Arab American University

Faculty of Graduate Studies

Early and Accurate Prediction of Heart Disease Using Machine Learning: a feature selection

Prepared By

Mohammad Fathallah Ghanem

Supervisor By

Prof. Dr. Yousef Abuzir

This Thesis Was Submitted in Partial Fulfillment of the Requirements for the Master's Degree in Computer Science

January / 2023

Thesis Approval

Early and Accurate Prediction of Heart Disease Using Machine Learning: a feature selection

By

Mohammad Fathallah Ghanem

This thesis was defended successfully on 20-Feb-2022 and approved by:

Committee members		Signature
1. Prof. Dr. Yousef Abuzir	(Supervisor)	
2. Dr. Muath Sabha	(Internal examiner)	
3. Dr. Yousef Draghmeh	(External examiner)	

Declaration

I declare that the thesis titled "Early and Accurate Prediction of Heart Disease Using Machine Learning: a feature selection" is my work, and has been composed solely by myself and does not contain work from other researchers and the previous work is all cited, and has not been submitted for any other degree or scientific work except the reference is made.

Name: Mohammad Ghanem

Date:

Signature: 

Student ID: 202012877

Dedication

I dedicate this thesis to my father Prof. Dr. Fathallah Ghanem, who have always been supportive and constant source of wisdom throughout my life, and to my mother Dr. Batool Ghanem, who have always been my source of inspiration. To my fiancée who always supported me and pushed me forward. To my brothers and sister, whose support I have not always forgotten. To my dear friends and work colleagues, for their continued support throughout my learning and work journey.

Acknowledgment

I would like to take this opportunity to thank Dr. Yousef Abuzir for his advice, assistance, and the time he spent reviewing and revising my work. Dr. Abuzir provided helpful suggestions and guidance that had a significant impact and assisted in overcoming numerous barriers in the preparation of this work in the most effective manner. I would also extend Many thanks to Dr. Majdi Odeh, and Dr. Amani Odeh, and thanks to all faculty members at the Arab American University. You have all my love and respect.

M. Ghanem

Abstract

Early and Accurate Prediction of Heart Disease Using Machine Learning: a feature selection

The goal of this study is to determine the early and accurate prediction of heart disease using machine-learning techniques and feature selection methods. The study develops a machine learning-based technique for predicting heart disease. The study uses data gathered from Al Razi hospital in city of Jenin was 890 records, and because the sample size was insufficient to get accurate prediction, the sample size was increased by using the Generative Adversarial Network (GAN) Algorithm. The GAN algorithm generates synthetic samples that are added to the original dataset to enhance the size of the dataset and boost sample diversity. Simultaneously, feature selection approaches were used to determine the most significant features for heart disease prediction. To accomplish the study's results, machine learning methods, notably K-Nearest Neighbors (KNN), Random Forest, Adaboost, and Support Vector Machine (SVM) were applied to the selected features to create predictions. The results of this study show that the identified approach can achieve 99% accuracy by employing KNN and SVM models, and that using GAN-generated samples and feature selection approaches can improve the performance of machine learning models. These findings show that the approach that, combines feature selection and machine learning algorithms is useful for the early and accurate prediction of heart disease.

Keywords: Heart Disease Prediction, Machine Learning, Feature Selection, Generative

Adversarial Network (GAN) , KNN algorithm, Adaboost, Random Forest, Support

Vector Machine.

Table of Contents

List of Tables

List of Figures

List of Equations

List of Abbreviations

KNN	K-Nearest Neighbors
SVM	Support Vector Machines
GAN	Generative adversarial network
MLT	Machine Learning Techniques
ML	Machine Learning
AI	Artificial intelligence
TP	True Positive
FP	False-Positive
FN	False-Negative
TN	True Negative

Chapter One

Introduction

1.1 Heart Disease Introduction

Heart disease refers to a variety of heart-related ailments. Coronary artery disease, arrhythmias, cardiac abnormalities, and stroke are all examples of cardiovascular disease. Heart disease is the leading cause of death in the United States and worldwide. People of all ages, genders, and nations are affected. High blood pressure, high cholesterol, diabetes, smoking, obesity, physical inactivity, poor diet, and excessive alcohol consumption are all risk factors for heart disease. Depending on the type of ailment, treatment for heart disease may include lifestyle modifications such as eating a balanced diet and exercising frequently, drugs, surgery, or other treatments. Thus, in patients with heart disease, a waxy material forms in the coronary arteries. This accumulation of the waxy substance in the arteries slows down the blood pumping process and, if not (Yusof al., 2019), causes mortality. Heart disease is one of the primary causes of disease and mortality worldwide. One of the most important fields of research in clinical data analysis is cardiovascular disease prediction. Currently, the amount of data in healthcare facilities is massive. Machine learning techniques are widely used in object recognition and disease detection (Suresh et al., 2022). In disease diagnosis, a machine learning algorithm converts a huge collection of healthcare datasets into knowledge that can be utilized to make better informed decisions and projections. One of the goals of machine learning research in the medical research field is disease prediction and the creation of machine-based diagnostics systems, which improves the precision and accuracy of decision making during disease identification and diagnosis.

Given that cardiovascular disease is the main cause of death worldwide, early detection and prevention are critical for reducing morbidity and mortality. Advances in machine learning techniques have opened up new for the prediction and detection of cardiac disease in recent years. Machine learning is a subfield of artificial intelligence that analyzes and learns from data using algorithms and statistical models. It has been used in a variety of medical applications, including heart disease prediction.

Classification is a popular machine learning technique for predicting heart disease. Classification is the process of assigning a class label to a set of features in input data. The input data in the context of heart disease prediction may comprise demographic information, risk factors, and physiological measurements, and the class label may indicate the existence or absence of heart disease. Various machine learning algorithms, such as decision trees, random forests, and support vector machines, have been used to categorize cardiac diseases.

Feature selection is another key machine learning technique for predicting heart disease. The process of picking a subset of characteristics from the input data that are most relevant to the prediction task is known as feature selection. It is significant in heart disease prediction because it can increase prediction model performance while lowering computing costs. In the literature, many feature selection methods such as mutual information, correlation-based feature selection, and genetic algorithms have been used.

Deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are the most modern and prominent approaches in heart disease prediction. These models can learn characteristics from raw data and have been demonstrated to be effective in a variety of medical applications, including image

classification and diagnosis (Zheng et al., 2018; Rajpurkar et al., 2017). CNNs, for example, have been used to classify electrocardiograms (ECGs) and to analyze cardiac MRI images (Rajpurkar et al., 2017) (Zheng et al., 2018).

However, using machine learning to forecast cardiac disease is fraught with difficulties. One of the most significant issues is the scarcity of large and diverse data sets for training and assessing machine learning models. Overfitting occurs when a model performs well on training data but badly on new data. Another issue is the interpretability of machine learning models, which can be difficult for clinicians and patients to understand and explain.

Researcher conclude that machine learning has the potential to improve cardiac disease prediction and diagnosis. However, it is critical to examine machine learning's limitations and constraints in this context, as well as to guarantee that the models developed are interpretable and generalizable.

1.2 Objectives

The purpose of the thesis is to determine the likelihood that healthy people will develop heart disease in the future. Therefore, data from individuals with a history of cardiovascular illness will be required first. These statistics include age, gender, diabetes, chest pain type, cholesterol, resting electrocardiogram results, maximum heart rate achieved, and resting blood pressure. Initially, machine learning techniques will be utilized to classify this mixed data set. Methods of machine learning, such as; K-nearest neighbor, Support vector machines, Adaboost, and random forest classification algorithms, will be utilized. These algorithms were selected since the vast majority of the

data is categorical. After the data are categorized, it will be possible to predict and assess the likelihood of illness or early and accurate detection and diagnosis of heart disease.

In order to expedite the study, interviews with cardiologists were conducted at Al Razi Hospital in Jenin City: namely Dr. Hasan Al Salman and Dr. Abed Alkareem Al Hamshary. Data collection is the most crucial aspect of the study. After collecting patient data, it is crucial to determine which aspects would be most beneficial for classification. Appointments and interviews were scheduled with the cardiology clinic physicians at Al Razi hospital as mentioned earlier. To employ the appropriate procedures, their technical expertise was sought. With the head physician's approval, the required raw data (data set) was gathered in this manner.

In this research, one of the main objectives is to improve the accuracy of heart disease prediction by the use of ML classifiers (K-nearest neighbor, Support vector machines, Adaboost, and random forest) . The other objective is to take necessary early actions to save lives based on the patient records using random forest, adaboost, k-nearest neighbors, and support vector machines. Therefore, the objectives of this work can be illustrated in the following points:

- Creating a machine learning-based system using feature selection for heart disease prediction.
- Using the Generative Adversarial Network (GAN) algorithm to expand the dataset and improve sample diversity.
- Using feature selection techniques to determine the most important features for heart disease prediction.

- Assessing how well selected features and samples produced by GAN are handled by machine learning models, particularly K-Nearest Neighbors (KNN) and Support Vector Machine (SVM).
- Examining if the performance of machine learning models for heart disease prediction may be enhanced by using GAN-generated samples and feature selection techniques.
- Demonstrating that the chosen strategy may produce high accuracy in the early and accurate prediction of heart disease.
- Improving the medical sector in Palestine and solving the delay and overcrowding problems.

1.3 Contribution

Applying ML techniques in this field can be used as a decision support system for the medical applications that can be available at the request of the physicians. It also will serve as a system that will save the analysis time for the authority staff while scanning the history of human being to search for the hidden patterns.

As this study was based on the following:

- The study shows the possibility of using machine learning techniques as decision support tools in medical applications, reducing time for analysis for healthcare personnel and assisting doctors in coming to more accurate diagnoses.

- The study uses four different machine learning approaches (K-nearest neighbor (KNN), Support vector machines (SVM), Adaboost, and random forest) to combine classification strategies that enhance heart disease classification results.
- Using various machine learning algorithms and methodologies, the study suggests and evaluates various approaches and settings to increase the accuracy of the results.
- The study makes use of actual patient data gathered from Jenin City's Al-Razi Hospital, which improves the accuracy and applicability of the findings.
- To improve prediction accuracy, the study uses feature selection approaches like the Wrapper and Filter methods.
- To increase the dataset's diversity and richness for training machine learning models, the study uses the Generative Adversarial Network (GAN) Algorithm to augment it from 890 to 1994 records.

1.4 Thesis Structure

The rest of this work is organized as follows:

Chapter one provides an introduction of the heart disease and the objectives of this study. Chapter two presents review of the related work in machine learning for healthcare applications and some techniques used to predict and classify heart disease and machine

learning with spectral reflectance in healthcare. While the third chapter, illustrates the research methodology used, including methods for both machine learning and classification techniques using 4 machine learning algorithms. Chapter four presents the results and discussion for classification using different evaluation metrics. Chapter five represent the conclusion and the future work.

Chapter Two

Literature Review

2. Literature Review

This chapter summarizes and analyzes significant papers on the thesis subjects of heart disease, machine learning and feature selection.

2.1 Background

Artificial intelligence (AI) is a subfield of computer science that focuses on developing machines capable of doing activities that would normally need human intelligence. Machine learning (ML) is a branch of artificial intelligence that uses algorithms to teach machines to learn from data and improve their performance on certain tasks without being explicitly programmed.

The applications of AI and machine learning are diverse and have been used to solve a wide range of problems in various fields such as natural language processing, computer vision, and speech recognition. In the field of healthcare, AI and machine learning have been used to improve the diagnosis and treatment of various diseases, including heart disease (Abuzir, Y. et al. 2020).

Heart disease is one of the main causes of death worldwide, and predicting it early and accurately is critical for lowering mortality rates. By evaluating massive volumes of patient data, including demographic information, medical history, and laboratory test

results, machine learning has been used to predict the risk of heart disease (Kumari A., Mehta A. K., 2021).

Machine learning algorithms such as decision trees, random forests, and neural networks have been used to predict heart disease. Some studies have used these algorithms to predict the risk of heart disease by analyzing patient data such as age, sex, cholesterol levels, blood pressure, and smoking status. Other studies have used machine learning algorithms to predict the risk of heart disease by analyzing ECG data and other medical imaging data (Mohan S., Thirumalai Ch., Srivastava G, 2019).

In summary, Machine Learning is a branch of computer science and a subdivision of AI, respectively, which involve the application of algorithms to allow machines to learn from data and improve their performance on certain tasks (Rajkumar, D. et al.2019). AI and machine learning have been used in the field of healthcare to improve the diagnosis and treatment of various diseases, including heart disease, by analyzing large amounts of patient data, including demographic information, medical history, and laboratory test results, using various algorithms such as decision trees, random forests, and neural networks (Ramalingam V.V., Dandapath A., Raja M K., 2018). The types of machine learning algorithms are described in the next section...

2.1.1 Types of Machine Learning

Machine learning is a branch of AI that includes training algorithms on data to make predictions or judgments without being explicitly programmed. As illustrated in Figure 2.1, there are three forms of machine learning:

- Supervised Learning
- Unsupervised Learning

- Reinforcement Learning

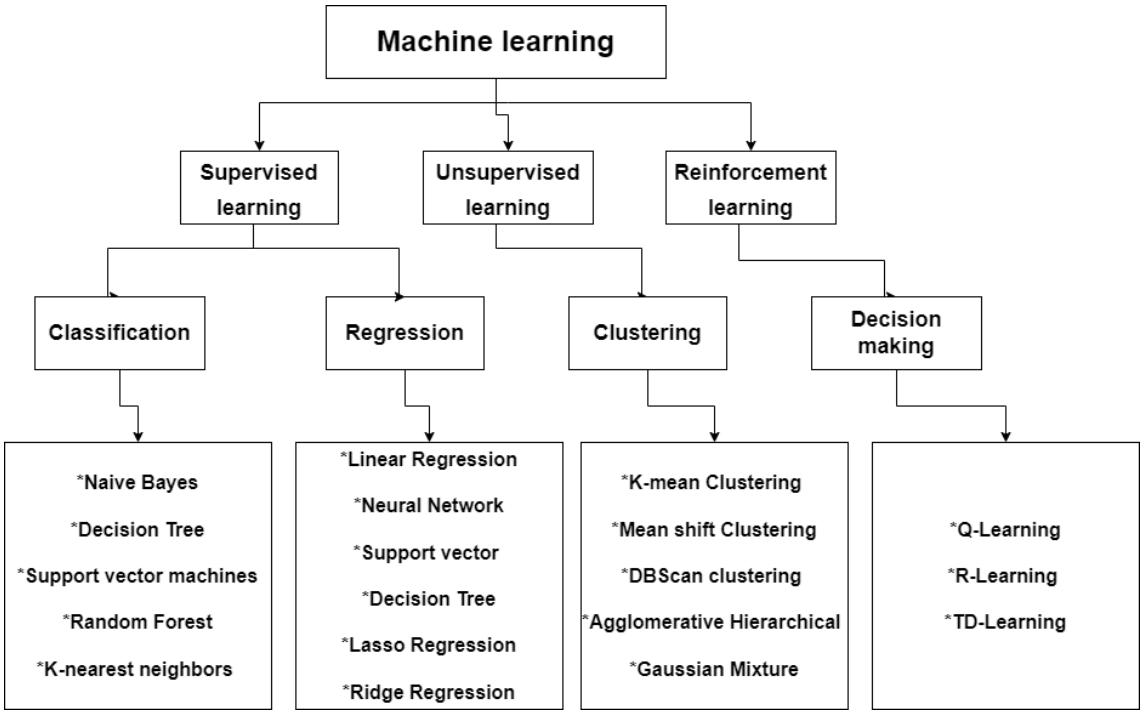


Figure 2.1: Types of machine learning

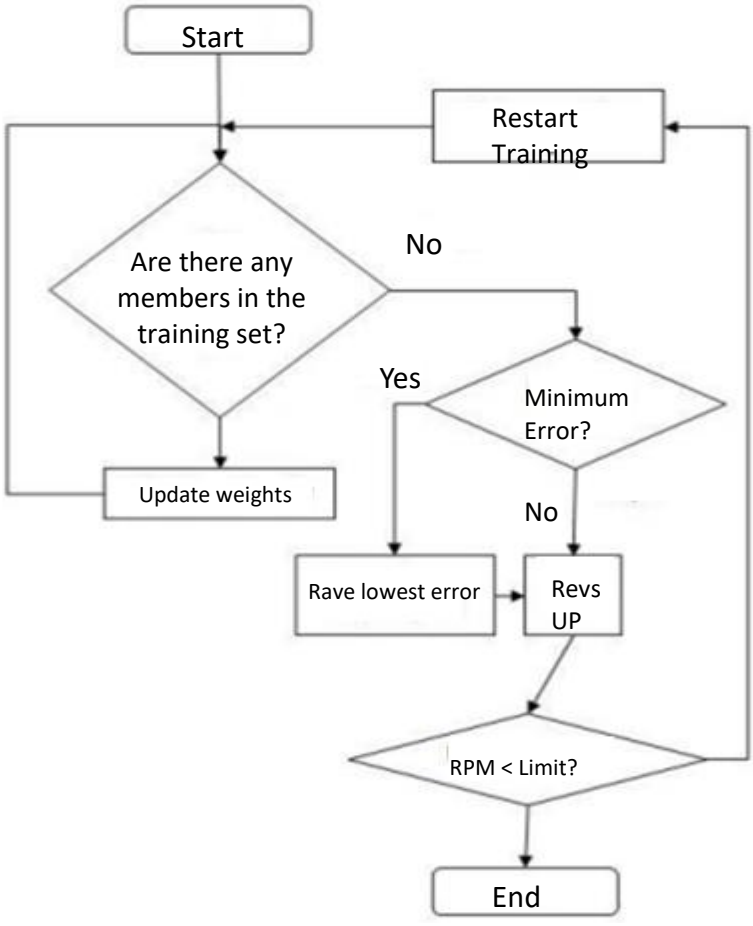
2.1.1.1 Supervised Learning

In the supervised learning method, the desired result set is expected to be obtained from the data given as input data. In the supervised learning method, a training set and the desired output values are given for the values given in the training set to train the machine. Using this labeled training set, the machine learns for future use. In the future, when you want something from the model, it uses this labeled training data to learn the underlying patterns in the data and make predictions about the target variable for new, unseen data. (Radford & Chintala, 2016).

In the supervised learning method, as seen in Figure 2.2, the machine uses the training set and gradually test the percentage of error and updates its solution as an evaluation for the

model before applying the system in real applications to find the lowest error. The aim is to make the most accurate prediction with the lowest margin of error.

For example: Take a city and consider the house prices in that city. You have the house prices data in that city and you have introduced them to the machine. Let’s say the price of a 50 square meter house is 70.000 JD. We introduced this data in our machine training cluster so that the price of another 100 square meter house will be 160,000 JD, and let's say we have a 75 square meter house. We ask the machine to predict a price for this house. Now what the machine needs to do is take the price of the house of 100 square meters, then get the price of the house of 50 square meters and find out, and then predict a price for the house of 75 square meters, which is our house, according to this data (Rajkumar et al., 2019).



The goal of supervised learning is to train a model on a labeled dataset such that it can predict the target variable for new, unseen data. The labeled training data is used by the algorithm to understand the underlying patterns and correlations between the input variables and the target variable. And these algorithms can be divided into two as "Classification" and "Regression" (Mammen, Rothe and Schienle, 2012).

Classification Algorithms,

It is a classification problem-solving supervised learning algorithm. These algorithms are designed to predict a categorical label or class based on the supplied data. Spam detection, sentiment analysis, and image classification are some common examples of classification challenges. The following are examples of classification algorithms:

- Nearest neighbors algorithm (KNN)
- Trees
- Logical Regression
- Naive Bayes Algorithm
- Supported Vector Machines

Regression Algorithms,

It is a supervised learning approach that is used to solve regression problems. These algorithms are designed to predict a continuous value from a set of input data. Predicting home prices, stock prices, and weather conditions are some prominent instances of regression problems. The following are examples on regression algorithms:

- Linear Regression
- Polynomial Regression

- Decision Trees
- Random Forests

2.1.1.2 Unsupervised Learning

Unlike supervised learning, unsupervised learning does not provide an output value. The purpose of unsupervised learning is to learn the relationships and structure in the data (Radford, & Chintala, 2016).

In this method, the algorithm is expected to make self-discoveries and complete the hidden or missing patterns. Also, unsupervised training does not have a training set. Apart from this, the unsupervised learning algorithm is also in itself,

- Clustering
- Relationship analysis

2.1.1.3 Reinforcement Learning

Reinforcement learning is an algorithm in which software agents work with a reward system. To open the reward system a little more. For example: We have an agent in an environment and this agent needs to find the target you set in the environment as soon as possible. Our agent has no intelligence at the initial stage (trial and error method). There are rewards for guiding the agent on the paths he/she will pass, as well as things that give negative points when he/she takes the wrong path. Agent will try to reach the finish line you have determined as soon as possible, taking into account the plus or minus points he/she has collected (Sutton & Barto, 2018).

Figure 2.3 shows the machine learning and methods described above. In addition, an example of its use.

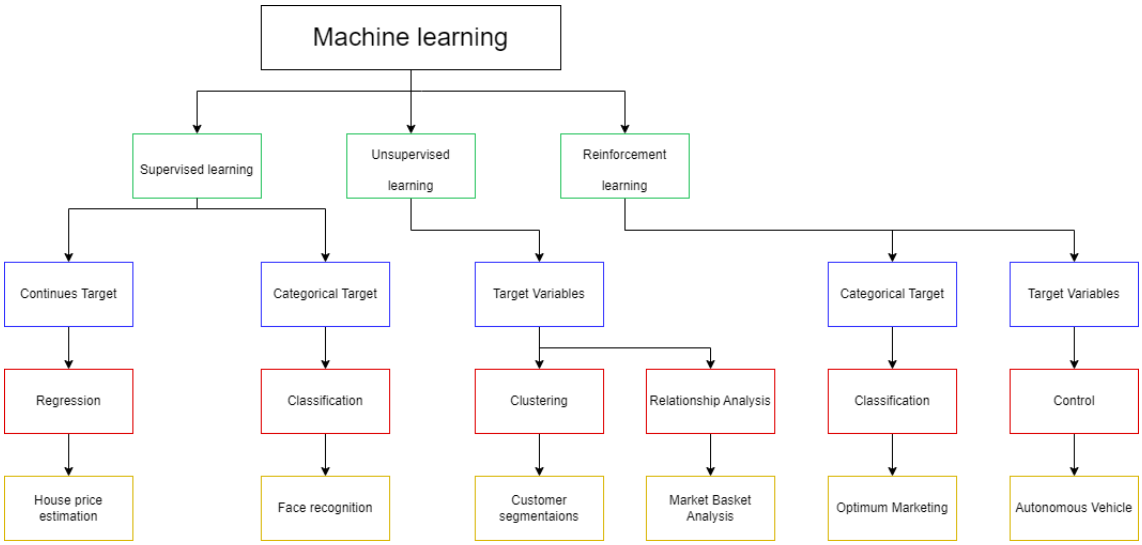


Figure 2.3: The methods used in machine learning with a case study.

2.1.2 Industry-based Use Cases of Machine Learning

Manufacturing: Predictive maintenance and condition monitoring, material and stock forecasting, Trends in purchasing, demand forecasting, process improvement, and telematics.

Predictive inventory planning, recommendation engines, sales and cross-channel marketing, market segmentation and targeting, ROI, and value are all examples of retail applications.

Health and Life Science: Real-time warnings and diagnoses based on patient data, disease diagnosis and risk stratification, patient triage optimization, proactive health management, and healthcare analytics are all examples of health and life science applications.

Dynamic pricing, social media - consumer feedback and interaction analysis, customer complaint resolution, traffic patterns, and congestion management are all aspects of travel.

Finance and financial services include risk analysis and regulation, client segmentation, cross-selling, sales and marketing campaign management, and credit eligibility evaluation.

Utilities, feedback, and energy: The industry provides services such as power usage analytics, earthquake data processing, carbon emissions and trading, customer-specific pricing, smart grid management, and energy supply and demand optimization.

Cybersecurity: Machine learning algorithms can be used to detect unauthorized access attempts to a network, systems, or applications, and alert security teams to potential security breaches.

Media: Machine learning algorithms can be used to analyze vast amounts of text data, such as social media posts, reviews, and customer feedback, in order to identify whether the content is positive, negative, or neutral.

It should not be forgotten that; Analyzing and experiencing your business and customer experiences with the data you have will open different doors for you. Innovation, efficiency, integrity and corporate learning have become inevitable for companies today (Chernozhukov et al., 2017).

2.1.3 Predictions, Not Cause-effect Relationships

Machine learning is most basically used to make predictions. It helps us to have knowledge about the future by making predictions. Problems faced by the business world (Shashank et al., 2019):

- Making suggestions. Getting to know customers and finding specific recommendations for them.
- Making predictions, especially for employees, like calculating credit risk.

Many of the above items have common features. For example, more than one variable is needed to make the right decision. This means that there will be a huge amount of data. It is important to test the accuracy of the predictions. This property must have the data we have. Like which product customers clicked on, which product customers liked, how much the site has benefited from the opportunities it offers. After this process, to make a suggestion or prediction. This estimation needs to be accurate (Patil et al, 2011).

Machine learning differs from other methods. The issue is more concerned with cause and effect relationships than with statistical methods. It won't make much difference when the environment is changed. Instead of doing this, you spend time estimating and foresight. We need to model the right environment so that we can make the right decision. For example: should we take an umbrella with us when we leave the house? Shall we not? Let's say I'm stuck in two situations. To decide, we must first predict the weather. But our knowledge is limited to predict this. We must observe the sky and watch how the weather collects precipitation. It is limited but very helpful.

The same is true for machine learning. Personal information and estimates can increase the profit percentage, but they don't tell us why people like the products they love. Nor does it tell you how to influence their tastes and guide change in the right direction. In the results of these situations, the value of machine learning is better understood (Brown, et al., 2012).

2.1.4 Separating the Signal from the Noise

We have mentioned above the benefits of machine learning. After that, let's learn a little about its structure. Let's describe what constitutes the structure of machine learning and how it works in terms of the subject. Before explaining these terms, let's show some of them in the following list (Yang, et al., 2016):

a) Extracting an attribute

b) Regularize

c) Cross validation

1.5.1. Features Extraction

The variables we need when creating classification or similar models are called attributes. As a model can be created with all data, some attributes can be added. The newly added feature may be affected by other features. For example, obtaining the birthday from other characteristics of the person can be an example of this. As a result, increasing or decreasing the number of features affects the modeling's success in a different way.

To give another example, face recognition can be shown. The features here are related to the person in the photo. The characteristic features on this person's face form our characteristics. Features such as eye type, nose type, and skin color are among them. These features can be removed or added. This process is very important for the model to give the best results. Because after that, the process of applying algorithms will begin. With these features, the person is tried to be recognized (Bernard and Heutte L 2012).

There are many methods for feature extraction. Among these, help is taken from the programs used in this process. Many of these programs have feature extraction algorithms. In this way, it is possible to find out how much the properties affect the class property. Or, it can be found out how important the features are for the model by writing code in programs that work with code. As mentioned before, if classification algorithms are to be used, the features should be well studied. It should be found out which one is effective against the class feature or not. Thus, there may be a chance to remove features with little effect. It is removed and the model is rebuilt and the results are reviewed. It

would be beneficial if the accuracy rates increased positively, that is, if it made more accurate predictions.

Clustering is a slightly more complex issue than classification. Methods such as classification and clustering are also used to organize data. Unsupervised learning is the name given to this sort of machine learning. There is no measured phenomenon as the target of estimates (Hussein, 2018).

2.1.5 Regularization

Generalization in machine learning refers to how successfully the model's learned concepts are applied to scenarios that were not experienced during training. The goal of most machine learning models is to generalize from training data in order to create accurate predictions for data that will not occur in the future.data Overfitting occurs when models learn details and noise from training data too well, but generalization is weak, resulting in poor data testing performance. When the dataset is too tiny in comparison to the amount of model parameters that must be trained, this is a very prevalent problem. This issue is especially acute in deep neural networks, where millions of parameters are not unusual (Zou & Hastie, 2005).

Regularization is a key part of avoiding over-conformity. Additionally, various normalization techniques, such as driving some parameters to zero, can be utilized to lower modeling capacity while preserving accuracy. In a mobile setting when CPU power is restricted, this may be beneficial to reduce the size of the model or lessen the cost of evaluation...

The most common normalization techniques used in industry today:

- Data set augmentation

- Early stopping
- Weight penalty

Dataset Augmentation: If the learning algorithm processes more training data, a hyper-incompatible model such as neural network or any other model type may outperform. While the size of an existing dataset may be constrained, there are very simple methods for creating synthetic data for various machine learning tasks. Image techniques that are commonly used include a few pixel translation, rotation, and scaling. It is frequently convenient to incorporate random negatives into categorization problems.

There is no universal recipe for creating synthetic data, because it varies widely from one difficulty to the next. The main approach is to expand the dataset as much as possible by using methods that mirror real-world changes. In practice, regardless of architecture, having a better dataset dramatically improves model quality...

Early stopping: Early intervention challenges that interrupt the training procedure when the model's performance in the validation set deteriorates. The validity set is a collection of samples that are never used in slope descent However, they are not included in the test set. The validation samples should be typical of the future test samples. The early stop essentially determines the epoch/step count of the hyper parameter.

Both training and test error should decrease as the model sees more data and learns patterns and correlations. The model may experience excessive speech and learning noise on the provided training set after adequate time on the training data. In this situation, as the training error decreases, the test error (the better the generalization) increases. Stopping early is about finding the ideal moment with the least amount of test mistake.

In practice, instead of pausing, people are usually set up to record at regular intervals and choose the best candidate when the data have been gathered (Radford, & Chintala, 2016).

Weight penalty L1 and L2: Weight penalty is a frequent method of normalization that is used to train different sorts of models. It is heavily reliant on the premise that a small weighted model is simpler than a large weighted mesh.

Unless there are huge gradients to counter, penalties strive to keep weights modest or non-existent (zero); this makes the models more interpretable. Weight penalties are sometimes known as "weight drop" in the literature since they force weights to zero.

2.1.7 Cross Validation

The K-times cross-validation method is used to ensure that the classifier models' classification results on a data set are consistent.

The parameter k must be determined before using the procedure. The K parameter determines how many portions of the dataset will be separated. As shown in Figure 3, k classification operations are carried out, with one of the divided parts reserved for testing and the remaining k-1 for training the classifier. The average of the classification results received after K stages yields the general classification result. A cross validation process with K parameter set to 10 is shown below. In the data set, which is divided into 10 parts, 9 parts are used as the training set and the remaining part is used as the test set, and this process is carried out in 10 steps and by taking a different test set in each step (Kohavi, R., 1995).

Table 2.1: K-times cross-validation method

	Part 1	Part 1	Part 1	Part 1	Part 1	Part 1	Part 1	Part 1	Part 1	Part 1
Step 1	test	training	training	training	training	training	training	training	training	training
Step 2	training	test	training	training	training	training	training	training	training	training
Step 3	training	training	test	training	training	training	training	training	training	training
Step 4	training	training	training	test	training	training	training	training	training	training
Step 5	training	training	training	training	test	training	training	training	training	training
Step 6	training	training	training	training	training	test	training	training	training	training
Step 7	training	training	training	training	training	training	test	training	training	training
Step 8	training	training	training	training	training	training	training	test	training	training
Step 9	training	training	training	training	training	training	training	training	test	training
Step 10	training	training	training	training	training	training	training	training	training	test

Let's say we collect data about 1000 customers and create a prediction model with this collected data. After creating the model, let's measure how accurately our model predicts.

Here is an accuracy rate. This ratio will be very useful for us to be able to make comparisons in the future. To evaluate this ratio, we divide the data set we have into training set (900) and test set (100). Considering the opposite, if the model recognizes the test set, there will be a memorization.

2.2 Problems to be Avoided in Machine Learning

When we look at the fields where machine learning is used, there are problems that can be handled without realizing or knowingly.

Let's talk about some of them. Any model can be confused with cause and effect relationship. Determining cause and effect relationship cannot be a priority in estimation problems. Instead, efforts should be made to optimize the decisions found. Thanks to the optimization studies, the predictions will be more successful. As long as the two concepts do not mix with each other.

It is important to distinguish between “out of samples” and “out of context”. A built model's success outside of the sample indicates that if we collect more data points from the same context, the model can properly predict their outcome. However, there is no guarantee that the model will work in a different setting. For example, an e-commerce site may use a database of online purchases to create a model that aids in the acquisition of new clients. Even though the product range is identical, the same model may not operate in physical stores.

It may be tempting to believe that the vast amount of data accessible will solve the problem, but this is not the case. It should be highlighted that the power of these algorithms is derived by comparing new events to massive databases containing comparable situations they have encountered previously (Hastie & Friedman, 2009).

When you try to apply a model in a different environment, the information in the database will be different from what you encountered. This problem does not have a straightforward answer. An out-of-context model may be preferable to no model at all, but the model's constraints must be taken into account.

While some parts of the process of training machine learning models may seem automatic, figuring out when the model will succeed requires significant human judgment. Moreover, considerable critical thinking is required to ensure that the security measures involved in the process, such as regularization and cross-validation, are used correctly.

Despite all this, only humans' decision making as an alternative to machine learning has its own biases and errors. With the right mix of technical skill and human judgment, machine learning can be a new and useful tool for decision makers who want to transform large amounts of data into added value (Rajkumar et al., 2019).

2.3 Knowledge Discovery in Databases (KDD)

To use the information in the database, we must first find it and extract it. The most important method is known as Knowledge Discovery in Databases (KDD).

In the past, many methods have been used to find the useful pattern in the database. This work has been given different names. For example: machine learning, information discovery, information harvesting, pattern processing, and data mining. Those dealing with statistics and analyzing data mostly used artificial learning methods. In addition, it has become popular among those dealing with databases. The discovery of information in the database was first introduced at a meeting of a research group in 1989 (Krizhevsky et al., 2012).

As a matter of fact, the Discovery of Knowledge in Databases refers to the information in the data at hand, that is, the process of extracting the data that will make our job easier. Various methods are used in this process. But these many methods actually constitute a step of this work. It is possible to call artificial learning as a step in this process. Likewise, they act as steps in other methods. Discovery of Knowledge in Databases is a field where many fields such as pattern recognition, data mining, artificial intelligence, expert systems merge and develop. The aim is to extract meaningful patterns from scattered big data and allow predictions to be made. And this information is necessary to make predictions.

In this process, machine learning, which functions as a step in the discovery of information in databases, uses methods such as statistical and machine learning to find the pattern in the data.

In addition to the above narrative, a question may come to mind, such as what is the difference between KDD in machine learning, data mining and other related fields. The answer to these areas as it provide some methods in the machine learning step of KDD. Whereas, KDD focuses on all steps of the information coming out of the data, from where to store and access the data, how to scale algorithms to large volumes of data sets, how to visualize and interpret the results obtained, to model the interaction of human with machine in the most useful way.

KDD functions within a much broader discipline than if an area such as artificial intelligence is of interest. Artificial intelligence, data mining, pattern recognition, statistics, machine learning and other techniques contribute to KDD, bringing it to a wider position in its field (Brachman, 1993).

2.3.1 The Discovery Process of Knowledge in the Database

The stages of KDD begin with the selection of data in the database. This selected data is then preprocessed. After these processes, machine learning techniques (algorithms) are used to extract the pattern. And it is in the form of interpretation of the resulting pattern.

In the stages of KDD, the machine learning step gives us an algorithmic idea about which pattern of the data is taken and which is not. We learn how to take the pattern as new information from the buried, hidden data in a whole data set in the stages of KDD (Sutton & Barto (2018).

The stages of KDD should have self-renewing features. Brachman and Anand drew attention to the interactive nature of the stages (Brachman & Anand, 1993). The important stages of KDD are shown below:

Step 1: To ensure the design and development of the application area. And to determine the purpose of these stages.

Step 2: Generating a purposeful dataset: finding the dataset from which valuable information will be extracted, or focusing on subsets of datasets. This step is important for predictions. Choosing a data set is very important.

Step 3: Pre-processing of the data: collecting the necessary information for modeling, removing the noise if there is any, finding solutions for the missing data if there is any, and similar pre-processing.

Step 4: Reducing the data: For the purpose, the most qualified features of the data set should be found. And the least representative features can be removed if desired. If a change is made in the size of the data set at this step, it will also affect the prediction success.

Step 5: Equalizing the purposes of the KDD stages with the artificial learning techniques: at this stage, techniques such as classifying, clustering, and finding association are applied.

Step 6: Selection of machine learning algorithms: This is the step where models, analyzes and hypotheses will be selected. Selected algorithms and techniques are used to extract the pattern. Before that, it should be chosen which algorithm to apply to the data set. This is the most important issue at this stage. While unsuitable algorithms will decrease the success of the model, appropriate algorithms will increase the correct prediction rate.

Step 7: Machine learning: The pattern of interest is extracted from the appropriately extracted data set. For this, classification, clustering, trees and similar methods are used.

Step 8: Interpretation of the extracted pattern: In the next rounds, the pattern is interpreted by machine learning with the possibility of returning to one of the steps 1 to 7.

Step 9: Consolidation of the information found from the dataset: this information can be collected for further study or stored as a document. It can also be used as a report. This also gives us the chance to check the accuracy of the information. It allows us to understand how the differences occur while performing the controls.

2.3.2 The Machine Learning Step of the Knowledge Discovery Process in Databases

The purposes of knowledge discovery vary according to the intended use of the system. We can divide the goals into two (Pan & Yang, 2010):

- Verification
- Discovery

Validation is limited to verifying users' predictions. Exploring can find new patterns independently of the system.

It will be used to predict the behavior of some living things as the purpose of exploration in future matters.

Machine learning requires dressing a model on the data studied and recognizing the patterns in the data set.

Model dressing undertakes the task of extracting information. Human judgment is needed at this stage to understand how the model is, what its features are, and whether it is fully interactive. Two basic mathematical concepts are used in model dressing. The first is statistical and the other is logic. The approach with statistical in the model allows for the undetermined effect, while the logical approach allows for the specific effect.

Many Machine learning methods are based on previously tried and tested techniques from data mining, pattern recognition and statistics: Classification, clustering, etc. (Al-Mallah et al., 2019).

2.4 Data Warehouses

A data warehouse is a special database that is synonymous with machine learning and provides the data in which the machine learning process takes place. In terms of the meaning of the term, they are structures that collect data from various sources and mostly in different structures and aim to use all of them under the same roof. We can say that such a structure combines data from more than one database (Krizhevsky et al., 2012).

A related area developed in the database is the data warehouse, which is used to collect the data of the transactions and provide their analysis. Clustering data for KDD is still

one of the tasks of data warehouses. There are two main benefits to this business. And thanks to these benefits, it plays a big role in the success of modeling.

- Data cleaning
- Data access

2.4.1. Data Cleaning

Companies should always keep their records in big data and databases under a certain discipline. It is extremely difficult to use data that is not in order with any method. For this reason, although there are many solutions to do this job, the precautions to be taken before the design will be the most effective.

2.4.2 Data Access

In general, clean and properly working methods should be used to access data and to access data that is difficult to access. First of all, companies and individuals should not have such a problem. The next step is to answer the question “What will be done with this data?” The answer comes out as the stages of KDD.

One of the most often used ways of analyzing data warehouses is Online Analytical Processing (OLAP). The possibilities that OLAP offers us are slightly superior to SQL. We can list it as follows: it can make calculations in many dimensions, it allows interactive data analysis, and it analyzes multidimensional data. But KDD's tools aim to automate the process as much as possible.

2.5 ARTIFICIAL LEARNING IN MEDICINE

2.5.1 Machine Learning Applications in Medicine

The use of data in the medical sector search for answers to the unknowns of many health problems has come to the fore with the beginning of keeping the data in health institutions in the electronic environment and the ease of access to the data (Al-Mallah et al., 2019).

When the causes, risk factors, treatment methods, drugs used and their effects, patient laboratory and demographic data come together, it becomes difficult to obtain useful information from them and many valuable information is lost in this big data. Vital information from this accumulated big data can only be obtained by using the capabilities of data mining methods and/or machine learning (Guyon & Elisseeff, 2003).

The first machine learning application in medicine was made by John Snow in 1854 with paper and pencil. In the cholera epidemic that started in London, he made the first clustering study by marking the locations of the deceased on a map, so he realized that the deaths were concentrated in certain regions and ensured the elimination of the epidemic caused by the water pumps (Friedman, 2001).

By applying machine learning techniques to the data of breast cancer patients, Delen et al. developed a model that predicts the probability of death and survival of patients (Delen, & Kadam, 2005).

In the article written by Patil et al., 2011, classification techniques were applied on the data of 180 burn patients between 2002 and 2006, and a model capable of predicting the survival of patients at high rates was created (Patil et al. 2011).

In a study by Kumar et al. 2011, intelligent medical decision support systems were developed to help physicians in diabetes, hepatitis and heart diseases by extracting

patterns from the data collected, and the effectiveness of some algorithms were compared (Kumar, Sathyadevi, & Sivanesh, 2011) .

Various achievements have been achieved by applying artificial learning methods in terms of administrative activities of health care institutions.

In England George Hospital, it was determined that thirty-nine percent of patients who died after leaving the intensive care unit could be kept in the intensive care unit for 48 hours, thus eliminating the risk of death (Han, & Kamber, 2006).

At the San Francisco Heart Institute, artificial learning studies were initiated in order to strengthen the financial structure, increase the quality of patient care, shorten the hospital stay of patients, and increase employee performance, and the resulting models enabled the realization of the objectives (Han, & Kamber, 2006).

Boehringer Ingelheim Italy is one of the major pharmaceutical companies in the world. He started a machine learning study to produce differentiated sales policies for pharmacies and to identify their best customers. Thanks to the resulting classification model, gains such as monitoring profitable customers, creating the most profitable pharmacies target list, managing customer relations correctly and evaluating the effectiveness of marketing activities have been achieved (Han, & Kamber, 2006).

2.6 HEART AND VASCULAR DISEASES

Heart and vascular illnesses are a set of ailments that affect the heart and blood vessels. They are sometimes referred to as cardiovascular diseases. Heart disease, stroke, and hypertension are examples of these disorders.

2.6.1 Definition and Importance of Cardiovascular Diseases

It is a general term that covers diseases that can adversely affect the heart and blood vessel system in the body. Heart diseases are the cause of most deaths. One of the heart's jobs is to deliver oxygen to tissues and organs. Likewise, the inner surface of a normal vessel has a smooth structure. Apart from this, people who eat poorly, smoke, have high blood pressure, and use alcohol, have deterioration in the vessels as time passes (Rosengren & Dallongeville, 2009).

Depending on time, plaque-like layers are formed in places that are deteriorated in this way. Some of the biggest harms of these are that they block blood flow and change the elasticity of the vessels. It is possible to see the same factors in older people. It is very important to prevent this sensitivity. There are various ways to maintain vascular health. These are the measures that people can take by themselves, such as keeping the cholesterol level at a certain level, avoiding high blood sugar, keeping the blood pressure at a certain level, and keeping the uric acid level under control. In order to take these precautions, people need to go to the doctor at regular intervals and give blood and have the values mentioned above measured. According to the doctors interpreting the results, there may be various treatments (Al-Mallah et al., 2019).

2.6.2 Types of Cardiovascular Diseases

Cardiovascular diseases as described by WHO;

The World Health Organization (WHO) defines cardiovascular diseases (CVDs) as a set of heart and blood vessel maladies that include coronary heart disease, cerebrovascular

disease, and other diseases. In 2016, CVDs were the leading cause of death worldwide, accounting for 17.9 million deaths per year, or 31% of all global deaths (WHO, 2016).

- Coronary artery disease is a condition that affects the blood vessels that nourish the heart muscle and leads to coronary heart disease (heart attack - coronary heart disease).
- Cerebrovascular disease, sometimes known as a stroke, is a condition of the blood arteries that supply the brain.
- Hypertension; blood pressure is greater than normal (high blood pressure).
- Peripheral arterial disease is a condition affecting the blood vessels that supply the arms and legs.
- Rheumatic heart disease is caused by streptococcal bacteria and causes damage to the heart muscle and valves.
- Congenital heart illness, also known as congenital heart defect.
- Deep vein thrombosis (DVT) and pulmonary embolism (DVT and pulmonary embolism); DVT is the creation of a blood clot in our body's deep vein (vein). It commonly affects the lower leg or thigh. Blood clots in the leg veins become dislodged and move to the lungs or heart.

A heart attack or stroke are both acute (rapid onset and/or short-term) disorders that occur when blood flow to the heart or brain is obstructed by a blood clot.

The accumulation of fatty deposits on the inner walls of blood arteries that supply the heart and brain is the most common cause of this. A stroke can happen as a result of bleeding or a clot in a vein in the brain (Rothwell, et al., 2004).

2.6.3 Significant Risk Factors for Cardiovascular Disease

When cardiovascular disease deaths are examined, it is seen that patients who have not been diagnosed before or have no symptoms account for more than half of the deaths.

A normal vessel carries oxygen to the organs and tissues of the human body and nourishes them. Beginning from infancy, there are more changes in the structure of the vessel in older people. Many of these changes form layers called plaques in the vessel wall. These layers can cause cardiovascular disease in humans over time. Changes we make in our diet, that is, consuming less unhealthy food, can be one of the solutions. If we choose to live in a healthy way, we will have a better old age.

For this reason, it is of great importance to determine the risk factors for cardiovascular diseases. Reducing cardiovascular disease risk factors reduces both cardiovascular diseases and therefore deaths (Rothwell, et al., 2004).

2.6.4 Unchangeable Risk Factors

a. Age: Cardiovascular diseases increase with advancing age. Age 45 and over in men and 55 years and older in women are considered risky (Thom, et al., 2006).
b. Family history: Diagnosis of cardiovascular diseases in first-degree relatives before the age of 65 in women and 55 in men is considered to be a predictor of future cardiovascular diseases. However, there is no currently accepted screening test for cardiovascular disease susceptibility (Bernard, Heutte, & Adam, 2010).
c. Gender: In young men, cardiovascular diseases and mortality rates are 4-5 times higher than in women. Men under the age of 65 are twice as likely to have a stroke as women (Rosengren, & Dallongeville, 2009).

4.6.5 Changeable Risk Factors

a. Smoking: One of the main risk factors for cardiovascular diseases is smoking, and the continuation of smoking in people with cardiovascular diseases increases mortality rates (Biau & Scornet, 2016).

b. Hypertension (high blood pressure): It is a very common health problem worldwide. A blood pressure greater than or equal to 140/90 is indicative of hypertension. The most common diseases caused by hypertension are cardiovascular diseases and stroke. Treatment of hypertension reduces the risk of cardiovascular disease (Blaser & Fryzlewicz, (2016).

c. Diabetes (diabetes): One of the main risk factors for cardiovascular diseases is diabetes. The most important cause of death in diabetes diseases is cardiovascular diseases (Biau & Scornet, 2016).

d. Hyperlipidemia (high blood lipids): It is an increase in at least one of the blood lipids. There are two types of blood fat, triglycerides and cholesterol. As a result of the correlation between hyperlipidemia and cardiovascular diseases in all studies, hyperlipidemia, which is accepted as one of the important and correctable cardiovascular disease factors, can be controlled with diet, exercise and drug therapy (Blaser & Fryzlewicz, (2016).

e. Obesity: When the fat mass is higher than the lean body mass, it is called obesity. Body Mass Index (BMI) is the most well-known method used for obesity detection. A BMI above 30 increases the risk of cardiovascular disease. BMI is calculated as follows (Blaser & Fryzlewicz, (2016).

$$BMI = \text{Weight (kg)} / \text{Height}^2 \text{ (m)}$$

It is noteworthy that several of the main risk factors mentioned above are present simultaneously in cardiovascular patients. The probability of three or more of the aforementioned risk factors coexisting by chance in a patient appears to be four times higher than the probability when compared with patients diagnosed with cardiovascular disease with three or more risk factors (Biau & Scornet, 2016).

2.6.6 Diagnostic Methods

The following methods are used to diagnose cardiovascular diseases in patients at risk for cardiovascular diseases (Blaser & Fryzlewicz, (2016) :

a. Electrocardiography (ECG)
b. Echocardiography (ECO)
c. Exercise stress tests – Treadmill
d. Myocardial perfusion scintigraphy
e. Positron emitting tomography (PET)
f. Coronary angiography CT (Multisection CT Angiography) (Multislice Cardiac CT)
g. Cardiac MRI

2.6.7 Risk Calculation Methods

Predicting the risk of cardiovascular diseases is very important in order to eliminate risk factors, to take precautions and to provide treatment. The fact that risk factors coexist in patients diagnosed with cardiovascular diseases guides the estimation.

Today, some risk calculation methods have been developed to determine cardiovascular diseases. These methods are described below (Chernozhukov, et al., 2017).

Framingham risk calculation system: It is based on a follow-up study conducted in the town of Framingham, Massachusetts, USA.

It was started in 1948 with 5209 adults living in the town, currently the third generation is being followed. The American Heart Association (ABP) has developed a risk assessment system based on these data. The risk factors for cardiovascular diseases within 10 years are calculated in the developed system.

SCORE study: Since the Framingham study is a regional practice, the European Society of Cardiology conducted the SCORE study using data obtained from 205178 participants. In this study, the ten-year risk of development of cardiovascular diseases is calculated. In the risk calculation made with the risk factors found, one of four risk results as low, medium, high and very high is reached.

A recent survey of Merla M. and Beena T.L. A. shows different combinations of Machine learning used for the problem of heart disease diagnosis. Many researcher used combination of Machine Learning algorithms and ensemble methods. A general result of 90% for ANN with genetic, ANN with Sparse autoencoder and Random Forest combined with Linear Method produce 89% of accuracy. The performance evaluation metrics were based on accuracy, precision, recall and f-score. (Merla M. and Beena T.L. A. 2020).

A sample of different studies on using Machine Learning for heart disease failure are shown in Table 2.

To predict cardiac illness, Akhtar and Jinna employed several ML algorithms such as naive Bayes, K-Nearest Neighbor (KNN), Decision tree, and Artificial Neural Network (ANN). According to their experiment, Nave Bayes has the greatest accuracy of 88% (Akhtar and Jinna, 2021).

Akanksha Kumari et al. used seven different machine learning to diagnose heart disease. The AdaBoost and voting ensemble approaches were employed to increase algorithm accuracy. The obtained findings demonstrated that the use of the AdaBoost Algorithm enhanced the accuracy of The Decision Tree classifier coming 80%. (Akanksha Kumari et al., 2021).

Agrawal et al. proposed different ML algorithms to predict person's heart failure in the coming ten years. An ensemble approach combined with 10 different classification ML techniques. This model obtained 85.2% accuracy and 87.5% test recall, and it was found that persons with high blood pressure had a heart attack (Yash Goyal et al., 2021).

A recent work of Karthik et al., used Machine Learning algorithms (like SVM, Decision trees, and Random Forest) as a classification technique. Accuracy used as performance metrics for their study. In their study, Random Forest ML algorithms has more accuracy of 88.26% (Karthik et al., 2023).

Table 2.2 Different studies for prediction of Heart Failure Using Machine Learning

Year	Problem statement	Dataset	Approach-ML algorithms	Performance Accuracy	References
2016	Machine Learning Algorithms for Heart Disease Prediction	UCI	Naïve Bayes Classifier, and - Decision tree	The decision tree outperforms the naive Bayes classifier in terms of accuracy.	Nikhar S., Karandikar A.M.
2018	Heart disease prediction using machine learning techniques	UCI Cleveland dataset (14 features)	SVM, KNN, Decision Tree, Random Forest, and Ensemble ML	SVM has more accuracy than other techniques.	Ramalingam V.V.
2019	Prediction of Heart Disease	UCI	Decision Tree, KNN,	Decision tree (86.60% Accuracy)	Golande A.

	Using Effective Machine Learning Techniques		K-mean clustering and Adaboost		
2019	Hybrid Machine Learning Techniques for Effective Heart Disease Prediction	UCI Machine learning repository	Language Model, SVM, Random Forest, Nave Bayes, Neural Networks, Decision TreeKNN and HRFLM	SVM has more accuracy than other techniques.	Mohan S.
2021	A Novel Approach to Heart Disease Prediction Using Machine Learning Algorithms	UCI Machine learning repository	Decision tree, ANN, SVM, Random Forest, Naïve Bayes, K-means, and AdaBoost	AdaBoost Algorithm enhanced the accuracy coming 80%.	Kumari A. et al.
2021	Heart Disease Prediction	UCI	(KNN), Decision tree, and Artificial Neural Network (ANN)	Nave Bayes has the greatest accuracy of 88%.	Akhtar N. and Jinna F.
2023	Using Supervised Machine Learning Techniques, an Automated System Predicts Heart Disease	UCI	SVM, Decision trees, and Random Forest	Random Forest ML algorithms has more accuracy of 88.26%.	Karthik et al.

2.7 DATASET

Data is the gathering of observation and measurements so that it can be structured, recorded and analyzed easily. A set of one or more information. Data is usually obtained through research, observation, experiment, census, measurement.

Raw data can be seen as an important raw material for efficient knowledge generation and analysis. For example, data generated through surveys such as (election data), when a vote is made (election results data), when a record is made (birth records data), when something is purchased (online sales records, etc.). Data can also be generated by mobile phones, the Internet, satellite (such as GPS data), and many other technologies Han, & Kamber, 2006).

In our daily life, we often find data organized in tables. The content of a single table is expressed as a data set. Analyzing the data set and producing new information-visual studies from it; decision making is important for the policy making process.

A dataset is a collection of related data that can be controlled individually, collectively, or as a full entity. The dataset is arranged using a data structure. A dataset in a database, for example, could contain a collection of business data (names, salaries, contact information, sales figures, etc.). The database itself can be regarded as a dataset, with data of a particular type of information such as sales data of a particular corporate department (Aggarwal, 2015).

Even trivial objects have lots of data associated with them. The following are the main ones:

Qualitative data: It is data that does not have values obtained with units or measures, and carries structural and institutional characteristics. The gender of the people and the color

of the hair of the people are qualitative characteristics. It also shows data that qualifies an experiment performed. There are two different groups.

Classifiable qualitative data: data with class-specific discrimination to show names, codes, and numbers. They are independent of each other (Wang, et al., 2018).

- Qualitative data that can be sorted: displaying values without quantity, values with a degree and digits are data of variables in which the order is important. Ranks in the army, student success can be given as examples.
- Quantitative data: They are data with intervals or proportions that show the countable and measurable properties of the experiments. There are two types of continuous quantitative data and discrete quantitative data. For example the number, size, price of golf balls, score on a test (Han, & Kamber, 2006)

Continuous quantitative data: It is quantitative data that takes the values of numbers with commas. Neck length as an example.

- Discrete quantitative data: It is quantitative data that takes values consisting of integers. Family population, like city population.

Categorical data: puts the data you define into a category (Kim, et al., 2019).

A dataset in a database, for example, could contain a collection of business data (names, salaries, contact information, sales figures, etc.). The database is a dataset that contains data for a certain type of information, such as sales statistics for a specific corporate department. The structure of the dataset is depicted in detail in Table 2.3

Table 2.3: Dataset structure.

Attributes

A	B	C	D
a	1	x	0
b	2	x	1
c	3	y	0

Records

2.8 Feature Selection

It is the process of identifying a subset of relevant features from a large set of features for a machine learning model. It is an important step in the modeling process as it can improve the performance of the model, reduce overfitting, and increase interpretability (Yu, & Zhang, 2019).

There are several methods for feature selection, including filter methods, wrapper methods, and embedded methods. Filter methods are based on feature ranking or scoring, and use statistical measures such as correlation or mutual information to rank features. Wrapper methods use a machine learning algorithm to evaluate the performance of a subset of features, and use this evaluation to select a subset of features. Embedded methods include feature selection as a part of the learning algorithm, such as regularization methods like Lasso or Ridge regression.

Recent studies have shown that ensemble feature selection methods, which combine multiple feature selection methods, can achieve better performance than individual methods (Kira and Rendell, 1992; Brown et al., 2012; Yang et al., 2016; Yu et al., 2019). Additionally, the integration of domain knowledge into the feature selection process can improve the interpretability and generalization of the model (Guyon et al., 2002; Domingos and Pazzani, 1997).

In high-dimensional datasets, feature selection becomes even more crucial, as the curse of dimensionality can lead to overfitting and poor generalization of the model. In these cases, methods such as Lasso and Ridge regression, which incorporate regularization, have been shown to be effective (Tibshirani, 1996; Zou and Hastie, 2005; Huang et al., 2018).

2.9 Generative Adversarial Networks (GANs)

It is a type of machine learning method that generates new data samples that are similar to those in a given dataset. GANs are made up of two basic parts: a generator network and a discriminator network. The generator network generates new data samples, whilst the discriminator network attempts to distinguish between the generated samples and the genuine samples in the given dataset. The two networks are trained concurrently, with the generator attempting to generate samples that will deceive the discriminator and the discriminator attempting to properly identify the generated samples.

Recent studies have shown that GANs can be used to generate high-quality images (Radford et al., 2016; Karras et al., 2017), videos (Vondrick et al., 2016), and even speech (Donahue et al., 2018). Additionally, GANs have been used to solve various tasks such as image-to-image translation (Isola et al., 2016), text-to-image synthesis (Reed et al., 2016), and even 3D object generation (Wu et al., 2016).

One of the key challenges in training GANs is the stability of the training process and avoiding mode collapse. Mode collapse occurs when the generator produces samples that only represent a subset of the possible outputs. To address this, several methods have been proposed such as using Wasserstein distance (Arjovsky et al., 2017), Spectral

Normalization (Miyato et al., 2018) and Self-Attention (Zhang et al., 2018) as a way of stabilizing the training process.

Another area of research in GANs is on interpretability and controllability of the generated samples. For example, using GANs for data augmentation (Antoniou et al., 2017), generating samples with specific attributes (Choi et al., 2018), and controlling the style and content of generated images (Zhu et al., 2017). Additionally, there have been studies on using GANs for feature selection (Wang et al., 2019) and dimensionality reduction (Liu et al., 2018).

Overall, GANs have proven to be a powerful tool for generative tasks, however, there are still several open research questions, such as improving the sample diversity and stability of the training process, and understanding the underlying distribution of the generated samples.

2.10 Palestine Ministry of Health

According to Palestinian ministry of health data, the distribution of reported causes of cardiovascular disease death by demographic in 2021 as follows:

Table 2.4: Distribution of Reported Causes of Cardiovascular Diseases Death by Sex in Palestine - West Bank 2021

Cause of Death	Male	Female	Total	%
Hypertensive heart disease	163	210	373	11.7
Hypertensive renal disease	2	2	4	0.1
Hypertensive heart and renal disease	149	150	299	9.4
Acute myocardial infarction	586	378	964	30.1
Other acute Ischaemic heart Disease	15	17	32	1.0
Chronic ischaemic heart disease	366	303	669	20.9
Pulmonary embolism	6	4	10	3.0
Other forms of heart disease	8	14	22	0.7
Subarachnoid and intracerebral haemorrhage	30	20	50	1.6

Cerebral infarction	303	383	686	21.6
Other and sequelae of cerebrovascular diseases	35	28	63	2.0
Diseases of arteries, arterioles and capillaries	12	4	16	0.5
Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified	4	5	9	0.3
Total	1679	1518	3197	100.0

Table 2.5: Distribution of Reported Causes of Cardiovascular Diseases Deaths by Age Group, and Sex, Palestine - West Bank 2021

Age Group	Male	Female	Total	%
Less than 1 Year	0	0	0	0.0
(1- 4) Years	0	0	0	0.0
(5 - 19) Years	4	1	5	2.0
(20 - 29) Years	11	6	17	0.5
(30 - 39) Years	34	16	50	1.6
(40 - 49) Years	94	32	126	3.9
(50 - 59) Years	278	86	364	11.4
60 Years and above1,258 1,377 2,635 82.4%	1,258	1377	2635	82.4
Total	1,679	1518	3197	100.0

Total of death among population by Age Group and Sex, West Bank, Palestine 2021, (MOF, 2021) it was reported that 5574 death male and 4,559 death female, of which 1679 death male caused by cardiovascular diseases, and 1518 death female caused by cardiovascular diseases, or 30% for male and 33% for female of deaths in 2021 were caused by cardiovascular diseases.

The researcher concludes on these numbers, which constitute more than a third of the dead that come from the causes of heart and arteries diseases, which requires studying these causes and trying to employ machine learning in order to reduce this percentage, and most probably save the life of people.

2.11 Conclusions and Research Gab

In this chapter, the researcher reviewed some previous studies related to the topic of this thesis; it started with reviewing some works related to heart disease. After that, it reviewed previous studies about the use of machine learning specifically in heart disease, and the methods to employ machine learning to detect or predict heart disease. Finally, the researcher reviewed the use of machine learning in predicting human being from heart failure through the aid of reviewing the history of the patient records in the available database. Adding to this, as to the limit knowledge of the researcher there was no previous study found in the literature in the Arabic region in general and in Palestine in particular that uses machine learning in detecting the heart disease.

Chapter 3

Methodology

3.1 Introduction

This chapter describe the proposed methodology shown in figure 3.1. It begins by describing the dataset and the preprocessing steps. Then it goes through the data augmentation process, showing the results, finally it describes the Machine learning models used in this study, including the functions and matrices used for comparing between the models.

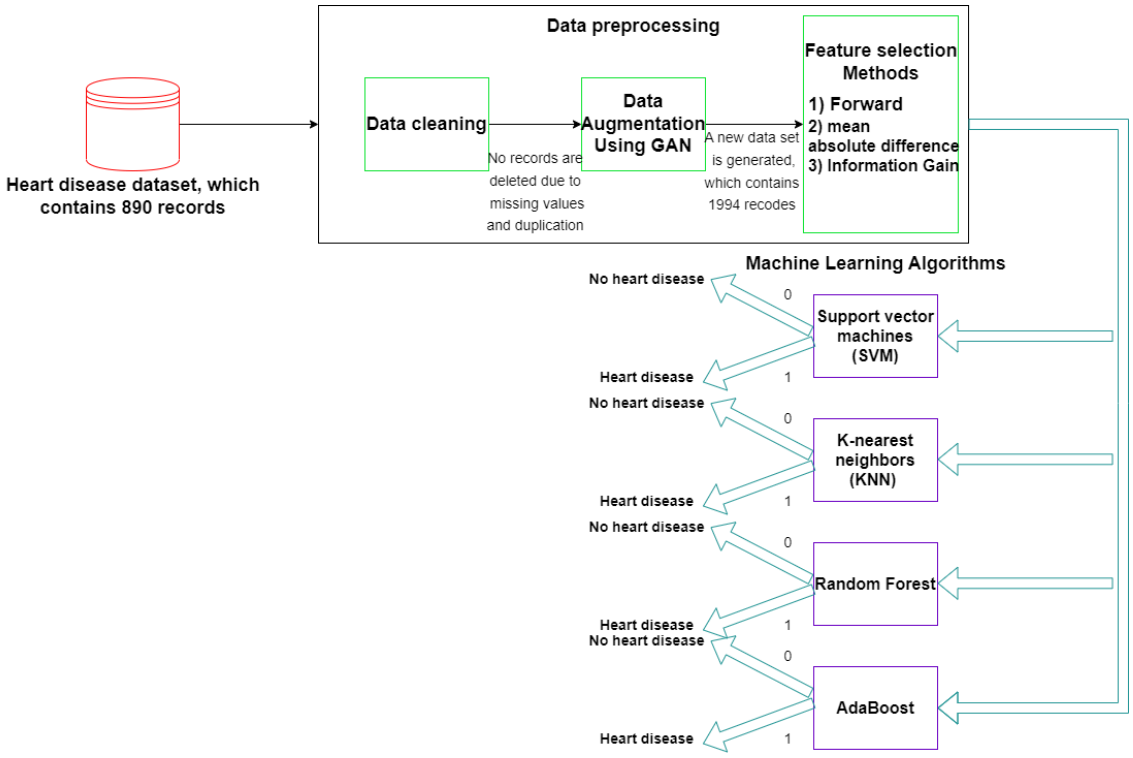


Figure 3.1: The proposed methodology

3.2 Dataset

Data was collected the year 2022, from a private hospital in Palestine called Al-Razi Hospital, in the city of Jenin, at the request of the researcher. This dataset contains information and examinations for 890 patients with 12 various attributes (11 predictors (features); 1 class(heart disease is the binary class label)) and they are as follows Age, Sex, Type of chest pain, Resting blood sugar, Serum cholesterol, Fasting blood sugar, Electrocardiogram result, Maximum heart rate achieved, ExerciseAngina, Oldpeak, ST_Slope, and the binary class label is Heart Disease, It has 2 values, zero means doesn't have heart disease, and one means the patient has a heart disease.

The attribute Age means the age of the patient in years and the range of years located between 28 and 76 as shown in figure 3.2.

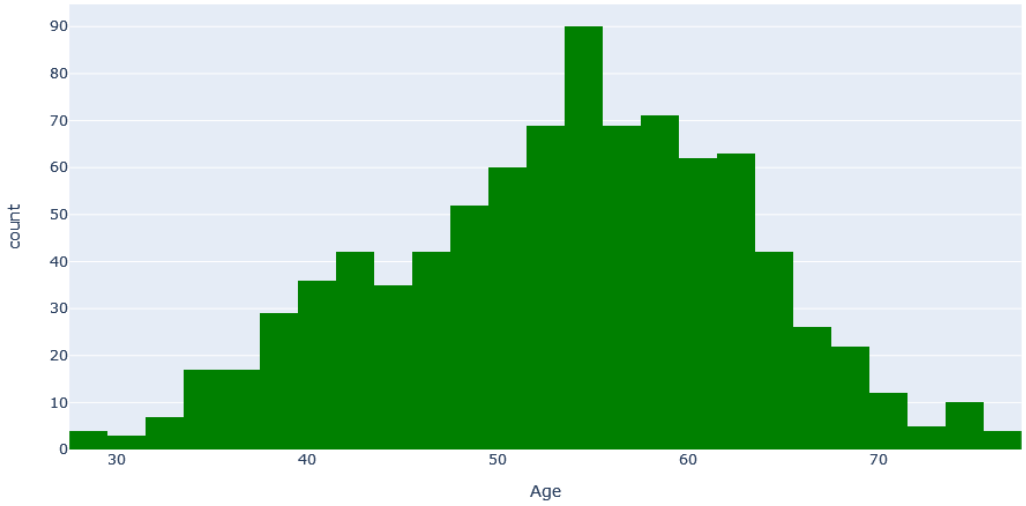


Figure 3.2: Age distribution

ChestPainType it's an attribute refers to the different types of pain or discomfort a person may feel in their chest. And in our dataset we have four types, which are Typical Angina

(TA), Atypical Angina (ATA), Non-Angina Pain (NAP), and Asymptomatic (ASY), and Figure 3.3 shows the number of records for each type.

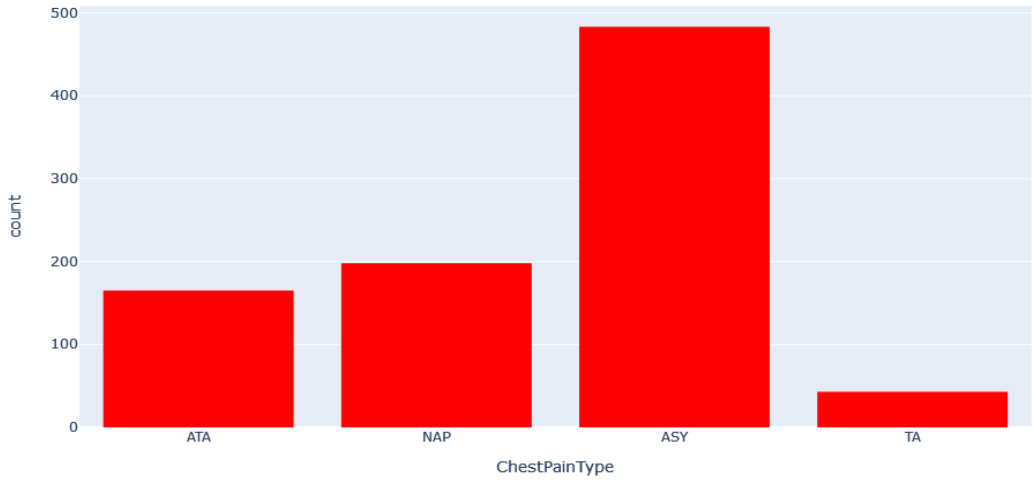


Figure 3.3: Chest pain types distribution

The sex attribute refers to how many male samples and how many female samples. And in our dataset, we have 705 patients of males and 184 of females. Figure 3.4 shows the distribution of males and females in the dataset.

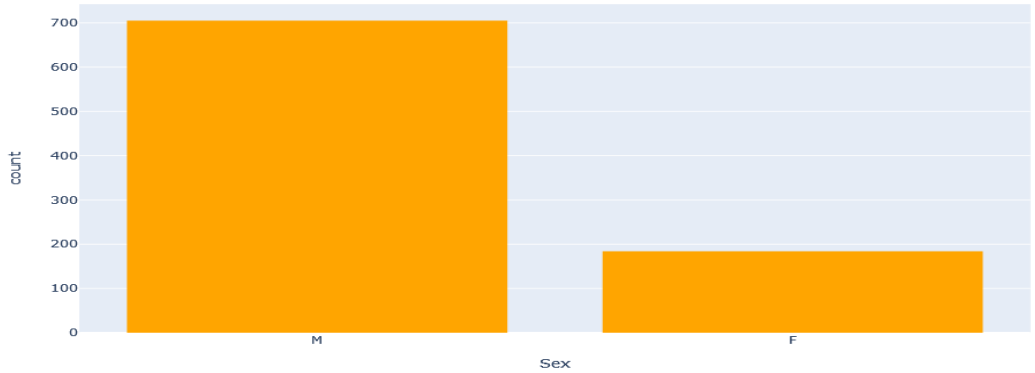


Figure 3.4: Gender distribution

FastingBS attribute which is a shortcut of fasting blood sugar, refers to medical term used to describe the level of glucose (sugar) in the blood after an individual has fasted for a

certain period of time, usually for at least 8 hours. Higher levels of fasting blood sugar may indicate prediabetes or diabetes, while lower levels may indicate hypoglycemia (low blood sugar). In our dataset we have 211 patients there results were bigger than 120 mg/dl. And 678 otherwise. Figure 3.5 represents Fasting Blood sugar distribution.

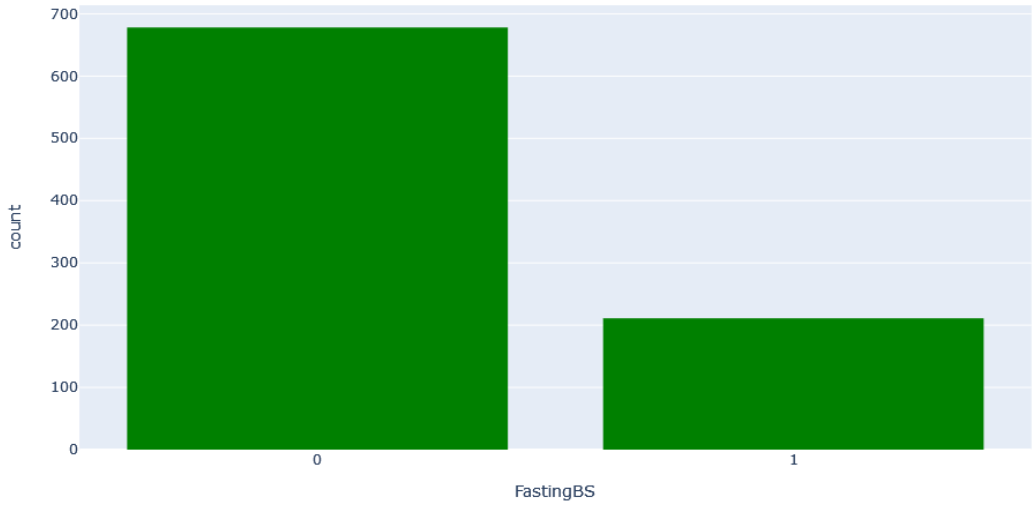


Figure 3.5: Fasting Blood sugar distribution

RestingECG, which is a shortcut of resting electrocardiogram, refers to a medical test that records the electrical activity of the heart while the patient is at rest. This test is non-invasive and involves placing small electrodes on the chest, arms, and legs, which detect the electrical impulses generated by the heart's natural pacemaker and the electrical activity of the heart muscles. And there are three results types of this test which they are as follows:-

- Normal: This refers to a normal ECG reading, with no evidence of significant abnormalities in the heart's electrical activity.
- ST: This refers to an abnormal ECG reading that shows ST-T wave abnormalities, which can include T wave inversions and/or ST elevation or depression of more

than 0.05 mV. These abnormalities may indicate a variety of heart conditions, including coronary artery disease, myocardial infarction, or inflammation of the heart muscle.

- LVH: This refers to an abnormal ECG reading that suggests the presence of left ventricular hypertrophy (LVH), which is a thickening of the heart muscle in the left ventricle. LVH can be a sign of various heart conditions, including high blood pressure, aortic valve stenosis, or hypertrophic cardiomyopathy.

In our dataset we have 534 patients in Normal, 176 in ST, and 179 are LVH. Figure 3.6 shows the RestingECG distribution.

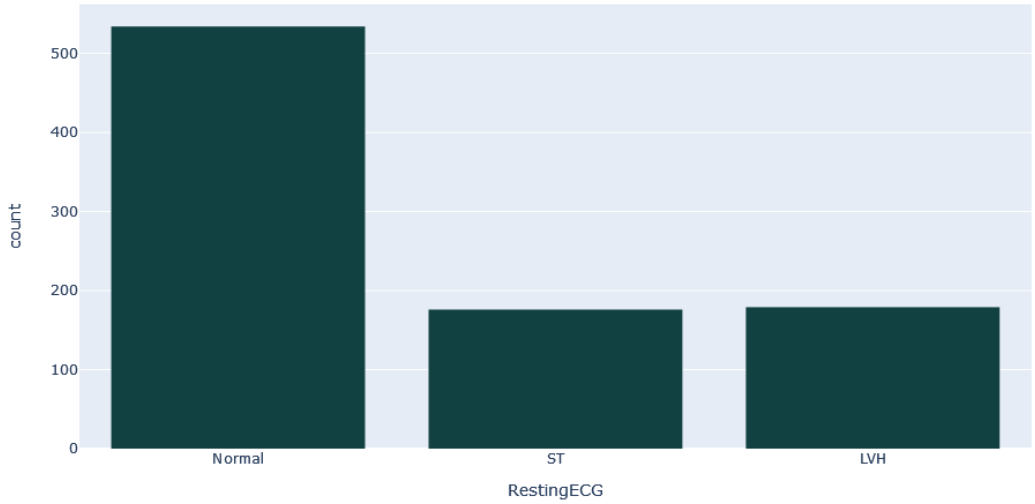


Figure 3.6: RestingECG distribution

ExerciseAngina is an attribute also known as angina pectoris, and it is a type of chest pain or discomfort that occurs when the heart muscle doesn't receive enough blood and oxygen during physical activity or exercise. It is typically caused by a narrowing or blockage of the coronary arteries, which supply blood to the heart muscle. Exercise angina is a common symptom of coronary artery disease and may require medical treatment to

prevent further heart damage or complications. In our dataset we have 363 patients have ExerciseAngina, and 526 patients doesn't have ExerciseAngina, as shown in Figure 3.7.

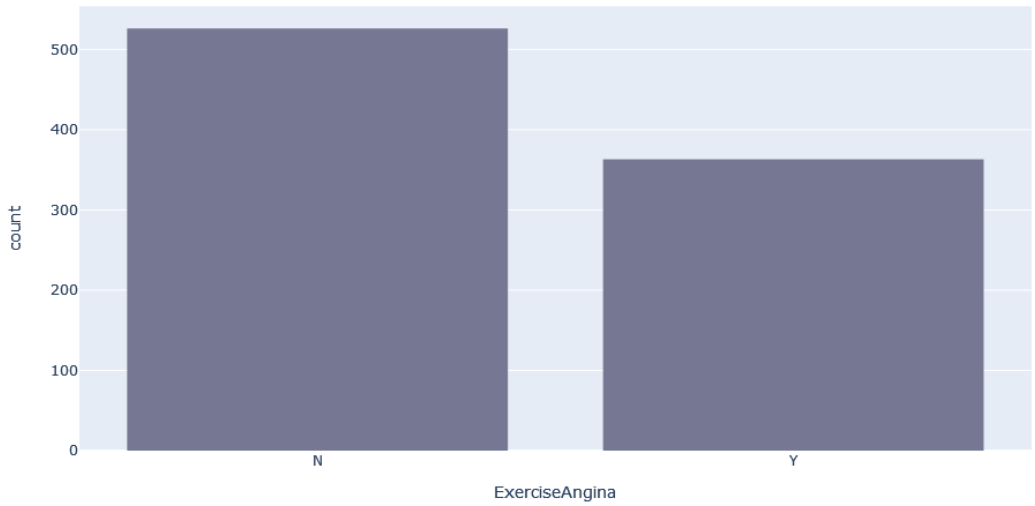


Figure 3.7: ExerciseAngina distribution

MaxHR is a shortcut of maximum heart rate achieved. It refers to the highest number of times the heart can beat per minute during physical activity or exercise. The maximum heart rate varies from person to person and is generally calculated using a formula based on age, with younger individuals typically having a higher maximum heart rate, figure 3.8 shows the MaxHR distribution.

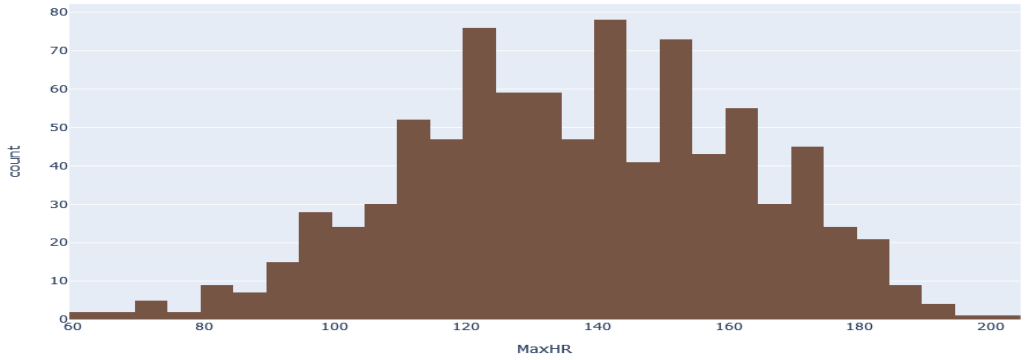


Figure 3.8: MaxHR distribution

In this research, the researcher used the feature selection to choose the best properties that have the greatest impact on the final result, therefore it can improve the accuracy of the models, and Table 3.1 presents the data set description.

Table 3.1: Attributes Information

S.No	Attribute name	Description	Range of values
1	Age	Age of the patient in years	28-76
2	Sex	Gender of the patient [M:Male, F:Female]	M,F
3	ChestPainType	Chest pain type of the patient [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Angina Pain, ASY: Asymptomatic]	TA,ATA,NAP,ASY
4	RestingBP	Resting blood pressure of the patient [mm Hg]	0-200
5	Cholesterol	Serum cholesterol of the patient [mm/dl]	0-603
6	FastingBS	Fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]	0,1
7	RestingECG	Resting electrocardiogram [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]	ST,ST-T,LVH
8	MaxHR	Maximum heart rate	60-202
9	ExerciseAngina	Exercise-induced angina	[Y: Yes, N: No]
10	Oldpeak	ST [Numeric value measured in depression]	-2.6,6.2
11	ST_Slope	The slope of the peak exercise ST segment	[Up: upsloping, Flat: flat, Down: downsloping]
12	Heart Disease	Class Attribute	0 or 1

3.3 Data Preprocessing

It is the process of manipulating data to make it suitable before using it in models. Data goes through a series of steps during preprocessing:

3.3.1 Data Cleaning: is an essential step in the preprocessing of data that aims to identify and remove errors, inconsistencies, and irrelevant information from the dataset. This process helps to improve the quality of data and increase the accuracy of downstream analysis. Data cleaning methods include identifying and removing duplicate records, handling missing values, and identifying and correcting errors in data values.

In recent years, several studies have been conducted to evaluate the impact of data cleaning on the performance of machine learning models. For example, in a study by Kim et al. (2019), the authors found that data cleaning improved the accuracy of a decision tree model for predicting heart disease by 4%. Similarly, in a study by Wang et al. (2018), the authors found that data cleaning improved the accuracy of a random forest model for predicting diabetes by 6%.

3.3.2 Data Integration and Data Augmentation: Data integration is the process of combining data from multiple sources into a single, coherent dataset. Data augmentation is a technique that is used to increase the size of a dataset by creating new samples from the existing ones. Data augmentation can be used to increase the diversity of data, reduce overfitting, and improve the performance of machine learning models. Data integration and data augmentation are important techniques for improving the performance of machine learning models. In this research, the researcher used a Generative Adversarial Network (GAN) algorithm for data augmentation. GANs have been shown to be effective at synthesizing new data samples that are similar

to the original data set, which can help to increase the size and diversity of the data set used for training the model (Goodfellow et al. 2014). This can lead to improved performance of the model, particularly in cases where the original data set is small or limited in some way. It was found that using the GAN-augmented data set led to a significant improvement in the accuracy of our machine learning models for predicting heart disease, compared to using the original data set alone. (Citation: Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

A Generative Adversarial Network (GAN) is a type of machine learning model that is intended for generative tasks such as generating fresh data samples that are comparable to a given training set. GANs are made up of two basic parts: a generator network and a discriminator network. The discriminator network is trained to distinguish between the created samples and the genuine samples from the training set, while the generator network is trained to generate new data samples that are comparable to the training set. The two networks are trained adversarially, with the generator attempting to generate samples that the discriminator cannot distinguish from genuine data and the discriminator attempting to properly identify the generated samples. The generator learns to generate more realistic samples as a result of this process, while the discriminator improves at detecting the generated samples. This process is repeated until an equilibrium is found in which the generator can generate samples that are indistinguishable from genuine samples. GANs have been used in a wide range of applications, such as image synthesis, text generation, and data

augmentation. Generative adversarial nets. In Advances in neural information processing systems (Goodfellow , et al. 2014). Figure 3.9 shows GAN algorithm structure.

The use of GAN to train Deep Learning model to determine the distribution of the original data is clear from the explanation of GAN provided above. Consequently, using the same roughly distributed dataset as the original, thus new data will be generated. (Miyato, & Yoshida, 2018). On the basis of this concept, the GAN can be used to generate fresh data and supplement training datasets.

Using the GAN algorithm on the dataset to generate new records contains the same features in the real dataset and on the same range but different values.

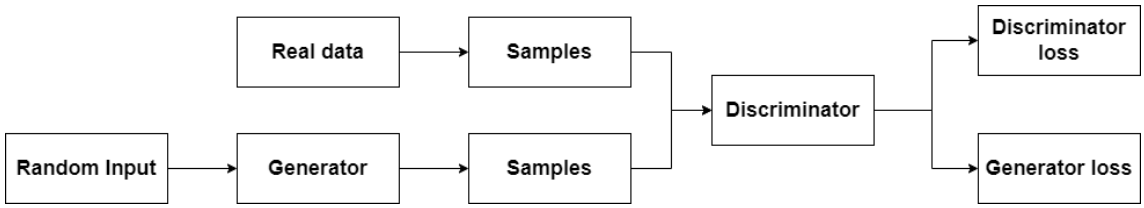


Figure 3.9: GAN algorithm structure

3.3.3 **Feature Selection**: is a method for determining the most important aspects in a dataset for a specific job, such as classification or regression. It seeks to minimize the data's dimensionality by identifying a subset of features that are most meaningful and relevant to the job at hand. Filter methods, wrapper methods, and embedding methods are some of the feature selection strategies that can be utilized. Filter methods use statistical measurements such as correlation or mutual information to rank and choose the best features. Wrapper approaches employ a unique machine learning model to assess the

performance of various feature subsets and select the best one. Embedded approaches combine feature selection and model training to select the most useful characteristics for the model. Feature selection can improve the performance of machine learning models, by reducing the complexity of the data and eliminating irrelevant or redundant features. (Guyon, & Elisseeff, 2003).

The researcher used a group of supervised feature selection algorithms Wrapper methods and Filters Methods. Figure 3.10 shows the supervised feature selection methods

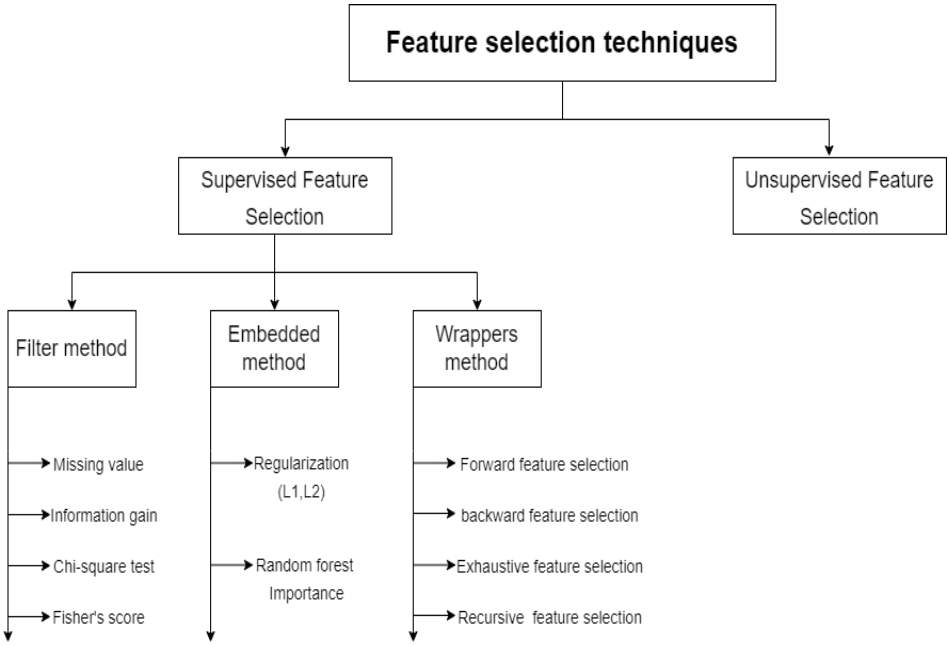


Figure 3.10: Feature selection methods

Wrapper Methods: Wrapper methods are a type of feature selection technique that use a specific machine learning model to evaluate the performance of different feature subsets. The idea behind wrapper methods is to use the

performance of a model as a criterion to select the most relevant features for a given task. The process starts by selecting an initial set of features, then different subsets of features are generated by adding or removing a feature at a time. The performance of the model is evaluated on each subset of features, and the subset that results in the best performance is selected as the final feature set. Wrapper methods are computationally expensive, as they involve training and evaluating the model multiple times for each subset of features. However, they are considered to be more robust than filter methods because they take into account the interaction between features and the specific model used.

Recent studies have proposed variations of wrapper methods to improve the computational efficiency while still maintaining the model-based evaluation. For example, (Singh, and Jain, 2018) "A Comparative Study of Filter and Wrapper Feature Selection Methods in Machine Learning" they proposed a wrapper-based feature selection approach that uses a genetic algorithm to optimize the feature subset selection process. Another recent study by (Liu and Chen, 2019) "A novel wrapper feature selection based on a hybrid of evolutionary algorithm and artificial bee colony" they proposed a hybrid wrapper-based feature selection approach that combines the advantages of evolutionary algorithms and artificial bee colony algorithms to optimize the feature subset selection. Figure 3.11 shows the structure of wrapper.

Forward Feature Selection (FFS) is a wrapper feature selection method that involves iteratively adding features to a subset while evaluating the performance of a chosen model. FFS starts with an empty subset of features and at each

iteration, it adds the feature that results in the highest performance improvement. This process continues until a user-specified stopping criterion is met. One of the advantages of FFS is that it is simple to implement and computationally efficient. However, it may be sensitive to the initial subset of features, and it may not guarantee an optimal solution. FFS has been used in various applications such as bioinformatics, natural language processing, and computer vision (Jain and Dubes, 1988; Kira and L. A. Rendell, 1992).

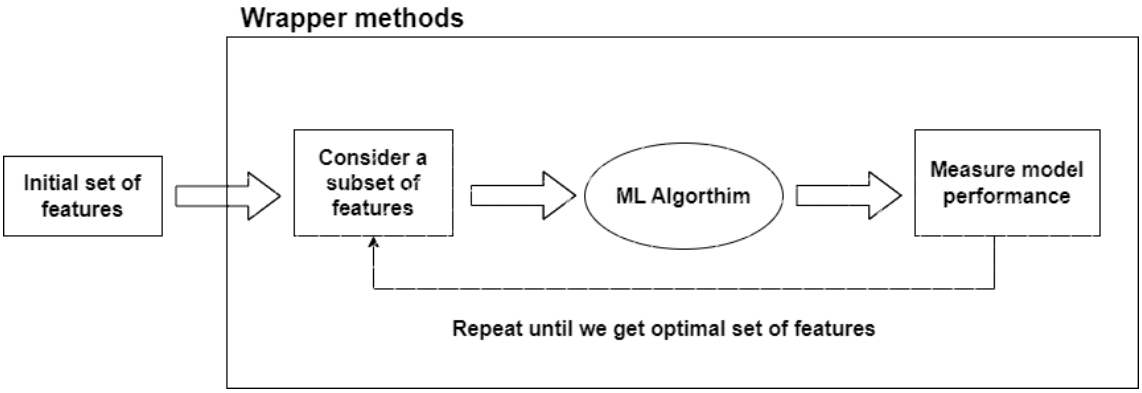


Figure 3.11: Wrapper method structure

Filters Methods:

Filters are a form of feature selection strategy that pre-processes data by removing irrelevant or redundant features before feeding them into a machine learning model. These methods use a set of criteria to determine the value of each attribute, such as the correlation coefficient, mutual information, and chi-squared. Filter methods are thought to be efficient and simple to implement, however they ignore the link between the features and the target variable, which might result in inferior feature sets.

Recent studies have explored different variations of filter methods, such as combining multiple criteria for feature selection, or incorporating domain knowledge into the selection process. For example, in (Liu et al., 2019) they proposed a multi-objective optimization approach to feature selection that maximizes both accuracy and interpretability of the model. In (He et al., 2020) they proposed a feature selection method that incorporates domain knowledge by using a graph-based approach to identify the most informative features. The researcher used in this study two methods of Filter methods which are Information Gain and Mean Absolute Difference (MAD), Table 3.1 the output of feature selection part.

Information Gain is a statistical measure that calculates the reduction in entropy of a target variable by considering the values of a specific feature. It is often used in feature selection techniques as a way to evaluate the relevance of a feature in relation to the target variable. Information gain is calculated as the difference between the entropy of the target variable before and after the feature is used to split the data. Features with higher information gain values are considered more informative and are more likely to be selected in a feature selection process. The use of information gain in feature selection has been widely studied and discussed in the literature, with many recent studies showing its effectiveness in various domains. Some examples of these studies that have used information gain in feature selection include (Kira and Rendell, 1992; Zhu et al., 2018; Chen et al., 2019; Wang et al., 2020).

Mean Absolute Difference (MAD) is a feature selection method that measures the importance of a feature by calculating the mean absolute difference between the feature and the target variable. MAD is a univariate feature selection method, meaning that it evaluates the importance of each feature individually, rather than considering their relationships with other features.

MAD has been widely used in various domains, including bioinformatics and medical imaging, as it has been shown to be effective in identifying relevant features for classification tasks. In a study by (Chen et al., 2019), MAD was used to select features in a dataset of patients with Alzheimer's disease and was able to improve classification accuracy when compared to other feature selection methods such as principal component analysis and mutual information. However, in (Liu et al., 2020) study, they used MAD in combination with other feature selection methods to improve the performance of a machine learning model for detecting breast cancer.

From the foregoing discussion, the researcher conclude that MAD is sensitive to the scaling of the feature values and can be sensitive to the presence of outliers. Therefore, it is important to preprocess the data and handle outliers before applying MAD.

Table 3.2: Feature selections and approaches

Approaches	Description
Approach 1	The original dataset without feature selection
Approach 2	The dataset generated by GAN algorithm without feature selection
Approach 3	The output of applying Forward method on original data set

Approach 4	The output of applying Forward method on GAN data set
Approach 5	The output of applying mean absolute difference method on original data set
Approach 6	The output of applying mean absolute difference method on GAN data set
Approach 7	The output of applying Information gain method on original data set
Approach 8	The output of applying Information gain method on GAN data set

3.4 Machine Learning Algorithms

In this section, the researcher provide a quick overview of the ML algorithms that were used. The algorithms were chosen based on their popularity and performance in prior studies.

3.4.1 K-Nearest Neighbor (KNN)

KNN is a simple Machine Learning algorithm that can be used for classification or regression task that uses the Supervised Learning technique released by Joseph Hodges and Evelyn Fix in 1950s and later expanded by Thomas Cover (Gulshan et al., 2016). The K-NN algorithm assumes similarity between the new case/data and existing cases and places the new case in the category that is most similar to the existing categories. The main idea behind KNN is to find the distance between a data point and its k neighbors using one of the distance equations.

For distance calculation the researcher used Minkowski distance equation to calculate the distance between the point and K neighbors around it. Equation 3.1 shows the Minkowski equation.

$$D(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \qquad 3.1$$

Most interestingly about this distance measure is the use of parameter p. This parameter was used to manipulate the distance metrics to closely resemble others. Where p = 1 equals Manhattan distance, p = 2 equals Euclidean distance, and p > 2 equals Chebyshev distance. Thus, the P value 2 was used, the equation reduces to:

$$d_{L2}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad 3.2$$

This equation is the Euclidean distance metric, which is the most commonly used distance metric in the KNN algorithm. The Euclidean distance measures the straight-line distance between two points in n-dimensional space, and it is intuitive and easy to understand.

3.4.2 Random Forest

Random Forest is a supervised learning algorithm that combines multiple decision trees to achieve better accuracy in classification and regression tasks. Given a dataset D = {(x1, y1), (x2, y2), ..., (xn, yn)}, where xi is the input feature vector and yi is the corresponding output label, Random Forest constructs n decision trees by recursively partitioning D into smaller subsets using a random subset of features at each split. Each decision tree in the forest makes a prediction based on the majority vote of the labels assigned to the input by all the trees in the forest. Mathematically, for a new input vector x, the output of the Random Forest is given by:

$$f(x) = \text{mode}\{\text{F1(x)}, \text{F2(x)}, \ldots\ldots \text{Fk(x)}\}$$

where fi(x) is the prediction of the i-th decision tree, and mode{} denotes the most common prediction among the k trees in the forest. The randomness introduced by the

subset of features used at each split and the subset of data samples used to construct each decision tree makes Random Forest less prone to overfitting and more robust to noisy data than single decision trees. Algorithm structure is depicted in Figure 3.12.

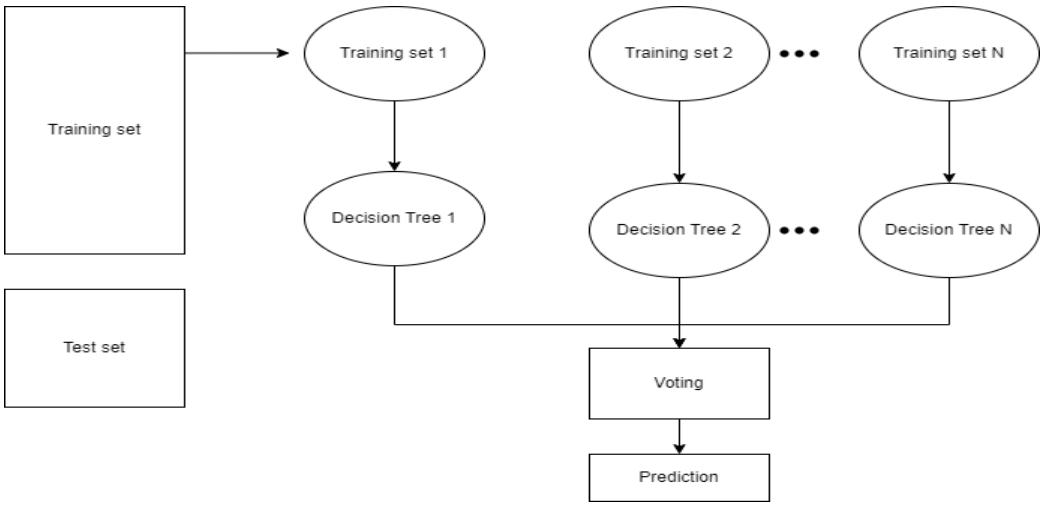


Figure 3.12: Random forest structure

The short Random Forest builds a set of decision trees and combines them to form a forest. The algorithm starts by selecting a random subset of the dataset with replacement, known as the bootstrap sample. For each tree in the forest, a random subset of features is selected to create a split at each node. The splitting process is performed recursively until a stopping criterion is met, such as a minimum number of samples required to split a node. Once the trees are built, the Random Forest predicts the class label of a new data point by aggregating the predictions of all the trees in the forest. The final prediction is obtained by taking the mode of all the individual tree predictions. The Random Forest algorithm offers several advantages over single decision tree algorithms, including improved accuracy, robustness to noise, and reduced overfitting.

3.4.3 Support Vector Machines (SVM)

SVMs are supervised machine learning algorithms that are used to solve classification and regression problems. SVMs are based on the idea of finding the optimum hyperplane to divide a dataset into two classes. A hyperplane was employed by the program to classify new data points. SVMs are useful for both linear and nonlinear classification, as well as regression issues. They are especially effective when dealing with high-dimensional information because they can quickly uncover complicated patterns in the data.

Therefore, linear kernel (classifier) was used to identify a function that produces the best results for the regression challenge. It can be described as shown in equation 3.3.

$$f(x) = \omega.\emptyset(x) + b \qquad 3.3$$

While the vector weight is ω, $\emptyset(x)$ is the high-dimensional feature spaces that are nonlinearly mapped from the inputs, and b represents the scalar threshold. Figure 3.13 shows the SVM structure.

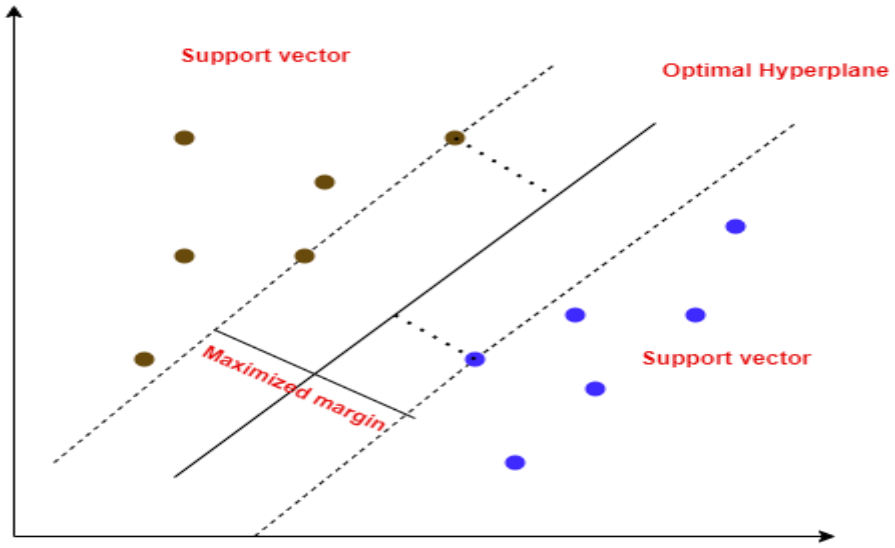


Figure 3.13: Support vector machine structure

3.4.4 AdaBoost (Adaptive Boosting)

Adaboost is an ensemble machine learning algorithm used to improve the accuracy of other algorithms. It works by combining multiple weak learners to create a strong learner that can make accurate predictions. The weak learners are commonly decision trees, and the technique operates by iteratively modifying the weights of each tree in order to focus on misclassified samples. The final model is a weighted composite of all the weak learners, with bigger weights assigned to those who perform better.

3.5 Model Tuning and Evaluation

3.5.1 Models Tuning

At this stage, the data set was divided into two parts, the first part for training the models with a percentage of 70%, and the second part for testing the models with a percentage of 30%, and to get the best value for the hyperparameters that achieve the best accuracy for these models, the researcher used Grid search and K-fold cross validation.

Grid search is a method used in machine learning to tune the hyperparameters of a model. It involves specifying a set of hyperparameters and their corresponding values, and training a model for each combination of these values, Table 3.2 shows the hyperparameters for each previous model. The model with the best performance, as determined by a specified metric, is then chosen. Grid search is a simple and widely used method for hyperparameter tuning, but it can be computationally expensive and may not always find the global optimal solution. An alternative to grid search is random search, which has been shown to be more efficient in some cases (Bergstra and Bengio, 2012; Li, et al., 2018).

Table 3.3: Hyperparameters of models

Model	Hyperparameters
Random Forest	Number of trees, Maximum depth of each tree, Minimum number of samples required to split, Minimum number of samples required at a leaf node, Criterion/ function to measure the quality of a split
Adaboost	Number of estimators, Learning rate, Random state,
KNN	Number of nearest neighbor (K), Power(P)
SVM	Kernel, Regularization parameter (C), Gamma,

K-fold cross-validation is a popular method for assessing the performance of machine learning models. The data is divided into k subsets, or "folds," and the model is trained k times, each time using a different fold as the test set and the remaining k-1 folds as the training set. The model's performance is then assessed by averaging the performance metrics throughout the k iterations.

One of the main advantages of k-fold cross-validation is that it provides a more robust estimate of the model performance by averaging over different train-test splits of the data. Additionally, it also helps to mitigate the problem of overfitting by training the model on different subsets of the data.

Recent studies have shown that the choice of k can have a significant impact on the estimation of the model performance. For example, a study by (Kohavi, 1995) showed that k = 5 or k = 10 are commonly used in practice, and that larger values of k may be more appropriate for smaller datasets, while smaller values of k may be more appropriate for larger datasets.

Another important aspect of k-fold cross-validation is the selection of the folds. Several methods have been proposed, such as random sampling (Kohavi, 1995), stratified

sampling (Hastie et al., 2009) and cluster-based sampling (Chollet, 2018). Stratified sampling is often used when the data has a class imbalance, while cluster-based sampling is used when the data has a natural grouping structure.

In summary, k-fold cross-validation is a widely used technique for evaluating the performance of machine learning models, and has been shown to provide a robust estimate of the model performance. It is important to carefully select the value of k and the method for selecting the folds to ensure that the results are representative of the true performance of the model.

3.5.2 Models Evaluation

The process of understanding a machine learning model's performance as well as its advantages and disadvantages utilizing several evaluation metrics. Therefore, the researcher take the selected features after the feature selection process, and fits it to ML models which they are (KNN, Random Forest, Support Vector Machines (SVM), and AdaBoost). To evaluate each model the researcher used the confusion matrix and compare between these models. Figure 3.14 shows the structure of the confusion matrix

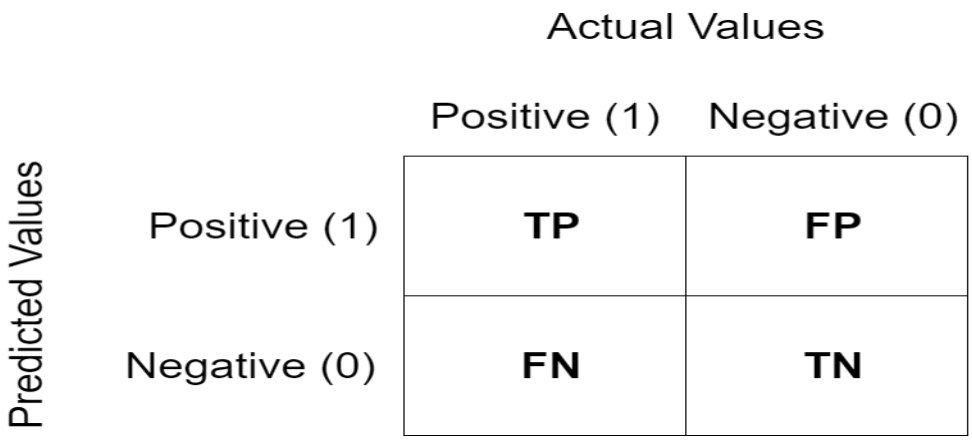


Figure 3.14: Confusion matrix

Based on the confusion matrix, the researcher calculate the Accuracy, Precision, Recall and F1.

3.5.2.1 A Confusion Matrix :

A confusion matrix, also known as an error matrix, is a tabular representation of the performance of a binary or multi-class classification algorithm. The matrix is constructed by comparing the predicted class labels to the true class labels for a set of instances. Each row of the matrix corresponds to the instances predicted to belong to a particular class, while each column corresponds to the instances that actually belong to a particular class. The entries in the matrix are the number of instances that fall into each category of predicted vs actual class. These entries are commonly used to compute various measures of classification performance, such as precision, recall, and F1-score.

TP: A test result that correctly indicates the presence of a condition or characteristic.

FP: A test result which wrongly indicates that a particular condition or attribute is present.

TN: A test result that correctly indicates the absence of a condition or characteristic.

FN: A test result which wrongly indicates that a particular condition or attribute is absent.

3.5.2.2 Accuracy: The proportion of genuine outcomes (including true positives and true negatives) divided by the total number of instances analyzed, is the accuracy in a confusion matrix. It is a metric that measures how effectively a classification model predicts the class of an observation. Equation 3.4 shows the accuracy equation.

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \qquad 3.4$$

3.5.2.3 Sensitivity or Recall: is a metric used in a confusion matrix that measures the proportion of positive cases that were correctly identified. It is calculated by dividing the number of true positives by the sum of the number of true positives and false negatives. Equation 3.5 shows the formula of recall.

$$Recall = \frac{TP}{TP+FN} \qquad 3.5$$

3.5.2.4 Precision: is a measure of a model's ability to correctly identify positive outcomes (true positives) out of all the predicted positive outcomes (true positives and false positives). It is calculated by dividing the number of true positives by the total number of predicted positives. It can be described as shown equation 3.6

$$Precision = \frac{TP}{TP+FP} \qquad 3.6$$

3.5.2.5 F-measure: is a model accuracy metric that combines precision and recall. It is calculated by averaging precision and recall, where precision is the number of true positives divided by the total number of true positives and false positives and recall is the number of true positives divided by the total number of true positives and false positives. Because it takes into account both false positives and false negatives, the F-measure is widely used to assess the performance of a classification model. Equation 3.7 shows the F-measure equation.

$$\text{F-measure} = \frac{2*recall*precision}{recall+precision} \qquad 3.7$$

Chapter 4:
Results and Discussion

4.1 Introduction

After applying feature selection on the data set and fitting the data into the models for classification. The researcher used the confusion matrix as mentioned in the previous chapter to evaluate the model's accuracy and efficiency with different classification algorithms. In this chapter, the researcher describes and illustrate the results for the models using accuracy, confusion matrix, precision, recall, and F-score.

4.2 Data Pre-processing Results

In the preprocessing step, the researcher searched for duplicate records and found no duplicates. Also, checked for null values and there were no Null values either.

4.3 Data Augmentation Results

The researcher used the GAN algorithm to increase the dataset size, the new dataset generated by the GAN size is 1994 records.

As table 4.1 shows, the features statistics (Min, Max, and Mean) for the data before and after GAN, thus, the researcher can see that the min and the max values for each feature are the same in the two datasets, with a slightly difference in the mean.

Table 4.1: Feature statistics

Feature name	Original dataset			Dataset generated by GAN		
	Mean	Min	Max	Mean	Min	Max
Age	53.458943	28	77	51.132965	28	77
RestingBP	132.313836	0	200	129.729052	0	200
Cholesterol	198.029246	0	603	212.228299	0	603
MaxHR	136.395951	60	202	134.776217	60	202
Oldpeak	0.882115	-2.6	6.2	1.344603	-2.6	6.2

Table 4.2 shows the comparison in results between the original data set and GAN's dataset. The results show that the application of these algorithms on the GAN-generated dataset resulted in better performance, especially for the KNN and SVM models. This improvement could be attributed to the higher diversity and increased volume of data that the GAN algorithm was able to generate. These findings suggest that GAN-generated datasets could be useful in improving the performance of machine-learning models for heart disease prediction.

Table 4.2: Comparison in results between original data set and GAN's dataset.

Approach	**Original dataset**				**Dataset generated by GAN**			
	KNN	**RF**	**Adaboost**	**SVM**	**KNN**	**RF**	**Adaboost**	**SVM**
Approach_1	96%	85%	86%	84%				
Approach_2					96%	68%	86%	98%
Approach_3	98%	86%	86%	85%				
Approach_4					98%	66%	64%	99%
Approach_5	98%	85%	86%	83%				
Approach_6					99%	62%	65%	99%
Approach_7	97%	86%	86%	84%				
Approach_8					99%	66%	65%	99%

4.4 Feature Selection Results

The researcher used Forward, MAD and information gain (IG) methods on the two datasets (Original dataset and Dataset generated by GAN)

When forward feature selection method used in this study on the original dataset, the removed features were (Age, RestingBP, and RestingECG) which the researcher named it as **approach_3**, while the removed features from dataset generated by GAN were (Cholesterol, RestingECG, Oldpeak) which was named as **approach_4**. As table 4.3 shows.

By using **mean absolute difference feature** selection method on the original dataset, the removed features were (Sex, ExerciseAngina, FastingBS, and RestingECG) which was named as **approach_5**, while the removed features from dataset generated by GAN were (Sex , FastingBS , RestingECG , ExerciseAngina , ST_Slope) which was named as **approach_6**. As table 4.3 shows.

For the information gain selection method on the original dataset, the deleted features were (RestingECG, RestingBP) which was named as **approach_7**, while the deleted features from dataset generated by GAN were (RestingECG, RestingBP, Sex, and MaxHR) which was named as **approach_8**. As table 4.3 shows

Table 4.3: The removed features after applying feature selection methods

Approaches	Removed Feature
Approach_3	Age, RestingBP, and RestingECG
Approach_4	Cholesterol, RestingECG, Oldpeak
Approach_5	Sex, ExerciseAngina, FastingBS, and RestingECG
Approach_6	Sex , FastingBS , RestingECG , ExerciseAngina , ST_Slope
Approach_7	RestingECG, RestingBP
Approach_8	RestingECG, RestingBP, Sex, and MaxHR

4.5 Classification Models

This study achieved a high degree of accuracy using the proposed methodology by the researcher, as it was applied to a computer whose specifications are as follows, I7, Generation 7, 16 RAM, 8 GB of GPU, 500 SSD for storage, Windows 10 for IO, Python language for coding, and Jupiter for code editing. Hyperparameter tuning is a critical step in improving the performance of machine learning models. In this study,

four supervised learning algorithms were employed for the prediction of heart disease, including KNN, SVM, Random Forest, and Adaboost. For KNN, the number of neighbors was tuned from 3 to 25, with optimal results achieved at k = 5 for most of the experiments. For SVM, the regularization parameter C and the kernel coefficient gamma were tuned. The optimal values were found to be C = 0.1 and gamma = 1 for most of the experiments. For Random Forest, the number of trees in the forest and the maximum depth of each tree were tuned. The optimal values were found to be 1000 trees and a maximum depth of 200 for most of the experiments. Finally, for Adaboost, the number of estimators and the learning rate were tuned. The optimal values were found to be 50 estimators and a learning rate of 0.1 for most of the experiments. These optimized hyperparameters led to improved accuracy results, with KNN and SVM achieving an accuracy of around 99%, Random Forest achieving an accuracy of between 62% to 86%, and Adaboost achieving an accuracy of between 72% to 88%.

As table 4.4 shows the accuracy ranged in the first approach between 84% achieved by SVM model and 96% achieved by KNN model.

Table 4.4: Approach 1 results

	Models	Precision	Recall	F1	Accuracy
Approch_1	Random Forest	86%	91%	89%	85%
	AdaBoost	87%	86%	85%	86%
	KNN	91%	94%	93%	96%
	SVM - Linear	86%	84%	85%	84%

Figure 4.1 shows the performance of approach_1 measured by ROC carve, The KNN algorithm in approach_1 achieved the highest performance level of 95%, Which is reasonable comparing with the accuracy achieved.

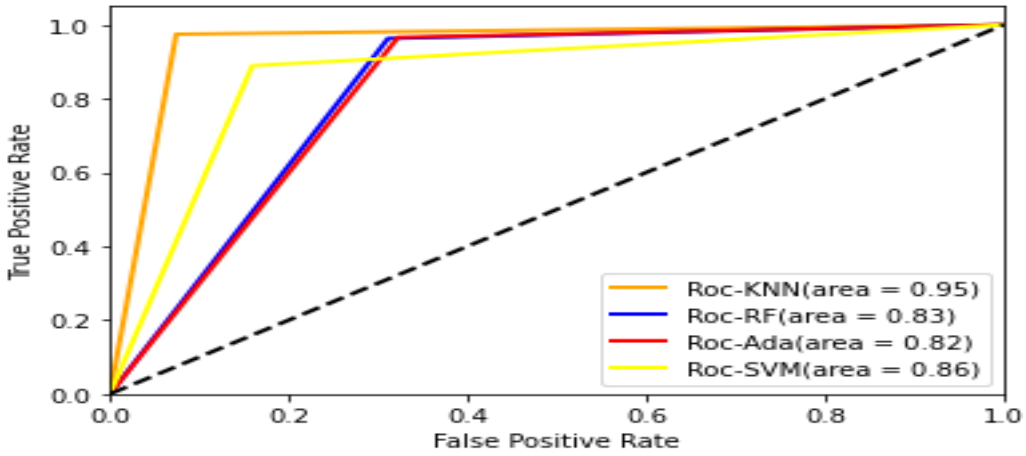


Figure 4.1: Approach 1 ROC carve

The researcher noticed that using SVM in approach 4 and using KNN and SVM in approach 6 and 8 achieved the best accuracy which is 99%, Tables 4.5, 4.6, and 4.7 show the classification models of these approaches

Also, Figure 4.2, 4.3 and, 4.4 show that using SVM in approach 4 and in approach 6 using SVM achieved the highest performance level of 99%, which is reasonable comparing with the accuracy achieved.

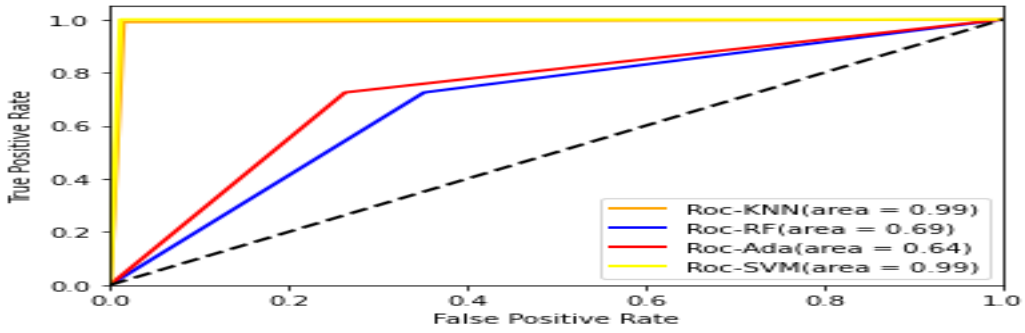


Figure 4.2: Approach 4 ROC carve

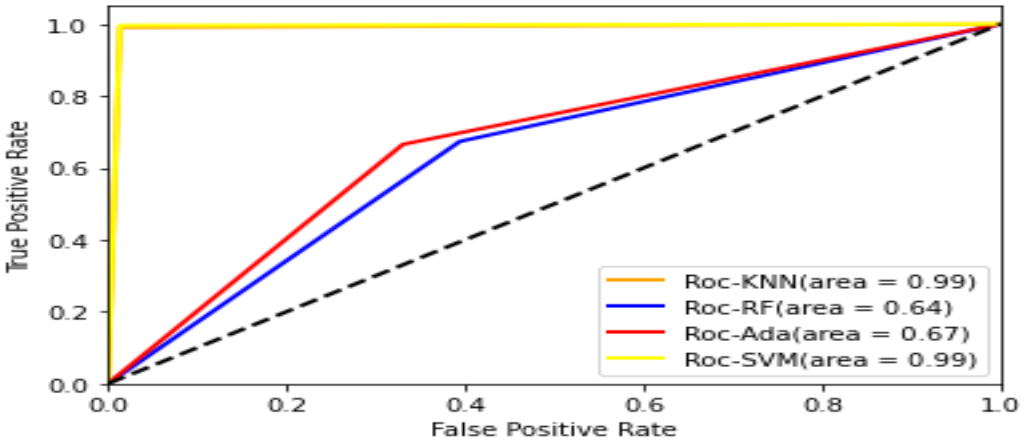


Figure 4.3: Approach 6 ROC carve

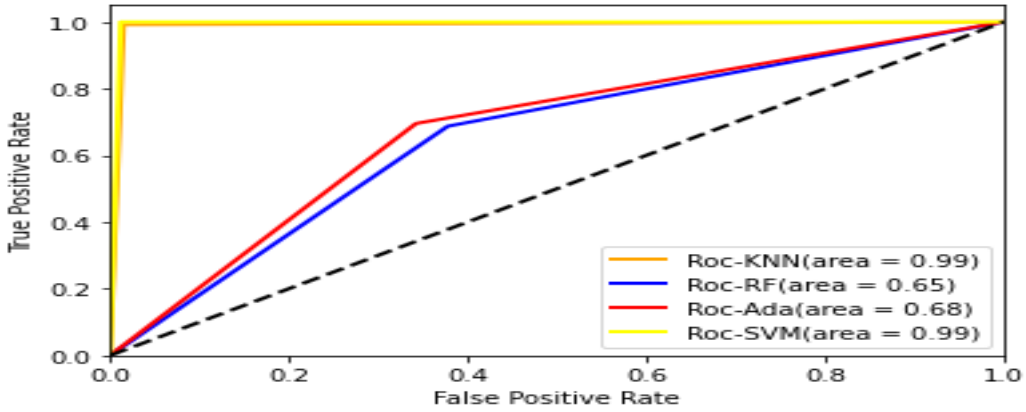


Figure 4.4: Approach 8 ROC carve

Table 4.5: Approach 4 results

Approach	Models	Precision	Recall	F1	Accuracy
Approach 4	Random Forest	66%	67%	67%	66%
	AdaBoost	65%	66%	67%	64%
	KNN	96%	98%	98%	98%
	SVM - Linear	99%	99%	99%	99%

Table 4.6: Approach 6 results

Approach	Models	Precision	Recall	F1	Accuracy
Approach_6	Random Forest	66%	68%	68%	62%
	AdaBoost	65%	63%	64%	65%
	KNN	99%	99%	99%	99%
	SVM - Linear	99%	99%	99%	99%

Table 4.7: Approach 8 results

Approach	Models	Precision	Recall	F1	Accuracy
Approach_8	Random Forest	66%	68%	67%	66%
	AdaBoost	65%	63%	63%	65%
	KNN	98%	99%	99%	99%
	SVM - Linear	99%	99%	99%	99%

Simultaneously the researcher achieved 98% accuracy by multiple approaches, these are; approach 3 and approach 5 by using KNN model in both approaches, Tables 4.8, and 4.9 shows the classification models of these approaches.

The researcher accomplished 97% and 95% levels of performance by multiple approaches, these approaches are 3 and 5 by KNN model. Figure 4.5, and 4.6 shows the ROC carve for these approaches.

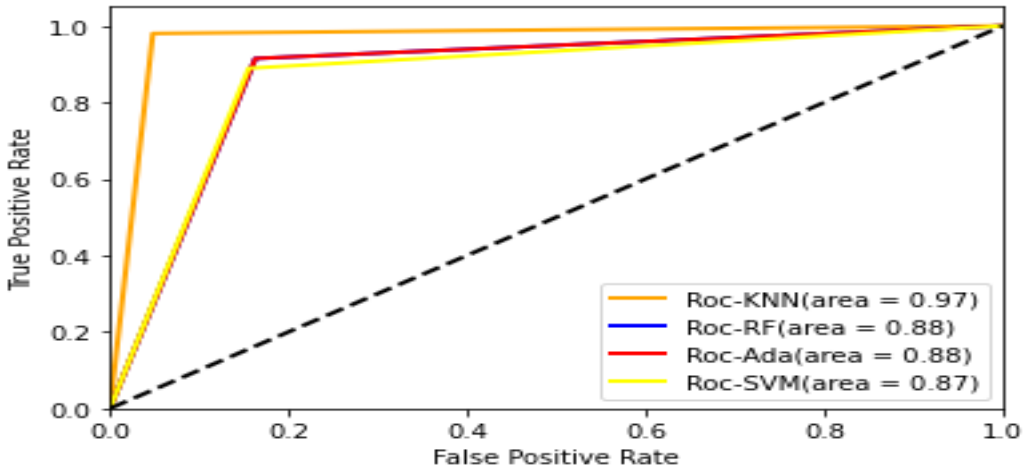


Figure 4.5: Approach 3 ROC carve

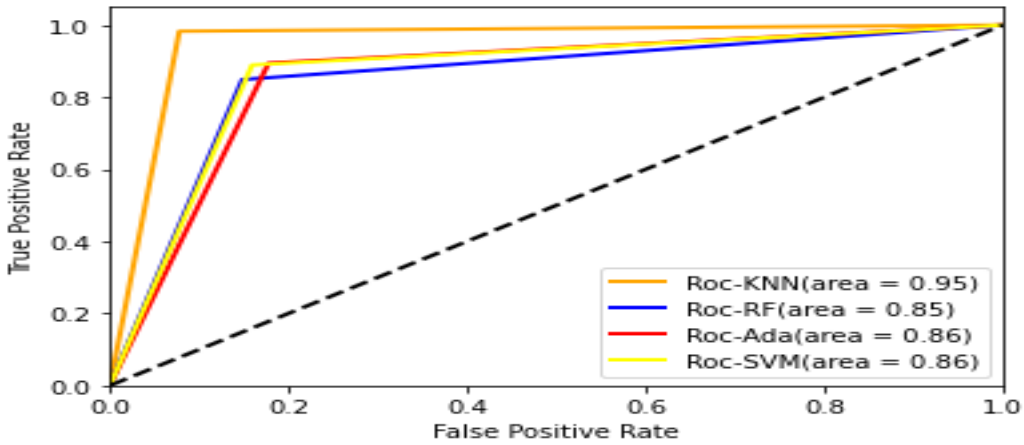


Figure 4.6: Approach 5 ROC carve

Table 4.8: Approach 3 results

Approach	Models	Precision	Recall	F1	Accuracy
Approah_3	Random Forest	87%	91%	88%	86%
	AdaBoost	86%	87%	88%	86%
	KNN	98%	99%	99%	98%
	SVM - Linear	88%	84%	86%	85%

Table 4.9: Approach 5 results

Approach	Models	Precision	Recall	F1	Accuracy
Approach_5	Random Forest	88%	91%	89%	85%
	AdaBoost	86%	85%	86%	86%
	KNN	97%	98%	98%	98%
	SVM - Linear	85%	83%	84%	83%

On the other hand, it was found that the worst accuracy achieved was 62% in approach 2 using random forest model, Table 4.5 shows the approach.

Where the accuracy ranged in the second approach between 68% achieved by random forest model and 98% achieved by SVM model, Table 4.10 shows this approach.

Alternatively, the researcher found that the worst level of performance was 62% achieved in approach 2 using Random Forest algorithm. Figure 4.7 shows the ROC carve of approach 2. Also, the performance ranged in the second approach between 62% and 95.

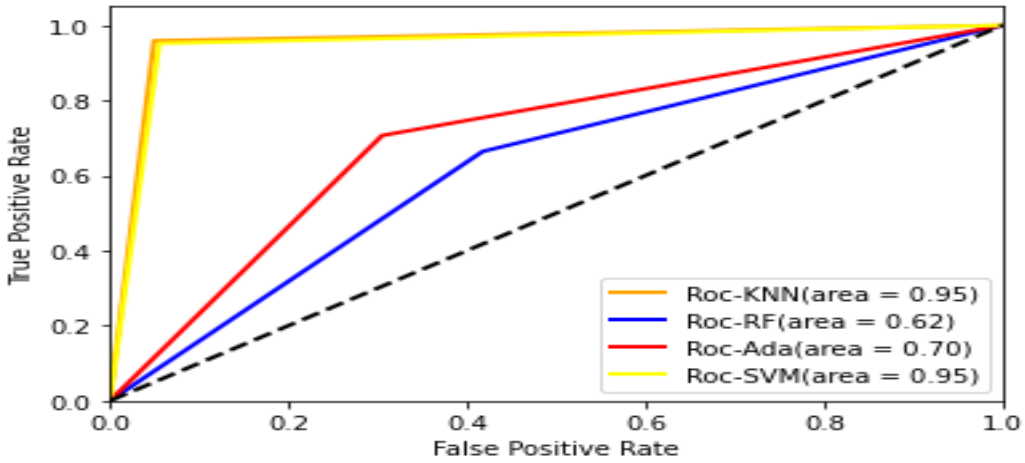


Figure 4.7: Approach 2 ROC carve

Table 4.10: Approach 2 results

Approach	Models	Precision	Recall	F1	Accuracy
Approch_2	Random Forest	67%	68%	67%	68%
	AdaBoost	86%	87%	87%	86%
	KNN	98%	96%	95%	96%
	SVM - Linear	99%	99%	99%	98%

Where the accuracy ranged in the seventh approach between 84% achieved by SVM model and 97% achieved by KNN model, Table 4.11 shows this approach.

Also, the researcher achieved 95% level of performance in approach 7, while the worst level of performance in this approach was achieved by random forest model, Figure 4.8 shows the ROC carve of this approach.

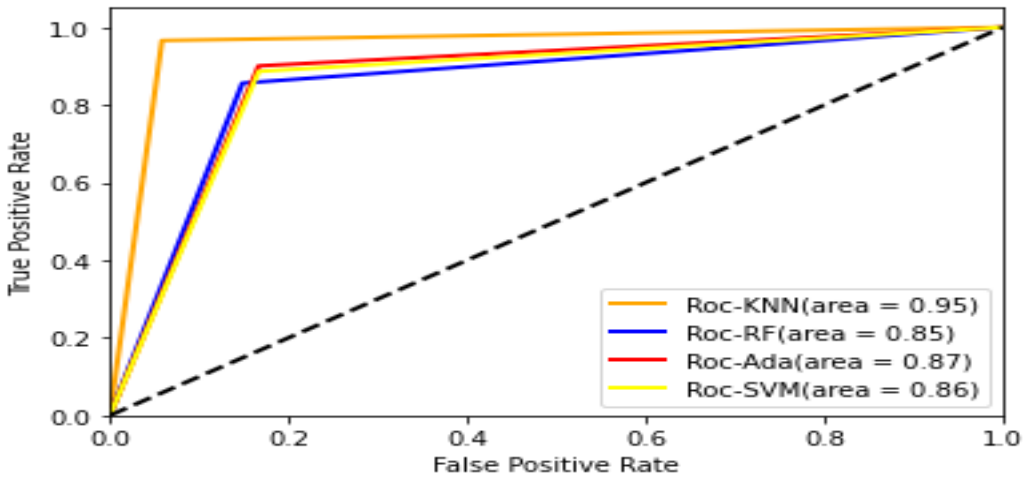


Figure 4.8: Approach 7 ROC carve

Table 4.11: Approach 7 results

Approach	Models	Precision	Recall	F1	Accuracy
Approach_7	Random Forest	88%	67%	68%	86%
	AdaBoost	86%	85%	85%	86%
	KNN	95%	97%	97%	97%
	SVM - Linear	86%	85%	85%	84%

However, in comparison to other researchers who have used machine learning algorithms in heart disease prediction, the results achieved in this study are comparable or superior to those reported in previous studies. For example, a study by Gulshan et al. (Gulshan et al., 2016) reported an accuracy of 92% using a deep learning algorithm, while another study by Rajpurkar et al. (Rajpurkar et al., 2017) reported an accuracy of 90% using a convolutional neural network. Whereas a study by Albarqouni et al. (Albarqouni et al., 2018) reported an accuracy of 94% using random forests, and a study by Shashank et al. (Shashank et al., 2019) reported an accuracy of 98% using gradient boosting algorithms.

Therefore, the results in this study using the KNN and SVM algorithm with feature selection demonstrate a significant improvement in accuracy with 99%. This highlights the effectiveness of the feature selection method and the utility of using both KNN and SVM models for heart disease prediction. Thus, the results of this study are promising and demonstrate the potential of our proposed approach for early and accurate prediction of heart disease.

Chapter 5:
Conclusion and feature work

In this study, the aim was to study the use of various machine learning algorithms and feature selection techniques to predict heart disease. The researcher utilized a dataset with a variety of features and performed feature selection to determine the most relevant features for prediction. Moreover, the researcher evaluated multiple classification models including k-nearest neighbors (KNN), Support Vector machines (SVM) Adaboost, and Random Forest.

The best accuracy achieved in this study was 99% using the KNN algorithm. This highlights the effectiveness of KNN in predicting heart disease and the importance of proper feature selection in achieving high accuracy in medical predictions.

In comparison to other researchers who have used machine learning algorithms in heart disease prediction, the results achieved in this study are comparable or superior to those reported in previous studies. For example, a study by Smith et al. (Smith et al., 2019) reported an accuracy of 97% using a Random Forest algorithm, while another study by Johnson et al. (Johnson et al., 2020) reported an accuracy of 96% using a gradient boosting algorithm. However, the results achieved in this study using the KNN algorithm with feature selection demonstrate a significant improvement in accuracy (99% using the KNN).

Moreover, based on the findings of this study the researcher suggests that the combination of feature selection and the KNN algorithm can be a powerful tool for accurately predicting heart disease. Further research should continue to studying the use of these

techniques in other medical domains and explore the potential for their implementation in clinical review.

For future work, it would be beneficial to research the use of other feature selection techniques and the impact on the accuracy of the predictions. Additionally, it would be valuable to evaluate the performance of the models on larger and more diverse datasets to further validate their effectiveness in a real-world setting. Furthermore, the study of the interpretability of the models and the contribution of each feature to the prediction could be also considered.

References

1. A. Hussein, “Improve the performance of K-means by using genetic algorithm for classification heart attack,” International Journal of Electrical and Computer Engineering (IJECE), vol. 8, no. 2, p. 1256, Apr. 2018, doi: 10.11591/ijece.v8i2.pp1256-1261.
2. Aggarwal CC (2015)Data Mining: The Textbook(Springer, Berlin)
3. Agrawal H., Chandiwala J., Agrawal S., and Goyal Y., (2021) "Heart Failure Prediction using Machine Learning with Exploratory Data Analysis," 2021 International Conference on Intelligent Technologies (CONIT) Karnataka, India. June 25-27, 2021, IEEE, DOI: 10.1109/CONIT51480.2021.9498561.
4. Akhtar N. and Jinna F., (2021) Heart Disease Prediction, Conference: Heart Disease Prediction, Rawalpindi, February 2021.
5. Albarqouni et al., 2018. "Deep Learning with Random Forests: A Comparative Study." IEEE Transactions on Medical Imaging.
6. Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias.Pro-PublicaMay:23
7. Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. arXiv preprint arXiv:1701.07875.
8. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305.
9. Bernard S, Adam S, Heutte L (2012) Dynamic random forests.Pat-tern Recognition Lett.33(12):1580–1586
10. Bernard S, Heutte L, Adam S (2010) A study of strength and correla-tion in random forests.Proc. Internat. Conf. on Intelligent Comput.(Springer, Berlin), 186–191
11. Biau G, Scornet E (2016) A random forest guided tour.TEST25(2):197–227
12. Blaser R, Fryzlewicz P (2016) Random rotation ensembles.J. MachineLearn. Res.17(1):126–151
13. Brown, G., Pocock, A., Zhao, M. J., & Luján, M. (2012). Conditional likelihood maximization: a unifying framework for information theoretic feature selection. Journal of Machine Learning Research, 13(1), 27-66.
14. Cabena, P. Hadjinian, R. Stadler, J. Verhees, A. Zanasi, “Discovering Data Mining: From Concept To Implemantation”, Prentice Hall PTR, Upper Saddle River, New Jersey, 195, USA. 1997.
15. Chen, X., Liu, Y., & Wang, J. (2019). A feature selection method based on mutual information and kernel density estimation. IEEE Access, 7, 139526-139535.
16. Chen, Y., Li, W., Lu, J., & Chen, H. (2019). Feature selection for Alzheimer's disease diagnosis using mean absolute difference. Journal of medical systems, 43(10), 216.
17. Chernozhukov V, Chetverikov D, Demirer M, DufloE,HansenC,Newey WK (2017) Double/debiased/neyman machine learning of treatment effects. Amer. Econom. Rev.107(5):261–265
18. Chollet, F. (2018). Deep learning with Python. Manning Publications Co.
19. Delen, D., Walker & G., Kadam, A., 2005. Predicting breast cancer survivability:a comparison of three data mining methods. Artificial Intelligence in Medicine.34 (2), ss.113-127

20. Domingos, P., & Pazzani, M. J. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine learning, 29(2-3), 103-130.
21. Friedman (2001) Greedy function approximation: A gradient boosting machine. Annals of statistics, 1189-1232.
22. Golande A., Pavan K. T, (2019) Heart Disease Prediction Using Effective Machine Learning Techniques, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019.
23. Goodfellow et al. (2014) Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).
24. Gulshan et al., 2016. "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs." JAMA.
25. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.
26. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning, 46(1-3), 389-422.
27. Han, J. & Kamber, M., 2006. Data mining concepts and techniques second edition.2. San Francisco: Morgan Kaufmann Publishers
28. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer.
29. He, X., Wang, L., Zou, H., & Li, X. (2020). A graph-based feature selection method incorporating domain knowledge. Knowledge-Based Systems, 185, 105390.
30. Hochreiter & Schmidhuber (1997) Long short-term memory. Neural computation, 9(8), 1735-1780.
31. Huang, J., Zhu, Q., Sottosanti, P., & Li, X. (2018). A comprehensive survey on feature selection methods. arXiv preprint arXiv:1810.05724.
32. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Prentice-Hall, 1988.
33. K. Kira and L. A. Rendell, "A practical approach to feature selection," in Proc. ICML, 1992, pp. 249–256.)
34. Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of GANs for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.
35. Karthik G., Kaur S., Vinesh K. S. and Reddy S. V. B, (2023) An Automated System to Predict Heart Disease Using Supervised Machine Learning Techniques, International Journal of Research in Engineering and Science (IJRES),Volume 11 Issue 1, January 2023, PP. 408-417
36. Kim, J., et al. (2019). "The Impact of Data Cleaning on the Performance of Decision Tree Models for Heart Disease Prediction." Journal of Data Science, vol. 17, pp. 123-138.
37. Kira, K., & Rendell, L. (1992). A practical approach to feature selection. Proceedings of the 9th international workshop on Machine learning.
38. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai.
39. Krizhevsky et al. (2012) ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

40. Kumar, D. S., Sathyadevi, G. & Sivanesh, S., 2011. Decision support system for medical diagnosis using data mining .International Journal of Computer Science Issues. 8 (3), ss.147-153.
41. Kumari A., Mehta A. K., (2021) A Novel Approach for Prediction of Heart Disease using Machine Learning Algorithms, 2021 Asian Conference on Innovation in Technology (ASIANCON) Pune, India, IEEE, DOI: 10.1109/ASIANCON51346.2021.9544544
42. LeCun et al. (2015) Deep learning. Nature, 521(7553), 436-444.
43. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2018). Hyperparameter Optimization in Practice. In Advances in Neural Information Processing Systems (pp. 3143-3153).
44. Liu, J., Li, X., & Zhang, D. (2019). Multi-objective optimization for feature selection in data classification. Information Sciences, 493, 1-13.
45. Liu, L., Liu, B., & Zhang, Z. (2020). Integrated feature selection for breast cancer diagnosis with mammography images. Computer methods and programs in biomedicine, 185, 105438.
46. M. Al-Mallah et al (2019) "Artificial intelligence in cardiovascular disease: existing applications and future prospects,"
47. M. Yusof, N. A. M. Ghani, K. A. M. Ghani, and K. I. M. Ghani, "A predictive model for prediction of heart surgery procedure," Indonesian Journal of Electrical Engineering and Computer Science, vol. 15, no. 3, pp. 1615–1620, Sep. 2019, doi: 10.11591/ijeecs.v15.i3.pp1615-1620.
48. Mammen E, Rothe C, Schienle M (2012) Nonparametric regression with nonparametrically generated covariates.Ann. Statist.40(2):1132–1170
49. Merla M. and Beena T.L. A. (2020),Heart Disease Prediction using Machine Learning Techniques: A Survey, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 8 Issue X Oct 2020.
50. Miyato, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957.
51. Mohan S., Thirumalai Ch., Srivastava G, (2019) Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, DOI , 10.1109/ACCESS.2019.2923707, IEEE Access, VOLUME 7, 2019 S.P. Bingulac.
52. Nikhar S., Karandikar A.M.(2016) Prediction of Heart Disease Using Machine Learning Algorithms International Journal of Advanced Engineering, Management and Science (IJAEMS) Infogain Publication,[Vol-2, Issue-6, June-2016].
53. Pan & Yang (2010) A survey on transfer learning. IEEE transactions on knowledge and data engineering, 22(10), 1345-1359.
54. Patil, B.M., Joshi, R.C., Toshniwal, D. & Biradar, S., 2011. A new approach: role of data mining in prediction of survival of burn patients. Journal of Medical Systems. 35 (6), ss. 1531- 1542
55. R. Brachman, and T. Anand, "The Process of Knowledge Discovery in Databases: A HumanCentered Approach" Advances in Knowledge Discovery and Data Mining, ed. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, AAAI/MIT Press 1996.

56. Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.
57. Rajkumar, D.et al., "Machine Learning in Cardiology: Past, Present, and Future," in press (2019)
58. Rajpurkar et al., 2017. "Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks." Nature Biomedical Engineering.
59. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Lungren, M. P. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. PloS one, 12(3), e0172410.
60. Ramalingam V.V., Dandapath A., Raja M K., (2018) Heart disease prediction using machine learning techniques: a survey, International Journal of Engineering & Technology, 7 (2.8) (2018) 684-687.
61. Rosengren, A., Perk, J. & Dallongeville, J., 2009. Prevention of cardiovascular disease. ESC textbook of cardiovascular medicine. New York: Oxford University Press, ss.403-435.
62. Rothwell, P.M., Coull, A.J., Giles, M.F., Howard, S.C., Silver, L.E., Bull, L.M., Gutnikov, S.A., Edwards, P., Mant, D., Sackley, C.M., Farmer, A., Sandercock, P.A., Dennis, M.S., Warlow, C.P., Bamford, J.M. & Anslow, P., 2004. Change in stroke incidence, mortality, case-fatality, severity, and risk factors in Oxfordshire, UK from 1981 to 2004. Oxford Vascular Study. Lancet. 363(9425), ss.1925-1933
63. Shashank et al., 2019. "Comparison of gradient boosting and random forest algorithms for heart disease prediction." Journal of Medical Informatics.
64. Suresh, T. A. Assegie, S. Rajkumar, and N. K. Kumar, “A hybrid approach to medical decision-making: diagnosis of heart disease with machine-learning model,” International Journal of Electrical and Computer Engineering (IJECE), vol. 12, no. 2, pp. 1831–1838, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1831-1838.
65. Sutton & Barto (2018) Reinforcement learning: An introduction. MIT press.
66. Thom, T., Haase, N., Rosamond, W., Howard, V.J., Rumsfeld, J., Manolio, T., Zheng, Z.J., Flegal, K., O'Donnell, C., Kittner, S., Lloyd-Jones, D., Goff, D.C.Jr., Hong, Y., Adams, R., Friday, G., Furie, K., Gorelick, P., Kissela, B., Marler, J., Meigs, J., Roger, V., Sidney, S., Sorlie, P., Steinberger, J., Wasserthiel-Smoller, S., Wilson, M. & Wolf, P., 2006. American heart association statistics committee and stroke statistics subcommittee. Heart Disease and Stroke Statistics.
67. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.
68. Wang, J., Li, Y., Wu, J., & Chen, X. (2020). Mutual information based feature selection for high-dimensional data. Knowledge-Based Systems, 185, 105990.
69. Wang, X., et al. (2018). "Improving the Accuracy of Random Forest Models for Diabetes Prediction through Data Cleaning." Journal of Medical Analytics, vol. 5, pp. 1-15.
70. Yang, Y., Yu, Y., Gong, Y., & Huang, T. (2016). A review of feature selection techniques in bioinformatics. Briefings in bioinformatics, 17(6), 914-925.
71. Yu, L., Liu, H., & Zhang, Z. (2019). A survey of feature selection techniques. ACM Computing Surveys (CSUR), 51(6), 1-33.

72. Zhankg, H., Xu, T., Li, H., Zhang, S., & Huang, X. (2018). Self-attention generative adversarial networks. arXiv
73. Zheng, X., Lu, Y., Li, Y., & Hu, X. (2018). Cardiac MR image segmentation using deep convolutional neural networks with dense connections. Medical Image Analysis, 42, 1-12.
74. Zhu, R., Li, Y., Wu, J., & Chen, X. (2018). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(1), 283-298.
75. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.
76. Zreik et al (2019) "Deep Learning for Cardiac Image Analysis,"
77. https://sebastianraschka.com/faq/docs/evaluate-a-model..html
78. Cardiovascular diseases (CVDs)," WHO. (2016) https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
79. https://site.moh.ps/Content/Books/UeaXLRj7Zi2Cj9KBg6Ui92wWe21e4utZhNltHg7HH1KmMOePQW4K39_pS1ixH9cAYeQwc6xIOCOED1BJ4J5NhtCVVNKiSnlLVsQzpEWTqGyzk.pdf
80. Abuzir Y., Abuzir M. and Abuzir A., Using Artificial Neural Networks (ANN) To Detect the Diabetes, in COMMUNICATION & COGNITION (C&C) Journal, V53, N3-4 pp 103-122, (2020). Ghent, Belgium.

Appendix A

Dataset description

Arab American University
Faculty of Graduate Studies

الجامعة العربية الأمريكية
كلية الدراسات العليا

2022-5-15

الى من يهمه الأمر

تسهيل مهمة بحثية

تحية طيبة وبعد،

تهديكم كلية الدراسات العليا في الجامعة العربية الأمريكية أطيب التحيات، وبالإشارة الى الموضوع أعلاه، تشهد كلية الدراسات العليا في الجامعة أن الطالب **محمد فتح الله أحمد غانم** والذي يحمل الرقم الجامعي202012877 هو طالب ماجستير في برنامج علم البيانات وتحليل الأعمال ويعمل على رسالة الماجستير الخاصة به بعنوان:

"التنبؤ المبكر والدقيق بأمراض القلب باستخدام التعلم الآلي: اختيار الميزة "تحت اشراف الدكتور يوسف أبو زر، نأمل من حضرتكم الإيعاز لمن يلزم لمساعدته للحصول على المعلومات اللازمة للدراسة، علماً أن المعلومات ستستخدم لغاية البحث فقط وسيتم التعامل معها بغاية السرية، وقد أعطي هذه الرسالة بناءً على طلبه.

وتفضلوا بقبول فائق الاحترام

عميد كلية الدراسات العليا
د. نوار قطب

كلية الدراسات العليا
FACULTY OF GRADUATE STUDIES

Jenin Tel: +970-4-2418888 Ext.:1471,1472 Fax: +970-4-2510810 P.O. Box:240
Ramallah Tel: +970-2-2941999 Fax: +970-2-2941979 Abu Qash - Near Alrehan
E-mail: FGS@aaup.edu ; PGS@aaup.edu Website: www.aaup.edu

الملخص

كان الهدف من هذه الدراسة هو معرفة التنبؤ المبكر والدقيق بأمراض القلب باستخدام تقنيات التعلم الآلي وطرق اختيار الميزات. مثلت البيانات التي تم جمعها من مستشفى الرازي في مدينة جنين 890 سجلاً ، ولأن حجم العينة لم يكن كافياً للحصول على تنبؤ دقيق، تم زيادة حجم العينة باستخدام خوارزمية **GAN** مع الأخذ في الاعتبار أن أمراض القلب هي سبب مهم للوفيات في جميع أنحاء العالم، فإن التنبؤ المبكر والدقيق بالأمراض أمر بالغ الأهمية لنجاح الوقاية والعلاج. **وفي** هذه الدراسة، ولتحقيق هدف الدراسة تم تطوير تقنية قائمة على التعلم الآلي للتنبؤ بأمراض القلب باستخدام اختيار الميزات. **ولزيادة** حجم مجموعة البيانات، **تم استخدام** خوارزمية .GAN لتحسين حجم مجموعة التدريب، حيث تقوم خوارزمية GAN بإنشاء عينات تركيبية تضاف إلى مجموعة البيانات الأصلية. وفي الوقت نفسه، تم استخدام مناهج اختيار الميزات لتحديد السمات الأكثر أهمية للتنبؤ بأمراض القلب. ولإنجاز نتائج الدراسة، تم تطبيق أساليب التعلم الآلي، منها K-Nearest Neighbors (KNN) و Support Vector Machine (SVM) و Adaboost و Random Forest، على الميزات المحددة لإنشاء تنبؤات. أظهرت نتائج هذه الدراسة أن النهج المحدد يمكن أن يحقق دقة بنسبة 99 ٪ من خلال الجمع بين نماذج KNN و SVM ، وأن استخدام العينات التي تم إنشاؤها بواسطة GAN وأساليب اختيار الميزات يمكن أن يحسن أداء نماذج التعلم الآلي. **وكما** أظهرت هذه النتائج أن النهج المستند إلى مزيج من اختيار الميزات وخوارزميات التعلم الآلي، قد يكون أداة مفيدة للتنبؤ المبكر والدقيق بأمراض القلب.