Prediction on Ischemic Heart Disease using Machine Learning Approaches

by

M. Raihan

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Biomedical Engineering

Khulna University of Engineering & Technology
Khulna 9203, Bangladesh

October, 2019

Declaration

This is to certify that the thesis work entitled "Prediction on Ischemic Heart Disease using Machine Learning Approaches" has been carried out by M. Raihan in the Department of Biomedical Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh. The above thesis work or any part of this work has not been submitted anywhere for the award of any degree or diploma.

27.10.2019

Signature of Supervisor

Raihan 27.10.19

Signature of Candidate

Approval

This is to certify that the thesis work submitted by M. Raihan entitled "Prediction on Ischemic Heart Disease using Machine Learning Approaches" has been approved by the board of examiners for the partial fulfillment of the requirements for the degree of Master of Science in Engineering (M. Sc. Eng.) in the Department of Biomedical Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh in October 2019.

BOARD OF EXAMINERS

1. 27/10/2019 — Chairman (Supervisor)

Dr. Muhammad Muinul Islam
Assistant Professor, Department of Biomedical Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh

2. 27/10/2019 — Member

Head of the Department
Department of Biomedical Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh

3. 27/10/2019 — Member

Dr. Mohiuddin Ahmad
Professor, Department of Electrical & Electronic Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh

4. 27.10.2019 — Member

Dr. Md. Shahjahan
Professor, Department of Electrical & Electronic Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh

5. 27.10.2019 — Member (External)

Dr. Md. Anisur Rahman
Professor, Computer Science and Engineering Discipline
Khulna University
Khulna 9208, Bangladesh

Acknowledgment

I express gratitude to the almighty ALLAH that the research study has been accomplished perfectly and successfully. I am very much delighted to present this research work on predicting Ischemic Heart Disease using different machine learning techniques.

I would like to express profound respect, deepest gratitude and hardest thanks to my thesis supervisor ***Dr. Muhammad Muinul Islam,*** Department of Biomedical Engineering, Khulna University of Engineering & Technology. The door to my supervisor's office was always open whenever I ran into trouble, spot or had a question about my research or writing. He consistently allowed this paper to be my own work but navigated me in the right direction whenever he thought I needed it.

I would also like to thank my committee members, ***Prof. Dr. Mohiuddin Ahmad*** and ***Prof***. ***Dr. Md. Shahjahan*** for their crucial suggestion on my research projects, and my external examiner ***Prof. Dr. Md. Anisur Rahman*** for his valuable time to read and give an opinion on my thesis work.

I would like to thank my respective teachers in the Department of Biomedical Engineering, Khulna University of Engineering & Technology for their valuable supports. I would also like to thank my seniors, friends, the staff of the department and everybody who was involved directly or indirectly to run this thesis work successfully. I would also like to thank ***Dr. Arun R More***, Former Facility Director of AFC Fortis Escorts Heart Institute, Khulna, Bangladesh and Founder of Rural Health Progress Trust, India for his supervision and inspiration.

Finally, I must express my very profound gratitude to my parents, my two elder sisters for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. Though my father died in the last moment of the research work he always gave me inspiration and prayed for my success. This accomplishment would not have been possible without them.

October, 2019 M. Raihan

Abstract

Ischemic heart disease (IHD) is a terrible experience that occurs when the flow of blood severely reduced or cut off due to plaque deposited on the inner wall of arteries that brings oxygen to the heart muscle, leads to the ischemic heart attack (IHA). Atherosclerosis i.e. plaque deposition on the inner wall of arteries is a silent process, has no critical symptoms to get a warning before IHD. For this reason, early detection is very important for the proper management of patients prone to IHD. In this thesis work, it was tried to predict IHD on the basis of patient history, symptoms and pathological findings of patients with heart disease using computational intelligence. Total 506 patient's data with a maximum of 151 features including historic, symptomatic and pathologic findings were collected from AFC Fortis Escort Heart Institute, Khulna, Bangladesh. First, it was tried to identify the significant risk factors of IHD i.e. the features which are significantly correlated with IHD by applying different feature selection techniques. Then IHD was predicted using significant risk factors by applying different classifier algorithms. The significant risk factors of IHD were determined by using Chi-Square correlation, Ranking the features based on information gain and Best First Search techniques. Among 151 collected features only 28 features showed high correlations with IHD based on 0.05 significance level and information gain 1% or above. 10-fold cross-validation technique was applied with different classification algorithms e.g. Artificial Neural Network (ANN), Bagging, Logistic Regression, and Random Forest to predict IHD using the most significant 28 risk factors. IHD prediction accuracy was observed ranges from 95.85% to 97.63% with different classifier algorithm. Random Forest showed the best prediction performance with an accuracy of 97.63%. The same processing technique and classification algorithms were applied to the Cleveland hospital dataset to validate our prediction approach. The observed IHD prediction accuracy was 80.46-83.77% without applying the proposed processing techniques, but the accuracy degraded to 79.80-81.46% applying the proposed processing techniques. The Cleveland hospital data contains 303 patients' data with only 13 features whereas the collected dataset contains 506 patient's data with 28 nicely correlated IHD risk factors. This is why the proposed method is not suitably applicable to Cleveland dataset.

Keywords- Machine Learning, IHD, Prediction, Artificial Neural Network, Bagging, Logistic Regression, Random Forest.

Contents

		PAGE
	Cover Page	i
	Declaration	ii
	Approval	iii
	Acknowledgment	iv
	Abstract	v
	Contents	vi
	List of Tables	ix
	List of Figures	x
	List of Abbreviations	xi
CHAPTER I	INTRODUCTION	1
	1.1 Ischemic Heart Disease	1
	1.2 State of the Art	2
	1.3 Objectives	4
	1.4 Thesis Organization	4
CHAPTER II	LITERATURE REVIEW	6
	2.1 Risk Factors	6
	2.2 Risk Factors for Ischemic Heart Disease	6
CHAPTER III	METHODOLOGY	18
	3.1 Work-flow of the Study	18
	3.2 Data Collection	18
	3.3 Simulation Tools and Software packages for Data Analysis	18
	3.4 Data Preprocessing	19
	3.5 Risk Factor Selection	19
	3.5.1 Best First Search (BFS)	20
	3.5.1.1 CfsSubsetEval (Correlation-based Feature Subset Selection)	20
	3.5.2 Ranker	20
	3.5.2.1 InfoGainAttributeEval (Information Gain Attribute Evaluation)	22
	3.5.3 Chi-square Correlation	22
	3.5.3.1 Finding P-Value	22
	3.6 Selection Process of Most Significant Risk factors based on Different Feature Selection Techniques	23

	3.7 Dataset Training and Testing	23
	3.7.1 Cross-validation	23
	3.7.2 10-fold cross-validation	23
	3.8 Prediction of IHD using Classification Algorithms	24
	3.9 Bagging	24
	3.9.1 Implementation of Bagging Algorithm in WEKA	25
	3.9.2 Advantages of Bagging	26
	3.9.3 Disadvantages of Bagging	26
	3.10 Logistic Regression	26
	3.10.1 Implementation of Logistic Regression Algorithm in WEKA	27
	3.10.2 Advantages of Logistic Regression	28
	3.10.3 Disadvantages of Logistic Regression	28
	3.11 Artificial Neural Network (ANN)	29
	3.11.1 Implementation of Artificial Neural Network (ANN) Algorithm in WEKA	29
	3.11.2 Advantages of Artificial Neural Network (ANN)	30
	3.11.3 Disadvantages of Artificial Neural Network (ANN)	30
	3.12 Random Forest	31
	3.12.1 Implementation of Random Forest Algorithm in WEKA	32
	3.12.2 Advantages of Random Forest	32
	3.12.3 Disadvantages of Random Forest	33
CHAPTER IV	RESULTS AND DISCUSSIONS	34
	4.1 Data Distributions of Collected Dataset	34
	4.2 Outcome of Feature Selection	34
	4.2.1 Outcome of Best First Search (BFS)	34
	4.2.2 Outcome of Ranker Algorithm	35
	4.2.3 Outcome of Chi-square Correlation	37
	4.2.4 Extracted Mostly Significant Features from the Dataset	37
	4.3 Prediction on Ischemic Heart disease	43
	4.3.1 Outcome of Bagging Algorithm	43
	4.3.2 Outcome of Logistic Regression Algorithm	43
	4.3.3 Outcome of the Random Forest Algorithm	44

4.3.4 Outcome of Artificial Neural Network (ANN) 45
4.4 Comparison of the Performance of Different Machine Learning Algorithms 46
4.5 Receiver Operating Characteristic (ROC) Analysis of Different Algorithms 52
CHAPTER V **CONCLUSION** 56
5.1 Conclusion 56
5.2 Limitations 56
5.3 Future Works 57
References 58
Appendix 70
List of Publications 88

List of Tables

Table No	Description	Page
3.1	Properties and Values of Bagging algorithm	25
3.2	Properties and Values of Logistic Regression algorithm	28
3.3	Properties and Values of ANN algorithm	28
3.4	Properties and Values of RF algorithm	32
4.1	Results for Best First Search	35
4.2	Results for Ranker algorithm	35
4.3	Results of Correlation algorithm	37
4.4	Significant Features List	38
4.5	Comparison between algorithms	46
4.6	Comparison of ANN algorithm's outcome between two datasets	48
4.7	Comparison of Bagging algorithm's outcome between two datasets	49
4.8	Comparison of Logistic Regression algorithm's outcome between two datasets	50
4.9	Comparison of Random Forest algorithm's outcome between two datasets	51
A1	Collected Demographic Features (Part 1)	72
A2	Collected Diagnostic Features (Part 2)	73
A3	Collected Patient's History Features (Part 3)	73
A4	Collected Symptoms Table (Part 4)	74
A5	Dataset with Mostly Significant Features (Partial)	85

List of Figures

Figure No	Description	Page
3.1	Working flow of the analysis	21
3.2	Decision Trees to Merge and to get an Ideal Outcome	31
4.1	Graphical view of Mean	39
4.2	Graphical view of Standard Deviation	40
4.3	Graphical view of some features and percentage	40
4.4	Graphical view of some features and percentage	41
4.5	Graphical view of some features and percentage	41
4.6	Graphical view of some features and percentage	42
4.7	Graphical view of some features and percentage	42
4.8	Graphical view of Bagging	43
4.9	Graphical view of Logistic Regression	44
4.10	Graphical view of Random Forest	45
4.11	Graphical view of ANN	45
4.12	Comparison of Algorithms	47
4.13	ROC Curve of Bagging algorithm (a) For yes Class (b) For no Class	52
4.14	ROC Curve of Logistic Regression algorithm (a) For yes Class (b) For no Class	53
4.15	ROC Curve of Random Forest algorithm (a) For yes Class (b) For no Class	54
4.16	ROC Curve of ANN algorithm (a) For yes Class (b) For no Class	55

List of Abbreviations

IHD	:	Ischemic heart disease
IHA	:	ischemic heart attack
AHA	:	American Heart Association
U.S.	:	United States
CA	:	Coronary Artery
HA	:	Heart attack
CD	:	Cardiovascular Diseases
HD	:	Heart Disease
IEHAPS	:	Intelligent and Effective Heart Attack Prediction System
MLPNN	:	Multilayer Perceptron Neural Network
RBF	:	Radial Basis Function
SCRL	:	Single Conjunctive Rule Learner
SVM	:	Support Vector Machine
ACS	:	Acute Coronary Syndrome
ML	:	Machine Learning
CHD	:	Coronary Heart Disease
HBP	:	High Blood Pressure
HTN	:	Hypertension
HR	:	Hear Rate
bpsystolic	:	Systolic Blood Pressure
bpdiastolic	:	Diastolic Blood Pressure
DLP	:	Dyslipoproteinemia
DM	:	Diabetes Mellitus
RBS	:	Random Blood Sugar
HDL	:	High Density Lipoprotein
EF	:	Ejection Fraction
LDL	:	Low Density Lipoprotein
ACS	:	Acute Coronary Syndromes
PAD	:	Peripheral Artery Disease
UMI	:	Unrecognized Myocardial Infraction
CSA	:	Controlled Substances Act
DCM	:	Dilated Cardiomyopathy
CHB	:	Complete Heart Block
SMR	:	Severe Mitral Regurgitation
SMS	:	Severe Multiple Sclerosis
Non-STEMI	:	Non-ST-Elevation Myocardial Infarction
ICM	:	Ischemic Cardiomyopathy

AMI	:	Acute Myocardial Infarction
PE	:	Pericardial Effusion
DAA	:	Dissecting Aortic Aneurysm
CKD	:	Chronic kidney disease
PCI	:	Percutaneous Coronary Intervention
ALVF	:	Acute Left Ventricular Failure
RA	:	Rheumatoid Arthritis
SVT	:	Supraventricular Tachycardia
IDCM	:	Idiopathic Dilated Cardiomyopathy
CCF	:	Negative Cardiac Failure
RHD	:	Rheumatoid Heart Disease
SB	:	Symptomatic Bradycardia
ESRD	:	End Stage Renal Disease
OMI	:	Old Myocardial Infarction
FVR	:	Feline Viral Rhinotracheitis
RTI	:	Respiratory Tract Infection
SCA	:	Survivor Cardiac Arrest
VHD	:	Valvular Heart Disease
RVI	:	Right Ventricular Involvement
AV	:	Atrioventricular Block
SK	:	Streptokinase
AKI	:	Acute Kidney Injury
RBBB	:	Right Bundle Branch Block
DVD	:	Double Vessel Disease
TMT	:	Treadmill Test
CAD	:	Coronary Artery Disease
SVD	:	Single Vessel Disease
TVD	:	Triple Vessel Disease
DVT	:	Deep Vein Thrombosis
AWMI	:	Anterior Wall Myocardial Infarction
PTCA	:	Percutaneous Transluminal Coronary Angioplasty
LAD	:	Left Anterior Descending
CHD	:	Coronary heart disease
LBBB	:	Left Bundle Branch Block
PPM	:	Pacemakers
COPD	:	Pulmonary Disease
DKA	:	Diabetic ketoacidosis
PAH	:	Phenylalanine Hydroxylase
POBA	:	Plain old balloon angioplasty

LRT	**:**	Leptin Replacement Therapy
TAVR	**:**	Trans catheter Aortic Valve Replacement
UTI	**:**	Urinary Tract Infection
LMS	**:**	Isolated left main stem
HBsAG	**:**	Hepatitis B Surface Antigen
CVD	**:**	Cardiovascular Disease
BPH	**:**	Benign prostatic hyperplasia
DM	**:**	Diabetes Mellitus
ECG	**:**	Electrocardiogram
CAG	**:**	Coronary Angiography
ANN	**:**	Artificial Neural Network
BFS	**:**	Best First Search
CfsSubsetEval	**:**	Correlation-based Feature Subset Selection
LR	**:**	Logistic Regression
RF	**:**	Random Forest
RBNN	**:**	Radial Basis Neural Network
TDNN	**:**	Time Domain Neural Network
SVM	**:**	Support Vector Machine
CNN	**:**	Convolution Neural Network
ROC	**:**	Receiver Operating Characteristic

CHAPTER I

INTRODUCTION

1.1 Ischemic Heart Disease

Ischemic Heart Disease (IHD) is a terrible experience that occurs when the flow of blood is severely cut off or reduced fully that brings oxygen to the heart muscle. Because of coronary arteries that supply the heart muscle with blood flow being narrowed from a buildup of cholesterol, fat and other substances a person could have experience of ischemic heart attack (IHA). It is the most common cause of death in most western countries. Ischemia means a "reduced blood supply". The coronary arteries supply blood to the heart muscle and no alternative blood supply exists, so a blockage in the coronary arteries reduces the supply of blood to the heart muscle. Most ischaemic heart disease is caused by atherosclerosis, usually, present even when the artery lumens appear normal by angiography. Initially, there is sudden severe narrowing or closure of either the large coronary arteries and/or of coronary artery end branches by debris showering downstream in the flowing blood. It is usually felt as angina, especially if a large area is affected. The narrowing or closure is predominantly caused by the covering of atheromatous plaques within the wall of the artery rupturing, in turn leading to a heart attack (Heart attacks caused by just artery narrowing are rare). A heart attack causes damage to the heart muscle by cutting off its blood supply.

As indicated by the most recent WHO information distributed in 2017 Coronary Heart Disease Deaths in Bangladesh achieved 112,791 or 14.31% of total death [2]. According to the American Heart Association (AHA), in the United States (U.S.) someone has IHA in about every 40 seconds that is really frightening [3]. The age balanced Death Rate is 108.99 per 100,000 of populace positions Bangladesh #104 on the world [2]. Nowadays, In Bangladesh CA is one of the biggest problems in the healthcare sector because 1.06 lakh deaths were caused by ischemic heart disease, 1.78 lakh by strokes and by hypertensive heart disease about 28,000, in 2013 [1]. The main reason of this condition is the food habits of Bangladeshi people [1]. They love to eat too oily, rich foods and they don’t have the habit of doing exercise though many of them are smokers [1].

Unfortunately, we don’t get a warning before IHA as this process (atherosclerosis) has no critical symptoms. Sometimes, when the coronary artery (CA) narrows, the heart expands itself to recoup. As CA becomes narrow, the flow of blood goes down and the supply of Oxygen also reduces. Collateral circulation is such a network that expands nearby blood vessels, it supports from IHA by supplying required blood to the heart. It can be also built after a Heart attack (HA) to help the heart muscle recover. Without knowing many people may have ischemic episodes which are called silent ischemia [4]. According to a new report, about 50% of U.S. adult peoples are affected by some kind of cardiovascular disease (CD) and excessive blood pressure which is also known as hypertension can lead to HA, heart failure and stroke. A statistic of 2016 shows that 48% of U.A. adults have a CD which is almost 121.5 million in numbers. This increase of CD driven to the change of definition for high blood pressure from 140/90 millimeters of mercury to 130/80 [5]. Women have different causes and risks than men in the case of HA for example, women have a 50% higher chance of HA than men due to depression. In 2014, about 50,000 women died from HA [4]. Every year 17.3 million people died from Heart Disease (HD) which ranked HD to number 1 position as a reason for the global cause of death [4], [6]. This report was extracted by compiling health data from more than 190 countries and it is expected that the number will rise to more than 23.6 million by 2030. Whereas stroke is the number 2 cause of death [6]. Though the number of people having recurrent and first strokes went up to 33 million by 2010, the number of deaths went down from 1990 to 2010 [7]. About 11.4 million people died in a range of 30 to 69 years when 15.9 million death people's age are 70 and older [7].

1.2 State of the Art

Many researchers have applied different Machine Learning techniques to find out the best result and identify new risk factors that caused different non-communicable diseases like heart attack [8]. For example, J48 and Naive Bayes have been used to detect the diagnosis and 94.2% accuracy has been found in the J48 algorithm, and in Naive Bayes, it was 82.6% [9]. Another technique has been proposed for detection and analysis of heart disease by utilizing the Naive Bayes algorithm, Support Vector Machine (SVM) and Naive Bayes algorithm provides 74% accuracy where on the other hand, SVM provides 94.60% [10]. A machine learning algorithm was proposed to predict diabetic disorder by using Naive Bayes and Decision trees. Naive Bayes gives 79.56% accuracy and the decision tree provides 76.95% accuracy [11]. A research study was performed on Coronary Heart Disease data about the Correlation of association rules and their application. They have performed the study on two different datasets namely Chinese medicine and Western medicine categories and the study has generated 6 rules for Chinese medicine when the Western medicine categories have generated 7 rules [12]. A technique was developed to predict the risk score of heart disease and got 98% accuracy in Naive Bayes and C4.5 [13]. An analysis was conducted to compare the performance of five systems that use data mining techniques and got 94% accuracy in SVM to predict Heart disease [14]. A study was performed on several significant algorithms like C4.5, Neural Network (Supervised & Unsupervised Learning), Naive Bayesian Classifier and Decision Tree Induction and acknowledged that despite having a large set of data it's not always good to get an accurate result and an intelligent system can reduce diagnosis cost for users [15]. Another intelligent recommendation system has been developed based on short-term risk HD prediction using Time Series Recommendation Algorithm and the accuracy of the system ranges from 75% to 100% [16]. Three algorithms named KNN, ANN, and SVM using voting methods to predict the disease and the accuracy for ANN was 87.50%, KNN and SVM obtained the same accuracy was 88.24% [17]. The dataset from Cleveland hospital contains 303 instances with 74 different attributes but among them, only 14 of them such as Age, Gender, fasting blood sugar Serum cholesterol, Chest pain type were used, Maximum heart rate, Resting Blood Pressure, Exercise induced angina, etc. were taken as referred by the other publishers. The accuracy of this system ranges from a minimum of 86.66% to a maximum of 95.55% for different sizes hidden layer [18]. SVM and Gain Ratio techniques have been used for feature selection and Random Forest and Naive Bayes have used to predict HD and the dataset used for this research includes features age, sex, trestbps, chol, cp, fbs, restecg, thalach, exang, oldpeak, thal, num, ca, slope [19]. An analysis has been performed on three available datasets Heart Disease Database, South African Heart Disease, and Z-Alizadeh Sani Dataset to predict HD and for different datasets, they have got different accuracy [20]. A study on various ML algorithms showed that it is tough to get the best performance by depending on a single algorithm whereas by combining multiple algorithms its possible to get the better result [21]. Different classification techniques like Single Conjunctive Rule Learner (SCRL), Radial Basis Function (RBF) and Support Vector Machine (SVM) have applied to predict HD and the accuracy are 83.82%, 69.96% and 84.15% respectively [22]. An Internet of thing (IoT) based system with 191 instances where each instance has several attributes such as age, sex, systolic and diastolic pressure, heartbeat, sugar etc. was developed for elder patients to monitoring health of the stroke patients and Naive Bayes, KNN and Tree based algorithms have used to predict the diagnosis [23]. An analysis with ML approaches was conducted by a research team to predict the risk of Acute Coronary Syndrome (ACS) [24]. They have analyzed the dataset in three seeds and the best Correctly Classified Instances achieved for AdaBoost is 75.49%, 76.28% for Bagging, 72.33% for K-NN, 75.30% for Random Forest, 72.72% for SVM [24]. Another analysis was conducted to determine the performances of several algorithms named Decision tree (J48), Naive Bayes, Random Forest, Adaboost, Bagging, Multilayer Perceptron, Simple Logistic to predict diseases using data mining techniques [25]. Around 768 instances with 500 tested negatives and 268 tested positives were used one of the experiments and the function *ReplaceMissingValues* in WEKA tool was used to handle missing data [26]. A survey showed that the accuracy of different data mining algorithms such as Naive Bayes, Decision Tree, Neural Network, K-Nearest Neighbour and Logistic Regression to predict heart disease depends on the number of risk factors and their types [27]. A smartphone based risk prediction application has developed to predict Heart Attack and different types of risk factors like hypertension, diabetes, dyslipidemia, smoking, family history, obesity, stress, etc. were collected from 506 different patients with three categories: low, medium and high [28]. The app was tested on 89 participants with having Acute Coronary Syndrome (ACS). About 83.9% of patients with high category had ischemic heart disease (IHD) while only 12.5% are in class low and on the other hand, those who have ACS 86.69% of them had high scores [28]. Real-time data of Bangladeshi patients can give an accurate prediction of having IHD and if the analysis has applied to any intelligent system such as website, mobile application, then better predictive performance can be produced.

1.3 Objectives

This thesis study's goal is to find the most significant risk factors of Ischemic Heart Disease of Bangladeshi patients and predict the IHD using different classification techniques.

Overall, the objectives of this thesis can be summarized as follows:

a. To identify the most significant risk factors of IHD.
b. To find the correlation between IHD and risk factors.
c. To predict Ischemic Heart Disease using mostly significant risk factors by applying different classification algorithms.

1.4 Thesis Organization

The remainder of the thesis is organized as the following:

Chapter 2 (Literature Review): In this chapter, different risk factors of IHD has been discussed with their reference value.

Chapter 3 (Methodology): The working flow of the thesis analysis has been discussed in this chapter. The analysis process includes the proposed techniques, feature selection, correlation, and classification.

Chapter 4 (Experimented Results and Discussions): In this chapter, we will cover the evaluation and results of our simulation model.

Chapter 5 (Conclusion): Finally, the thesis is concluded in this chapter with suggestions for future research.

CHAPTER II

LITERATURE REVIEW

2.1 Risk Factors

A risk factor is any feature, symptoms of an individual that increases the possibility of developing a disease. Several symptoms, food habits, lifestyle, age, family history, diabetes, smoking can increase the risk of chances of getting a heart attack. These are called risk factors. Some risk factors are not controllable and lead to an ischemic heart attack.

2.2 Risk Factors for Ischemic Heart Disease

With the help of doctors of AFC Fortis Escorts Hospital Khulna, Bangladesh a questionnaire has been prepared and attached in the appendix section. Each patient has a unique MRD no to identify that patient's conditions. Around 5 demographic risk factors such as age, sex, height, weight and BMI are leads to IHD. Total 12 diagnostic risk factors such as heart rate, systolic blood pressure, diastolic blood pressure, hemoglobin, creatinine, RBS, total cholesterol, HDL, troponin, ejection fraction, triglycerides, LDL and patient's history like smoking duration, smoking per day, hypertension duration, diabetes mellitus, etc. can lead to IHD.

Most of the people die due to heart disease at age 65 [29]. Although It is risky for both gender, women have more possibility to have heart disease [29]. Around 72% of people who are in business professionals have a possibility to have heart disease [30]. Modern women have professional and housewife responsibilities, consume an excess of fat and carbohydrates, smoke, do not exercise regularly and do not have enough time to rest [31]. Salesman over 45 age has more possibility to have heart disease. [32]. Around 82.1% of farmers have heart disease a stroke or diabetes [33]. Clinicians have suspected that retired people can be affected by heart disease [34]. Unemployment people have about 50% chance to die by having heart disease, according to study, in 2017 [35]. Lawyers have a chance to fall in heart attach after age 30 [36]. The prevalence of CVD risk factors was high and featured hypertension (48.5%) and 18.7% were at high risk of a heart attack or stroke within 10 years [37]. The younger people who are in the banking profession are usually getting admission to the hospital due to heart-related issues. [38]. Polices have greater chance to increase of the prevalence of CHD risk factors depending on age, and a large number of

CHD risk factors was significantly higher in the group of people over 45 year of age [39]. Adults aged 18 to 24 years have at least 1 coronary heart disease (CHD) risk factor. The number of risk factors is related to the extent of atherosclerosis [40]. During delivery, about 1,061 women had a heart attack who are involved with labor works, during their pregnancy 922 had heart attacks and after giving birth 2,390 women had a heart attack [41]. According to researchers, for each two and a half inches shorter, the possibility of having coronary heart disease increases 13.5 percent [42]. For age, 1 to 10 years' normal weight varies from 7.6 to 33.8 kg, for age 10 to 20 years its 37.6 to 71.8 kg and from 20 to 75 years' weight range is 71.8 to 74 kg depending on their heights. Depending on heights and age BMI can be higher or lower [43]. The ideal BMI ranges from 18.5 to 24.9 [44]. A normal heart rate range is 60 to 100 beats per minute [45]. Normal systolic blood pressure is 120 and normal diastolic blood pressure is 80 [45]. Any duration is harmful for heart. The more time anyone is smoking the more possibility to have heart disease. There is no such limitation to smoke which is good for heart [46]. It is always harmful for heart either anyone has it a little bit or huge. Due to excessive strain, the heart veins can be narrowed and fat can be increased which are not good for heart [47]. It is a range of disorders of lipoprotein lipid metabolism which includes abnormally high and low lipoprotein densities. Abnormalities in the composition of the lipoprotein particles [48]. A reading of more than 200 mg/dL (11.1 mmol/L) after two hours indicates diabetes. A reading between 140 and 199 mg/dL (7.8 mmol/L and 11.0 mmol/L) defines prediabetes [49]. Normally heart beat is from 60 to 100 times per minute. The rate may drop below 60 beats per minute who exercise continuously or have medicines. If your heart rate is over 100 beats per minutes, this is called tachycardia [50]. Syncopal episodes usually last only seconds or minutes [51]. Chest pain can lasts more than 5 days in IHD [52]. For men, 13.5 to 17.5 grams per deciliter and for women, 12.0 to 15.5 grams per deciliter is the normal range [53]. The normal serum creatinine range is 0.6–1.1 mg/dL for women and 0.7–1.3 mg/dL for men [54]. RBS normal range should be between 80 mg/dl and 130 mg/dl [55]. When the Low-Density Lipoprotein (LDL)-cholesterol is over 4.5 mmol/L you are at high risk; when it's between 3.5 and 4.5 mmol/L, you are at borderline-high risk. Ideally your LDL-cholesterol should be below 3.5mmol/L [56]. High Density Lipoprotein (HDL) cholesterol levels greater than 60 milligrams per deciliter (mg/dL) are high. If HDL cholesterol level is less than 40 mg/dL then it is low and it is not good [57]. The normal reference range is listed as 0.00 – 0.40 [58]. The normal percentage of Ejection Fraction (EF) is 55 [59]. It is a type of fat by which we can measure the cholesterol level in blood. Excessive Triglycerides increases the chance of heart attack [60]. High Low-Density Lipoprotein (LDL) level is very bad for heart [60]. Acute Coronary Syndromes (ACS) reduces blood flow in heart functions and indicates a risk of heart attack. It often causes chest pain [61]. Old Myocadiac Infarction (OMI) generally called heart attack which occurs when blood flow down or stops to a certain part of heart and causes damage to the heart muscle [61]. Inferior Wall ST Segment Elevation Myocardial Infarction (IWMI) can cause heart failure, irregular heart beat, cardiogenic shock and cardiac arrest. Most of the time IWMI occurs due to coronary artery disease (CAD) [62]. Peripheral Artery Disease (PAD) may decrease blood flow to heart and brain, as well as legs [63]. People with Stable Angina may have chest pain. It might experience during dealing with stress [64]. Unstable Angina is one kind of chest pain which occurs at rest or with stress or exertion. It can lead to heart failure as well as heart attack [64]. Unrecognized Myocardial Infraction (UMI) is known to constitute a substantial. CAD is the infarction (MI) directs clinical management and affects prognosis [65]. Central Sleep Apnea, Mixed Sleep-disordered Breathing, and Systolic Heart Failure [66]. Dilated Cardiomyopathy (DCM) is a condition when the hear becomes enlarged and It can’t pump blood properly [67]. Complete Heart Block (CHB) is a rare disorder which has an incidence of about 1 in 22,000 live births. It may be related with high morbidity and mortality [68]. If the damage is extensive enough, a heart attack can cause sudden and severe mitral valve regurgitation. Abnormality of the heart muscle (cardiomyopathy). Over time, certain conditions, such as high blood pressure, can cause your heart to work harder, gradually enlarging your heart's left ventricle. Severe Multiple Sclerosis (SMS) can affect cardiovascular function in different of ways which lead to abnormalities in blood pressure, heart rhythm, heart rate, left ventricular systolic function etc. [69]. Non-ST-Elevation Myocardial Infarction (Non-STEMI) is a type of heart attack which stands for non-ST-elevation myocardial infarction. ST refers to the ST segment [70]. The implantable cardiac monitor (ICM) is a small device implanted subcutaneously into the left side of the chest, offering three to four years of battery life. ICMs have the advantage of long-term arrhythmia monitoring while also allowing for patients to self-capture and record symptomatic events [71]. Acute Myocardial Infarction (AMI) is usually the result of a blockage in one or more of the coronary arteries [72]. Due to limited amount of space in the pericardial cavity, fluid accumulation leads to an increased intrapericardial pressure which can negatively affect heart function. EF with pressure to affect heart function is known as cardiac tapenade [73]. Dissecting Aortic Aneurysm (DAA) occurs if a weak spot in the wall of your aorta begins to bulge that can occur anywhere in the aorta [74]. It reduces the flow of blood in heart. As a result, it may cause heart related issues.

Diabetes and high blood pressure are the two leading causes of Chronic kidney disease (CKD) which increases risk for heart attack, heart failure and stroke [75]. It is an excess strain resulting damage from high blood pressure causes the coronary arteries serving to become narrowed from plaque [46]. Percutaneous Coronary Intervention (PCI) known as angioplasty with stent is a non-surgical process which uses a thin flexible tube to place a tiny structure known as a stent to open up blood vessels in heart that was narrowed by plaque buildup [76]. Acute Left Ventricular Failure (ALVF) is actually not a disease but a process, when diastolic heart failure happens, the left ventricle has grown stiff or thick. It failed to fill the lower left chamber of heart properly and as a result it reduces the amount of blood pumped out to the body [77]. Lung complications from Rheumatoid Arthritis (RA) can be dangerous and fatal. Although it is known as RA, it not only affects joints but also tissues, eyes, heart and lungs [78]. Supraventricular Tachycardia (SVT) is a condition which doesn't cause certain death. It will damage the heart very slowly in terms of time which is really serious [79]. Idiopathic Dilated Cardiomyopathy (IDCM) affect mitochondrial activity and myocardial metabolism [80]. Although there have been current improvements in heart failure treatment, researchers say the prognosis for people with the disease is still bleak, with about 50% having an average life expectancy of less than five years. For those with advanced forms of heart failure, nearly 90% die within one year [81]. Rheumatoid Heart Disease (RHD) is caused by rheumatic fever which is an inflammatory disease that can affects a number of connective tissues, specially in the heart, joints, skin, or brain. It takes years to build and can result in heart failure [82]. It is a condition when heart rate becomes slower than normal heart rate. Symptomatic Bradycardia (SB) is a serious problem as if heart doesn't response well then, the flow of Oxygen will also be reduced [83]. When patients living with kidney failure choose to forgo dialysis, how long they can live depends on the amount of kidney function they have, how severe their symptoms are and their overall medical condition [84]. Atrial fibrillation (AF or A-fib) is an abnormal heart rhythm characterized by rapid and irregular beating of the atria. ... The disease is associated with an increased risk of heart failure, dementia, and stroke. Feline Viral Rhinotracheitis (FVR) is a type of supraventricular tachycardia [85]. Respiratory infections can trigger a heart attack. Researchers of University of Sydney finds the danger of a heart attack is expanded 17-overlap in the week following a respiratory disease, for example, flu or pneumonia [86].

Hemorrhagic Shock is a clinical disorder coming about because of diminished blood volume brought about by blood misfortune, which prompts decreased cardiovascular yield and organ perfusion [87]. Cardiogenic shock is a condition where your heart all of a sudden can't siphon enough blood to address your body's issues. The condition is regularly brought about by a serious heart attack [88]. Aortic Stenosis occurs if heart's aortic valve narrows down. It reduces the blood flow [89]. Coronary Artery Bypass Graft (CABG) is a kind of surgery which improves flow of blood to the heart [90]. Valvular Cardiomyopathy is characterized by damage of a defect in one of the four heart valves: the mitral, aortic, tricuspid or pulmonary [90]. The average survival rate is 10.6% and survival with good neurologic function is 8.3%. About one in three victims survives when the SCA occurs [91]. Valvular Heart Disease (VHD) is defined by the damage in one of the four heart valves [92]. Right Ventricular Involvement (RVI) complicates up to 40% of inferior STEMIs [93]. Pneumonia is a special problem to heart patients. Even, the flu may cause complications, including bacterial pneumonia, or the worsening of chronic heart issues. It is a lung infection that prevents your lungs from getting enough oxygen in the blood, creating a strain on the heart [94]. First-degree atrioventricular block (AV block), or PR prolongation, is a disease of the electrical conduction system of the heart in which the PR interval is lengthened beyond 0.20 seconds [95]. Sinus rhythm is defined by the presence of correctly oriented P waves on the ECG. Sinus rhythm is essential, but not enough, for normal electrical activity in the heart [96]. The epicardial coronary arteries are analyzed before expanding the heart chambers. The left and right coronary arteries emerge from their individual aortic sinus [97] [98]. It is a thrombolytic medication and enzyme. It is utilized to separate clumps sometimes of heart attack, aspiratory embolism, and blood vessel thromboembolism [99]. It occurs if the heart stops beating. It is often caused by an electrical issue which causes the heart muscle to beat ineffectively [100]. Ventricular tachycardia (V-tach or VT) is a type of regular, fast heart rate that arises from improper electrical activity in the ventricles of the heart. Ventricular tachycardia may result in cardiac arrest and turn into ventricular fibrillation [101]. Hypothyroidism affects the heart in a number of ways. Insufficient thyroid hormone represents heart rate. It also prevents the arteries from becoming more elastic; blood pressure increases in order to circulate blood [102]. Ventricular Bigeminy is one of the heart rhythm issues in which there are looped rhythms heart beats, one is longer and another is shorter [103]. Vertigo is also known by orthostatic hypotension which decrease the flow of blood. Atrial Fibrillation is an irregular heart beat that can lead to stroke, heart failure and other heart-related issues [104]. Acute Kidney Injury (AKI) occurs during heart failure that has been labelled cardiorenal syndrome type-1 [105]. Right Bundle Branch Block (RBBB) is a heart block in the right bundle part of the electrical conduction framework. During a correct pack part block, the right ventricle isn't straightforwardly initiated by driving forces going through the right group branch. It is a cardiac stress test. A male patient walks on a stress test treadmill to have his heart's function checked [106]. CAD is the most common kind of heart disease. It happens when the arteries which supply blood to heart muscle become narrowed [107]. Single Vessel Disease (SVD) is also called coronary microvascular disease or small vessel heart disease. It's frequently analyzed after a specialist discovers practically zero narrowing in the primary supply routes of your heart, regardless of having side effects that propose coronary illness [108]. Blockage of this vessel can have severe consequences. Patients having triple vessel disease, meaning that 3 large vessels have blockages from atherosclerotic plaques. The left ventricle also has decreased contractile function [109]. Deep Vein Thrombosis (DVT) occurs if blood clumps together to form a clot, normally in the legs. The stress is that if this coagulation turns out to be free, it can go through your circulation system and become stopped in your lung. At the point when this occurs, the blood to your lungs is blocked [110]. The coronary arteries take oxygen rich blood specifically to your heart muscle. When these arteries become blocked or narrowed due to a buildup of plaque, the blood flow to your heart can decrease significantly or stop completely. This can cause a heart attack. Percutaneous Transluminal Coronary Angioplasty (PTCA) is a non-surgical procedure that relieves narrowing and obstruction of the arteries to the muscle of the heart (coronary arteries). This allows more blood and oxygen to be delivered to the heart muscle. This term is used because the left main coronary and/or the left anterior descending supply blood to large areas of the heart. This means that if these arteries are abruptly and completely occluded it will cause a massive heart attack that wills likely lead to sudden death [111]. Congenital coronary fistulas are rare and generally isolated. These anomalies are described as a connection between one or various coronary arteries and a cardiac chamber or large vessel. However, we found no cases in the literature describing a fistulous tract that connected a coronary artery with the aorta [112]. Congestive hepatopathy, also known as nutmeg liver and chronic passive congestion of the liver, is liver dysfunction due to venous congestion, usually due to congestive heart failure [113]. Dyslipoproteinemia also referred to as dyslipidemia, encompasses a range of disorders of lipoprotein lipid metabolism that include both abnormally high and low lipoprotein concentrations, as well as abnormalities in the composition of these lipoprotein particles [114]. Renal artery stenosis is the narrowing of one of the renal arteries, most often caused by atherosclerosis or fibromuscular dysplasia. This narrowing of the renal artery can impede blood flow to the target kidney, resulting in renovascular hypertension – a secondary type of high blood pressure [115]. A number of disorders can display psychotic symptoms, including: Schizophrenia - a serious mental health disorder affecting the way someone feels, thinks, and acts. Bipolar psychosis - individuals have the symptoms of bipolar disorder (intense highs and lows in mood) and also experience episodes of psychosis [116]. In hypothyroidism, an underperforming thyroid gland makes insufficient thyroid hormone, which may affect nearly every organ in the body, including the heart. In heart failure patients, it's found that both hypothyroidisms overall and subclinical hypothyroidism increased the risk of death [117]. It is a major cause of illness and death. Coronary heart disease (CHD) normally happens when cholesterol accumulates on the artery walls, creating plaques. The arteries narrow, reducing blood flow to the heart. Examples of coronary heart disease include angina and heart attack [117]. Mild heart attack either occurs when there is a blockage in one of the smaller arteries of the heart or the blockage does not completely stop the flow of blood to the heart muscle. Mild heart attacks may also cause damage to a smaller portion of the heart, less of which is permeant. Urosepsis is sepsis with a source localized to the urinary tract (or male genital tract, e.g. prostate) [118]. Left Bundle Branch Block (LBBB) is a cardiac conduction abnormality seen on the ECG [119].

A pacemaker is a device that sends small electrical impulses to the heart muscle to maintain a suitable heart rate or to stimulate the lower chambers of the heart (ventricles). A pacemaker may also be used to treat fainting spells (syncope), congestive heart failure and hypertrophic cardiomyopathy [120]. The role of anemia in the progression of congestive heart failure. The anemia itself can worsen cardiac function, both because it causes cardiac stress through tachycardia and increased stroke volume, and because it can cause a reduced renal blood flow and fluid retention, adding further stress to the heart [121]. This occurs when low oxygen levels due to Chronic Obstructive Pulmonary Disease (COPD) cause a rise in blood pressure in the arteries of the lungs, a condition known as pulmonary hypertension. This increase in pressure places excess strain on the heart's right ventricle as it works to pump blood through the lungs [122]. Electrolyte imbalance. Substances in your blood called electrolytes such as potassium, sodium, calcium and magnesium help trigger and conduct the electrical impulses in your heart. Electrolyte levels that are too high or too low can affect your heart's electrical impulses and contribute to arrhythmia development. Ischemic hepatitis is damage throughout the liver caused by an inadequate blood or oxygen supply. The liver diseases affecting the heart include complications of cirrhosis such as hepatopulmonary syndrome, Porto pulmonary hypertension, pericardial effusion, and cirrhotic cardiomyopathy as well as noncirrhotic cardiac disorders such as high-output failure caused by intrahepatic arteriovenous fistulae [123]. Most patients came from emerging countries (75%). Mitral stenosis (MS) with or without mitral regurgitation (MR) was present in 273 women, isolated MR in 117. Hospital admission occurred in 23.1% of the women with MS, and the main reason was heart failure (mild MS 15.8%, moderate 23.4%, severe 48.1%; $P<0.001$) [52]. Metabolic encephalopathy is a problem in the brain. It is caused by a chemical imbalance in the blood. The imbalance is caused by an illness or organs that are not working as well as they should. It is not caused by a head injury [124]. The causes of musculoskeletal pain are varied. Muscle tissue can be damaged with the wear and tear of daily activities. Trauma to an area (jerking movements, auto accidents, falls, fractures, sprains, dislocations, and direct blows to the muscle) also can cause musculoskeletal pain. A buildup of plaque (atherosclerosis) inside the artery wall reduces blood flow to the brain. Atherosclerosis that is severe enough to cause symptoms carries a high risk of stroke and can lead to brain damage and death [125]. This can lead to loss of blood flow if left untreated. The arteries that supply blood to the bowel may become narrowed or blocked from cholesterol buildup. When this happens in the arteries to the heart, it causes a heart attack. When it happens in the arteries to the intestine, it causes intestinal ischemia [126]. The metabolic changes associated with DKA have the potential to cause significant cell damage, illness or death. Some of the possible complications of DKA are listed below: Complications from associated illnesses. For example, infection, stroke, and heart attacks are possible. Pulmonary Hypertension patients develop inflammation and mutations in the cells that line the pulmonary arteries. As a result, the heart cannot properly pump blood through the pulmonary arteries into the lungs, which increases the pressure in the arteries [127]. The group most at risk for hyponatremia in heart failure is female geriatrics with low body mass [127]. The group most at risk for hyponatremia in heart failure is female geriatrics with low body mass [128]. Symptotic Nodal bradycardia with stpm in which the stimulus of the heart's contraction arises in the atrioventricular node or common bundle and able to show noticeable symptoms. Asymptotic Nodal bradycardia with stpm in which the stimulus of the heart's contraction arises in the atrioventricular node or common bundle but fails to show noticeable symptoms.

The syndrome affects the bones, eyes, skin, lungs, and nervous system along with the heart and blood vessels [128]. Cough-variant asthma is a type of asthma in which the main symptom is a dry, non-productive cough. (A non-productive cough does not expel any mucus from the respiratory tract.) People with cough-variant asthma often have no other "classic" asthma symptoms, such as wheezing or shortness of breath [129].

Plain old balloon angioplasty (POBA), where the lumen stenosis of an artery has been treated by balloon dilatation only, without applying a stent. POBA - Where the Arts Live, a non-profit, online arts community, which preserves and display the creative legacies of 20th and 21st century artists [130]. Although it is usually difficult to single out hyponatremia as the cause of conduction defects which usually occur in the presence of cardiac disease, potent medications or other electrolyte abnormalities, we suggest that hyponatremia may play a role in the pathogenesis of conduction defects in the diseased heart [131].

Bronchial asthma is a medical condition which causes the airway path of the lungs to swell and narrow. Due to this swelling, the air path produces excess mucus making it hard to breathe, which results in coughing, short breath, and wheezing. The disease is chronic and interferes with daily working [132]. If anemia is severe, fainting may occur. Other symptoms include brittle nails, shortness of breath, and chest pains. Blood oxygen levels can be so low that a person with severe anemia can have a heart attack [133].

Transcatheter aortic valve replacement (TAVR) is safe in low-risk patients with symptomatic severe aortic stenosis [134]. The role of anemia in the progression of congestive heart failure. The anemia itself can worsen cardiac function, both because it causes cardiac stress through tachycardia and increased stroke volume, and because it can cause a reduced renal blood flow and fluid retention, adding further stress to the heart [135].

Typically, Urinary Tract Infection (UTI) cause urinary symptoms, such as pain or burning during urination. Some mild bladder infections may go away on their own within a couple of days. Kidney infections and UTIs that are complicated by other factors require longer treatment. Complications of UTIs are not common but do occur [136].

Pleural effusion in heart failure results from increased interstitial fluid in the lung due to elevated pulmonary capillary pressure. Pleural effusion in heart failure was also believed to be transudative only occurring as a result of systemic factors [137].

The inflammation associated with psoriasis can increase the risk of heart attack and stroke. That's because inflammation can damage arteries. This results in the formation of blockages or plaques (different from psoriasis skin plaques) inside the blood vessels that supply the heart muscle [138]. Isolated left main stem (LMS) disease occurs in 9% of patients with coronary artery disease (CAD). It is more commonly accompanied with atherosclerotic lesions in other coronary arteries. CABG has been the gold standard for treating unprotected LMS disease [139]. Hypoxic Ischemic Encephalopathy. Hypoxic-ischemic encephalopathy is the most common cause of neonatal seizures, and results from a lack of oxygen and blood flow to the newborn during the perinatal period. Perinatal asphyxia, more appropriately known as hypoxic-ischemic encephalopathy (HIE), is characterized by clinical and laboratory evidence of acute or subacute brain injury due to asphyxia. The primary causes of this condition are systemic hypoxemia and/or reduced cerebral blood flow (CBF) (see the image below). Birth asphyxia causes 840,000 or 23% of all neonatal deaths worldwide [140]. Complications of CKD and heart disease. This can increase your chances of having a heart attack. High blood pressure: Damaged kidneys may release too much of an enzyme called renin, which helps to control blood pressure. This increases the risk for heart attack, congestive heart failure and stroke.

A positive Hepatitis B Surface Antigen (HBsAG) test result means that you are infected and can spread the hepatitis B virus to others through your blood. anti-HBs or HBsAb (Hepatitis B surface antibody) - A "positive" or "reactive" anti-HBs (or HBsAb) test result indicates that a person is protected against the hepatitis B virus [141].

Cardiovascular disease (CVD) is a class of diseases that involve the heart or blood vessels. CVD includes coronary artery diseases (CAD) such as angina and myocardial infarction (commonly known as a heart attack). Coronary artery disease, stroke, and peripheral artery disease involve atherosclerosis. Benign prostatic hyperplasia (BPH), also called prostate enlargement, is a noncancerous increase in size of the prostate. Symptoms may include frequent urination, trouble starting to urinate, weak stream, inability to urinate, or loss of bladder control. One of the major risk factors of cardiovascular diseases is arteriosclerosis which is secondary to the excess of LDL cholesterol. Dyslipidemia is a metabolic abnormality leading to a persistent increase in the plasmatic concentration of cholesterol and triglycerides. If you have a family history of cardiovascular disease, you have an increased risk of developing cardiovascular diseases such as coronary heart disease, angina, heart attack, heart failure and stroke [142].

Over time, smoking contributes to atherosclerosis and increases your risk of having and dying from heart disease, heart failure, or a heart attack. Compared with non-smokers, people who smoke are more likely to have heart disease and suffer from a heart attack. Smoking damages the lining of your arteries, leading to a buildup of fatty material (atheroma) which narrows the artery. This can cause angina, a heart attack or a stroke. The carbon monoxide in tobacco smoke reduces the amount of oxygen in your blood [143]. Diabetes can cause blood sugar to rise to dangerous levels. Medications may be needed to manage blood sugar. Smoking puts individuals, whether or not they have diabetes, at higher risk for heart disease and stroke. Regular exercise also improves factors linked to cardiovascular health, resulting in lower blood pressure, healthier cholesterol levels, and better blood sugar regulation. Stress may affect behaviors and factors that increase heart disease risk: high blood pressure and cholesterol levels, smoking, physical inactivity and overeating. Your body releases adrenaline, a hormone that temporarily causes your breathing and heart rate to speed up and your blood pressure to rise [144].

A heart attack occurs when blood flow to a part of the heart is blocked, usually by a blood clot. Without oxygenated blood, the heart muscle begins to die. A stroke is a brain attack, cutting off vital blood flow and oxygen to the brain. Stroke happens when a blood vessel feeding the brain gets clogged or bursts [145]. Medical drug history is very important in diagnosing the actual disease. Apart from preventing prescription errors, proper medication histories are very useful in detecting proper disease. Dyspnea (shortness of breath) is a common symptom affecting as many. With many different underlying conditions, it can be occurred and some can be life-threatening. Palpitations can be a sign of a severe problem, such as an overactive thyroid gland (hyperthyroidism) or an abnormal heart rhythm (arrhythmia). Fast heart rate causes Arrhythmias (tachycardia) and slow heart rate causes bradycardia or an irregular heart rhythm [146]. Temporary loss of consciousness usually related to inadequate amount of blood flow to the brain is called syncope. It occurs when blood pressure is too low (hypotension) and the heart doesn't pump enough oxygen to the brain [147]. Vomiting occurs as one of the symptoms in case of Heart Attract specially for women than men. In case of heart disease, radiation means to spread the pain in left arm or any part of the body [148]. Chest pain symptoms can range from a mild sensation, to a severe pain. In older people heart attacks are more common than younger people, but they can also occur in people of any age [149].

An electrocardiogram (ECG) is a diagnostic test which measures the electrical activity of heart to show whether or not it is functioning normally. It will be used to record the heart's rhythm and activity on a moving strip of paper or a line on a screen [150]. A coronary angiogram is a procedure that uses X-ray imaging to see your heart's blood vessels. The test is generally done to see if there's a restriction in blood flow going to the heart [151].

CHAPTER III

METHODOLOGY

3.1 Work-flow of the Study

At first, the Questionnaire has been prepared as added in the appendix section and verified by the cardiologist of AFC Fortis Escorts Hospital. Data has been saved to the excel sheet very carefully. Missing information has been filled out with Mean, Median and Mode. Best First Search, Ranker algorithm and Chi-square correlation have been applied to identify the most significant risk factors. To train the dataset cross-validation techniques have been applied and four classification algorithms such as Bagging, Logistic Regression, Artificial Neural Network (ANN) and Random Forest have been applied to determine different statistical metrics and compared the performances among them. The overall work-flow of this research study has been shown in Figure 3.1.

3.2 Data Collection

Data has been collected from AFC Fortis Escort Heart Institute, Khulna, Bangladesh. Around 506 patient’s data with 151 features have been collected and added in the appendix and the agreement documents have been added to the appendix section. Four categories of data have been collected. Some are diagnostic, some are patients history, some are demographic and some symptomatic data.

3.3 Simulation Tools and Software packages for Data Analysis

WEKA (Waikato Environment for Knowledge Analysis) was developed at the University of Waikato, New Zealand. It is a software written in Java for ML study. WEKA comes with a GUI that makes it easy to visualize datasets and train and compare different classifiers. It is a handy tool to have for ML analysis. WEKA contains several ML algorithms that help to study on data mining [152] [153].

IBM SPSS Statistics is the world’s leading statistical tools. Compared to other data analysis packages it’s easier to use has a lower total cost of ownership and it covers every step of the entire analytical process from planning all the way to development. It is available with the required function to perform needed analysis for study and organizations. It empowers us to clarify the relationship between variables create clusters identify trends and make predictions these are powerful. We can quickly access, manage and analyze any kind of data set including survey data, corporate databases, and medical data. It works with all common data types of external programming languages. It helps analysist to analysis on the dataset with its build-in library functions which include different ML algorithms and it also shows a graphical view of information [154].

3.4 Data Preprocessing

Data Preprocessing is very essential to deal with real-time data. In real-time data, lots of irregularities can happen. For example, we notice lots of missing information in the data. In this thesis, the dataset has been preprocessed using WEKA (version 3.8.3). Some unnecessary columns have been removed like name, contact info, address, etc. For filling up missing data, a popular function named **ReplaceMissingValue** (Mean and Mode) has been used. **ReplaceMissingValue** works based on mean and median. It replaces both nominal and numerical missing values of a dataset [155]. It skips the outcome column automatically. The filter may result in additional data to the console if **debug** set to true and the filter's capabilities are not checked except it is built if **doNotCheckCapabilities** set. A function **ignoreClass** by using that before applying the filter the index of the class will remain unset. Then the function **Randomize** has been used to shuffle the order randomly instance of skipping it [156]. **RandomSeed** function will generate a random number of seeds. Then the new dataset has saved for further process.

3.5 Risk Factor Selection

As we know, in day to day life our food habits are changing. We have adopted a new lifestyle in this 21^{st} era. We need to identify which risk factors are more significant to arise IHD. All refactors are not significant and we need to identify their significances.

For selecting risk factors, a couple of algorithms have used such as Best First Search (BFS) Ranker and Chi-square correlation. For BFS **CfsSubsetEval** need to be used and for Ranker the **InfoGainAttributeEval** function has used in WEKA. The brief about the algorithms have been given below:

3.5.1 Best First Search (BFS)

BFS is an algorithm to search and find out the most promising node by traversing every node of a graph. It uses a technique known as a heuristic to determine an ideal option and to verify the option [157]. WEKA tool has been applied to use the BFS for selecting features from the dataset. WEKA has a build-in class named **BestFirst** [158] which helps to select features by evaluating features [159].

3.5.1.1 CfsSubsetEval (Correlation-based Feature Subset Selection)

Another function in WEKA named **CfsSubsetEval** which evaluates the value of a subset of features by considering the individual prescient capacity of each element alongside the level of repeated value between them. Subsets of attributes that are exceptionally related with the class while not having a high relationship are preferred [160]. It helps to select the important features from a dataset and also the related subsets of the features.

3.5.2 Ranker

As we mentioned earlier, we need to find the significances between the risk factors and ranking the risk factors that can be a good approach. Ranker is a scheme of ranking for each feature. It is not a method for searching. It ranks each feature sequentially based on their evolutions and it also removes the lower-ranked attributes ranked by itself. It is possible to set a custom threshold for deleting the lower-ranked features. It is also possible to specify specific features that must need to be there in the list whatever it is in lower or upper rank.

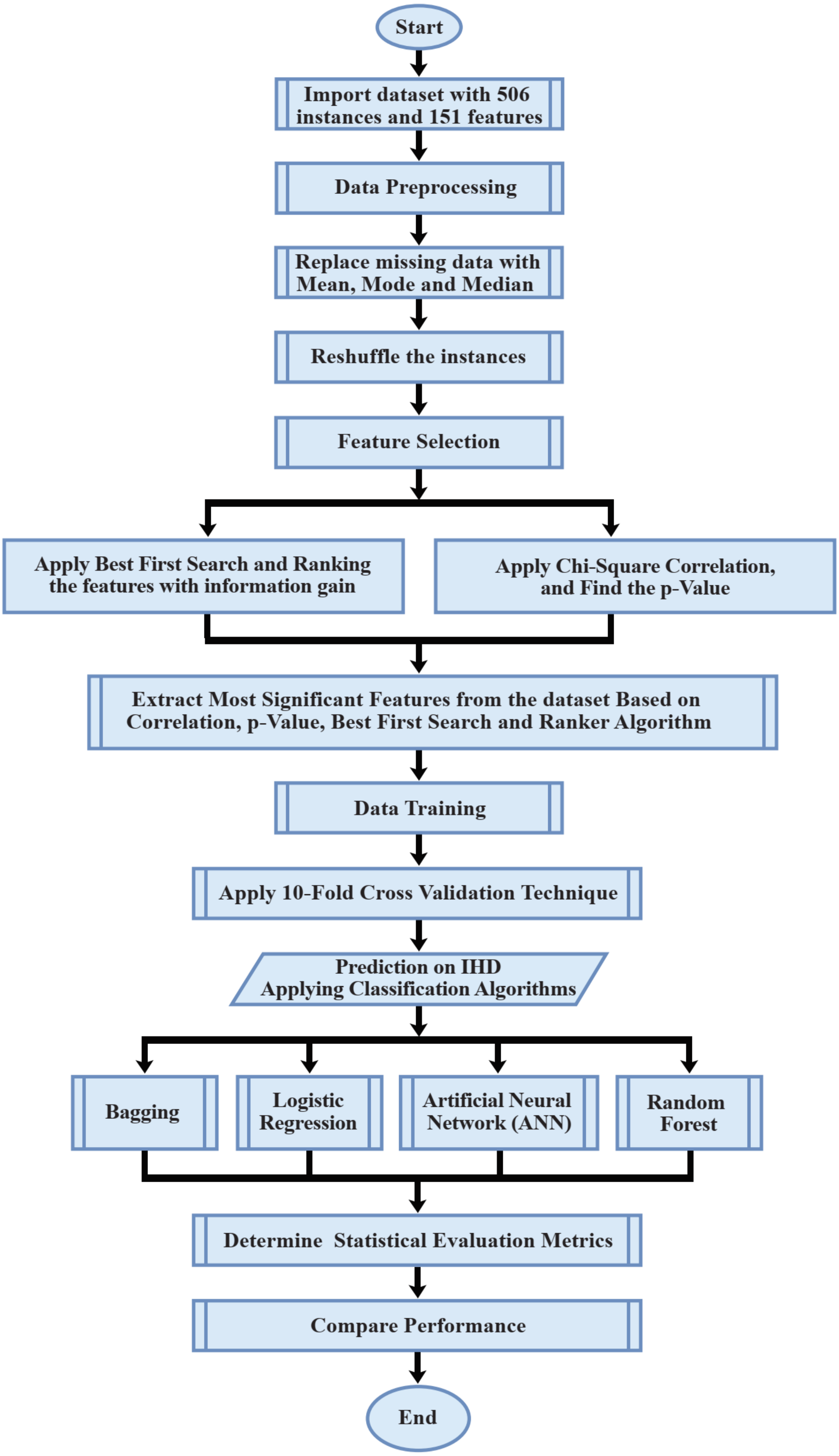

Figure 3.1: Working flow of the analysis

3.5.2.1 InfoGainAttributeEval (Information Gain Attribute Evaluation)

With respect to the class InfoGainAttributeEval measures, the information gain to evaluates the worth of an attribute. InfoGain(Class,Attribute) = H(Class) - H(Class | Attribute) [161]. It helps to determine the most correlated features in my dataset.

3.5.3 Chi-square Correlation

A chi-square test is a statistic that defines the relationships between variables. It uses for testing the dependency of features in a dataset [162]. This technique has been used by using IBM SPSS software to find out the dependent and independent attributes of the dataset. To perform this operation, the following steps will be followed:

1. Create crosstab and then perform a chi-square correlation. Options to click to perform this are, **Analyze > Descriptive Statistics > Crosstabs**.
2. A window will pop up with two options: Row(s) and Column(s).
3. Then the statistics window has opened with 15 different checkmark options.
4. Cells window also have some checkmark options. Among the three options namely observed, expected and unstandardized can be selected (optional) for the Chi-square test of independence.
    a. **Observed**: It is enabled by default and which defines the number of observations of a defined cell.
    b. **Expected**: Expected number of observations of the cell.
    c. **Unstandardized Residuals**: The executed "residual" value as observed minus expected.

3.5.3.1 Finding P-Value

IBM SPSS will be used to find out the P-Value. It defines the probability or chance of rejecting a null hypothesis if it is true. SPSS statistical software has packages that can make the calculation of any parameter. It has two hypotheses. One is a null hypothesis that can be composite or simple and another one is the alternate hypothesis. The significant level will be defined as 5%.

3.6 Selection Process of Most Significant Risk factors based on Different Feature Selection Techniques

The outcome of the BFS has been compared with Ranker and Chi-square correlation. Some features have been discarded if it is present in BFS but not highly ranked or correlated. Similarly, some features have been added if it is highly correlated and ranked based on information gain but it is not present in BFS. Here, the significance of each feature has been checked with the rest other two feature selection techniques. This way, the most significant features can be identified.

3.7 Dataset Training and Testing

For training the preprocessed dataset, Percentage Split and Cross-Validation methods can be used. Percentage split is a way to split a dataset into two-part namely training data and testing data respectively. It randomly re-sample the original data in n%. For example, 75% of the total rows of the dataset can be selected a training data and the rest 25% can be selected as test data [163].

3.7.1 Cross-validation

Checking the stability of a model is always important for a data scientist as it is always important to check the validity of models. Cross-validation is such a validation method to measure how the outcome of a statistical model will simplify to an independent dataset. It is also known as out of sample testing or rotation estimation. It is a deliberate method for doing rehashed holdout that really enhances it by diminishing the fluctuation of the gauge. It basically used in predictive models to know how accurately the model is predicting the outcome classes. It helps to get the accuracy of a system and to know which model can perform better. It helps to avoid overfitting and underfitting as well [164].

3.7.2 10-fold cross-validation

10-fold cross-validation is a validation technique that came from K-Fold Cross-validation where K is a simple parameter where K has a specific value. For example, for this study, the value of K has been set to 10 which means, K=10 and that is known as 10-Fold Cross-Validation. In this way, data has been divided into 10 parts. Different techniques on different datasets have proved that the 10-Fold Cross-validation is the best option to apply on the dataset to get estimate error and it also proves that the stratification improves outcomes slightly. So, it is enough to divide data into 10 parts. That is the reason to use K=10 or 10-fold cross-validation where almost all class attributes are illustrated in almost the right proportion. The procedure has been repeated 10 times and averages the results to achieve an optimal result. It is a standard way to repeat 10-fold cross-validation 10 times which involves the learning algorithm invoking 100 times in the dataset. So, 10-fold cross-validation has been used in this research study to get an evolution result and estimate the error. Finally one more time the classification has done to get an actual classifier [165].

3.8 Prediction of IHD using Classification Algorithms

Classification Algorithms that are used in ML and their categories are [166]:

I. Logistic Regression (LR)
II. Random Forest
III. Artificial Neural Networks
IV. Bagging

3.9 Bagging

Bagging performs as a technique for expanding accuracy. It is connected to learning plans for a numeric forecast. Assume that, patient suppose you might like to have a diagnosis made based on your indications. Rather than asking a specialist, you may ask several. If a certain diagnosis of analysis happens more than some other, you may pick this as the final or best diagnosis. The last finding diagnosis is depending on a majority vote, where each specialist gets an equivalent vote. Now by a classifier of each specialist, and you have the fundamental thought behind bagging. Naturally, a larger part vote made a substantial category of specialists might be more dependable than a greater part vote made a little category [167].
Theoretically, bagging can be characterized as follows:

(I) Build a bootstrap sample $M \times i = (Y \times i, X \times i)$ $(i = 1, \ldots, n)$ according to the empirical distribution of the pairs $Mi = (Yi, Xi)$ $(i = 1, \ldots, n)$.

(II) Compute the bootstrapped predictor $\hat{\theta} \times n\,(x)$ by the plug-in principle; i.e., $\hat{\theta} \times n\,(y) = hn(M \times 1, \ldots, M \times n)$, where $\hat{\theta}n(y) = hn(M1, \ldots, Mn)$.

(III) The bagged predictor is $\hat{\theta}_n; B(y) = E \times [\hat{\theta} \times n\,(y)]$.

3.9.1 Implementation of Bagging Algorithm in WEKA

Bagging as a classifier can erase variance. Bagging is also worked for regression. In this research, Naive Bayes as the base learner has used and WEKA software tool because of a variety of learning algorithms and implement the algorithm in this software tool. Because it is a modern technique to verify accuracy. The button has been chosen and selects “Bagging” under the “meta” group [168]. The seed value 1 is used and a number of iterations are 10. The number of execution slots 1 is used for ensemble construction. The batch size is 100.

Ultimately, the quality of the bags can be indicated in the **numIterations** parameter. The default value is 10, despite the fact that it is entirely expected to utilize esteems in the hundreds or thousands. Continue increasing the value of **numIterations** until no longer see an improvement in the model, or run out of memory. Bagging stands for bootstrap aggregation. The property of bagging value is shown in Table 3.1:

Table 3.1: Properties and Values of Bagging algorithm

Property	Value
bigSizePercent	100
batchSize	100
catOutOfBag	False
Classifier	Naive Bayes
Debug	False
Do Not Check Capabilities	False
numDecimalPlaces	2
numExecutionSlots	1
numIterations	10
outputOutOfBagComplexityStatistics	False
printClassifiers	False
representCopiesUsingWeights	False
Seed	1
storeOutOfBagPredictions	False

3.9.2 Advantages of Bagging

The advantage of bagging that gives the ensemble learning and their multiple fragile learners outperform to one powerful learner [169]. Bagging decrease variance and avoid over-fitting [170]. It can be worked as a regression including also classification. It refines the stability including the accuracy of ML algorithms used in statistical regression and also classification. The advantage of bagging does not concentrate on any specific occasion of the training information – Therefore, over-fitting is less susceptible when applied to noisy data. Bagging is robust to noise and outliers.

3.9.3 Disadvantages of Bagging

The disadvantage of the bagging is losing the interpretability of the model. There can be an issue of big bias if not modeled correctly. Another is that when bagging gives better accuracy, it is costly in terms of computation and not be desirable depending on the use case. It can gently reduce the performance of stable formulas, for example, K-nearest neighbors.

3.10 Logistic Regression

Logistic Regression is one of the popular algorithms among ML algorithms which uses for binary classification. It is a statistical strategy of investing a dataset in which there is more than one independent variable that determines an outcome. The result is estimated with a dichotomous variable. It only includes information coded as 1 and 0 [171]. Logistic regression is named for the task used at the core of the technique, the logistic function. Logistic Regression could help to use to identify whether the student passed or failed. The identifications are discrete. We can also see probability scores underlying the model's classifications.

The basic formula for the postulated linear model is:

$$g(P(x)) = \alpha + \beta y1 + \gamma y2$$

In the formula,

Link function = g()

The expected value of the target variable = E(y)

α, β and γ = Values which are to be predicted

3.10.1 Implementation of Logistic Regression Algorithm in WEKA

LR accepts the input variables are numeric and have a Gaussian distribution. WEKA implementation has been adapted to support multi-class classification problems. But after choosing the selection of Logistic under the functions group the algorithm can run for a fixed number of iterations, but by default will run until it is estimated that the algorithm has converged. The algorithm for Logistic Regression (LR) classifier works with the class which follows the multinomial logistic regression method along with the ridge estimator. Binomial logistic regression was used in this research study. The batch size of LR was **100**, the value of the ridge was **1.0E-12.** The parameter matrix which has to be classified using the LR algorithm can be expressed as (1).

$B = m \times (k-1)$ ………………………….. (1)

Where B represents the parameter matrix, k is the number of classes, n denotes the number of instances, and m stands for the number of attributes. The possibility of any class j except the last one to be classified with the LR algorithm can be expressed using (2) while the last class has the possibility as stated in (3). Xi represents the fraction of cells in j for a patient i.

$P(y_i) = \frac{e^{\sum_{j=1}^{k-1} y_{iB_j}}}{1+e^{\sum_{j=1}^{k-1} y_{iB_j}}}$ ………… (2)

$P'(Y_i) = \frac{1}{1+e^{\sum_{j=1}^{k-1} y_{iB_j}}}$ ……….. (3)

$M = -\sum_{i=1}^{n}\left[\sum_{j}^{k-1}\left(Y_{ij} + In\left(P_j(Y_i)\right)\right) + \left(1 - \sum_{j=1}^{k-1} P_j(Y_i)\right)\right] + ridge \times B^2$ ……….. (4)

The log-likelihood of negative multinomial is expressed in (4). In general, L is minimized as much as possible so that B can be determined. We have used a slightly modified LR algorithm so that the instances can be classified using some weighted values. Moreover, the missing values have been replaced using a customized filter [172]. The property of bagging value is given in Table 3.2:

Table 3.2: Properties and Values of Logistic Regression algorithm

Property	Value
batchSize	100
Debug	False
Do Not Check Capabilities	False
Maxlts	-1
numDecimalPlaces	3
Ridge	1.E-12
UseConjugateGradientDescent	False

3.10.2 Advantages of Logistic Regression

It is a generally used procedure since it is extremely proficient, does not require an excessive number of computational resources, it's outstandingly interpretable, it doesn't require information highlights to be scaled, it doesn't require any tuning, it's anything but difficult to regularize, and it yields well-adjusted anticipated probabilities. Strategic relapse improves when you evacuate ascribes that are inconsequential to the yield variable just as characteristics that are so like one another. Another bit of leeway of calculated relapse is that it is unbelievably simple to actualize and productive to prepare. Due to its straightforwardness and the way that it tends to be actualized moderately simple and fast, Logistic Regression is likewise a decent pattern that you can use to quantify the exhibition of other progressively complex Algorithms. The Logistic Regression calculation is a straightforward relapse calculation that can outline an N-dimensional sign to a 1-dimensional sign.

3.10.3 Disadvantages of Logistic Regression

A drawback of logistic regression that can't take care of non-direct issues with calculated relapse since it's choice limit is straight. It is that the elucidation is progressively troublesome in light of the loads is multiplicative and not added substance. Strategic relapse can experience the ill effects of complete partition. On the off chance that there is an element that would impeccably isolate the two classes, the Logistic Regression model can never again be prepared. This is on the grounds that the weight for that element would not join, on the grounds that the ideal weight would be vast.

3.11 Artificial Neural Network (ANN)

Neural networks are a group of calculations that displayed freely after the human mind, that are intended to perceive designs. They translate the tactile of information through a sort of machine observation. The examples they perceive are numerically contained in vectors, into which all certifiable information. ANN is a computational model relies upon the structure and elements of natural neural systems. Data that courses through the system influence the structure of the ANN in light of the fact that it changes - or learns, in light of that information and yield. ANN a computational nonlinear model that is generally used in Machine Learning and is considered to be a prominent component of futuristic Artificial Intelligence [173]. The backpropagation algorithm is a part of ANN where inputs and desired outputs are known. The output received from the algorithm compares with the targeted output and if the targeted output achieved then well otherwise the system sends feedback with the error to evaluate the desired result. By using the Activation Function, it is possible to convert an input signal of a node in n ANN to an output signal. And, in the next level, the output will be used as input. So, the Activation function basically enables the input signal if the archived output is not equal to the desired output.

3.11.1 Implementation of Artificial Neural Network (ANN) Algorithm in WEKA

ANN classifier algorithm works based on feed-forward and back-propagation neural networks. At first, the information has been transmitted from the input layer to the output through several intermediate layers called the hidden layers and interlayer information transfer involves several weighted paths whose weights have been predefined randomly at the beginning. The properties for ANN has shown in Table 3.3. The weights have been updated to optimum values by applying the back-propagation neural network on the basis of the sigmoid function. These property values for batch size, debug, and checking capability have been kept unchanged. The decaying property has been disabled in order to prevent a decrease in learning rate which has been kept constant as 0.1 with momentum 0.2 and the training period has been limited to 500 epochs. Normal to binary filter has been implemented with the normalizing feature enabled with a view to normalizing all attributes in the dataset including the numeric and nominal ones. Multilayer perception has been used for the ANN classifier, the number of hidden layers is more than one and we have two hidden layers with 16 and 14 nodes.

Table 3.3: Properties and Values of ANN algorithm

Property	Value
GUI	False
autoBuild	True
batchSize	100
Debug	False
Decay	False
Do Not Check Capabilities	False
hiddenLayer	16.14
learningRate	0.1
Momentum	0.2
nominalToBinaryFilter	True
normalizeAttributes	True
normalizeNumericClass	True
numDecimalPlaces	2
Reset	True
Seed	0
trainingTime	500
validationSetSize	0
ValidationTheshold	20

3.11.2 Advantages of Artificial Neural Network (ANN)

ANNs have the ability to learn and display non-linear and complex connections, which is extremely significant in real life, a huge number of connections between inputs and outputs are non-linear also complex. ANNs can generalize—after gaining the initial inputs and their connections, it can deduce unseen connections on unseen information as well, thus making the model generalizes and predict on unseen information. Artificial neural networks learn occasions and settle on choices by remarking on comparative occasions and have the ability to make ML. ANNs have the numerical quality that can perform more than one work in the meantime. Information for example in traditional programming is stored on the whole system, not on a database [174].

3.11.3 Disadvantages of Artificial Neural Network (ANN)

As per their structure, ANN requires processors with parallel handling power. Hence, the realization of the equipment of hardware is dependent. This is the most important issue of ANN which is the unexplained behavior of the network. At the time of ANN produces a probing solution, it does not give a clue as to why and how. This decreases faith in the network. There is no particular guideline for determining the structure of artificial neural networks. It can work with numerical information. But the problems have to be translated into numerical values before being introduced to ANN. The display mechanism to be determined here will directly influence the performance of the network. The presentation component to be resolved here will straightforwardly impact the exhibition of the system. This relies upon the client's capacity. The system is decreased to a specific estimation of the blunder on the example implies that the planning has been finished. This regard does not give us perfect results.

3.12 Random Forest

Random Forest is easy to use the ML that produces, even without hyper-parameter tuning and a great result. It is also used because of its simplicity and the fact that it can be used for both classification and regression tasks. Random Forest or Random decision forests correct for decision trees nature of overfitting to their training set. Random Forest algorithm is a supervised classification algorithm and flexible. There is an immediate connection between the number of trees in the timberland and the outcomes it can get the bigger the number of trees is given the more exact outcome [175].

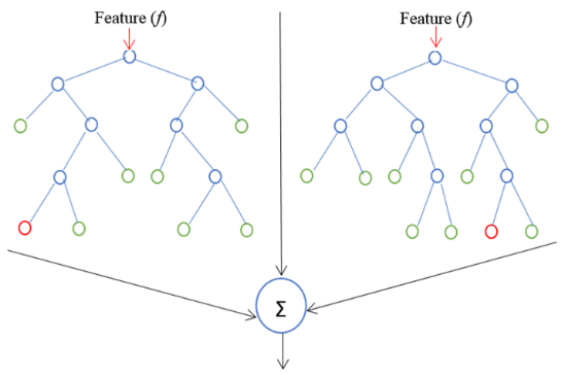

Figure 3.2: Decision Trees to merge and to get an ideal outcome

So, it works based on several generated decision trees. It merges the decision trees to get an optimal result. Figure 3.2 illustrates the process of generating decision trees and evaluating the outcome by merging the trees together.

3.12.1 Implementation of Random Forest Algorithm in WEKA

Random Forest (RF) of the ensemble classifier pursues the decision tree model of rules for classification analysis. In the season of preparing, it is conducted by building a multitude of decision trees. The properties of random forest has shown in Table 3.4. The resultant is the mode of the classes for grouping. Random Forest is very simple to utilize approach. The batch measure was 100 and the number of implementation slots was 1 and iterations were 100.

Table 3.4: Properties and Values of RF algorithm

Property	**Value**
bagSizePercent	100
batchSize	100
breeakztiesRandomly	False
calcOutOfBag	False
computeAttributeImportance	False
Debug	False
doNotCheckCapabilities	False
maxDepth	0
numDecimalPlaces	2
numExecutionSlots	1
numFeatures	0
numIterations	100
outputOutOBagComplexityStatistics	False
prinClassifiers	False
Seed	1

3.12.2 Advantages of Random Forest

The advantage of prototypes is processed that give data about the connection between the variables and the classification. It registers proximities of Random Forest between pairs of cases that can be used in locating outliers, clustering, or give interesting views of the data [176]. The capabilities of the above can be extended to outlier detection, leading to unsupervised clustering, and unlabeled data. It offers a technique for identifying variable interactions. The advantage to using a Random Forest over a decision tree: It reduces over-fitting and additionally accurate.

3.12.3 Disadvantages of Random Forest

Random forests have been detected the team for some datasets with noisy classification tasks. For information and categorical variables with different levels, random forests are biased in favor of those attributes with more levels. Therefore, the variable significance scores from random forest are not dependent for this type of data. The major limitation of the Random Forests algorithm is may make slow-going for real-time prediction because of a large number of trees. It’s more complex and difficult to visualize the model or comprehend why it predicted something [177]. It’s so difficult to implement and computationally high-cost.

CHAPTER IV

RESULTS AND DISCUSSIONS

4.1 Data Distributions of Collected Dataset

The dataset contains a total 506 unique instances where each instance has maximum of 151 historic, symptomatic and pathologic features with one outcome class. A total of 6 demographic features such as Age, Sex, Profession, etc. have shown in appendix Table A1. In appendix Table A2 some diagnostic features such as Heart Rate. Systolic Blood Pressure, Diastolic Blood Pressure, Creatinine, Ejection Fraction, etc. have shown. Here we see that the average Heart Rate of the patients is 82.85. The patient's history and their symptoms have shown in appendix Table A3 to A4.

In this collected data, around 88 features have a low frequency (less than 20). Around 9 categories of a heart attack such as ACS, STEMI, NSTEMI, Unstable Angina, UMI, AMI, OMI, AWMI, IWMI have been collected which leads to IHD.

4.2 Outcome of Feature Selection

For determining significant risk factors to IHD we applied 3 approaches e.g. Best First Search algorithm, Ranker algorithm, and Chi-square Correlation. The results are described in the preceding subsections:

4.2.1 Outcome of Best First Search (BFS)

Table 4.1 shows the results of different attributes for BFS. In this feature selection technique, 22 features have been selected out of 152 features including IHD. In these 22 features we can see that one demographic features, two diagnostic features, and nineteen symptoms has been selected. BFS algorithms generated total 1058 subsets.

Table 4.1: Results for Best First Search

Selected Features			
Instances	**Features**	**Search Direction**	**Total number of subsets evaluated**
506	152	Bi-Directional	1058
sex	non-stemi	valvular-cardiomyopathy	stroke
bp-systolic	AMI	AWMI	chest-pain
ACS	PCI	PTCA	troponin
IWMI	rhenumated-arthitis	severe-musculoskeletal-pain	EF
PAD	DVD	bronchial-asthma	
severe-multiple-sclerosis	thrombolysed-stk	pleural-effusion	

4.2.2 Outcome of Ranker Algorithm

InfoGainAttributeEval is a class which have been used for the Ranker algorithm. It evaluated the significance of an attribute by calculating **Information Gain** of the attributes. By determining Information gain it helped to select the most co-related attributes. Table 4.2 represents the outcomes of various features for the Ranker algorithm. we can see that around 33 features are highly ranked at 1% information gain. The highest information gain value is 0.88417 and lowest information value is 0.01028 out of these 33 risk factors.

Table 4.2: Results for Ranker algorithm

Feature Name	Information Gain	Rank	Feature Name	Information Gain	Rank
ACS	0.8841705634	1	syncope	0.0088995882	35
EF	0.1024586208	2	atrial-fibrillation	0.0076178667	36
troponin	0.1003215585	3	DCM	0.0076178667	37
non-stemi	0.0786177118	4	CKD	0.007355043	38
PTCA	0.0689170536	5	PAH	0.0068467254	39
DVD	0.058163819	6	UTI	0.0068467254	40
profession	0.0507358281	7	valvular-cardiomyopathy	0.0068467254	41
chest-pain	0.0506771806	8	SVT	0.0068467254	42
AMI	0.0500164805	9	RTI	0.0068467254	43
CAD	0.0493060846	10	OMI	0.0062041019	44

Feature Name	Information Gain	Rank	Feature Name	Information Gain	Rank
SVD	0.0470669215	11	ICM	0.0062041019	45
sex	0.0467425202	12	ALVF	0.0049647108	46
CAG	0.0458161627	13	FVR	0.004887474	47
ECG	0.04277936	14	PPM	0.004887474	48
rhenumated-arthitis	0.034772016	15	tobaco	0.0048433689	49
sweating	0.0294097517	16	hypothyroidism	0.0034168067	50
Smoking	0.0287504477	17	PAD	0.0034168067	51
unstable-angina	0.0280243869	18	pericardial-effusion	0.0034168067	52
bp-systolic	0.0258334005	19	anaemic-heart-failure	0.0034168067	53
stroke	0.0247435082	20	mild-pr	0.0034168067	54
thrombolysed-stk	0.0221163671	21	electrolyte-imbalance	0.0034168067	55
IWMI	0.0221163671	22	urosepsis	0.0034168067	56
creatinine	0.0200910111	23	normal-coronasies	0.0034168067	57
PCI	0.0178153534	24	VHD	0.0034168067	58
AWMI	0.0156808306	25	hemorrhagic-shock	0.0034168067	59
vomiting	0.0142190784	26	aortic-stenosis	0.0034168067	60
bronchial-asthma	0.0137463524	27	RHD	0.0034168067	61
patients_status	0.0128909897	28	dissecting-aortic-aneurysm	0.0034168067	62
radiation	0.0128909897	29	metabolic-encephalopathy	0.0034168067	63
normal-epicardial-coronary-artery	0.012285016	30	moderate-ms	0.0034168067	64
cardiogenic-shock	0.0118361734	31	hypothyrodism	0.0034168067	65
palpitation-duration	0.010508016	32	severe-musculoskeletal-pain	0.0034168067	66
severe-multiple-sclerosis	0.0102898691	33	hyponatremia	0.0034168067	67
palpitation	0.0095547347	34	pleural-effusion	0.0034168067	68

4.2.3 Outcome of Chi-square Correlation

The features also have been selected based on the correlation between features in my dataset. Some of the features have been found are highly correlated and some are not on a 5% significant level **($p = 0.05$)**. The highly correlated attributes have been dropped from the dataset. Table 4.3 shows the results for Correlation. Around 43 features have been accepted at 5% significant level.

Table 4.3: Results of Correlation algorithm

Sl. No.	Features Name	P-Value	Sl. No.	Features Name	P-Value
1	SVD	0.0001	23	bpsystolic	0.003
2	sex	0.0001	24	palpitationduration	0.005
3	ACS	0.0001	25	PCI	0.006
4	unstableangina	0.0001	26	radiation	0.006
5	nonstemi	0.0001	27	patients_status	0.006
6	AMI	0.0001	28	palpitation	0.007
7	rhenumatedarthitis	0.0001	29	severemultiplesclerosis	0.008
8	DVD	0.0001	30	syncope	0.009
9	CAD	0.0001	31	cardiogenicshock	0.01
10	PTCA	0.0001	32	AWMI	0.01
11	smoking	0.0001	33	DCM	0.015
12	stroke	0.0001	34	atrialfibrillation	0.015
13	sweating	0.0001	35	CKD	0.016
14	chestpain	0.0001	36	SVT	0.031
15	EF	0.0001	37	RTI	0.031
16	ECG	0.0001	38	valvularcardiomyopathy	0.031
17	CAG	0.0001	39	PAH	0.031
18	IWMI	0.002	40	UTI	0.031
19	normalepicardialcoronary artery	0.002	41	creatinine	0.031
20	thrombolysedstk	0.002	42	FVR	0.05
21	bronchialasthma	0.002	43	PPM	0.05
22	vomiting	0.002			

4.2.4 Extracted Mostly Significant Features from the Dataset

In BFS algorithms, a total of 22 significant features has been selected. But out of these 22 features only 18 features have chosen as most significant based on other two feature selection techniques. As the PAD features have not highly ranked (rank: 51, information gain value: 0.00341) and correlated (p-value: 0.128), this feature is discarded for further analysis. Similarly, severe-musculoskeletal-pain (rank: 66, information gain: 0.00341, p-value: 0.128), troponin (rank:3, information gain: 0.10032, p-value: 0.646) and pleural-effusion (rank: 68, information gain: 0.00341, p-value: 0.128) has been discarded from the outcome of BFS. The troponin feature has ranked high based on information gain but it can be discarded if we compare with the outcome of rest two feature selection techniques. Rest 10 features out of 28 have chosen based on high information gain and p-value. Those features which have lower information gain less than 1%, can be discarded easily. For this reason, in the ranker technique, only 33 features have been selected based on high information gain.

The outcomes of BFS, Ranker and Correlation algorithms have been merged to get the most significant features and after merging the algorithms 28 significant features have been found. They are sex, bp-systolic, ACS, IWMI, unstable-angina, severe-multiple-sclerosis, non-stemi, AMI, PCI, rhenumated-arthitis, DVD, thrombolysed-stk, valvular-cardiomyopathy, CAD, SVD, AWMI, PTCA, bronchial-asthma, smoking, stroke, sweating, vomiting, chest-pain, EF, ECG, CAG, patients_status. All the selected significant features have been shown in Table 4.4.

Table 4.4: Significant Features List

Serial No.	Features Name	Type	BFS	Ranker	P-Value
1	ACS	Symptoms	☑	0.88417	0.0001
2	EF	Diagnostic	☑	0.10245	0.0001
3	Non-stemi	Symptoms	☑	0.07861	0.0001
4	PTCA	Symptoms	☑	0.06891	0.0001
5	DVD	Symptoms	☑	0.05816	0.0001
6	Chest-pain	Symptoms	☑	0.05067	0.0001
7	AMI	Symptoms	☑	0.05001	0.0001
8	CAD	Symptoms	☒	0.04930	0.0001
9	SVD	Symptoms	☒	0.04706	0.0001
10	Sex	Demographic	☑	0.04674	0.0001
11	CAG	Symptoms	☒	0.04581	0.0001
12	ECG	Symptoms	☒	0.04277	0.0001
13	Rhenumated-arthitis	Symptoms	☑	0.03477	0.0001
14	Sweating	Symptoms	☒	0.02940	0.0001

Serial No.	Features Name	Type	BFS	Ranker	P-Value
15	Smoking	Symptoms	☒	0.02875	0.0001
16	Unstable-angina	Symptoms	☒	0.02802	0.0001
17	Bp-systolic	Diagnostic	☑	0.02583	0.003
18	Stroke	Symptoms	☑	0.02474	0.0001
19	IWMI	Symptoms	☑	0.02211	0.002
20	Thrombolysed-stk	Symptoms	☑	0.02211	0.002
21	Creatinine	Diagnostic	☒	0.02009	0.031
22	PCI	Symptoms	☑	0.01781	0.006
23	AWMI	Symptoms	☑	0.01568	0.01
24	Vomiting	Symptoms	☒	0.01421	0.002
25	Bronchial-asthma	Symptoms	☑	0.01374	0.002
26	Patients_status	Symptoms	☒	0.01289	0.006
27	Servere-multiple-sclerosis	Symptoms	☑	0.01028	0.008
28	Valvular-cardiomyopathy	Symptoms	☑	0.00684	0.031

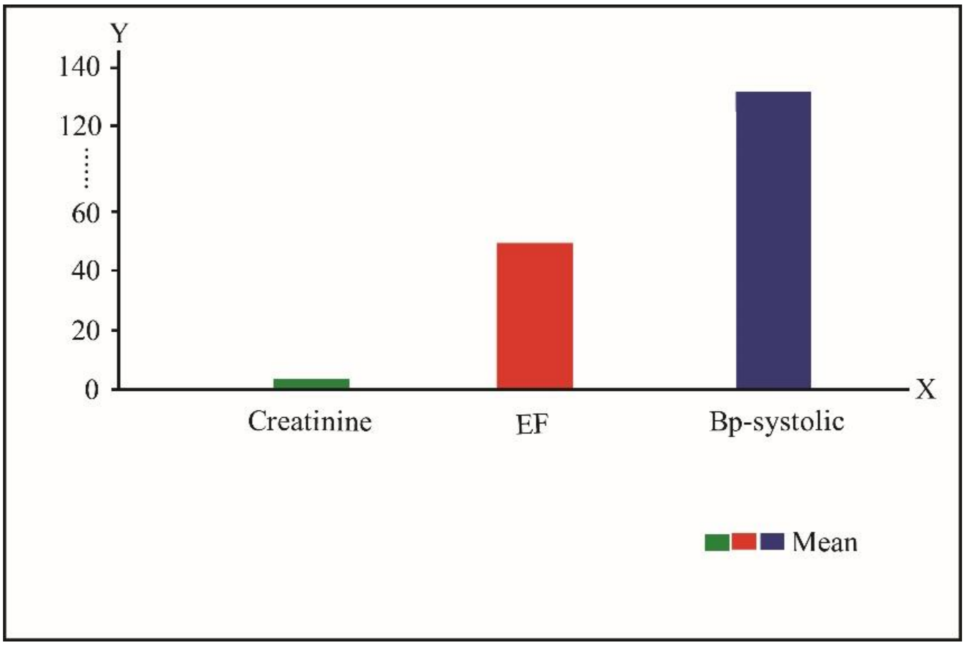

Figure. 4.1: Graphical view of Mean

Figure. 4.1 reveals the graphical view of the value of Mean for different attributes in the bar graph. The values are represented in Y-axis and the various features are in X-axis. The mean of creatinine is 1.718 and for EF 49.709. Mean of systolic blood pressure is 130.7.

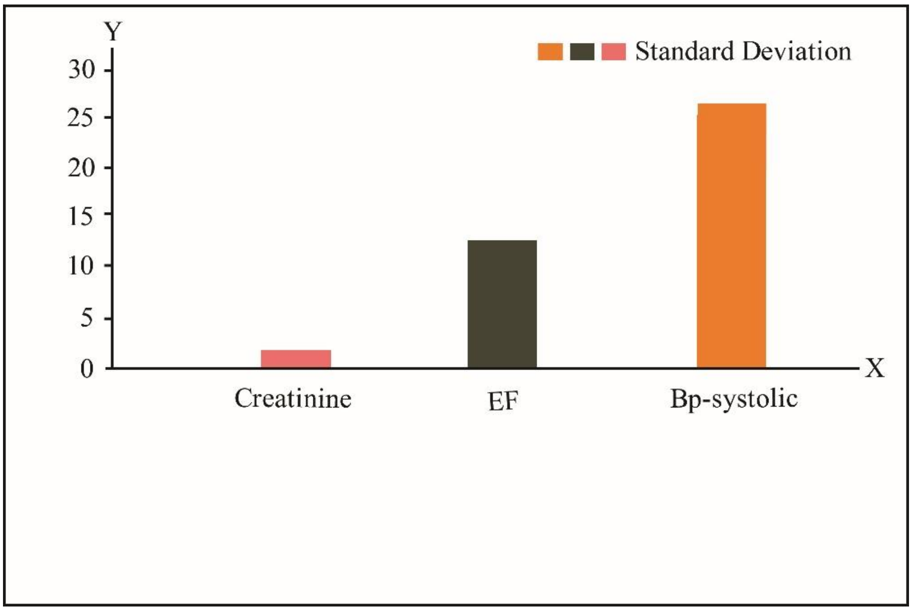

Figure. 4.2: Graphical view of Standard Deviation

Figure. 4.2 shows the graphical view of the value of Standard Deviation for different attributes in the bar graph. The values are represented in Y-axis and the various features are in X-axis. The standard deviation of creatinine is 1.651, for EF 12.278 and 27.754 is for systolic blood pressure.

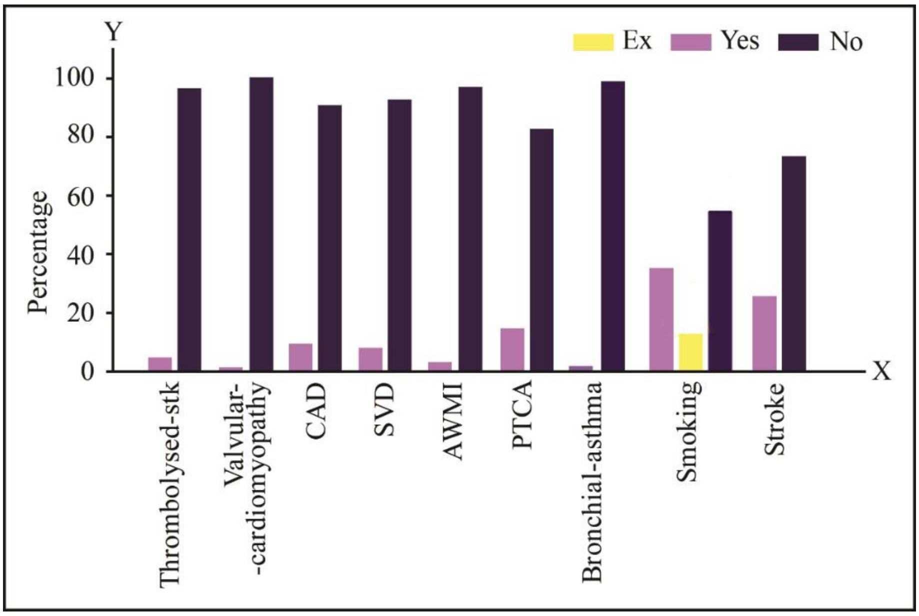

Figure. 4.3: Graphical view of some features and percentage

Figure. 4.3 shows the graphical view of several attributes and their percentage in the bar graph. The percentages are represented in Y-axis and the various attributes are in X-axis.

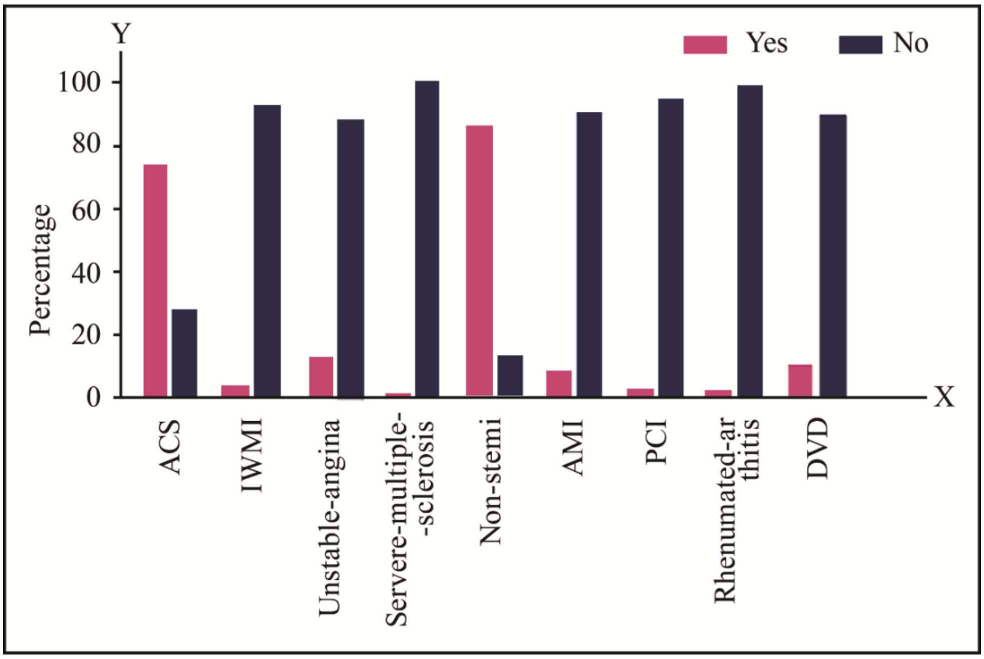

Figure. 4.4: Graphical view of some features and percentage

Figure. 4.4 shows the graphical view of several attributes and their percentage in the bar graph. The percentages are represented in Y-axis and the various attributes are in X-axis.

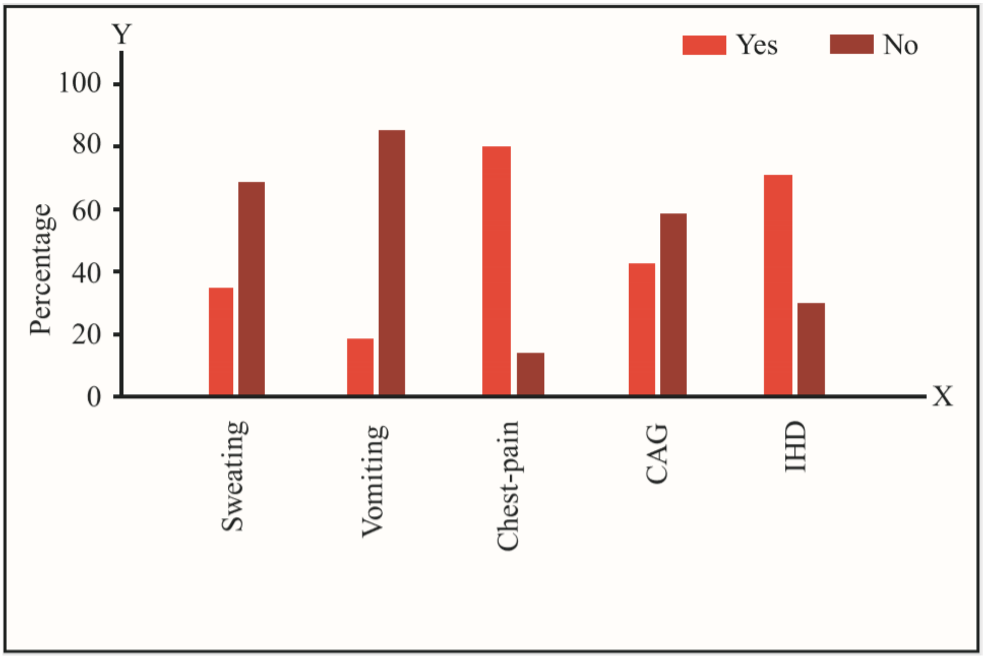

Figure. 4.5: Graphical view of some features and percentage

Figure. 4.5 shows the graphical view of several attributes and their percentage in bar graph. The percentages are represented in Y axis and the various attributes are in X axis.

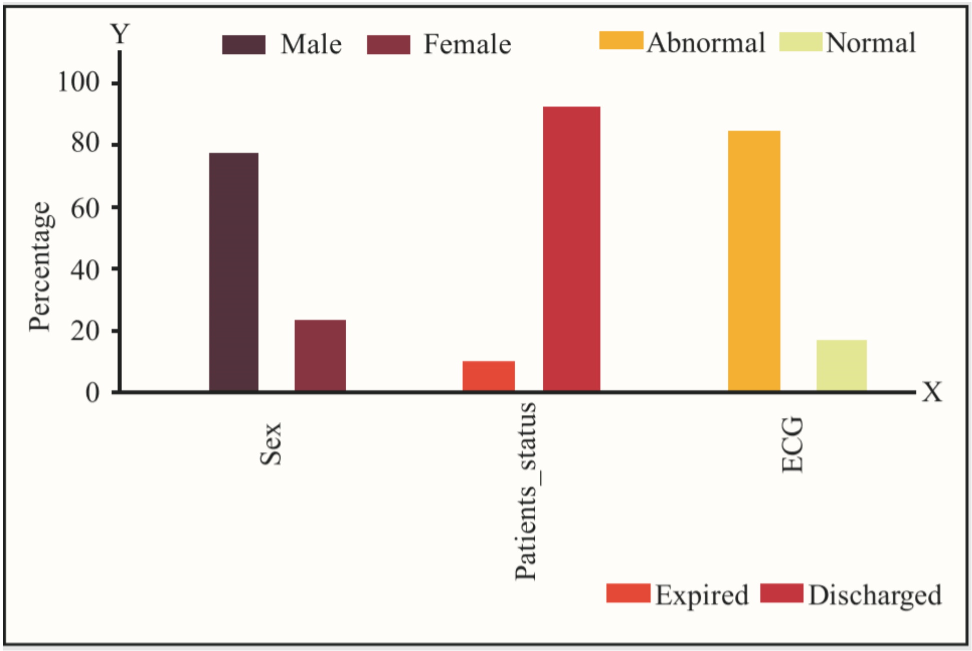

Figure. 4.6: Graphical view of some features and percentage

Figure. 4.6 shows the graphical view of several attributes and their percentage in the bar graph. The percentages are represented in Y-axis and the various attributes are in X-axis.

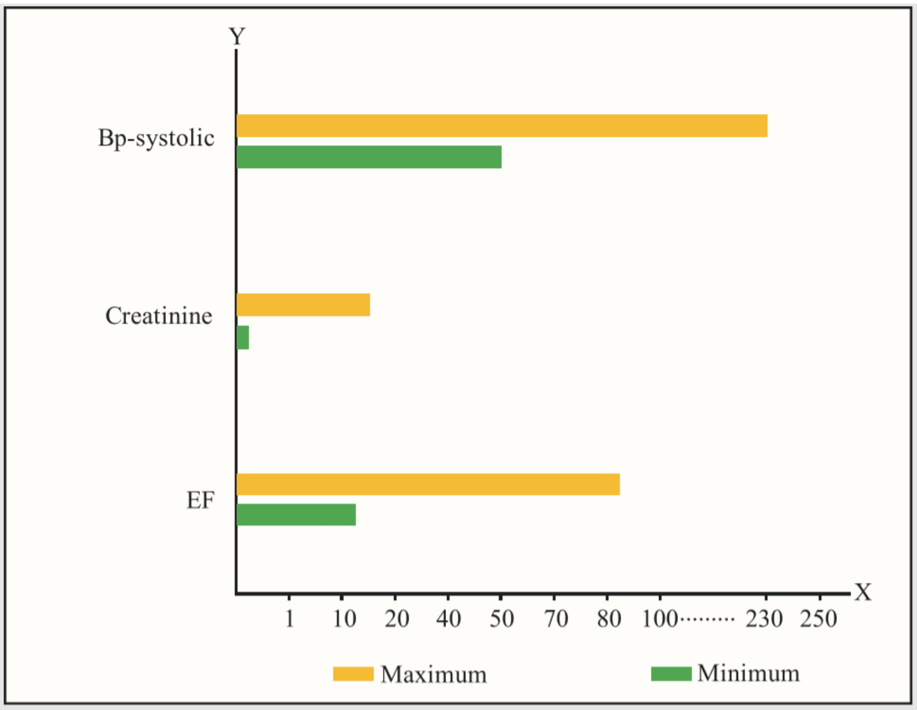

Figure. 4.7: Graphical view of some features and percentage

Figure. 4.7 shows the graphical view of several attributes and their maximum and minimum value in bar graph. The values are represented in Y axis and the various attributes are in X axis.

4.3 Prediction on Ischemic Heart disease

Four ML algorithms have been used for classification. They are: Bagging, Logistic Regression, Artificial Neural Network and Random forest.

4.3.1 Outcome of Bagging Algorithm

The Figure. 4.8 shows the graphical view of the outcomes of the Bagging algorithm where the percentages are represented in the Y-axis and the various factors are on the X-axis. Naive Bayes was used as a base learner. The accuracy or Correctly classified instance for Bagging has got is 96.4427% when the value of Kappa statistics is 0.9154. Mean absolute error and root mean squared error are 0.0509 and 0.1725 respectively. In addition, the value of TP and FP are 0.964 and 0.052 and the Precision, Recall, F-Measure is same at 0.964. And, the value of MCC, ROC Area and PRC Area are respectively 0.915, 0.986 and 0.987.

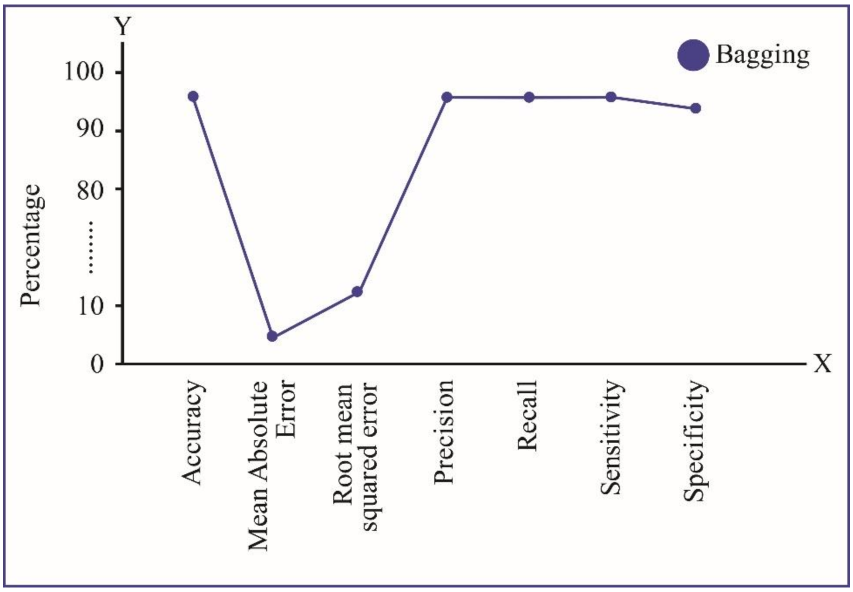

Figure. 4.8: Graphical view of Bagging

4.3.2 Outcome of Logistic Regression Algorithm

The Figure. 4.9 shows the graphical view of the outcomes of the Logistic Regression algorithm. The percentages are represented in Y-axis and the various factors are in X-axis.

Logistic Regression (LR) is the second algorithm in the study that has been used as a classification algorithm. The accuracy of LR has achieved is 96.64%. Kappa statistics, mean absolute error and root mean squared error are respectively 0.9196, 0.0426 and 0.1731. Furthermore, the values of TP, FP are 0.966, 0.055 and Precision, Recall are the same 0.966. F-Measure, MCC are 0.966 and 0.920 when ROC and PRC Area are 0.964 and 0.966.

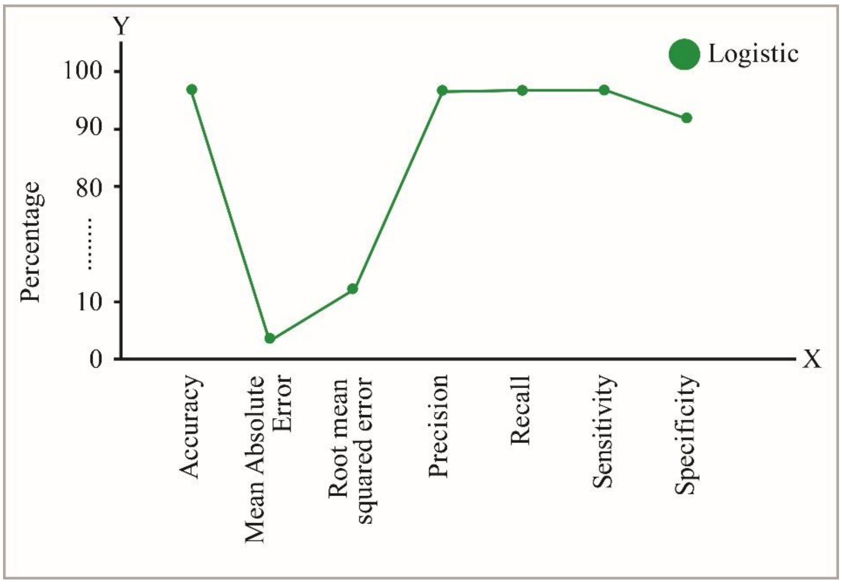

Figure. 4.9: Graphical view of Logistic Regression

4.3.3 Outcome of the Random Forest Algorithm

The Figure. 4.10 shows the graphical view of the outcomes of the Random Forest algorithm. The percentages are represented in the Y axis and the various factors are in the X axis. Random Forest (RF) is another algorithm that has been used as a classification algorithm. The accuracy of RF has achieved is 98.6285%. Kappa statistics, Mean Absolute Error and Root Mean Squared Error are respectively 0.9425, 0.076 and 0.1566. Furthermore, the values of TP, FP are 0.976, 0.055 and Precision is 0.977. F-Measure, MCC are 0.976 and 0.944 when ROC and PRC Area are the same at 0.98.

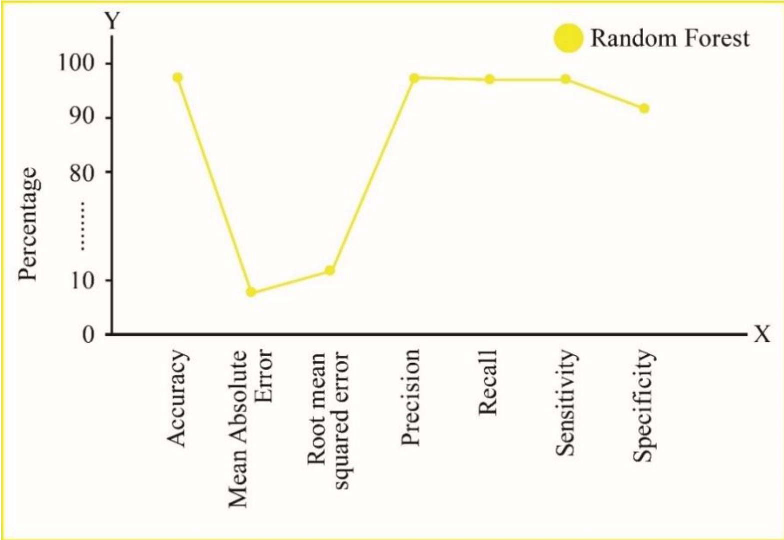

Figure. 4.10: Graphical view of Random Forest

4.3.4 Outcome of Artificial Neural Network (ANN)

The Figure. 4.11 shows the results of ANN in a graphical view. It shows the outcomes of different factors. ANN is another algorithm has been used as a classification algorithm in the research. The accuracy has received for ANN is 95.8498% and the Kappa statistics is 0.9018. Mean absolute error and root mean squared error are 0.0404 and 0.1812 when the TP and FP are respectively 0.958 and 0.055. Precision, Recall, and F-Measure is 0.96 when the MCC, ROC Area, and PRC Area are 0.902, 0.986 and 0.989.

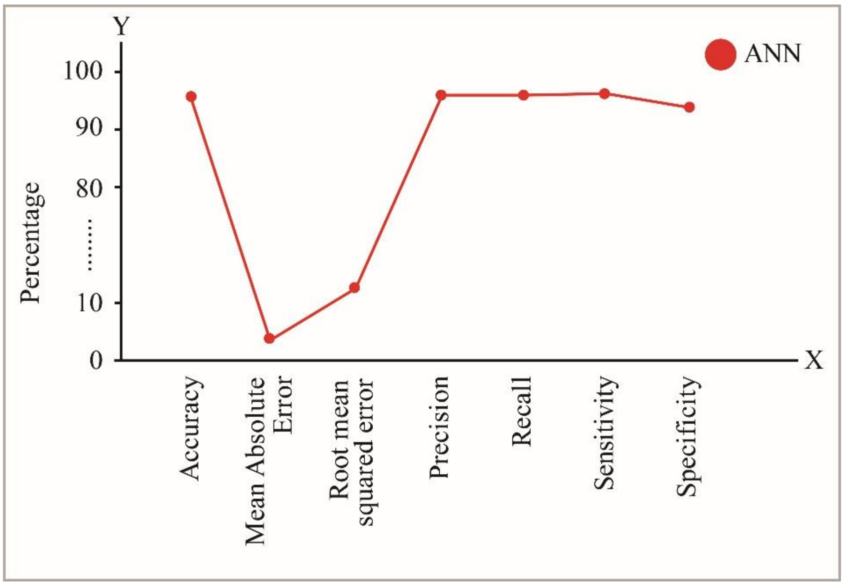

Figure. 4.11: Graphical view of ANN

4.4 Comparison of the Performance of Different Machine Learning Algorithms

Comparison between the classification algorithms have been shown in Table 4.9 where the highest accuracy 97.63% have got in RF and the accuracy of Bagging and LR are about to same as 96.44% and 96.64% respectively when the accuracy for ANN is the lowest accuracy among the algorithms for the dataset have been used in the study.

The values of Kappa statistics also differ from one another for the algorithms. The value of Kappa statistics for ANN and LR about to same are 0.918 and 0.954 and for Bagging and RF are 0. 9154 and 9425. It is seen that the best algorithm for this study is Random Forest and it has an accuracy of 97.6285% which is better than all other algorithms applied to the collected dataset. As it is observed from this table, when the accuracy increases the kappa statistic also increases whereas the root means squared error and the root relative squared error decrease. The sensitivity of ANN is lower comparatively Random Forest where the specificity of ANN is higher than Random Forest. As Random Forest is an ensemble learning technique it produces many models where ANN one. Depending on different learning rates, the number of hidden layers the accuracy of ANN can be varied comparatively Random Forest. But sometimes it can lead to overfitting and underfitting problems.

Table 4.5: Comparison between algorithms

Evaluation Metrics	**Algorithms Names**			
	ANN	**Bagging**	**LR**	**RF**
Accuracy (%)	95.85	96.44	96.64	**97.63**
Incorrectly Classified Instances (%)	4.15	3.55	3.36	**2.37**
Kappa Statistic	0.9018	0.9154	0.9196	**0.9425**
Mean Absolute Error	0.040	0.050	0.042	0.076
Root Mean Squared Error	0.181	0.172	0.173	**0.156**
Root Relative Squared Error (%)	39.46	37.55	37.67	**34.09**
True Negative Rate/ Specificity (%)	93.464	93.464	92.81	92.156
Precision (Weighted Avg.)	0.959	0.964	0.966	**0.977**
Recall/Sensitivity (Weighted Avg.) (%)	95.8	96.4	96.6	**97.6**
F-Measure (Weighted Avg.)	0.959	0.964	0.966	0.976
Matthews Correlation Coefficient (MCC) (Weighted Avg.)	0.902	0.915	0.920	**0.944**
Receiver Operating Characteristics (ROC) Area (Weighted Avg.)	0.986	0.986	0.964	0.986
Precision Recall Curve (PRC) Area (Weighted Avg.)	0.989	0.987	0.966	0.988

Figure. 4.12 represents the graphical view of the algorithms used in this study. It shows some important factors like accuracy, mean absolute error, root means absolute error, precision, recall, specificity, and sensitivity.

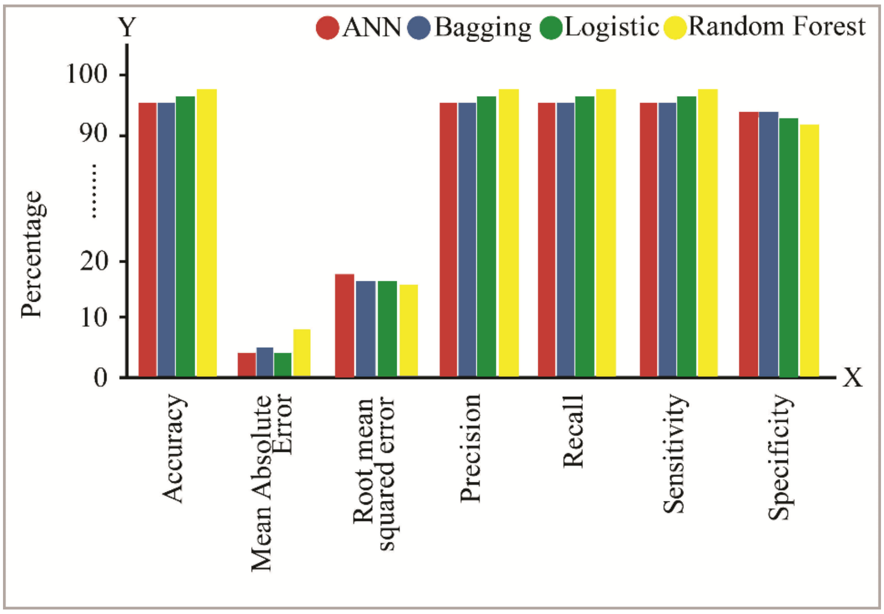

Figure. 4.12: Comparison of Algorithms

The same analysis has been performed for the four-classification algorithm using another heart disease dataset named Cleveland hospital dataset. The dataset has been collected from the UCI Machine Learning Repository. This dataset has 303 instances with 13 features [18], [22] [25]. The results of this analysis have been displayed in Table 4.6 to Table 4.9.

The Artificial Neural Network Algorithm has been applied to the Cleveland dataset and shown in Table 4.6. In our dataset the number of instances were 506 and 28 features. We have obtained 81.457 % accuracy with proposed techniques in Cleveland hospital data. On the other hand, we have achieved 95.85% accuracy in ANN with our processed dataset.

Table 4.6: Comparison of ANN algorithm's outcome between two datasets

Evaluation Metrics	Datasets		
	Our Dataset	Cleveland Dataset without processed technique	Cleveland Dataset with processed technique
Total Number of Instances	506	303	303
Number of Features	28	13	13
Accuracy	95.85%	80.4636 %	81.457 %
Incorrectly Classified Instances	4.15	19.5364 %	18.543 %
Kappa statistic	0.9018	0.6074	0.6272
Mean Absolute Error	0.040	0.2022	0.193
Root mean squared error	0.181	0.4184	0.413
TP Rate	93.464	0.805	0.815
Precision (Weighted Avg.)	0.959	0.805	0.815
Recall (Weighted Avg.)	95.8	0.805	0.815
F-Measure (Weighted Avg.)	0.959	0.805	0.815
MCC (Weighted Avg.)	0.902	0.608	0.627
ROC Area (Weighted Avg.)	0.986	0.865	0.875
PRC Area (Weighted Avg.)	0.989	0.861	0.870

In Table 4.7 the Bagging Algorithm has been applied to the two Cleveland dataset, one is processed and other is without following the proposed techniques. With the processed technique in Cleveland dataset 81.7881% accuracy have achieved but on the other hand without process it performs slightest better results with 83.4437% accuracy. With AFC Fortis hospital's data 96.44% accuracy has been obtained.

Table 4.7: Comparison of Bagging algorithm's outcome between two datasets

Evaluation Metrics	**Datasets**		
	Our Dataset	**Cleveland Dataset without processed technique**	**Cleveland Dataset with processed technique**
Total Number of Instances	506	303	303
Number of Features	28	13	13
Accuracy	96.44%	83.4437 %	81.7881 %
Incorrectly Classified Instances	3.55	16.5563 %	18.2119 %
Kappa statistic	0.9154	0.6657	0.6325
Mean Absolute Error	0.050	0.1887	0.1936
Root mean squared error	0.172	0.3594	0.3691
TP Rate	93.464	0.834	0.818
Precision (Weighted Avg.)	0.964	0.834	0.818
Recall (Weighted Avg.)	96.4	0.834	0.818
F-Measure (Weighted Avg.)	0.964	0.834	0.818
MCC (Weighted Avg.)	0.915	0.666	0.633
ROC Area (Weighted Avg.)	0.986	0.903	0.898
PRC Area (Weighted Avg.)	0.987	0.903	0.898

In Table 4.8 the Logistic Regression Algorithm has been applied to the two Cleveland dataset, one is processed and other is without following the proposed techniques. With the processed technique in Cleveland dataset 82.7815 % accuracy have achieved but on the other hand without process it performs slightest better results with 83.7748 % accuracy. With AFC Fortis hospital's data 96.64 % accuracy has been obtained.

Table 4.8: Comparison of Logistic Regression algorithm's outcome between two datasets

Evaluation Metrics	Datasets		
	Our Dataset	Cleveland Dataset without processed technique	Cleveland Dataset with processed technique
Total Number of Instances	506	303	303
Number of Features	28	13	13
Accuracy	96.64 %	83.7748 %	82.7815 %
Incorrectly Classified Instances	3.36	16.2252 %	17.2185 %
Kappa statistic	0.9196	0.6722	0.6512
Mean Absolute Error	0.042	0.2242	0.2258
Root mean squared error	0.173	0.3529	0.3563
TP Rate	92.81	0.838	0.828
Precision (Weighted Avg.)	0.966	0.838	0.829
Recall (Weighted Avg.)	96.6	0.838	0.828
F-Measure (Weighted Avg.)	0.966	0.837	0.827
MCC (Weighted Avg.)	0.920	0.673	0.653
ROC Area (Weighted Avg.)	0.964	0.898	0.896
PRC Area (Weighted Avg.)	0.966	0.887	0.885

In Table 4.9 the Random Forest Algorithm has been applied to the two Cleveland dataset, one is processed and other is without following the proposed techniques. With the processed technique in Cleveland dataset 79.8013 % accuracy have achieved but on the other hand without process it performs slightest better results with 82.4503 % accuracy. With AFC Fortis hospital's data 97.63% accuracy has been obtained.

Table 4.9: Comparison of Random Forest algorithm's outcome between two datasets

Evaluation Metrics	**Datasets**		
	Our Dataset	**Cleveland Dataset without processed technique**	**Cleveland Dataset with processed technique**
Total Number of Instances	506	303	303
Number of Features	28	13	13
Correctly Classified Instances	97.63%	82.4503 %	79.8013 %
Incorrectly Classified Instances	2.37	17.5497 %	20.1987 %
Kappa statistic	0.9425	0.6455	0.5911
Mean Absolute Error	0.076	0.2649	0.2736
Root mean squared error	0.156	0.352	0.3655
Root relative squared error	92.156	70.6189 %	73.3278 %
TP Rate	0.977	0.825	0.798
Precision (Weighted Avg.)	0.976	0.825	0.798
Recall (Weighted Avg.)	0.944	0.825	0.798
F-Measure (Weighted Avg.)	0.986	0.824	0.797
MCC (Weighted Avg.)	0.988	0.646	0.593
ROC Area (Weighted Avg.)	97.63	0.907	0.887
PRC Area (Weighted Avg.)	2.37	0.909	0.887

4.5 Receiver Operating Characteristic (ROC) Analysis of Different Algorithms

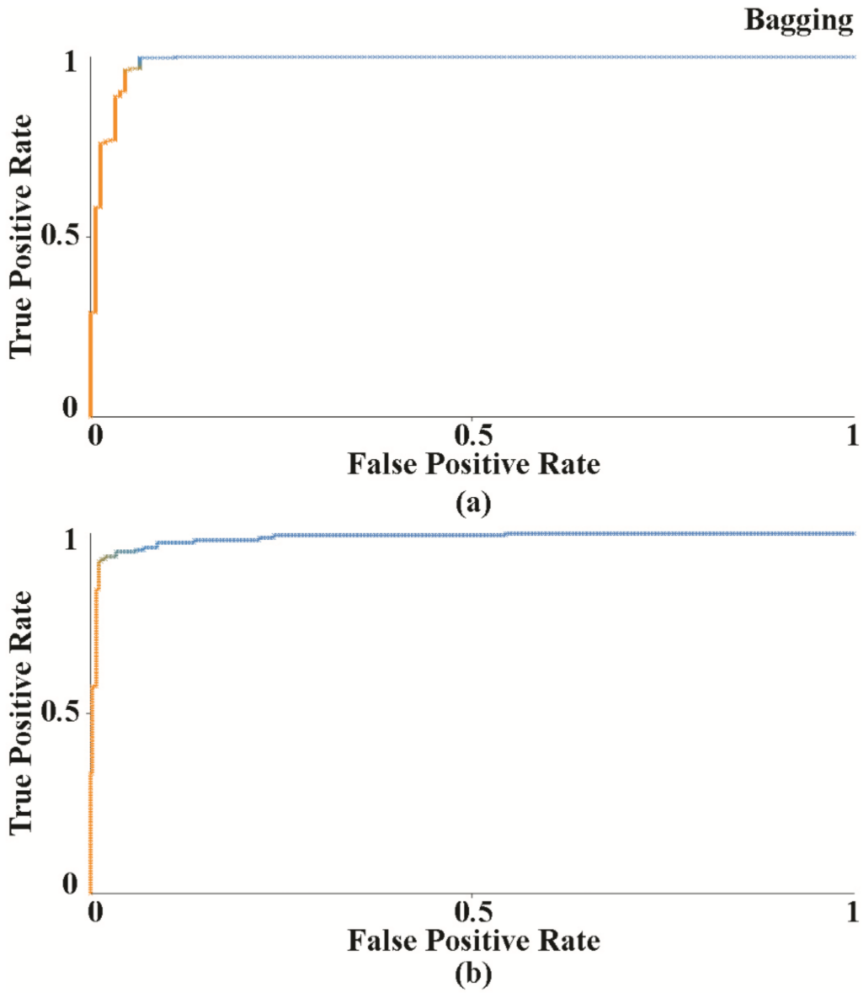

Figure. 4.13: ROC Curve of Bagging algorithm (a) For yes Class (b) For no Class

The ROC curves for both yes and no cases of Bagging algorithm has been shown in Figure. 4.13 where the True Positive Rate has been plotted on Y-axis and False Positive Rate on X-axis.

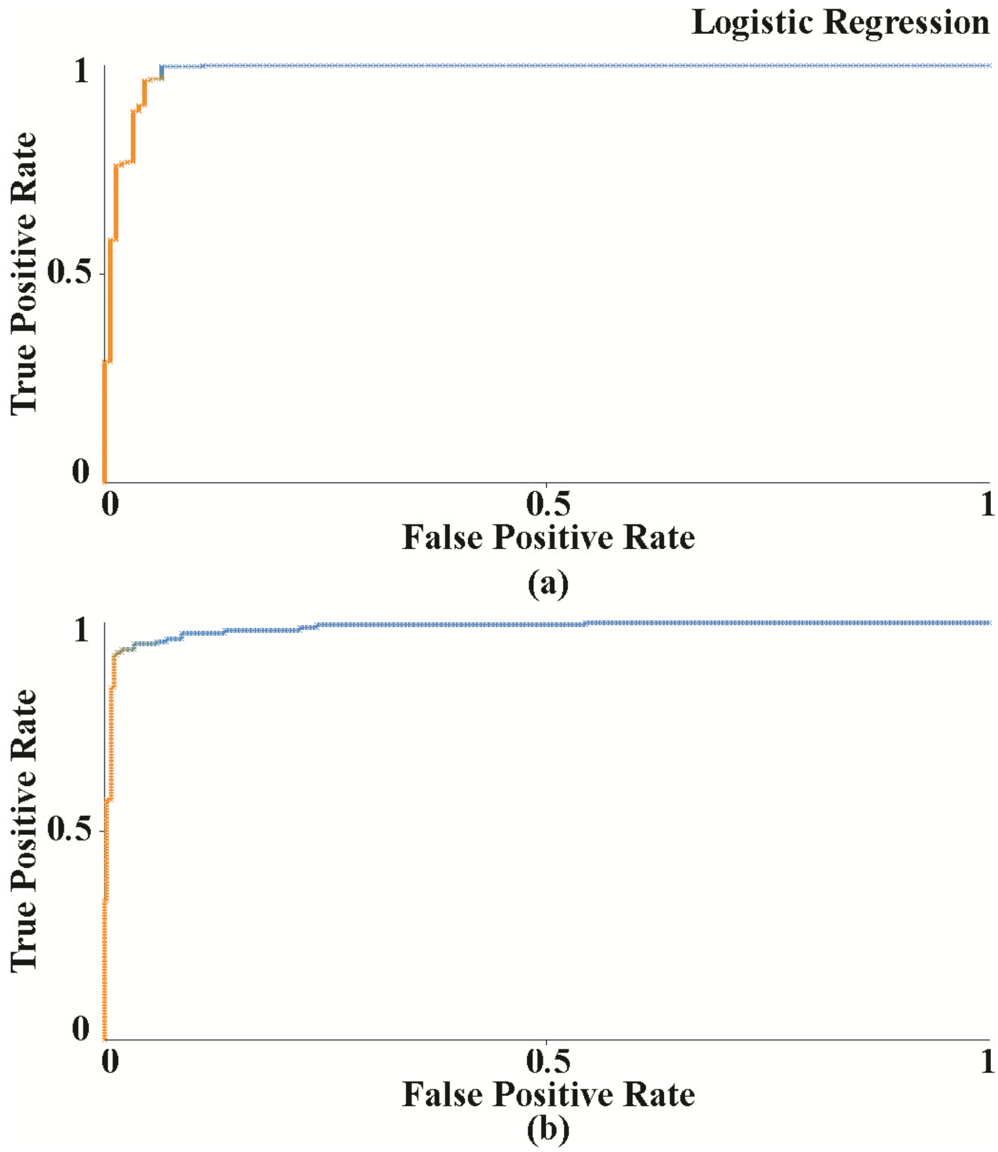

Figure. 4.14: ROC Curve of Logistic Regression algorithm (a) For yes Class (b) For no Class

Similarly, Figure. 4.14 represents the ROC curves of Logistic Regression and it also shows both cases yes and no. And, it was also plotted in the same manner as previous.

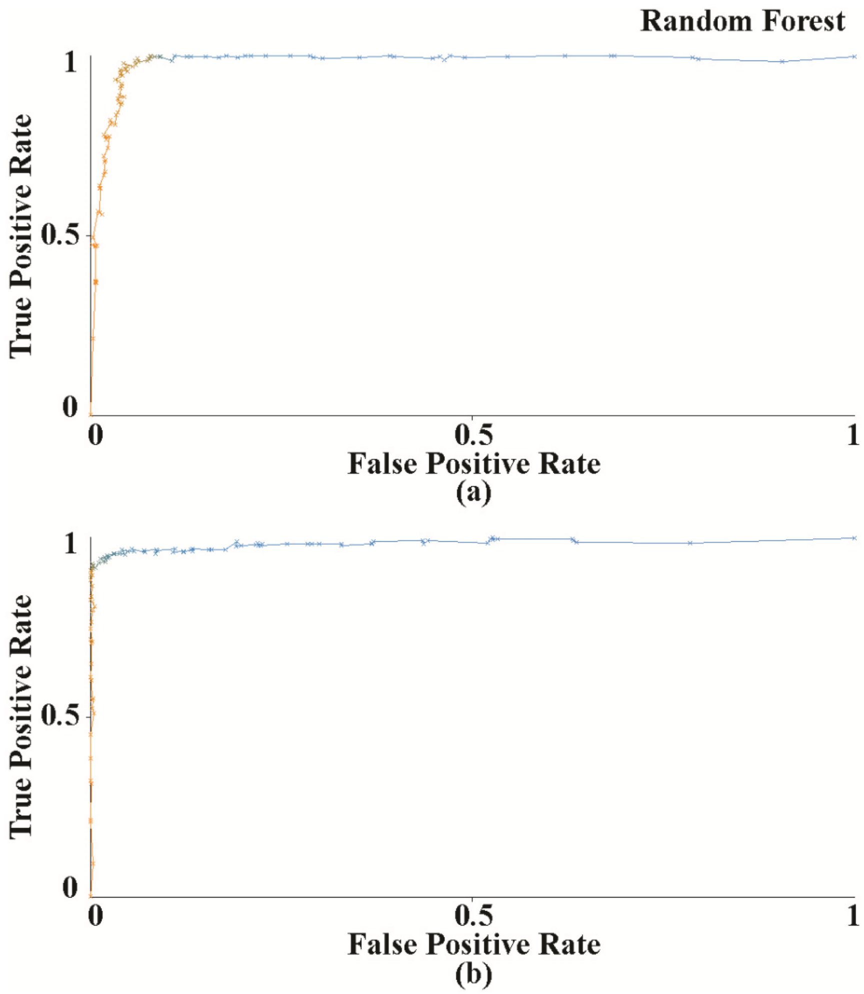

Figure. 4.15: ROC Curve of Random Forest algorithm (a) For yes Class (b) For no Class

Again, Figure 4.15 shows the ROC Curves for the Random Forest algorithm where the first on is for yes and the second one is for no.

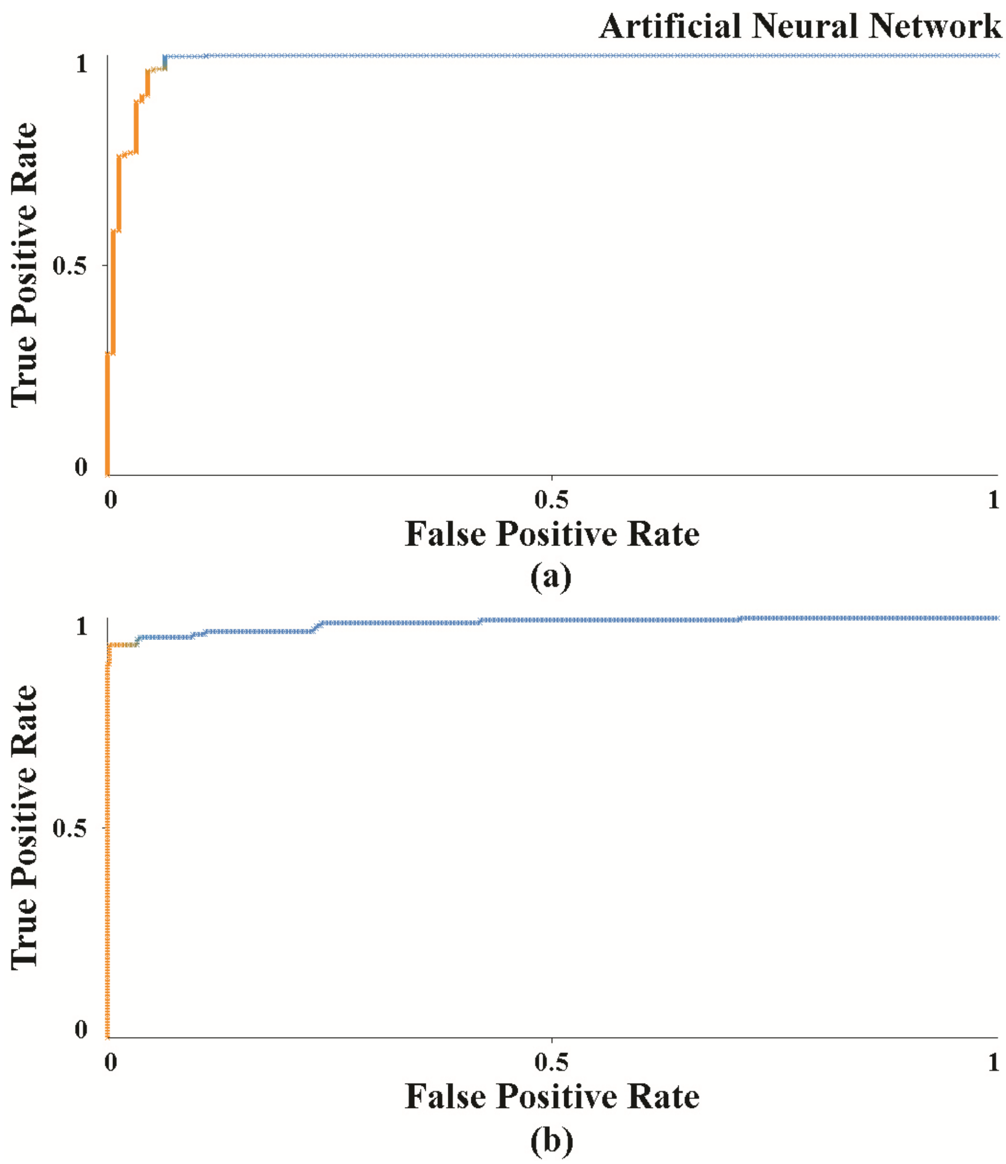

Figure. 4.16: ROC Curve of ANN algorithm (a) For yes Class (b) For no Class

The ROC curves for both yes and no cases of ANN algorithm has been shown in Figure. 4.16 where the True Positive Rate has been plotted on Y-axis and False Positive Rate on X-axis.

CHAPTER V

CONCLUSION

In section 5.1 the concluding remarks of this research have described and shown limitations in 5.2. Finally, future work has elaborated in 5.3.

5.1 Conclusion

In this work, the patient's data with heart disease from local hospitals was used and tried to predict IHD by applying computer intelligence. Around 506 patients' data were analyzed with 151 features categorized as demographic data, patient history, symptoms, and diagnosis. In this thesis, 28 mostly significant features were selected by using feature selection techniques and several classification techniques have been applied. The results for Artificial Neural Network (ANN), Bagging, Logistic Regression (LR) and Random Forest (RF) were very impressive. For ANN 95.85% accuracy has obtained, for Bagging 96.44%, for LR 96.64% and for RF 97.63% accuracy. So, the Random Forest algorithm gives the best performance among them. After that, the results have been compared based on different parameters of the algorithms such as kappa statistics, mean absolute error, root mean squared error, TP rate, FP rate, MCC, F-Measure, etc. By comparing all these statistical metrics, it was found that the Random Forest algorithm has given the best performance among the algorithms. The proposed technique has been applied on Cleveland hospital dataset to validate our model. The accuracy of the algorithms lies between 80.46% to 83.77% without applying the proposed processing techniques. The best result with 81.457% accuracy can be shown by applying the proposed processing techniques Artificial Neural Network. The Cleveland hospital data contains only 303 patient's data with 13 features.

5.2 Limitations

There were several limitations to perform the study. The number of instances were not so adequate. The study was performed based on the collected dataset from AFC Fortis Escorts Hospital, Khulna, which contains only 506 instances and each instance has 151 features. There are several effective methods for data preprocessing. Some typical methods like Mean, Mood, and Median were used in this research study. The data preprocessing technique needs to be improved more efficiently.

5.3 Future Works

The number of instances in the dataset needs to be increased to make a more effective model. There are some useful techniques such as Regression Analysis, Bayesian Belief, Naive Bayes, etc. that can be used for data preprocessing. There are several advanced Ensemble Learning algorithms, Radial Basis Neural Network (RBNN), Time Domain Neural Network (TDNN), Convolution Neural Network (CNN) as well as Support Vector Machine (SVM) with multi-class and binary Kernel that will be implemented. Finally, an expert system will be developed based on the best model achieved from this research study.

References

[1] "Coronary Heart Disease in Bangladesh", World Life Expectancy, 2019. [Online]. Available: https://www.worldlifeexpectancy.com/bangladesh-coronary-heart-disease. [Accessed: 15- Mar- 2018].

[2] "Overcoming a Fitness Plateau", www.heart.org, 2019. [Online]. Available: https://www.heart.org/en/healthy-living/fitness/staying-motivated/overcoming-a-fitness-plateau. [Accessed: 05- Mar- 2018].

[3] P. Palma, "Heart diseases now a plague in Bangladesh", The Daily Star, 2015.

[4] "Silent Ischemia and Ischemic Heart Disease", www.heart.org, 2019. [Online]. Available: https://www.heart.org/en/health-topics/heart-attack/about-heart-attacks/silent-ischemia-and-ischemic-heart-disease. [Accessed: 07- Mar- 2018].

[5] "Cardiovascular diseases affect nearly half of American adults, statistics show", *www.heart.org*, 2019. [Online]. Available: https://www.heart.org/en/news/2019/01/31/cardiovascular-diseases-affect-nearly-half-of-american-adults-statistics-show. [Accessed: 09- Mar- 2018].

[6] "Heart attacks in women undertreated, experts say", *www.heart.org*, 2019. [Online]. Available: https://www.heart.org/en/news/2018/05/01/heart-attacks-in-women-undertreated-experts-say. [Accessed: 10- Mar- 2018].

[7] "American Heart Association statistical report tracks global figures for first time", *www.heart.org*, 2014. [Online]. Available: https://www.heart.org/en/news/2018/05/01/american-heart-association-statistical-report-tracks-global-figures-for-first-time. [Accessed: 12- Mar- 2018].

[8] K. Shailaja, B. Seetharamulu and M. Jabbar, "Machine Learning in Healthcare: A Review", in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2018, pp. 910 - 914.

[9] K. Williams, P. Adebayo Idowu, J. Ademola Balogun and A. Ishola Oluwaranti, "Breast Cancer Risk Prediction Using Data Mining Classification Techniques", *Transactions on Networks and Communications*, vol. 3, no. 2, pp. 1-11, 2015.

[10] G. Parthiban and S. K. Srivatsa, "Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients", *International Journal of Applied Information Systems*, vol. 3, no. 7, pp. 25-30, 2012. Available: 10.5120/ijais12-450593 [Accessed 20 March 2018].

[11] A. Iyer, J. S and R. Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques", *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 1, pp. 01-14, 2015.

[12] Z. Lin, W. Yi, M. Lu, Z. Liu and H. Xu, "Correlation Research of Association Rules and Application in the Data about Coronary Heart Disease", in *2009 International Conference of Soft Computing and Pattern Recognition*, Malacca, Malaysia, 2009, pp. 143 - 148.

[13] M. jabbar, P. Chandra and B. Deekshatulu, "Prediction of risk score for heart disease using associative classification and hybrid feature subset selection", in *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, Kochi, India, 2019, pp. 628 - 634.

[14] E. AbuKhousa and P. Campbell, "Predictive data mining to support clinical decisions: An overview of heart disease prediction systems", in *2012 International Conference on Innovations in Information Technology (IIT)*, Abu Dhabi, United Arab Emirates, 2012, pp. 267 - 272.

[15] K. Rajmohan, I. Paramasivam and S. Sathyanarayan, "Prediction and Diagnosis of Cardio Vascular Disease -- A Critical Survey*", 2014 World Congress on Computing and Communication Technologies*, 2014. Available: 10.1109/wccct.2014.74 [Accessed 20 April 2018].

[16] R. Lafta, J. Zhang, X. Tao, Y. Li and V. Tseng, "An Intelligent Recommender System Based on Short-Term Risk Prediction for Heart Disease Patients", in *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Singapore, Singapore, 2015, pp. 102 - 105.

[17] T. Mahboob, R. Irfan and B. Ghaffar, "Evaluating ensemble prediction of coronary heart disease using receiver operating characteristics", in *2017 Internet Technologies and Applications (ITA)*, Wrexham, UK, 2017, pp. 110 - 115.

[18] T. Karayılan and Ö. Kılıç, "Prediction of heart disease using neural network", in *2017 International Conference on Computer Science and Engineering (UBMK)*, Antalya, Turkey, 2017, pp. 719 - 723.

[19] K. Pahwa and R. Kumar, "Prediction of heart disease using hybrid technique for selecting features", in *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, Mathura, India, 2017, pp. 500 - 504.

[20] F. Babič, J. Olejár, Z. Vantová and J. Paralič, "Predictive and Descriptive Analysis for Heart Disease Diagnosis", *in 2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Prague, Czech Republic, 2019, pp. 155 - 163.

[21] M. Kirmani, "Cardiovascular Disease Prediction Using Data Mining Techniques: A Review", *Oriental Journal of Computer Science & Technology*, vol. 10, no. 2, pp. 520-528, 2017.

[22] S. Pouriyeh, S. Vahid, G. Sannino, G. Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease", in *2017 IEEE Symposium on Computers and Communications (ISCC)*, Heraklion, Greece, 2019, pp. 204 - 207.

[23] R. Ani, S. Krishna, N. Anju, M. Aslam and O. Deepa, "Iot based patient monitoring and diagnostic prediction tool using ensemble classifier", in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Udupi, India, 2019, pp. 1588 - 1593.

[24] M. Raihan, M. Islam, P. Ghosh, S. Shaj, M. Chowdhury, S. Mondal and A. More, "A Comprehensive Analysis on Risk Prediction of Acute Coronary Syndrome Using Machine Learning Approaches", in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2018, pp. 1 - 6.

[25] D. Kinge and S. Gaikwad, "Survey on data mining techniques for disease prediction", *International Research Journal of Engineering and Technology (IRJET)*, vol. 05, no. 01, pp. 630-636, 2018.

[26] N. Kaur, V. Gupta, S. Kataria, "An Efficient Hybrid Classifier to Improve the Health Prediction using Data Mining", *International Journal of Research in Electronics and Computer Engineering*, vol. 06, no. 03, pp. 1171-1174, 2019.

[27] B. Rathnayakc and G. Ganegoda, "Heart diseases prediction with Data Mining and Neural Network Techniques", in *2018 3rd International Conference for Convergence in Technology (I2CT)*, Pune, India, 2018, pp. 1 - 6.

[28] K. More, M. Raihan, A. More, S. Padule and S. Mondal, "Smart phone based "heart attack" risk prediction; innovation of clinical and social approach for preventive cardiac health", *Journal of Hypertension*, vol. 36, p. e321, 2018.

[29] "Understand Your Risks to Prevent a Heart Attack", *www.heart.org*, 2016. [Online]. Available: https://www.heart.org/en/health-topics/heart-attack/understand-your-risks-to-prevent-a-heart-attack. [Accessed: 08- Sep- 2018].

[30] M. Marcus, "Some jobs seem riskier when it comes to heart health", *CBS NEWS*, 2016. [Accessed: 9- Sep- 2018].

[31] M. Solimene, "Coronary heart disease in women: a challenge for the 21st century", *Clinics*, vol. 65, no. 1, pp. 99-106, 2010. Available: 10.1590/s1807-59322010000100015 [10- Sep- 2018].

[32] C. Moran, "Farmers 7 times more likely to die from heart disease", *Agriland.ie*, 2015. [Online]. Available: https://www.agriland.ie/farming-news/farmers-7-times-more-likely-to-die-from-cardiovascular-disease/. [Accessed: 11- Sep- 2018].

[33] E. González, "Retiring may predispose to fatal heart attack", *JAMA: The Journal of the American Medical Association*, vol. 243, no. 1, p. 13, 1980. Available: 10.1001/jama.1980.03300270007004 [Accessed 12 September 2018].

[34] "Unemployment associated with 50% higher risk of death in heart failure patients", *Escardio.org*, 2019. [Online]. Available: https://www.escardio.org/The-ESC/Press-Office/Press-releases/unemployment-associated-with-50-higher-risk-of-death-in-heart-failure-patients. [Accessed: 13- Sep- 2018].

[35] "Amiram Elwork | Lawyers With Depression", *Lawyerswithdepression.wordpress.com*, 2009. [Online]. Available: https://lawyerswithdepression.wordpress.com/tag/amiram-elwork/. [Accessed: 02- Oct- 2018].

[36] E. Laurence, J. Volmink, T. Esterhuizen, S. Dalal and M. Holmes, "Risk of cardiovascular disease among teachers in Cape Town: Findings of the South African PaCT pilot study", *South African Medical Journal*, vol. 106, no. 10, p. 996, 2016. Available: 10.7196/samj.2016.v106i10.10869 [Accessed 05 October 2018].

[37] M. Selby-Green, "'If you keep working, you will die': London bankers in their 20s and 30s are having more heart attacks, doctors say", *Business Inside*, 2018. [Accessed 06 October 2018].

[38] M. Shiozaki, N. Miyai, I. Morioka, M. Utsumi, S. Hattori, H. Koike, M. Arita and K. Miyashita, "Job stress and behavioral characteristics in relation to coronary heart disease risk among Japanese police officers", *Industrial Health*, vol. 55, no. 4, pp. 369-380, 2017. Available: 10.2486/indhealth.2016-0179 [Accessed 7 October 2018].

[39] J. Arts, M. Fernandez and I. Lofgren, "Coronary Heart Disease Risk Factors in College Students", *Advances in Nutrition*, vol. 5, no. 2, pp. 177-187, 2014. Available: 10.3945/an.113.005447 [Accessed 8 October 2018].

[40] "Heart Attacks During and After Pregnancy", *Healthline*, 2019. [Online]. Available: https://www.healthline.com/health-news/women-having-heart-attacks-during-after pregnancy#1. [Accessed: 9- Oct- 2018].

[41] R. Rettner, "Why Short People May Have Higher Risk of Heart Disease", *Live Science*, 2015. [Online]. Available: https://www.livescience.com/50429-short-height-heart-disease-genes.html. [Accessed: 10- Oct- 2018].

[42] R. Santos, J. Mota, D. Santos, A. Silva, F. Baptista and L. Sardinha, "Physical fitness percentiles for Portuguese children and adolescents aged 10–18 years", *academia*, 2014. [Accessed: 11- Oct- 2018].

[43] "What is the body mass index (BMI)?", *nhs.uk*, 2016. [Online]. Available: https://www.nhs.uk/common-health-questions/lifestyle/what-is-the-body-mass-index-bmi/. [Accessed: 20- Oct- 2018].

[44] B. Gholipour, "What Is a Normal Heart Rate?", *Live Science*, 2018. [Online]. Available: https://www.livescience.com/42081-normal-heart-rate.html. [Accessed: 21- Oct- 2018].

[45] K. Bjartveit and A. Tverdal, "Health consequences of smoking 1-4 cigarettes per day", *Tobacco Control*, vol. 14, no. 5, pp. 315-320, 2005. Available: 10.1136/tc.2005.011932 [Accessed 22 October 2018].

[46] “How High Blood Pressure Can Lead to a Heart Attack", *www.heart.org*, 2016. [Online]. Available: https://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure/how-high-blood-pressure-can-lead-to-a-heart-attack. [Accessed: 23- Oct- 2018].

[47] J. B. Halter, J. G. Ouslander, S. Studenski, K. P. High, S. Asthana, M. A. Supiano and C. Ritchie., *Hazzard's Geriatric Medicine and Gerontology, 7e,* 7th ed. New York, N.Y.: McGraw-Hill Education LLC., 2017. [Accessed: 24- Oct- 2018].

[48] "Type 2 diabetes - Diagnosis and treatment - Mayo Clinic", *Mayoclinic.org*, 2019. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/diagnosis-treatment/drc-20351199. [Accessed: 25- Oct- 2018].

[49] M. MacGill, "Heart rate: What is a normal heart rate?", *Medical News Today*, 2017. [Online]. Available: https://www.medicalnewstoday.com/articles/235710.php. [Accessed: 26- Oct- 2018].

[50] "Syncopal Episode: Causes, Symptoms, and Care", *Healthline*, 2019. [Online]. Available: https://www.healthline.com/health/syncopal-episode. [Accessed: 27- Oct- 2018].

[51] "Angina Pectoris (Stable Angina)", *www.heart.org*, 2019. [Online]. Available: https://www.heart.org/en/health-topics/heart-attack/angina-chest-pain/angina-pectoris-stable-angina. [Accessed: 28- Oct- 2018].

[52] "Hemoglobin test", *Mayoclinic.org*, 2019. [Online]. Available: https://www.mayoclinic.org/tests-procedures/hemoglobin-test/about/pac-20385075. [Accessed: 29- Oct- 2018].

[53] "Creatinine — Serum | Medical Tests | UCSF Benioff Children's Hospital", *Ucsfbenioffchildrens.org*, 2019. [Online]. Available: https://www.ucsfbenioffchildrens.org/tests/003475.html. [Accessed: 30- Oct- 2018].

[54] "Random Blood Sugar (RBS) Test Lab Test", *Portea.com*, 2018. [Online]. Available: https://www.portea.com/labs/diagnostic-tests/random-blood-sugar-rbs-test-151/. [Accessed: 01- Nov- 2018].

[55] "Non-HDL Cholesterol (Non-HDL-C)", *Doc's Opinion*, 2013. [Accessed: 2- Nov-2018].

[56] "HDL Cholesterol: The Good Cholesterol", *WebMD*, 2005. [Accessed: 3- Nov- 2018].
[57] D. Ahmed, "Troponin Levels – The Heart Attack Blood Test", *MyHeart.net*, 2014. [Accessed: 6- Nov- 2018].
[58] “Ejection Fraction", *HRS,* 2019. [Online]. Available: https://www.hrsonline.org/Patient-Resources/The-Normal-Heart/Ejection-Fraction. [Accessed: 07- Nov- 2018].
[59] "The Danger of High Triglycerides for People with Heart Disease", *Cardiosmart.org*, 2016. [Online]. Available: https://www.cardiosmart.org/News-and-Events/2016/05/The-Danger-of-High-Triglycerides-for-People-with-Heart-Disease. [Accessed: 08- Nov- 2018].
[60] "Acute coronary syndrome", *Mayo Clinic*, 2019. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/acute-coronary-syndrome/symptoms-causes/syc-20352136. [Accessed: 09- Nov- 2018].
[61] "Myocardial infarction", *En.wikipedia.org,* 2019. [Online]. Available: https://en.wikipedia.org/wiki/Myocardial_infarction. [Accessed: 10- Nov- 2018].
[62] “Peripheral artery disease (PAD)", *Mayoclinic.org*, 2018. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/peripheral-artery-disease/symptoms-causes/syc-20350557. [Accessed: 11- Nov- 2018].
[63] “Unstable Angina", *Healthline*, 2019. [Online]. Available: https://www.healthline.com/health/unstable-angina. [Accessed: 11- Nov- 2018].
[64] Han W. Kim, Igor Klem, Dipan J. Shah, Edwin Wu, Sheridan N. Meyers, Michele A. Parker, Anna Lisa Crowley, Robert O. Bonow Robert M. Judd, and Raymond J. Kim, "Unrecognized Non-Q-Wave Myocardial Infarction: Prevalence and Prognostic Significance in Patients with Suspected Coronary Disease", *PLoS Medicine*, vol. 6, no. 4, p. e1000057, 2009. Available: 10.1371/journal.pmed.1000057 [Accessed 12 November 2018].
[65] R. Khayat, A. Pederzoli and W. Abraham, "Central Sleep Apnea in Heart Failure", *Uscjournal.com*, 2009. [Online]. Available: https://www.uscjournal.com/articles/central-sleep-apnea-heart-failure. [Accessed: 12- Nov- 2018].
[66] "Dilated cardiomyopathy", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Dilated_cardiomyopathy. [Accessed: 13- Nov- 2018].
[67] D. Friedman, J. Glickstein, L. Duncanson and J. Buyon, "A review of congenital heart block", *PubMed Central (PMC)*, 2003. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3232542/. [Accessed: 14- Nov- 2018].
[68] T. Kaplan, A. Berkowitz and M. Samuels, "Cardiovascular Dysfunction in Multiple Sclerosis", *The Neurologist*, vol. 20, no. 6, pp. 108-114, 2015. Available: 10.1097/nrl.0000000000000064 [Accessed 15 November 2018].
[69] D. Ahmed, "What’s a NSTEMI? Non ST Segment Myocardial Infarction", *MyHeart.NET*, 2015.
[70] C. Sandesara, R. Gopinathannair and B. Olshansky, "Implantable Cardiac Monitors: Evolution Through Disruption", *The Journal of Innovations in Cardiac Rhythm Management*, vol. 8, no. 9, pp. 2824-2834, 2017. Available: 10.19102/icrm.2017.080903 [Accessed 16 November 2018].

[71] "Atherosclerosis", *www.heart.org*, 2019. [Online]. Available: https://www.heart.org/en/health-topics/cholesterol/about-cholesterol/atherosclerosis. [Accessed: 17- Nov- 2018].
[72] R. Shabetai, "Pericardial effusion: haemodynamic spectrum", *Heart*, vol. 90, no. 3, pp. 255-256, 2004. Available: 10.1136/hrt.2003.024810 [Accessed 18 November 2018].
[73] "Aortic dissection", *Mayo Clinic*, 2019. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/aortic-dissection/symptoms-causes/syc-20369496. [Accessed: 19- Nov- 2018].
[74] "Heart disease & chronic kidney disease (CKD)", *Kidneyfund.org*, 2019. [Online]. Available: http://www.kidneyfund.org/kidney-disease/chronic-kidney-disease-ckd/complications/heart-disease/. [Accessed: 20- Nov- 2018].
[75] "Percutaneous coronary intervention", *Heart and Stroke Foundation of Canada*, 2019. [Online]. Available: https://www.heartandstroke.ca/heart/treatments/surgery-and-other-procedures/percutaneous-coronary-intervention. [Accessed: 21- Nov- 2018].
[76] "Congestive Heart Failure (CHF)", *Healthline*, 2019. [Online]. Available: https://www.healthline.com/health/congestive-heart-failure#prevention. [Accessed: 22- Nov- 2018].
[77] "Rheumatoid arthritis", *Mayo Clinic*, 2019. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/rheumatoid-arthritis/symptoms-causes/syc-20353648. [Accessed: 23- Nov- 2018].
[78] "Supraventricular Tachycardia Ablation - Cardiovascular Centre", *Cardiovascularcentre.com.au*, 2019. [Online]. Available: http://www.cardiovascularcentre.com.au/supraventricular-tachycardia-ablation. [Accessed: 24- Nov- 2018].
[79] Frustaci A, Magnavita N, Chimenti C, Caldarulo M, Sabbioni E, Pietra R, Cellini C, Possati GF, Maseri A., "Marked elevation of myocardial trace elements in idiopathic dilated cardiomyopathy compared with secondary cardiac dysfunction", *Journal of the American College of Cardiology*, vol. 33, no. 6, pp. 1578-1583, 1999. Available: 10.1016/s0735-1097(99)00062-5 [Accessed 25 November 2018].
[80] J. Warner, "Heart Failure Patients Too Optimistic", *WebMD*, 2008. [Online]. Available: https://www.webmd.com/heart-disease/heart-failure/news/20080603/heart-failure-patients-too-optimistic. [Accessed: 26- Nov- 2018].
[81] "Rheumatic Heart Disease", *Johns Hopkins Medicine Health Library*, 2019. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/rheumatic-heart-disease. [Accessed: 27- Nov- 2018].
[82] "Bradycardia Causes, Signs & Symptoms, and Diagnosis | Baptist Health", *Baptist Health*, 2019. [Online]. Available: https://www.baptisthealth.com/pages/services/heart-care/conditions/bradycardia.aspx. [Accessed: 28- Nov- 2018].
[83] *If You Choose Not to Start Dialysis Treatment*. New York: National Kidney Foundation, 2008.
[84] “Atrial fibrillation", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Atrial_fibrillation. [Accessed: 30- Nov- 2018].
[85] "Respiratory infection can increase the risk of a heart attack", *The University of Sydney*, 2017. [Online]. Available: https://sydney.edu.au/news-opinion/news/2017/05/16/respiratory-infection-can-increase-the-risk-of-a-heart-attack.html. [Accessed: 01- Dec- 2018].

[86] R. Klabunde, "Cardiovascular Physiology Concepts", *Cvphysiology.com*, 2014. [Online]. Available: https://www.cvphysiology.com/Blood%20Pressure/BP031. [Accessed: 02- Dec- 2018].

[87] "Cardiogenic shock", *Mayo Clinic*, 2017. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/cardiogenic-shock/symptoms-causes/syc-20366739. [Accessed: 03- Dec- 2018].

[88] "Aortic valve stenosis", *Mayo Clinic*, 2018. [Online]. Available: 60. https://www.mayoclinic.org/diseases-conditions/aortic-stenosis/symptoms-causes/syc-20353139. [Accessed: 4- Dec- 2018].

[89] "Coronary Artery Bypass Grafting", *Nhlbi.nih.gov*, 2014. [Online]. Available: https://www.nhlbi.nih.gov/health-topics/coronary-artery-bypass-grafting. [Accessed: 04- Dec- 2018].

[90] "AHA Releases 2015 Heart and Stroke Statistics", *Sca-aware.org*, 2014. [Online]. Available: https://www.sca-aware.org/sca-news/aha-releases-2015-heart-and-stroke-statistics. [Accessed: 05- Dec- 2018].

[91] "Valvular Heart Disease", *Johns Hopkins Medicine Health Library*, 2019. [Online]. Available: https://www.hopkinsmedicine.org/heart_vascular_institute/conditions_treatments/conditions/valvular_heart_disease.html. [Accessed: 06- Dec- 2018].

[92] N. Kakouros and D. Cokkinos, "Right ventricular myocardial infarction: pathophysiology, diagnosis, and management", *Postgraduate Medical Journal*, vol. 86, no. 1022, pp. 719-728, 2010. Available: 10.1136/pgmj.2010.103887 [Accessed 7 December 2018].

[93] "Flu and Pneumonia Prevention", *www.heart.org*, 2015. [Online]. Available: https://www.heart.org/en/health-topics/consumer-healthcare/flu-and-pneumonia-prevention. [Accessed: 8 December 2018].

[94] "First-degree atrioventricular block", *En.wikipedia.org*, 2018. [Online]. Available: https://en.wikipedia.org/wiki/First-degree_atrioventricular_block. [Accessed: 09- Dec- 2018].

[95] "Sinus rhythm", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Sinus_rhythm. [Accessed: 10- Dec- 2018].

[96] D. Duncker, R. Bache, D. Merkus and M. Laughlin, "Exercise and the Coronary Circulation", *Muscle and Exercise Physiology*, pp. 467-503, 2019. Available: 10.1016/b978-0-12-814593-7.00022-0 [Accessed 11 December 2018].

[97] "Dissociated vertical deviation", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Dissociated_vertical_deviation. [Accessed: 12- Dec- 2018].

[98] "Streptokinase", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Streptokinase. [Accessed: 13- Dec- 2018].

[99] "Cardiac Arrest", *The Heart Foundation*, 2019. [Online]. Available: https://www.heartfoundation.org.au/your-heart/heart-conditions/cardiac-arrest. [Accessed: 14- Dec- 2018].

[100] "Ventricular tachycardia", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Ventricular_tachycardia. [Accessed: 15- Dec- 2018].

[101] "Thyroid hormone: How it affects your heart - Harvard Health", *Harvard Health*, 2019. [Online]. Available: https://www.health.harvard.edu/heart-health/thyroid-hormone-how-it-affects-your-heart. [Accessed: 16- Dec- 2018].

[102] "Bigeminy", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Bigeminy. [Accessed: 17- Dec- 2018].

[103] "What is Atrial Fibrillation (AFib or AF)?", *www.heart.org*, 2016. [Online]. Available: https://www.heart.org/en/health-topics/atrial-fibrillation/what-is-atrial-fibrillation-afib-or-af. [Accessed: 18- Dec- 2018].

[104] A. S. Maisel, N. Katz, H. L. Hillege, A. Shaw, P. Zanco, R. Bellomo, I. Anand, S. D. Anker, N. Aspromonte, S. M. Bagshaw Tomas Berl, I. Bobek, D. N. Cruz, L. Daliento, A. Davenport, M. Haapio, A. A. House, S. Mankad, P. McCullough, Alexandre Mebazaa, Alberto Palazzuoli, Piotr Ponikowski, Federico Ronco, Geoff Sheinfeld, Sachin Soni, Giorgio Vescovo, Nereo Zamperetti and Claudio Ronco., "Biomarkers in kidney and heart disease", *Nephrology Dialysis Transplantation*, vol. 26, no. 1, pp. 62-74, 2010. Available: 10.1093/ndt/gfq647 [Accessed 19 December 2018].

[105] "Cardiac stress test", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Cardiac_stress_test. [Accessed: 20- Dec- 2018].

[106] "Coronary Artery Disease | CAD | MedlinePlus", *Medlineplus.gov*, 2019. [Online]. Available: https://medlineplus.gov/coronaryarterydisease.html. [Accessed: 21- Dec- 2018].

[107] "Small vessel disease", *Mayo Clinic*, 2018. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/small-vessel-disease/symptoms-causes/syc-20352117. [Accessed: 22- Dec- 2018].

[108] "Deep Vein Thrombosis", *Cardiosmart.org*, 2019. [Online]. Available: https://www.cardiosmart.org/Heart-Conditions/Deep-Vein-Thrombosis. [Accessed: 23- Dec- 2018].

[109] "Heart Attack (Myocardial Infarction) | Cleveland Clinic", *Cleveland Clinic*, 2019. [Online]. Available: https://my.clevelandclinic.org/health/diseases/16818-heart-attack-myocardial-infarction. [Accessed: 23- Dec- 2018].

[110] R. Sánchez-Soriano, C. Chamorro-Fernández, R. Raso-Raso, J. Valencia, V. Mainar and G. Grau-Jornet, "Patient With Angina and “Congenital Bypass”. A New Case of Aortocoronary Fistula", *Revista Española de Cardiología (English Edition)*, vol. 66, no. 10, pp. 825-827, 2013. Available: 10.1016/j.rec.2013.04.017 [Accessed 24 December 2018].

[111] "Congestive hepatopathy", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Congestive_hepatopathy. [Accessed: 25- Dec- 2018].

[112] Sophia Morel, Jade Leahy, Maryse Fournier, Benoit Lamarche, Carole Garofalo, Guy Grimard, Floriane Poulain, Edgard Delvin, Caroline Laverdière, Maja Krajinovic, Simon Drouin, Daniel Sinnett, Valérie Marcil and Emile Levy., "Lipid and lipoprotein abnormalities in acute lymphoblastic leukemia survivors", *Journal of Lipid Research*, vol. 58, no. 5, pp. 982-993, 2017. Available: 10.1194/jlr.m072207 [Accessed 27 December 2018].

[113] "Renal artery stenosis", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Renal_artery_stenosis. [Accessed: 28- Dec- 2018].

[114] "Frequently Asked Questions about Schizophrenia", *Brain & Behavior Research Foundation*, 2019. [Online]. Available: https://www.bbrfoundation.org/faq/frequently-asked-questions-about-schizophrenia. [Accessed: 29- Dec- 2018].

[115] G. FAUNTLEROY, "Hypothyroidism and its Effects on the Heart | Endocrine News", *Endocrine News*, 2013. [Online]. Available: https://endocrinenews.endocrine.org/hypothyroidism-and-the-heart/. [Accessed: 30- Dec- 2018].

[116] C. Nordqvist, "Coronary heart disease: What you need to know", *Medical News Today*, 2018. [Online]. Available: https://www.medicalnewstoday.com/articles/184130.php. [Accessed: 31- Dec- 2018].

[117] K. Scotland and D. Lange, "Prevention and management of urosepsis triggered by ureteroscopy", *Research and Reports in Urology*, vol. 10, pp. 43-49, 2018. Available: 10.2147/rru.s128071 [Accessed 1 January 2019].

[118] "Left bundle branch block", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Left_bundle_branch_block. [Accessed: 02- Jan- 2019].

[119] "Permanent Pacemaker", *Cleveland Clinic*, 2019. [Accessed: 03- Jan- 2019].

[120] S. DS, W. D, I. A and J. Nephrol, "The role of anemia in the progression of congestive heart failure. Is there a place for erythropoietin and intravenous iron? - PubMed - NCBI", *Ncbi.nlm.nih.gov*, 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/15593047. [Accessed: 04- Jan- 2019].

[121] "The Relationship between COPD and Heart Problems", *Lung Institute*, 2019. [Online]. Available: https://lunginstitute.com/blog/the-relationship-between-copd-and-heart-problems/. [Accessed: 05- Jan- 2019].

[122] J. Naschitz, G. Slobodin, R. Lewis, E. Zuckerman and D. Yeshurun, "Heart diseases affecting the liver and liver diseases affecting the heart", *American Heart Journal*, vol. 140, no. 1, pp. 111-120, 2000. Available: 10.1067/mhj.2000.107177 [Accessed 6 January 2019].

[123] I. van Hagen et al., "Pregnancy Outcomes in Women with Rheumatic Mitral Valve Disease", *Circulation*, vol. 137, no. 8, pp. 806-816, 2018. Available: 10.1161/circulationaha.117.032561 [Accessed 7 January 2019].

[124] H. Staff, "Learning About Metabolic Encephalopathy", *Myhealth.alberta.ca*, 2018. [Online]. Available: https://myhealth.alberta.ca/Health/aftercareinformation/pages/conditions.aspx?hwid=abo4850. [Accessed: 08- Jan- 2019].

[125] "Intracranial artery stenosis", *MAYFIELD Brain & Spine*, 2018. [Online]. Available: https://mayfieldclinic.com/pe-intracranialstenosis.htm. [Accessed: 09- Jan- 2019].

[126] "Small intestinal ischemia and infarction: MedlinePlus Medical Encyclopedia", *Medlineplus.gov*, 2019. [Online]. Available: https://medlineplus.gov/ency/article/001151.htm. [Accessed: 11- Jan- 2019].

[127] H. Adrogué, "Hyponatremia in Heart Failure", *Methodist DeBakey Cardiovascular Journal*, vol. 13, no. 1, pp. 40-40, 2017. Available: 10.14797/mdcj-13-1-40 [Accessed 12 January 2019].

[128] J. Stephenson, "Symptomatic bradycardia 1: physiology and causes", *Nursing Times*, 2008. [Online]. Available: https://www.nursingtimes.net/clinical-archive/cardiovascular/symptomatic-bradycardia-1-physiology-and-causes/1872019.article. [Accessed: 13- Jan- 2019].

[129] “Heart Disease and Marfan Syndrome", *WebMD*, 2019. [Online]. Available: https://www.webmd.com/heart-disease/guide/marfan-syndrome#1. [Accessed: 14- Jan- 2019].
[130] "Cough-Variant Asthma", *WebMD*, 2019. [Online]. Available: https://www.webmd.com/asthma/guide/cough-variant-asthma#1. [Accessed: 15- Jan- 2019].
[131] A. Shehab, N. Gupta and S. Mushahawar, "The Reduction in Total Mortality with Drug-Eluting Stents: Does Emperor has New Clothes?", *Interventional Cardiology Journal*, vol. 02, no. 03, 2016. Available: 10.21767/2471-8157.100032 [Accessed 16 January 2019].
[132] M. Mouallem, E. Friedman, Y. Shemesh, H. Mayan, R. Pauzner and Z. Farfel, "Cardiac conduction defects associated with hyponatremia", *Clinical Cardiology*, vol. 14, no. 2, pp. 165-168, 1991. Available: 10.1002/clc.4960140214 [Accessed 18 January 2019].
[133] "Bronchial Asthma", *Medanta.org*, 2017. [Online]. Available: https://www.medanta.org/bronchial-asthma/. [Accessed: 19- Jan- 2019].
[134] "What Causes Anemia?", *Healthline*, 2019. [Online]. Available: https://www.healthline.com/symptom/anemia. [Accessed: 19- Jan- 2019].
[135] R. Waksman et al., "Transcatheter Aortic Valve Replacement in Low-Risk Patients: One-Year Results from the LRT Trial", *JACC: Cardiovascular Interventions*, 2019. Available: 10.1016/j.jcin.2019.03.002 [Accessed 20 January 2019].
[136] A. Manolis et al., "Erythropoietin in Heart Failure and Other Cardiovascular Diseases: Hematopoietic and Pleiotropic Effects", *Current Drug Target -Cardiovascular & Hematological Disorders*, vol. 5, no. 5, pp. 355-375, 2005. Available: 10.2174/156800605774370326 [Accessed 21 January 2019].
[137] J. McIntosh, "What to know about urinary tract infections", *Medical News Today*, 2018. [Online]. Available: https://www.medicalnewstoday.com/articles/189953.php. [Accessed: 22- Jan- 2019].
[138] E. Ekpe, I. Essien and U. Idongesit, "Significant pleural effusion in congestive heart failure necessitating pleural drainage", *Nigerian Journal of Cardiology*, vol. 12, no. 2, p. 106, 2015. Available: 10.4103/0189-7969.152020 [Accessed 23 January 2019].
[139] M. Theobald, "Psoriasis and Heart Disease: The Hidden Connection", *EverydayHealth.com*, 2017. [Online]. Available: https://www.everydayhealth.com/psoriasis/living-with/psoriasis-heart-health-untold-story/. [Accessed: 24- Jan- 2019].
[140] K. Patel, "Left main stem Disease- An evolving conundrum", *British Cardiovascular Society*, 2016.
[141] S. Zanelli, "Hypoxic-Ischemic Encephalopathy", *Medscape*, 2018.
[142] [6]"Hepatitis B Blood Tests", *Hepb.org*, 2019. [Online]. Available: http://www.hepb.org/prevention-and-diagnosis/diagnosis/hbv-blood-tests/. [Accessed: 25- Jan- 2019].
[143] "Family history", *Bhf.org.uk*, 2019. [Online]. Available: https://www.bhf.org.uk/informationsupport/risk-factors/family-history. [Accessed: 26- Jan- 2019].

[144] "Risk Factors", *Hear Research Australia*, 2019. [Online]. Available: https://www.heartresearch.com.au/heart-disease/risk-factors/. [Accessed: 27- Jan- 2019].
[145] "Stress and Heart Health", *www.heart.org*, 2018. [Online]. Available: https://www.heart.org/en/healthy-living/healthy-lifestyle/stress-management/stress-and-heart-health. [Accessed: 27- Jan- 2019].
[146] "Learn the difference between a heart attack and stroke?", *Inovaheart.org*, 2019. [Online]. Available: https://www.inovaheart.org/smarthearts/heart-attack-vs-stroke. [Accessed: 28- Jan- 2019].
[147] "Heart palpitations", *Mayo Clinic*, 2018. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/heart-palpitations/symptoms-causes/syc-20373196. [Accessed: 28- Jan- 2019].
[148] "Syncope (Fainting)", *www.heart.org*, 2017. [Online]. Available: https://www.heart.org/en/health-topics/arrhythmia/symptoms-diagnosis--monitoring-of-arrhythmia/syncope-fainting. [Accessed: 29- Jan- 2019].
[149] "Chest pain", *Healthdirect.gov.au*, 2017. [Online]. Available: https://www.healthdirect.gov.au/chest-pain. [Accessed: 29- Jan- 2019].
[150] "Electrocardiogram (ECG) and high blood pressure", *Bloodpressureuk.org*, 2019. [Online]. Available: http://www.bloodpressureuk.org/BloodPressureandyou/Medicaltests/ECG. [Accessed: 30- Jan- 2019].
[151] "Coronary angiogram", *Mayo Clinic*, 2018. [Online]. Available: https://www.mayoclinic.org/tests-procedures/coronary-angiogram/about/pac-20384904. [Accessed: 31- Jan- 2019].
[152] A. Unnam, "WEKA – Graphical User Interference Way to Learn Machine Learning", *Analytics Vidhya*, 2019. [Online]. Available: https://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/weka-gui-learn-machine-learning/. [Accessed: 14- May- 2018].
[153] I. Witten, E. Frank, M. Hall and C. Pal, *Data Mining Practical Machine Learning Tools and Techniques*, 4th ed. Morgan Kaufmann, 2016, pp. 166-580.
[154] "SPSS Software", *Ibm.com*, 2019. [Online]. Available: https://www.ibm.com/analytics/spss-statistics-software. [Accessed: 15- May- 2018].
[155] "ReplaceMissingValues", *sourceforge*, 2019.
[156] "Controlling Random Number Generation- MATLAB & Simulink Example", *Mathworks.com*, 2019. [Online]. Available: https://www.mathworks.com/help/matlab/examples/controlling-random-number-generation.html. [Accessed: 17- May- 2018].
[157] “Best-first search", *cs.washington*, 2019.
[158] "BestFirst", *sourceforge*, 2019.
[159] R. Kohavi and G. John, "Wrappers for feature subset selection", *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273-324, 1997. Available: 10.1016/s0004-3702(97)00043-x [Accessed 22 May 2018].
[160] "CfsSubsetEval", *sourceforge*, 2019.
[161] "InfoGainAttributeEval", *sourceforge*, 2019.

[162] "Using Chi-Square Statistic in Research - Statistics Solutions", *Statistics Solutions*, 2019. [Online]. Available: https://www.statisticssolutions.com/using-chi-square-statistic-in-research/. [Accessed: 30- May- 2018].

[163] "Statistics - Resampling through Random Percentage Split [Gerardnico]", *Gerardnico.com*, 2018. [Online]. Available: https://gerardnico.com/data_mining/validation_set. [Accessed: 08- Jun- 2018].

[164] G. Drakos, "Cross-Validation", *Towards Data Science*, 2018. [Online]. Available: https://towardsdatascience.com/cross-validation-70289113a072. [Accessed: 20- Aug- 2018].

[165] J. Brownlee, "A Gentle Introduction to k-fold Cross-Validation", *Machine Learning Mastery*, 2018.

[166] M. Sidana, "Types of classification algorithms in Machine Learning", *Medium*, 2017. [Online]. Available: https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14. [Accessed: 22- Aug- 2018].

[167] U.., "Classification-algorithms", *edureka*, 2019

[168] H. Kandan, "Bagging the skill of Bagging (Bootstrap aggregating).", *Medium*, 2018. [Online]. Available: https://medium.com/@harishkandan95/bagging-the-skill-of-bagging-bootstrap-aggregating-83c18dcabdf1. [Accessed: 28- Aug- 2018].

[169] "LogisticRegression", *nickgillian*, 2014.

[170] "The Logistic Regression Algorithm", *machinelearning-blog*, 2019.

[171] F. Schoonjans, "Logistic regression", *MedCalc*, 2019. [Online]. Available: https://www.medcalc.org/manual/logistic_regression.php. [Accessed: 14- Jan- 2019].

[172] J. Brownlee, "How to Use Classification Machine Learning Algorithms in WEKA", *Machine Learning Mastery*, 2016.

[173] S. Sharma, "Artificial Neural Network (ANN) in Machine Learning", *Datasciencecentral*, 2017. [Online]. Available: https://www.datasciencecentral.com/profiles/blogs/artificial-neural-network-ann-in-machine-learning. [Accessed: 30- Sep- 2018].

[174] J. Mahanta, "Introduction to Neural Networks, Advantages and Applications", *Towards Data Science*, 2017.

[175] S., "How Random Forest Algorithm Works in Machine Learning", *Medium*, 2017.

[176] “Random Forest algorithm", *amateurdatascientist*, 2019.

[177] N. Donges, "The Random Forest Algorithm", *Medium*, 2018.

Appendix

Data Collection Sheet

Risk Factors

Date: Registration No: MRD No:
Time: Patient Name :
Age: Sex: Contact:
Address :
Profession:
Date of Admission: Time of Admission:
D/O Discharge: BP: …………………………………….mm of Hg
Height: ……………………..cm Weight: …………………………………Kg
BMI: ………………………………………kg/m^2 HR: ……………/min

Diagnosis 1	
Diagnosis 2	
Diagnosis 3	
Diagnosis 4	
Diagnosis 5	
Diagnosis 6	

History	Yes/No	Duration	Remarks
Family History of Cardiovascular Disease			
Smoking			
H/O Tobacco Chewing			
Hypertension			
Dyslipidemia (DLP)			
Diabetes Mellitus (DM)			
Physical Inactivity			
Psychological			

Stress			
H/O Stroke			

PAST HISTORY:

CAG	CABG	Heart Attack	Psychological Disorder

Presentation	Yes/No	Duration	Remarks
Chest Pain			
Dyspnea			
Palpitation			
Syncope			
Nausea			
Sweating			
Vomiting			

Investigations	Value		Value
Hb %		Troponin I	
Sr. Creatinine		EF	
RBS	moml/L	Triglyceride	
Total Cholesterol		LDL	
HDL		ECG	
SPO_2			

Others:

	STEMI	NSTEMI	Remarks
TIMI Score			
Coronary Angiogram (CAG)			
Percutaneous Coronary Angioplasty (PTCA)			

Outcome: ……………………………………………………………………………………………………

For Technical Use:

Table A1: Collected Demographic Features (Part 1)

Feature Names	**Subcategories**	**Data Distribution**
		Mean ± Standard Deviation
Age	Minimum:15 Maximum:100	57.206 ± 13.121
Sex	Male	381
	Female	125
Profession	Business	365
	Housewife	53
	Salesman	1
	Farmer	10
	Retired	2
	Unemployed	5
	Lawyer	2
	Teacher	9
	Docwriter	1
	shopkeeper	4
	Private	2
	Banker	6
	Govt	3
	Police	4
	Student	2
	Labour	1
	Doctor	3
	Carpainter	2
	Pharmacist	1
	Advocate	3
	Professor	2
	Officer	1
	Engineer	2
	Service	2
	Driver	1
height	Minimum: 136 Maximum:190	162.753 ± 4.821
weight	Minimum: 45 Maximum:115	66.204 ± 5.72
BMI	Minimum: 16.598 Maximum:37.949	24.961 ± 1.932

Table A2: Collected Diagnostic Features (Part 2)

Feature Names	**Data Distribution**
	Mean ± Standard Deviation
Heart Rate (HR)	82.855 ± 26.015
Systolic Blood Pressure (bpsystolic)	130.7 ± 27.754
Diastolic Blood Pressure (bpdiastolic)	83.442 ± 14.818
Hemoglobin	12.27 ± 2.022
Creatinine	1.718 ± 1.651
Random Blood Sugar (RBS)	9.69 ± 7.306
Total cholesterol	167.429 ± 5.431
High Density Lipoprotein (HDL)	36.263 ± 1.025
Troponin	3.059 ± 4.171
Ejection Fraction (EF)	49.709 ± 12.278
Triglycerides	133.557 ± 6.103
Low Density Lipoprotein (LDL)	94.757 ± 3.547

Table A3: Collected Patient's History Features (Part 3)

Feature Names	**Data Distribution**
	Mean ± Standard Deviation
Smoking duration (in days)	6647.909 ± 1505.21
Smoking per day	16.542 ± 1.747
Tobacco duration (in days)	4912.742 ± 672.846
Hypertension (HTN) duration (in days)	2493.291 ± 1260.221
Dyslipoproteinemia (DLP) duration (in days)	1490.417 ± 136.78
Diabetes Mellitus (DM) duration (in days)	3358.152 ± 1270.055
Dyspnea duration (in days)	85.228 ± 156.522
Palpitation duration	35.143 ± 9.846
Syncope duration	2 ± 0.063
Chest pain duration	730.462 ± 142.925

Table A.4: Collected Symptoms Table (Part 4)

Feature Names	Subcategories	Data Distribution
Acute Coronary Syndromes (ACS)	yes	353
	No	153
Inferior Wall ST Segment Elevation Myocardial Infarction (IWMI)	yes	21
	No	485
Stable Angina	yes	22
	No	484
Unstable Angina	yes	61
	No	445
Non-ST-Elevation Myocardial Infarction (Non-STEMI)	yes	70
	No	436
Acute Myocardial Infarction (AMI)	yes	46
	No	460
Peripheral Artery Disease (PAD)	yes	1
	No	505
Unrecognized Myocardial Infraction (UMI)	yes	1
	No	505
Hypertension (HTN)	yes	352
	No	154
Acute Left Ventricular Failure (ALVF)	yes	31
	No	475
Cardiogenic Shock	yes	26
	No	480
Double Vessel Disease (DVD)	yes	53
	No	453
Thrombolysed Stroke	yes	21
	No	485
Treadmill Test (TMT)	yes	27
	No	479
Controlled Substances Act (CSA)	yes	10
	No	496
Coronary Artery Disease CAD)	yes	56
	No	450
Single Vessel Disease (SVD)	yes	54
	No	452

Feature Names	Subcategories	Data Distribution
Percutaneous Transluminal Coronary Angioplasty (PTCA)	yes	88
	No	418
Dyslipoproteinemia	yes	66
	No	440
Complete Heart Block (CHB)	yes	16
	No	490
Family History	yes	152
	No	354
Dilated Cardiomyopathy (DCM)	yes	5
	No	501
Smoking	yes	173
	No	270
	Ex	63
Tobacco	yes	108
	No	398
Diabetes Mellitus (DM)	yes	222
	No	284
Severe Mitral Regurgitation (SMR)	yes	1
	No	505
Physical Activity	yes	24
	No	482
Psychological Stress	yes	108
	No	398
Stroke	yes	129
	No	377
Severe Multiple Sclerosis (SMS)	yes	3
	No	503
Drug History	yes	391
	No	115
Dyspnea	yes	205
	No	301
Palpitation	yes	49
	No	457
Ischemic Cardiomyopathy (ICM)	yes	6
	No	500
Syncope	yes	24
	No	482

Feature Names	Subcategories	Data Distribution
Sweating	yes	178
	No	328
Vomiting	yes	105
	No	401
Radiation	yes	38
	No	468
Chest Pain	yes	392
	No	114
Pericardial Effusion (PE)	yes	1
	No	505
Electrocardiogram (ECG)	Abnormal	410
	normal	96
Coronary Artery Graft Coronary Angiography (CAG)	yes	234
	No	272
Patients Status	Expired	38
	discharged	468
Dissecting Aortic Aneurysm (DAA)	yes	1
	No	505
Anterior Ischemia	yes	1
	No	505
Chronic kidney disease (CKD)	yes	9
	No	497
Percutaneous Coronary Intervention (PCI)	yes	17
	No	489
Rheumatoid Arthritis (RA)	yes	10
	No	496
Supraventricular Tachycardia (SVT)	yes	2
	No	504
Idiopathic Dilated Cardiomyopathy (IDCM)	yes	1
	No	505
Negative Cardiac Failure (CCF)	yes	1
	No	505
Rheumatoid Heart Disease (RHD)	yes	1
	No	505

Feature Names	Subcategories	Data Distribution
Symptomatic Bradycardia (SB)	yes	3
	No	503
End Stage Renal Disease (ESRD)	yes	1
	No	505
Old Myocardial Infarction (OMI)	yes	6
	No	500
Feline Viral Rhinotracheitis (FVR)	yes	4
	No	502
Respiratory Tract Infection (RTI)	yes	2
	No	504
Hemorrhagic Shock	yes	1
	No	505
Aortic Stenosis	yes	1
	No	505
Coronary Artery Bypass Graft (CABG)	yes	15
	No	491
Survivor Cardiac Arrest (SCA)	yes	3
	No	503
Valvular Heart Disease (VHD)	yes	1
	No	505
Right Ventricular Involvement (RVI)	yes	1
	No	505
Pneumonia	yes	1
	No	505
Atrioventricular block	yes	1
	No	505
Sinus rhythm	yes	1
	No	505
Normal Epicardial Coronary Artery	yes	11
	No	495
Surviving Cardiac Arrest	yes	3
	No	503
Ventricular Tachycardia	yes	1
	No	505
Hypothyroidism	yes	1
	No	505

Feature Names	Subcategories	Data Distribution
Ventricular Bigeminy	yes	1
	No	505
Vertigo	yes	3
	No	503
Atrial Fibrillation	yes	5
	No	501
Acute Kidney Injury (AKI)	yes	10
	No	496
Right Bundle Branch Block (RBBB)	yes	1
	No	505
Valvular Cardiomyopathy	yes	2
	No	504
Triple Vessel Disease (TVD)	yes	3
	No	503
Deep Vein Thrombosis (DVT)	yes	2
	No	504
Anterior Wall Myocardial Infarction (AWMI)	yes	15
	No	491
Left Anterior Descending (LAD)	yes	1
	No	505
Congenital RCA Aorta Fistula	yes	1
	No	505
Congestive Hepatitis	yes	2
	No	504
Renal Artery Stenosis	yes	3
	No	503
Psychosis	yes	1
	No	505
Hypothyroidism	yes	1
	No	505
Normal Coronasies	yes	1
	No	505
Mild Pulmonary Regurgitation (PR)	yes	1
	No	505
Urosepsis	yes	1
	No	505

Feature Names	Subcategories	Data Distribution
Left Bundle Branch Block (LBBB)	yes	4
	No	502
Ermanent Pacemakers (PPM)	yes	4
	No	502
Anaemic Heart Failure	yes	1
	No	505
Chronic Obstructive Pulmonary Disease (COPD)	yes	3
	No	503
Electrolyte Imbalance	yes	1
	No	505
Ischemic Hepatitis	yes	1
	No	505
Moderate Multiple Sclerosis	yes	1
	No	505
Metabolic Encephalopathy	yes	1
	No	505
Severe Musculoskeletal Pain	yes	1
	No	505
Vertebral Artery Stenosis	yes	1
	No	505
Subacute Intestinal Obstruction	yes	1
	No	505
Diabetic ketoacidosis (DKA)	yes	1
	No	505
Pulmonary Arterial Hypertension Phenylalanine Hydroxylase (PAH)	yes	2
	No	504
Hypokalemia	yes	1
	No	505
Dyselectrolytemia	yes	4
	No	502
Symptotic Nodal Bradycardia with STMP	yes	1
	No	505
Asymptotic Nodal Bradycardia with STMP	yes	1
	No	505

Feature Names	Subcategories	Data Distribution
Marfan Syndrome	yes	1
	No	505
Cough Variant Asthma	yes	1
	No	505
Plain Old Balloon Angioplasty (POBA)	yes	1
	No	505
Hyponatremia	yes	1
	No	505
Bronchial Asthma	yes	4
	No	502
Severe Anaemia	yes	1
	No	505
Leptin Replacement Therapy (LRT)	yes	1
	No	505
Anaemia	yes	1
	No	505
Urinary Tract Infection (UTI)	yes	2
	No	504
Pleural Effusion	yes	1
	No	505
Psoriasis	yes	1
	No	505
Left main stem (LMS) disease	yes	1
	No	505
Hypoxic Ischemic Encephalopathy	yes	1
	No	505
Renal Impairment	yes	3
	No	503
Hepatitis B Surface Antigen (HBsAG)	yes	2
	No	504
Cardiovascular Disease (CVD)	yes	2
	No	504
Benign Enlargement of Prostate (BEP)	yes	2
	No	504
Dyslipidemia	yes	1
	No	505

Sample Scanned Copy of the Datasheet

✓

RISK FACTORS FOR CAD

25

Date : 28.04.2016 Reg No : 1504656 MRD No : 655 Time : 8:15 pm

Patient's Name : Mr. Sudhir Das Age : 80 Years Sex : Male

Adress : Digholia, Khulna. Contact : 01735475810

Profession : N/A Date Of Admission : 31.10.2015

Time Of Admission : 3.40 pm D/O Discharge : 05.11.2015

Hight : 150 cm Weight : 65 kg BMI : .... kg/m²

HR : 110 /min BP : 100/50 mm of Hg

Diagnosis 1	AMI (Anterior)
Diagnosis 2	CAD - SVD
Diagnosis 3	S/P Primary PCI
Diagnosis 4	PTCA, CAD
Diagnosis 5	
Diagnosis 6	

RISK FACTORS:	Yes/No	Duration	Remarks
Family History Of Cardiovascular Disease	No	—	
Smoking	No	—	
H/O Tobaco Chewing	No	N/A	
Hypertension	No	N/A	22 medicin S
Dyslipidemia (DLP)	No	N/A	
Diabetes Mallitus (DM)	No	N/A	
Physical Inactivity	No	—	
Psychological Stress	No	—	
H/O Stroke	No	—	

PAST HISTORY :

CAG	CABG	Heart Attack	Psychological Disorder
No	No	No	No

Presentations	Yes/No	Duration	Remarks
Chest Pain	Yes	—	
Dyspnea	No		
Palpitation	No		
Syncope	No		

Investigations	Value		Value
Hb %	12.8 gm/dL	Troponin I	N/A
Sr. Creatinine	1.5 mg/dL	EF	45 %
RBS	8 moml/L	Triglyceride	N/A
Total Cholestrol	N/A	LDL	N/A
HDL	N/A	ECG	SR, T↓ I, aVL

Others :

TIMI Score	STEMI	NSTEMI	Remarks
Coronary Angiogram (CAG)	Yes		
Percutaneous Coronary Angioplasty (PTCA)	Yes		LAD

Outcome : ....

For Technical Use :

Agreement Copy

An initiative by AFC HEALTH LTD. A Concern of Active Fine Chemicals Limited | Fortis Escorts HEART INSTITUTE

Date: 28/03/2016

To,

Mr Saikat Mondal,

Asst Professor & Supervisor,

Computer Science and Engineering Discipline, Khulna University.

This is in reference to the letter received, requesting collaboration for the research on "Smartphone based Heart diseases Prediction using clinical data and Data Mining Approaches", as a project of Department of Computer Science and Engineering, Khulna University, Khulna, Bangladesh.

From Mr Saikat Mondal, Mr Raihan and Mr Shagor, we understood that aim of this study is to design a mobile application which will help the general population to suspect any heart ailment they may have as existing or have chances to develop in future.

We, AFCH Fortis Escorts Heart Institute, Khulna, agree to collaborate with Khulna University with following terms and conditions.

1. AFCH Fortis Escorts Heart Institute (AFCHFEHI) will provide clinical information.

2. The data analysis and correlation between the risk factors and diagnosis will be done by the doctors at AFCH FEHI.

3. Patients personal information will be kept completely confidential and by no means it will be disclosed to the researchers.

4. Researchers (From University & AFCH FEHI) will not be allowed to take any scan or photocopy of the clinical records without permission.

5. No medical records will be allowed to carry out of hospital.

6. AFCH FEHI will have its right to share its name with the name of application.

7. The end product will be solely for the purpose of research project and it will not be utilized for commercial purposes without consent from AFCH FEHI.

A-17 Majid Sarani, Sonadanga, Khulna | ☎ +88 041 730009, 733703 | fehikhulna@afchealthbd.com

An initiative by AFC HEALTH LTD. A Concern of Active Fine Chemicals Limited | Fortis Escorts HEART INSTITUTE

8. Any commercial interest generating out of the app will be shared with AFCH FEHI and Researchers cannot make any commercial gains without written permission from AFCH FEHI.

9. Any research paper, published in any journal related to this research will include names of the consultants and researchers from AFCH FEHI.

10. AFCH FEHI has right to publish papers; publish abstracts or any scientific literature out of the research in collaboration with Khulna University but not independently.

11. AFCH FEHI will have right to mention hospitals name in the app which may suggest the users to go for their health checkups in AFCH FEHI Hospitals.

13. AFCH FEHI will have access to the technical details of app however, AFCH FEHI agrees, no changes will be done without mutual consent.

14. The expenses occurring towards the research will be borne by department / university.

15. No Financial disclosures allowed from either side.

16. Both institutes preserve the right to publish research papers / research data at local / national / international level.

Accepted

12.05.2016

Assistant Professor
Computer Science & Engineering Discipline
Khulna University, Khulna-9208, Bangladesh

28/3/16.

Dr Arun Ramrao More

MBBS, PGDCC (Bangalore), MBA
Consultant Cardiologist, Non Invasive Cardiology, FEHI- New Delhi.
Facility Director - Head Management
AFCH Fortis Escorts Heart Institute,
Khulna, Bangladesh.
e-mail: arunmoregsa@gmail.com
arunmore@afchealthbd.com

DR. ARUN R MORE
MBBS, PGDCC (Bangalore)
Consultant Cardiologist (FEHI NEW DELHI)
Facility Director
Fortis Escorts Heart Institute Khulna.

A-17 Majid Sarani, Sonadanga, Khulna | +88 041 730009, 733703 | fehikhulna@afchealthbd.com

Table A5: Dataset with Mostly Significant Features (Partial)

sex	**bp-systolic**	**ACS**	**IWMI**	**unstable-angina**	**severe-multiple-sclerosis**	**non-stemi**	**AMI**	**PCI**	**rhenumated-arthitis**	**DVD**	**thrombolysed-stk**	**valvular-cardiomyopathy**	**CAD**	**SVD**	**AWMI**	**PTCA**	**bronchial-asthma**	**smoking**	**stroke**	**sweating**	**vomiting**	**chest-pain**	**creatinine**	**EF**	**ECG**	**CAG**	**patients_status**	**IHD**
male	110	yes	no	no	no	yes	no	no	no	no	no	no	no	no	no	no	no	ex	no	no	no	yes	1.3	50	normal	yes	discharged	yes
male	110	yes	no	no	no	yes	no	no	no	no	no	no	no	no	no	no	no	no	no	yes	no	yes	1	25	abnormal	no	discharged	yes
male	160	no	no	yes	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	yes	no	yes	1	59	abnormal	no	discharged	no
male	120	yes	no	no	no	no	yes	no	no	no	no	no	no	no	no	no	no	yes	no	no	yes	yes	1.724718	49.17756	abnormal	no	expired	yes
male	160	yes	no	no	no	no	no	no	no	no	no	no	no	no	yes	yes	no	ex	no	no	no	yes	1.6	30	abnormal	yes	discharged	yes
male	140	yes	no	no	no	no	no	no	no	no	no	no	no	no	no	yes	no	yes	no	yes	no	yes	0.8	49.17756	abnormal	yes	discharged	yes
female	131.4692	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	yes	3.1	49.17756	abnormal	no	discharged	no
male	120	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	yes	1.724718	49.17756	abnormal	no	discharged	no
female	220	yes	no	no	no	no	no	yes	no	no	no	no	no	no	no	no	no	no	no	yes	no	yes	1	69	abnormal	no	discharged	yes

female	female	male	female	male	male	male	male	male	male	male
160	140	160	130	150	80	131.4692	90	120	160	150
yes	no	no	no	no	no	yes	no	yes	yes	no
no	no	no	no	no	no	no	no	no	no	no
no	no	yes	yes	yes	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	yes	yes	no
no	no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no	no
yes	no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	yes	no	no	no	no
no	no	no	no	no	no	no	no	no	yes	no
no	no	no	no	no	no	no	no	no	no	no
no	no	ex	no	yes	no	yes	yes	no	no	ex
no	no	no	no	no	yes	no	no	yes	yes	no
no	no	no	no	yes	no	yes	yes	no	no	no
no	no	no	no	yes	yes	no	yes	no	no	no
yes	no	yes	yes	no	yes	yes	yes	yes	yes	yes
1	1.724718	1	1.9	1.6	1.9	1.724718	1.7	1.4	1.8	0.9
54	62	59	52	44	49.17756	49.17756	53	50	35	65
abnormal	normal	abnormal	abnormal	abnormal	abnormal	abnormal	abnormal	abnormal	abnormal	normal
yes	no	no	no	no	no	no	no	no	no	yes
discharged	discharged	discharged	discharged	discharged	discharged	expired	discharged	discharged	discharged	discharged
yes	no	no	no	no	no	yes	no	yes	yes	no

male	male	male	female	male	male	female	male	male	male
120	110	160	210	131.4692	140	150	131.4692	130	130
yes	no	yes	yes	yes	yes	yes	yes	yes	no
yes	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	yes	no	no	no
no	no	no	no	no	no	no	no	no	no
yes	no	no	no	no	no	yes	yes	no	no
no	no	no	no	no	no	no	no	no	no
yes	no	no	no	no	yes	no	yes	no	yes
no	no	yes	no	no	no	no	no	no	yes
yes	no	no	no	no	no	yes	no	no	yes
no	no	yes	yes	yes	no	yes	no	no	no
yes	yes	yes	yes	yes	yes	yes	yes	no	yes
1.3	1	1.9	1.5	2.6	1.724718	0.8	1	1.4	0.8
49.17756	49.17756	50	35	49.17756	25	60	42	49.17756	50
abnormal	abnormal	abnormal	abnormal	abnormal	abnormal	abnormal	abnormal	abnormal	abnormal
yes	no	yes	yes	yes	no	yes	yes	yes	no
discharged	discharged	discharged	discharged	expired	discharged	discharged	discharged	discharged	discharged
yes	no	yes	yes	yes	yes	yes	yes	yes	no

List of Publications

[1]. **M. Raihan**, M. Islam, P. Ghosh, S. Shaj, M. Chowdhury, S. Mondal and A. More, "A Comprehensive Analysis on Risk Prediction of Acute Coronary Syndrome Using Machine Learning Approaches", in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2018, pp. 1-6.

[2]. **M. Raihan**, P. K. Mandal, M. Islam, T. Hossain, P. Ghosh, S. Shaj, A. Anik, M. Chowdhury, S. Mondal and A. More, “Risk Prediction of Ischemic Heart Disease Using Artificial Neural Network”, in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox’s Bazar, Bangladesh, 2019, pp. 1-5.