or FLDA method are applied, better results compared to the raw data set are obtained with the kNN classifier, on both data sets. Also, in our experiments with kNN, we observed that when the number of features is smaller, the obtained results are better. Unfortunately, kNN performed lower than other tested classifiers.

Dt	Combination of Feature Selection	Att	SN (%)	Pre (%)	FM	AUC	Acc (%)
Z-Alizadehsani	CS+GR+RF	5	92.87	85.57	0.841	0.891	84.86
	CS+SVM+DK	9	87.85	90.62	0.888	0.843	84.41
	RF	9	92.55	86.28	0.829	0.858	84.15
	Average of All Possible FS Comb.	-	88.52	79.59	0.830	0.750	75.97
	Probabilistic FS of KNN 1	8	87.40	91.13	0.890	0.894	84.77
	Probabilistic FS of KNN 2	7	87.87	88.41	0.879	0.863	82.80
	Probabilistic FS of KNN 3	10	90.67	85.47	0.875	0.843	82.14
	Average of Probabilistic FS	-	90.16	74.85	0.815	0.619	71.10
	FLDA	1	88.09	94.81	0.866	0.923	86.54
	-	55	87.53	73.10	0.796	0.500	68.04
UCI Cleveland	CS+RF+SVM	5	81.37	85.93	0.825	0.903	84.41
	CS+GR+RF	6	82.36	83.59	0.822	0.892	83.77
	CS	8	80.65	84.35	0.817	0.905	83.74
	Average of All Possible FS Comb.	-	65.49	73.23	0.685	0.774	72.66
	FLDA	1	76.97	81.78	0.782	0.883	81.05
	-	-	62.14	65.74	0.635	0.699	68.00

Dt: Data set, Att: Number of Attributes included, SN: Sensitivity, Pre: Precision, FM: F-Measure, AUC: Area Under Curve, Acc: Accuracy

Table 5.2.1.1 Performance evaluations of kNN classifier on two CVD data sets when different feature selection techniques are applied

5.2.2 Logistic Regression

Logistic Regression (LR) classifier is applied on two different data sets, i.e., Z-Alizadehsani and UCI Cleveland. Using LR classifier, 91.11% accuracy, 94.39% sensitivity, 93.65% precision, 0.938 F-Measure, and 0.954 AUC values are achieved on the Z-Alizadehsani data set. The same classifier achieved 83.11% accuracy, 79.34% sensitivity, 83.78% precision, 0.811 F-Measure, and 0.902 AUC on the UCI Cleveland data set. Table 5.2.2.1 shows in detail the best results obtained with the LR classifier, sorted with respect to the accuracy values in various situations. Among different probabilistic feature selection runs, the results with the top three accuracies are listed in Table 5.2.2.1. LR classifier achieved one of the best performance results in both CVD data sets. The LR classifier performed poor results when the feature size is less. There is no significant difference between the results obtained with the raw data set and the data set after applying the feature selection method.

Dt	Combination of Feature Selection	Att	SN (%)	Pre (%)	FM	AUC	Acc (%)
Z-Alizadehsani	SVM	19	94.39	93.65	0.938	0.954	91.11
	SVM	18	93.91	93.62	0.935	0.953	90.76
	SVM	21	93.46	93.90	0.935	0.953	90.75
	Average of All Possible FS Comb.	-	92.36	90.87	0.909	0.936	87.85
	Probabilistic FS of LR 1	23	93.50	92.29	0.927	0.947	89.51
	Probabilistic FS of LR 2	16	93.44	92.24	0.926	0.941	89.43
	Probabilistic FS of LR 3	25	93.00	92.62	0.926	0.950	89.41
	Average of Probabilistic FS	-	90.56	84.43	0.871	0.844	80.76
	FLDA	1	1	75.57	0.857	1	75.78
	-	55	92.09	91.14	0.915	0.924	87.76
UCI Cleveland	RF+SVM+DK	10	79.34	83.78	0.811	0.902	83.11
	GR+RF+SVM+DK	10	79.34	83.78	0.811	0.903	83.11
	IG+RF+SVM+DK	10	78.62	83.73	0.807	0.903	82.78
	Average of All Possible FS Comb.	-	75.89	82.40	0.783	0.895	81.06
	FLDA	1	68.84	91.03	0.774	0.922	82.37
	-	55	78.51	82.26	0.796	0.901	81.74

Dt: Data set, Att: Number of Attributes included, SN: Sensitivity, Pre: Precision, FM: F-Measure, AUC: Area Under Curve, Acc: Accuracy

Table 5.2.2.1 Performance evaluations of LR classifier on two CVD data sets when different feature selection techniques are applied

5.2.3 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) classifier is applied on two different data sets, i.e., Z-Alizadehsani and UCI Cleveland. Using LDA classifier, 90.15% accuracy, 92.57% sensitivity, 93.83% precision, 0.930 F-Measure, and 0.924 AUC values are achieved on the Z-Alizadehsani data set. The same classifier achieved 83.41% accuracy, 78.51% sensitivity, 84.52% precision, 0.809 F-Measure, and 0.901 AUC on the UCI Cleveland data set. Table 5.2.3.1 shows in detail the best results obtained with the LR classifier, sorted with respect to the accuracy values in various situations. Among different probabilistic feature selection runs, the results with the top three accuracies are listed in Table 5.2.3.1. LDA classifier achieved one of the best performance results on both CVD data sets, with the ensemble (IG, SVM, BS, DK) feature selection method. When LDA classifier is applied, the effect of feature selection methods on the performance results is higher on the Z-Alizadehsani data set, compared to the UCI Cleveland data set.

Dt	Combination of Feature Selection	Att	SN (%)	Pre (%)	FM	AUC	Acc (%)
Z-Alizadehsani	**IG+SVM+BS+DK**	24	92.57	93.83	0.930	0.924	90.15
	SVM	20	92.55	93.82	0.930	0.941	90.13
	SVM	19	92.07	94.40	0.930	0.942	90.13
	Average of All Possible FS Comb.	-	90.84	91.41	0.906	0.924	87.58
	Probabilistic FS of LDA 1	24	90.71	94.51	0.924	0.920	89.47
	Probabilistic FS of LDA 2	24	91.10	94.13	0923	0.925	89.38
	Probabilistic FS of LDA 3	23	91.19	93.61	0.922	0.918	89.16
	Average of Probabilistic FS	-	88.77	85.53	0.867	0.841	80.55
	FLDA	1	90.47-	96.26	0.887	1	89.55
	-	55	87.92	91.97	0.897	0.902	85.79
UCI Cleveland	**IG+SVM+BS+DK**	11	78.51	84.52	0.809	0.901	83.41
	SVM+BS	10	77.80	84.48	0.806	0.901	83.08
	GR	8	78.62	83.49	0.806	0.907	82.77
	Average of All Possible FS Comb.	-	73.37	83.29	0.779	0.896	81.06
	FLDA	1	77.80	85.80	0.812	0.922	83.74
	-	55	74.37	83.29	0.779	0.896	81.06

Dt: Data set, Att: Number of Attributes included, SN: Sensitivity, Pre: Precision, FM: F-Measure, AUC: Area Under Curve, Acc: Accuracy

Table 5.2.3.1 Performance evaluations of LDA classifier on two CVD data sets when different feature selection techniques are applied

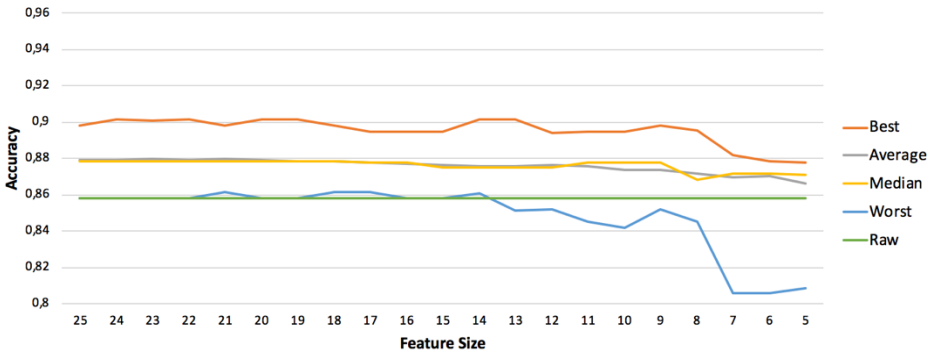

Figure 5.2.3.1 The accuracy of LDA classifiers when different numbers of features are used on Z-Alizadehsani data set

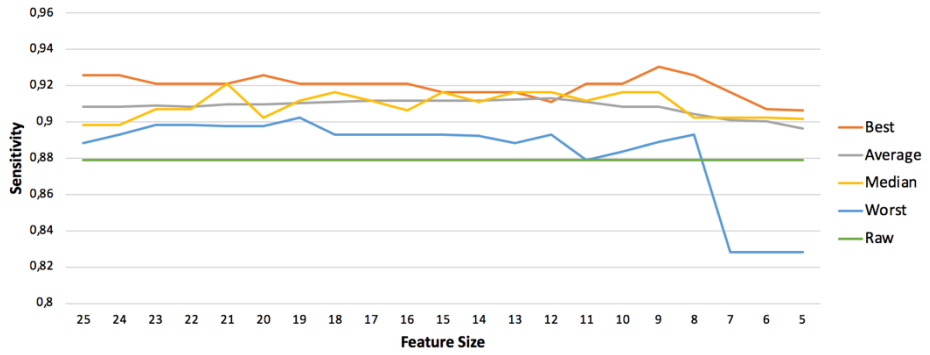

Figure 5.2.3.2 The sensitivity of LDA classifiers when different numbers of features are used on Z-Alizadehsani data set

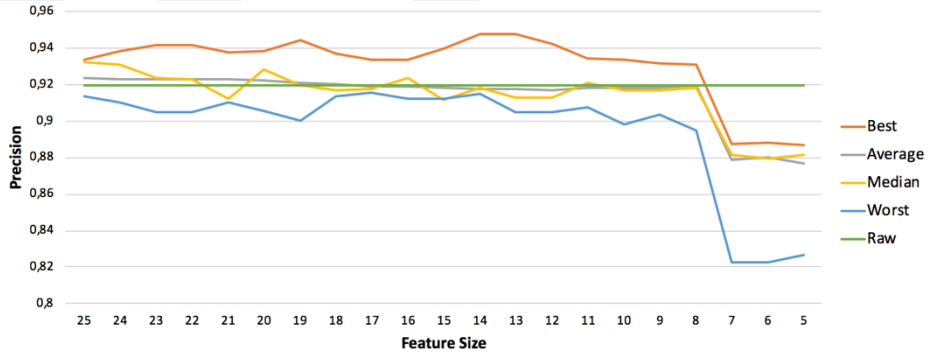

Figure 5.2.3.3 The precision of LDA classifiers when different numbers of features are used on Z-Alizadehsani data set

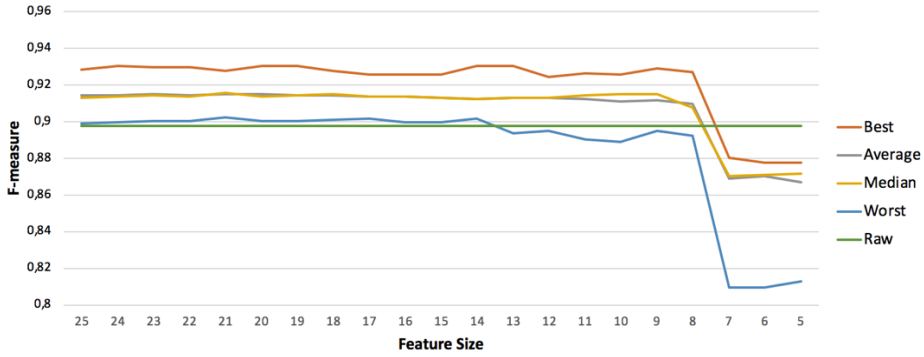

Figure 5.2.3.4 The F-measure of LDA classifiers when different numbers of features are used on Z-Alizadehsani data set

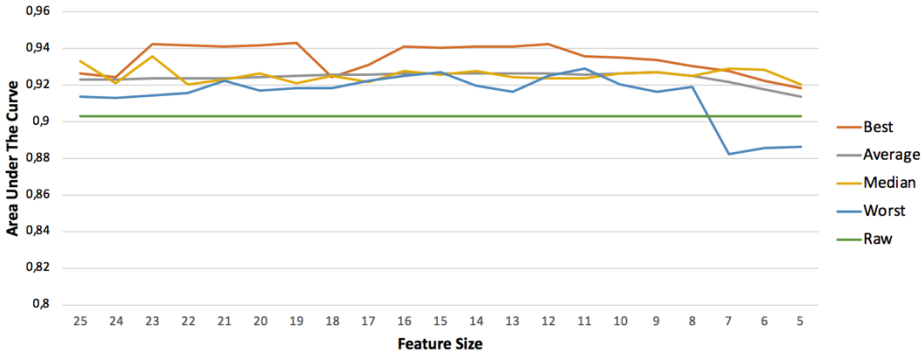

Figure 5.2.3.5 The AUC of LDA classifiers when different numbers of features are used on Z-Alizadehsani data set

Figure 5.2.3.1 plots the accuracies of LDA classifiers when different numbers of features are used on the Z-Alizadehsani data set. When the feature size is less than 14, the accuracy values decrease. Especially when the feature size is less than nine, the accuracy values decrease dramatically. Similarly, Figures 5.2.3.2, 5.2.3.3, 5.2.3.4 plot the sensitivity, precision, and F-measure of LDA classifiers respectively, when different numbers of features are used on Z-Alizadehsani data set. These plots have similar patterns with accuracy plots. As shown in Figure 5.2.3.5, the number of features size does not affect AUC performance until the feature size is less than 12. Once the feature size is less than 12, AUC values decrease. The LDA classifier is observed to perform poor results when the feature size is small. When the number of features is more than 15, the performance results of LDA classifier on CVD data set are stable.

5.2.4 Naïve Bayes

Naïve Bayes (NB) classifier is used in two different data sets, which are Z-Alizadehsani and UCI Cleveland. Using NB classifier, 87.42% accuracy, 90.17% sensitivity, 92.66% precision, 0.911 F-Measure, and 0.920 AUC values are achieved on the Z-Alizadehsani data set. The same classifier achieved 82.78% accuracy, 80.82% sensitivity, 81.67% precision, 0.811 F-Measure, and 0.892 AUC on the UCI Cleveland data set. Table 5.2.4.1 shows in detail the best results obtained with the NB classifier, sorted with respect to the accuracy values in various situations. Among different probabilistic feature selection runs, the results with the top three accuracies are listed in Table 5.2.4.1. When NB classifier is applied, the effect of feature selection methods on the performance results is much higher on the Z-Alizadehsani data set, compared to the UCI Cleveland data set (there were almost no improvement on the performance results by the FS methods).

Dt	Combination of Feature Selection	Att	SN (%)	Pre (%)	FM	AUC	Acc (%)
Z-Alizadehsani	RF+SVM+BS+DK	8	90.17	92.66	0.911	0.920	87.42
	RF+SVM+BS+DK	9	89.69	92.48	0.908	0.920	87.05
	IG	10	91.19	91.06	0.909	0.923	86.86
	Average of All Possible FS Comb.	-	57.12	94.96	0.680	0.914	66.90
	Probabilistic FS of NB 1	25	86.99	92.18	0.892	0.901	85.15
	Probabilistic FS of NB 2	5	95.38	85.52	0.901	0.863	84.91
	Probabilistic FS of NB 3	11	90.73	88.90	0.896	0.905	84.87
	Average of Probabilistic FS	-	43.39	90.05	0.527	0.815	55.82
	FLDA	1	90.47	96.26	0.887	1	89.55
	-	55	67.87	73.39	0.796	0.856	73.39
UCI Cleveland	RF+SVM+BS+DK	7	80.82	81.67	0.811	0.892	82.78
	IG+SVM	10	81.53	81.06	0.811	0.892	82.75
	BS	8	80.76	81.11	0.807	0.896	82.43
	Average of All Possible FS Comb.	-	79.31	80.41	0.796	0.888	81.54
	FLDA	1	77.80	85.80	0.812	0.922	83.74
	-	55	80.00	79.76	0.796	0.892	81.40

Dt: Data set, Att: Number of Attributes included, SN: Sensitivity, Pre: Precision, FM: F-Measure, AUC: Area Under Curve, Acc: Accuracy

Table 5.2.4.1 Performance evaluations of NB classifier on two CVD data sets when different feature selection techniques are applied

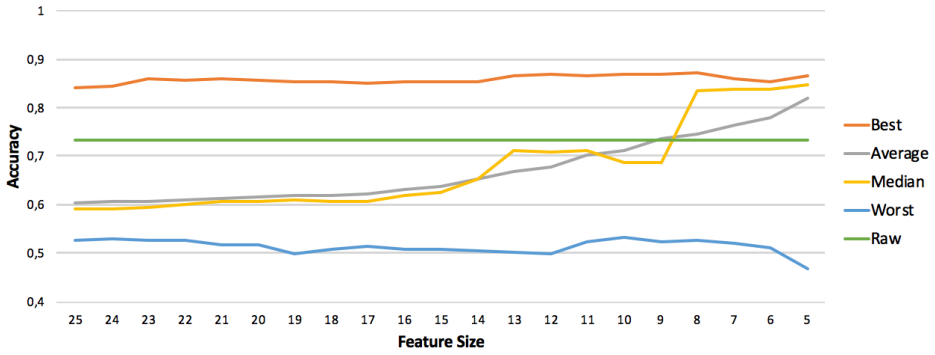

Figure 5.2.4.1 The accuracy of NB classifiers when different numbers of features are used on Z-Alizadehsani data set

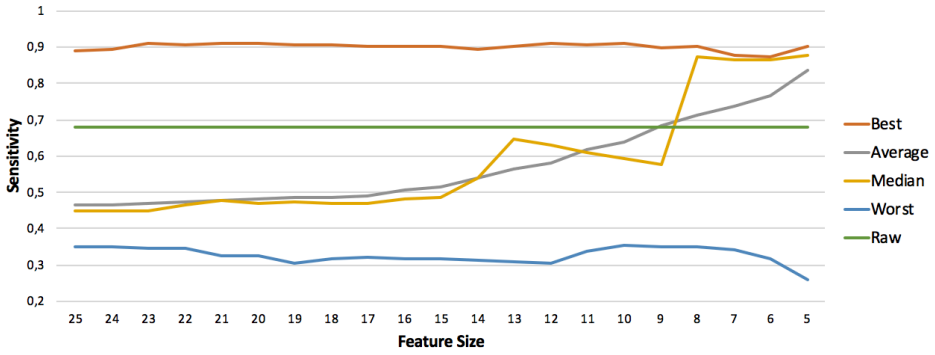

Figure 5.2.4.2 The sensitivity of NB classifiers when different numbers of features are used on Z-Alizadehsani data set

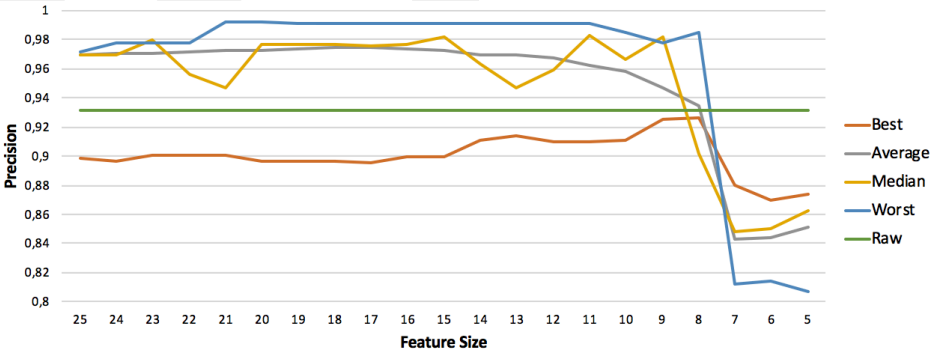

Figure 5.2.4.3 The precision of NB classifiers when different numbers of features are used on Z-Alizadehsani data set

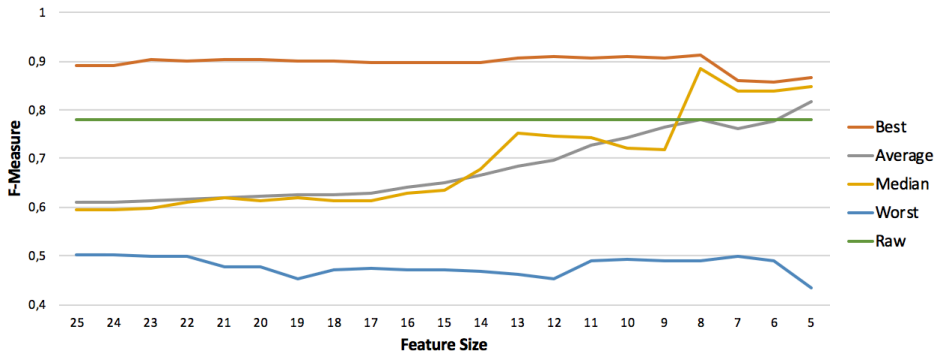

Figure 5.2.4.4 The F-measure of NB classifiers when different numbers of features are used on Z-Alizadehsani data set

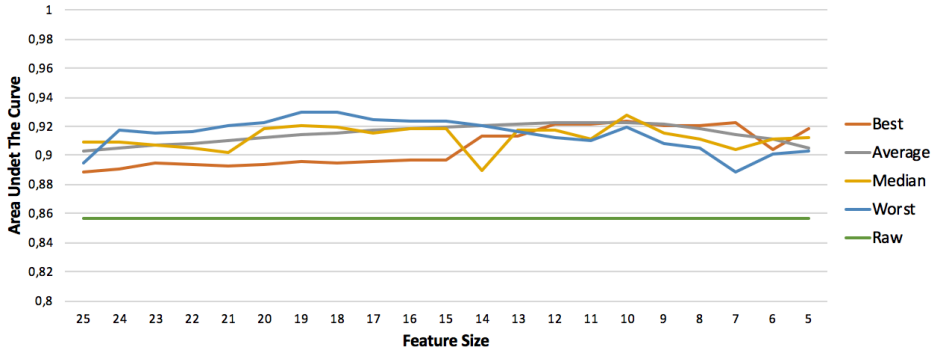

Figure 5.2.4.5 The AUC of NB classifiers when different numbers of features are used on Z-Alizadehsani data set

Figure 5.2.4.1 plots the accuracies of NB classifiers when different numbers of features are used on the Z-Alizadehsani data set. The NB classifier can produce good results when fewer features are used, and when the number of feature increases, the performance results go down. Similarly, Figures 5.2.4.2, 5.2.4.3, 5.2.4.4 plot the sensitivity, precision, and F-measure of NB classifiers respectively, when different numbers of features are used on Z-Alizadehsani data set. These plots have similar patterns with accuracy plots. As shown in Figure 5.2.4.5, the number of features size does not affect AUC. Once the feature size is less than 15, average of performance metrics values are increasing. The NB classifier could perform poor results when the number of features is increasing.

5.2.5 Support Vector Machine

Support Vector Machine (SVM) classifier is applied on two different data sets, i.e., Z-Alizadehsani and UCI Cleveland. Using SVM classifier, 91.13% accuracy, 93.48% sensitivity, 94.37% precision, 0.937 F-Measure, and 0.950 AUC values are achieved when kernel is linear, and c is 1 on the Z-Alizadehsani data set. The same classifier achieved 83.10% accuracy, 80.05% sensitivity, 83.13% precision, 0.811 F-Measure, and 0.911 AUC on the UCI Cleveland data set. Table 5.2.5.1 shows in detail the best results obtained with the SVM classifier, sorted with respect to the accuracy values in various situations. Among different probabilistic feature selection runs, the results with the top three accuracies are listed in Table 5.2.5.1. When SVM classifier is used, the feature selection methods did not affect the performance results significantly on both CVD data sets.

Dt	Combination of Feature Selection	Att	SN (%)	Pre (%)	FM	AUC	Acc (%)
Z-Alizadehsani	SVM	19	93.48	94.37	0.937	0.950	91.13
	SVM+BS	21	93.00	94.37	0.934	0.949	90.80
	SVM+BS	25	93.48	93.89	0.935	0.952	90.79
	Average of All Possible FS Comb.	-	91.18	91.24	0.910	0.933	87.93
	Probabilistic FS of SVM 1	25	92.96	93.32	0.929	0.929	90.01
	Probabilistic FS of SVM 2	25	91.64	94.13	0.927	0.918	89.77
	Probabilistic FS of SVM 3	25	93.07	92.44	0.926	0.909	89.51
	Average of Probabilistic FS	-	90.23	84.68	0.868	0.834	80.44
	FLDA	1	92.38	95.55	0.913	1	90.28
	-	55	90.71	91.46	0.909	0.932	87.14
UCI Cleveland	RF	10	80.05	83.13	0.811	0.911	83.10
	RF+BS+DK	9	80.10	83.18	0.810	0.910	83.10
	CS	10	80.00	83.11	0.812	0.909	83.09
	Average of All Possible FS Comb.	-	75.97	82.23	0.783	0.898	80.99
	FLDA	1	78.57	85.29	0.814	0.922	83.75
	-	55	77.85	82.63	0.794	0.899	81.42

Dt: Data set, Att: Number of Attributes included, SN: Sensitivity, Pre: Precision, FM: F-Measure, AUC: Area Under Curve, Acc: Accuracy

Table 5.2.5.1 Performance evaluations of SVM classifier on two CVD data sets when different feature selection techniques are applied

5.2.6 Multilayer Perceptron

Multilayer Perceptron (MLP) classifier is applied on two different data sets, i.e., Z-Alizadehsani and UCI Cleveland. Using MLP classifier, 91.78% accuracy, 93.50% sensitivity, 95.14% precision, 0.941 F-Measure, and 0.956 AUC values are achieved when alpha is 1, hidden layer is 50, max iteration is 1000 and activation is tanh on the Z-Alizadehsani data set. The same classifier achieved 85.47% accuracy, 82.96% sensitivity, 86.22% precision, 0.839 F-Measure, and 0.911 AUC on the UCI Cleveland data set. Table 5.2.6.1 shows in detail the best results obtained with the MLP classifier, sorted with respect to the accuracy values in various situations. Among different probabilistic feature selection runs, the results with the top three accuracies are listed in Table 5.2.6.1. When MLP classifier is used, feature selection methods improved the performance results significantly on both CVD data sets. Among all other tested classifiers, the best performance results are obtained with MLP classifiers on both CVD data sets.

Dt	Combination of Feature Selection	Att	SN (%)	Pre (%)	FM	AUC	Acc (%)
Z-Alizadehsani	SVM	21	93.50	95.14	0.941	0.956	91.78
	SVM+BS	25	93.48	94.83	0.939	0.952	91.45
	IG+BS+CMIM+DK	25	94.04	93.86	0.938	0.956	91.20
	Average of All Possible FS Comb.	-	90.69	90.47	0.899	0.928	86.55
	Probabilistic FS of MLP 1	25	94.41	93.53	0.937	0.941	91.14
	Probabilistic FS of MLP 2	14	92.53	91.59	0.918	0.914	88.41
	Probabilistic FS of MLP 3	16	92.57	91.33	0.918	0.924	88.17
	Average of Probabilistic FS	-	88.47	84.54	0.861	0.833	79.70
	FLDA	1	90.47	95.24	0.881	1	88.58
	-	55	88.89	89.68	0.891	0.917	84.54
UCI Cleveland	CMIM	9	82.96	86.22	0.839	0.911	85.47
	IG+SVM+BS	9	83.68	85.77	0.842	0.911	85.45
	CS+GR+RF+BS+CMIM	9	82.96	85.67	0.836	0.901	85.11
	Average of All Possible FS Comb.	-	79.17	82.23	0.800	0.898	81.89
	FLDA	1	80.76	83.28	0.815	0.922	83.43
	-	55	78.62	82.30	0.795	0.886	81.42

Dt: Data set, Att: Number of Attributes included, SN: Sensitivity, Pre: Precision, FM: F-Measure, AUC: Area Under Curve, Acc: Accuracy

Table 5.2.6.1 Performance evaluations of MLP classifier on two CVD data sets when different feature selection techniques are applied

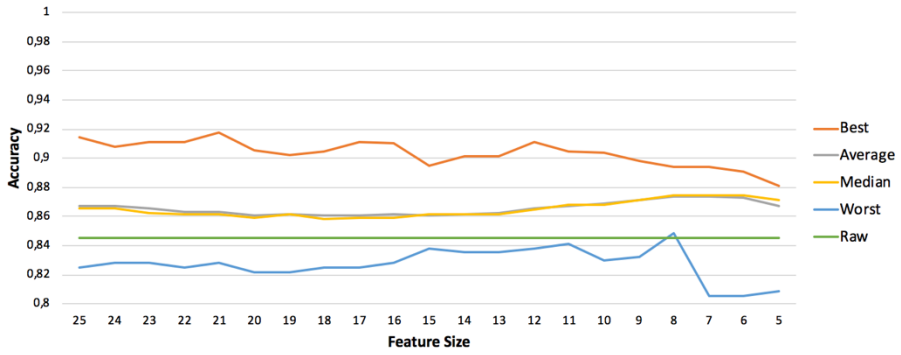

Figure 5.2.6.1 The accuracy of MLP classifiers when different numbers of features are used on Z-Alizadehsani data set

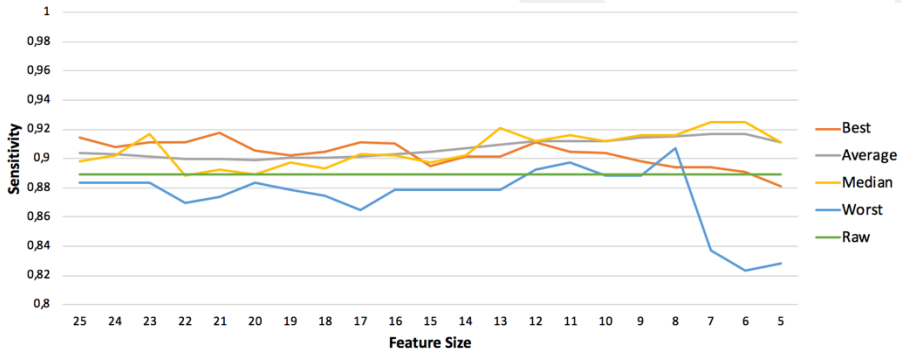

Figure 5.2.6.2 The sensitivity of MLP classifiers when different numbers of features are used on Z-Alizadehsani data set

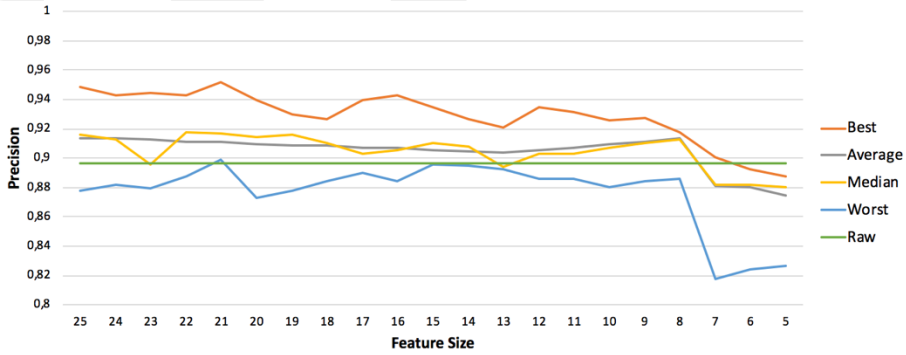

Figure 5.2.6.3 The precision of MLP classifiers when different numbers of features are used on Z-Alizadehsani data set

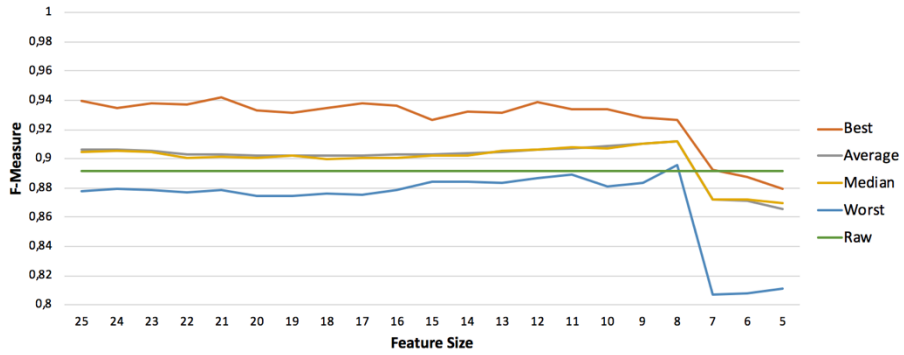

Figure 5.2.6.4 The F-measure of MLP classifiers when different numbers of features are used on Z-Alizadehsani data set

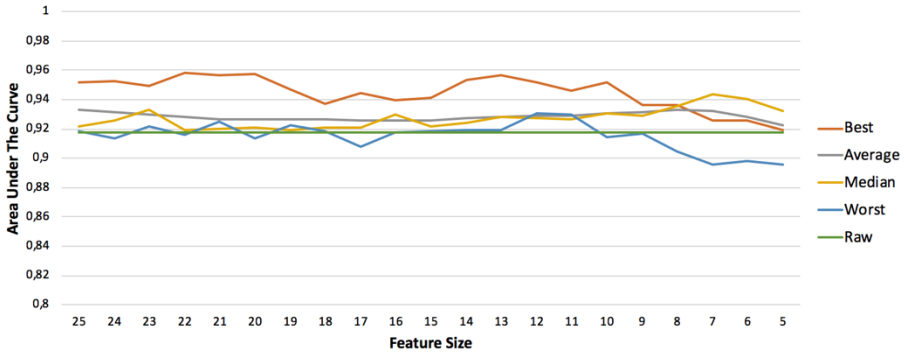

Figure 5.2.6.5 The AUC of MLP classifiers when different numbers of features are used on Z-Alizadehsani data set

Figure 5.2.6.1 plots the accuracies of MLP classifiers when different numbers of features are used on the Z-Alizadehsani data set. When the feature size decreases, accuracy results decrease. Especially when the feature size is 7 the accuracy results start to decrease dramatically. These features need to be examined, the irrelevant feature needs to be identified and removed from the data set, if this situation is observed in other combinations as well. Similarly, Figures 5.2.6.2, 5.2.6.3, 5.2.6.4 plot the sensitivity, precision, and F-measure of MLP classifiers respectively, when different numbers of features are used on Z-Alizadehsani data set. These graphics are almost similar to the accuracy graphic. In Figure 5.2.6.5, the change in the number of features did not affect AUC values until the feature size is less than 10. The MLP classifier could perform poor results when the feature size is dramatically reduced.

5.2.7 Random Forest

Random Forest (RF) classifier is applied on two different data sets, i.e., Z-Alizadehsani and UCI Cleveland. Using RF classifier, 90.49% accuracy, 92.59% sensitivity, 92.59% precision, 0.927 F-Measure, and 0.937 AUC values are achieved on the Z-Alizadehsani data set. The same classifier achieved 84.72% accuracy, 80.00% sensitivity, 81.67% precision, 0.812 F-Measure, and 0.903 AUC on the UCI Cleveland data set. Table 5.2.7.1 shows in detail the best results obtained with the RF classifier, sorted with respect to the accuracy values in various situations. Among different probabilistic feature selection runs, the results with the top three accuracies are listed in Table 5.2.7.1. When RF classifier is applied, feature selection methods did not affect considerably the performance results in both CVD data sets. When the RF classifier is used, the worst performance results are obtained when FLDA methods is applied as dimension reduction technique on both CVD data sets.

Dt	Combination of Feature Selection	Att	SN (%)	Pre (%)	FM	AUC	Acc (%)
Z-Alizadehsani	CS+SVM+BS	12	92.59	92.59	0.927	0.937	90.49
	SVM+BS+CMIM+DK	21	96.77	88.69	0.920	0.933	90.15
	SVM+BS+DK	25	94.55	89.94	0.917	0.943	90.12
	Average of All Possible FS Comb.	-	91.56	89.58	0.899	0.921	86.47
	Probabilistic FS of RF 1	13	94.93	90.09	0.924	0.916	90.13
	Probabilistic FS of RF 2	25	93.96	91.23	0.913	0.924	89.11
	Probabilistic FS of RF 3	25	93.50	90.73	0.914	0.933	88.83
	Average of Probabilistic FS	-	89.47	83.51	0.861	0.822	79.48
	FLDA	1	63.80	84.39	0.635	0.851	62.08
	-	55	94.43	87.60	0.902	0.920	87.22
UCI Cleveland	IG+RF+SVM	10	80.00	81.67	0.812	0.903	84.74
	BS+CMIM	12	77.74	81.86	0.800	0.894	84.74
	CS+GR+CMIM	12	78.46	85.98	0.793	0.904	84.44
	Average of All Possible FS Comb.	-	77.76	80.83	0.788	0.891	80.83
	FLDA	1	68.18	72.15	0.689	0.846	72.29
	-	55	77.96	81.14	0.792	0.892	82.71

Dt: Data set, Att: Number of Attributes included, SN: Sensitivity, Pre: Precision, FM: F-Measure, AUC: Area Under Curve, Acc: Accuracy

Table 5.2.7.1 Performance evaluations of RF classifier on two CVD data sets when different feature selection techniques are applied

5.2.8 Ensemble Methods

Voting classifier is used on two different CVD data sets, i.e., Z-Alizadehsani and UCI Cleveland. Hard voting and soft voting techniques are tested separately on both data sets. The classifiers are chosen according to the obtained results. The best three and best five classifiers are utilized in hard and soft voting methods. Voting classifier is applied on two different data sets, i.e., Z-Alizadehsani and UCI Cleveland. Using Voting classifier, 91.11% accuracy, 91.11% sensitivity, 91.70% precision, 0.910 F-Measure, and 0.958 AUC values are achieved on the Z-Alizadehsani data set. The same classifier achieved 84.13% accuracy, 84.13% sensitivity, 85.02% precision, 0.836 F-Measure, and 0.913 AUC on the UCI Cleveland data set. Table 5.2.8.1 shows in detail the best results obtained with the Voting classifier, sorted with respect to the accuracy values in various situations. Among different probabilistic feature selection runs, the results with the top three accuracies are listed in Table 5.2.8.1. The voting classifier could not achieve better performance results than the Voting classifier.

Dt	Combination of Feature Selection	V	At	SN (%)	Pre (%)	FM	AUC	ACC (%)
Z-Alizadehsani	SVM + BS	S3	22	93.93	93.75	0.938	0.955	91.13
	SVM	S3	22	93.93	93.89	0.935	0.957	91.11
	SVM+DK	S3	16	93.91	93.91	0.937	0.952	91.09
	Average of All Possible FS Comb.	S	-	91.89	91.73	0.916	0.939	88.01
	SVM+BS	H3	22	93.48	94.88	0.940	-	91.45
	SVM+DK	H3	16	93.91	93.91	0.937	-	91.08
	SVM+BS	H3	20	93.48	94.07	0.935	-	90.81
	Average of All Possible FS Comb.	H	-	91.86	92.00	0.917	-	88.22
Cleveland	CS+GR+SVM+BS+CMIM+DK	S3	10	81.53	84.74	0.827	0.919	84.80
	CS	S3	10	81.53	84.74	0.826	0.917	84.79
	CS+IG+GR+BS+CMIM+DK	S3	11	79.34	84.81	0.820	0.912	84.47
	Average of All Possible FS Comb.	S	-	77.17	82.94	0.794	0.904	81.88
	CS+IG+GR+SVM+DK	H5	12	77.80	84.10	0.800	-	83.43
	CS+IG+SVM+CMIM	H5	12	79.34	84.67	0.813	-	83.43
	CS+GR+BS+DK	H5	12	79.34	84.67	0.813	-	83.43
	Average of All Possible FS Comb.	H	-	76.17	82.83	0.787	-	81.43

Dts: Data set, V: Voting Type, H: Hard Voting, S: Soft Voting, 3 or 5 following S and H: the number of classifiers used in ensemble classifier, At: Number of Attributes included, SN: Sensitivity, Pre: Precision, FM: F-Measure, Auc: Area Under Curve, Acc: Accuracy

Table 5.2.8.1 Performance evaluations of ensemble classifiers on two CVD data sets when different feature selection techniques are applied

Chapter 6

Discussions

In this thesis, two publicly available CVD data sets including 303 samples are used. While the Z-Alizadehsani data set includes 55 features, Cleveland data set consists of 14 features. Feature selection techniques, dimension reduction, and classification algorithms are used to diagnose the CVD using stratified 10-fold cross-validation. To facilitate CVD diagnosis, several studies in literature analyzed different CVD data sets, applied different pre-processing steps and classification methods. Existing CVD diagnosis data sets are different from each other in terms of feature characteristics, sample sizes, feature sizes, and ethnicity of the samples. Therefore, to compare our model with existing studies, the two widely used CVD diagnosis data sets are selected. Also, in terms of completeness, these two data sets have lesser numbers of missing data.

There is no internationally recognized standard machine learning approach for CVD diagnosis. Although some studies have presented satisfactory performance results for the diagnosis of CVD on a particular data set, these models did not perform well on different CVD data sets. One of the goals of this thesis is to develop a single classifier that performs well on two different publicly available CVD data sets. In our study, we generate 92.895 classification models as following: There are 255 different combinations for 8 different feature selection methods. The numbers of tested features are 21 and 8, for Z-Alizadehsani and Cleveland data sets, respectively. We used 11 different classifiers including 7 single and one ensemble classifier with 4 different variations (S3, S5, H3, H5, where S and H indicate soft and hard voting, respectively, and the number following S and H indicates the number of classifiers used in ensemble classifier).

When the proposed ensemble feature selection method is applied on Z-Alizadehsani data set, 58.905 models (255 * 21 * 11) are generated. When it is applied on UCI Cleveland data set, 22.440 models (255 * 8 * 11) are generated. When the probabilistic feature selection method is applied on Z-Alizadehsani data set, 11.550 models (50 * 8 * 11) are generated. Among these 92.895 different classification models, the best performance result of each classifier is presented in Table 6.1. Except for NB and KNN classifiers, the results of other classifiers are very close to each other. Once the proper features are selected, each classifier generated significant performance results for CVD diagnosis. The results were evaluated in two different ways: (i) achieve good performance results, and (ii) achieve good performance results on both data sets when the same subset of feature selection methods are used as ensemble FS method.

Dt	Classifier (CFS)	Att	SN (%)	Pre (%)	FM	AUC	Acc (%)
Z-Alizadehsani	MLP (SVM)	21	93.50	95.14	0.941	0.956	91.78
	H3 (SVM+BS)	22	93.48	94.88	0.940	-	91.45
	SVM (SVM)	19	93.48	94.37	0.937	0.950	91.13
	LR (SVM)	19	94.39	93.65	0.938	0.954	91.11
	RF (CS+SVM+BS)	12	92.59	92.59	0.927	0.937	90.49
	LDA (IG+SVM+BS+DK)	**24**	**92.57**	**93.83**	**0.930**	**0.924**	**90.15**
	NB (RF+SVM+BS+DK)	8	90.17	92.66	0.911	0.920	87.42
	KNN (CS+GR+RF)	5	92.87	85.57	0.841	0.891	84.86
Cleveland	MLP (CMIM)	9	82.96	86.22	0.839	0.911	85.47
	S3 (CS+GR+SVM+BS+CMIM+DK)	10	81.53	84.74	0.827	0.919	84.80
	RF (IG+RF+SVM)	10	80.00	81.67	0.812	0.903	84.74
	KNN (CS+RF+SVM)	5	81.37	85.93	0.825	0.903	84.41
	LDA (IG+SVM+BS+DK)	**11**	**78.51**	**84.52**	**0.809**	**0.901**	**83.41**
	LR (RF+SVM+DK)	10	79.34	83.78	0.811	0.902	83.11
	SVM (RF)	10	80.05	83.13	0.811	0.911	83.10
	NB (SVM)	7	80.82	81.67	0.811	0.892	82.78

Dt: Data set, CFS: Combination of Feature Selection, Att: Number of Attributes included, SN: Sensitivity, Pre: Precision, FM: F-Measure, AUC: Area Under Curve, Acc: Accuracy

Table 6.1 For each tested classifier, the best performance results of two different CVD data sets

MLP classifier achieved the best performance results in both CVD data sets with different sets of selected features. In the Z-Alizadehsani data set, MLP

classifier that uses an ensemble of CS, BS, CMIM, and DK feature selection techniques resulted in 91.44% accuracy, 91.44% sensitivity, 92.00% precision, 0.914 F-Measure, and 0.944 AUC. In UCI Cleveland data set, MLP classifier that uses an ensemble of CS, IG, RF, SVM, CMIM feature selection techniques resulted in 85.79% accuracy, 85.79% sensitivity, 86.60% precision, 0.856 F-Measure, and 0.901AUC. With a single model, one of the best performance results is obtained in both data sets using LDA classifier and ensemble (IG, SVM, BS, DK) feature selection method. In the Z-Alizadehsani data set, this model achieved 90.13% accuracy, 90.13% sensitivity, 91.08% precision, 0.900 F-Measure, and 0.942 AUC. The Cleveland data set the same model achieved 83.41% accuracy, 83.41% sensitivity, 83.35% precision, 0.832 F-Measure, and 0.901 AUC. Overall, NB has the worst performance in both data sets.

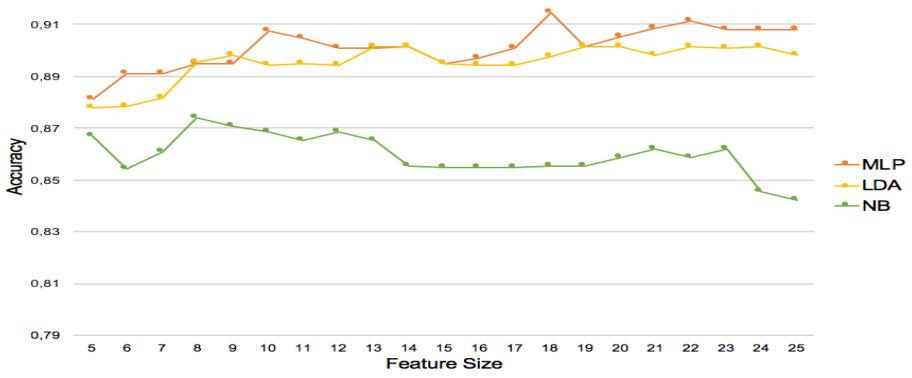

Figure 6.1 Comparison of the accuracies that are obtained with MLP, LDA and NB classifiers using different numbers of features on Z-Alizadehsani data set

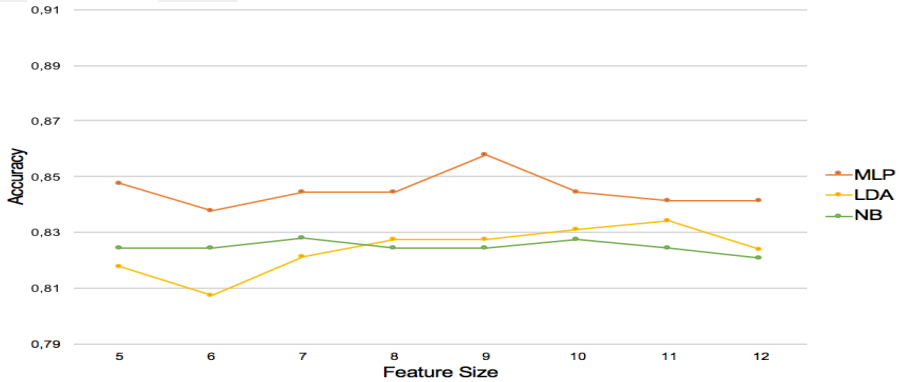

Figure 6.2 Comparison of the accuracies that are obtained with NLP, LDA and NB classifiers using different numbers of features on UCI Cleveland data set

Figure 6.1 and Figure 6.2 compare the accuracies that are obtained with MLP, LDA and NB classifiers using different numbers of features on Z-Alizadehsani and UCI Cleveland data sets, respectively. As shown in Figure 6.1, for the Z-Alizadehsani data set, MLP and LDA classifiers perform better than the NB classifier when different numbers of features are tested. As shown in Figure 6.2, for the UCI Cleveland data set, MLP classifier performs better than the LDA and NB classifiers when different numbers of features are tested. For the UCI Cleveland data set, the NB classifier performs the second best, when less than 7 features are used. When more than 7 features are included, LDA performs the second best, on the same data set.

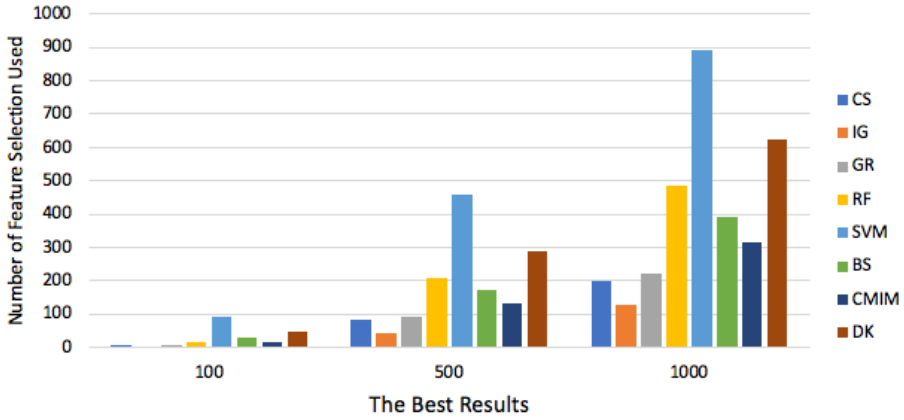

Figure 6.3 Frequencies of the feature selection methods in the top 100, top 500 and top 1000 best performing models (in terms of accuracy) on Z-Alizadehsani data set

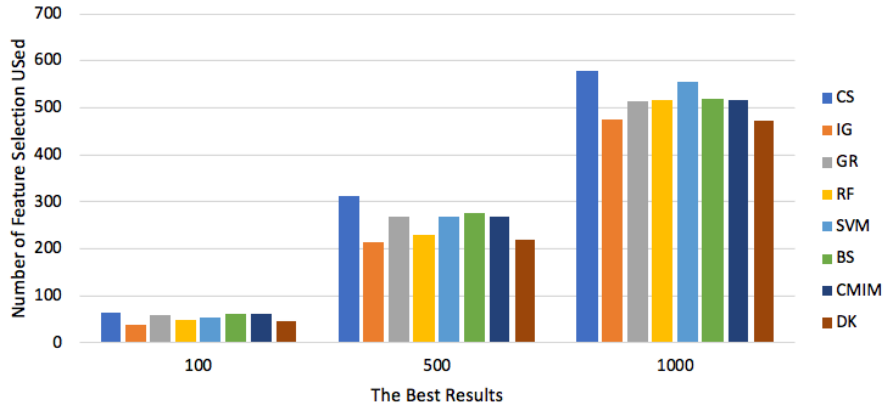

Figure 6.4 Frequencies of the feature selection methods in the top 100, top 500 and top 1000 best performing models (in terms of accuracy) on UCI Cleveland data set

We listed the top 1000, top 500 and top 100 accuracy values among 92.895 different classifiers and analyzed the frequencies of different feature selection methodologies in these lists. When we evaluate these frequencies on the Z-Alizadehsani data set, we observe that SVM, DK, and RF feature selection techniques are more frequently included in the ensemble feature selection method of high scoring classifiers, as shown in Figure 6.3. In the UCI Cleveland data set, the frequencies of different feature selection methods are very close to each other, as shown in Figure 6.4. Possible reasons might be due to the UCI Cleveland data set does not contain enough features, samples, or could be because of the data characteristics.

Reference	Data set	Method	Accuracy (%)
El Biary et al [16]	UCI Cleveland	DT	78.54
Alizadehsani et al [14]	Z-Alizadehsani	Bagging	79.54
Kumari et al [38]	UCI Cleveland	SVM	80.06
Chitraet et al [39]	UCI Cleveland	SVM	82.00
Shouman et al [12]	UCI Cleveland	DT	84.10
Palaniappan et al [40]	UCI Cleveland	KNN	85.55
Frantistec et al [18]	Z-Alizadehsani	DT	86.67
Alizadehsani et al [13]	Z-Alizadehsani	SMO	92.09
Reed et al [22]	UCI Cleveland	RF	92.16
Anbarisi et al [11]	UCI Cleveland	DT	99.20
Proposed Single Model	UCI Cleveland	LDA	**83.41**
	Z-Alizadehsani	LDA	**90.15**

Table 6.2 The performance evaluation of the proposed method with existing studies

In this thesis, a novel ensemble method for the feature selection process is proposed, and the performance comparisons are presented. In our experiments, the accuracy of the proposed model is slightly lower than the existing models, as shown in Table 6.2. However, most of the existing studies do not give adequate information on cross-validation, data splitting process, or training-test sets that might drastically affect the performance results. We believe that our model could overperform other studies if we also apply pre-processing procedures, as

mentioned in these papers. We would like to note that our primary goal is to offer an adaptive method that can work on several data sets without additional analysis and pre-processing. The model is tested on two different data sets and shown to perform well and generate satisfactory results in terms of accuracy, sensitivity, precision, F-measure, and AUC.

Chapter 7

Conclusions and Future Prospects

7.1 Conclusions

With the development of machine learning and data mining techniques, it becomes possible to diagnose Cardiovascular Diseases (CVD) at a lower cost using biochemical values. This thesis aims to develop a computational Cardiovascular Disease diagnosis model via incorporating domain knowledge. Two publicly available data sets from the UCI Machine Learning Repository are used. Weka and Python programing language are used to apply data mining and machine learning algorithms to classify UCI Cleveland and Z-Alizadehsani data sets via stratified-10-fold cross-validation.

There is no internationally recognized standard machine learning approach for CVD diagnosis. Although some studies have reported satisfactory performance results for the CVD diagnosis on a particular CVD data set, these models do not perform well on different CVD data sets. We attempt to create a single model that achieves satisfactory results on different data sets. In this thesis, for CVD diagnosis problem, different computational feature selection (FS) methods, domain knowledge-based FS method, ensemble FS method, dimension reduction, and different classification algorithms have experimented. The ensemble feature selection method included all possible combinations of eight different FS methods and different numbers of features. For the classification task, seven single classifiers and one ensemble classifier method are tested. Although none of the existing studies present a detailed performance evaluation, in this

study, the performance results of the proposed method are presented with several evaluation metrics, e.g. accuracy, sensitivity, precision, F-measure, and AUC.

In our experiments with stratified-10-fold cross-validation, in a single model with an ensemble (IG, SVM, BS, DK) feature selection method, the LDA classifier achieved one of the best performance results in both CVD data sets. This model resulted in 90.13% accuracy, 90.13% sensitivity, 91.08% precision, 0.900 F-Measure, and 0.942 AUC, on the Z-Alizadehsani data set. The same model achieved 83.41% accuracy, 83.41% sensitivity, 83.35% precision, 0.832 F-Measure, and 0.901 AUC on the Cleveland data set.

The main contribution of this thesis is our proposed ensemble feature selection method. By analyzing the patient's physical and biochemical values with data mining and machine learning techniques, this thesis contributes to society via enabling CVD diagnosis in a more economical and efficient way. It is noteworthy to state that our primary goal is to offer an adaptive approach that can work on several data sets without additional analysis. The proposed model is tested on two different data sets and shown to generate sustainable performance results in terms of accuracy, sensitivity, precision, F-measure, and AUC. Our proposed model could be easily adapted to current practice. Via contributing to early CVD diagnosis, this model finally aims to reduce mortality.

7.2 Future Prospects

In future work, we would like to examine other CVD diagnosis data sets with our proposed classification model. We could incorporate additional machine learning techniques and deep learning to further improve performance results with a single model. In addition to the features that are selected by cardiologists, it has been observed in this thesis that some features that are not suggested by medical experts also have positive effects on performance results. These features need to

be examined by medical specialists and could be potentially added to the significant features list of CVD diagnosis. These features could be proposed as potential biomarkers to be added to FHS. Additionally, there is no publicly available CVD data set on the Turkish population. Via collaborating with Turkish cardiologists, we would like to contribute to the development of such a population-specific data set, apply our model on this data set and observe whether there are Turkish population-specific features for CVD diagnosis.

BIBLIOGRAPHY

[1] Cardiovascular disease (CVDs), https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (13.11.2019).

[2] Coronary heart disease (CHD), https://www.bhf.org.uk/information support/conditions/coronary-heart-disease (13.11.2019).

[3] Know the difference, https://www.nhlbi.nih.gov/sites/ default/files/media/ docs/Fact_Sheet_Know_Diff_Design.508_pdf.pdf (13.11.2019).

[4] Mortality due to Cardiovascular Disease and Diabetes, http://chartsbin.com/view/2621 (13.11.2019).

[5] Cardiovascular disease diagnosis, https://www.news-medical.net/health/ Cardiovascular-Disease-Diagnosis.aspx (13.11.2019).

[6] Alizadehsani, Roohallah, "Machine learning-based coronary artery disease diagnosis: A comprehensive review", Computers in biology and medicine, 111, page no (2019).

[7] M. Akay, “Noninvasive diagnosis of coronary artery disease using a neural network algorithm”, Biol. Cybern, 67, 361–367 (1992).

[8] Polat, Kemal, Seral Şahan, Salih Güneş, "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing", Expert Systems with Applications, 32.2, 625-631 (2007).

[9] Palaniappan, Sellappan, R. Awang, "Intelligent heart disease prediction system using data mining techniques", IEEE/ACS international conference on computer systems and applications, volume no, 108-115 (2008).

[10] R. Das, I. Turkoglu, A. Sengur. "Effective diagnosis of heart disease through neural networks ensembles" Expert systems with applications, 36.4, 7675-7680 (2009).

[11] M. Anbarasi, E. Anupriya, N. C. S. N. Iyengar. "Enhanced prediction of heart disease with feature subset selection using genetic algorithm", International Journal of Engineering Science and Technology, 2.10, 5370-5376 (2010).

[12] M. Shouman, T. Turner, R. Stocker. "Using decision tree for diagnosing heart disease patients" Proceedings of the Ninth Australasian Data Mining Conference, Australian Computer Society, 121 (2011).

[13] R. Alizadehsani, Roohallah, "Diagnosis of coronary artery disease using cost-sensitive algorithms." Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on. IEEE (2012).

[14] R. Alizadehsani, J. Habibi, Z. Alizadeh Sani, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, F. Khozeimeh, F. Alizadeh-Sani, "Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiography features", Res. Cardiovasc. Med., 2, 133–139 (2013).

[15] B. Subanya, R. R. Rajalaxmi. "Artificial bee colony-based feature selection for effective cardiovascular disease diagnosis.", International Journal of Scientific & Engineering Research, 5.5, 606-612 (2014).

[16] El-Bialy, Randa, "Feature analysis of coronary artery heart disease data sets", Procedia Computer Science, 65, 459-468 (2015).

[17] Verma, Luxmi, Sangeet Srivastava, P. C. Negi, "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data.", Journal of medical systems, 40.7, 178 (2016).

[18] Babič, František, "Predictive and descriptive analysis for heart disease diagnosis." Computer Science and Information Systems (FedCSIS), volume no, 155-163 (2017).

[19] B. Kolukisa, H. Hacilar, G. Goy, M. Kus, B. Bakir-Gungor, A. Aral, V.C. Gungor, "Evaluation of classification algorithms, linear discriminant analysis and a new hybrid feature selection methodology for the diagnosis of coronary artery disease", IEEE Int. Conf. Big Data (Big Data), 2232–2238 (2018).

[20] Kolukısa, Burak, et al. "Diagnosis of Coronary Heart Disease via Classification Algorithms and a New Feature Selection Methodology." *International Journal of Data Mining Science* 1.1, 8-15 (2019)

[21] Kolukisa, Burak, et al. "Coronary Artery Disease Diagnosis Using Optimized Adaptive Ensemble Machine Learning Algorithm." (2020)

[22] Reddy, N. Satish Chandra, "Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction", International Journal of Innovative Computing, 9.1 (2019).

[23] Detrano, Robert, "International application of a new probability algorithm for the diagnosis of coronary artery disease" American Journal of Cardiology, 64.5, 304-310 (1989).

[24] J. Han, J Pei, "Data Mining Concepts and Techniques", Morgan Kaufmann (2011).

[25] J. Hurwitz, D. Kirsch, "IBM Limited Edition" (2018).

[26] Kira, Kenji, Rendell, Larry "The feature selection problem: Traditional methods and a new algorithm.", AAAI-92 Proceedings, 2 ,129-134 (1992).

[27] Kira, Kenji, A. Rendell. "A practical approach to feature selection", Machine Learning Proceedings, Morgan Kaufmann, 249-256 (1992).

[28] I. Kononenko, E. Šimec, M. Robnik-Šikonja. "Overcoming the myopia of inductive learning algorithms with RELIEFF." Applied Intelligence, 7.1, 39-55, (1997).

[29] Guyon, Isabelle, "Gene selection for cancer classification using support vector machines", Machine learning, 46.1-3, 389-422 (2002).

[30] S. Fong, R.P. Biuk-Aghai, R.C. Millham. "Swarm Search Methods in Weka for Data Mining", Proceedings of the 2018 10th International Conference on Machine Learning and Computing. ACM, 122-127 (2018).

[31] Framingham Heart Study, https://www.framinghamheartstudy.org/ (13.11.2019).

[32] Cardiovascular risk, https://www.nhlbi.nih.gov/health-topics/assessing-cardiovascular-risk (13.11.2019).

[33] C. Chen, "Computer Vision in Medical Imaging: World Scientific Publishing Company", World Scientific (2014).

[34] Support vector machine, https://scikit-learn.org/stable/modules/svm.html (13.11.2019).

[35] Multi-layer perceptron, https://scikit-learn.org/stable/modules/neural_networks_supervised.html (13.11.2019).

[36] L. Rokach, "Ensemble-based classifiers", Artificial Intelligence Review,33.1-2, 1-39 (2010).

[37] R. Polikar, "Ensemble based systems in decision making." IEEE Circuits and systems magazine, 6.3, 21-45 (2006).

[38] Kumari, M., Godara, S. "Comparative study of data mining" (2011).

[39] Chitra, R., and V. Seenivasagam. "Heart disease prediction system using supervised learning classifier." Bonfring International Journal of Software Engineering and Soft Computing 3.1, 01-07 (2013).

[40] Palaniappan, Sellappan, and Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques." 2008 IEEE/ACS international conference on computer systems and applications, IEEE (2008).

[41] Guyon, I., Gunn, S., Nikravesh, M. and Zadeh, L., "Feature Extraction: Foundations and Applications", Springer (2006).

[42] Jensen, Richard, and Qiang Shen, "Computational intelligence and feature selection: rough and fuzzy approaches" John Wiley & Sons (2008).

[43] Bolón-Canedo, Amparo Alonso-Betanzos, "Ensembles for feature selection: a review and future trends.", Information Fusion, 52, 1-12 (2019).

[44] Quinlan, J. Ross. "Induction of decision trees." Machine learning 1.1, 81-106 (1986).

[45] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." Journal of machine learning research, 2825-2830 (2011).

[46] Karaboga, Dervis, "An idea based on honey bee swarm for numerical optimization" Technical report-tr06, Erciyes university, engineering faculty, computer engineering Department (2005).

[47] Khader, Ahamad Tajudin, Mohammed Azmi Al-betar, and A. Awadallah Mohammed. "Artificial bee colony algorithm, its variants and applications: a survey." (2013).

APPENDIX

Probabilistic FS	Attributes
KNN 1	St Elevation, Typical Chest Pain, Dyspnea, BBB, HTN, Region RWMA, K, Weak Peripheral Pulse
KNN 2	LowTH Ang, CR, Q Wave, LVH, Region RWMA, EX, Smoker, Typical Chest Pain
KNN 3	Dyspnea, Age, CRF, Airway disease, Typical Chest Pain, LowTH Ang, Tinversion, Atypical, LVH, St Elevation

FS: Feature Selection

Table 9.1 The list of selected attributes in the top 3 scoring probabilistic feature selection method when KNN classifier is used

Probabilistic FS	Attributes
LR 1	Neut, CHF, Diastolic Murmur, FH, Tinversion, BP, ESR, BBB, Lung rales, Region RWMA, Current Smoker, EF-TTE, Age, Edema, DM, DLP, Typical Chest Pain, Airway disease, LVH, Q Wave, Na, St Elevation, St Depression
LR 2	Region RWMA, Dyspnea, Na, LDL, LowTH Ang, Q Wave, HTN, CRF, DM, Age, EF-TTE, Typical Chest Pain, Diastolic Murmur, Sex, FBS, Tinversion
LR 3	FBS, Region RWMA, Lymph, TG, Typical Chest Pain, EX, Smoker, Nonanginal, Tinversion, CRF, EF-TTE, WBC, Age, Dyspnea, Diastolic Murmur, Sex, BBB, DM, Weight, BMI, LowTH Ang, HTN, Q Wave, Obesity, LDL, HB

FS: Feature Selection

Table 9.2 The list of selected attributes in the top 3 scoring probabilistic feature selection method when LR classifier is used

Probabilistic FS	Attributes
LDA 1	St Elevation, BUN, DLP, BP, Lymph, FH, Sex, BBB, Typical Chest Pain, Q Wave, HB, Region RWMA, Airway disease, Age, VHD, Systolic Murmur, HTN, Poor R Progression, Edema, Current Smoker, ESR, K, TG, Nonanginal
LDA 2	PR, Region RWMA, Age, Q Wave, FH, Typical Chest Pain, St Depression, DM, Edema, HTN, DLP, St Elevation, BBB, Current Smoker, PLT, BUN, Sex, Exertional CP, LowTH Ang, Weight, LDL, HB, Atypical, Weak Peripheral Pulse
LDA 3	Age, St Elevation, Typical Chest Pain, EF-TTE, DLP, Region RWMA, Diastolic Murmur, HTN, BMI, Nonanginal, CRF, St Depression, Weak Peripheral Pulse, Dyspnea, LowTH Ang, FH, PLT, Current Smoker, ESR, Length, BP, BUN, BBB

FS: Feature Selection

Table 9.3 The list of selected attributes in the top 3 scoring probabilistic feature selection method when LDA classifier is used

Probabilistic FS	Attributes
NB 1	Lung rales, Systolic Murmur, Tinversion, VHD, Region RWMA, Airway disease, Typical Chest Pain, ESR, Sex, Current Smoker, K, Atypical, FH, LDL, Dyspnea, DM, HTN, Exertional CP, Nonanginal, Age, Na, TG, Obesity, Weight, CVA
NB 2	Age, Atypical, EF-TTE, Nonanginal, Tinversion
NB 3	Typical Chest Pain, Region RWMA, DM, CVA, BUN, Sex, DLP, Age, Nonanginal, LDL, Tinversion

FS: Feature Selection

Table 9.4 The list of selected attributes in the top 3 scoring probabilistic feature selection method when NB classifier is used

Probabilistic FS	Attributes
SVM 1	St Depression, Poor R Progression, HDL, BMI, Length, St Elevation, BP, TG, HTN, Lung rales, EF-TTE, HB, Tinversion, Age, VHD, Region RWMA, BUN, Obesity, DLP, CHF, K, Airway disease, CRF, Atypical, Typical Chest Pain
SVM 2	Region RWMA, HDL, Tinversion, Typical Chest Pain, EF-TTE, HTN, BP, VHD, DLP, Diastolic Murmur, St Depression, Age, ESR, Sex, BBB, CVA, Current Smoker, Nonanginal, Weight, Q Wave, Atypical, K, BMI, Systolic Murmur, Length
SVM 3	St Depression, LowTH Ang, Systolic Murmur, Age, Length, Poor R Progression, Current Smoker, Tinversion, CR, HTN, TG, LVH, CHF, Lymph, Weak Peripheral Pulse, Typical Chest Pain, ESR, EF-TTE, WBC, Weight, Q Wave, Lung rales, Na, Edema, BMI

FS: Feature Selection

Table 9.5 The list of selected attributes in the top 3 scoring probabilistic feature selection method when SVM classifier is used

Probabilistic FS	Attributes
MLP 1	St Elevation, Neut, Airway disease, HTN, Weight, Poor R Progression, St Depression, Region RWMA, Q Wave, Age, Tinversion, VHD, Systolic Murmur, DLP, DM, FH, Atypical, PR, BBB, TG, Dyspnea, Typical Chest Pain, Diastolic Murmur, K, Nonanginal
MLP 2	Typical Chest Pain, EF-TTE, HTN, Diastolic Murmur, Systolic Murmur, CHF, Region RWMA, HDL, Q Wave, Age, BP, ESR, BMI, BBB,
MLP 3	Q Wave, Poor R Progression, ESR, HTN, St Depression, EF-TTE, DM, Age, DLP, BBB, Nonanginal, Obesity, Typical Chest Pain, Region RWMA, BMI, LVH

FS: Feature Selection

Table 9.6 The list of selected attributes in the top 3 scoring probabilistic feature selection method when MLP classifier is used

Probabilistic FS	Attributes
RF 1	EF-TTE, BP, CHF, Age, BBB, Current Smoker, Typical Chest Pain, ESR, Region RWMA, Tinversion, HDL, K, FBS
RF 2	CHF, St Elevation, HTN, DM, Age, Current Smoker, Tinversion, Obesity, Typical Chest Pain, VHD, LowTH Ang, Region RWMA, LDL, Dyspnea, Exertional CP, BMI, EX, Smoker, Q Wave, Poor R Progression, FH, Thyroid Disease, CR, Atypical, Weight, Lymph
RF 3	FBS, Region RWMA, Lymph, TG, Typical Chest Pain, EX, Smoker, Nonanginal, Tinversion, CRF, EF-TTE, WBC, Age, Dyspnea, Diastolic Murmur, Sex, BBB, DM, Weight, BMI, LowTH Ang, HTN, Q Wave, Obesity, LDL, HB

FS: Feature Selection

Table 9.7 The list of selected attributes in the top 3 scoring probabilistic feature selection method when RF classifier is used