6-2022

ROBUST DETECTION OF CORONARY HEART DISEASE USING MACHINE LEARNING ALGORITHMS

Anas Haitham Domyati

جامعة الإمارات العربية المتحدة
United Arab Emirates University

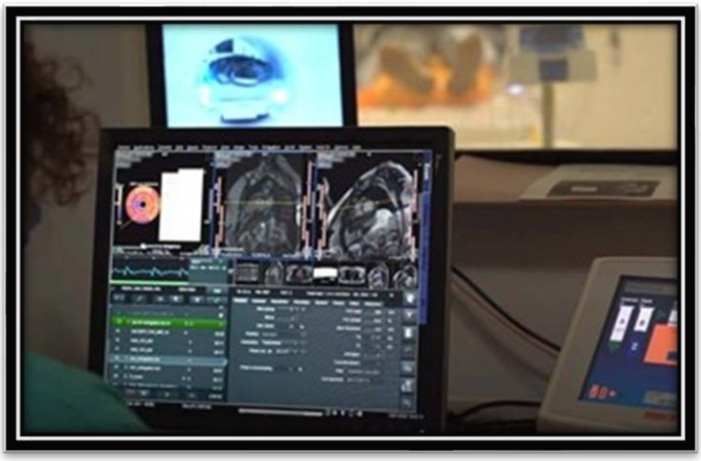
MASTER THESIS NO. 2022:33

College of Engineering

Department of Electrical and Communication Engineering

ROBUST DETECTION OF CORONARY HEART DISEASE USING MACHINE LEARNING ALGORITHMS

Anas Haitham Domyati



June 2022

United Arab Emirates University

College of Engineering

Department of Electrical and Communication Engineering

ROBUST DETECTION OF CORONARY HEART DISEASE USING MACHINE LEARNING ALGORITHMS

Anas Haitham Domyati

This thesis is submitted in partial fulfilment of the requirements for the degree of Master of Science in Electrical Engineering

June 2022

United Arab Emirates University Master Thesis
2022: 33

Cover: Heart Disease detection using machine learning algorithms

(Photo: By Mark Greaves)

Print: University Print Service, UAEU 2022

Declaration of Original Work

I, Anas Domyati, the undersigned, a graduate student at the United Arab Emirates University (UAEU), and the author of this thesis entitled "*Robust Detection of Coronary Heart Disease Using Machine Learning Algorithms*", hereby, solemnly declare that this thesis is my own original research work that has been done and prepared by me under the supervision of Dr. Qurban Memon, in the College of Engineering at UAEU. This work has not previously formed the basis for the award of any academic degree, diploma or a similar title at this or any other university. Any materials borrowed from other sources (whether published or unpublished) and relied upon or included in my thesis have been properly cited and acknowledged in accordance with appropriate academic conventions. I further declare that there is no potential conflict of interest with respect to the research, data collection, authorship, presentation and/or publication of this thesis.

Student's Signature:

Date: 20-5-2022

Approval of the Master Thesis

This Master Thesis is approved by the following Examining Committee Members:

1) Advisor (Committee Chair): Qurban A. Memon
   Title: Associate Professor
   Department of Electrical Engineering
   College of Engineering

   Signature _____ Date 25-06-2022

2) Member: Atef A. Abdrabou
   Title: Associate Professor
   Department of Electrical Engineering
   College of Engineering

   Signature _____ Date 25-06-2022

3) Member (External Examiner): Mahmud Iwan Solihin
   Title: Associate Professor
   Department of Electrical and Computer Engineering
   College of Engineering, UCSI University, Malaysia

   Signature _____ Date 25.06.2022

This Master Thesis is accepted by:

Acting Dean of the College of Engineering: Professor Mohamed H. Al-Marzouqi

Signature Mohamed AlMarzouqi Date July 04, 2022

Dean of the College of Graduate Studies: Professor Ali Al-Marzouqi

Signature Ali Hassan Date July 04, 2022

Abstract

Predicting whether or not someone will get heart or cardiac disease is now one of the most difficult jobs in the area of medicine. Heart disease is responsible for the deaths of about one person per minute in the contemporary age. Processing the vast amounts of data that are generated in the field of healthcare is an important application for data science. Because predicting cardiac disease is a difficult undertaking, there is a pressing need to automate the prediction process to minimize the dangers that are connected with it and provide the patient with timely warning. The chapter one in this thesis report highlights the importance of this problem and identifies the need to augment the current technological efforts to produce relatively more accurate system in facilitating the timely decision about the problem. The chapter one also presents the current literature about the theories and systems developed and assessed in this direction.

This thesis work makes use of the dataset on cardiac illness that can be found in the machine learning repository at UCI. Using a variety of data mining strategies, such as Naive Bayes, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), and Random Forest, the work that has been reported in this thesis estimates the likelihood that a patient would develop heart disease and can categorize the patient's degree of risk. The performance of chosen classifiers is tested on chosen feature space with help of feature selection algorithm. On Cleveland heart datasets of heart disease, the models were placed for training and testing. To assess the usefulness and strength of each model, several performance metrics are utilized, including sensitivity, accuracy, AUC, specificity, ROC curve and F1-score.

The effort behind this research leads to conduct a comparative analysis by computing the performance of several machine learning algorithms. The results of the experiment demonstrate that the Random Forest and Support

Vector machine algorithms achieved the best level of accuracy (94.50% and 91.73% respectively) on selected feature space when compared to the other machine learning methods that were employed. Thus, these two classifiers turned out to be promising classifiers for heart disease prediction. The computational complexity of each classifier was also investigated. Based on the computational complexity and comparative experimental results, a robust heart disease prediction is proposed for an embedded platform, where benefits of multiple classifiers are accumulated. The system proposes that heart disease detection is possible with higher confidence if and only if many of these classifiers detect it. In the end, results of experimental work are concluded and possible future strategies in enhancing this effort are discussed.

Keywords: Heart disease detection, Robust detection, Machine learning algorithm.

Title and Abstract (in Arabic)

الاكتشاف المتين لأمراض القلب التاجية باستخدام خوارزميات التعلم الآلي

الملخص

يعد التنبؤ بما إذا كان شخص ما سيصاب بأمراض القلب أو القلب الآن من أصعب الوظائف في مجال الطب. أمراض القلب مسؤولة عن وفاة حوالي شخص واحد في الدقيقة في العصر المعاصر. تعد معالجة الكميات الهائلة من البيانات التي يتم إنشاؤها في مجال الرعاية الصحية تطبيقًا مهمًا لعلوم البيانات. نظرًا لأن التنبؤ بأمراض القلب هو مهمة صعبة ، فهناك حاجة ملحة لأتمتة عملية التنبؤ لتقليل المخاطر المرتبطة بها وتزويد المريض بالتحذير في الوقت المناسب. يسلط الفصل الأول في تقرير الأطروحة هذا الضوء على أهمية هذه المشكلة ويحدد الحاجة إلى زيادة الجهود التكنولوجية الحالية لإنتاج نظام أكثر دقة نسبيًا في تسهيل اتخاذ القرار في الوقت المناسب بشأن المشكلة. من ناحية أخرى ، يقدم الفصل الثاني الأدبيات الحالية حول النظريات والأنظمة التي تم تطويرها وتقييمها في هذا الاتجاه.

تستخدم هذه الأطروحة مجموعة البيانات الخاصة بأمراض القلب التي يمكن العثور عليها في مستودع التعلم الآلي في .UCI باستخدام مجموعة متنوعة من استراتيجيات استخراج البيانات ، مثل Naive Bayesو Decision Tree و Support Vector Machine (SVM) و K-Nearest Neighbor (K-NN)و Random Forest ، فإن العمل الذي تم الإبلاغ عنه في هذه الأطروحة يقدر احتمالية ذلك قد يصاب المريض بأمراض القلب ويمكنه تصنيف درجة مخاطر المريض. يتم اختبار أداء المصنفات على مساحة ميزة محددة من خلال خوارزميات اختيار الميزة المختلفة. تم تدريب النماذج واختبارها على كليفلاند. يتم استخدام العديد من مقاييس الأداء وهي: الدقة ، والحساسية ، والنوعية ، ومنحنى AUC ، و F1-Score ، و MCC ، و ROC لمراقبة فعالية وقوة كل نموذج. الغرض من الجهد وراء هذا البحث هو تقديم تحليل مقارن عن طريق حساب أداء العديد من خوارزميات التعلم الآلي. أظهرت نتائج التجربة أن خوارزميات Random Forest و Support Vectorقد حققت أفضل مستوى من الدقة (94.50٪ و 91.73٪) على مساحة الميزة المختارة. عند مقارنتها بأساليب التعلم الآلي الأخرى المستخدمة. وهكذا ، تبين أن هذين المصنفين من المصنفات الواعدة للتنبؤ بأمراض القلب. كما تم التحقيق في التعقيد الحسابي لكل مصنف. استنادًا إلى التعقيد الحسابي والنتائج التجريبية المقارنة ، يُقترح تنبؤ قوي بأمراض القلب لمنصة مضمنة ، حيث يتم تجميع فوائد كلا المصنفين. يقترح النظام أن أمراض القلب موجودة إذا وفقط إذا اكتشفها

كلا المصنفين. في النهاية ، يتم الانتهاء من نتائج العمل التجريبي ومناقشة الاستراتيجيات المستقبلية المحتملة لتعزيز هذا الجهد

مفاهيم البحث الرئيسية: اكتشاف الكائن , التصنيف , تعلم الآلة , التنبؤ بأمراض القلب.

Acknowledgements

First and foremost, I would like to praise Allah the Almighty, the Most Gracious, and the Most Merciful for His blessing given to me during my study and in completing this thesis. May Allah's blessing goes to His final Prophet Muhammad (peace be upon him), his family, and his companions.

I wish to express my sincere gratitude to Dr. Qurban Memon, thesis supervisor for providing me with all the necessary facilities and valuable guidance and encouragement. that was required for this research.

Finally, I thank my loving wife, son, and parents for being my support system. They were always helping me directly or indirectly in making this project a success.

Dedication

To my beloved parents, wife, son, and family

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CVD	Cardiovascular disease
DT	Decision tree
FN	False Negative
FP	False Positive
KNN	K-nearest neighbors
MLP	Multi-layer perceptron
ML	machine learning
NB	Naive Basis
RF	Random forest
SVM	Support vector machine
TP	True Positive
TN	True Negative

Chapter 1

Chapter 1: Introduction

1.1 Overview

Heart disease is regarded as one of the world's most dangerous and life-threatening chronic diseases. In heart illness, the heart generally fails to deliver enough blood to other regions of the body to allow them to operate normally. Heart failure is caused by the narrowing and occlusion of coronary arteries. The supply of blood to the heart is controlled by coronary arteries. According to a recent poll, the United States tops the list of countries having a high number of patients with heart disease [1]. The most prevalent signs of coronary disease include weakness in the body, swollen feet, shortness of breathing and weariness, among others. Smoking, a poor diet, high blood pressure along with high cholesterol, fitness issue with a lack of exercise, and other factors can all contribute to an increased risk of heart disease. There are various types of heart disease, the most common being coronary artery disease (CAD), which can cause chest pain, stroke, and heart attack. Congestive heart failure, (birth related) congenital heart disease, and cardiovascular issues are all examples of other kinds of heart disease.

Traditional investigative procedures were initially utilized to detect cardiac illness, but they were discovered to be complex. Heart disease identification and treatment are extremely difficult to diagnose and treat in developing nations due to a lack of medical diagnostic instruments and medical experts. Heart disease is a deadly disease, which is on the rise in both developing and developed states. According to research document published by World Health Organization [2], 17.90 million people died from cardiovascular disease in 2016. This sum accounts for around a one-third of whole world deaths. According to research [3], 0.2 million Pakistanis are affected by this deadly disease each year.

According to a survey published by the European Society of Cardiology (ESC), 26.5 million persons have cardiac disease, with 3.8 million diagnosed each year. Patients with heart disease die in 50-55 percent of cases within the first three years, and heart disease treatment accounts for around 4% of the yearly healthcare expenditure [3]. Traditional invasive approaches for detecting heart disease relied on a patient's medical history, physical test results, and a doctor's investigation of pertinent symptoms [4]. Among the standard procedures, angiography is considered one of the most precise approaches for diagnosing cardiac problems. Angiography, on the other hand, has a number of drawbacks, including a hefty price tag and a wide range of adverse effects. Therefore, clear and appropriate diagnosis of cardiac disease is deemed necessary and critical to preventing further damage to the patient.

Due to risk misclassification, roughly 20% of patients with high-risk of cardiovascular disease are underestimated. Typical diagnostic approaches are based on medical specialists' personal experiences and knowledge of the condition, which increases the likelihood of mistakes, delays appropriate treatment, lengthens treatment periods, and raises expenses significantly. Therefore, a clear and appropriate diagnosis of cardiac disease is deemed necessary and critical to preventing further damage to the patient.

Looking at previous recent studies, it appears that machine learning approaches have found a way to help in this for predicting and classifying heart disease patients using a number of attributes. Individual and ensemble methods involving machine learning are reported in literature recently that explore models for early diagnosis of heart disease and the one that is more accurate. An intelligent system that builds database of such patients and can help facilitate the doctor in predicting the level of the disease has been termed need of the hospitals world-wide. In the next sub-section, a comprehensive survey of literature is presented to highlight recent interest in this direction.

1.2 Literature Review

To overcome issues in traditional invasive-based methods to detect of heart problem, the researchers have tried developing various non-invasive and smart healthcare related systems that are based on predictive capability of machine learning techniques such as Random Forest, Support Vector Machine (SVM) Nave Bayes (NB), K-Nearest Neighbor (KNN), and Decision Tree (DT), among others. As a result, the number of people dying from heart disease has dropped. In the literature, researchers have regularly used the Cleveland heart disease dataset. In this context, the authors [5] employed a logistic regression classification system to diagnose cardiac disease and achieved a 77.1 percent accuracy. Authors in [6] presented the use of neural networks to diagnose and predict blood pressure and heart disease, among many other features. The authors created an adaptive neural network that used the given disease attributes to produce an output that was carried out by an output perceptron and nearly included 120 hidden layers, which is the basic and related technique to ensure an acceptable result of heart disease if the model is used. In another work [7] employed Naive Basis (NB) and Decision Tree for heart disease diagnosis and prediction with reasonable accuracy. With NB, accuracy estimated was 82.7 percent, and with DT, it remained 80.4 percent. Using ANN, the authors [8] employed the adaptive fuzzy ensemble approach to predict cardiac disease. An ensemble model involving ANN for the prediction of heart related disease has also been proposed by authors in [9], for effective diagnosis of cardiac disease.

The authors [10] have developed a non-invasive model to predict coronary artery disease with good results in terms of various performance metrics. The approach used three classifiers for stenosis detection in three arteries. The authors in [11] study a variety of machine learning approaches, which may be used for classification of cardiac disease. In another comparative work [12], the authors develop a model based on machine

learning, which combines five (5) different methods with higher accuracy and improved performance in predicting the heart failure. Similarly, the authors [13] have introduced a system for predicting risk factors related to hear disease. It uses Nave Bayesian (NB) approaches for categorization of data and the AES Advanced Encryption Standard (AES) algorithm for transporting data safely for illness prediction. The approach illustrates multiple knowledge abstraction techniques using data mining methods adopted for heart disease prediction. The output reveals that the established diagnostic system effectively assists in predicting risk factors concerning heart diseases. The authors [14] explored Naive Bayes and support vector machine for classification on UCI dataset for predicting cardiac disease. The measures deployed for performance analysis were root mean squared error, sum of Squared Error, mean absolute error, and it was determined from this work that support vector machine outperformed Naive Bayes with respect to classification accuracy. The Multi-Layer Perceptron (MLP) Neural Network has also been investigated in [15] for predicting cardiac illness using 14 significant attributes. The sensitivity, specificity, precision, and accuracy were used on UCI dataset to evaluate the performance of the system with a high accuracy of 93.39%. The use of various machine learning algorithms have motivated many to explore its effective strengths in many applications [16-19] including healthcare.

The combination of different machine learning has been explored in [20], where the authors developed a hybrid neural network consisting of ANN and fuzzy neural network for building a heart disease classification system resulting in an accuracy of 82.4% in comparison to previous studies. On same lines, the authors [21] present a decision support system involving SVM and ANN (trained on back propagation algorithm) with three layers for diagnosis of heart disease to achieve an accuracy of 80% in heart disease diagnosis. A survey work has been conducted in [22] that looks at several classification

algorithms to predict the risk level of patients based on various parameters. The algorithms included Naive Bayes, KNN, Decision tree, and ANN, and accuracy was estimated for a variety of variables. The authors found that the accuracy of the risk increases with number of parameters. A hybrid recommender system [23] has also been proposed to detect heart disease with promising results. The classification approach employs multiple kernel learning with adaptive neuro-fuzzy inference system based deep learning with two-fold approach. Another hybrid system in the form of a framework has been proposed in [24], where authors develop a machine learning based system to detect heart disease using a heart disease dataset. The framework uses various machine learning algorithms, there feature selection methods and seven classifier evaluation metrics to distinguish between healthy person and a person with heart problems with accuracy of 86%.

Data mining methods, according to [25], might be used to diagnose cardiac disease. The authors employ a data mining tool that included a collection of machine learning algorithms, namely Naive Bayes, decision tree, and bagging. Eleven attributes are finally used for diagnosis of heart disease. With ten-fold validation, the system generated an unbiased estimate of prediction models, with an accuracy of 82.31% for Naive Bayes, 84.35% for Decision Tree, and 85.03% for bagging. A deep learning network has also been investigated by authors [26], where design, evaluation and optimization of various architectures is conducted on publicly available database for heart disease detection using k-way cross validation and Matthews correlation coefficient (MCC). The architecture is claimed for 99% accuracy and 0.98 MCC. In another work [27], the authors suggested a system using data mining techniques in conjunction with the MapReduce algorithm. This system takes a large set of instances as an input. The performance of this system is calculated by parallel implementation and accuracy achieved is 98%

suggesting that the Mapreduce technique could be used to accurately predict HD risks in the clinic.

The deep learning technique [28] is employed for predicting four cardiac abnormalities. First, the transfer learning technique is explored based on low-scale pre-trained deep neural networks. Second, a new Convolutional Neural Network (CNN) framework is developed for abnormality prediction in heart. Third, the models that are pretrained and CNN model are put in place for extraction of features to be used in K-NN, SVM, RF, NB, and DT algorithms. The experimental results suggest that 98.23% accuracy, 98.22% recall, 98.31% precision, and 98.21% F1 score. In another work, authors [29] use a two-step method for the training dataset, followed by (1) division into two groups (2) enrich one by variational autoencoders. After that, two deep networks start learning from these groups separately. The authors use dataset that is collected from a survey used by Korean National Health and Nutritional organization. Two experiments were done, where the experiments compared proposed method with NB, Random Forest, K-NN, DT, SSVM, and Adaptive Boosting, with performance accuracy of 89.2%, specificity of 84.0%, precision of 91.1%, recall of 92.0%, f-score of 91.5%, and AUC of 88.2%.

For low resource setups, a CardioXNet, approach is proposed [30] that employs light and end-to-end architecture to enable automatic detection of five (5) classes of cardiac auscultation, involving two phases of learning namely, sequence residual learning and representation learning. The results achieved show that the architecture provides performance in all the metrics vis-à-vis previous methods with accuracy of 99.60%, precision of 99.56%, recall of 99.52% and F1- score of 99.68%. The authors claim that the proposed approach is suitable for CVD screening using mobile phones. In another study [31], the authors focus on every aspect of the performance. For this purpose, an improved deep belief network is proposed, where the features

that do not contribute enough are removed, and stacked genetic algorithm is used for optimum performance evaluated by accuracy, sensitivity, specificity, precision, F1 score, and Matthew's correlation coefficient. The results show accuracy of heart disease prediction by about 94.61%.

Chapter 2

Chapter 2: Materials and Methods

2.1 Preprocessing

The set of data is incomplete if there is missing data. When no data value is saved for a variable in an observation, it is referred to as missing values or missing data in statistics. Blank/dashes are used to symbolize these missing values. The most common cause of missing values is respondents forgetting, refusing, or failing to answer particular questions. Sensor failure, data loss while transferring, Internet connection disruption, and incorrect mathematical computations, such as dividing by zero, are all possible causes. When missing values are included in a dataset, it is difficult to determine whether they will affect the results or not. Each variable in a dataset may have a modest number of missing replies, but the missing data may be large when combined. Because of the missing data, the analysis may run but the results may not be statistically significant. It is more useful to replace missing values with a user constant or the mean value for research purposes than to remove those observations from the dataset. The nominal attributes 'ca' and 'thal' in the Cleveland heart dataset had missing values, which were substituted with the user constant based on the majority mark. In 176 of the 299 observations, the attribute 'ca' contains four missing values, with the value 0 being the majority mark. Meanwhile, the attribute 'thal' has two missing values and the majority mark of 3 in 166 of the 301 observations. As a result, the missing values in 'ca' and 'thal' were replaced by the corresponding majority marks 0 and 3, respectively, to ensure that the dataset was complete.

2.2 Dataset and features

Only 14 of the 303 samples in the Cleveland heart dataset with 76 attributes are judged more valuable for research and experimental purposes. The attribute descriptions for the Cleveland heart dataset are listed in Table 1. In

Cleveland dataset's numeric characteristics, there are no missing values. Out of the total cases in the dataset, the 165 cases are with heart disease and 138 without heart disease.

Table 1: Attributes found in heart dataset of Cleveland

S. No	Attribute	Code	Classes or Description
1	Age	Age	Age in years (26 to 88)
2	Sex	Sex	Gender instance (0 = Female, 1 = Male)
3	Chest Pain Type	CPT	Chest pain type (1: typical angina, 2: atypical angina, 3: non- anginal pain, 4: asymptomatic)
4	Rest Blood Pressure	RBP	Resting blood pressure in mm Hg [94, 200]
5	Serum Cholestoral	SCH	Serum cholesterol in mg/dl[126, 564]
6	Fasting Blood Sugar	FBS	Fasting blood sugar > 120 mg/dl (0 = False, 1= True)
7	Resting electrocardiographic results	RECG	Resting ECG results (0: normal, 1: ST-T wave abnormality, 2: LV hypertrophy)
8	Thallium scan	Thal	Defect types: value 3: normal, 6: fixed defect, 7: irreversible defect
9	Exercise-induced angina	EIA	Exercise induced angina (0: No, 1: Yes)
10	Old peak	OPK	ST depression induced by exercise relative to rest [0.0, 62.0]
11	Slope	PES	Slope of the peak exercise ST segment (1: up-sloping, 2: flat, 3: down-sloping)
12	Major Vessels	VCa	Number of major vessels colored by fluoroscopy (values 0 - 3)
13	Max Heart Rate	MHR	Maximum heart rate achieved [71, 202]
14	Class/Target	Class	Diagnosis of heart disease (1: Unhealthy, 2: Healthy)

The analysis of the dataset based on these features is shown in Figure 1 (with 9 features), and Figure 2 (with 5 features). All these figures suggest that dataset is balanced.

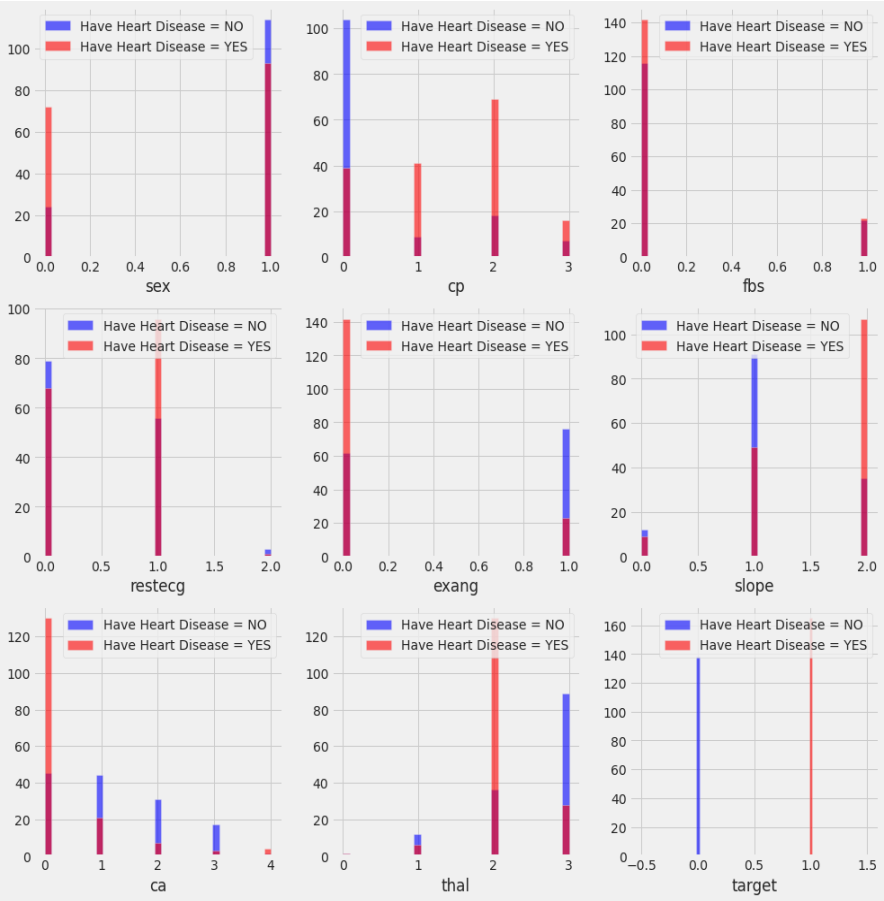


Figure 1: Summary of heart disease patients according to features-1 [32]

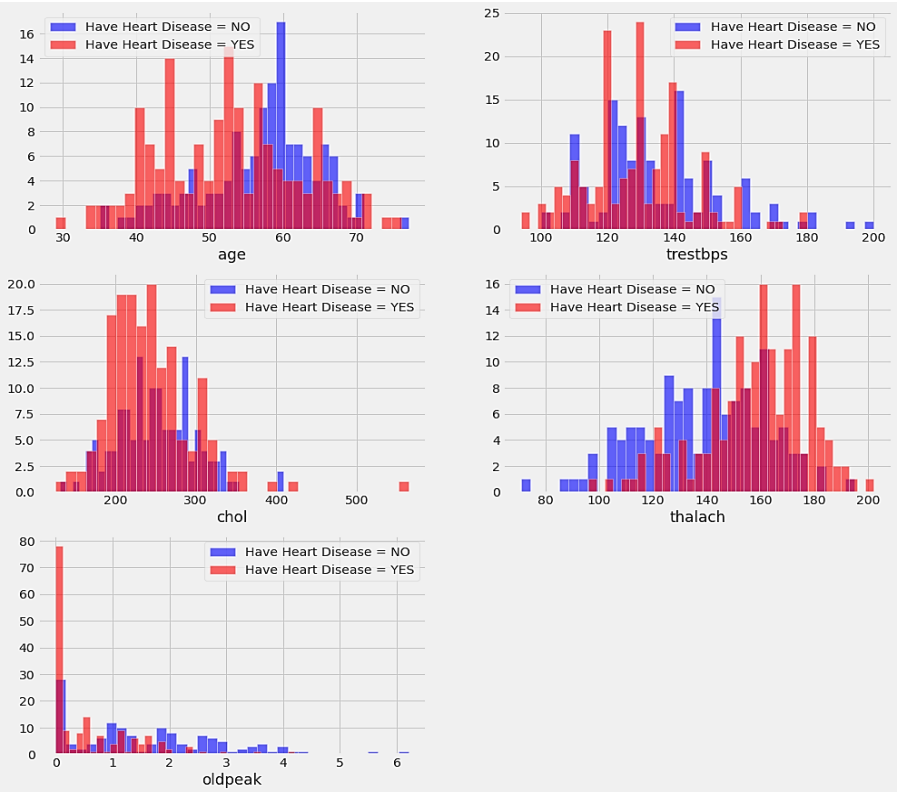


Figure 2: Summary of heart disease patients according to features-2 [32]

2.3 Feature selection

From all the attributes found in a dataset, the selection strategy for notable features selects a best sub-space of features. This is crucial because irrelevant dataset properties may cause classification performance to degrade. Using a feature selection strategy, classification algorithms become more efficient while also taking less time to execute. In the literature, four feature selection techniques have been identified. Each one is discussed in further detail below.

2.3.1 Fast correlation-based filter (FCBF)

The method for selecting features employs a search technique that works sequentially. It picks complete features first and then utilizes symmetric uncertainty to determine the features' dependence on one another and their

effect on the intended output label. Following that, it employs a backward sequential search approach to identify the most critical attributes. FCBF excels on datasets with a large number of dimensions. Symmetrical Uncertainty (SU) is a normalized information theoretic measure that calculates the relationships between features using entropy and conditional entropy values. If X is defined as a random variable with P(X) as its probability, then entropy of X is defined as:

$$H(x) = -\sum P(x_i)\log_2(P(x_i)) \qquad \text{Equation 1}$$

The conditional entropy or conditional uncertainty of a random variable X in the presence of another random variable Y, and is equal to the average of X's conditional entropies over Y:

$$H(X|Y) = -\sum P(y_j)\sum P(x_i \mid y_j)\log_2(P(x_i y_j)) \qquad \text{Equation 2}$$

$$SU(X, Y) = 2\,\frac{H(x)-H(x|y)}{H(x)+H(y)} \qquad \text{Equation 3}$$

A SU value of 1 implies that the value of another feature can be completely predicted using that feature, while a value of 0 shows that two characteristics are completely independent. Both characteristics have symmetric SU values. While nominal features are required to compute SU values, continuous features may also be employed provided their values are discretized adequately.

2.3.2 Minimal redundancy maximum relevance (mRMR)

mRMR employs a heuristic technique to identify the most critical attributes with the least amount of redundancy and the greatest amount of relevance. It chooses just those characteristics that are valuable and relevant to the goal. The technique analyses each characteristic individually and then calculates its pairwise redundancy related to other features due to its heuristic nature.

For problems with a large number of domain features, the mRMR feature selection strategy is ineffective [33].

The algorithm begins with an empty set, weighs features using mutual information, and then utilizes a forward selection approach combined with a search strategy for identifying the optimal features sub-set. It involves *k* parameter that causes its termination, when the specified feature subset has *k* features. Mutual Information [33] is a symmetrical reference to knowledge and a theoretical measure, which quantifies the information amount about one random variable that may be gleaned by viewing another random variable. The mutual information between features f_i and f_j is provided by:

$$I(f_i; f_j) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \qquad \text{Equation 4}$$

where x denotes the range of potential values for f_i and *y* denotes the range of possible values for f_j.

2.3.3 Least absolute shrinkage and selection operator (LASSO)

LASSO selects features using the absolute value of updated coefficient of features. The zero becoming values are truncated from the subset of features when the coefficient values of features are modified. LASSO outperforms using low values of feature parameter. The subset of characteristics will be chosen based on their high coefficient values, while the other features will be discarded. Certain irrelevant properties with relatively higher coefficient values can also be chosen and incorporated in feature subset [34]. The six most profound qualities that have a high connection with the objective are listed, along with their associated scores as determined by the LASSO feature selection method. Feature selection using LASSO is based on estimating the coefficients Pi ∈ R Li×D of a linear regression of position onto features according to [31], where l represents a hyperparameter that is set, and L_1-

norm term on right-hand side is a regularization parameter. Any element within *P*i with zero value shows that the corresponding features do not play a part in estimation of the position. Therefore, we mark the rows where absolute values cross a set threshold and consider the corresponding features relevant [35].

2.3.4 Relief

Relief is built on the notion of instance-based learning, which weights each feature according to its importance. The weight assigned to each property indicates the ability to distinguish across class values. Weights are assigned to qualities, and those with a weight greater than a user-specified cut-off are picked as the final subset [36] The relief feature selection method prioritizes the features that have the greatest impact on the target. The technique works by randomly picking cases from the training data. For each sampled instance, the closest instance of the same class (nearest hit) and the opposite class (nearest miss) are discovered. The value of an attribute's weight is changed based on how well its values separate the sampled instance from its closest miss and hit. It will be given a high weight if an attribute distinguishes between instances belonging to different classes and has the same value for examples belonging to the same class.

Each feature in a dataset is given a weight using this approach. Following that, the weights can be gradually increased or lowered. The objective is to make sure that critical features carry a high weight while the other features carry a lower weight. The Relief determines feature weights using techniques similar to those used in K-NN. Kira and Rendell demonstrated the algorithm of well-known feature selection approach [37]. Relief searches for two closest neighbors: one in the opposite class, designated as closest miss M, and one in the same class, designated as closest hit H. Based on R_i, M, and H values, it adjusts the W [A] consistency computation for feature A. If there is a

significant discrepancy between instances R_i and H, it is undesirable, and the value of performance (W[A]) is decreased. If the difference between M and R_i for attribute A is considerable, however, A can be utilized for differentiating across classes, increasing the W[A] weight. This procedure will be repeated m number of times, where m represents a user-defined variable. The relief feature selection algorithm may be stated as follows [38]:

Algorithm:

Input: Attribute values and the class for each training instance

Output: Predictions for attributes

Pseudocode:

a. Initialize all weights W[A] to be zero

b. For 1 to *m*, where *m* is user defined

c. Randomly choose an instance R_i

d. Find NearestHit (NH) and NearestMiss (NM)

e. for A from 1 to *a*, do the following

f. W[A] = W[A] – diff (A, R_j, NH) /m+diff(A,R*i*, NM)/m

g. Stop

The above algorithm may be stated as follows: An instance R_i at random is chosen once the weights are initialized to zero. After this, two nearest neighbors are found: one from the same class as R_i, known as nearest hit (NH), and one from a different class, known as nearest miss (NM). Next, the weights of all attributes are modified in accordance with the values of *R*i, NM, and NH. This entire procedure runs in cyclic manner based on a user-defined parameter *m*.

In order to see that this feature selection algorithm fits the requirement of this thesis work, processing time is computed to compare different algorithms for a sample of classifiers. The results are discussed in a section after selected classifiers are discussed.

2.4 Training and cross validation

The Cleveland dataset contains 304 instances, each of which has a unique set of 13 characteristics in addition to the labels and is chosen for training. The set of data is divided into categories based on the presence or absence of cardiac disease. The Hungarian dataset contains 1025 cases, 525 of which are positive, and 500 of which are negative. Both datasets have identical attribute descriptions, and both contain comparable attributes.

In K-Fold cross-validation, the entire dataset is partitioned into k pieces. The first k-1 data portions are used for training, while the rest of them are for testing. This procedure is carried out k times. For CV, a variety of k values have been used in numerous research. Because of its great reproducibility, k = 10 is used for experimental work in this case. At each iteration of ten-fold CV, 90% of the data is utilized to train the model and 10% is used to test the model. Finally, the final result is obtained by taking the mean of the findings from each phase.

2.5 Classification approaches

The classification is a kind of supervised learning approach, which is used to forecast result based on previously collected data. Below, we discuss commonly used approaches.

2.5.1 Bayes net

The Bayesian network is a prediction model that is probabilistic. Bayesian networks build from such probabilistic distributions and use probability principles to forecast and diagnose. Both discrete and continuous variables are supported by Bayesian networks. The network is shown as a variables set with relationships that are conditional and represented by acyclic directed graphs. Connections between nodes in Bayesian network represent dependent traits, while disconnected nodes reflect conditionally independent features.

Assume that X is a piece of evidence that is contingent on *n* characteristics ($X = A_1, A_2, \ldots A_n$). Assume H is a hypothesis that the evidence is of type C, P(H|X) denotes the probability of the hypothesis H given the evidence X, and P(X|H) represents the posterior probability of X given H. As indicated in equation, the posterior probability may be computed using the following equation:

$$P(H \mid X) = P(X \mid H)\, P(H)/P(X) \qquad \text{Equation 5}$$

where P(H) denotes the probability, that the hypothesis is right. The probability of the evidence is denoted by P(X). If the hypothesis is correct, P(X|H) denotes the probability of evidence being present, whereas P(H|X) represents the probability of hypothesis being present if evidence is there.

2.5.2 Naive bayes

The Bayesian classification tool is based on Bayes theorem and is also known as Naive Bayes classifier. It is probability-based classifier that is a subset of the Bayesian network. All characteristics in the Naive Bayes network are conditionally independent. As a result, adjustments to one feature have no effect on another. The Naive Bayes technique is well suited for categorizing large datasets. Conditional independence is used in the classifier algorithm. The value of an attribute is unconnected to the values of other characteristics in a class, according to conditional independence.

Assume D is a set of training data with their corresponding class labels. There are n characteristics in each tuple in the collection., denoted by X = (A1, A2, …..,,An). Assume m classes represented as C1, C2,... Cm. This classifier tool predicts that a tuple X belongs to the class with the highest posterior probability, and is conditioned on X. The Naïve Bayes classifier shows that

the tuple X belongs to the class Ci if and only if

$$P(Ci \mid X) > P(Cj \mid X) \quad for\ 1 \leq j \leq m, j \neq i$$

Equation 6

As a result, $P(Ci \mid X)$ is maximized. The maximum posteriori hypothesis refers to the class Ci for which $P(Ci \mid X)$ is maximized. Bayes' theorem states,

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)}$$ Equation 7

If the attribute values are conditionally independent, then

$$P(X \mid C) = \prod^{n} P(X_k|C_i)$$ Equation 8

where Xk refers to the value of attribute Ak for tuple X.

Bayesian classifiers are successful in that they are highly scalable, have the lowest classification error rate, and does not require huge data to make accurate classifications.

2.5.3 Random forest

Random forest is a classification system that is based on trees. The program builds a forest with a huge number of trees, as the name suggests. It's an ensemble algorithm with many algorithms. It generates a collection of decision trees using randomly selected subset of training data. It then continues the procedure with many samples selected randomly and decides by majority vote. Although the random forest method is good at coping with missing variables, it is likely to be affected by overfitting. Sometimes appropriate parameter adjustment may be utilized to manage overfitting. To achieve optimum outcome, the random forest classifier constructs integral number of numerous decision trees. It generally relates tree learning to bootstrap aggregation.

Let the given data be XD {x1, x2, x3; : : : : : : ; xn) with known responses YD { x1, x2, x3; : : : : : : ; xn} with an upper limit of B and a lower limit of b D 1. The prediction of sample x0 is made by averaging predictions from each individual x0 tree using the equation:

$$j = \frac{1}{B}\sum_{b=1}^{B} f(x') \qquad \text{Equation 9}$$

The random forest (RF) classifier is a combination of several distinct tree predictors, and is often employed for analysis of huge data. It represents a learning technique that groups, uses regression, and other functions in the ensemble.

2.5.4 Boosting

Boosting is an ensembling algorithm. Boosting divides the original dataset into several subgroups. A (subset) fraction of the data is used to train the classifier, which results in a set of moderately performing models. The components that the previous model couldn't reliably categorize are used to create new subsets. Then, using a cost function, the assembly process improves their performance by integrating the weak models. The boosting method combines predictions that belong to different types, where models are weighted according to their performance.

2.5.5 Bagging

Additionally, bagging is referred to as a bootstrap aggregation. Certain patterns from the training set are randomly selected and replaced using bagging. With the exception of a few omissions and duplicates, the freshly built training set is likely to have exact same amount of patterns as the previous set for training. Bootstrap replicate refers to the fresh training set. Bagging is the process of extracting bootstrap samples from the data and then trains the classifier on each sample. Each classifier's voting is aggregated, and the result of classification is selected by majority vote or average. Bagging

has been shown to be an effective method of optimizing the performance of a poor classifier. Bagging reduces prediction variance by generating numerous datasets with replacement from samples taken randomly from the original dataset. The final aggregate classifier can be written as:

$$\tilde{f}(x) = sign\left(\sum_{i=1}^{T} f(x_i)\right) \qquad \text{Equation 10}$$

or based on majority vote:

$$\tilde{f}(x) = sign\left(\sum_{i=1}^{T} sign\,(f(x_i))\right) \qquad \text{Equation 11}$$

where T is number of sets/groups, and f (x_i) are the outputs.

2.5.6 Multilayer perceptron

Artificial neurons typically are utilized in numerous layers of the multilayered perceptron method, including hidden layers. These methods are used to address binary classification problems. In a perceptron, each neuron carries activation function. These neural networks have been programmed into algorithms. They employ artificial neuron is dubbed as Perceptron. The activation function translates each neuron's weighted inputs and decreases the layer count to two. A perceptron learns by altering its weights.

For supervised learning, back-propagation (BP) algorithm is used for different classifiers. After determining the architecture of an MLP, the network's weights must be generated through a training procedure. The training is done using a training pattern and the intended output of the network. The approach is one of many gradient descent techniques that are used to lower the performance function by updating the network weights along the performance function's negative gradients. BP's operations are divided into two segments. To begin, the network input layer is supplied with a training input pattern. The input pattern is then transmitted through the network's layers until the output layer produces an output. If the actual output does not match the expected output, adjusting the network's weights as a

result of the method an error is computed and sent backwards across the network from the output to the input layer.

BP may be implemented in two methods. The first is by incremental training, which computes the gradient and updates the weights after each input is applied to the network. Second, using batch mode, in which all inputs are applied to the network prior to updating the weights. To determine the change in weights and biases, the gradients obtained at each training example are combined together. For systems with a greater number of training samples, incremental training is often considered to be more efficient and also quicker. When the BP method is used, the total squared error may be used to determine the network's performance. This error is calculated for each training pattern, and the back-propagation algorithms seek to reduce the criterion's average value over all training patterns. This average represents the Mean Squared Error (MSE) between the actual and intended outputs, as defined in the formula below:

$$MSE = \frac{1}{2}\sum_{k=1}^{nk}(t_k - a_k)^2 \qquad \text{Equation 12}$$

The input dataset is partitioned into training dataset (80%) and 20% of the test dataset (10%). A dataset for training is a collection of data meant for training a model. The testing dataset is meant for evaluating the performance of trained model.

2.5.7 Decision tree

With only two number classes, the decision tree (DT) algorithm is one of the most powerful and well-known prediction devices [39]. In a decision tree structure, each internal node represents a property being tested, each branch represents an outcome, and each leaf represents a different class. An inverse-tree progression approach is commonly used in 'learning' based on decision trees. The approach may be used for solving regression and classification

problems. From the root node, the tree grows by choosing the 'Best Feature/Attribute' from a list of attributes, then splits. The 'Best Attribute' is frequently determined by calculating two additional metrics: 'Entropy,' and 'Information Gain,' and is the one that contains the most helpful data. The term "Entropy" refers to the degree to which a dataset is homogenous, whereas "Information Gain" refers to the pace at which the entropy of attributes increases or decreases [40], and is given by

$$I(p(y_1), p(y_2), p(y_3) \dots \dots p(y_k) = -\sum_{i=1}^{k} p(y_i)\, log_2 p(y_i)$$

Equation 13

where yi are possible features and p(yi) are corresponding probabilities.

2.5.8 Support vector machine

Support vector machines (SVMs) are outstanding but versatile machine learning techniques that may be utilized for regression and classification. When compared to other machine learning algorithms, SVMs feature a distinct execution approach. They have recently gained popularity due to their ability to cope with a wide range of continuous and categorical data. SVM's purpose is to divide datasets into classes so that the most extreme peripheral hyperplane can be found. First SVM will produce hyperplanes iteratively that isolates the classes in the most ideal manner. Second, the hyperplane that successfully separates the classes will then be chosen.

The support vector machine algorithm for binary classification is one-vs.-one, and is mathematically defined as [41]:

$$Q = \frac{1}{2} \|w\|^2 - \sum_i \propto_i \; [y_i(\overline{w}.x_i + b) - 1]$$

Equation 14

where $\propto_i$, x_i, y_i, b, w denote constant(s), input, outpu7t class (+1/-1), bias, and final weights once Q is set to minimum [41], and:

$$\sum_i \propto_i y_i = 0\,,\; where\; 0 \leq \propto_i \leq C\,;$$

Equation 15

where C being a constant, is satisfied. The resulting vector of weights and the decision for class "+" is calculated as [41]:

$$\bar{w} = \sum_i \propto_i \ [y_i(x_i)]; \ (\sum_i \propto_i y_i \bar{x}_i . \bar{u} + b) \geq 0,$$

Equation 16

For multi-SVM, the coding matrix is K×Q, where K represents the classes number and Q is the number of learners. The algorithm executes in a way such that all data are compared two at a time until each class becomes unique.

2.6 Processing Time

In this section, comparative results are presented for selected classifiers versus feature selection algorithm during training with k=10 cross validation. The selected classifiers picked are random forest, support vector machine, K-NN and Naïve Bayes classifier since each of these have distinct classification approach. The results presented in Table 2 will enable corresponding experimental work in chapter 3.

Table 2: Processing time (sec) for six feature selection versus classifiers

	LASSO	FCBF	mRMR	Relief
Support vector machine (Linear)	10.18	18.23	18.24	0.003
Random forest	2.2	14.34	15.125	1.1
Naïve Bayes classifier	1.59	34.11	34.11	6.59
K-NN (K=7)	10.15	24.5	29.43	0.002

The table shows that Relief selection algorithm takes least time for all classifiers. However, it should be noted that training is done offline and thus it has lower significance compared to testing/run time.

2.7 Evaluation Metrics

The authors [38] provided measures based on the spatial intersection of ground-truth and system-generated results and then produced multiple performance metrics, which were then averaged for all sampled frames. The examples of various detector performance keys precision, recall, and precision The True Positive is the number of good (safe and uncluttered) detection found by the algorithm. The amount of non-good detection mistakenly identified as excellent detection by the algorithm, and the number of good detections overlooked by the system, are known as false positives and false negatives, respectively.

The other evaluation metrics may be briefly defined. The confusion matrix contains the values True positive, True negative, False negative, and False positive, and the metrics list includes sensitivity, accuracy, F1-score, specificity, recall, Mathew Correlation-coefficient (MCC), AUC-score, and ROC curve, all of which are calculated from the confusion matrix. The evaluation matrices are described below:

Accuracy: This represents the model's overall performance and may be computed using the formula below.:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100$$

Equation 17

Specificity: This is the proportion of newly categorized healthy individuals to whole number of healthy individuals. Following formula is used for calculating specificity:

$$Specificity = \frac{TN}{TN+FP} * 100$$

Equation 18

Sensitivity (Recall): This is the proportion of newly diagnosed heart patients to the overall number of heart patients. It can be calculated as:

$$Sensivity = \frac{TP}{TP+FN} * 100$$

Equation 19

Precision: Precision is defined as the ratio of the actual positive score to the positive score predicted by the classification model, and it is determined using the formula:

$$Precision = \frac{TP}{TP+FP} * 100$$

Equation 20

F1-score: F1 is measure of both recall precision. The value ranges from 0 to 1. The value closer to 1, the better the classification model performance. It can be calculated as:

$$F1 = \frac{2*(Precision*Recall)}{recision+Recall}$$

Equation 21

MCC: MCC is a correlation coefficient that compares actual and projected outcomes. MCC produces numbers ranging from -1 to 1 and can be calculated as follows:

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Equation 22

Finally, we look at AUC (Area Under the Curve), which summarizes a classifier's ROC. The classification algorithm performance is closely related to AUC, i.e., the higher the AUC value, the better the classification algorithm's performance. The ROC (receiver operating characteristic) curve consists of the TP rate as the y-axis and FP rate as the x-axis with the area under the ROC curve (AUC) being calculated to show the performance of the classifier.

2.8 Proposed Methodology

The methodology of the proposed approach can be described in stages. These include pre-processing, features selection, training, classification, and performance evaluation. The framework adopted in this thesis is displayed in Figure 1, and individual parts are discussed below.

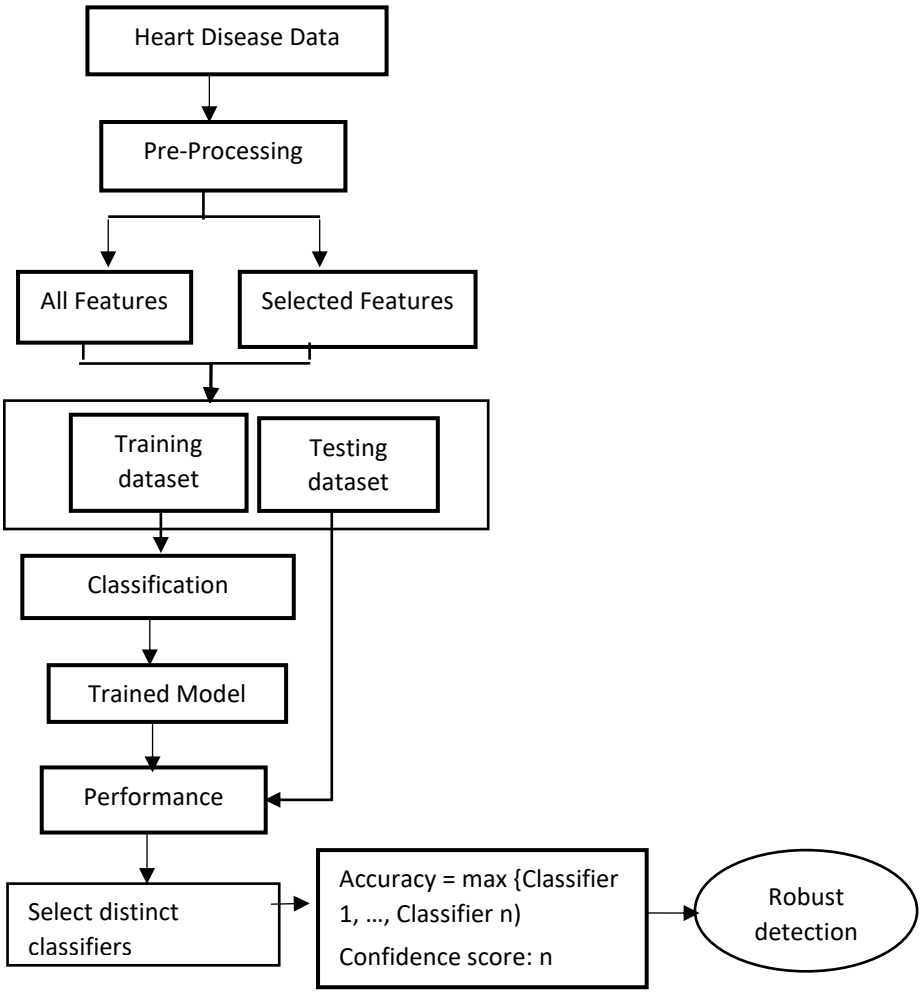


Figure 3: Proposed methodology for robust detection

The Figure 2 depicts the suggested methodology's experimental procedure. As a preliminary step, the Cleveland heart disease dataset was downloaded in.csv format from the UCI machine learning repository. We examined the attributes, value ranges, kinds, and other statistical data after entering the data into the software application. The next step was to pre-process the data, which comprised tasks such as locating missing values in the dataset and replacing them with the user constant or the mean value, depending on the kind of attribute to improve the performance of machine learning classifiers. Following that, cross-validation and a number of machine learning approaches were used to categorize the data.

As discussed in section 2.1, data pre-processing is to be conducted in transforming the raw data fit for processing. The steps that need to be conducted are removal of missing values, min–max scalar and standard scalar on the dataset to make classification more meaningful. After this step, features are to be selected to generate optimal feature sub-space, as it may occur that the inclusion of irrelevant features may degrade the classification performance. The feature selection helps to identify the most important features of heart diseases, and build a relatively better model by increasing learning, and reducing training time. The algorithm selected to select optimal features is relief feature selection algorithm. The steps needed to accomplish this have been stated in section 2.3.4. For binary classes and L_2 distance to calculate difference, this weight update formula may also be written as:

$$W[A] = W[A] - \left(x_j - nearHit_j\right)^2 + \left(x_j - nearMiss_j\right)^2$$

Equation 23

where nearMiss is the closest different-class instance and nearHit is the closest same-class instance. The Manhattan distance (L_1) is used in this work to find these instances rather than the Euclidean distance (L_2). Finally, the high score features are chosen for development of the model. The ideal number of features is identified for each feature subset. The search starts with

a five-feature subset, and the subset is improved ($n = 6$) in the second iteration, with the highest feature significance. This approach continues to rank features up to n = 10 to create distinct feature subsets.

After optimal features are selected, the next step to be followed is training and classification. For robustness in results, different tests should be conducted. In this thesis, this is equivalent to using multiple classifiers. Thus, testing becomes more focused and helps determine whether or not a system generates results with confidence. In hospitals, typically, a specialist or a doctor views medical test reports before making any judgement. If number of distinct test reports indicate the same results, the confidence of the doctor improves to make judgement in that direction. Thus, there is a direct relationship. More, the number of distinct classifiers detect in the same direction, the higher the confidence. The maximum accuracy, thus obtained would be:

$$\left.\begin{array}{c} Accuracy = \max\ (classifier\ 1, classifier\ 2, \ldots, classifier\ n) \\ where \quad n = confidence\ score \end{array}\right\}$$

Equation 24

where classifiers used in the equation are those, which detected in the same direction (either true or false). Thus, a system running on a processor with two cores with each core capable of two threads, can test four classifiers and generate result with a maximum confidence of four and accuracy as per equation 24.

2.9 Computational Complexity

How much time of a computational platform or how many resources does a model or an algorithm consume is also important for timely detection and prediction of a developed system. There are two types of computational complexities, namely run time/space complexity that refers to time it takes to complete operations during run time, and training complexity that refers to

time it takes to complete training of a model for all inputs. The focus is on run time/space complexity as it was assumed that training platform will have enough resources and that training is done offline. In other words, the preference is on timely performance over cost incurred during training. For this purpose, the relative computational cost of each algorithm, planned for performance evaluation in the next section, is presented below. To tabulate the complexity of machine algorithms, let us define the variables that are used to estimate the complexity of a machine algorithm, as follows:

Number of training examples = n , Number of neighbors/support vectors/decision trees = k

Dimensions in the data = d , Classes = c

Table 3: Training, run time complexity and processing time of used algorithms

No.	**ML Algorithm**	**Training time complexity**	**Run time/space complexity**	**Processing time (s)**
1	Naïve Bayes	O ($n * d$)	O ($c * d$)	34.10
2	Support Vector Machine	O (n^2)	O ($k * d$)	14.13 (RBF)
3	K-nearest neighbor	O ($k * n * d$)	O ($n * d$)	18.22
4	Decision Tree	O (n *log(n)* d)	O (maximum depth of the tree)	24.4
5	Random Forest	O (n *log(n) *d * k)	O (depth of tree* k)	20.91

It is obvious from the table that training time complexities vary but run time complexity of random forest, K-NN and support vector machine are lower than all other machine learning algorithms. It should be noted, here, that processing time depends on the given parameters used in a particular algorithm and the machine used to run the classifier for testing. Thus, the processing time may vary depending on the actual machine used.

In the next chapter, the experimental results based on discussion in this section are presented.

Chapter 3

Chapter 3: Experimental Results

In this section, we will present experimental results obtained using the proposed approach and a number of well-known classifiers implemented on heart disease dataset. As stated in earlier Chapter 2, the dataset employed for this study is the Cleveland dataset. The classifier list implemented in this experiment include Random Forest, Decision Tree, Naive Bayes, SVM, and KNN. The measures used to evaluate the algorithm's performance include Accuracy score, Precision (P), Recall (R), and F-measure. The precision metric measures the accuracy of affirmative analysis. In simulations, the pre-processed dataset is utilized to conduct experiments, and the afore-mentioned classification approaches are investigated, and evaluation metrics measured.

Confusion Matrix describes the model's performance. Using the confusion matrix parameters, the previously indicated performance measures are calculated. Table 3 displays the confusion matrix parameters for various methods. Table 4 displays the accuracy and other metrics scores calculated for classification algorithms. The performance of all classifiers was thus to be determined by an empirical evaluation of their corresponding success rates. The training, validation and testing was done using Google Collaboratory in python language. The hardware machine specification of google colab were as follows:

Model name	: Intel(R) Xeon(R) CPU @ 2.30GHz
CPU MHz	: 2299.998
Cache size	: 46080 KB
Cpu cores	: 2
RAM	: 12GB
GPU	: Nvidia K80 / T4

GPU Memory Clock	0.82GHz / 1.59GHz
Performance	4.1 TFLOPS / 8.1 TFLOPS
Max execution time	12 hours
Max idle time	90 min

Table 4: Values obtained for confusion matrix using different algorithms

	Algorithm	True Positive	False Positive	False Negative	True Negative
1	NB	21	6	3	31
2	Linear SVM	21	5	3	30
3	K-NN	22	5	4	30
4	DT	25	2	4	30
5	RF	22	5	6	28

Table 5: Classification results on full feature space

	Algorithm	Accuracy	Precision	Recall	F-1 Score
1	Naive Bayes (NB)	86.7	83.7	91.1	87.3
2	Linear SVM	81.97	90.0	88.0	89.0
3	K-NN	86.81	88.0	88.0	88.0
4	DT	81.97	84.5	82.3	83.5
5	RF	90.16	93.7	88.2	90.16

The Table 5 suggests that random forest outperformed all classifiers with respect to accuracy parameter, whereas random forest and support vector machine together performed equally well based on precision, recall and F1-score parameters and thus outperformed other classifiers. In contrast, NB performs the worst with respect to sensitivity, specificity, and accuracy, while NB and DT reported weaker performance compared to the other classification algorithms. As such random forest and SVM (and K-NN to some extent) distinguish themselves as robust classifiers when compared to other models evaluated on the dataset utilized in this experiment. On one hand, SVM attempts to optimize separation between the two classes, whereas random forest attempts to construct classes statistically. Some of the previously listed studies have tried to construct an ensemble of classifiers, but for the purpose of comparing them only. The research in this thesis has concluded that heart illness may be detected from findings of SVM, random forest and any other distinct classifier like K-NN (all executed in parallel) to generate a confidence score (as in Equation 24), and maximum accuracy of 90.16% using full feature space. The integration of classifiers improves the confidence of the medical specialist in true positives and true negatives of the results.

The simulation was repeated based on Relief sub-feature space. The results are shown in Table 6. The measurements based on Table 6 indicate that RF classifier performance is greatly impacted by employing sub-feature spaces of varied lengths. All of these accomplishments are attributed to the best (feature) selection strategy that not only minimizes the feature space but improves the classification ability as well. Thus, the similar conclusive approach as represented in equation (24) can also be made for classifier performance based on selected feature subspace.

Table 6: Classification Results on selected feature space

	Algorithm	Accuracy	Precision	Recall	F-1 Score
1	Naive Bayes (NB)	86.7	85.7	91.9	88.2
2	Linear SVM	91.73	92.3	88.2	89.1
3	K-NN	87.11	86.06	84.41	87.2
4	DT	84.98	85.9	89.16	84.3
5	RF	94.50	93.7	89.2	91.16

A comparison may also be made between the proposed approach and other contemporary machine learning techniques reported in the literature. The findings indicate that our suggested model has a higher success rate than other models in the literature. The results thus accumulated are shown in Table 7.

Table 7: Proposed system compared to existing methods using a heart disease dataset.

	Research contribution	Algorithm	Accuracy
1	Cited in reference [24]	Hybrid framework	86.00
2	Cited in reference [43]	HRFLM	88.70
3	Cited in reference [44]	ANFIS	91.00
4	Cited in reference [3]	ANN-fuzzy-AHP	91.10
5	Cited in reference [45]	Stacked SVM	91.11
6	This thesis work	RF boosted with Relief	94.50

Chapter 4

Chapter 4: Discussion and Conclusions

With the rising number of deaths due to heart disease, it has become necessary to design a system that can efficiently and precisely anticipate heart disease. The goal of the research was to discover the best effective machine learning system for detecting cardiac problems. In this thesis, the importance of early coronary disease detection was discussed and the need to bolster the current technological efforts to produce more accurate system in facilitating the whole detection process was also highlighted. In the same direction, current literature about heart disease detection was also surveyed and weakness of current solutions were identified.

This thesis study made use of the heart disease dataset that can be publicly found in the machine learning repository at UCI. Using a variety of data mining strategies, such as Naive Bayes, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), and Random Forest, this research estimated the likelihood that a patient would develop heart disease and thus categorize the patient's degree of risk. The classifiers performance was evaluated using an approach for selection of features on both the complete and selected feature space. The selected models were trained and tested on the Cleveland heart disease dataset. Several performance metrics were employed to measure effectiveness of heart disease detection process. The computational complexity of each machine learning model and its performance were chosen to build robustness in the proposed approach.

The results of the experiment demonstrate that the Random Forest was the top performer with accuracy of 94.5% followed closely by Support Vector machine algorithm with accuracy of 91.7% on selected feature space. Thus, these two top performing classifiers turned out to be promising classifiers for heart disease prediction. The SVM, on one hand, attempts to optimize separation between two classes, whereas random forest constructs classes

statistically. Since the testing of these classifiers is typically conducted on an embedded system for real time prediction, the characteristics of these classifiers can be combined to generate robustness in heart disease detection. For a robust heart disease prediction in an embedded platform, the proposed system stated that the heart disease exists if and only if majority of classifiers detect it, and thus provide a confidence of that majority vote.

As far as future direction in this research is concerned, a number of strategies can be thought about. The first and foremost is to implement a detection system for heart illness in real time. This direction can be easily facilitated by deploying currently available embedded system that can implement both classifiers in parallel.

Another direction is to facilitate the hospital cardiac team through online social network, similar to works in [46-47] where patient data can be transported to the hospital system for real time monitoring. This study may also be improved by using a larger heart disease repository to improve training results which may lead to better detection results.

References

1. Groff, E. R., & La Vigne, N. G. (2002). Forecasting the future of predictive crime mapping. Crime Prevention Studies, 13, 29-58.

2. Inés, A., Domínguez, C., Heras, J., Mata, E., & Pascual, V. (2021). Biomedical image classification made easier thanks to transfer and semi-supervised learning. Computer methods and programs in biomedicine, 198, 105782.

3. Samuel, O. W., Asogbon, G. M., Sangaiah, A. K., Fang, P., & Li, G. (2017). An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. Expert Systems with Applications, 68, 163-172.

4. Olaniyi, E. O., Oyedotun, O. K., & Adnan, K. (2015). Heart Diseases Diagnosis Using Neural Networks Arbitration. International Journal of Intelligent Systems & Applications, 7(12).

5. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., ... & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American journal of cardiology, 64(5), 304-310.

6. Parthiban, L., & Subramanian, R. (2008). Intelligent heart disease prediction system using CANFIS and genetic algorithm. International Journal of Biological, Biomedical and Medical Sciences, 3(3).

7. Palaniappan, S., & Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. In 2008 IEEE/ACS international conference on computer systems and applications (pp. 108-115). IEEE.

8. Paul, A. K., Shill, P. C., Rabin, M., Islam, R., & Murase, K. (2018). Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease. Applied Intelligence, 48(7), 1739-1756.

9. Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. Expert systems with applications, 36(4), 7675-7680.

10. Alizadehsani, R., Hosseini, M. J., Khosravi, A., Khozeimeh, F., Roshanzamir, M., Sarrafzadegan, N., & Nahavandi, S. (2018). Non-invasive detection of coronary artery disease in high-risk patients based on the stenosis prediction of separate coronary arteries. Computer methods and programs in biomedicine, 162, 119-127.

11. Golande, A., & Pavan Kumar, T. (2019). Heart disease prediction using effective machine learning techniques. International Journal of Recent Technology and Engineering, 8(1), 944-950.

12. Alotaibi, F. S. (2019). Implementation of machine learning model to predict heart failure disease. International Journal of Advanced Computer Science and Applications, 10(6).

13. Repaka, A. N., Ravikanti, S. D., & Franklin, R. G. (2019, April). Design and implementing heart disease prediction using naives Bayesian. In 2019 3rd International conference on trends in electronics and informatics (ICOEI) (pp. 292-297). IEEE.

14. Lutimath, N. M., Chethan, C., & Pol, B. S. (2019). Prediction of heart disease using machine learning. International Journal of Recent Technology and Engineering, 8(2), 474-477.

15. Subhadra, K., & Vikas, B. (2019). Neural network based intelligent system for predicting heart disease. International Journal of Innovative Technology and Exploring Engineering, 8(5), 484-487.

16. Memon, Q., & Khan, S. (2001). Camera calibration and three-dimensional world reconstruction of stereo-vision using neural networks. International Journal of Systems Science, 32(9), 1155-1159.

17. Memon, Q. A., & Laghari, M. S. (2006). Building relationship network for machine analysis from wear debris measurements. International Journal of computational intelligence, 3(2).

18. Valappil, N. K., & Memon, Q. A. (2021). CNN-SVM based vehicle detection for UAV platform. International Journal of Hybrid Intelligent Systems, (Preprint), 1-12.

19. Memon, Q. (2019). On assisted living of paralyzed persons through real-time eye features tracking and classification using Support Vector Machines. Medical Technologies Journal, 3(1), 316-333.

20. Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. Expert systems with applications, 35(1-2), 82-89.

21. Gudadhe, M., Wankhade, K., & Dongre, S. (2010, September). Decision support system for heart disease based on support vector machine and artificial neural network. In 2010 International Conference on Computer and Communication Technology (ICCCT) (pp. 741-745). IEEE.

22. Thomas, J., & Princy, R. T. (2016, March). Human heart disease prediction system using data mining techniques. In 2016 international conference on circuit, power and computing technologies (ICCPCT) (pp. 1-5). IEEE.

23. Manogaran, G., Varatharajan, R., & Priyan, M. K. (2018). Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system. Multimedia tools and applications, 77(4), 4379-4399.

24. Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems, 2018.

25. Chaurasia, V., & Pal, S. (2014). Data mining approach to detect heart diseases. International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol, 2, 56-66.

26. Tomov, N. S., & Tomov, S. (2018). On deep neural networks for detecting heart disease. arXiv preprint arXiv:1808.07168.

27. Nagamani, T., Logeswari, S., & Gomathy, B. (2019). Heart disease prediction using data mining with mapreduce algorithm. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN, 2278-3075.

28. Abubaker, M. (2022). Detection of Cardiovascular Diseases in ECG Images Using Machine Learning and Deep Learning Methods. IEEE Transactions on Artificial Intelligence.

29. Amarbayasgalan, T., Pham, V. H., Theera-Umpon, N., Piao, Y., & Ryu, K. H. (2021). An efficient prediction method for coronary heart disease risk based on two deep neural networks trained on well-ordered training datasets. IEEE Access, 9, 135210-135223.

30. Shuvo, S. B., Ali, S. N., Swapnil, S. I., Al-Rakhami, M. S., & Gumaei, A. (2021). CardioXNet: A novel lightweight deep learning framework for cardiovascular disease classification using heart sound recordings. IEEE Access, 9, 36955-36967.

31. Ali, S. A., Raza, B., Malik, A. K., Shahid, A. R., Faheem, M., Alquhayz, H., & Kumar, Y. J. (2020). An optimally configured and improved deep belief network (OCI-DBN) approach for heart disease prediction based on Ruzzo–Tompa and stacked genetic algorithm. IEEE Access, 8, 65947-65958.

32. Predicting heart disease using machine learning. (2022). Retrieved 30 June 2022, from https://www.kaggle.com/code/faressayah/predicting-heart-disease-using-machine-learning

33. Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, 27(8), 1226-1238.

34. Zhou, C., & Wieser, A. (2018, January). Jaccard analysis and LASSO-based feature selection for location fingerprinting with limited computational complexity. In LBS 2018: 14th International Conference on Location Based Services (pp. 71-87). Springer, Cham.

35. Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(3), 273-282.

36. Chikhi, S., & Benhammada, S. (2009). ReliefMSS: a variation on a feature ranking ReliefF algorithm. Int. J. Bus. Intell. Data Min., 4(3/4), 375-390.

37. Kira, K., & Rendell, L. A. (1992, July). The feature selection problem: Traditional methods and a new algorithm. In Aaai (Vol. 2, No. 1992a, pp. 129-134).

38. Muhammad, Y., Tahir, M., Hayat, M., & Chong, K. T. (2020). Early and accurate detection and diagnosis of heart disease using intelligent computational model. Scientific reports, 10(1), 1-17.

39. Thenmozhi, K., & Deepika, P. (2014). Heart disease prediction using classification with different decision tree techniques. International Journal of Engineering Research and General Science, 2(6), 6-11.

40. Azad, C., & Jha, V. K. (2015). Genetic algorithm to solve the problem of small disjunct in the decision tree based intrusion detection system. International Journal of Computer Network and Information Security, 7(8), 56-71.

41. Cortes, V. (1995). Cortes C., Vapnik V. Support-vector networks, Machine learning, 20(3), 273-297.

42. Bashir, F., & Porikli, F. (2006, June). Performance evaluation of object detection and tracking systems. In Proceedings 9th IEEE International Workshop on PETS (pp. 7-14).

43. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 7, 81542-81554.

44. De Silva, A. M., & Leong, P. H. (2015). Grammar-based feature generation for time-series prediction (pp. 978-9812874108). Berlin, Germany: Springer.

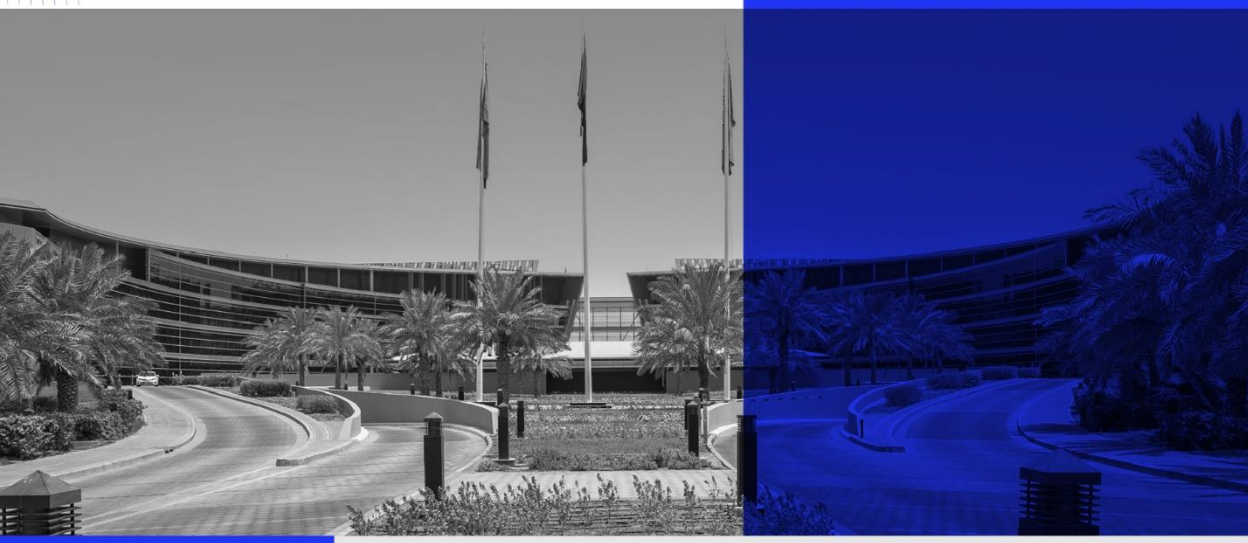
45. Ali, L., Niamat, A., Khan, J. A., Golilarz, N. A., Xingzhong, X., Noor, A., ... & Bukhari, S. A. C. (2019). An optimized stacked support vector machines based expert system for the effective prediction of heart failure. IEEE Access, 7, 54007-54014.

46. Memon, Q. (2013). Smarter Healthcare Collaborative Network. Building Next-Generation Converged Networks: Theory.

47. Memon, Q. A., & Mustafa, A. F. (2015). Exploring mobile health in a private online social network. Int. J. Electron. Heal., 8(1), 51-75.

List of Publications

Domyati. A, and Memon. Q, (2022). "Robust Detection of Cardiac Disease Using Machine Learning Algorithms", *International Conference on Control and Computer Vision* (ICCCV), Xiamen University, China,



UAE UNIVERSITY MASTER THESIS NO. 2022:33

The contribution of the current work is to facilitate diagnose the heart disease based on contemporary machine learning algorithms. The performances of the classifiers are tested on feature spaces selected through various feature selection algorithms. The models were trained and tested on the Cleveland heart disease datasets. Several performance measures are used to observe the effectiveness of the selected models. It is found out that SVM and random forest achieved very promising results with both full feature space and selected feature space, specifically with relief feature selection algorithm compared to other ML algorithms implemented.

Anas Haitham Domyati received his Master of Science in Electrical Engineering from the Department of Electrical Engineering, College of Engineering at UAE University, UAE. He received his BSc from the Department of Electronics and Communication Engineering, College of Engineering at UAE University, UAE.