SCHOOL OF GRADUATE STUDIES

EARLY DETECTION OF HEART DISEASE: ENHANCING PREDICTION THROUGH MACHINE LEARNING TECHNIQUES

MSC THESIS

SIRAGE TEMAME AREB

JANUARY, 2024

WOLKITE, ETHIOPIA

Wolkite University

School of Graduate Studies

Early Detection of Heart Disease: Enhancing Prediction through Machine Learning Techniques

A Thesis Submitted to the School of Graduate Studies, in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science and Engineering (Specialization: Computer Science)

Sirage Temame

Major Advisor: Mesfin Abebe (Ph.D)

Co-Advisor: Jemal Ahmed

January, 2024

Wolkite, Ethiopia

APPROVAL SHEET

SCHOOL OF GRADUATE STUDIES

WOLKITE UNIVERSITY

"Early Detection of Heart Disease: Enhancing Prediction through Machine Learning Techniques"

Submitted by:-

Mr. Sirage Temame Areb	_____	_____
Name of Student	Signature	Date

Approved by:-

1. Dr. MESFIN ABEBE		05/02/2024
Name of Major Advisor	Signature	Date

2. Mr. JEMAL AHMED	_____	_____
Name of Co-Advisor	Signature	Date

3. _____	_____	_____
Name of Chairman, DGC	Signature	Date

4. _____	_____	_____
Name of Dean, SGS	Signature	Date

WOLKITE UNIVERSITY

SCHOOL OF GRADUATE STUDIES

We hereby certify that we have read and evaluated this Thesis **"Early Detection of Heart Disease: Enhancing Prediction through Machine Learning Techniques"** prepared under our guidance by Sirage Temame Areb. We recommend that the Thesis shall be submitted as a fulfilling the requirements for the award of a MSc. degree in Computer Science and Engineering.

1. Dr. MESFIN ABEBE  05/01/2024

   Major Advisor Signature Date

2. _____ _____ _____

   Co-Advisor Signature Date

As members of board of Examiners of Masters of Science Thesis open defense examination, we have read and evaluated this Thesis prepared by Sirage Temame Areb and examined the candidate. We hereby certify that, the thesis is accepted for fulfilling the requirements for the award of the degree of Masters of Science (MSc) in Computer Science and Engineering (specialization: computer Science).

1. Baye Y. (Ph.D.)_____  01-05-2024

   Name of the External Examiner Signature Date

2. _____ _____ _____

   Name of Internal Examiner Signature Date

3. _____ _____ _____

   Name of Chairman Signature Date

Final approval and acceptance of the Thesis is contingent up on the submission of its final copy to the Council of Postgraduate Program (CPGS) through the candidate's department or school graduate committee (DGC or SGS).

DECLARATION

I confirm that this thesis is my original work by signing below. I have prepared, gathered, analyzed, and compiled this thesis in accordance with all ethical and technical scholarly standards. All academic material incorporated into the Thesis has been acknowledged via citation. This thesis is turned in to Wolkite University as a partial fulfillment of the requirements for a Master's Degree. The thesis is placed in the Wolkite University Library and is accessible to users in accordance with the library's policies. I hereby declare that this thesis has not been submitted for consideration for any academic degree, diploma, or certificate to any other University or Institution. It may use brief quotes from this thesis without obtaining permission as long as you provide due credit to the original author. The Head of the Department may grant permission for extensive quotes from this Thesis or for its reproduction in whole or in part if the use of the material is believed to be in the interest of research. But in every other case, agreement needs to be held from the Thesis Author.

Name:-_____

Signature:-_____

Date:-_____

Department:-_____

ACKNOWLEDGEMENTS

First of all I would like to thanks my creator and sustainer Allah for all his blessings from the start to the end of this research work. Next to that I am exceptionally grateful to my main advisor Dr.Mesfin Abebe for his friendly approach, commitment to advice and guide, invaluable advice, very critical comments and continuous supports that greatly improved this thesis work. Similarly, I would like to thanks my Co-advisor Mr. Jemal Ahmed for his friendly advice, support and guide, comments appreciations in order to complete my thesis work throughout the time of study. I would also like to thanks some of my friends and staff members in Wolkite University specialized and teaching hospital, the coordinators in outpatient ward staff members, especially Dr. Mustejab, for his special advice and commitment to support me in giving real and detail information in the process of collecting data.

LIST OF ACRONYMS

AB	Adaptive Boost Algorithms
AB-WAE	Accuracy Based Weighted Aging Classifier Ensemble
AIDS	Acquired Immune Deficiency Syndrome
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
BADT	Back ward selection Adaptive Boosting and Decision Tree (Sequential)
BANB	Back ward selection Adaptive Boosting and Naïve Bayes (Sequential
BBDT	Backward selection Bagging and Decision Tree (Sequential)
BBNB	Backward selection Bagging and Naïve Bayes (Sequential)
BC	Breast Cancer
BFS	Backward Feature Selection
BN	Bayes Net
Ca	Colored Vessels
CART	Classification and Regression Tree
CFS	Chi-square Feature Selection
CHD	Congenital Heart Disease
CHUC	University Hospital Center For Coimbra
CLBSHS	Cleveland Long Beach Switzerland and Hungarian
CM	Confusion Matrix
CPU	Central Processing Unit
CSA	Central Statistics Agency
CSV	Comma Separated Values
CV	Cross Validation
CVD	Cardio Vascular Disease
DT	Decision Tree
EDHS	Ethiopian Demographic and Health Survey
FADT	Forward selection Adaptive Boosting and Decision Tree (Sequential)
FANB	Forward Selection Adaptive Boosting and Naïve Bayes (Sequential)

FBDT	Forward Selection Bagging and Decision Tree (Sequential)
FBNB	Forward selection Bagging and Naïve Bayes (Sequential)
FFS	Filter Feature Selection Method
FN	False Negative
FP	False Positive
FS	Feature Selection Method
GB	Gradient Boosting
GBS	Giga Bites
GHz	Giga Hearth
HD	Heart Disease
HIV	Human Immune Virus
HRFLM	Hybrid Random Forest and Linear Model
HTML	Hyper Text Markup Language
IGFS	Information Gain Feature Selection
IHD	Ischemic Heart Disease
KNN	K Nearest Neighbors
LASSO	Least Absolut Shrinkage and Selection Operator
LM	Linear Modeling
LR	Logistic Regression
MDP	Markov Decision Process
ML	Machine Learning
MLP	Multi-Layer Perceptron
NB	Naïve Bayes
NCM	Normalized Confusion Matrix
OADT	Optimization Selection Adaptive Boosting and Decision Tree
OANB	Optimization Selection Adaptive Boosting and Naïve Bayes
OBDT	Optimization Selection Bagging and Decision Tree
OBNB	Optimization Selection Bagging and Naïve Bayes
PART	Projective Adaptive Resonance Theory
PN	Prime Number
PS	Percentage Splitting

RF	Random Forest
RFEM	Recursive Feature Elimination Method
RHD	Rheumatic Heart Disease
SFFS	Sequential Forward Feature Selection
SMOTE	Synthetic Minority Oversampling Technique
SVC	Support Vector Classifier
SVM	Support Vector Machine
TASH	Tikur Anbesa Specialized Hospital
TN	True Negative
TP	True Positive
UCI	University of California Irvine
VC	Voting Classifier
WFS	Wrapper Feature Selection Method
WHO	World Health Organization
WKUTRH	Wolkite University Teaching and Referral Hospital
XGB	Extra Gradient Boosting Algorithms

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Heart disease is the abnormal health condition that influences parts of the heart and all its parts. World Health Organization is assured that the disease is one of the leading killer disease of the worldwide population. The prevalence of the disease is also increasing through developing countries like Ethiopia. Machine Learning (ML) is one of the key technique in the management and processing of a huge number of health data's and it supports in diagnosis and prediction of disease at early stages. The main objective of this study is developing an early detection of Heart Disease enhancing prediction through ML technique; such as Random forest (RF), K Nearest Neighbor, Support vector Machine (SVM), Gradient Boosting (GB) and Voting Classifier with two Feature Selection (FS) methods, of Chi-Square (CFS) and Sequential Forward Feature Selection (SFFS) methods. The data used for the experimentation purpose was collected from Public repositories and Local Hospitals. Thus, these datasets are accessed to develop the proposed model in combined and in separate way. Before FS methods are performed, all the ML algorithms are applied for the three imbalanced and balanced HD datasets. Then after, the two FS methods are applied with ML techniques on the three imbalanced and balanced datasets. Models are evaluated through different model evaluation metrics with two data splitting technique namely Percentage Splitting (PS) and 10-Fold-Cross Validation (10-F-CV) techniques and finally different results are registered. Thus, before FS methods are applied on the balanced datasets, ***SVM and GB*** *achieved a good accuracy score of* ***99.2%*** *using PS and similarly after FS technique is applied,* ***RF with CFS*** *achieved a better accuracy score of* ***99.5%*** *using PS for the combined dataset. Finally,* ***RF with CFS*** *model is saved and deployed with in flask server to show the prototype of the prediction model and this may help users and experts to detect and appropriate prevention of the disease at early stage.*

Keywords: - *Chi-square feature selection, Heart Disease detection****,*** *Machine Learning*

CHAPTER ONE

INTRODUCTION

1.1. Background of the Study

The heart is the most important organ of the human body because it pumps our blood and circulates to the entire body. This organ is encircled by double-layered tissue membranes and shielded by the rib cage. The heart is a four chambered organ that uses five different types of blood vessels such as veins, arteries, capillaries, arterioles, and venules, to divide blood into oxygenated and deoxygenated circulation. The heart is the main organ of the human body since it siphons or pumps our blood and flows to the whole body [1]. The word Heart disease refers to a broad range of illnesses. These medical conditions refer to the pathological states that directly affect the heart and all of its components. In every days life of human being Heart Disease (HD) is the main cause of death in the whole world. World Health Organization (WHO) predicted that at least 12 million deaths occurred worldwide. In every year due to HD greater than 80% of deaths occurred in the world. WHO predicted that in future almost 23.6 million peoples will die because of HD[2]. HD is a common deadly condition and it current the number one killer disease of the global population. The other report of WHO states that cardiovascular (CVD) kills 17.9 million peoples every year accounting for about 32% of the world death. This report also states that HD and stroke are the leading causes of CVD accounting about an approximate number of 85% of deaths, where the age of the people was under 70 and accounting one-third of all premature deaths [3]. This report also states that, 17.3 million deaths caused by HD in 2008, approximately 6.2 million were due to stroke, and predicted that 7.3 million people deaths were due to coronary heart disease (CHD). WHO also estimated that approaches to 23.6 million people will die due to HD and stroke-related disease by 2030[3]. Another report shows that almost one-third of the population of the world were died in developing countries in 2010 [4]. Additional report in [5] states that Cancer, Chronic respiratory Disease, CVD and Diabetes mellitus diseases are on increasing and the leading risks of human health and development. These causes accounts around 35 million deaths each year and 85 % are in developing countries including Ethiopia [5]. WHO additionally reported

that, in 2014, around 30% of the people of Ethiopia are died due Non-Communicable Diseases (NCD) of which CVD contributes 9% [5].

Another study from 2021 indicates that the primary causes of CVD deaths in Ethiopia were various forms of HD, with approximately 170 Ethiopians dying every day. Ischemic heart disease (IHD) accounted for 45% of these deaths, heart stroke for 34% and hypertensive HD for approximately 11% [6]. All the reports of these studies shows that the prevalence of HD in Ethiopia is highly increasing with respect to the prevalence of the disease around the world. Similar to diseases of the circulatory system, high blood pressure, smoking, diabetes, and physical inactivity are just a few of the numerous causes of CVD, which includes HD. Up until now, research has always been concentrated on strategies to lower the number of deaths from cardiac related illnesses. Approximately 90% of cardiac related illnesses can be prevented, according to studies [7], [8].

As we know, any medical conditions or any disease have its own risk factors, so cardio vascular disease including heart disease also have its own risk factors. These includes age, Sex, smoking family history, poor-diet, cholesterol level, High blood pressure, physical inactivity, obesity and alcohol are considered to be the risk factors for heart disease(HD) and other hereditary risk factors such as high blood pressure and diabetes had leads to heart disease, however, some of the risk factors are controllable [8].Heart disease has a wide range of symptoms; some may experience weariness or fatigue and chest discomfort, while nearly 50% of people have no symptoms at all until they have a heart attack. WHO has statistical data indicating that a high risk of abnormalities or permanent disability can arise from CVD in a number of men and women[9].

There are various kinds of HD conditions, some of them are as follows:- Coronary or valvular refers to damage to the blood pumping vessels that prevents blood from reaching the heart, An elevated blood pumping condition toward arterial walls is indicated by hypertension. Cardiac arrest is a sign of an unexpected malfunction with the heart's operation and awareness. Heart failure is the result of the heart failing to pump blood. An arrhythmia is an abnormal heartbeat, such as one that is too rapid or too slow. A peripheral artery denotes a situation where blood is being pumped to the limbs from constricted vessels. A stroke is characterized by disruption of the blood flow to the brain. Congenital refers to an anomaly that exists in the heart prior to

birth. Heart disease can also result from any combination of the aforementioned few reasons.[10]. Since HD has a complicated nature, it must be managed carefully because failing to do so could harm the heart or result in an early death. So, proper and accurate detection and diagnosis of the HD risk in an expected patient is necessary for reducing their associated risks of severe heart issues and improving security of heart [11]. Thus, in this study, we have tried to include several heart conditions, such as heart failure-, chronic heart failure, congenital HD, ischemic heart condition, heart stroke and rheumatic heart condition are few of these conditions.

The goal of Machine learning (ML), a branch of research that lies at the nexus of computer science, artificial intelligence (AI), and statistics, is to extract knowledge from data. ML is sometimes referred to as statistical learning or predictive analytics [12]. The primary objective of ML is teaching machines or computers to complete undertakings tasks by giving them several models with the idea of how to do or not complete a certain tasks [13]. One of the common application of ML is the prediction of an outcome based upon existing data [12]. ML is a branch of data mining that efficiently deals with large-scale, well-formed data sets. Various ML algorithms like Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), K Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF) and Ensemble technique XG Boost. compared to finding the most accurate model [14]. In the medical field, ML can be used to diagnose, detect and predict various diseases [15]. According to reports, machine learning (ML) is a rapidly expanding field in the medical diagnosis sector, where computer analysis can reduce manual error and improve accuracy. With computer and machine learning techniques, the diagnosis of a disease is extremely reliable. ML concepts are used to predict diseases like diabetes, liver disease, heart disease, and tumors. Regression algorithms like RF, Lasso, and LR were employed in the medical field, and ML algorithms like DT, NB, and SVM are accessible. ML is a field of AI and it can be subdivided in to supervised ML, Unsupervised ML, Semi supervised ML, Ensemble learning and Reinforcement learning.

This field is becoming a very interesting and applicable research area in the fields of medication and health sectors in several disease prediction and detection using patient history data with a highly performing classifier. Predictive analysis with the help of efficient and numerous ML techniques helps to predict the disease faster, accurately and correctly helping

to in treating of patients within a few second and reduce waiting time[15]. Basically this research paper focused on heart patients history data records collected from Public online repository named as Heart Disease Dataset | Kaggle holding four separated datasets in combined form, namely cleveland_hungary_switzerland_and_long_beach and local dataset's collected from Wolkite university referral hospital. Actually public datasets collected from public online repositories have 14 basic features such as Age, Chest pain type, Rest blood pressure, Sex, Fasting blood sugar, Cholesterol, Rest Eco cardiograph, Thalach or Maximum heart rate, Exercise induced angina, Slope of the peak exercise, Oldpeak, Ca or number of major vessels and Thal or Defect type are all used to classify the patient as having HD or not having HD. As most of local patient datasets found in local health institutions in Ethiopia are not properly configured and found in manualized and paper format. We have tried to configure the collected local hospital HD patient datasets in the standard form of public datasets collected from in order to get similar setup. We have applied and compared several classifiers with FS methods based on their prediction performance. Finally the better model is recommended for the prediction of HD.

1.2. Motivations of the Study

Despite being preventable, cardiovascular disease (CVD) accounts for approximately 31% of all deaths worldwide, with over 3 million deaths occurring in people under the age of sixty. More than 80 percent of deaths linked to CVD occurred in low- and middle-income nations, such as Ethiopia. The global disease burden report from 2015 states that in many of the world's poorer regions, a greater percentage of deaths are attributable to cardiovascular disease (CVD) due to population growth and aging. Compared to Western and Southern Sub-Saharan Africa, the disease is more common in Eastern and Central Sub-Saharan Africa [5]. According to a 2014 WHO report, non-communicable diseases claimed the lives of about 30% of Ethiopians, with cardiovascular disease accounting for 9% of these deaths [5]. In Ethiopia there is almost a little study is conducted to predict HD risk but related works such as a study proposed by Hana et.al [14], was aims in developing and recommending a ML technique for the prediction of Chronic Kidney Disease (CKD) and also in Ethiopian local hospitals, there is almost a little ML based predicting models in which helps health experts use to classify the patients record having the expected disease or not faster. As we have mentioned before this topic HD is very

risky. However, studies show that about 90% of HD can be prevented [7]. ML classifiers can analyze medical data to predict the severity of the disease correctly by simply inserting common clinical predictors or features of patient history data. Here we have be inspired to work on the improvement of HD detection using ML techniques.

1.3. Statements of the Problem

Non-communicable diseases (NCDs) are largely to blame for the deaths and disabilities of millions of people worldwide, irrespective of population size, economic development, or social progress [16]. For a number of years, low-income nations like Ethiopia had not regarded NCD as a serious problem [16]. One of the NCDs that kills millions of people worldwide, including a high death rate in Ethiopia, is HD and the global disease burden report from 2015 states that in many of the world's poorer countries, including Ethiopia, the proportion of deaths due to CVD has increased due to population growth and aging [5]. Additionally according to a systematic review conducted in Ethiopia in 2021, found that the prevalence of CVD ranges from 7.2% to 24% [5], which shows the risk of the disease is highly prevalent in poorer regions of the country. According to reports, the high mortality rate and prevalence of HD in recent decades pose a serious threat to people's health [7]. Misdiagnosis and time-consuming with manual analysis of large HD-related datasets are the drawbacks [7] in the prevalence of the disease, this means being manual paper based analysis of patient’s data may increase the risk of the diseased patient because this may increase the detection and prediction time period.

Nowadays, the majority of medical diagnoses and treatments in Ethiopia are using antiquated or traditional laboratory testing and ineffective healthcare systems that manually process and produce clinical trials' outcomes [17]**.** Because of this, it is quite uncommon in Ethiopia to be able to access prompt and trustworthy medical care. In developing countries like Ethiopia early prevention of disease is almost none because of lack of medical resources [16], lack of health professionals specialized on that fields, lack of early diagnosis and treatment ordering and also lack of heart disease predicting tools helping the health care experts. According to a number of recent researches to overcome such limitations, ML approaches have demonstrated promising outcomes in the diagnosis and detection of cardiac conditions[17] and other various diseases [15], Specifically, complex nonlinear relationships can be learned from the training datasets using supervised learning approaches and algorithms and these algorithms are also

comparing in different areas such as disease diagnosis, detection or prediction and finds the most accurate models [14]. Thus, Various ML algorithms including Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), K Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF) and Ensemble technique XG Boost are compared in different disease diagnosis, detection and prediction areas. For example, C.Beulah Christalin Latha [18] has proposed several techniques and proved that an Ensemble technique (VC) as a good model for the prediction of HD risk. Ibomoiye Domor et.al [22], Saiyed Faiayaz Waris and S. Koteeswaran [9], Senthilkumar Mohan et.al [19], G. Ramesh et.al [20], Mohammad Shafenoor et.al [21], Hidayet Tacki [22], Devansh Shah et.al [23] and Nikhil Bora [24], all was proposed different Supervised and Ensemble ML based strategies to handle the risks of HD and proved that ABWAE, IKNN, HRFLM, RF & SVM, VC, SVM-linear, KNN and RF were the respective and the best performance achiever algorithms with and without applying FS methods. However, as it is observed in those works, no local HD datasets was included in all those materials and the performance they achieved using these algorithms with FS methods is not that much good. In addition to this, certain supervised ML and EL algorithms have demonstrated diagnostic and predictive accuracy in health care areas that surpasses of highly skilled medical professionals. However, as noted in [16] and [17], very few studies have been done on the application of ML for the prediction of HD in Ethiopia for a variety of reasons, including a lack of local and suitably configured datasets [17].

In other way, various risk factors are also linked to HD, so early detection of the diseases risk factor is very crucial for prompt management of the disease, necessitating the use of accurate, dependable, and reasonable approaches[25]. When a favorable feature combination is also missing or the algorithms are not used appropriately, the effectiveness of ML algorithms used for CVD prediction is also greatly reduced. Thus, Finding the crucial feature combination method, that works best with the top-performing algorithm is therefore crucial [26]. Therefore, quick and efficient methods for HD detection and prediction methods are needed to reduce death and disability from heart disease. This is especially important in poor nations like Ethiopia, where there is a dearth of cardiac specialists and a high prevalence of the disease with incorrect diagnoses. Therefore, by learning about the condition from patient data, ML can be a useful tool to help doctors diagnosing the disease. Thus, creating and building ML model that can identify and forecast HD at an early stage is necessary. The proposed ML technique

can support physicians to reduce the difficulties, bias as well as other related burdens of medical diagnosis. In general, the contributions of this work is developing a reliable model for early detection of HD; enhancing prediction performance through different supervised and Ensemble ML techniques with two different FS methods, this can help for the identification of the most better ML model with the most informative features for the prediction of HD through the given separated and combined datasets. This will supports health care physicians and ill patients in timely detection of the disease and helps in ordering treatments early in accurate and reliable manner.

1.4. Research Questions

In order to address the problem the following two research questions must be answered:-

Q1. Which machine learning technique is better to predict heart disease?

Q2. Which feature selection method can be best to decrease the dimensionality of the dataset and enhance the classifier performance?

1.5. Objective of the Study

1.5.1. General Objective

The general objective of this study is to develop an early detection of Heart Disease: enhancing prediction through ML technique.

1.5.2. Specific Objectives

In order to achieve the general objective of this study there are different specific objectives mentioned as follows:-

- ✓ To configure necessary data in appropriate way.
- ✓ To identify relevant features for dataset preparation.
- ✓ To develop ML model to detect and predict HD in all available features.
- ✓ To design HD prediction models with two feature selection method.
- ✓ To evaluate and validate the performance of the proposed models.
- ✓ To identify the best ML model and FS method for heart disease prediction.
- ✓ To develop a prototype system that can help for heart disease prediction.

1.6. Scope and Limitation of the Study

1.6.1. Scope of the Study

Although there are several machine learning approaches applied to predict the severity of several diseases such as breast cancer disease, chronic kidney disease, diabetics and acute appendicitis. This study covers a limited number of supervised ML algorithms with few improving FS methods to predict HD. The focus of this study is proposing models for health care institutions in Ethiopia. The HD data which is used to develop the proposed model is collected from public repositories named as cleveland_hungary_switzerland_and_long_beach found in Heart Disease Dataset | Kaggle address and the local hospital heart patient datasets collected from hospitals such as Wolkite University referral hospital heart patient record data. The datasets collected from both in Public repositories and in Wolkite university hospital are used for the development of the model. The model development process was done in two ways. The first one is preprocessing the individual datasets and then combining the two separated datasets and preprocess. Finally developing the individual model in order to make the study more trustful**.** The Local dataset is configured and prepared in the form of public dataset format, having 14 features with categorical and numerical forms of features. The machine learning algorithms used to develop the prediction model are the same for public, local datasets and the combined datasets. SVM, Gradient Boosting (GB), KNN, RF and Voting Classifier (VC) are used as ML algorithms for model development.

1.6.2. Limitations of the Study

This research passes several limitation in the study process. The first very challengeable issue is that we have faced in research process is collecting data in local health institutions in Ethiopia. Some of those health institutions in Ethiopia are not recommend to give the necessary data as fear of privacy and accountability of patient record. Even if they recommend to get the data, it is not well ordered in very good format. The other challenge behind collecting local datasets is, as almost no qualitative data is available in a better software format and almost all health institutions are using traditional shelf of file cabinet or paper folder format. The other limitation that we have been challenged here is as no standard of selecting features related with HD in local hospitals in Ethiopia, beside that of a little researches or studies are done using ML and HD. In addition to these limitations the challenge that we have faced with no timely

contribution of research fee for the study process from the regarding institution, even if the research budget is very interesting issue in conducting a certain study in such program. The other and the main issue that we have faced when the experimental implementation was done is the performance of the personal computer being very poor, made the model development and deployment process very bored. It is because running python programs in such less performing device locally is not recommended.

1.7. Significance of the Study

Heart disease is not only an individual or family problem but also national problem and becoming global problem. Therefore, identifying HD and predicts its risk factors at early stage, supports to limit the risk of the problem early. There is a very wide significance in developing the proposed work in supporting several individuals and groups. The first main significance of this study is helping physicians in providing necessary features which can provides accurate prediction of the disease on the patient and helps experts in early decision making. This produces high accuracy and less classification error rate and prevent the problem easily. Secondly, it provides patient beneficiary improvement through several directions, because if the physician or the health care professionals applied the model effectively, this helps in preventing and controlling of the problem easily and minimize clinical impact outcomes in the patient through reduction of heart failure which leads to the mortality of the expected patient. The other beneficiary in detecting and prediction of HD over a few key parameters can save lives and seeks to achieve greater accuracy than the current methodologies or approaches. Thus, the suggested work also have the advantage of producing a more effective algorithm that can quickly and accurately determine whether a patient is expected to have heart disease or not through the prototype. Additionally, this work will have a little advantage in lowering the cost of the institution's health resources and each patient's budget. Finally, in other researcher side and in system developer side the proposed work will supports and recommends a system developers and researchers to develop better work for future regarding early detection of heart disease through enhanced ways and helps them as a referencing material.

1.8. Organization of the Thesis

This Thesis work is organized in six main chapters and the figure shown in fig 1.8.1 shows the outline of the thesis in an easy to follow manner, this helps readers to give a way of the general appearance of the whole work.

Chapter One: Introduction of the Study:-This chapter describes the Background of the study, motivation of the study, statements of the problem, the research questions to be answered, the general and specific objectives, the scope and limitations, the significance and Organization of the thesis work. Chapter Two: Literature Review and Related Works:-This chapter describes all the literatures reviewed, providing the concepts & information's about heart and HD current status, the ML algorithms and FS methods and other works related with HD are also discussed. Chapter Three: Research Methodology:-This chapter discusses all the research methodologies concerning data sources and data collections, the developing process methods of the prediction models like data preprocessing methods, FS methods, methods of model performance evaluations metrics the implementation and deployment environment are also discussed. Chapter Four: The Implementation:-This chapter focusses on the discussion of implementation and experimentation, such as data preprocessing implementation, ML model development implementation, feature selection implementation, model performance evaluation testing implementation. Chapter Five: Results Discussions: - discusses all the data collection and data preprocessing results, model building and evaluation results on the original full featured datasets and the results of feature selection methods are discussed. Chapter six: Conclusion, Recommendation and Future Works:-This chapter mainly focusses on the conclusion of the study, the recommendation and the future recommendations regarding the study area for researchers or developers. The figure shown in fig 1.8.1 below shows the outline of the thesis in an easy to follow manner.

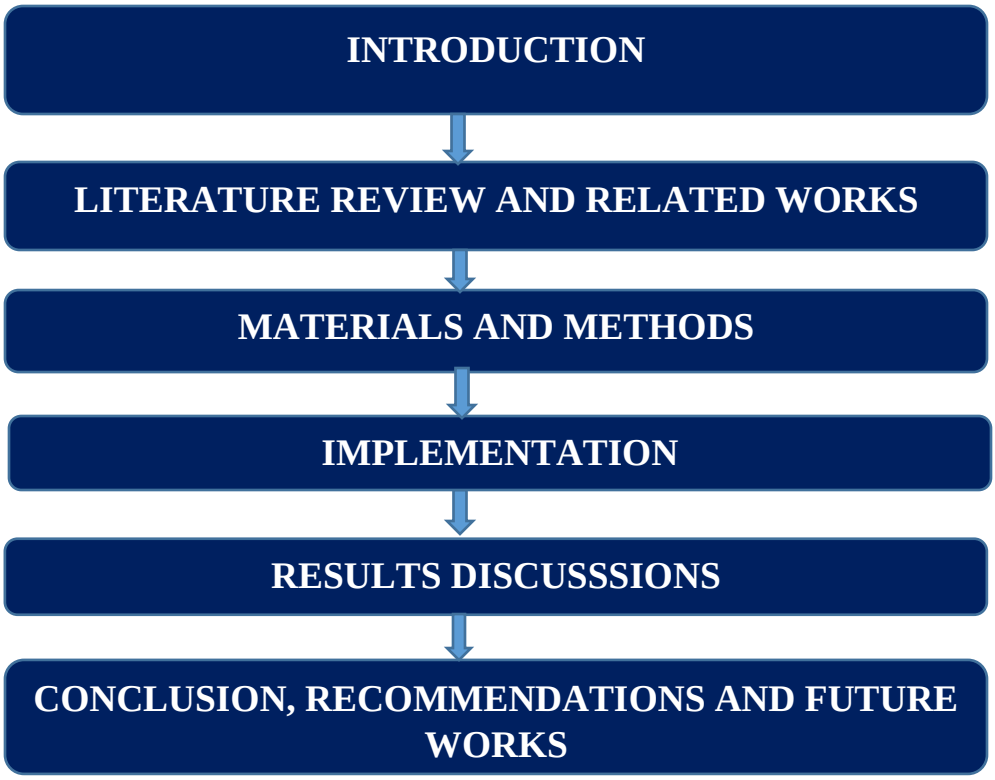

Figure 1.8.1.Organization of the thesis

CHAPTER TWO

LITERATURE REVIEW AND RELATED WORKS

In this chapter we have tried to review all the related literature works on the concepts of heart disease and machine learning. Firstly we have discussed the concepts behind the prevalence of heart disease around the world including Ethiopia. Some of the main types of heart disease and its risk factors are also discussed. Secondly, we have also discussed ML and types of Machine Learning such as Supervised ML, Unsupervised ML, Reinforcement Learning, Ensemble learning and Semi-supervised learning concepts. We have also discussed the relationship between ML and health care. Finally, some related works are also discussed with their findings.

2.1. Heart Disease and Its Prevalence

The heart, which pumps blood throughout the body, is the most vital organ in humans. This organ is encircled by two layers of tissue membranes and shielded by the rib cage. The four chambered organ that makes up the heart divides blood into oxygenated and deoxygenated [1]. The human heart is roughly the size of a fist, according to experts, and it contains five major types of blood vessels: arteries, veins, capillaries, arterioles, and venules [3]. Even though the heart is a vital and significant organ in the human body, there are some conditions that can affect it, ranging from minor damage to the patient's death from a known illness. Heart disease can be called silent killer disease around all over the world because its prevalence is highly distributed everywhere in the world including developed and under developing nations including Ethiopia. Every year a million peoples are dying of this disease. If heart disease is not prevented early by predicting the main contribution factors and ordering treatments for the patient who is expected to be diseased, this leads the condition highly risky and more prevalent.

2.1.1. Prevalence of Heart Disease around the World

Heart disease is a term that assigns to a large number of medical conditions or disease such as rheumatic HD, ischemic HD, congenital HD that are more related to heart. These medical conditions describe the strange health conditions that directly influence the heart and all its parts. This disease is a common fatal disease and it is currently the number one killer of the global population [3].

In day todays human life HD is becoming the major cause of deaths in the world. The WHO has estimated and reported that 12 million death occurred worldwide. They have also predicted that in the future almost 23.6 million peoples will die due to this disease and each year due to over 80% of deaths in world are because of HD[2]. According to the WHO report CVD causes 17.9 million people die every year accounting for about 32% of the world's deaths including HD. This report also stated that stroke and HD are the leading causes of CVD, accounting for approximately 85% of deaths, where people under the age of 70 account for one-third of all premature deaths[7]. The report also proved that 17.3 million deaths caused by HD in 2008, accounting approximately 6.2 million death were due to stroke and an estimation of 7.3 million people were due to coronary HD. WHO predicted that around 23.6 million people will die due to HD and stroke-related disease in future 2030 [27].

Generally speaking, the prevalence of CVD is rapidly aggregating in the world and consequently it is currently considered the leading cause of death in both developing and developed countries [28]. CVD was responsible for the death of 17.9 million people worldwide in 2016, accounting for 31% of all global deaths[28]. Additionally, the prevalence of CVD increased from 257 million peoples from 1990 to 550 million in 2019 [28] and the number of associated deaths shows a steady increase from 12.1 million in 1990 to 18.6 million in 2019 [29]. All the reports of these studies shows that the prevalence of HD around the world starting from the very early time to now a days is highly increasing.

2.1.2. Prevalence of Heart Disease in Ethiopia

The livelihood of developing countries like Ethiopia are affected by different conditions such as disease. Heart disease is one of these challenging conditions in Ethiopia and many researches are assuring that the prevalence of heart disease challenges the livelihood of peoples in developing countries like Ethiopia and it is also becoming the silent killing disease. As WHO and other different study reports show that the disease affects the life style of many lives in the world specially affects poor or developing countries including Ethiopia. Here we have included these reports in short notes.

A report shows that almost one third of the world population died in developing countries by 2010 [4] due to CVD. The other report shows in [4], CVD contributes high percentage of prevalence together with other health conditions. As CAD (Coronary Artery Disease), chronic

respiratory disease, Cancer and diabetes mellitus are on rising and the leading risks to human health and development. Additionally, it results in roughly 35 million deaths annually, of which 85% occur in developing nations like Ethiopia [4]. According to WHO estimation, approximately 9% of all deaths in Ethiopia in 2012 were caused by CVD which includes heart disease and another report in 2014, The WHO estimated that NCD claimed the lives of almost 30% of Ethiopians, with CVD accounting for 9% of these deaths[4]. An additional systematic review conducted in Ethiopia reveals that the prevalence of CVD, which includes HD, ranges from 7.2% to 24%. In year 2017, 2,838,764 individuals in Ethiopia were affected by CAD; of these, 33.7% had rheumatic heart disease (RHD), 22.5% had IHD, and 11.4% had an ischemic stroke [5]. Another study conducted in 2021 shows that there are different types of HD including IHD having 45%, heart stroke accounting 34% and hypertensive HD around 11% were the major leading causes of CVD deaths in Ethiopia with that 170 Ethiopians are die each day [5]. All the reports of these studies shows that the prevalence of HD in developing countries like Ethiopia is highly increasing with respect to the prevalence of the disease around the world.

2.2. Types of Heart Disease and Its Risk Factors

Many health conditions or diseases have their own types and these types have their own risk factors which increase the risk of the disease in the patient and this may help health experts differing these conditions with other health condition. Heart also have its own types of health condition that can affect the life of a person having the disease.

2.2.1. Types of Heart Disease

There are several types of HD which are expected as challenging millions of people in the world including Ethiopia. From these the following are few of these types of heart disease [9], Such as, rheumatic heart disease, ischemic heart disease, congenital heart disease, heart failure or coronary heart disease , stroke, hypertensive heart diseases and inflammatory heart disease and etc. Studies shows that most of the types of cardio vascular diseases including heart conditions can be prevented [6], if prevention and treatments are taken early before the disease reached to high level of risk. Even if there are several types of heart conditions, in this study, we have tried to include several heart conditions in this study such as heart failure, chronic heart failure, congenital HD, Ischemic heart condition, heart stroke and rheumatic heart

condition are few of these conditions. Thus, the next section is discussing some of these types of heart disease which are also discussed in most of the studies cited in this paper.

Ischemic heart disease (IHD): - is a types of heart disease, which occurs as a result of a restricted blood supply to the heart muscle. In more than 95% of cases, the cause of IHD is coronary blood flow reduction caused by coronary artery atherosclerosis, therefore the term "Coronary HD" is often used to describe this illness [30]. It is also defined as coronary HD or coronary artery disease, that is the eventual manifestation of myocardial dysfunction [31]. It can also said to be the supplying of blood and oxygen to a portion of the myocardium in insufficient way and this can typically happen when there is an inequity between myocardial oxygen supply and request [32]. Some of the main risk factors for IHD are high blood cholesterol, high blood pressure, tobacco usage, physical inactivity, unhealthy diet, diabetes, advancing ages, inheritance or genetics disposition and etc. In other hand other risk factors like poverty, low educational status, poor mental health or depression, inflammation and blood clotting disorders are some of modified risk factors[33].

Stroke: - strokes is another types of HD or heart condition caused by damaging or disruption of the supplying of blood to the brain. Stroke can be a result from either ischemic stroke (embolic stroke or thrombotic stroke) or rupture of a blood vessel (hemorrhagic stroke[33]. WHO explains as stroke is a clinical disorder in which a rapidly developing clinical signs of focal or global disturbance of cerebral function. If stroke is durable more than 24 hours leads to death or causing damage to the brain tissue with no apparent cause of vascular origin[33], [34]. Some of the major risking factors for stroke includes high blood pressure, rhythm or atrial fibrillation disorder, high blood cholesterol, smoking tobacco, diets which are unhealthy, physical inactivity, advancing ages and diabetes [33].

Rheumatic heart disease (RHD): -is a types of HD and defined as a chronic cardiac condition or disease with an infective etiology, causing high disease burden in low-income settings. Affected personnel of RH conditions are young people and associated sickness or mortality is highly increasing. However, Comparatively RHD is neglected due to the population involved and its lower incidence relative to other HD [35]. The disease can be caused by acute rheumatic fever which causes joint pain, skin changes, fever and sometimes abnormal movements [33].

Congenital heart disease (CHD):- is malformations or the disordering of heart structures existing at birth that may be caused by genetic factors or by adverse exposures during gestation and happens mostly in child hood while in birth time and develops to a person's life. It can be also renamed as a range birth defects affecting the usual workings of the heart. Such as holes in the heart, abnormal valves and abnormal heart chambers [33]. CHD disease also can be defined as a lifelong disease that results from a heart defect or structural anomaly at birth [36]. Maternal alcohol consumption, drug use during pregnancy (e.g., thalidomide, warfarin), maternal infections (e.g., rubella), inadequate nutrition or insufficient folate intake, strong blood link between parents, or consanguinity are some of the key risk factors for Congenital heart disease (CHD)[33]. Thus, with in several heart conditions, in this study we have tried to include several heart conditions, such as heart failure or coronary HD, chronic heart failure, congenital HD, ischemic heart condition, heart stroke and rheumatic heart condition are few of these conditions.

2.2.2. Risk Factors of Heart Disease

Risk factor is something that increase the chance of getting a disease. Although there are several risk factors for the development of HD, Certain risk factors are modifiable through alterations in lifestyle and medical interventions, whilst other risk factors are immutable. An individual's chances of acquiring HD increase with the number of risk factors they possess. As we have previously defined that every types of heart disease have their own risk factors, we have also mentioned the main, modified and non-modified risk factors of HD in the next section. These general risk factors are also included in the lists of the risk factors of HD in studies done in Ethiopia as mentioned in [37].

Mainly risk factors can be divided into two main categories [38]. The first risk factor is the major risk factor, which contributes the main risk of the disease. Such major risk factors includes high blood pressure, abnormal blood lipid, tobacco smoking, physical inactivity, overweight with obesity, unhealthy diet, diabetes mellitus, socioeconomic status, stress and alcohol intake are some of the major risk factors for different types of heart disease [38]. The second one is modifiable risk factors, which can be modified, changed or treated in some ways, such as low socio economic status, mental ill health, physical inactivity, uncontrolled diabetics, obesity, psychosocial stress, smoking, alcohol usage, use of certain medication, lipoprotein

and left ventricular hypertrophy [38]. The other risk factor which are non-modifiable risk factor and cannot be modified or updated by the person's life style, include advancing age, heredity or family history or genetics, menopause, gender and ethnicity or race are contributing for the high risk of HD to be increase [38]. However, The most important risk factors of HD and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol [39]. These all modifiable and non-modifiable risk factors contributes the main risk for the contribution of additional risk for the HD in the expected patient [39].

2.2.3. Treatments and Prevention of Heart Disease

Some of the risk factors for HD can be changed or modified while others cannot be changed or modified. Thus, early prevention of the risk factors of HD and taking treatments at an early time will reduce the risks for exposure to the disease. When an expected patient is going to visit health institutions or hospitals they must keep what the health experts or doctors recommended in order to prevent the disease early and the patient to get healthy treatments for his or her life. The treatment of the risk factors of the disease before being at high rate of risk will be the other main issue the expected patient to keep it up. Such that Cessation of tobacco use, reduction of salt in the diet, eating more fruit and vegetables, doing regular physical activity and avoiding harmful use of alcohol together with the help of the health experts [31], [35]. Consumption of healthy diets, avoiding or reducing overweight or obesity and treatments of risk factors including appropriate therapy according to local guidelines for diabetics, hyperlipidemia and hypertension, considering aspirin and statins shows to reduce or modify the risk of CAD including HD [31]. Identifying those persons, expected to be at higher risk of CVD or HD and ensuring if they receive appropriate treatment can also prevent premature deaths or high rate of morbidity. In developing countries like Ethiopia there must be some equipment's and medicines related with such diseases including HD, because access of non-communicable disease medicines and basic health technologies in all primary health care facilities are very essential to ensure that those in need receive treatment and counselling very early [38].

2.3. Machine Learning

Extracting knowledge from data is the focus of machine learning (ML), which is also a branch of study that lies at the convergence of computer science, artificial intelligence, and statistics.

Statistical learning or predictive analytics are the terms, which describes ML [12]. The goal of ML is to teach machines or software's to carry out tasks by providing them with a couple of examples in the concepts of how to do or not do a task [13]. The investigation of computer and internet technology has leads to many discoveries in the fields of science, law, business and medicine[19]. As ML is said to be a sub field of AI that supports different institutions keeping and managing their huge number of data in a computerized system or formats, Currently most institutions and organizations worldwide including Ethiopia are applying computerized systems, aims to save time, human resource, money, different equipment's and other tangible and intangible resources resulting these organizations more effective. Additionally, machine learning (ML) is quickly developing as the most attractive field of study due to its ability to handle huge amounts of data and computer systems' capacity to automatically increase previously acquired knowledge by finding new information without explicit or specific programming[40]. The general term that can describe ML is the training and testing of machine or computer based algorithms in the way how human being can learns and understands.

2.3.1. Types of Machine Learning

According to different experts and scholars ML can be subdivided in to different learning approaches such as Supervised, Unsupervised, Reinforcement and Semi-Supervised learning. Now we have to discuss them in the next section.

2.3.1.1. Supervised Machine Learning

Supervised ML is a types of ML in which the Machine accepts known label of input data and correct output labeled data with some true prediction is expected. This types of learning aims to get about representing function to represent input variable called independent variables with output variables called dependent variables based on a sample input-output pair [41]. The most common supervised ML tasks can be sub divided as classification tasks, that classify the data or regression tasks, that fits the data [41]. Under the principles of a task-driven approach, supervised ML is conducted when certain objectives are determined to be achieved from a given set of inputs. The rules of supervised ML may be applied to unseen label of data with unknown label of output data. Some of the supervised ML classification algorithms are LR, DT, KNN, NB and SVM.

2.3.1.2. Unsupervised Machine Learning

Unsupervised ML is another types of ML in which a model is trained using unlabeled sets of data without a need of human interference and allowed to act on that data without any supervision. This type of ML works on the concepts of data-driven process [41]. The main goal of Unsupervised ML is to get the original structure of dataset and grouping or clustering the data according to similarities. The most common types of unsupervised ML are association and clustering rules including the tasks of density estimation, K-means clustering, feature Learning, dimensionality reduction, anomaly detection and association rule findings[41].

2.3.1.3. Reinforcement Learning

Reinforcement Learning (RL) is a ML approach where a machine, software agent, or computer automatically assesses an ideal behavior in a specific context or environment in order to increase its effectiveness[41]. This process of learning rule mainly defined as a subdomain of ML where an agent learns from the environment and performs a desirable action called reward while if it undesirable action it will be penalized. RL can be categorized as positive or negative reinforcement and important to resolve problems in process of decision making and sometimes in a sequence of decisions like thermostat, automatic trader, robotic walking or movements, automatic chess playing, go player and automatic vehicles driving[19]. The most common examples of RL is Markov Decision Process (MDP).

2.3.1.4.Semi Supervised Learning

Semi Supervised Learning (SSL) defines the combination or the hybridization of both Supervised and Unsupervised Learning in which labeled data and unlabeled data could be included and combined, then the approach will be applied to find some knowledge from that data. SSL lies in the middle of supervised and unsupervised learning[41]. Therefore, an SSL's ultimate objective is to provide a prediction result that is superior to the one that is generated using labeled data from the model. Some of the tasks that could be done in semi SSL are text classification, labeling data, machine translation and fraud detection [41].

2.3.1.5. Ensemble Learning

Ensemble Learning (EL) is another types of ML in which a combinations of different models, experts or classifiers are joint to solve a particular computationally intelligence problem [43]. In this learning approach multiple base classifiers are trained and combines their prediction in

to a single output in order to achieve a better performance. EL can also improve the classification and prediction performance of all the base classifier models and reduce the weakest or very poor models. The main purposes of EL are to enhance model performance or lessen the possibility of selecting a poor model by accident [43]. Some examples of EL are voting classifier, Bagging, Boosting (Gradient Boosting and Adaptive Boosting) and stacking. Most of the researchers are using different ML techniques specially classification techniques in order to predict and diagnosis disease [16]. The priority learning algorithms that we have gave to review for this research study are ML techniques specially supervised classification algorithms such as SVM, KNN, RF and EL techniques such as GB and VC are tried to be applied.

2.4. Machine Learning and Health Care Institution

Health Care Institution is a preliminary and big institution for one country especially for developing countries like Ethiopia to keep the health care of the countries people or citizen. Countries like Ethiopia are getting their internal source of economy from agriculture with a very poor access to safe water, poor house holding or shelter, poor personal hygiene or environmental sanitation, lack of better food and health service. Healthcare is a significant sector that provides millions of people with value-based treatment, and it is also quickly rising to the top of the revenue earning charts in many nations[44]. So better health care infrastructure is a preliminary thing for developing countries like Ethiopia. One of the common and very important issue in keeping health of individuals in one country must be an early prevention of disease but sometimes preventing disease may become very challenging issues. However if it is impossible to diagnose and predict the disease early, it is impossible to prevent a disease early. An early diagnosis of a disease will reduces tangible and intangible resources of an individuals and health care institutions of that country including Ethiopia. Early prediction and diagnosis of a disease helps to reduce the expected individual health risks and this also helps the concerned health experts to decide on some issues, such as making the diagnosis of the disease in the patient simple and a timely ordering of treatments or interventions.

Generally speaking, early prediction of a disease using a disease predicting tools helps the risky individual patients and the health care experts. ML provides a technique on how to apply ML algorithms within an organization and evaluate the efficacy, suitability and efficiency of

ML applications [15]. ML can also apply in different tasks in health care area such as disease diagnosis and prediction, manufacturing and drug discovery, medical data imaging and diagnosis [19]. This technique used to predict a certain disease expectation if a person have or not, in this case HD expectation in a patient. Within this study we had used different labeled dataset in order to predict HD using few supervised classification and EL algorithms.

2.4.1. Machine Learning and Disease Prediction

ML is becoming an important technique in disease prediction and diagnosis and it uses patient history data in order to predict the disease. ML has various application areas and health care is one of these ML application area [15]. Predictive analysis of a disease with the help of several ML algorithms helps us to predict the disease and helps in treating the patients in an effective manner[15]. Many researchers in several studies are applying ML approaches to predict disease, especially some of supervised classification algorithms have a greater contribution in disease prediction including HD. As we have discussed in the previous section of this study supervised ML tasks can be subdivided in to classification and regression tasks and these tasks uses already labeled data to predict the input and an output data. Although many researchers uses more than two different algorithms in order to predict HD on their study for the reason of computation, comparison reason and making highly performing techniques or algorithms [19], we have tried to discuss and apply some of these supervised classification algorithms such as SVM, KNN, RF and other EL techniques such as GB and VC in order to apply on HD datasets in this study.

Support Vector Machine (SVM):- is a types of supervised algorithm which is applied for regression and classification tasks. This algorithm allows data classification, learning and prediction. SVM uses some kinds of hyperplane or set of hyperplane to classify a large amounts of datasets and it allows to classify data's that are linear and non-linear using kernel-trick [41]. The hyperplanes are decision boundaries that helps to separate the data points [19]. In each class, the hyperplane with the largest distance from the closest training data points will have the strongest separation. The classifier's generalization or classification error decreases with increasing margin[41]. SVM can be utilized in several application areas such as handwriting recognition, image classification and face detection and in the fields of medical diagnosis and weather prediction [36].

K-Nearest Neighbor (KNN):- is the other types of Supervised ML classification algorithm which works on the concepts of selection of the most NN called K. KNN is an instance based learning or non-generalizing learning algorithm [41] and it keeps track of every instance in N-dimensional space that matches the training set. KNN also uses data's in the way of similarity measuring techniques named as Euclidean or Manhattan distance measures [41], [36].

Random Forest (RF):- is a very popular types of EL algorithm or technique for the classification of tasks and it is chosen for the purpose of reducing overfitting and improves the prediction accuracy in the training sets of data. This technique also allows an extra randomness by using several decision tree in parallel fitting technique called parallel ensemble method [41]. In different dataset sample, RF uses majority voting method or average of the outcomes of each tree. It sometimes works or trained on the concepts of bagging and random FS methods.

Gradient Boosting (GB):- is another family of EL called boosting technique, which allows or converts weak classifiers in to strong ones by growing or adding the ensemble of predictors and using their corrections to fit the new predictor to the remaining mistakes of the old ones. GB is also called an EL algorithm that generates a final model based on a series of individual base classifiers or to combine many simple models, which are known to be weak learners [46].

Voting Classifier (VC): - is another types of EL technique used for classification or regression tasks. This technique will combines different algorithms that are trained with similar datasets and finds the prediction by considering majority voting concepts. VC trains the base classifiers in the ensemble and predicts an output class based on the highest prediction accuracy of the classifiers. The main aim is creating a single model which trains the model and predicts output based on their combined majority of voting for each output class. Voting Classifier can be subdivided in to Hard voting, in which the class with the highest majority of votes that is, the class with the highest probability of being predicted by each of the base classifiers and soft voting, in which the winner class will have the highest probability averaged by all of the classifiers. Some of the advantages and disadvantages of the applied ML algorithms in this study's experimental work are summarized as table 2.4.1 below.

Table 2.4.1.1. Advantages and disadvantages of machine learning algorithms

No	***Classifiers***	***Advantages of classifiers***	***Disadvantages of classifiers***
1.	K Nearest Neighbor (KNN)	NO training period and New data can be added, that doesn't impact the accuracy, very easy and simple, no assumptions, Variety of distance criteria to be chosen, Use for regression & classification problems.	May not work well for larger and higher dimensional datasets, needs feature scaling, Sensitive to noisy data, missing values and outliers.
2.	Support Vector Machine (SVM)	More effective in high dimensional space than the number of samples, Relatively memory efficient, If the classes in the data points are well separated works really better, Works appropriately for both regression and classification tasks.	May not be suitable for larger datasets, takes more time to train comparatively, may not perform well if there is more noise, and does not have a probabilistic explanation because data points are positioned above and below the classifying hyperplane.
3.	Random Forest (RF)	As is works on the concepts of EL techniques reduces overfitting, higher accuracy achievement, Works well for both categorical and continuous values, Solves both regression and classification problems, handles missing values automatically, No feature scaling required as no distance calculation is applied, very stable and less impacted by noise.	Can take longer training period as compared to some other classifiers, may happen complexity because of a lot of trees are created, More computational power and resources may require.
4.	Gradient Boosting (GB)	More accurate comparably, Trains faster in larger datasets, Some of them handle missing values natively.	Sometimes prone to overfitting, can be computationally expensive and take long time in low performance CPUs.

5.	Voting Classifier (VC)	Not hindered by larger errors or misclassification from one model, One model's poor performance can be countered by another model's excellent performance.	Voting may inherently become more computationally challenging since it needs the use of multiple models and Creating, training, and deploying such models may become more expensive and coasty.

2.5. Feature Selection Methods

FS method is an important concept in ML in which important features or attributes are selected in order to develop a very good model, to predict the output and reduce the number of independent variables called input variables, in developing a predictive model [16]. The major goal of FS method in ML is reduction of unwanted and noisy data's, removing and reduction of non-important and unwanted data features, reduce the dimension of the data to know the most important features, getting a better prediction accuracy model and etc. [16]. FS technique also involves the preprocessing of the given datasets in order to remove unwanted and redundant features in the original datasets to get adequate classifying [19]. FS also have a very important concept in health care and medical fields, in order to select very relevant and important features from patient's history data in an expected patient. This technique can also be utilized in areas of prediction of diseases such as prediction of diseases such as HD, prediction of hypertensive patient's data, prediction of kidney disease, diabetics' patient's record prediction and prediction of breast cancer [19]. Several researchers are applying different types of FS method on their study in order to remove redundant and unwanted features and tried to improve the prediction accuracy of models on their datasets. These main types of FS methods are shown in figure 2.5.1 below as Filter feature selection (FFS), Wrapper feature selection (WFS) technique and Embedded feature selection (EFS) technique. Now, we have discussed these three main types of FS methods in the next section of this study.

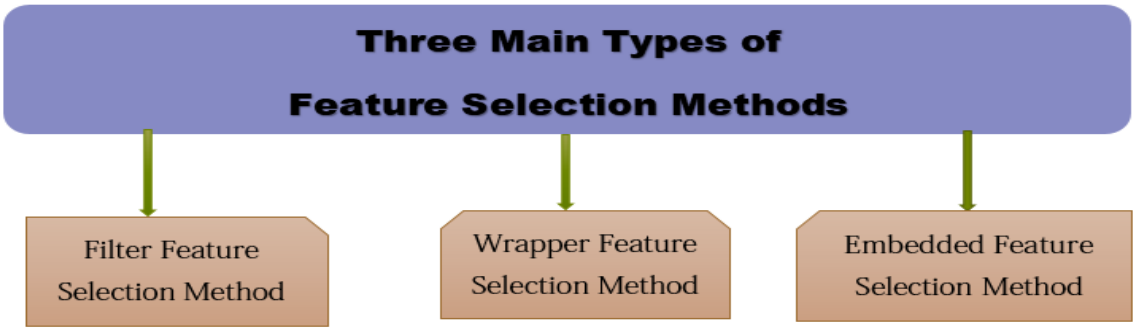

Figure 2.5.1. Types of feature selection method

Filter Feature Selection Method (FFSM):- are a types of FS method that returns feature sets which are independent towards any ML classifier or algorithm and it is highly generalizable method [47]. Studies show that FFS performs with independent of any predicting models or classifiers algorithms [47], this means that model could be applied after a certain relevant features are selected using the method. FFS methods also uses some statistical methods to know the dependencies between input or independent variables and output or dependent variables. Sometimes FFS is fast methods, in terms of time complexity. Some usual examples of FFS methods includes ANOVA, Chi-Square, Correlation, Information gain and etc.[48].

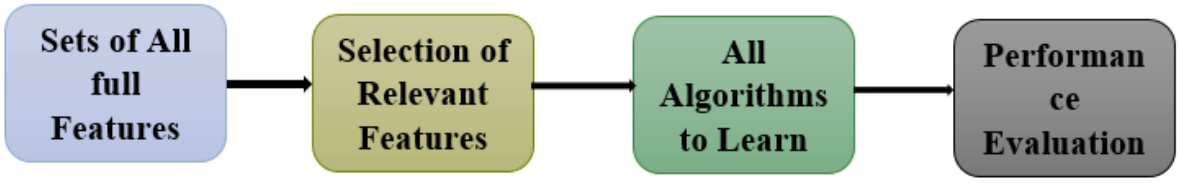

Figure 2.5.2. Filter Feature Selection Method

Wrapper Feature Selection Method (WFS): - is another types of FS method, which selects and trains ML algorithms or classifiers during selection of relevant features sets using greedy search algorithm concepts. Wrappers methods uses predetermined ML classifiers or algorithms for the selection of subsets of features evaluation by performance such as accuracy. This process makes the final feature subsets that are selected and correlated with the chosen relevance measures [48]. Studies shows that wrapper method evaluates on a specific ML

algorithms to find optimal or better features[49]. WFS method achieves a superior performance of accuracy but it may takes high computation time. Some of common WFS methods are FFS, BFE and Step-Wise selection [50].

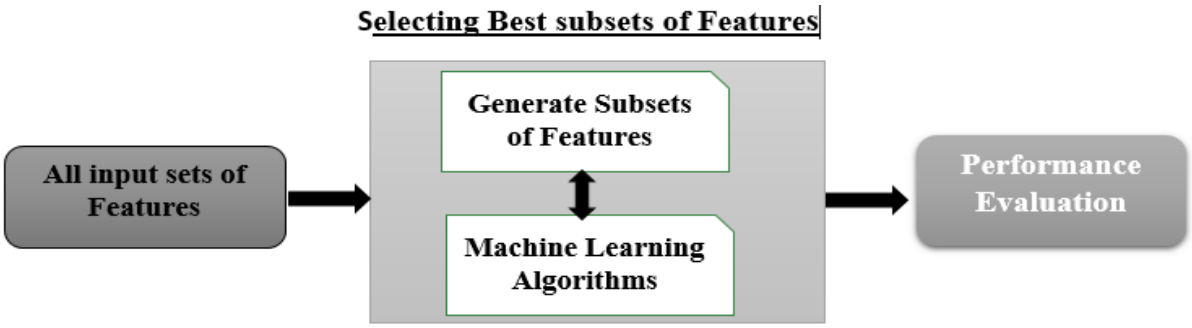

Figure 2.5.3. Wrapper Feature Selection Method

Embedded Feature Selection Method (EFS): - is another kind of FS method having an average time complexity between FFS and WFS methods and combines the benefits of wrapper and filter methods by including the interaction of feature sets but sometimes takes reasonable computational costs at a lowest computational cost compared with Wrapper approaches [19]. Some of the common examples of EFS method are LASSO (Least Absolut Shrinkage, and Selection Operator) Regularization, Elastic Net, Weighted NB, ,Feature Importance, and Ridge-regression etc. [19], [50]. EFS method have ML algorithms with FS which is encapsulated and embeds or fixes features during model development process [50]. In this proposed study we have tried to apply WFS method specifically SFFS method and other FFS method named as Chi-Square FS (CFS) method, in order select the most useful attributes or features together with some of supervised ML algorithms such as SVM, KNN, RF and EL techniques such as GB and VC algorithms, in order to have a better and enhanced ML algorithm in detecting HD.

2.6. Related Works

ML is becoming a greater field in every day's activity of life of human being. Several researchers are applying various ML techniques in several areas. This field also becomes parts of activities especially in contributing relevancy in diagnosis of human and plants disease such as prediction and diagnosis of CKD, [16], HD prediction [24], Complicated appendicitis prediction [51], Diabetes prediction, BC prediction [49], HIV AIDS prediction [52] and Liver

disease diagnosis, prediction and etc. The next literatures are some of the reviewed related works which are more related with ML techniques. HD related works with FS method especially WFS and FFS methods are also reviewed and discussed here.

Dibaba Adeba [16] proposed a CKD prediction using a ML techniques to predict the severity of KD as multi classed (not chronic disease, mild, moderate, severe) and binary classed (the absence or presence of CKD). In this work three ML classification techniques were used RF, SVM and DT and determine the best performance model by training all classifiers using local hospital patient history dataset collected from St.Paulo's hospital. In addition to ML algorithms they have applied two FS techniques in combination with the mentioned ML algorithms in order to enhance the performance of the utilized models, these are ANOVA (a type of FFS) and Recursive FEM(a type of WFS) to increase the performance of the algorithm mentioned and 10-F-CV was applied. Comparatively RF classifier with RFE method achieved good performance accuracy and F1- scores of 99.8% and 99.8% respectively for binary classes and accuracy and F1-scores of 79% and 77.9% respectively for multiclass with this classifier RF was achieved. Finally they have recommended the proposed severity prevention model of CKD based on the severity of the disease to provide good prevention, treatment and diet recommendation for health care experts.

C. Beulah Christalin Latha et.al [25] proposed an ensemble classification method for prediction of HD risk. In this proposed work before applying ensemble technique they have used six classifiers (Bayes Net, NB, RF, and C4.5/Quinlan, PART (Projective Adaptive Resonance Theory and Multi-Layer Perceptron (MLP))) for the Cleveland HD dataset classifications individually. From this MLP, PART and C45 found to be weak classifiers while NB, RF and BN was achieved a better performance compared with others. When the HD dataset is traind and tested using the six proposed individual classifiers they all achieved an accuracy ranges from 75.58 % to 83.17 %. After they have achieved an accuracy with the individual classifier they have applied an ensemble technique in order to improve the accuracy of weak algorithms by combining several classifiers on HD datasets .The data set collected from UCI repository address of UCI Machine Learning Repository: Heart Disease Data Set named as Cleveland heart disease data set with 303 instances with 14 features having 8 categorical and 6 numerical. A comparative analytical approach was done to determine how the ensemble technique can be

applied for improving the prediction performance of different weak classifiers in the ensemble technique applied on the HD dataset. The ensemble techniques applied on the weak classifiers to analyze the data were bagging, boosting, and majority voting and stacking. When bagging was used it was improved by 6.92%, when boosting was used the accuracy was improved by 5.94%, while weak classifiers was ensemble using voting classifier the performance was increased by a value of 7.26% and after they have used an EL technique of stacking they were achieved an accuracy of 6.93%. After all the Ensemble techniques are applied, EL of majority voting using NB, BN, RF and MLP have been chosen as a best accuracy achiever with 7.26% , having a specified accuracy of 85.48%. It was tried to enhance the performance of the process using FS implementation method called Brute force method which is applied to limit the lower bound with a minimum of 3 attributes. All of the possible combinations of 3 attributes from the 13 feature were selected and each combination was tested with the classifier algorithm techniques. However the performance achieved using classifiers with BFFS method was not better than the accuracy achieved by the majority voting ensemble. Majority voting was achieved an accuracy from 75.58%-83.17% to 85.48% compared to other ensembles without using BFFS method. So the researcher decided that majority voting had been the best ensemble technique to predict HD.

Ibomoiye Domor Mienye et.al [28] proposed an EL ML method which was thoughts as improved ensemble technique for the prediction of HD risk. This proposed technique performs random partitioning of the original dataset into smaller sub-sets using a mean-based splitting or partitioning technique. Every partition was then modeled using classification and regression tree (CART) algorithms after a simple randomization was introduced and applied. The accuracy-based weighted aging classifier ensemble (AB-WAE), which is a modification of the weighted aging classifier ensemble (WAE), was used to create the homogeneous ensemble from the various CART models. To do the experiment and to evaluate the pre-proposed model, Cleveland HD dataset obtained from UCI repository with in an address of UCI Machine Learning Repository: Heart Disease Data Set was used. The data set have 303 instances with 14 features and Framingham heart disease dataset obtained from Framingham heart study dataset | Kaggle having 4238 instances with 16 features was used. Lastly after they have compared classifiers proposed in excluding of their work, the proposed work was achieved an accuracy of 93% and 91% with Cleveland and Framingham HD datasets respectively through the

ensemble technique. The study was aimed to evaluate the efficacy of the proposed method by examining various performance metrics, including sensitivity, precision, and F1 score.

Saiyed Faiayaz Waris et.al [9] proposed a study to save life of heart disease patient by predicting HD over a few significant parameters of the heart and it was aimed to obtain more accuracy than the existing methodologies or techniques in prediction of HD. This approach is an improved form of KNN and New-Novel-K=K+PN classifiers. The Novel KNN aims to get more accuracy compared with the existing KNN approach. This paper work tried to improve accuracy by reducing the classification error rate. The reason behind taking PN is, its powerfulness in set-righting any situations. To do this experiment Cleveland HD dataset having 303 instances with 14 features are used and an accuracy of 88% and 93% is achieved with KNN and Novel KNN respectively.

Senthilkumar Mohan et.al [19] Proposed HRFLM(hybrid of RF and LM) method which was aims in finding significant features or attributes and resulting in improving the accuracy of prediction for CVD specially HD. The prediction model also introduces the different combinations of features using DT entropy based splitting and FS technique. The model developed selects all the features without any restriction with several classifiers. In order to achieve the experiment, Cleveland HD dataset found on the UCI repository having 303 records with 14 attributes are accessed and the experiment was done using an *R* studio rattle plate form. R studio is an easy to use visual representation of the datasets and also used to perform HD dataset classification. Before FS were applied, all individual classifiers such as RF, LM and DT were trained and tested in order to check performance with all features of Cleveland HD dataset. After classifiers are applied on the HD datasets, to find out the best classification performance, it was also applied DT entropy based FS method with the classifiers to get better accuracy. The results shows that RF and LM achieves best accuracy and the error rate for RF comparing to all other datasets partitioned before was 20.9% and the error rate for LM method comparing with all other datasets partitioned was 9.1%, that was the best compared to DT and RF methods. Then they have combined RF and LM as hybrid model in order to achieve improved and highest accuracy. Their experiment in the proposed hybrid HD prediction model with RF and LM achieved a performance level with a high accuracy of 88.7% and a less error rate of 9.1 % was achieved.

Majdi R. Alnowami et.al [49] proposed a WFS approach to investigate potential biomarkers of breast for early detection of breast cancer (BC). The data which is used in the study process was collected from the department of Gynecology in the University-Hospital-Center-For-Coimbra (CHUC) between 2009 and 2013 and the data collected holds 64 women were as BC patients (Cases) and 52 women as healthy (Control) volunteers. The proposed study attempted to investigate the impact of utilizing one or more distinct biomarkers, which is not only signal the presence of cancer but also the degree to which a patient responds to suggested treatment as predictors of BC. Before using FS method they have used all the features with the selected classifiers. Then in order to achieve the performance of the proposed work they have used a specified technique of WFS method called SBFS method to select features with integration of different classifiers such as SVM, RF and DT. Four features were selected as important features for BC Biomarker in BC Screening and detection using SVM classifier that shows the optimal set of biomarkers and the features having the highest misclassification rate was removed for n-1 feature subsets. Several widely used statistical measures was used in order to evaluate classifiers performances of the proposed model. Lastly, the results shows that sensitivity of 94%, a specificity of 90% and confidence interval of 95 for AUC of 89%, 98% and an accuracy performance of 92% were achieve.

Daniel Mesafint Belete et.al [52] proposed WFS technique on EDHS (Ethiopian Demographic and Health Survey) of HIV/AIDS datasets. The prosed work was used to predict individual status or test out come from HIV AIDS data set collected from EDHS and found in central statistics agency (CSA). The data set has 78,877 instances of data with 25 known attributes or features and having two classes from which 55,209 instances are from one class and 23,668 instances are from another class having nominal and numerical values. In this proposed paper three common WFS method such as SBFS, SFFS) and Exhaustive RFE are applied in order to classify the data set with other seven ML algorithms namely RF, SVM, KNN, LR, NB, Ada-Boosting (AB) and GB . In this experiment the three WFS methods with in every 20, 15 and 10 features in each round was applied in cooperation with all the algorithms and achieved their own classification performance. They actually used five classification evaluation metrics such as accuracy, precision, recall, F1-Score and ROC curve. Lastly RF achieves an accuracy of 94% with RFS, GB achieves an accuracy of 90% with RFS and KNN achieves an accuracy of

90% with RFS. That all was achieved the highest accuracy using recursive feature elimination method of WFS compared with other algorithms.

G. Ramesh et.al [20] proposed an improved accuracy of heart attack risk prediction based on information gain (IG) FS method which was assumed to boost a ML classifier performance. IGFS in this experiment uses all the features and makes dimensionality reduction on the HD datasets accessed. Cleveland HD dataset collected in UCI repository having 303 instances and 14 features with the last feature used as the target class was used for the experiment. All the Classifier techniques such as RF, KNN, NB, DT, LR and SVM are employed with this IGFS methods (IGs). After they have applied different performance evaluation techniques and achieved a better accuracy using SVM and RF. An accuracy of 87.67% was achieved without applying FS and an accuracy of 88.9823% was achieved using SVM with FS and 88.9812 % was achieved using RF with FS. The least performance accuracy was achieved by DT.

Mohammad Shafenoor Amin et.al [21] aims to propose an identification of significant features and data mining methods in prediction of HD. Techniques introduced here tried to improve the prediction accuracy in CVD. The Cleveland HD data sets having 303 instances and 14 features and Statlog HD data sets found in an address named as UCI Machine Learning Repository: Statlog (Heart) Data Set having 270 instances with 14 features was used for the experiment. The model proposed was developed using seven classification algorithms such as DT, NB, ANN, KNN, SVM, LR and Voting (Hybrid of NB and LR). All the classifiers are applied on the dataset before FS method was performed and individual classifiers get their performance on prediction. After this BFFS method was applied to restrict the lower boundary to make as a minimum of 4 or 3 features. The process was repeated as a subset containing a minimum of three (3) or four (4) features with that of the seven data mining classifier techniques on Cleveland heart disease data sets. The process was performed in the concept of 2^n-1 feature combination selection method. Then all the models were evaluated using performance evaluation metrics such as accuracy, precision and F-measures. Lastly as their experiment it was observed that nine (9) best features are selected as a very significant features and three (3) main classifiers such as NB, SVM and VC were chosen as a best predictor models or classifiers. A higher performing data mining algorithm having an accuracy of 87.4 % was achieved by VC (ensemble of NB and LR) in predicting HD data sets used after FS was applied.

Hidayet Takci et.al [22] proposed an improved heart attack prediction by using ML techniques combined with FS method. The aim of the paper is identifying the best performing ML algorithm and the best performing FS method in order to predict heart attack. Statlog HD data sets having 270 instances with 13 features each extracted from 76 features were accessed in order to achieve their experiment with that of several ML techniques and FS methods. The experiment was performed using at least twelve ML techniques such as C45., Binary LR, SVM with Sigmoid Kernel, C-RT, SVM with Linear kernel, SVM with Polynomial Kernel, KNN, SVM with RBF Kernel, ID3, NB, Multinomial LR and MLP and other four different types of FS methods from two categories of FS such as stepwise regression (back ward logic and forward logic from WFS methods) and Fisher Filtering and Relief-F (from FFS methods) method were used. Accuracy and ROC curve were also compared. The different concept which is comparing the proposed work with others is that the algorithms used was divided in to four categories such as regression analysis, DT, SVM and others. Before applying FS method, each ML algorithms were tried to evaluate their performance of accuracy, ROC curve and processing time on the HD data set. After this process FS method were applied and the highest accuracy were registered on feature sets extracted from relief-F FS method. The experiment of the work shows that SVM in the form of linear kernel becomes the best performing ML algorithm in combination with relief-F FS method achieving a best and higher accuracy of 84.81 % to predict heart attack using Statlog HD dataset.

Rattanawadee Panthong et.al [53] proposed WFS method in order for dimensionality reduction based on EL technique for different types of separated datasets. The dataset used for the experiment was collected from UCI repository and thirteen (13) datasets with different numbers of features or attributes having different size of dimensions were accessed. The study applied three WFS methods such as SBFS, SFFS and Optimization Selection (Evolutionary) based EL techniques such as AB and Bagging. Two other classifier techniques were used such as NB and DT as subset evaluators. The used Optimization Selection was in the form of four approaches such as OBDT (Optimization selection, Bagging and DT), OBNB (Optimization Selection, Bagging and NB), OADT (Optimization Selection, AB and DT) and OANB (Optimization Selection, AB and NB). More adaptive methods such as BBDT (SBFS, Bagging with DT), FBDT (SFFS, Bagging with DT), BBNB (SBFS, Bagging with NB), FBNB (SFFS, Bagging with NB), BANB (SBFS, AB with NB), BADT (SBFS, AB with DT), FADT (SFFS,

AB with DT) and FANB (SFFS, AB with NB) were also applied. Lastly the Bagging techniques using DT which SFFS method with Bagging, and DT (FBDT) achieved a highest performance with that of WFS method called SFFS of an accuracy of 89.60 % for four data sets compared with others.

Devansh Shah et.al [23] Proposed HD prediction methodology using ML which was aims to envision the probability in developing HD in a specified patient through computerized predicting technique. The paper also tried to depict which attribute or feature provides the higher precision value. This proposed work presents various ML approaches such as NB, DT, RF and KNN algorithms in order to create a model with a maximum performance accuracy to classify the Cleveland HD datasets found in UCI repository having 303 instances of heart patient's record with 14 features or attributes. The highest accuracy was achieved using KNN, achieving an accuracy of 90.789% to predict HD.

Nikhil Bora et.al [24] proposed a HD prediction methodology using ML. The main objective behind this study is that, how efficiency rate is enhanced in HD prediction. In this paper several ML algorithms such as LR, NB, SVM, KNN, XGB and RF were used. In order to conduct the proposed work, two separate datasets such as Cleveland HD data sets and CLbSHS was used. The first HD data set was collected from UCI repository called Cleveland HD dataset having 303 instances of data with 14 features and the second datasets was collected form Kaggle web site which is a combination of five known HD data sets named as Cleveland-Long Beach-Switzerland-Hungarian and Stalog HD datasets having 1190 instances with 12 features. Then after they have used these ML classification approaches it was achieved an accuracy of 92% using SVM through Cleveland HD data sets and RF achieves a higher accuracy of 94.12% with the second data set. After they have achieved such performance on individual data sets, then the data has been combined in order to propose the combination of dataset results. Actually the results of the combined datasets performance achieves an accuracy of 93.31 % using RF.

Hana Engidasew [14] proposed a ML based CKD prediction model. In the proposed work they have tried to classify the existence of CKD in a patient. In order to achieve this work they have tried to access a data set collected from local hospital found in Addis Ababa, Ethiopia named as Tikur Anbesa Specialized Hospital (TASH). The data having 2135 instances of patient

records with 21 features and they have tried to preprocess well and accessed for the experiment. The proposed work experiment was applied four types of ML algorithms such as MLP-ANN, SVM, DT and RF. A higher performance accuracy of 99% using MLP ANN with Pearson's Correlation FS technique was achieved to predict CKD.

Yap Bee Wah et.al [48] introduced FS method in the case of wrapper and filter approach for maximizing of classification accuracy. This approach mainly compares filter and wrapper approaches in classifier performance applied on different disease data sets simulated. In order to perform the experiment, simulation was done on R and three real world datasets were applied such as Bc Wisconsin datasets found in Breast Cancer Wisconsin (Diagnostic) Data Set | Kaggle having 669 BC patient data instances with 9 features, Spam base email datasets found in an address Spam Mails Dataset | Kaggle having an instances of 4601 with 57 different attributes or features and Pima Indians Diabetes datasets found in an address Pima Indians Diabetes Database | Kaggle having 768 instances with 8 features which are collected from UCI ML repository. LR was used as ML classification algorithm in order to evaluate the performance on WFS method. In the proposed work, other FS methods such as FFS method (named as IG and correlation based FS methods) and WFS methods (named as SBFS and SFFS) were applied on the three real world data sets collected from UCI ML repository data base. The results shows that the wrapper method called SFFS and SBFS were better than the FFS method in performance evaluation resulted in simulation on R (an open source programming language). The experimental work shows that the pima Indians diabetes datasets results with an accuracy of 83.94% was achieved with no FS, BC Wisconsin results with a higher accuracy of 97 % with SFFS and SBFS (by selecting 8 relevant features only) while Spam-Base-Email dataset results with a higher accuracy of 93% with SBFS method (by selecting 46 relevant features only from 57 full features). The table 2.6.1. Shows some of the summary of related works included above.

Table 2.4.1.1. Some of Summary of Related works

No	***The Title***	***Author name***	***Techniques used***	***Dataset used and Accuracy achieved***	***Limitation of the paper and suggested concepts***

1.	Improving the Accuracy of prediction of HD-risk based on Ensemble classification Technique.	C.Beulah Christalin Latha and S.Carolin Jeeva (2019) [17].	BN, NB, RF, C4.5/Quinlan, PART, MLP, BFFS, M-VC, Bagging, Boosting	Cleveland HD datasets, VC achieves accuracy of 85.48%	Used Small datasets collected in Public repositories only and Local datasets are not included.
2.	An improved EL approach for the prediction of HD risk	Ibomoiye Domor et.al (2020) [28].	CART and Ensemble of ABWAE are used.	Cleveland & Framingham datasets, 93% & 91% achieved for Cleveland and Framingham.	No clear data preprocessing methods applied for Cleveland datasets, No FS methods applied.
3.	HD Early Prediction Using a novel ML method called Improved KNN classifier in python.	Saiyed Faiayaz Waris, S. Koteeswaran (2021) [9].	K NN and Improved form of KNN= K+PN	Cleveland HD datasets, KNN and Improved KNN achieves an accuracy of 91% and 93%.	Smaller datasets without clear FS method was used.
4.	Effective HD Prediction Using Hybrid ML Techniques	Senthilkumar Mohan et.al (2019) [19]	Firstly RF, LM and DT was used, HRFLM and DT entropy based FS and splitting method.	Cleveland datasets and an accuracy of 88.7% using HRFLM.	Smaller non-local dataset, no clear data preprocessing, Simply removal of missing values and Requires advanced knowledge on quantitative and

					statistical data analysis and as the data increase complexity also increase.
5.	Improving the accuracy of heart attack risk prediction based on IGFS technique.	G. Ramesh et.al (2020) [20].	RF, KNN, NB, DT, LR, SVM and IGFS method was applied.	Cleveland HD datasets, a score of 88.982% and 88.981% achieved using RF and SVM with IGFS method.	Smaller datasets with No clear data preprocessing method are done for Cleveland data.
6.	Identification of significant features and data mining techniques in predicting HD.	Mohammad Shafenoor et.al (2018) [21].	DT, NB, NN, KNN, SVM, LR, Voting (NB and LR) and BFFS was applied.	Cleveland and Statlog HD datasets used, 87.4% was achieved using VC.	Simply removing of missing values for Cleveland datasets and the feature which contribute best performance is not clearly identified.
7.	Improvement of heart attack prediction by the FS methods.	Hidayet Tacki (2018) [22].	C45, Bi-LR, SVM(Sigmoid, L, Poly, RBF), C-RT, KNN, ID3, NB, Multinomial-LR and MLP. BFS & FFS & Relief-F from FFS.	Statlog HD dataset, SVM with linear kernel achieves a highest score of 84.81 % with relief-F FS method.	The data splatted to 90%/10%, this may affect the accuracy of testing datasets, The dataset used is still online datasets and very small in number.

8.	HD Prediction using ML Techniques	Devansh Shah et.al (2020) [23].	DT, NB, KNN and RF	Cleveland HD dataset and a higher score of 90.789% was achieved using KNN.	No FS methods, Small online data with no data preprocessing.
9.	Using ML To Predict HD	Nikhil Bora (2021) [24].	LR, NB, SVM, KNN and An ensemble method of XGB was used.	Cleveland & CLbSHS HD, SVM scores 92% with Cleveland & RF scores 94.12% with CLbSHS dataset & RF scores 93.1% with in combined datasets.	No FS method applied with no clear data preprocessing methods for Cleveland datasets.

Almost all of the above reviewed related works shows that although several works are done related with HD using ML. However, it is not that much enough in including FS methods, especially WFS and FFS methods. Almost all reviewed studies data collection method shows the datasets accessed for model development and testing were from different online Public repositories, rather collecting real patient data's from health institutions or hospitals for configuring it for their model implementation was not observed. Although several studies are done related with ML technique, most of these works related with HD prediction are done outside the country, Ethiopia. This shows almost there is a little studies are conducted in Ethiopia related with ML techniques and HD. There for we have been motivated and proposed to access local hospital HD dataset to develop the proposed model for this study. In addition to this, the collected local HD dataset is used in combination with other public datasets in order to develop a better model and this process increase the prediction trustfulness of the model for local patient data in standard of public dataset format. The study also includes FS method, in order to get more relevant features that affect the prediction result positively. So the researcher's aim is getting the more relevant features that makes a better and effective prediction performance accuracy using best ML algorithms with a good FS method.

CHAPTER THREE

MATERIALS AND METHODS

This chapter discusses all research activities, methods or techniques, procedures or materials in the process of early detection of Heart Disease (HD) enhancing prediction through ML technique and FS methods. As we have discussed in previous chapter ML have a greater impact in the area of health care as a disease diagnosis and prediction method. Especially in health care institutions a large number of data cannot be managed by human being but ML techniques can be very appropriate for solving such issues. In this section, all the methods and techniques used in the proposed study are discussed. Firstly, all the data sources and data collection methods are discussed. Secondly, the architecture of the proposed model with all the dataset preprocessing methods in python programming language are discussed, As such data preprocessing methods are very useful step in the process of unselecting features that are not relevant and using a better features for developing the proposed model. In this study data preprocessing method is a very crucial part of the proposed early detection of HD model development process. The other phase is the FS method phase which is used to select very useful features that are parts of HD dataset features. The ML model development phase explains all the ML techniques that are used to develop the proposed model and detects HD and predicts its class for the given separated datasets early. Model performance evaluation metrics such as accuracy, precision, recall, F-score, sensitivity, specificity and finally the implementation and deployment environments are also discussed.

3.1. Data Source

The data source to conduct the proposed study was collected from two main sources. The first source of data is collected from public dataset repositories such as UCI repositories having a website address of Heart Disease Dataset | Kaggle. These repositories are the main source of HD datasets that are configured properly for the purpose of researches conducted by several researchers publicly. These datasets are the main datasets used for the purpose of model development. The second main source of data was local health institutions found in Ethiopia, specially Wolkite University Teaching and Referral Hospital (WKUTRH). The hospital launched giving service for the near community starting from 2012 until now and gives service

for the near community in different fields of cases such as CVD inpatient and outpatient cases including different HD conditions.

3.2. Data Collection

Both the data's collected from Public repositories and WKUTRH are secondary patient record data's which are collected without direct communication with the patient. In order to consider different features to predict the presence or absence of HD, we have tried to use different data collecting methods, such as interviewing some of the concerning health experts on that area. WKUTRH health experts who are working in inpatient ward and outpatient ward class's helps us in collecting relevant data. Reviewing different Public and local resources on the area of HD and related works was also another source of data. However, collecting and reviewing local dataset resources is not that much satisfy able, so considering publicly available datasets standard format have a better role in configuring local datasets used. The HD data which was collected in WKUTRH is the patient history data stored in data record class of the hospital before these days and these data's are recorded manually on paper based format. Thus, this dataset collected was configured in the form of Public dataset standard format with an instances 774 with 14 features including 1 target categorical value. While, the other dataset available in public repository address of Heart Disease Dataset | Kaggle, considered is named as Cleveland_Hungary_Switzerland_and_Long_Beach has instances of 1025 with 14 features including 1 target class. Both public and local datasets contains their own categorical and numerical features with the mentioned target class.

Through this process, first, the data collected from public repositories and local hospital are applied for model development and evaluated through different model evaluation metrics and then after, the combination of the two datasets are used for model development and finally performance have been measured and evaluated. Through the process of combining the two HD datasets into a single merged dataset has a number of advantages. One of these advantages, is the ability to analyze a more comprehensive dataset that incorporates data from both public and local sources. Combining those datasets also enables us to determine which attribute is more pertinent when the number of data instances increases, as well as qualities and correlations that might not be visible when looking at individual datasets holistically. Combining local and public HD datasets can lowers the possibility of making decisions based

on partial or incomplete information, which contributes to more informed decision-making. Furthermore, expanding the amount of data with more instances and attributes can improve the quality of information that is available for analysis using the improved data. Combining the two identically designed datasets also has the benefit of enabling cross-validation information sharing, which increases data reliability by allowing information from one dataset to be cross-checked or validated in another. Ultimately, merging the Local and Public databases enhances performance and contributes to better data quality and more efficient decision-making processes.

3.3. Dataset Preparation and Description of Features

In this study HD detection at early time through prediction of the disease using ML techniques and FS method is proposed. Dataset preparation and feature description is an important process in preparing relevant features and know a features having a specified description. Although there are additional public HD datasets available in online repositories, we have also used local hospital HD datasets in Ethiopia. This is because using local hospital data's have its own advantages. As the study is done in Ethiopia the developed model will helps different concerned experts and users and the study will be more trust full. The dataset collected from local hospital that is used for building the model was collected from WKUTRH heart patient history data and the researcher prepared it in an appropriate form. The standard used for preparing local datasets was the standard format of the HD data's found in Heart Disease Dataset | Kaggle repository, named as cleveland_hungary_switzerland_and_long_beach. Both datasets are prepared in the form of CSV (comma separated values) format in order to make the data appropriate for machine readable in python programming plate form. The public dataset and the local dataset holds 14 features such as age, sex, chest pain type (Cp), resting blood pressure (Trestbp), serum cholesterol (Chol), fasting blood sugar (Fbs), resting Eco-cardio graphic results (restecg), maximum heart rate achieved (thalach), exercise induced angina (Exang), ST depression exercise (Oldpeak), slope of the peak exercise ST segments (Slope), number of major vessels colored by fluoroscopy (Ca), types of defects (thal) and the target class attribute. The description of these features are tabulates as table 3.3.1 below.

Table 2.4.1.1. Description of Heart Disease Dataset

No	***Attribute name***	***Description of features (attributes name)***	***Range of Datasets***	***Type of the attribute***
1.	Age	Age of the person in years	7 to 100	Numeric
2.	Sex	Gender of the person having 1= Male , 0 = Female	0, 1	Categorical
3.	Cp	Chest pain Type having 0=Typical angina, 1=atypical angina, 2=non-angina pain, 3=asymptomatic	0,1,2,3	Categorical
4.	Trestbps	Resting Blood Pressure in mm Hg	17 to 225	Numeric
5.	Chol	Serum Cholesterol in mg/dl	126 to 564	Numeric
6.	Fbs	Fasting Blood Sugar in mg/dl	0,1	Categorical
7.	RestEcg	Resting Echocardiographic Results	0,1	Categorical
8.	Thalach	Maximum Heart rate Achieved	40 to 202	Numeric
9.	Exang	Exercise Induced Angina	0,1	Categorical
10.	Oldpeak	ST Depression induced by Exercise	0 to 6.2	Numeric
11.	Slope	Slope of the peak Exercise ST segment	0,1,2	Categorical
12.	Ca	Number of Major Vessels Colored by Fluoroscopy	0 to 3	Numeric
13.	Thal	Defect Type having 0= normal, 1= Fixed defect, 2= Reversible defect and 3=non reversible defect	0,1,2,3	Categorical
14.	Target	The binary target class attribute	0 or 1	Categorical

3.4. Heart Disease Prediction Modeling

3.4.1. Architecture of Heart Disease Prediction model

After the collection of necessary datasets of HD from different sources and configuring them properly to get the most relevant feature for the prediction of HD, the model was developed using five different ML algorithms namely KNN, SVM, RF, GB and VC and evaluation metrics are compared for them. The overall architecture of the proposed HD prediction model is shown in the figure 3.4.1 below.

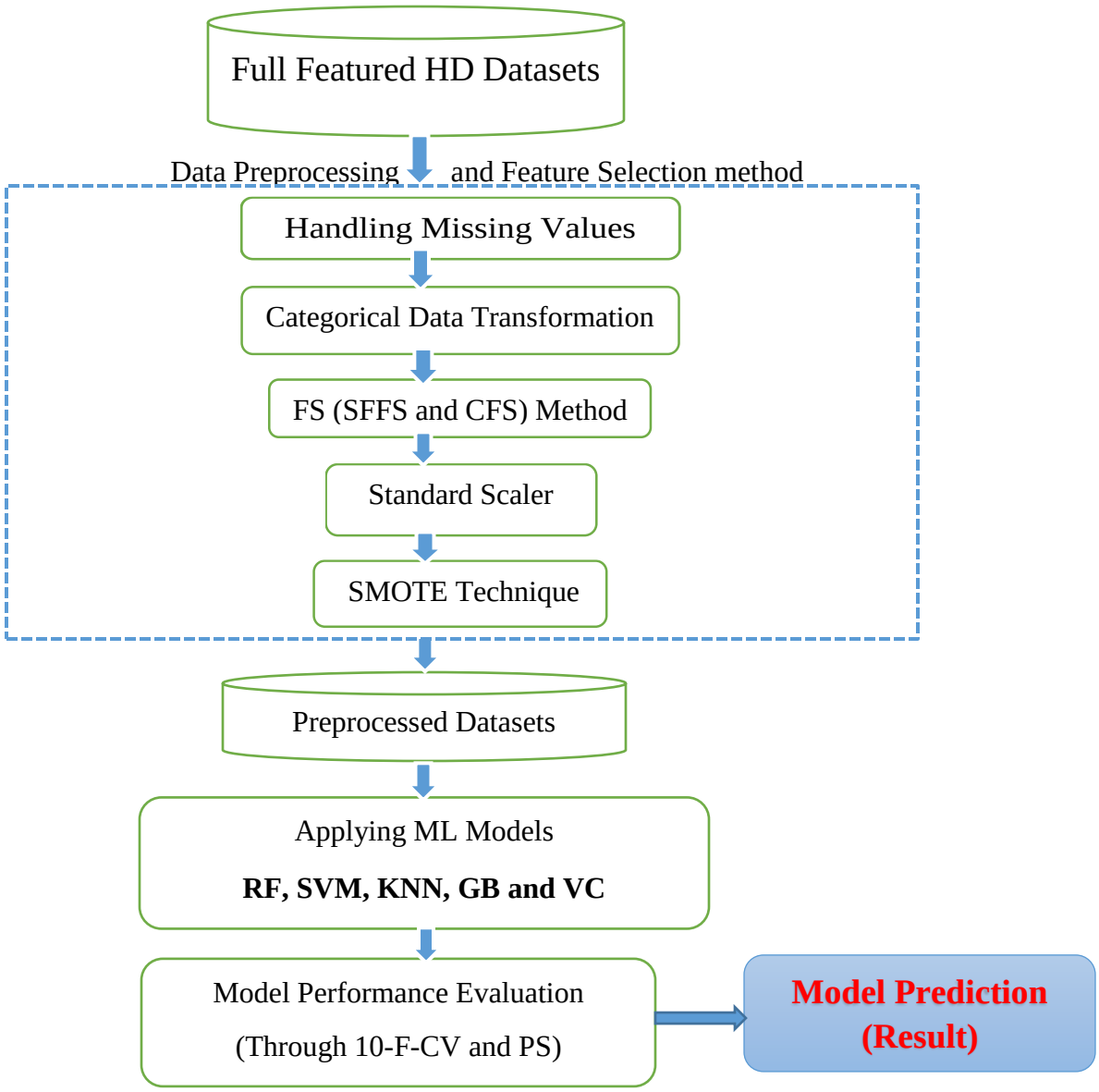

Figure 3.4.1. The proposed heart disease prediction architecture

3.4.2. Data Preprocessing

Data is the most important parts in ML, AI and data analysis [1]. Data can have several issues in every day's activity such as missing values, outliers, categorical and numerical values in a mixed formats. Thus, after datasets was collected and prepared, data preprocessing methods such as noisy data cleaning, handling missing values, handling categorical data and FS methods have been done in order to get a very clean and correct datasets to get better early detection and predicting HD ML model.

3.4.2.1. Noisy Data Cleaning

In this study's experimental work noisy data or outliers cleaning was appropriately done in the data preprocessing section because cleaning noisy data or outlier is a crucial part in data preprocessing step before the process of model building phase.

3.4.2.2. Handling Missing Values

Some datasets can have their own missing values through different cases such writing important diagnostic information in easy hand writing format [14] or missed through the case of privacy purpose or missed by the case of missing on the data base. Thus, handling missing values was performed for the separated datasets and unfortunately those datasets has no missing values.

3.4.2.3. Handling Categorical Data (Data Transformation)

Here in this experimental work, with in data transformation step we have tried to convert all the nominal data in to numerical format using dummy variables encoding because Data transformation is used to transform data's which exists in different format in to machine understandable and required format such as categorical data transform in to numerical data format. In this case, 'sex' having a categorical or nominal values of 'male'=1 and 'female'=0, 'chest pain type' having values of 'typical'=0, 'atypical'=1, 'non-angina pain'=2 and 'asymptomatic'=3, 'Fasting Blood Sugar' having nominal values of 1='True' and 0='False', 'Exercise Induced Angina' having a categorical values of 1='yes' and 0='no', 'Number of major blood vessels' having a categorical values of '0-3', 'thal' or 'defect type' having a value 0='normal' or 1='fixed defect' or 2='reversible defect', 3='non-reversible defect' and the last one is the 'target' having values 1='diseased' and 2='not diseased'.

3.4.2.4.Data Resampling

In this study's experimental work, the most common data resampling technique named as SMOTE (Synthetic Minority Over Sampling Technique) was applied to balance the class distribution of instances, to reduce biases and overfitting that happens in the case of majority data classes.

3.4.2.5.Feature Scaling

Feature scaling has been performed to make standardize the non-dependent variables of the datasets in to a specified range, i.e. in range between [-1 and 1]. In this work we have put the

numerical features in to the same scale and same range, to handle a variable dominates on other variable. In some ML algorithms feature scaling is mandatory such as in KNN, SVM, DNN or ANN [16], [14]. The descriptive feature mean in this study is scaled to 0 to a standard deviation of 1 using a typical scaling technique. The standardization equation of feature scaling is shown in equation 3.4.1 below:-

$$X' = \frac{X - mean(X)}{a}$$ *3.4.1. The standard feature scaling equation*

Where, X′ = New Value

X = Original Value

Mean (x) = Mean Value

a = Standard Deviation

3.4.2.6. Feature Selection Method

The main goal behind FS method is minimizing the dimensionality of the features or attributes and maximize the performance of prediction and prevents overfitting [48] [16]. Several researchers in different studies are using different FS methods in order to make their model more accurate in different datasets including HD datasets. In this study's experimental work, two FS method named as SFFS from WFS method and Chi-Square from FFS method was performed. The FS flow diagram for the prediction of HD shows in the figure 3.4.2.

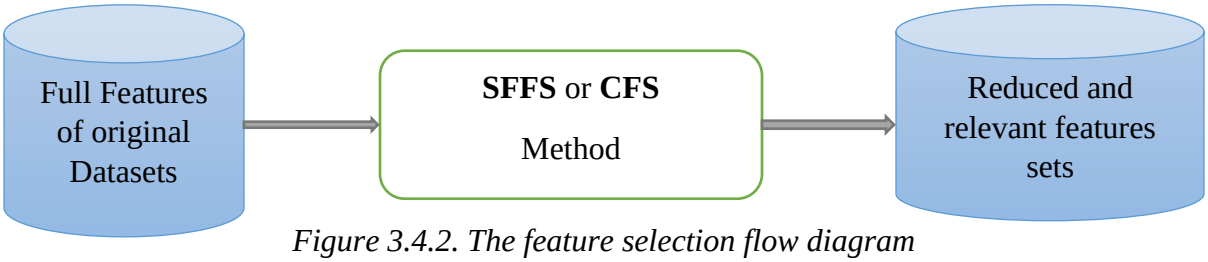

Figure 3.4.2. The feature selection flow diagram

Sequential Forward Feature Selection (SFFS): - is one of the FS method, applied to select features from an empty set of features to all other feature sets in a specified dataset iteratively. It adds features iteratively in order to check weather accuracy is improved or not. This procedure was go on until new features are added to check weather enhance the developed

model's performance. The performance of each added features are evaluated using PS and CV method. SFFS methods applied in this work shown as in the diagram 3.4.3 below:-

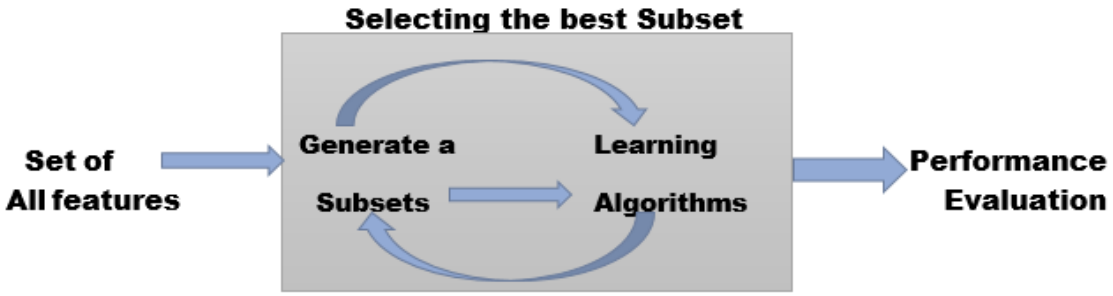

Figure 3.4.3.Sequential forward feature selection method architecture

Chi-Square Feature Selection (CFS):- In this study's experimental work CFS method was the second applied FS method which works by selecting the best features based on the univariate statistical tests named as Chi-square. In this process CF among the features and the target variable can be observed through the existence of relationship between features. Thus, CFS method was applied and different features having the higher values of chi-square are selected.

3.4.3. Machine Learning Model

The model building stage of the proposed study shows the early detection and prediction of HD using ML technique and FS method based on already classified or categorized data. These ML algorithms include classification techniques such as KNN and SVM and other EL techniques such as RF, VC and GB algorithms, considering a combination of SFFS and CFS method, through different performance evaluation metrics. All these ML classification and EL algorithms are chosen in order to get better model development for the prediction of HD datasets. These algorithms were popular in some of the previous works such as in [19], [25], [23], [9],[54],[55] and [16] and they were also better performed compared with other ML algorithms in predicting diseases. As discussed in section 3.2 of this chapter, two datasets are used for the experimental work i.e. WKURTH HD datasets collected from local hospital and CLbSH HD dataset collected from kaggle, then preprocessed separately and in combined ways to get very clean datasets. Finally models are evaluated through two different model evaluation techniques namely 10-F-CV and PS methods (data splitting in to 80% for training and 20% for

testing) for the individual datasets. The proposed early detection of HD enhancing prediction through ML model flow diagram for those separated datasets is shown in the figure 3.4.4.

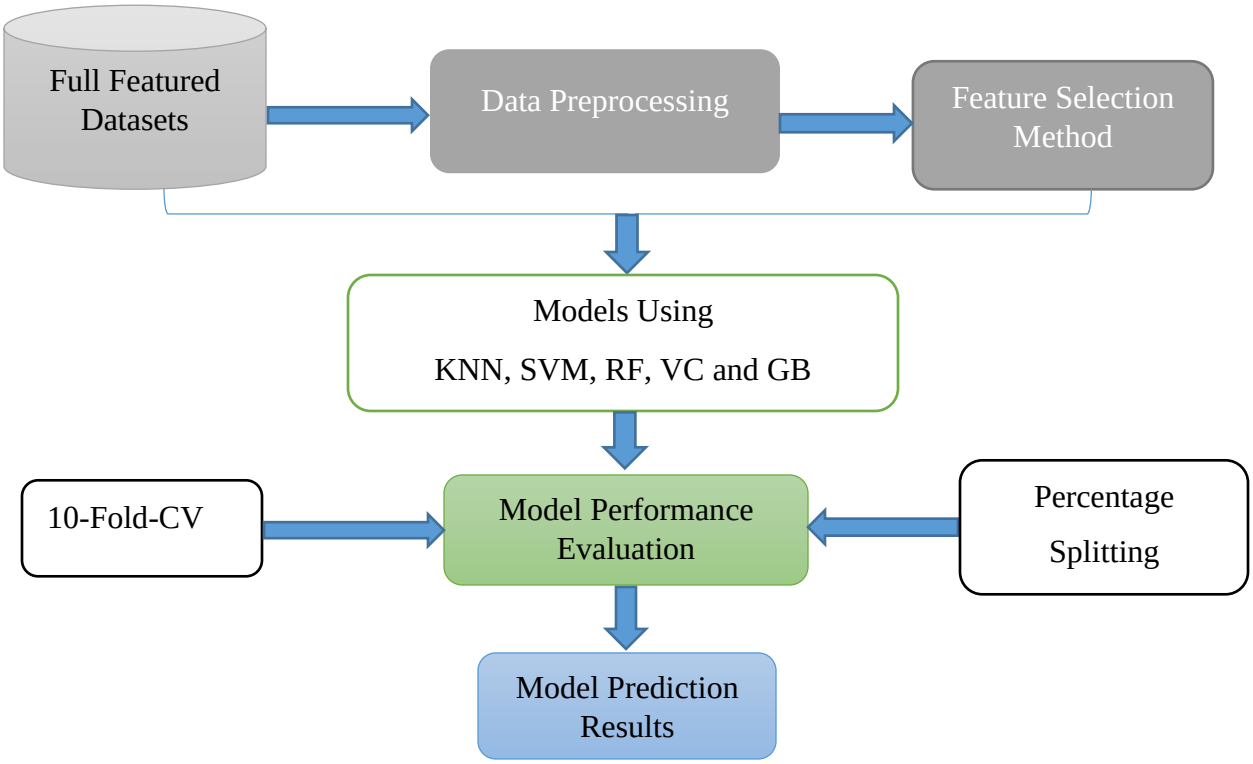

Figure 3.4.4. Heart disease prediction model flow diagram

The diagram shown in figure 3.4.4, shows the process of model development of HD prediction. In this study to develop HD prediction model five ML algorithm were applied such as KNN, SVM, RF, VC and GB algorithms. KNN is very simple and easy to implement and it uses different neighbor distance criteria. KNN requires no training period and it doesn't impact the accuracy. SVM is another ML classifier to develop the proposed HD prediction model on this study. It could be classified as linear, non- linear, polynomial, radial basic function or sigmoid SVM and it uses a certain kinds of hyperplane or sets of hyperplane to classify linear and nonlinear datasets. In this study we have tried to apply SVM with a linear-kernel trick for classifying both HD dataset separately and in combination form. This study's experimental set up also applies another ML model named as RF, which is a stable and achieves a higher accuracy. RF is also a very highly popular types of ML technique for classification of different tasks and allows an extra randomness by using several DT in parallel fitting technique called parallel ensemble technique. The other ML classifier and EL algorithm applied to develop HD

prediction model on both HD datasets individually and in combined ways was GB algorithm. GB is a types of EL technique which boosts or allows to convert weak classifiers in to strong ones. GB can also trains faster in a largest datasets and allows a predictors adding in to an ensemble and each of them corrects its predecessor predictor and it is proposed in this work to get the enhanced model. The last EL technique used to develop the proposed enhanced HD prediction model is VC. This classifier works by combining different classifiers for the same datasets by considering majority voting (MV). In MVC the base classifiers applied for model development are KNN, GB and RF. Finally, in order to get the most enhanced model for the detection of HD, ML algorithms i.e. RF, KNN, SVM, GB and VC (Ensemble of KNN, GB and RF) are applied for both datasets individually and in combined form.

3.5. Evaluation of Models

3.5.1.Model Prediction Performance Evaluation

Performance Evaluation is a process of evaluating models performance through different metrics and it is an essential phase in developing the most accurate model and very essential to another newly input datasets which are unseen. Performance evaluation technique could help to know which algorithm best fits the given datasets in order to solve a given problem. Mainly there are three methods of model performance evaluation techniques these are hold-out, percentage splitting and CV[16]. The hold-out evaluation technique is an evaluation technique that provides an unbiased estimates of learning performance. This technique is also used to check how much well is a ML model performs on unseen datasets and it uses training, validation and testing data's that are randomly split. Hold-out model performance evaluation technique is simple, flexible and faster than CV technique [16]. The CV model performance evaluation is a procedural step of splitting or partitioning the original datasets in to sample datasets and evaluating the ML model using another dataset samples to know the performance of that model. CV method is the most popular ways to avoid over fitting and it can be applied based on three ways: - validation, leave one out CV and K-F-CV [9]. The other splitting technique is percentage splitting which splits the whole datasets in to training and testing by using percent (%). In the proposed study we have to apply the two types of model validation evaluation methods, i.e. the K-F-CV (10-F-CV) which works by splitting the datasets in to K equal partitions called folds. In every step of 10-F-CV, 9 folds are utilized to train the model,

and one fold is used for testing. The second model evaluation technique to be applied is the percentage splitting technique. In this technique 80% of the individual dataset used for training and 20% for testing the model and finally these techniques are applied with FS method.

3.5.2. Models Performance Evaluation Metrics

Classifiers can use different prediction performance measurement methods in model developing process, in order to evaluate different models performance. Some of the model performance evaluation metrics used are accuracy, precision, recall, F1-score, sensitivity, specificity and other fundamental confusion matrices terms of measurement such as True Positive TP, True Negative TN, False Positive FP and False Negative FN.

3.5.2.1. Confusion Matrices

Confusion matrices (CM) is a types of performance evaluation metrics used here, showing us in the form of NxN matrix structure and used to evaluate the performance of a classification model where N is the number of classes to be predicted [56]. CM helps to understand how much the correct and non-correct prediction is done by the classifier. It also used to find the correctness of the model predicted by considering TP, FP, TN and FN and helps us in measuring the values of accuracy, precision, recall, f1-score specificity and sensitivity. This study mainly considers binary class target values named as case or control considering CM shown in figure 3.5.1 below.

		ACTUAL VALUES	
		Positive	**Negative**
PREDICTED VALUES	**Positive**	**TP**	**FP**
	Negative	**FN**	**TN**

Figure 3.5.1. Confusion Matrices of Binary Class

There are Four basic terms in Confusion Matrix:-

- ✓ TP: - The case of both Predicted value and Actual values being True.
- ✓ TN: - The case of both Predicted values and Actual values being False.
- ✓ FP: - The case of Predicted values being True & the Actual output values being False.
- ✓ FN: - The case of predicted values being False while the Actual o/p values being True.

3.5.2.2. Accuracy

Accuracy defines the capability of a ML techniques to predict the given datasets target class depending on the datasets independent classes correctly. It implies how much the predicted value is close to the definite or the actual value. Accuracy is also named as classification accuracy that is the division ratio of the correct prediction value to the total number of predicted value made by the prediction model [20]. Classification accuracy can be computed using the equation 3.5.1.

$$Accuracy=\frac{TN+TP}{TP+FP+TN+FN} \qquad \textit{3.5.1. Accuracy Calculating equation}$$

Where, TN+TP= Number of correct prediction

TP+FP+TN+FN= the total number of predictions made

3.5.2.3. Precision

Precision is defines as the true or the positive samples correctly identified in the actual class[16] of the prediction model. Precision is the ratio of TP values in the sample to the total positive samples predicted by the classifier. Precision can be calculated by equation 3.5.2.

$$Precision = \frac{TP}{TP + FP} \qquad \textit{3.5.2. Precision calculating equation}$$

3.5.2.4. Recall

Recall implies the TP values rates classified correctly[16]. Recall is the ratio of the number of TP values in the sample to the summation of TP samples and FN samples in the predicted data. The recall can be calculated in equation 3.5.3 below.

$$Recall = \frac{TP}{TP + FN} \qquad \textit{3.5.3. Precision calculating equation}$$

3.5.2.5.F1-Score

F1-score or F-measure is the best measure for the test accuracy of the developed model. It is the harmonic mean of Recall and Precision [57]. The higher the F1-score the better will be the performance of the developed model. The equation for F-measure shows in equation 3.5.4 below.

$$F1_{score} = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad \textit{3.5.4. F1-score calculating formula}$$

3.5.2.6. Sensitivity

Sensitivity implies the TP rate and it is the ratio of the TP actual samples to the summation of TP and FN values in a given data [8]. It also implies the positive samples are identified as positive or correct with respect to all positive sample data's given. Sensitivity can be calculated using the equation 3.5.5 below:-

$$Sensitivity = \frac{TP}{TP + FN} \qquad \textit{3.5.5. Sensitivity score calculating equation}$$

3.5.2.7. Specificity

Specificity implies the TN rate and it's the ratio of the TN actual values or samples to the summation of the TN values and FP samples in a given data [8]. Specificity can be the number of actual negative values or samples that are identified correctly with in the given data. Specificity can be calculated using the equation 3.5.6 below:-

$$Specificity = \frac{TN}{TN + FP} \qquad \textit{3.5.6. Specificity calculating equation}$$

3.6. Implementation Environment of the Prediction Model

In this study's experimental work to implement the model proposed, different implementation environment packages and libraries with different versions are used for the purpose of development stages from the scratch to the prototype development in Python. Python programming language is used to develop the model. Python is a versatile or general purpose programming language that can be used to create a variety of ML and AI projects, including mobile and web applications. Python programming language also helps to process ML images, texts and numbers using different formats through simple and easy ways. The Anaconda navigator used for this implementation had a version of 2.3.1. Which is also used to install and

launch different packages and libraries such as Jupiter note book, Sickie-learn, Pandas, Numpy, Matplotlib, Seaborrn and Flask server. Table 3.6.1 below shows all the versions and description of all the libraries, packages and tools used to develop the model.

Table 3.5.2.7.1. Tools and Package used for Implementation

Tools and Packages	***Version***	***Description of the Tools, Packages or library***
Anaconda Navigator	2.3.1	Support in launching applications and manages different packages, libraries, and environments for ML code building.
Jupiter Notebook	6.4.12	An open source web application used to create and share a computational documents or note book files and it is also used in ML for processing of data and model development.
Python	3.8.5	It is a general purpose and high level programming language for the development of several ML applications.
Sickie-Learn	1.0.2	A free ML library software module and used for developing several ML regression and classification tasks.
Pandas	1.4.4	Is a python programming language library software & used for data manipulation & analysis and helps to load data from external sources.
Numpy	1.21.5	An open source programming library and used for processing categorical or numerical data's.
Matplotlib	3.5.2	ML python programming library for the purpose of plotting of graphs for the usage of different results visualization.
Seaborne	0.11.2	It is a python library for the purpose of visualization of random distribution of statistical data.
Mlxtend	0.21.0	Is a ML extension library for the proposed model and for the day-to-day data science tasks
Flask(Server)	1.1.2	A frame work written in a python and used in building web application in python & applied here for the implementing and deployment of the proposed model prototype.

Microsoft Word	Word 2013	Is an application software and used to write the documentation of the study proposed.
Microsoft Excel	Excel 2013	An application software, used for configuration of the datasets used, .CSV file format preparation.

3.7. Deployment Environment

The experimental work of section 3.6 of this study, the packages and the tools are implemented and deployed on personal laptop computer having a computer properties with processor of Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz - 2.70 GHz, Installed RAM-4.00 GB (3.79 GB usable), 1000 GB HD storage and System type of Microsoft windows 10 prof with 64-bit operating system, x64-based processor.

CHAPTER FOUR

IMPLEMENTATION

This chapter mainly discusses the experimental implementation for the early detection of HD enhancing prediction through ML technique. Separated but similarly configured datasets such as public and local HD datasets was used separately and in combined ways for the purpose of implementation of the model development process. Five different ML algorithms are applied before and during FS method. Thus, all the ML model development implementations, data preprocessing implementations, the FS implementations and model evaluation implementations codes are included under this chapter.

4.1. Machine Learning Model Development Implementation

4.1.1.Dataset Preprocessing Implementation

In this study python programming language implementation plate form was used to apply all the ML codes developed for the prediction of HD, launching from the dataset preprocessing to the model development stage. In this stage after datasets are loaded, preprocessing methods such as handling missing values, categorical data transformation, feature scaling, balancing the imbalanced datasets using SMOTE technique and FS methods are done. Before loading the individual HD datasets separately the researcher imported different libraries for both local and public datasets in separated python implementation code, in order to develop the proposed early detection of HD. The python code for importing different libraries and packages are shown in Appendix C1. These all python libraries and packages are used and imported in first phase of model development process and some of these libraries and packages are also described in section 3.6 of this study. Here the three separated HD datasets are also loaded using the sample implementation code shown in Appendix C2.

4.1.1.1. Handling Missing Values Implementation

Handling missing values was discussed in the previous section 3.4.2.2, there is no missing values are detected in both local and public HD datasets. However the researcher applied an implementation code for checking if there are some missing values. The implementation code for checking missing values is shown in the Appendix C3.1.

4.1.1.2. Categorical Data Transformation Implementation

In this process all the categorical features in those of the HD datasets are converted to numeric values i.e. 0 and 1 format transformation, in order to have similar format of numeric values. The python sample code shown in Appendix C3.2 is the categorical data transformation implementation from categorical string values to numeric value by importing one hot encoder () method.

4.1.1.3. Applying Data Balancing techniques for Imbalanced data

Since the public and the local HD datasets having a little imbalanced target class values, we have applied SMOTE (Synthetic Minority Over Sampling Technique), in order to balance the binary target class values and this technique is applied in python implementation code shown in Appendix C3.3.

4.2. Feature Selection Implementation

As discussed in the previous sections of this study, FS implementation is a technique or a method to select the most relevant features from a subsets of original features. CFS and SFFS method are implemented in python implementation code. These methods used in order to improve the prediction performance of models selected by reducing some of irrelevant features from the original features of the proposed HD datasets.

4.2.1.Chi-Square Feature Selection Implementation

The implementation code for the CFS method has been implemented by using a package named as feature_selection through importing SelectKBest method. The implementation code for CFS method is shown in Appendix C4.1.

4.2.2.Sequential Forward Feature Selection Implementation

SFFS method is implemented in python plate form and applied by selecting features iteratively from an empty sets of features to other successive feature sets in the datasets until a better performance is achieved. The implementation code for SFFS method is computed in the sample python code shown in Appendix C4.2.

4.3. Machine Learning Model Implementation

In order to develop the proposed model the researcher applied five ML and EL algorithms. Before applying different models scikit_learn library package was installed and other different libraries were imported. Mlxtend frame work was also installed in order to import SFFS method from feature_selection module. All those packages, modules, libraries and other model evaluation metrics are implemented to develop the proposed model. The sample implementation code for these ML algorithms including RF, SVM, KNN, GB and VC are shown in Appendix C5 and Appendix C6. The first classifier was RF, which was implemented using a method named as RandomForestClassifier() with an ensemble module of Sklearn package, fitting different trees with n_esimator with a class of balanced values. The second classifier applied for model development is SVM. This classifier was implemented by using SVC () method found in the sklearn package, by importing linear-Kernel SVM (LinearSVC). The other classifier applied for the prediction of HD model development is KNN. KNN classifier was implemented in python plate form using KNeighborsClassifier() method and the classifier method found in sklearn package and imported as KNeighborsClassifier. The fourth ensemble ML classifier that we have applied is GB algorithm. The implementation code was applied on python plate form using GradientBoostingClassifier() method. GB is found in the sklearn package with in ensemble module and applied as GradientBoostingClassifier. The last and the final EL based classifier applied to do the experimental implementation of the model proposed is named as VC, which is an ensemble of n_estimator_classifiers of the base classifiers. This classifier is found in sklearn package and ensemble module imported as VotingClassifier in python plate form using VotingClassifier(estimators=classifiers) method. The implementation code for all these models is shown in sample code found in Appendix C5.

4.4. Model Testing and Evaluation Implementation

Model testing and evaluation is a very interesting concept in ML model development implementation stages. In this study we have applied several model evaluation metrics such as confusion matrix, precision, recall, accuracy, F1_score, sensitivity and specificity. As we have discussed before we have implemented model testing and evaluation metrics with two dataset splitting and evaluation techniques applied for the three separated datasets individually. These technique are PS technique (80% for training and 20% for testing) and 10-F-CV technique.

The 10-F-CV technique is found in a sklearn package with in model selection module and imported as cross_val_score using the applied fitted model. The implementation code for 10-F-CV and PS techniques are shown in the python code shown in Appendix C5 and C6 respectively.

4.5. Prototype System

The prototype system is a user interface form integrated saved and deployed on the flask web server. First, the web app python code (API) was created to load the better model, to input user data from the created HTML template form and to make the prediction. The HTML (Hyper Text Markup Language) template for the front end allows users to insert or input HD predicting features and displays the result if an expected patient have a disease or not. The user interface form helps to insert the values of features to predict HD is shown as in figure 4.5.1 below. As we have discussed in section 3.5.2, different model evaluation metrics are predicted and results are registered. This prototype could be helps physicians to insert independent features to predict the dependent target class as diseased or not. This interface holds 12 minimized and independent features with 1 dependent target feature. The main aim of this work is predicting HD in an expected patient, in order to detect the disease early. This prototype could supports users or experts classify the outcome as diseased or not. The sample code for integrating the model with in Flask server is shown in Appendix C7.

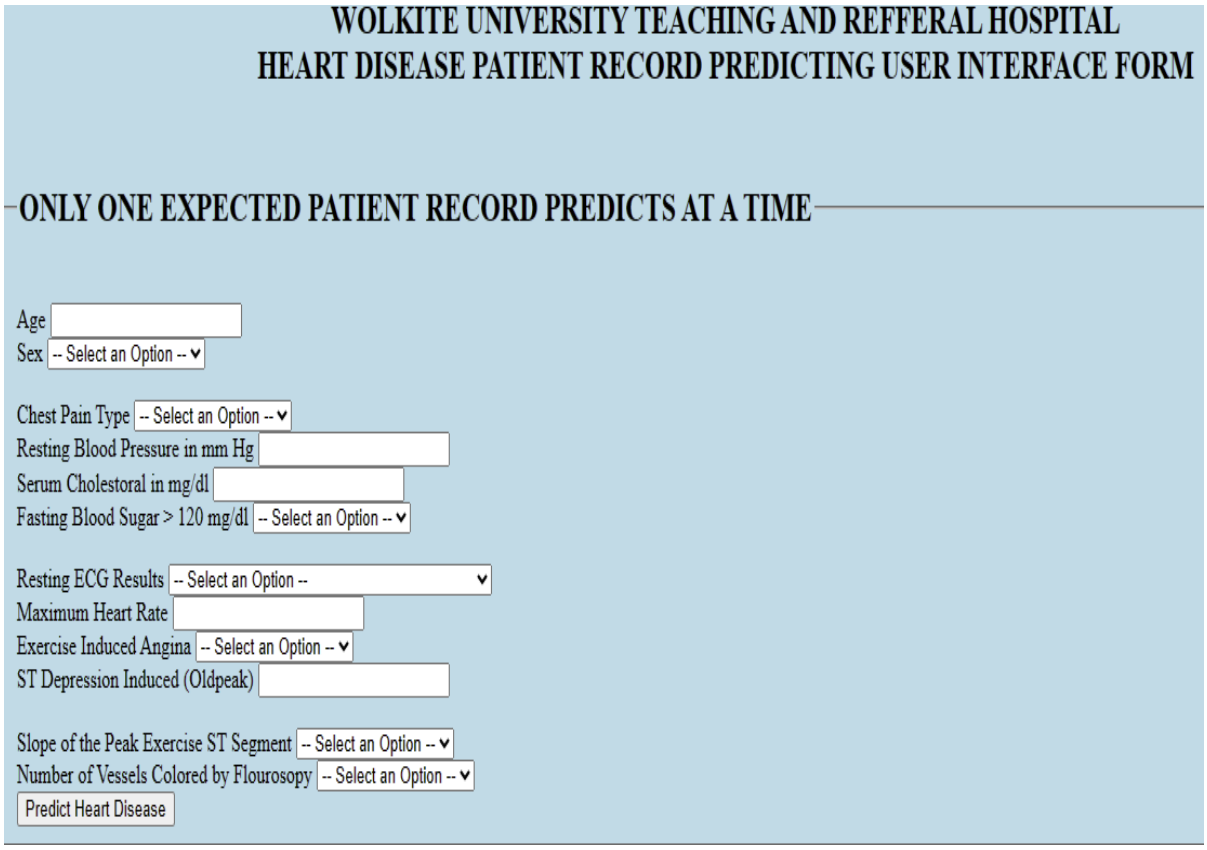

Figure 4.5.1. User Interface for data Input

CHAPTER FIVE

RESULT AND DISCUSSION

In this chapter we have tried to discuss the results or the outcome of the implementation or the experimentation work using five ML techniques with FS and with full features to develop the proposed early detection of HD enhancing prediction through ML techniques. We have discussed the data collection and preprocessing results, the model building and evaluation results of the original datasets and all the results of FS method are also discussed.

5.1. Dataset Collection and Data Preprocessing Result

As we have discussed in the previous chapter of this study, the researcher have used two separated but similarly configured, i.e. public and local HD datasets. The first HD dataset used for the prediction model development were the public HD datasets found in Heart Disease Dataset | Kaggle repository, having 1025 instances with 14 features. In order to keep similar setup and to have no discriminated and biased sample through different sample size with Local datasets, we have taken random number of samples similar with local datasets with an instances of 774 with 14 features. The instance of this public HD datasets in this experimental work were categorized as 390 instances as heart diseased patient and having 384 instances as non-heart diseased. The second local (WKURTH) HD datasets having an instances of 774 with 14 relevant features was also applied. After we have implemented all the data preprocessing method no data was dropped as no missing values or instances were exist. So after data preprocessing method the datasets was still the same to instances of 774 with instances 415 with HD and 359 instances of expected patients with no HD. After we have get all the results of these two separated datasets, we have combined the two original HD datasets in to one and preprocessed to develop an HD detection model. This dataset named as the combined HD dataset having instances of 1799 with 14 features. After the datasets is preprocessed, it is assured that the dataset have 941 instances of HD patients and 858 instances with no HD. Finally all the non-balanced binary class distribution of the two separated and the individually combined HD datasets are shown in the figure 5.1, 5.2 and 5.3 below respectively:-

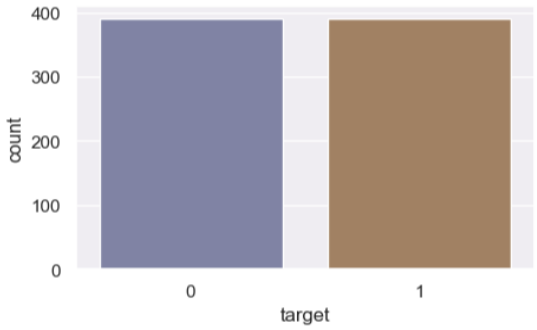

Figure 5.1.1. The binary class distribution of Public heart disease dataset

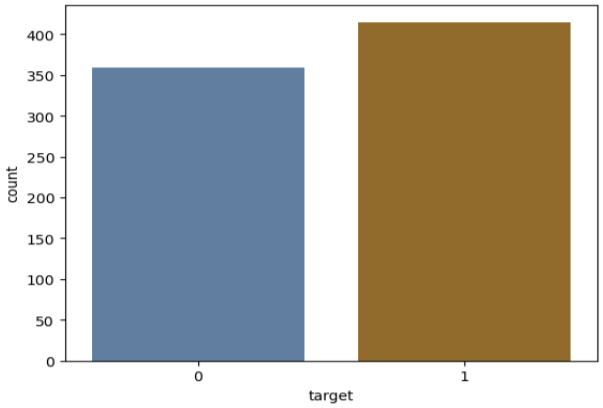

Figure 5.1.2. The binary class distribution of Local WKUTRH HD dataset

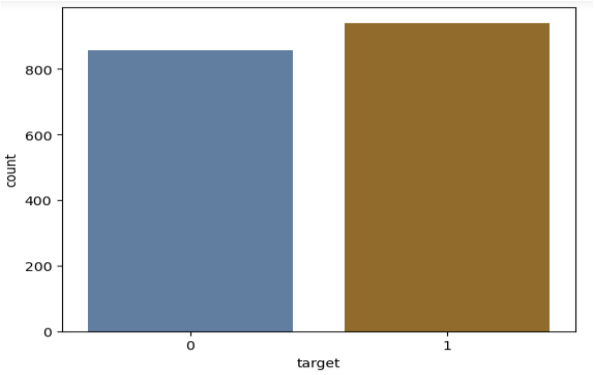

Figure 5.1.3. The binary class distribution of the combined HD dataset

In the first figure 4.6.1 above, the binary class distribution of the public HD dataset showing 50.4% (390 with HD instances) and 49.6% **(**384 with no HD instances), while the second figure 4.6.2, Shows the binary class distribution of local WKUTRH HD datasets showing 53.6% (415

instances with HD) and 46.4% (359 instances with no HD). The last figure 4.6.3 above, shows the binary class distribution of the combination of the two separated datasets i.e. the combination of local datasets and public datasets having a class distribution of 52.3 % (941 instances with HD) and 47.7% (858 instances with no HD). Since all the three HD datasets with a binary class distribution are not balanced, data resampling technique named as SMOTE (Synthetic Minority Oversampling) technique was applied to balance the class distribution of instances. In order to reduce the bias of imbalanced classes, we have to convert the minority class distribution in to majority classes. So after applying SMOTE technique on the proposed three separated datasets, the datasets had been balanced.

The first dataset having the same set up with local dataset have instances of 774 with 390 HD and 384 with no HD instances is balanced using SMOTE technique to 390 diseased and 390 non-diseased instances with a total balanced instances of 780. The second dataset having 774 instances with 415 HD and 359 with non-HD instances is balanced using SMOTE technique to 415 diseased and 415 non-HD with a total balanced instances of 830. The last combined dataset having 1799 instances with 941 instances with HD and 858 non-HD was being balanced using SMOTE technique to 941 diseased to 941 non-diseased instances with a total of balanced instance of 1882. The experimentation results of balanced class distribution of the three separated HD datasets is shown in figures 5.4, 5.5 and 5.6 below.

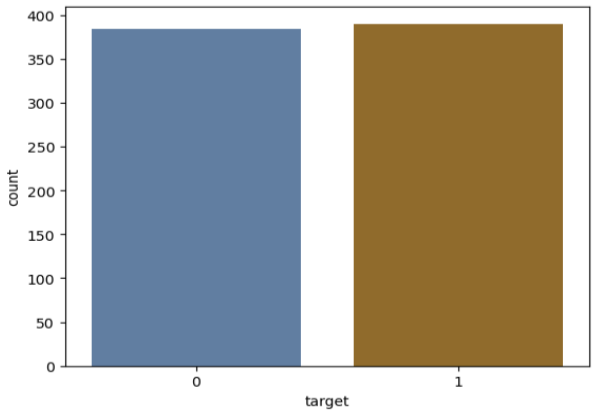

Figure 5.1.4. Public Heart disease datasets balanced target class value

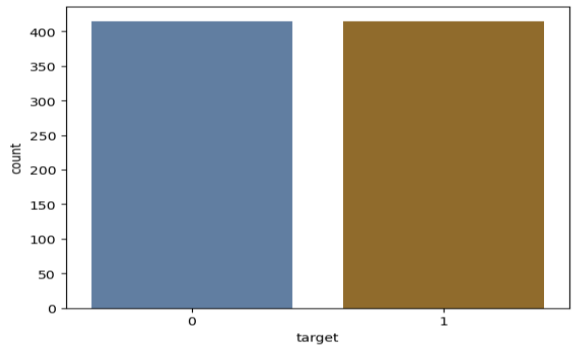

Figure 5.1.5. WKUTRH Heart disease datasets balanced target class value

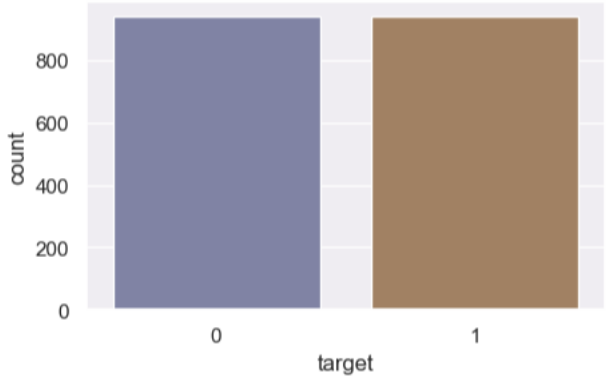

Figure 5.1.6. The Combined Heart disease dataset balanced target class value

The three figures above shows the class distribution of the three separated HD datasets after balancing or SMOTE techniques is applied, in order to balance the imbalanced heart disease dataset target classes.

5.2. Model Building and Evaluation Results on the Original Datasets

The model building stage for the HD detection and prediction is implemented using five ML algorithms i.e. RF, SVM, KNN, GB and VC. After all the data preprocessing methods are done, 10-F-CV and PS technique with ML algorithms are applied for the three separated datasets before and during FS methods. Models are evaluated through different model evaluation metrics for the individually balanced and non-balanced HD datasets. The next table shows the performance evaluation metrics results of each of the ML models for the original and separated HD datasets during PS and 10-F-CV method. Such model performance

evaluation matrices includes are Accuracy, Precision, Recall, F1-score, Sensitivity and Specificity and finally the results are shown in table 5.1 below:-

Table 4.2.2.1.The PS and 10-F-CV evaluation results of the three original HD datasets (balanced and non-balanced datasets)

Separated HD Datasets	***ML Model***	***Resampling***	***Splitting through***	***Performance Evaluation Metrics***					
				Accuracy	***Precision***	***Recall***	***F1-score***	***Specificity***	***Sensitivity***
1. International HD Dataset	RF	Before	80%/20%	99.0	98.7	98.7	98.7	98.7	98.7
			10-F-CV	98.3	98.8	98.0	98.4	98.5	98.0
		After	80%/20%	**98.8**	98.8	98.8	98.8	98.8	98.8
			10-F-CV	**98.3**	99.0	**98.3**	98.5	98.4	98.2
	KNN	Before	80%/20%	86.5	85.2	**88.5**	86.8	84.4	88.5
			10-F-CV	84.7	83.8	88.1	85.8	86.7	88.0
		After	80%/20%	80.9	80.0	80.3	80.7	80.0	82.3
			10-F-CV	85.2	84.1	87.6	85.6	86.9	87.6
	SVM	Before	80%/20%	85.2	82.3	89.7	85.9	80.5	90.0
			10-F-CV	85.5	83.1	91.8	86.8	88.2	91.7
		After	80%/20%	85.8	86.8	83.5	85.2	80.0	82.3
			10-F-CV	85.0	82.3	89.5	85.7	87.2	90.0
	GB	Before	80%/20%	97.4	96.3	98.7	97.5	96.1	98.7
			10-F-CV	96.3	96.4	96.5	96.4	94.8	96.5
		After	80%/20%	93.2	89.5	97.5	93.3	89.2	97.5
			10-F-CV	96.5	96.2	97.0	96.5	96.0	97.1
	VC	Before	80%/20%	97.4	96.3	98.7	97.5	96.1	98.7
			10-F-CV	96.6	96.8	96.8	96.8	96.6	96.8
		After	80%/20%	94.4	91.7	97.5	94.5	91.6	97.5
			10-F-CV	96.5	96.1	97.0	96.5	95.9	97.0
2. Local WKURTH HD Dataset	RF	Before	80%/20%	98.7	99.6	97.6	98.8	99.6	97.6
			10-F-CV	97.3	97.9	97.1	97.5	97.8	97.0
		After	80%/20%	**98.8**	98.8	98.8	98.8	98.8	98.8
			10-F-CV	98.0	98.4	97.1	98.0	98.4	97.0
	KNN	Before	80%/20%	97.4	98.8	96.4	97.6	98.6	96.4
			10-F-CV	94.7	96.9	93.2	95.0	95.6	93.1
		After	80%/20%	98.8	99.7	97.6	98.8	99.6	97.6
			10-F-CV	94.8	98.2	91.3	94.5	97.6	91.3
	SVM	Before	80%/20%	98.7	98.8	98.8	98.8	98.6	98.8
			10-F-CV	98.2	98.8	97.8	98.3	98.1	97.8
		After	80%/20%	**99.2**	99.4	98.8	99.4	99.6	97.6
			10-F-CV	**98.2**	99.0	97.3	98.2	98.2	97.3
	GB	Before	80%/20%	96.8	97.6	96.4	97.0	97.2	96.4
			10-F-CV	97.2	97.4	97.3	97.3	96.9	97.2

		After	80%/20%	**99.2**	98.6	99.5	99.4	98.8	99.4
			10-F-CV	**98.2**	98.6	97.8	98.2	98.2	97.8
	VC	Before	80%/20%	98.0	99.6	96.4	98.2	99.8	96.4
			10-F-CV	97.5	98.6	96.8	97.7	97.8	96.7
		After	80%/20%	98.8	98.8	98.8	98.8	98.8	98.8
			10-F-CV	**97.8**	98.8	96.9	97.8	98.2	97.0
3. Combined HD Dataset	RF	Before	80%/20%	98.1	99.5	96.9	98.2	99.4	96.9
			10-F-CV	98.7	99.2	98.4	98.8	98.8	98.4
		After	80%/20%	**98.1**	99.5	97.1	98.3	99.4	97.1
			10-F-CV	**98.5**	99.4	98.5	98.2	99.1	98.4
	KNN	Before	80%/20%	88.9	91.0	88.3	89.6	89.6	88.3
			10-F-CV	85.4	86.7	85.5	86.0	86.8	85.5
		After	80%/20%	86.2	90.5	83.5	86.9	89.5	83.5
			10-F-CV	85.1	86.7	83.4	84.8	87.0	83.3
	SVM	Before	80%/20%	89.4	90.3	90.3	90.3	88.4	90.3
			10-F-CV	85.9	86.0	87.9	86.9	86.5	88.0
		After	80%/20%	84.4	88.1	82.5	85.2	89.5	83.5
			10-F-CV	85.7	85.1	87.6	86.2	86.0	87.6
	GB	Before	80%/20%	94.7	96.8	93.4	95.1	96.3	93.4
			10-F-CV	93.6	94.2	93.6	93.8	94.4	93.5
		After	80%/20%	93.4	95.9	91.7	93.8	95.3	91.7
			10-F-CV	94.2	94.9	93.4	94.1	94.3	93.4
	VC	Before	80%/20%	95.8	97.4	94.9	96.1	97.0	94.9
			10-F-CV	94.6	94.9	94.8	94.8	94.4	95.0
		After	80%/20%	94.7	97.0	93.2	95.0	96.5	93.2
			10-F-CV	95.2	96.0	94.5	95.2	95.5	94.5

The table 5.1, shows the PS and the 10-F-CV performance evaluation results of the three individually balanced and non-balanced HD datasets before FS method on the original datasets. Comparing the three HD datasets, different results are registered before and after resampling these datasets. Firstly, before balancing the first HD datasets the highest accuracy score was achieved using RF 99.0% by PS method in the first dataset, RF and SVM achieved a performance of 98.7% using PS in the second dataset and an accuracy of 98.7% achieved by RF using 10-F-CV in the combined datasets. However, in order to get unbiased performance of models, the results of balanced dataset was also evaluated. Then, after balancing the first dataset the highest accuracy score of 98.8% and 98.3% was achieved by RF using PS and 10-F-CV respectively. While both SVM and GB achieved a better accuracy score of 99.2% and 98.2% using PS and 10-F-CV respectively in the second local dataset. Similarly in the

combined dataset the highest and better performance achiever model was RF with accuracy scores of 98.1% and 98.5% using PS and 10-F-CV technique respectively.

When we compare all algorithms before FS is applied, the highest and the better accuracy score of **99.2%** was achieved in the **second** dataset using **SVM and GB** with PS. In addition to accuracy score other metrics was also evaluated such as precision, recall, f1-score, and specificity and sensitivity scores. So the better performance achiever models of SVM and GB has also achieved a good performance in others evaluation metrics. Thus, **SVM** achieved a Precision, Recall, F1-score, Sensitivity and Specificity score of 99.4%, 98.8%, 99.4%, 99.6% and 97.6% respectively. **GB** achieved a precision, recall, f1-score, sensitivity and specificity score of 98.6%, 99.5%, 99.4%, 98.8% and 99.4% respectively.

5.3. Results of Feature Selection Method

The FS process is applied using two FS methods, i.e. CFS and SFFS method. CFS selects the most importance features among the given original features with the target variables, observed through the existence of relationship between them. In other hand, in the SFFS method every features of the datasets are computed sequentially with the proposed ML algorithm and performance is evaluated. The SFFS process was done on the three separated HD datasets in parallel with the proposed ML algorithms that are RF, SVM, KNN, GB and VC algorithms. In the next table we have summarized the selected features of the three separated datasets with in the different FS methods with the proposed ML algorithms for binary class distribution.

Table 4.2.2.1. The selected features for the three datasets before and after resampling

Datasets Used	**Types of FS Used**	**ML** Algorisms	Original Number of Features	Selected Number of Features Before Resampling		Selected Number of Features After Resampling	
				(for PS method)	(for 10-F-CV splitting method)	(for PS method)	(for 10-F-CV splitting method)
1. Public HD Datasets	**CFS**	RF	14	13	13	13	13
		KNN	14	13	13	13	13
		SVM	14	13	13	13	13
		GB	14	13	13	13	13
		VC	14	13	13	13	13

	SFFS	RF	14	12	12	12	11
		KNN	14	6	6	6	7
		SVM	14	8	9	11	11
		GB	14	7	11	12	10
		VC	14	8	9	10	10
2. Local Heart Disease Datasets	**CFS**	RF	14	12	12	12	12
		KNN	14	12	12	12	12
		SVM	14	12	12	12	12
		GB	14	12	12	12	12
		VC	14	12	12	12	12
	SFFS	RF	14	12	5	3	13
		KNN	14	5	5	5	5
		SVM	14	12	9	10	12
		GB	14	5	7	11	10
		VC	14	5	5	10	8
3. Combined Heart Disease Datasets	**CFS**	RF	14	13	13	13	13
		KNN	14	13	13	13	13
		SVM	14	13	13	13	13
		GB	14	13	13	13	13
		VC	14	13	13	13	13
	SFFS	RF	14	4	12	13	13
		KNN	14	7	5	8	7
		SVM	14	9	10	11	8
		GB	14	11	11	12	11
		VC	14	12	12	12	12

5.3.1. Model Results Evaluation for Individual HD Datasets during FS

After we have implemented model evaluation for original individual HD datasets different results are registered. In order to minimize the individual HD dataset features and to achieve a better performance, applying FS method is better. FS method is a good solution to reduce dimensionality of a datasets with unselecting non-relevant features by strengthening the performance of ML algorithms. One of the major aim behind this work is minimizing the dimensionality of the separated HD dataset attributes or features and maximizes the prediction performance of models by preventing overfitting problems. During SFFS and CFS method the models applied are also evaluated using PS and 10-F-CV method. As we have evaluated models through performance evaluation metric for the original balanced and non-balanced dataset, here we have also applied and evaluated such performance evaluation metric including accuracy, sensitivity, specificity, precision, recall and F1_score. All the results are putt in table 5.3.a below:

Table 5.3.1.1.a. The PS and 10-F-CV evaluation results for the three HD datasets After FS method (balanced and non-balanced datasets)

Separated HD Dataset	*FS Method Used*	*ML Model used*	*Resampling*	*Selected Feature*	*Splitting through*	*Performance Evaluation Metrics*					
						Accuracy	*Precision*	*Recall*	*F1-score*	*Specificity*	*Sensitivity*
1. Public HD Dataset	CFS Method	RF	Before		80/20%	99.4	99.7	98.8	99.4	99.8	98.8
				13	10-F-CV	99.4	99.5	99.2	99.4	99.3	99.1
			After		80/20%	**98.1**	99.6	96.6	98.2	99.6	96.6
					10-F-CV	**99.2**	99.7	98.5	99.2	99.5	98.5
		KNN	Before		80/20%	83.2	86.4	82.4	84.3	84.3	82.4
				13	10-F-CV	81.4	80.0	84.4	82.1	80.2	84.4
			After		80/20%	84.6	88.9	82.8	85.7	87.0	82.8
					10-F-CV	81.3	80.0	83.6	81.7	81.1	83.5
		SVM	Before		80/20%	87.1	87.4	89.4	88.4	84.3	89.4
				13	10-F-CV	84.9	82.0	90.0	85.7	81.8	90.0
			After		80/20%	88.5	92.6	86.2	89.3	87.0	82.8
					10-F-CV	85.1	82.1	90.3	85.9	82.2	90.2
		GB	Before		80/20%	97.4	99.8	95.3	97.6	99.8	95.3
				13	10-F-CV	97.3	97.9	96.7	97.3	97.8	96.7
			After		80/20%	96.8	98.8	95.4	97.1	98.6	95.4
					10-F-CV	97.4	97.7	97.2	97.4	97.8	97.1
		VC	Before		80/20%	98.1	99.2	96.5	98.2	99.2	96.5
				13	10-F-CV	97.8	97.9	98.0	97.9	96.7	98.0
			After		80/20%	95.5	98.8	93.1	95.9	98.6	93.1
					10-F-CV	97.7	97.7	97.7	97.7	97.7	97.5
	SFFS Method	RF	Before	11	80/20%	99.3	99.5	99.5	96.1	99.2	99.4
				12	10-F-CV	96.7	98.5	97.2	96.9	97.4	97.1
			After	**12**	80/20%	**99.3**	99.3	99.4	99.3	99.1	99.3
				11	10-F-CV	**97.5**	97.5	98.1	97.7	97.6	98.0
		KNN	Before	6	80/20%	87.6	84.7	87.7	88.9	83.8	87.7
				6	10-F-CV	86.1	82.1	87.5	87.7	81.8	87.4
			After	6	80/20%	87.4	85.6	92.3	88.1	87.2	92.3
				7	10-F-CV	86.3	84.8	93.5	87.3	82.8	93.4
		SVM	Before	8	80/20%	85.8	83.1	93.2	87.9	82.8	93.0
				9	10-F-CV	85.6	83.3	92.0	87.1	83.1	92.0
			After	11	80/20%	86.3	82.3	91.7	87.1	87.1	91.8
				11	10-F-CV	86.3	80.9	92.0	87.1	79.2	91.8
		GB	Before	7	80/20%	99.0	99.1	99.4	99.1	98.8	99.3
				11	10-F-CV	95.2	95.1	95.4	94.8	94.6	95.4
			After	12	80/20%	99.1	98.8	99.4	99.1	98.0	99.5
				10	10-F-CV	95.7	95.9	96.9	95.7	95.8	97.0

		VC	Before	8	80/20%	99.6	99.6	99.5	96.3	99.6	99.4
				9	10-F-CV	96.1	97.2	96.6	96.3	97.0	96.5
			After	10	80/20%	**99.3**	99.4	99.4	99.5	99.1	99.3
				10	10-F-CV	96.3	95.9	97.2	96.3	96.1	97.1
2. Local Hospital HD Dataset	CFS Method	RF	Before		80/20%	98.1	98.8	97.6	98.2	98.6	97.6
				12	10-F-CV	97.2	98.1	96.6	97.3	98.0	96.5
			After		80/20%	**99.4**	98.8	99.5	99.4	98.9	99.5
					10-F-CV	97.8	98.6	97.1	98.0	98.4	97.0
		KNN	Before		80/20%	96.8	97.6	96.4	97.0	97.2	96.4
				12	10-F-CV	95.3	97.6	93.7	95.5	97.4	93.6
			After		80/20%	94.6	96.1	92.4	94.2	96.6	92.4
					10-F-CV	95.7	98.0	93.2	95.5	97.8	93.1
		SVM	Before		80/20%	98.1	97.6	98.8	98.2	97.2	98.8
				12	10-F-CV	98.2	98.8	97.8	98.3	98.2	97.8
			After		80/20%	**97.6**	95.2	98.5	97.5	96.6	92.4
					10-F-CV	**98.3**	98.8	97.8	98.3	97.8	97.8
		GB	Before		80/20%	97.4	97.6	97.6	97.6	97.2	97.6
				12	10-F-CV	97.3	97.4	97.6	97.5	96.9	97.5
			After		80/20%	98.2	96.3	98.8	98.1	96.6	98.8
					10-F-CV	98.0	98.3	97.6	98.0	98.0	97.5
		VC	Before		80/20%	98.7	99.6	97.6	98.8	99.6	97.6
				12	10-F-CV	97.4	98.6	96.6	97.6	98.4	96.6
			After		80/20%	**99.4**	98.8	99.6	99.4	98.9	99.6
					10-F-CV	**97.7**	98.6	96.9	97.7	97.8	97.0
	SFFS Method	RF	Before	3	80/20%	99.5	99.5	99.5	98.8	99.4	99.5
				5	10-F-CV	98.5	98.2	98.5	98.6	98.3	98.5
			After	12	80/20%	**99.3**	99.4	99.3	99.2	99.3	99.2
				13	10-F-CV	**99.0**	99.1	98.8	98.9	98.9	98.8
		KNN	Before	5	80/20%	98.2	98.2	98.2	98.3	98.2	98.2
				5	10-F-CV	98.2	98.2	98.2	98.3	98.4	98.2
			After	5	80/20%	98.0	98.3	98.5	98.0	98.2	98.4
				5	10-F-CV	98.1	98.5	98.5	98.0	98.0	98.4
		SVM	Before	12	80/20%	98.1	99.4	97.3	98.2	99.0	97.2
				9	10-F-CV	97.4	97.6	94.0	97.1	97.6	94.1
			After	10	80/20%	98.2	99.0	97.9	98.2	98.8	98.0
				12	10-F-CV	97.7	98.2	97.3	97.7	97.8	97.2
		GB	Before	7	80/20%	99.5	99.5	99.5	99.5	99.6	99.4
				5	10-F-CV	98.5	98.5	98.2	98.6	98.8	98.0
			After	11	80/20%	**99.3**	99.3	99.3	99.4	99.4	99.3
				10	10-F-CV	98.6	98.2	98.8	98.6	98.2	98.9
		VC	Before	5	80/20%	99.5	99.5	99.3	99.6	99.6	99.3
				5	10-F-CV	98.5	98.5	98.2	98.6	98.4	98.1

			After	10	80/20%	**99.3**	99.3	99.1	99.2	99.1	99.0
				8	10-F-CV	**99.0**	98.5	98.8	99.1	98.6	98.8
3. Combined HD Dataset	CFS Method	RF	Before		80/20%	98.6	99.5	98.0	98.7	99.4	98.0
				13	10-F-CV	98.7	99.0	98.5	98.8	98.6	98.5
			After		80/20%	**99.5**	99.4	99.4	99.4	99.5	99.4
					10-F-CV	**98.4**	98.4	98.4	98.5	98.3	98.4
		KNN	Before		80/20%	85.6	88.7	84.2	86.4	87.2	84.2
				13	10-F-CV	84.7	87.0	83.9	85.2	86.9	84.0
			After		80/20%	89.2	89.4	86.9	88.1	91.1	86.9
					10-F-CV	84.6	86.8	82.3	84.3	85.0	82.3
		SVM	Before		80/20%	88.6	90.6	88.3	89.4	89.0	88.3
				13	10-F-CV	85.5	86.0	87.2	86.5	85.8	87.1
			After		80/20%	87.0	83.9	89.1	86.4	91.1	88.9
					10-F-CV	85.2	84.7	87.0	85.7	84.8	86.9
		GB	Before		80/20%	93.3	93.0	94.9	94.0	91.5	94.9
				13	10-F-CV	92.8	93.0	93.6	93.2	92.9	93.7
			After		80/20%	95.2	92.4	97.7	95.0	93.1	97.7
					10-F-CV	93.2	93.0	93.7	93.3	93.3	94.0
		VC	Before		80/20%	95.0	94.5	96.5	95.5	93.3	96.4
				13	10-F-CV	94.3	93.7	95.7	94.7	92.8	95.6
			After		80/20%	96.0	93.5	98.3	95.8	94.1	98.3
					10-F-CV	94.2	93.5	95.0	94.2	93.7	94.8
	SFFS Method	RF	Before	4	80/20%	99.5	99.5	99.5	98.8	99.6	99.2
				12	10-F-CV	99.2	99.5	99.1	99.2	99.6	99.0
			After	13	80/20%	**99.2**	99.3	99.1	99.4	99.2	99.1
				13	10-F-CV	**98.7**	99.1	98.8	98.7	98.7	99.0
		KNN	Before	7	80/20%	89.9	88.2	93.2	90.4	88.2	93.0
				7	10-F-CV	88.8	87.2	93.9	89.3	90.1	94.0
			After	9	80/20%	89.4	86.2	93.1	89.7	87.8	93.2
				8	10-F-CV	89.1	86.2	93.1	89.3	88.2	93.2
		SVM	Before	9	80/20%	85.9	85.0	91.3	86.9	85.8	91.3
				10	10-F-CV	85.3	83.8	91.3	86.2	88.5	91.3
			After	11	80/20%	84.9	84.1	92.2	85.3	86.8	92.1
				8	10-F-CV	84.9	83.9	92.2	85.0	85.4	92.2
		GB	Before	11	80/20%	97.0	96.3	97.9	97.1	97.2	98.0
				11	10-F-CV	94.6	94.1	95.2	94.6	94.5	95.1
			After	12	80/20%	96.8	96.8	96.9	96.8	96.6	97.0
				11	10-F-CV	94.0	93.8	94.4	93.6	94.0	94.5
		VC	Before	12	80/20%	98.9	98.8	99.2	98.9	98.2	99.0
				12	10-F-CV	97.0	98.1	97.7	97.1	98.1	97.7
			After	12	80/20%	98.8	98.7	99.5	98.5	98.5	99.4
				12	10-F-CV	96.8	97.6	95.5	96.8	97.6	95.5

Table 5.3.a, above shows the PS and the 10-F-CV evaluation results of the three balanced and non-balanced HD datasets during applying two FS method. During FS technique for the first public HD dataset before resampling technique was applied, SFFS with VC (selecting 8 features) using PS achieved a highest accuracy of 99.6% and CFS with RF also achieved a highest accuracy score of 99.4% using 10-F-CV. However, after applying resampling technique, CFS with RF (selecting 13 features) using PS and 10-F-CV method achieved a better accuracy of **98.1%** and **99.2%** respectively. While in other hand, SFFS with RF (selecting 12 features) and SFFS with VC (selecting 10 features) also achieved highest and equal accuracy score of **99.3%** using PS and at the same time SFFS with RF (selecting 11 features) also achieved a good accuracy score of **97.5%** using 10-F-CV, comparing all algorithms for SFFS method for this Public dataset. Similarly other performance evaluation metric was also evaluated and different results are registered. However, the Better accuracy score achiever methods **(SFFS with RF and VC) also** achieved better precision, recall, f1-score, sensitivity and specificity scores, i.e. SFFS with RF achieved 99.3%, 99.4%, 99.3%, 99.1% and 99.3% while SFFS with VC achieved 99.4%, 99.4%, 99.5%, 99.1% and 99.3% of precision, recall, f1-score, specificity and sensitivity scores using PS method respectively.

During these two FS process time, for the second local HD dataset before resampling technique was applied, SFFS with RF (selecting 3 features), SFFS with GB(selecting 7 features) and SFFS with VC (selecting 5 features) all achieved an accuracy scores of 99.5% and 98.5% using PS and 10-F-CV method respectively. However, after applying resampling technique, different performance results are registered, i.e. both CFS with RF (selecting 12 features) and CFS with VC (selecting 12 features) using PS method achieved a better accuracy scores of **99.4%** and CFS with SVM (selecting 12 features) achieved a good accuracy scores of **98.3%** using 10-F-CV method. In other hand SFFS with RF (selecting 12 features), **SFFS with GB** (selecting 11 features) and **SFFS with VC** (selecting 10 features) all achieving a respective and a good scores of **99.3%** using PS technique, while **SFFS with RF** (selecting 13 features) and SFFS with VC (selecting 8 features) both achieving a good scores of **99.0%** using 10-F-CV. In addition to accuracy score evaluation, other performance evaluators such as Precision, Recall, F1-score, Sensitivity and Specificity are also evaluated and better results are achieved i.e. **CFS with RF** achieved respective scores of 98.8%, 99.5%, 99.4%, 99.5% and 98.9%. While for **CFS with VC** achieved scores of 98.8%, 99.6%, 99.4%, 99.6% and 98.9%.

In similar way, when FS methods were applied on the combined HD dataset before and after resampling the dataset, different results are achieved. First, before resampling technique were applied on the combined dataset during FS, SFFS with RF (selecting 4 and 12 features) using PS and 10-F-CV achieved a respective and highest scores of 99.5% and 99.2%. However, after applying resampling technique on this dataset, CFS with RF (selecting 13 features) achieves an accuracy of **99.5%** and **98.4%** using PS and 10-F-CV method respectively. In other hand for this combined dataset, SFFS with RF (selecting 13 features for PS and 10-F-CV) achieved good accuracy scores of **99.2%** and **98.7%** respectively. Similarly, comparing through these applied algorithms, other performance evaluation techniques such as Precision, Recall, F1-Score, Sensitivity and Specificity are also evaluated for the combined dataset and a better results for **CFS with RF** (selecting 13 features) of 99.4%, 99.4%, 99.4%, 99.4% and 99.5% respectively was achieved using PS.

Finally, comparing all the ML algorithms for these three balanced HD datasets using two FS method namely **CFS and SFFS** method, different performance evaluation results are achieved. The higher accuracy score **using PS** was achieved by **CFS with RF** (selecting 13 features) with scores of **99.5%,** achieved in the combined HD dataset. Similarly the highest accuracy score using **10-F-CV** was achieved by **CFS with RF** (selecting 13 features) with in the first HD dataset achieving a good accuracy score of **99.2%.** However, in other hand some ML algorithms with the two FS method, achieves a lowest and lesser performance score compared with other ML techniques applied on the three separated HD datasets. Such as KNN and SVC achieved a lowest scores with in the first and on the combined datasets, achieving less than 90%, however in the second datasets a little good scores are registered comparing in the first and on the combined datasets achieved.

As discussed in the previous section of this study, other model performance evaluation metrics such as CM besides sensitivity, specificity, precision, recall, F1_score and others terms such as TP, TN, FP and FN values was also evaluated. The results of all CM of each model according to the given separated HD datasets are evaluated and the results are computed. However, we have only put the CM results of the combined HD dataset models using CFS method applying PS, which have achieved a better accuracy score. In this HD dataset the CM and the normalized confusion matrix (NCM) results, showsing how classifiers correctly

classified each class with prediction errors using CFS methods are shown as figure 5.7 below, which shows the CM and the NCM of the classifiers in the proposed combined and balanced HD dataset, using CFS method:-

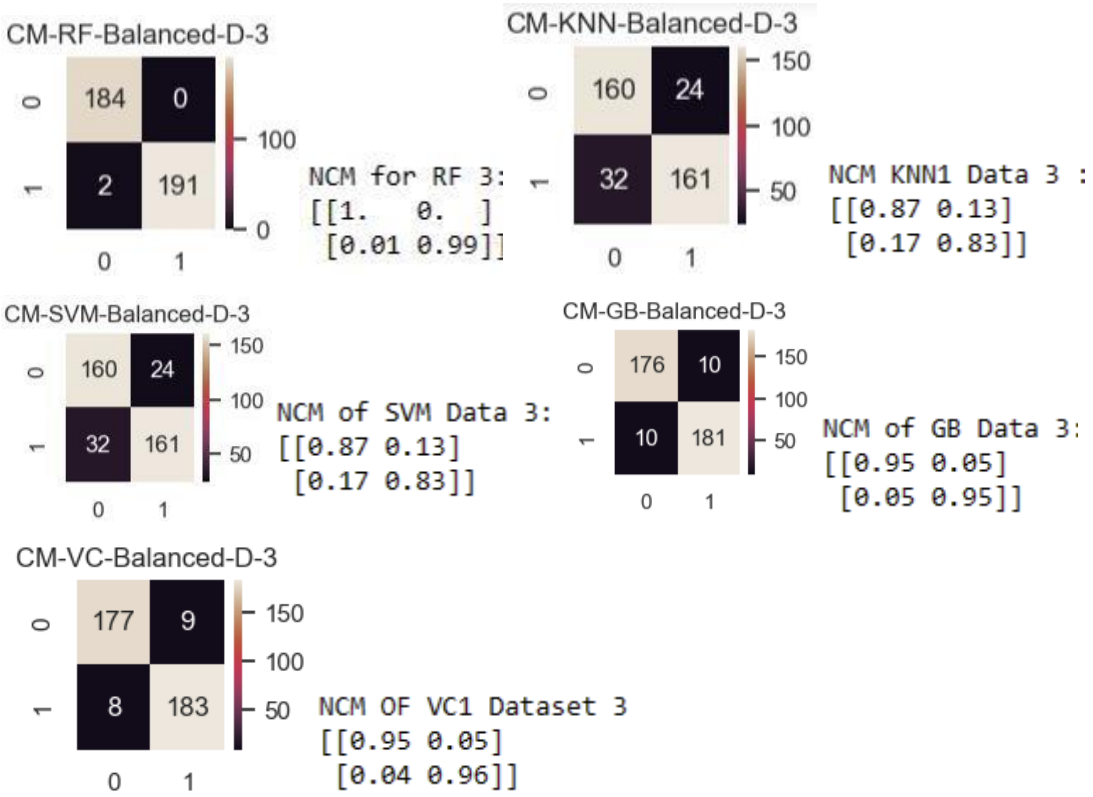

Figure 5.3.1. The CM and NCM of the Five Classifiers Using CFS on dataset 3

Figure 5.7, above shows the NCM and the CM of RF, KNN, SVM, GB and VC after we have applied Chi-Square feature selection (CFS) method on the dataset 3 (Combined HD dataset). RF classifier classifies 100% of not-diseased and 99.0% of diseased correctly and 0% of not diseased as diseased and 1% of diseased as not diseased are misclassified after we have applied CFS on the proposed dataset. Both KNN and SVM classifies 87.0% of not-diseased and 83.0% of diseased correctly and 13% of not diseased as diseased and 17% of diseased as not diseased are misclassified. Similarly, GB also classifies 95.0% of not-diseased and 95.0% of diseased correctly and 5% of not diseased as diseased and 5% of diseased as not diseased are misclassified, after we have applied CFS method on the proposed combined HD dataset. Finally, comparing all these classifiers through this dataset, RF was achieved a better normalized CM of 99.5%.

As we have discussed in the previous sections of this study, we have proposed an early detection of HD enhancing prediction performance through ML technique, by applying different ML and FS method on two different but similarly configured Public and Local HD datasets. First, the public HD datasets having 1025 instances is reduced to the same number of sample to local datasets in order to keep similar setup with local datasets having 774 instances, with both having the same number of 14 features. Then, both local and public HD datasets are combined to get the third dataset in order to check if the ML models prediction performance have been enhanced. However, we have used the original number of sample size of public datasets of 1025 instances to combine with local datasets. So the combined dataset have an imbalanced instances of 1799 with 14 features and finally balanced to sample size of 1882 instances.

Within this study we have applied five ML algorithms and two FS methods. The developed models was validated and evaluated using PS and 10-F-CV techniques. Model performance evaluation metrics such as accuracy, Precision, Recall, F1-score, Sensitivity and Specificity are also evaluated through those techniques. Initially ML techniques such as RF, KNN, SVM, GB and VC are applied on the three imbalanced HD datasets before FS techniques using PS and 10-F-CV. Then after, two FS methods such as CFS and SFFS methods are applied for the three initially imbalanced but later individually configured and balanced datasets, this is because of selecting the most relevant features for the prediction of HD. Different number of features are selected with in the three separated datasets using the two FS methods. Using CFS method for all proposed classifiers before and during resampling technique, 13, 12 and 13 features are selected for the three public, local and combined datasets respectively.

Similarly, for the three HD datasets applying SFFS method with all the proposed ML techniques before and after resampling, different features are selected to predict the disease in the given datasets. For the first balanced HD dataset applying SFFS method with PS technique RF selects 12, KNN selects 6, SVM selects 11, GB selects 12, and VC selects 10 features. While for this balanced dataset applying 10-F-CV technique RF selects 11, KNN selects 7, SVM selects 11, GB selects 10, and VC selects 10 features. For the second HD dataset using SFFS method and PS technique RF selects 12, KNN selects 5, SVM selects 10, GB selects 11, and VC selects 10 features. While using 10-F-CV technique RF selects 13, KNN selects 5,

SVM selects 12, GB selects 10, and VC selects 8 features. For the third HD dataset using SFFS method and PS technique RF selects 13, KNN selects 9, SVM selects 11, GB selects 12, and VC selects 12 features. While using 10-F-CV technique RF selects 13, KNN selects 8, SVM selects 8, GB selects 11, and VC selects 12 features.

When the two FS methods namely CFS and SFFS methods are applied, there are some features which were selected frequently in these datasets. In CFS method age, cp, chol, thalach, trestbps, exange, oldpeak, slope and ca features are selected frequently in all datasets. At the same time, using SFFS method sex, cp, chol, fbs, restecg, thalach, exange, oldpeak, slope, ca and thal features are selected frequently. Hence, from all the given datasets, seven (7) features were selected frequently using the two FS methods such as cp, chol, thalach, exange, oldpeak, slope and ca to predict HD. Specifically, for the combined HD dataset, the highest performance achiever model was RF (before FS) and RF with CFS and SFFS (during FS), thus different features are selected before and after FS method is applied. When using SFFS method, for the combined dataset with PS, all the 13 features were selected, however, performance is lower than CFS method. While, during CFS method was applied with PS, 12 features (age, sex, cp, trestbps, rest-ecg, chol, thalach, exange, oldpeak, slope, ca and fbs) were selected for the combined datasets by removing "thalasimia" feature. These 12 features are very relevant and important features to predict HD for the combined datasets and which are selected using RF model with CFS method. Appendix B, tables 0.1, 0.2, 0.3 and 0.4 shows the selected features using CFS and SFFS methods from the given 14 features in all the three HD datasets.

Although several studies are done related with ML technique and disease prediction, most of these works related with HD detection and prediction are done outside the country, Ethiopia. This shows almost there is a little studies are conducted in Ethiopia related with ML techniques with FS and HD. However, related works are done in the prediction of other diseases such as CKD [16] and heart related cases such as a literature in [25] proposed an ensemble classification method for prediction of HD risk and MV achieving an accuracy of 85.48% was the highest of all the applied classifiers and the researcher proposed MV technique for the prediction of HD risk. The proposed work in [21], proposed an EML method to improve ensemble technique for the prediction of HD risk and AB-WAE was achieved a better accuracy of 93% and 91% with in two different public datasets. A paper in [9] also proposed a study to

save life of HD patient by predicting HD over a few significant parameters of the heart and KNN and Novel-KNN methods achieved a respective accuracy of 88% and 93%[47] also proposed HRFLM method aims in finding significant features or attributes and resulting in improving the accuracy of prediction for CVD specially HD and achieved a better accuracy of 88.7% using HRFLM method. The other paper referred in [48] proposed an improved accuracy of heart attack risk prediction based on IGFS method assuming to boost a ML classifier performance and a better accuracy using SVM with FS achieved an accuracy of 88.9812% and RF with FS achieved an accuracy of 88.9812%. Similarly a paper in [21] proposed an identification of significant features and data mining methods in prediction of HD and finally a higher performing data mining algorithm having an accuracy of 87.4% was achieved by VC (ensemble of NB and LR) in predicting Cleveland HD datasets used after FS was applied. In the same way a paper work in [50] proposed an improved heart attack prediction by using ML techniques combined with FS method aiming in identifying the best model and best FS method, the experiment of this work shows that SVM(linear-kernel) achieves a better result in combination with relief-F FS method with accuracy scores of 84.81% to predict heart attack applied on Statlog HD dataset. It was also a paper referred in [23], proposed HD prediction method using ML, aiming in envision of the probability in developing HD in a specified patient through computerized predicting technique, finally the highest performance was achieved by using KNN, achieving an accuracy of 90.789 % to predict the Cleveland HD datasets. The last paper summarized here is[44], which proposed a HD prediction methodology using ML, two separated datasets such as Cleveland HD datasets having 303 instances and CLbSHS having 1190 instances with were used and different results achieved, however after combining these two datasets a better accuracy score of 93.31% was achieved using RF. Most of the above HD related literatures are done outside the country Ethiopia. In most of the works related with HD prediction outside Ethiopia, local datasets are not included and the FS they applied is not clearly identified and etc.

The proposed study here, which introduces a Detection of HD enhancing prediction performance through ML algorithms such as RF, KNN, SVM, GB and VC with two FS methods namely CFS and SFFS. To do the experiment two different HD datasets i.e. Local and Public HD datasets are employed separately and in combined ways. First, the experiment was done on the original and imbalanced datasets and then after done on the balanced datasets.

To validate the different model performance metric, two validation and splitting techniques such as PS and 10-F-CV techniques was applied on the two separated and combined datasets. Finally after balancing these three HD datasets, the two FS methods are applied and a better performing algorisms are registered. As balanced dataset can have a better effects in ML model, in case of reducing overfitting and biases, we have used balanced HD datasets before and after FS methods are applied and Performance Evaluation metrics such as Accuracy, Precision, Recall, F1-score, Sensitivity and Specificity are also evaluated and different results are registered.

Before FS methods are applied for the three balanced HD datasets, both SVM and GB achieved an accuracy of **99.2%** using PS, registered with in the second local HD dataset. Then after FS method are applied on the three balanced HD datasets, RF (selecting 13 features) with CFS achieved an accuracy of **99.5%** on the combined dataset and **99.2%** on the first dataset using PS and 10-F-CV respectively. While in other hand, SFFS with RF and SFFS with VC also achieved a highest accuracy of **99.3%** using PS in the first and second datasets and similarly both SFFS with RF and SFFS with VC achieved a better accuracy score of 99.0% using 10-F-CV method. So comparing all the results of ML algorithms without and with FS method, **SVM and GB** achieved a better accuracy score of **99.2%** using PS with in the second dataset and **CFS with RF** (selecting 13 features) have achieved a better an accuracy score of **99.5%** using **PS** with in the combined dataset**.** However different performance evaluation metrics such as Precision, Recall, F1-score, Sensitivity and Specificity scores are also evaluated and a better results are registered. Comparison between the proposed HD detection and prediction model and other previously related literatures on the prediction of HD are shown in the table 5.4.b.

Table 5.3.1.2.b. Comparison between the proposed models with the previous works

Reference	***ML Models Applied***	***Feature Selection method Used***	***The Model achieving the best accuracy score***
[17]	BN, NB, RF, C45, PART, MLP, MVC, Bagging and Boosting	BFFS method	85.48 % by MVC

[21]	CART, Ensemble of AB-WAE	No FS method	91% for Framingham and 93% for Cleveland
[9]	KNN and Improved KNN=K+PN	No FS method	91% by KNN and 93% by improved KNN
[47]	RF and LM separately, HRFLM	DT entropy based FS	88.7% by HRFLM
[55]	RF, KNN, NB, DT, LM, SVM.	IGFS method	88.982% by RF-IGFS and 88.981% by SVM-IGFS.
[55]	DT, NB, NN, KNN, SVM, LM and VC(NB and LM)	BFFS method	87.4% by VC
[55]	C45. Binary LR, SVM (Sigmoid, Linear, RBF, polynomial), CRT, KNN, ID3, NB, Multinomial LR and MLP	SBFS and SFFS, Filtering and Relief-F from FFS method	84.81% by Linear-SVM with Relief-F FS method.
[55]	DT, NB, KNN and RF	No FS method	90.789% by KNN
[55]	LM, NB, SVM, KNN and Ensemble of XGB.	No FS method	93.1% by RF for combined dataset.
The Proposed Work	**RF, KNN, SVM, GB and VC**	**SFFS** and **CFS**	Before FS**, SVM** and **GB** achieved accuracy of **99.2% &** After FS, **CFS with RF** achieved **99.5%,** both using PS method.

Both the two original dataset of the proposed work has 14 features including the result of the target value. However, the model which we recommended for the final model development holds 12 features of independent variables with 1 target values, using CFS with RF (selecting 13 features by unselecting thalasimia feature) for balanced datasets using PS method. Before applying FS method, SVM and GB achieved a good scores of 99.2% and finally after applying FS method RF with CFS have achieved a highest and a better accuracy scores of 99.5% using PS method.

CHAPTER SIX

CONCUSION, RECOMMENDATION AND FUTURE WORKS

In this chapter, the proposed early detection and prediction of HD model development using different ML technique and feature selection methods have been concluded. Recommendations and some other future directions for research works have been explained.

6.1. Conclusion

In this study, detection of HD enhancing prediction through ML technique is proposed in order to predict the disease early and prevent further distribution of the disease in the expected patient and helps health care experts or professionals to make early decision making. To choose the most pertinent features for the HD prediction, two distinct FS methods were used. The prototype of the better model was deployed to the flask server to help users or health care experts in early prediction of the disease in the expected patient faster.

Two similarly configured HD datasets are used separately and in combined ways in order to develop the proposed model. Before applying FS methods different data preprocessing methods are done in ML tools namely python. Unfortunately there were no missing values was observed, when data visualization techniques are done for these datasets. Other preprocessing methods such as categorical data transformation, feature scaling and data balancing techniques was applied for all the datasets. As there were a little data class imbalance in these datasets, data balancing technique namely SMOTE technique was applied for the three separated HD datasets. First we have used similarly configured balanced dataset instances of 780 instances for Public datasets and 830 instances for local datasets are used for the individual model development. Finally a total of balanced and combined instances of 1882 HD dataset was also used for the model development. The model development stages was done using five ML algorithms such as RF, KNN, SVC, GB and VC before and after CFS and SFFS methods are applied. The performance evaluation metrics are done through PS and 10-F-CV techniques on the three balanced and imbalanced datasets and finally different results are achieved for those datasets.

In this case, Before FS methods are applied for the three balanced HD datasets, **SVC** and **GB** achieved an accuracy scores of **99.2%** each using PS with in the second dataset. While after FS method is applied **RF with CFS (**selecting 13 features), achieves a better accuracy score of **99.5%** within the combined dataset using PS. In addition to this, other evaluation metrics was also evaluated and better results are registered. Finally **SVC and GB** before FS and **RF with CFS** was Chosen as better performance achieve using PS with in different datasets. As a result **RF with CFS** have been deployed on the flask server, which helps users or experts to predict HD in an expected patient record.

6.2. Recommendation

Heart disease is becoming a very challenging condition around the world including Ethiopia. It is also becoming a silent killer disease, which kills one-third of the world population in developing countries like Ethiopia. Some ML tools or techniques are not much accessible in Ethiopia. However, ML have a greater contribution in developing a model which supports health care experts in diagnosis and prediction of a certain diseases. Thus, the application of ML in disease detection and prediction especially HD is very interesting research area, because it provides tools to predict the disease and helps to stop its progress in an early stage. As it is very clear that there is a little integration of technology with health care systems in Ethiopia, it causes a great deal of stress for medical personnel and makes it challenging for researchers to conduct their work effectively. There was a challenge in data collection in local hospitals that is why every data is recorded in a manual paper based format, this leads to missing relevant records or features in every patient history. In local hospital we have also observed that experts are not using their computer as a decision support system, they are simply using papers only for decision making system. This leads other experts or researchers in difficulty of recognition of improper handwriting. The other recommendation for medical experts is that, if there is a standard software format for the recording of relevant features may help other users or academicals students or practitioners and researchers to find the data easily.

6.3. Future Works

In this study's experimental work some of the supervised and EL techniques are used to develop the proposed early detection of HD enhancing prediction through ML model with FS methods, however it may become better to see the performance of other deep learning or unsupervised learning approaches with a hybrid of other FS methods such as WFS methods with FFS methods. Secondly, the local HD dataset's found in hospitals in Ethiopia is manual paper based, which is another challenge in developing a good and accurate model. This is because the data is very poor, not well organized and not real-time data's. Thus, it is also better to use local datasets which are greater in number and organized in a better software standard format than this proposed dataset and future model can be developed using Image processing and deep learning approaches for the combination of chest-X-ray image datasets, Ecg datasets and balanced text datasets proposing (multi-prediction-mode). The other challenge due to different environmental and other additional high level risk factors of the disease in Ethiopia such as HD occurred due to diabetes and some other features which are not included in this research work can have additional contribution in the prediction of this disease using AI approaches. The other issue behind the reduction in performance of models having negative contribution might be the low in quality of the local dataset and the lower in performance of the materials, tools and deployment environment might have its own effect, such as low performance personal computer and attributes which are not relevant for HD prediction. In future works researchers can add a better contribution by fixing the issue behind the quality of local datasets, differing features that must be included to predict HD disease by adding the performance of the deployment tools and environment. As a researcher we are recommending other researchers can also study the severity level prediction of HD using DL and AI, through prediction of different target classes such as not-diseased, mild, moderate, severe and other classes. Finally, even if we have achieved a better performing model for this proposed work, it might be proved a better performance can be achieved using other techniques or additional local and balanced datasets.

REFERENCES

[1] Yash Jayesh Chauhan, "Cardiovascular Disease Prediction using Classification Algorithms of Machine Learning," *Int. J. Sci. Res.*, vol. Volume 9, no. Isuue 5 May, pp. 194–200, 2020, doi: 10.21275/SR20501193934.

[2] K. Vembandasamy, R. Sasipriya, and E. Deepa, "Heart Diseases Detection Using Naive Bayes Algorithm," *Int. J. Innov. Sci. Eng. Technol.*, vol. 2, no. 9, pp. 441–444, 2015.

[3] M. N. Uddin and R. K. Halder, "An ensemble method based multilayer dynamic system to predict cardiovascular disease using machine learning approach," *Informatics Med. Unlocked*, vol. 24, p. 100584, 2021, doi: 10.1016/j.imu.2021.100584.

[4] S. Bashir, U. Qamar, and M. Y. Javed, "An Ensemble based Decision Support Framework for Intelligent Heart Disease Diagnosis," 2014.

[5] D. A. Angaw, R. Ali, A. Tadele, and S. Shumet, "The prevalence of cardiovascular disease in Ethiopia : a systematic review and meta - analysis of institutional and community - based studies," pp. 1–9, 2021.

[6] D. Yadeta, W. Walelgne, J. M. Fourie, W. Scholtz, O. Scarlatescu, and G. Nel, "Cardiovascular Topics Ethiopia Country Report PASCAR and WHF Cardiovascular Diseases Scorecard project," vol. 32, no. 1, pp. 37–46, 2021, doi: 10.5830/CVJA-2021-001.

[7] D. Zhang *et al.*, "Heart Disease Prediction Based on the Embedded Feature Selection Method and Deep Neural Network," vol. 2021, no. Ml, 2021, doi: https://doi.org/10.1155/2021/6260022 Research.

[8] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics Med. Unlocked*, vol. 16, no. July, p. 100203, 2019, doi: 10.1016/j.imu.2019.100203.

[9] S. F. Waris and S. Koteeswaran, "Heart disease early prediction using a novel machine learning method called improved K-means neighbor classifier in python," *Mater. Today Proc.*, no. xxxx, pp. 1–7, 2021, doi: 10.1016/j.matpr.2021.01.570.

[10] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, “A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms,” *Hindawi*, vol. Volume 201, p. 21, 2018, doi: https://doi.org/10.1155/2018/3860146 Research.

[11] V. Ravikumar and M. Bhavani, “EFFECTIVE HEART DISEASE PREDICTION USING HYBRID MACHINE LEARNING,” *JES J. Eng. scince*, vol. 12, no. 12, pp. 273–285, 2021.

[12] S. and A. C. M. Guido, *Introduction to Machine Learning with Python*. 2016.

[13] W. Richert and L. P. Coelho, *Building Machine Learning Systems with Python*. 2013.

[14] H. E. Taye and A. Science, “MACHINE LEARNING BASED CHRONIC KIDNEY DISEASE PREDICTION MODEL Adama , Ethiopia,” 2021.

[15] S. D. D. Kriti Gandhi, Mansi Mittal, Neha Gupta, “Disease Prediction using Machine Learning,” *Int. J. Res. Appl. Sci. Eng. Technol. ijraset Cite*, vol. 8, no. June 2020, VI, pp. 1–12, 2022, doi: http:/doi.org/10.22214/ijraset.2020.6077.

[16] Dibaba Adeba Debal, “Chronic Kidney Disease Prediction Using Machine Learning Techniques, Adama , Ethiopia,” 2021.

[17] S. N. Kassaye, K. Kakeba, B. G. Ieee-member, and A. Dessie, “Rheumatic Heart Disease Detection Using Machine Learning Techniques,” *2021*, pp. 1–20.

[18] I. D. Mienye, Y. Sun, and Z. Wang, “An improved ensemble learning approach for the prediction of heart disease risk,” *Informatics Med. Unlocked*, vol. 20, p. 100402, 2020, doi: 10.1016/j.imu.2020.100402.

[19] S. Mohan, C. Thirumalai, and G. Srivastava, “Effective Heart Disease Prediction using Hybrid Machine Learning Techniques,” *IEEE Access*, vol. PP, p. 1, 2019, doi: 10.1109/ACCESS.2019.2923707.

[20] G. Ramesh, K. Madhavi, P. D. Kumar, J. Somasekar, and J. Tan, “Improving the accuracy of heart attack risk prediction based on information gain feature selection technique,” *www.elsevier.com/locate/matpr Improv.*, no. xxxx, 2020, doi:

10.1016/j.matpr.2020.12.079.

[21] M. S. Amin, Y. K. Chiam, and K. D. Varathan, “Identification of significant features and data mining techniques in predicting heart disease,” *Telemat. Informatics*, 2018, doi: 10.1016/j.tele.2018.11.007.

[22] H. Takci, “Improvement of heart attack prediction by the feature selection methods,” *Turkish J. Electr. Eng. Comput. Sci. · January 2018*, no. January 2018, 2020, doi: 10.3906/elk-1611-235.

[23] D. Shah, S. Patel, and S. Kumar, “Heart Disease Prediction using Machine Learning Techniques,” *Springer Nature, Comput. Sci. 1345*, p. 6, 2020, doi: https://doi.org/10.1007/s42979-020-00365-y.

[24] N. Bora, “Using Machine Learning to Predict Heart Disease,” *Calif. STATE Univ. SAN MARCOS Proj.*, 2021.

[25] D. Shah, S. Patel, and S. Kumar, “Heart Disease Prediction using Machine Learning Techniques,” 2020, doi: https://doi.org/10.1007/s42979-020-00365-y.

[26] M. S. Amin, Y. K. Chiam, and K. D. Varathan, “Identification of significant features and data mining techniques in predicting heart disease,” *Telemat. Informatics*, vol. 36, pp. 82–93, 2019, doi: 10.1016/j.tele.2018.11.007.

[27] M. N. Uddin and R. K. Halder, “An Ensemble Method Based Multilayer Dynamic System to Predict Cardiovascular Disease Using Machine Learning Approach,” *Informatics Med. Unlocked*, p. 100584, 2021, doi: 10.1016/j.imu.2021.100584.

[28] N. Baeradeh, M. G. Johari, L. Moftakhar, and R. Rezaeianzadeh, “The prevalence and predictors of cardiovascular diseases in Kherameh cohort study : a population - based study on 10 , 663 people in southern Iran,” pp. 1–12, 2022.

[29] M. Brodmann *et al.*, “Global Burden of Cardiovascular Diseases,” vol. 76, no. 25, 2020, doi: 10.1016/j.jacc.2020.11.010.

[30] M. Kasprzyk, B. Wudarczyk, R. Czyz, L. Szarpak, and B. Jankowska-polanska, “Ischemic heart disease – definition , epidemiology , pathogenesis , risk factors and

treatment," vol. 2020, no. 6, pp. 358–360, 2020, doi: 10.25121/PNM.2018.31.6.358.

[31] N. H. Lung, "Risk Factors for Coronary Heart Disease," pp. 1–8, 2011, doi: 10.1161/CIR.0b013e3182009701.

[32] F. O. R. C. Diseases, "KENYA NATIONAL GUIDELINES FOR CARDIOVASCULAR DISEASES MANAGEMENT, DIVISION OF NON-COMMUNICABLE DISEASES MINISTRY OF HEALTH," 2020.

[33] J. W. von Goethe, "Types of cardiovascular," p. 1774, 2002.

[34] A. A. Puthenpurakal, "Stroke 1: definition, burden, risk factors and diagnosis," *Nurs. Pract. Rev.*, vol. 113, no. 11, pp. 44–48, 2017.

[35] J. M. Katzenellenbogen, A. P. Ralph, R. Wyber, and J. R. Carapetis, "Rheumatic heart disease : infectious disease origin , chronic care approach," *Katzenellenbogen al. BMC Heal. Serv. Res. 17793*, no. 2017, pp. 1–16, 2018, doi: 10.1186/s12913-017-2747-5.

[36] R. M. H. D. N. P. Cpnp-ac, "Pediatric Patients With Congenital Heart Disease," *TJNP J. Nurse Pract.*, vol. 15, no. 1, pp. 118–124, 2019, doi: 10.1016/j.nurpra.2018.10.017.

[37] A. Workina, A. Habtamu, T. Diribsa, and F. Abebe, "Knowledge of modifiable cardiovascular diseases risk factors and its primary prevention practices among diabetic patients at Jimma University Medical Centre : A cross- sectional study," *PLOS Glob. PUBLIC Heal.*, pp. 1–11, 2022, doi: 10.1371/journal.pgph.0000575.

[38] G. Uchida *et al.*, "Risk factors of Heart Disease," *J. Japanese Soc. Gastroenterol.*, vol. 114, no. 10, pp. 1819–1828, 2017, doi: 10.29309/tpmj/2009.16.04.2730.

[39] M. Mohammadnezhad, T. Mangum, W. May, J. J. Lucas, and S. Ailson, "Common Modifiable and Non-Modifiable Risk Factors of Cardiovascular Disease (CVD) among Pacific Countries," pp. 153–170, 2016, doi: 10.4236/wjcs.2016.611022.

[40] E. by J. sen Series, Intechopen, *Machine Learning Algorithms, Models and Applictions*, vol. 7. 2021.

[41] I. H. Sarker, "Machine Learning : Algorithms , Real - World Applications and Research Directions," *SN Comput. Sci. 2160*, 2021, doi: https://doi.org/10.1007/s42979-021-

00592-x.

[42] I. H. Sarker, “Machine Learning : Algorithms , Real - World Applications and Research Directions,” *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[43] B. Mahesh, “Machine Learning Algorithms - A Review,” *Int. J. Sci. Res.*, no. October 2020, 2020, doi: 10.21275/ART20203995.

[44] J. M. Chatterjee, “Machine Learning with Health Care Perspective: Machine Learning and Healthcare,” *Springer*, vol. 13, no. January 2020, pp. 1–11, 2021, doi: 10.1007/978-3-030-40850-3.

[45] N. Gupta and S. Dhall, “Disease Prediction using Machine Learning,” no. June 2020, 2022.

[46] V. Ravikumar and M. Bhavani, “EFFECTIVE HEART DISEASE PREDICTION USING,” *JES J. Eng. scince*, vol. 12, no. 12, pp. 273–285, 2021.

[47] U. F. Njoku and A. Abelló, “Impact of filter feature selection on classification : an empirical study,” vol. 3130, 2022.

[48] Y. B. Wah, Ibrahim, Nurain, and H. A. Hamid, “Feature selection methods : Case of filter and wrapper approaches for maximising classification accuracy SCIENCE & TECHNOLOGY Feature Selection Methods : Case of Filter and Wrapper Approaches for Maximising Classification Accuracy,” no. January, 2018.

[49] M. R. Alnowami, F. A. Abolaban, and E. Taha, “A wrapper-based feature selection approach to investigate potential biomarkers for early detection of breast cancer,” *J. Radiat. Res. Appl. Sci.*, vol. 15, no. 1, pp. 104–110, 2022, doi: 10.1016/j.jrras.2022.01.003.

[50] V. Verma, “A comprehensive guide to Feature Selection using Wrapper methods in Python,” 2020.

[51] C. Kang, “Preoperative prediction of complicated appendicitis using machine learning method,” *Res. Sq.*, pp. 1–20, 2020, doi: https://doi.org/10.21203/rs.3.rs-27341/v1

License:

[52] D. M. Belete and D. H. Manjaiah, “Wrapper Based Feature Selection Techniques On EDHS-HIV / AIDS Dataset,” *Eur. J. Mol. Clin. Med.*, vol. 07, no. 08, 2020, 2020.

[53] R. Panthong and A. Srivihok, “Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm,” *Procedia - Procedia Comput. Sci.*, vol. 72, pp. 162–169, 2015, doi: 10.1016/j.procs.2015.12.117.

[54] A. Garg, B. Sharma, and R. Khan, “Heart disease prediction using machine learning techniques,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012046.

[55] M. N. Uddin and R. K. Halder, “Jo ur na l P re r f,” *Informatics Med. Unlocked*, p. 100584, 2021, doi: 10.1016/j.imu.2021.100584.

[56] A. Suresh, “58 What is a confusion matrix.” p. 20, 2020.

[57] H. V. Garlapati, “Model Evaluation techniques.” 2021.

APPENDIXES

Appendix A: Datasets Feature Descriptions

- **Age**: - is the age of the patient in years (Numeric)
- **Sex: -** is the gender of the patient [1: Male and 0: Female] (Nominal).
- **Chest pain type**: - is the types of the chest pain experienced by the patient experienced as [0: typical angina, 1: atypical angina, 2: non-angina pain, 3: asymptomatic] (Nominal).
- **trestbps**:-is the level of blood pressure at resting mode in mm/hg (Numerical).
- **Cholesterol (chol)**:- is the serum cholesterol in mg/dl (Numeric).
- **Fasting blood sugar (fbs):-** is the blood sugar level on fasting > 120 mg/dl represents as [1: True, 0: False] (Nominal).
- **Resting Ecg:** - is the results of electrocardiogram results while at rest represented as [0: Normal, 1: Abnormal in ST-T wave, 2: left ventricular hypertrophy] (Nominal).
- **Max Heart Rate (Thalach):-** is the maximum heart rate achieved (Numeric).
- **Exang**: - is angina induced by exercise representing [0: No and 1: Yes](Nominal).
- **Old-peak**: - is ST depression induced by exercise relative to rest (Numeric).
- **Slope**: - slope of the peak exercise ST segment representing [0: normal, 1: upsloping, 2: flat, and 3: down sloping] (Nominal).
- **Ca**: - is the number of major vessels colored by fluoroscopy having values from 0 to 3 (Nominal).
- **Thal**: - shows the defects type representing [0: normal, 1: fixed, 2: reversible, 4: non-reversible] (Nominal).
- **Target**:- the class of the target variable to be predicted as 1: patient and 0: Normal

Appendix B: Selected Features using CFS and SFFS

Table 5.3.1.1: Selected features using CFS for the Heart disease datasets

Public HD Dataset	Local HD Dataset	Combined HD Dataset
Age, sex, cp, trestbps restecg, chol, thalach, exang, oldpeak, slope, ca, thal	Age, cp, trestbps, fbs, chol, thalach, exang, oldpeak, slope, ca, thal	Age, sex, cp, trestbps, restecg, chol, thalach, exang, oldpeak, slope, ca, fbs

Table 5.3.1.2. Selected features using SFFS for Public HD dataset

Classifier	Using	Selected Features Using SFFS for Public HD Datasets after resampling
RF	10-FCV	Sex, cp, trestbps, chol, fbs, restecg, thalach, exang, pldpeak, slope, ca
	80/20 %	Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, pldpeak, slope, thal
KNN	10-FCV	Sex, cp, oldpeak, slope, ca, thal, restecg
	80/20 %	Cp, oldpeak, slope, ca, thal, resecg
SVC	10-FCV	Sex, cp, chol, thalach, exang, oldpeak, slope, ca, thal
	80/20 %	Age, sex, cp, trestbps, fbs, thalach, exang, oldpeak,slope,ca,thal
GB	10-FCV	Age, sex, cp, trestbps, chol, fbs, restecg, oldpeak, ca, thal
	80/20 %	Age, sex, cp, trestbps, chol, fbs, thalach, exang, oldpeak, slope, ca, thal
VC	10-FCV	Age, sex, cp, trestbps, chol, restecg, thalach, oldpeak, ca, thal
	80/20 %	Age, sex, cp, trestbps, chol, restecg, thalach, oldpeak, slope, ca

Table 5.3.1.3. Selected features using SFFS for Local HD dataset

Classifier	Using	Selected Features Using SFFS for Local HD Datasets after resampling
RF	10-FCV	Age, sex, cp, chol, exang, trestbps, thalach, slope, thal, oldpeak, fbs, restecg, ca
	80/20 %	Trestbps, thalach, oldpeak
KNN	10-FCV	Cp, fbs, exang, oldpeak, slope
	80/20 %	Cp, fbs, exang, oldpeak,thal
SVC	10-FCV	Sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal

	80/20 %	Sex, cp, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal
GB	10-FCV	Age, cp, chol, fbs, resecg, thalach, exang, oldpeak, ca, thal
	80/20 %	Age, sex, trestbps, chol, cp, thalach, exang, oldpeak, fbs, thal, restecg
VC	10-FCV	Age, cp, fbs, thalach, exang, oldpeak, ca, thal
	80/20 %	Age, Sex, cp, thalach, exang, oldpeak chol, fbs, restecg, thal

Table 5.3.1.4 Selected features using SFFS for combined HD dataset

Classifier	**Using**	**Selected Features Using SFFS for combined HD Datasets after resampling**
RF	10-FCV	Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal
	80/20 %	Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal
KNN	10-FCV	Cp, fbs, exang, oldpeak, slope, ca, thal
	80/20 %	Sex, cp, fbs, exang, oldpeak, slope, ca, thal
SVC	10-FCV	Age, sex, cp, trestbps, exang, oldpeak, slope, ca
	80/20 %	Sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca
GB	10-FCV	Age, sex, cp, chol, fbs, thalach, exang, oldpeak, slope, ca, thal
	80/20 %	Age, sex, cp, trestbps, chol, fbs, restecg, exang, oldpeak, slope, ca, thal
VC	10-FCV	Sex, cp, tresbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal
	80/20 %	Age, sex, cp, trestbps, chol, fbs, restecg, exang, oldpeak, slope, ca, thal

Appendix C: Sample Code

Appendix C1: Sample code for importing different libraries

```
#Importing all Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import pickle
from sklearn.model_selection import train_test_split
from sklearn.model_selection import
cross_val_score,cross_val_predict
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import
RandomForestClassifier,RandomForestRegressor
from sklearn.ensemble import AdaBoostClassifier,
GradientBoostingClassifier, VotingClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import LinearSVC,SVC
from mlxtend.feature_selection import
SequentialFeatureSelector
from sklearn.feature_selection import VarianceThreshold
from mlxtend.plotting import plot_sequential_feature_selection
as plot_sfs
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.metrics import precision_score, recall_score,
plot_roc_curve, roc_auc_score
#change the direction of the file location
import os
os.chdir("C:\\Users\\SIRILA\\Desktop\\pgclass\\New folder")
import warnings
warnings.filterwarnings('ignore')
['__output__.json', '__notebook__.ipynb']
```

Appendix C2: Sample Code for loading Different datasets

```
#Loading the International heart Disease Data sets
data =
pd.read_csv("1_Cleveland_Hungary_Switzerland_and_Long_Beach.cs
v")
# Loding the Local Hospital WKURTH Heart Data sets
data = pd.read_csv("2_Wku_HD_Last.csv")
# the Combined WKU and International HD data sets
data =
pd.read_csv("3_Combined_WKU_and_International_HD_data.csv")
```

Appendix C3: Sample Code for Data Preprocessing

Appendix C3.1: Handling Missing Value Implementation Sample Code

```
# Checking if there is missing values and sum up these total
missing values
isnull_number=[]
for i in data.columns:
    x=data[i].isnull().sum()
    isnull_number.append(x)
pd.DataFrame(isnull_number,index = data.columns, columns =
["Total Missing Values"])
data.isnull().sum()
```

Appendix C3.2: Categorical Data transformation Implementation Sample Code

```
# Categorical data labeling using one hot encoding to encode
dummy variables
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
label_encoder_x= LabelEncoder()
x[:, 0]= label_encoder_x.fit_transform(x[:, 0])
#Encoding for dummy variables
onehot_encoder= OneHotEncoder()
x= onehot_encoder.fit_transform(x).toarray()
data_copy = pd.get_dummies(data_copy,columns =
catagoric_var[:-1],drop_first =True)
```

Appendix C3.3: SMOTE Data class Balancing technique applied

```
from imblearn import under_sampling, over_sampling
from imblearn.over_sampling import SMOTE
from collections import Counter
smote=SMOTE()
X_resampled, Y_resampled=smote.fit_resample(x,y)
print(sorted(Counter(Y_resampled).items()),Y_resampled.shape,X
_resampled.shape)
```

Appendix C3.4: Feature Scaling for numerical and continuous feature values

```
from sklearn.preprocessing import StandardScaler
StandardScaler = StandardScaler()
columns_to_scale=['age','resting_blood_pressure','cholesterol'
,'max_heart_rate_achieved','st_depression']
data_copy[columns_to_scale]=
StandardScaler.fit_transform(data_copy[columns_to_scale ])
```

Appendix C4: Sample Code for Feature selection implementation

Appendix C4.1: chi² statistical test to select best of features from HD datasets

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
x= data.iloc[:,0:13] # independent columns
y= data.iloc[:,-1]# target column
#Aplly SelectKBest class to Select best top best feature and
order them accordingly
# select 13 feature for the international heart disease
datasets
ordered_rank_features = SelectKBest(score_func=chi2, k=13)
ordered_feature =ordered_rank_features.fit(x,y)
fit=ordered_rank_features.fit(x,y)
```

Appendix C4.2: A Sample Code for Applying SFFS method for HD datasets

```
from mlxtend.feature_selection import
SequentialFeatureSelector
from sklearn.ensemble import RandomForestRegressor,
RandomForestClassifier
from sklearn.metrics import roc_auc_score
feature_selector1 =
SequentialFeatureSelector(RandomForestClassifier(n_estimators=
100, random_state=50, n_jobs=-1),
           k_features= (1,13),  forward=True,
           floating=False,  verbose=2,
           scoring='accuracy', cv=10)
# before applying resampling technique and with cv=10
features =
feature_selector1.fit(np.array(train_features.fillna(0)),
train_labels)
```

Appendix C5: Sample Code for Building Models Using Percentage Splitting

```
# Random Forest Classifier Instantiating for balancing
datasets
From sklearn.ensemple import RandomForestClassifier
randfor_resampled=RandomForestClassifier(n_estimators=100,max_
features='auto', class_weight='balanced', random_state=0)
randfor_resampled.fit(X_train_resampled, Y_train_resampled)
Y_pred_rf_resampled =
randfor_resampled.predict(X_test_resampled)
print(Y_pred_rf_resampled)
```

Gradient Boosting model

```
from sklearn.ensemble import AdaBoostClassifier,
GradientBoostingClassifier
gb_resampled=GradientBoostingClassifier()
gb_resampled=train_model(X_train_resampled,Y_train_resampled,X
_test_resampled,Y_tes_resampled,GradientBoostingClassifier)
gb_resampled.fit(X_train_resampled,Y_train_resampled)
y_pred_gb_resampled = gb_resampled.predict(X_test_resampled)
print(y_pred_gb_resampled)
Y_pred_rf_resampled =
randfor_resampled.predict(X_test_resampled)
print(Y_pred_rf_resampled)
```

Appendix C6: Sample code for building models using 10-F-CV

 # Random Forest model

```
randfor_resampled= RandomForestClassifier(n_estimators=100)
accuracy_score_randfor_resampled=cross_val_score(randfor_resam
pled,x1,y1,cv=10)
accuracy_score_randfor_resampled
accuracy_score_randfor_resampled.mean()
# Gradient Boosting model
gb_resampled= GradientBoostingClassifier(random_state = 1)
accuracy_score_gb_resampled=cross_val_score(gb_resampled,x1,y1
,cv=10)
accuracy_score_gb_resampled
accuracy_score_gb_resampled.mean()
y_pred_gb_resampled = gb_resampled.predict(X_test_resampled)
print(y_pred_gb_resampled)
Y_pred_rf_resampled =
randfor_resampled.predict(X_test_resampled)
print(Y_pred_rf_resampled)
```

Appendix C7: Sample Code for Model Saving

```
# Sample Code for Saving the model RF with CFS
# Saving the model as serialized object pickle
import pickle
with
open('1_Cleveland_Hungary_Switzerland_and_Long_Beach.pkl',
'wb') as file:
pickle.dump(randfor_resampled, file)
```

Appendix C8: Sample Code for Integrating ML Model with the Flask Server

```
# Detection of Heart Disease and enhancing prediction through
ML techniques web app python sample code
```

```
Import  numpy as np
Import pickle
From flask import Flask, request, render_template
# 1. Load ML model
model =
pickle.load(open('3_Combined_WKU_and_International_HD_data.pkl
', 'rb'))
# 2. Create application
app = Flask(__name__)
# 3. get user input from html/ Bind home function to URL
@app.route('/')
def home():
    return render_template('Heart Disease Classifier.html')
# 4. Bind predict function to URL/ make prediction
@app.route('/predict', methods =['POST'])
def predict():
    # Put all form entries values in a list
    features = [float(i) for i in request.form.values()]
    # Convert features to array
    array_features = [np.array(features)]
# 5.getting the prediction result/ Predict features
    prediction = model.predict(array_features)
    output = prediction
    # Check the output values and retrive the result with html
tag based on the value
    if output == 1:
        return render_template('Heart Disease
Classifier.html',
                               result = 'Hi User ,The Patient
is likely to have Heart Disease!')
    else:
        return render_template('Heart Disease
Classifier.html',
                               result = 'Hi User, The Patient
is not likely to have Heart Disease!')
if __name__ == '__main__':
    # To on debug mode and Run the Application
    app.debug = True
    app.run()
```