Syrian Arab Republic
Ministry of Higher Education
Syrian Virtual University

الجمهوريّة العربيّة السوريّة
وزارة التعليم العالي
الجامعة الافتراضيّة السوريّة

Studying the cases of heart and arteries nutrition for cardiac patients in the Syrian community using bioinformatics tools

A thesis submitted as a fulfilment of requirements for Master's degree in Bioinformatics

By

Basma Alashoush/basma_160658

Supervised by

Dr. Yanal Ahmad Alkudsi

2023-2024

Contents

Table of Abbreviations:

Abbreviation	Meaning
AI	Artificial intelligence
CVD	Cardiovascular diseases
CAD	Coronary artery disease
CHD	Coronary heart disease
PAD	Peripheral artery disease
MI	Myocardial infarction
ML	Machine learning
WHO	World Health Organization
LDL	low-density lipoproteins
DL	Deep learning
SAGE	Serial analysis of gene expression
ESTs	Expressed sequences tags
MPSS	Massively parallel signature sequencing
SVM	Support vector machine
DT	Decision tree
RF	Random forest

Table of figures:

Table of tables:

Abstract:

Cardiovascular diseases (CVD) are one of the most causes of death worldwide. Although of many habits like smoking and comorbidities are considered as a risk factors for developing CVD, poor eating habits should be taken into consideration.

Bioinformatics tools provide powerful computational methods for analyzing CVD data. Therefore, the aim of this study was to develop a machine learning models to predict the deferent CVD to make the right decision in the protocol treatments of patients.

The dataset was collected from Al-WATANI hospital in Sweida, which included patients demographic, comorbidities, and dietary. The dataset was further split into training (60%) and test (40%) sets for building model and evaluating.

Our study included 183 patients of which 111 patients were with hypertension, 33 patients with Infarction, 20 patients with congestive heart failure, and 19 patients with arrhythmia. Moreover, the accuracy of the algorithms varied, with support vector machine achieving the lowest accuracy of 71.62%, while it increased remarkably to 91.74% when applying the balanced weights. We also found that decision tree and random forest (tree depth=5) achieved the same accuracies of 85.14%. However, when increasing the depth of the trees to 10, 15, or 20, the accuracy increased to 87.84% and remained steady.

These models demonstrated high accuracy and reliability, making them valuable tools for clinical decision-making.

1.1 Introduction:

Cardiovascular diseases are the leading cause of death worldwide, killing more than 17 million people annually, which exceeds global cancer mortality rates. Poor eating habits are a major risk factor. Despite these alarming statistics, most research and funding in bioinformatics and computational biology to date has mostly focused on cancer research, with a relatively modest footprint in cardiovascular disease. According to the World Health Organization, ischemic heart disease and stroke have remained the leading causes of death globally over the past 15 year.

Bioinformatics tools provide powerful computational methods for analyzing complex biological data, enabling researchers to uncover new insights into the molecular mechanisms underlying CVD. The programmatic need for bioinformatics measurement and awareness of state-of-the-art tools for conducting CVD research will align with multiple areas of expertise (Such as single-cell sequencing techniques, long-read mapping, 3D genome visualization, etc.). This makes cardiac informatics research a truly multidisciplinary initiative to dissect the molecular mechanisms underlying complex cardiovascular disease traits.

Understanding the relationship between nutrition and cardiovascular health is crucial for developing effective prevention strategies and personalized interventions. There is experimental, epidemiological, and clinical evidence demonstrating an association between diet and increased risk of cardiovascular disease (CVD). While nutritionally poor diets can have a significant negative impact on cardiovascular health, dietary interventions using specific nutrients or functional foods are effective components in terms of prevention. It is estimated that dietary factors may be responsible for approximately 40% of all cardiovascular diseases. In fact, in one of the seminal studies on modifiable risk factors and heart health (INTERHEART study), more than 90% of all myocardial infarction cases were attributed to preventable environmental factors with nutrition identified as an important determinant of CVD.

1.2. Research's question:

How can bioinformatic tools and techniques be utilized to improve the understanding of the relationship between nutrition and heart health in cardiac patients.

1.3. Problems:

Cardiovascular diseases are the leading cause of death worldwide, and the limited understanding of how specific dietary components affect cardiovascular health is one of the biggest problems facing patients as well as the lack of comprehensive knowledge regarding the molecular pathways involved in the development of cardiovascular diseases. Identifying potential biomarkers or genetic variants associated with the risk of cardiovascular disease is one of the challenges facing the health sector, but it aims to detect the disease and diagnose the condition. As well as providing practical nutritional recommendations for heart patients.

The field of cardiac informatics is still in its early days and with significant opportunities to leverage cutting-edge data science techniques and machine learning (ML) methodologies, cardiac bioinformatics is best positioned to address domain-specific research questions by developing clinical applications to enhance computationally intensive tasks such as Those in medical imaging, cardiovascular disease risk prediction modeling, are among other active research areas. For example, current methods for imaging CVD calcification are mostly limited to advanced calcification and miss clinically relevant early microcalcifications, creating an unmet need for the implementation of advanced imaging tools and artificial intelligence to improve diagnosis and risk assessment.

Theoretical review

1.4. Cardiovascular System

The cardiovascular system consists of the heart, blood vessels, and blood. Its primary function is to transport nutrients and oxygen-rich blood to all parts of the body and to carry deoxygenated blood back to the lungs. Abnormalities or injuries to any or all parts of the cardiovascular system can result in serious health complications. Common conditions that can affect the cardiovascular system include coronary artery disease, heart attack, high blood pressure, and stroke.[1]

It consists of the following organs and tissues:

The heart: A muscular pump that forces blood around the body.

A closed system of blood vessels: These vessels include:

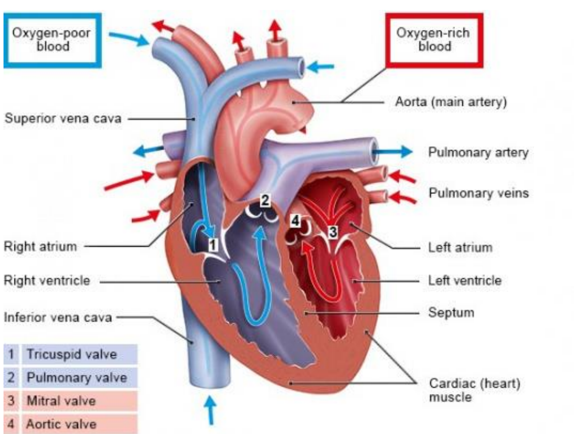
Arteries: Vessels that carry blood away from the heart.

Veins: Vessels that bring blood back to the heart.

Capillaries: Tiny vessels that branch off from arteries to deliver blood to all body tissues Trusted Source [1].

1.5. How the Heart Works:

The heart is a hollow muscle. A wall through the middle (known as the septum) divides it into two halves. Each half has two chambers called the atrium and ventricle. The left ventricle pumps oxygen-rich blood out of the heart and into the body (systemic circulation) through an artery called the aorta. The first blood vessels that branch off from the aorta are the coronary arteries. There are 4 heart valves between the right atrium and right ventricle (Tricuspid valve), the left atrium and left ventricle (Mitral valve), and where the blood leaves the heart through the arteries (Pulmonary valve, Aortic valve). They ensure that the blood flows in the right direction and doesn't flow back. Put simply, the valves of the heart function like one-way gates [2]



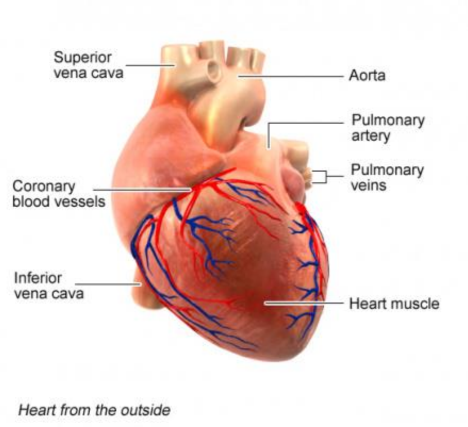


Figure (1): The flow of blood in the heart

Figure (2): The heart from outside

1.6. Cardiovascular Disease (CVD):

The cardiovascular system consists of the heart and blood vessels.[3].A wide array of problems can arise within the cardiovascular system, a few of which include endocarditis, rheumatic heart disease, and conduction system abnormalities.

Cardiovascular disease (CVD) or heart disease includes diseases like:[4,5,6]

1.6.1. Coronary artery disease (CAD):

Sometimes referred to as Coronary Heart Disease (CHD), is characterized by atherosclerosis in coronary arteries and can be asymptomatic results from decreased myocardial perfusion that causes angina, myocardial infarction (MI), and/or heart failure. It accounts for one-third to one-half of the cases of CVD.

1.6.2. Coronary heart disease (CHD):

Is a slowly developing chronic disease that mainly results from a progressive narrowing of blood vessels that supply the myocardium with oxygenated blood, giving rise to ischemia at times of increased oxygen demands .As clinical result includes an inadequate ejection of blood from the heart (heart failure), irregular cardiac rhythms (arrhythmias), or acute coronary syndromes (ACSs) such as myocardial infarctions and unstable angina, which are often followed by sudden cardiac death.

1.6.3. Peripheral artery disease (PAD):

Particularly arterial disease involving the limbs that may result in claudication.

1.6.4. Aortic atherosclerosis:

Including thoracic and abdominal aneurysms.it occurs when plaque builds up and hardens inside your arteries. Plaque consists of cholesterol, calcium and other fatty substances which can partially or completely block blood flow to your heart.[7]

1.6.5. Myocardial infarction (MI):

Is a myocardial injury and necrosis due to formation of plaques in the interior walls of the arteries resulting in reduced blood flow to the heart and injuring heart muscles because of lack of oxygen supply.[8]

1.7. Prevalence of CVD:

According to the World Health Organization (WHO) Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. The incidence of CAD is observed to rise with age, regardless of gender. In the ONACI registry in France, the incidence of CAD was about 1% in the 45 to 65 age group, which increased to about 4% as the age group reached 75 to 84 years.[9]

In 2017 to March 2020 data, the prevalence of CVD (comprising CHD, HF, stroke, and hypertension) in adults ≥20 years of age was 48.6% overall (127.9 million in 2020) and increases with age in both males and females. HD and stroke currently claim more lives each year than cancer and chronic lower respiratory disease combined. In 2020, 207.1 of 100 000 people died of HD and stroke basis on 2020 mortality data,

In 2020, 19.05 million deaths were estimated for CVD globally, which amounted to an increase of 18.71% from 2010. The age-standardized death rate per 100 000 population was 239.80, which represents a decrease of 12.19% from 2010. Overall, the crude prevalence of CVD was 607.64 million cases in 2020, an increase of 29.01% compared with 2010. However, the age-standardized prevalence rate was 7354.05 per 100 000, an increase of 0.73% from 2010.[10]

1.8. Blood vessels

Blood vessel are composed of three layers. The outer layer is mainly connective tissue that gives structure to the vessels. The middle layer is smooth muscle that contracts and dilates to control blood flow and blood pressure. The inner lining is a thin layer of endothelial that in a healthy state is smooth and responsive.[11]

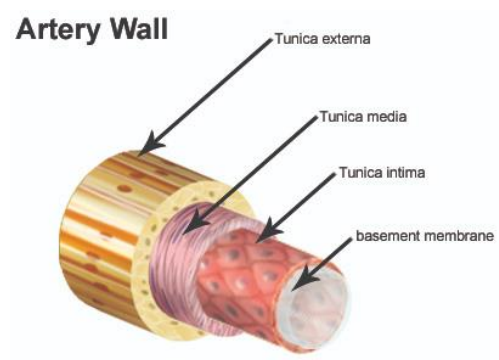


Figure (3): Artery wall

1.9. Pathophysiology of CAD:

Atherosclerosis is a condition that affects the arteries and the aorta, and can lead to diseases due to reduced or blocked blood flow from narrowing of the blood vessels. Atherosclerosis is caused by a low-grade inflammation of the intima (inner lining) of medium-sized arteries, which is worsened by the common risk factors such as high blood pressure, high cholesterol, smoking, diabetes, and genetics. In coronary atherosclerosis, the inner layer of the coronary arteries becomes thicker over time, which can eventually shrink the space inside the artery to different extents. Atherosclerosis can cause acute syndromes such as AMI and SCD, especially in the proximal parts of the main coronary arteries where the artery splits and the flow changes [12].

Atherosclerosis leads to the formation of fatty streaks, low-density lipoproteins (LDLs) under the endothelial layer of the vessel wall, and an inflammatory response that involves macrophages, foam cells, and fibrous plaques [13][14].

As the inflammation progresses, the deeper layers can die, triggering more macrophage recruitment that can harden and turn into atherosclerotic plaques.[15] Chronic heart disease such as Ischemia happens when a plaque grows enough to interfere with the blood supply to the tissues (usually >70% stenosis). Atherosclerosis plaque can remain stable and cause narrowing in the arteries, or it can rupture and cause a blood clot. This is the main reason for the high rates of death and disability from atherosclerosis, such as MI, stroke [15].

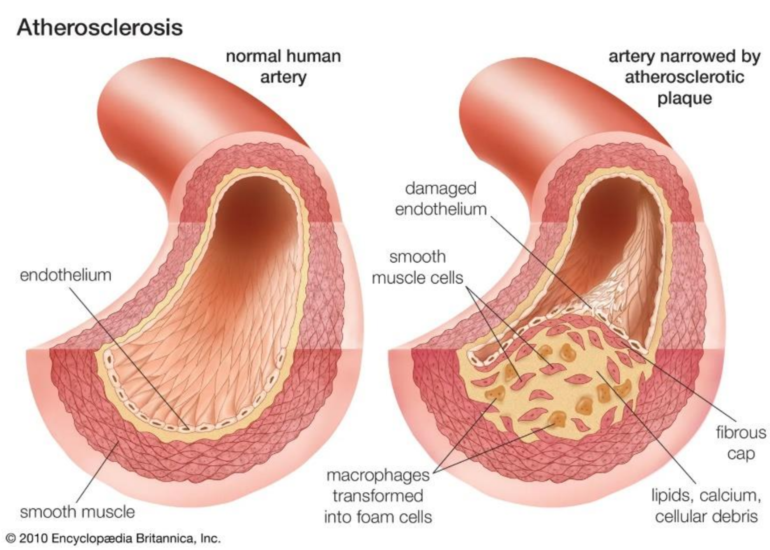


Figure (4): A normal artery compared with an artery affected by atherosclerosis.[16]

1.10. Risk Factors of Heart disease:

There are several factors that increase the risk of heart disease like family history of cardiovascular disease, smoking, stress, obesity, nutrition, and high blood cholesterol levels, particularly in association with LDLs, and Socioeconomic status

Men develop atherosclerosis more often than women, and individuals with diabetes mellitus have a significantly higher incidence of the disease.[17,18]

1.11. Therapeutics:

Atherosclerosis poses significant risks to health, but these can be mitigated through the use of certain medications. Statins are one example, as they have the ability to lower cholesterol and fat levels in the bloodstream. Other drugs, like β-Blockers administration to CAD patients was found to reduce heart rate and metabolic requirements to the myocardium, thereby preventing ischemia, and Calcium antagonists are drugs used for coronary vasospasm to gain relief from symptoms only, in addition to β-blockers therapy, also including anticoagulants and aspirin These drugs could alter platelet function by inhibiting the activity of an enzyme necessary for the production of prostaglandins, known as cycloxygenase in order to prevent blood clots from forming.

In some cases, surgical intervention is a procedure used during a heart attack to quickly open a blocked artery and reduce the amount of damage to the heart may be necessary to remove blockages caused by atheroma in large arteries such as the aorta or carotids. Synthetic materials can be used to replace the obstructed sections. Alternatively, atherectomy may be performed to remove fatty deposits from the carotid circulation. This procedure involves using a catheter to insert a tiny knife into the vessel and carefully remove the plaques.[19]

1.11. The role of diet and nutrition in Cardiovascular Disease (CVD):

Nutrition is one of the major modifiable risk factors for CVD, which is the leading cause of death and disability worldwide. Dietary patterns, foods, and nutrients can influence various aspects of CVD, such as blood pressure, cholesterol levels, inflammation, oxidative stress, endothelial function, thrombosis, and cardiac arrhythmias. Therefore, understanding the effects of nutrition on CVD is essential for developing effective prevention and treatment strategies.

A report from the Global Burden of Disease Study stated that an estimated 22.4% of all male deaths and 20.7% of all female deaths in 2015 were attributable to poor dietary factors [20]. Between 2002 and 2012, diet-associated cardiometabolic death rates decreased for polyunsaturated fats (−20.8%), nuts and seeds (−18.0%), and sugar-sweetened beverages (−14.5%) but increased for sodium (5.8%) and unprocessed red meats (14.4%).[21]

Nutrition can affect the occurrence and development of atherosclerosis by influencing the levels and oxidation of lipoproteins, the inflammatory response, and the endothelial function of the arterial wall.Some nutrients, such as phytosterols, omega-3-polyunsaturated fatty acids, polyphenols, vitamins, and fiber, have been shown to have beneficial effects on atherosclerosis by lowering cholesterol, improving lipid profile, reducing oxidative stress, and modulating inflammation [22]. Other nutrients, such as saturated fat, trans-fat, sodium, and sugar, have been shown to have detrimental effects on atherosclerosis by increasing cholesterol, blood pressure, glycemic load, and inflammation [23].

Nutrition can also affect the risk and prognosis of heart failure by influencing the cardiac structure and function, the neurohormonal activation, the oxidative stress, and the metabolic status of the heart. Some dietary patterns, such as the Mediterranean diet, the DASH diet, and the plant-based diet, have been associated with lower risk and better outcomes of heart failure by providing antioxidants, anti-inflammatory agents, and cardioprotective nutrients. Other dietary patterns, such as the Western diet, the high-sodium diet, and the high-glycemic diet, have been associated with higher risk and worse outcomes of heart failure by inducing obesity, hypertension, diabetes, and cardiac remodeling [24,25].

The mortality of heart failure patients affected by Nutritional status influencing the energy balance, the muscle mass, the immune function, and the quality of life. Malnutrition can lead to a significant decrease in life expectancy in heart failure patients by causing cachexia, sarcopenia, infection, and depression. Obesity can lead to a significant increase in life expectancy in heart failure patients by providing a metabolic reserve, a hemodynamic advantage, and a hormonal protection. However, the association between obesity and malnutrition, also known as the obesity paradox, is complex and controversial, and may depend on the severity and duration of heart failure [22].

1.12. Dietary Patterns and CVD:

Dietary patterns are the combinations and quantities of foods and nutrients consumed over time. They reflect the overall quality of the diet and its impact on health outcomes. Several dietary patterns have been associated with lower CVD risk, such as the Mediterranean diet, the Dietary Approaches to Stop Hypertension (DASH) diet, the Portfolio diet, and the vegetarian or vegan diet. These dietary patterns share some common features, such as high intake of fruits, vegetables, whole grains, nuts, seeds, legumes, and plant oils, and low intake of red and processed meats, refined grains, added sugars, and sodium. They also provide adequate amounts of fiber, antioxidants, phytochemicals, omega-3 fatty acids, and other beneficial nutrients that may protect against CVD.

A meta-analysis of 56 prospective studies involving 4.3 million participants and 81,964 CVD events found that adherence to a healthy dietary pattern was associated with a 21% lower risk of CVD, a 25% lower risk of coronary heart disease (CHD), and a 14% lower risk of stroke. Another meta-analysis of 32 prospective studies involving 1.9 million participants and 134,627 CVD events found that adherence to a Mediterranean diet was associated with a 10% lower risk of CVD, a 14% lower risk of CHD, and a 6% lower risk of stroke. A systematic review and meta-analysis of 13 randomized controlled trials (RCTs) involving 2,528 participants found that the DASH diet lowered systolic blood pressure by 5.5 mm Hg and diastolic blood pressure by 3.0 mm Hg compared with the control diet. A meta-analysis of 10 RCTs involving 1,159 participants found that the Portfolio diet lowered low-density lipoprotein (LDL) cholesterol by 17% compared with the control diet. A meta-analysis of 96 observational studies and 44 RCTs involving 1.6 million participants and 83,867 CVD events found that vegetarian and vegan diets were associated with lower blood pressure, lower LDL cholesterol, lower body mass index (BMI), and lower CVD risk compared with omnivorous diets.[26]

1.13. Bioinformatics:

bioinformatics targets to develop methodology and analysis tools to explore large volumes of biological data, helping to store, organize, systematize, annotate, visualize, query, mine, understand, and interpret complex data volumes. It uses conventional, modern computer science and cloud computing, statistics, and mathematics, as well as pattern recognition, reconstruction, machine learning, simulation and iterative approaches, and molecular modeling/folding algorithms.The emergence and advances of the bioinformatics field, however, are tightly associated with the computerized programming and software developments needed for the handling and structural and functional analysis of large volumes of molecular sequences of DNA, RNA, proteins, and metabolites. [27,28]

Bioinformatics tools are very important to analyze gene and protein expression profiles. Large-scale sequencing of cDNA libraries has generated large volumes of serial analysis of gene expression (SAGE), expressed sequences tags (ESTs), massively parallel signature sequencing (MPSS), transcriptome profiling, or RNA-Seq, and various applications of multiplexed in-situ hybridization (microarray) profile data.[29]

1.14. Cardioinformatics:

it is a term that refers to the nexus of bioinformatics and precision cardiology, which aims to integrate and analyze various types of omics data, such as genomics, transcriptomics, proteomics, metabolomics, and epigenomics, to understand the molecular mechanisms, biomarkers, and therapeutic targets of cardiovascular disease. Cardioinformatics also involves the use of machine learning, artificial intelligence, and big data techniques to discover novel insights and patterns from complex and heterogeneous data sources.[30]

1.15. Artificial Intelligence (AI)

Artificial intelligence (AI) was a groundbreaking innovation for humankind, which created a new realm of possibilities. Alan Turing first defined artificial intelligence in 1950 [31]. AI is achieved by examining how the human brain functions, and how humans acquire, decide, and act while tackling a problem, and then using the findings of this examination as a foundation for creating smart software and systems [32].In this AI technique, computers learn from past experiences and data. The volume of data is growing fast, so there is a demand to manage the data effectively. Sometimes, it is very challenging for humans to manually obtain valuable information from raw data because of their variability and ambiguity. This is where machine learning comes in handy. The aim is to discover hidden patterns in the data and forecast new data. AI can use very intricate nonparametric models from a huge amount of data, unlike simple parametric models that need a suitable-sized data set used in statistics [34]. AI can enhance various aspects of patient care, such as medical imaging quality, early diagnosis, prognosis prediction, risk stratification, patient data analysis, personalized treatments and more [31]. AI has many subfields, such as machine learning (ML), deep learning (DL), and computer vision.

1.16. Deep learning (DL):

DL is one of the amazing developments in machine learning. It was coined in 2006 and based on the human brain's neural networks. It is a data-processing approach that employs multiple layers. The layers can be thought of as receiving weighted input, applying mostly nonlinear functions to it, and passing the output to the next layer.[35]

1.17. Machine learning (ML):

ML is one of the ways or means to achieve artificial intelligence, where a computer model is able to acquire new abilities and knowledge to perform useful tasks. The main idea of ML is to learn from data in order to make predictions or decisions based on the given task. After the ML algorithm is trained with data, the ML model will receive an input. The output will be a predictive model, based on the data that trained the model [35,36]. 14 ML concentrates on the learning by creating algorithms that can be explicitly coded using known features. In ML, there are four widely used learning methods, each suitable for solving different tasks: supervised, unsupervised, semi supervised, and reinforcement learning.

● Supervised ML: The most common supervised learning tasks are classification (determining the group a new measurement belongs to) and regression (estimating a continuous value of a new observation). It is helpful if the task at hand requires the input data to be classified into predefined classes or make predictions. The basic steps of supervised machine learning are:[35,33,31]

1. obtain a dataset and divide it into separate training, and test datasets
2. use the training datasets to teach a model the relationship between features and target
3. assess the model via the test dataset to see how well it predicts.

● Unsupervised learning: aims to find patterns in a dataset and assign individual instances in the dataset to those categories. Some of the most common unsupervised learning tasks are clustering, association, and anomaly detection [34].

● Semi Supervised Learning: Semi supervised learning can be considered as the "middle ground" between supervised and unsupervised learning and is especially useful for datasets that have both labeled and unlabeled data

● Reinforcement learning: is the method of training an algorithm for a specific task where there is no single correct answer, but a desired outcome [31].

1.18. Description for algorithms

1.18.1. support vector machine (SVM)

It is used for classification and regression; it is a popular ML approach. SVM was introduced by Vapnik in the late twentieth century. For unlabeled data, supervised ML algorithms are unable to perform. Using a hyperplane to find the clustering among the data, SVM can categorize unlabeled data. However, SVM output is not nonlinearly separable.
To overcome such problems, selecting appropriate kernel and parameters is two key factors when applying SVM in data analysis.

1.18.2. Logistic Regression:

Logistic Regression: is a classification algorithm where the goal is to find a relationship between features and the probability of a particular outcome.
It is the best regression analysis to use when the dependent variable or response variable is binary. It uses a sigmoidal curve to estimate class probability.

It works by combining the input variable (X) in a linear form and using coefficients to predict an output variable (Y) which is a binary value of 0 or 1. The logistic regression technique models the chance of an outcome based on the individual characteristics or input variables (X). It is represented mathematically as follows:

$$\log_{10}\frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n,$$

where π indicates the probability of an event, β represents estimated parameter values or regression coefficients associated with the variables via maximum likelihood estimation, and x indicates the parameter variables.

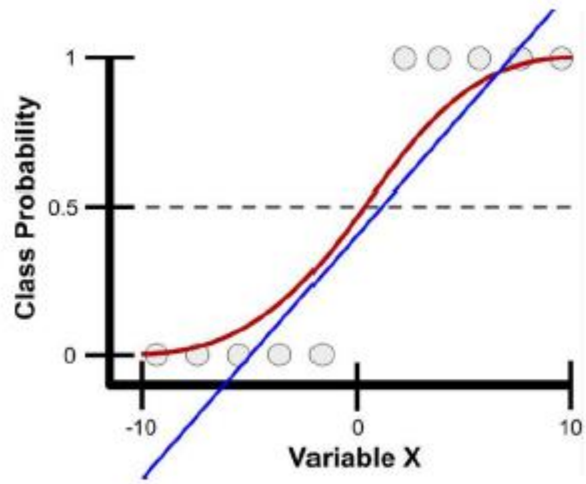


Figure (5) : illustrate logistic regression

1.18.3. Decision Trees

A decision tree is a supervised learning technique, primarily used for classification tasks. In DT models, the attribute may take on various values known as classification trees; leaves indicate distinct classes, whereas branches reflect the combination of characteristics that result in those class labels. It begins with a root node, the first decision point for splitting the dataset, and contains a single feature that best splits the data into their respective classes. Each split has an edge that connects either to a new decision node that contains another feature to further split the data into homogenous groups[36].

1.18.4. Random forest

Random forest is a supervised machine learning algorithm that constructs several decision trees. The final decision is made based on the majority of the decision tree. It suffers from low bias and high variance. RF converts high variance to low variance, Random forest is an extension of this DT, known as an ensemble method, that produces multiple decision trees. Rather than using every feature to create every decision tree in a random forest, a subsample of features is used to create each decision tree. Trees then predict a class outcome, and the majority vote among trees is used as the model's final class prediction [37].

1.19. Performance Evaluation

To maximize the chance of generalizability to the performance of the algorithm on unseen data, the training dataset is usually split into a slightly smaller training dataset and a separate validation dataset. Performance of a learned model can be evaluated in a number 21
of ways, but is most commonly evaluated based on prediction accuracy (classification) or error (regression)
Model performance is monitored via some form of accuracy on the training and testing datasets during. So long as the accuracy of the model on the training set (X%) and validation set (Y%) are increasing and converging after each training iteration, the model is considered to be learning

1.19.1. Accuracy (Acc)

The accuracy denotes total correctly identifying instances among all of the instances. as eq:

$$ACC = \frac{T_p + T_N}{T_p + T_N + F_p + F_N}$$

Eq1: $$\textbf{TPR /Recall / Sensitivity} = \frac{\textbf{TP}}{\textbf{TP + FN}}$$

Precision is measured as the proportion of precisely predicted to all expected positive observations.
Eq2:

$$\textbf{\textit{Precision}} = \frac{\text{TP}}{\text{TP + FP}}$$

False Positive Rate(FPR): defined as

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$FPR = 1 - \text{Specificity}$$

$$= \frac{FP}{TN + FP}$$

$$\textit{F - measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

1.19.2. Recall

Represents the rate of values that measures positive records that the classifier correctly predicted. Moreover, it is called true positive rate (TPR) or sensitivity. Thus, recall is calculated as shown in Eq1

1.19.3. Confusion Matrix

It is a summary of predictions that are correct and incorrect per class. It shows the ways in which your classification model is confused when it makes predictions.

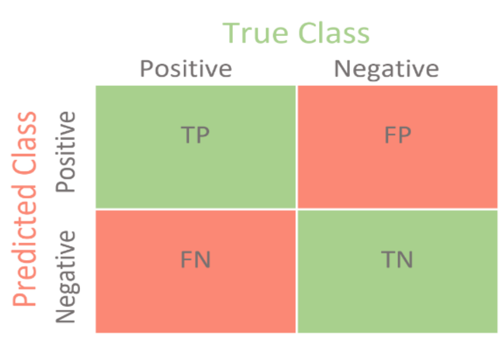


Figure (6): box of confusion matrix

1.20. AI and Cardiology

The creation of new medical devices enables the acquisition of new knowledge in the area of disease diagnosis. One of the most effective ways to rapidly diagnose diseases is to use computer-aided decision making, i.e., machine learning to derive knowledge from data. Progress in diagnostic applications of artificial intelligence (AI) for heart diseases.

Material and methods

2. Material and methods:

2.1. Study sample:

Our dataset was collected from Al-WATANI hospital in Sweida, which included patients demographic (age and gender), comorbidities (diabetes, high lipids and cholesterol, previous heart surgery, and familial heart disease), dietary habits (eating salt, and eating vegetables and fruits), and other information as shown in table 1.

The final dataset in the (.xlsx) file consists of 183 samples with 13 features (11 nominal features and 2 numerical features) and a single output variable (1= hypertension, 2= Infarction, 3= congestive heart failure, 4= arrhythmia) (table 1).

Table 1: Features in the dataset.

Features	Description	
Gender	Nominal	0= female and 1= male
Age	Nominal	0= age between 30-40 years old 1= age between 40-50 years old 2= age between 50-60 years old 3= age between 60-70 years old 4= age between 70-80 years old 5= >80 years old
Weight	Nominal	0= 30-40 Kg 1= 40-50 Kg 2= 50-60 Kg 3= 60-70 Kg 4= 70-80 Kg 5= >80 Kg
Systolic_pressure	Numerical	
Diastolic_pressure	Numerical	
Smoking	Nominal	0= I don’t smoke

		1= less than one packet 2= up to two packets 3= 3 packets or more
Diabetes	Nominal	0= I don't have diabetes 1= I have diabetes
Eating_salt	Nominal	0= I don't eat salt 1= light 2= moderate 3= a lot
Eating_vegetables_and_fruits	Nominal	1= moderate 2= a lot
Familial_heart_disease	Nominal	0= no 1= yes 3= maybe
Previous_heart_surgery	Nominal	0= no 1= yes
High_lipids_and_cholesterol	Nominal	0= no 1= yes
Diagnosis	Nominal	1= hypertension 2= Infarction 3= congestive heart failure 4= arrhythmia

2.2. Practical study and results:

We employed Google Colab for writing scientific codes in Python, to develop various models. so we worked with the data that had been processed in the first step.

Before working we imported many libraries using the following codes:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

There weren't any missing values and to make sure we used the code:

```
df.iloc[:,:].isna().sum()
```

and the result was as follow:

```
gender                          0
age                             0
weight                          0
Systolic_pressure               0
Diastolic_pressure              0
Smoking                         0
Diabetes                        0
Eating_salt                     0
Eating_vegetables_and_fruits    0
Familial_heart_disease          0
Previous_heart_surgery          0
High_lipids_and_cholesterol     0
Diagnosis                       0
dtype: int64
```

Then we run some descriptive codes and chart codes to take an insight to the data features:

We used the "value_counts()" code to count variables and "countplot()" code to draw charts.

There were 101 females and 82 males as shown in the result:

```
0      101
1       82
Name: gender, dtype: int64
```

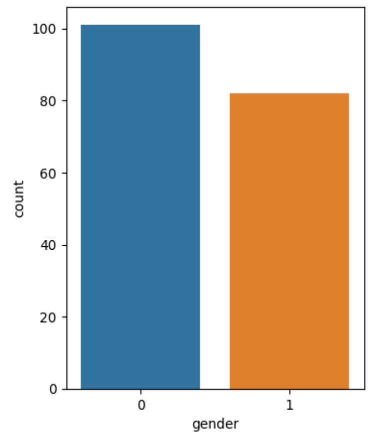


Figure (7): Gender distribution in the study sample.

The largest number of patients were among the age group (50-60 years old) (51 patients), followed by the age group (60-70 years old) (43 patients), as shown in the following result:

```
2     51
3     43
1     28
4     27
0     25
5      9
Name: age, dtype: int64
```

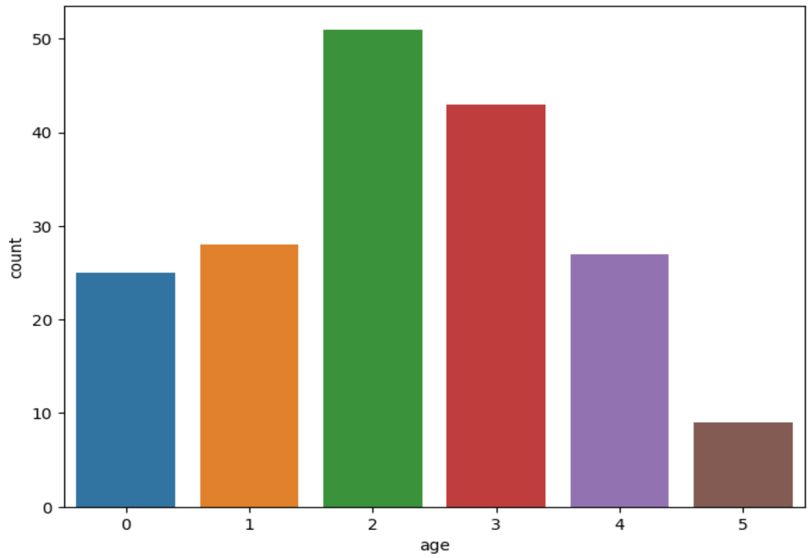


Figure (8): Age distribution in the study sample.

As for numerical variable we characterized it using mean, standard deviation, and other characterization criteria as shown in the following table.

Table (2): Pressure distribution in the study sample.

	Systolic_pressure	Diastolic_pressure
count	183.000000	183.000000
mean	13.595628	9.387978
std	1.579533	1.184732
min	12.000000	8.000000
25%	12.000000	8.000000
50%	13.000000	9.000000
75%	14.000000	10.000000
max	18.000000	13.000000

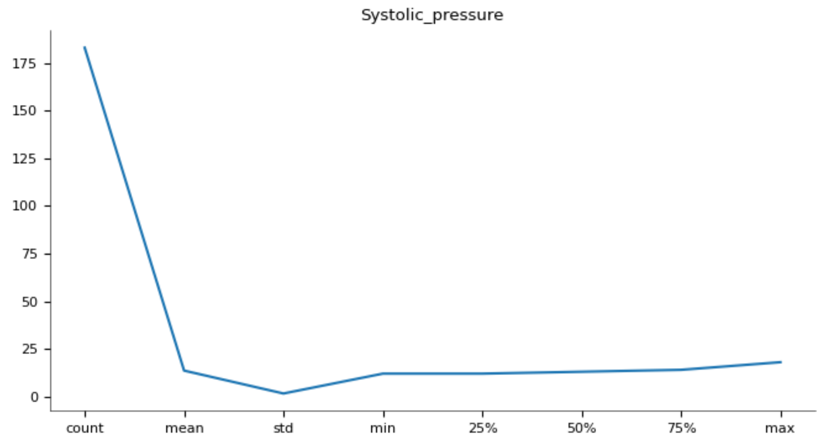


Figure (9): Systolic pressure distribution in the study sample.

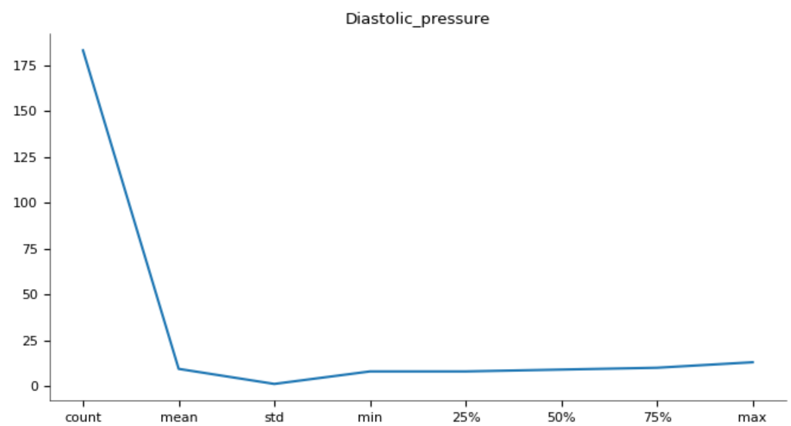


Figure (10): Diastolic pressure distribution in the study sample.

As for the outcome which is the diagnosis, there were 111 patients with hypertension, 33 patients with Infarction, 20 patients with congestive heart failure, and 19 patients with arrhythmia.

```
1    111
2     33
3     20
4     19
Name: Diagnosis, dtype: int64
```

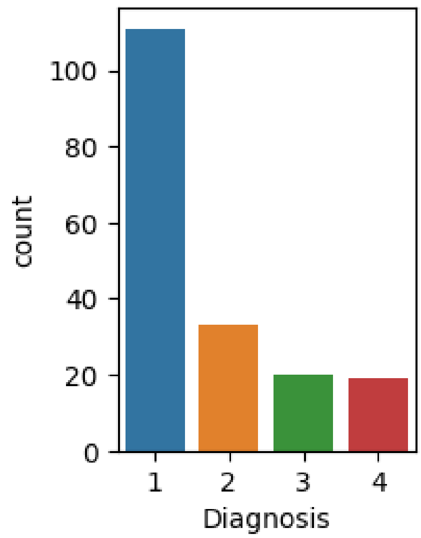


Figure (11): The outcome distribution in the study

As for dietary habits, we visualize the relationship between eating salt and the outcome by drawing the following chart:

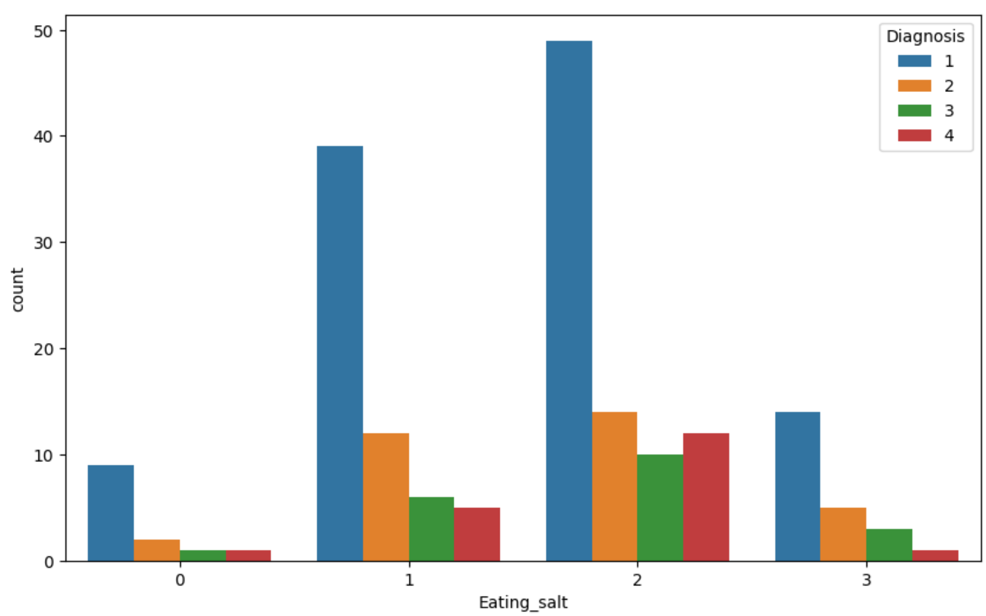


Figure (12): The relationship between eating salt and the outcome.

It can be clearly seen from the chart that the group that did not consume salt was the least susceptible to the four diseases, while the other groups that consumed salt, whether in low, medium or high quantities, were more susceptible to all diseases, most notably high blood pressure.

This is consistent with the fact that salt intake is a primary cause of high blood pressure, while other diseases can be a subsequent chronic consequence of high blood pressure. This explains why high blood pressure is the dominant disease among those who eat salt.

To visualize the relationship between eating vegetables and fruits and the outcome we drew the following chart:

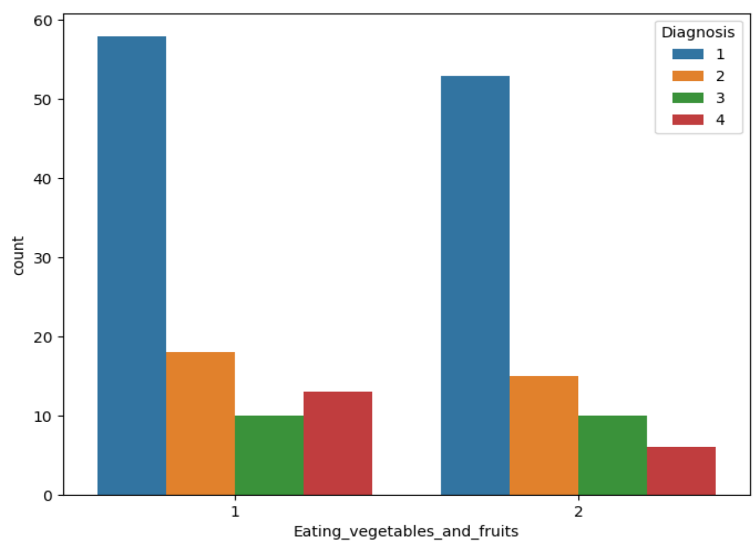


Figure (13): The relationship between eating vegetables and fruits and the outcome.

There was no significant difference that could be observed from the chart between the two groups, and this may be due to the presence of other dietary habits that must be taken into consideration, such as the amount of food rich in cholesterol.

After that, a heat map was generated to illustrate the correlation among the study variables.

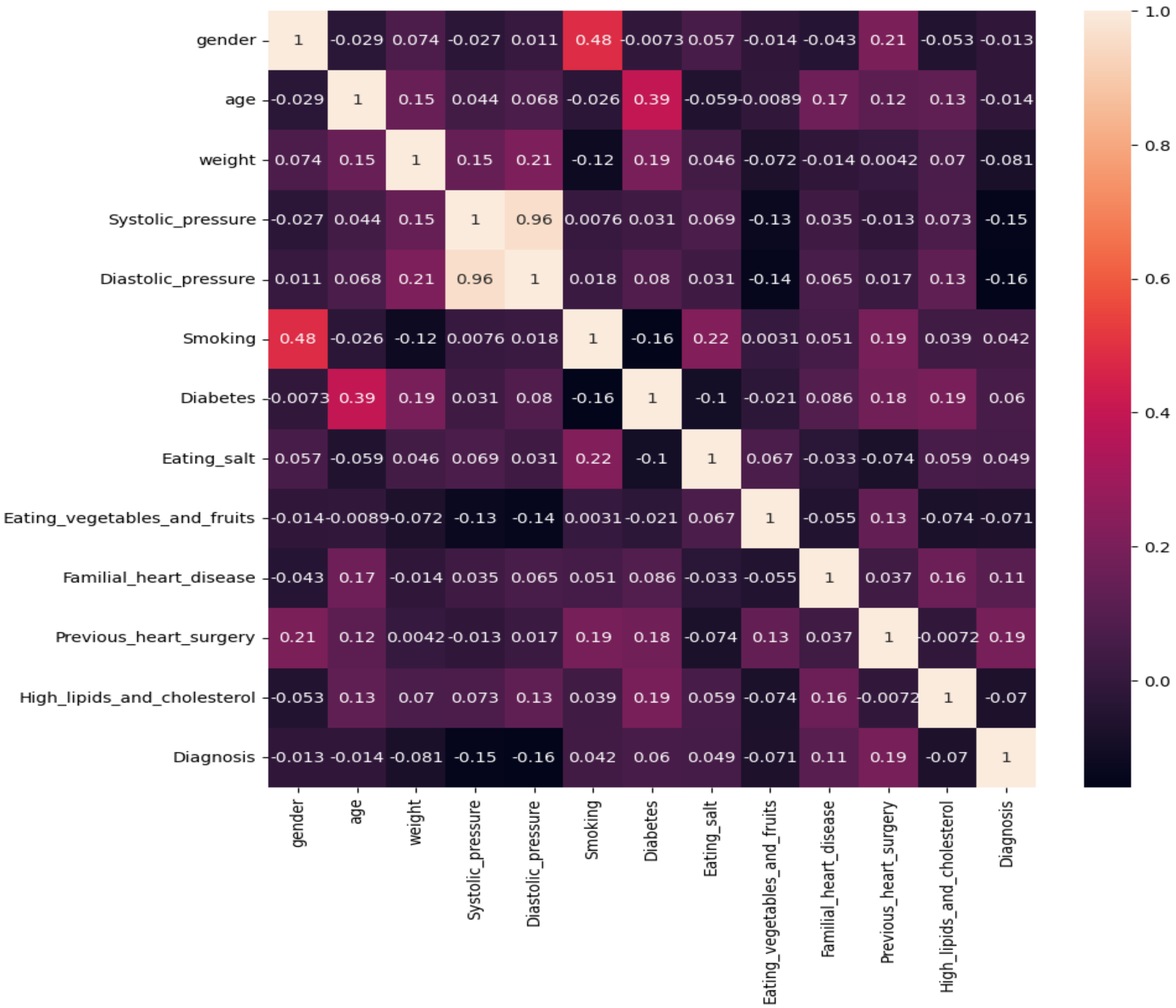


Figure (14): Heat map showing the correlation among the study variables.

As we can see there was a strong correlation between diastolic and systolic pressure. Therefore, we deleted the “diastolic_pressure” feature to prevent the algorithm from being biased by this variable. For deleting the variable we used the “del()”code.

We also performed data standardization, which is a technique used to ensure that variables with different scales do not disproportionately influence the results. By standardizing the data, all variables were transformed to a common scale with a mean of 0 and a standard deviation of 1. For standardization we used the “StandardScaler()”code.

Data was divided into a training set comprising 60% of the total data, and a test set comprising 40% of the total data using the following code:

```
X_train,X_test, y_train,y_test= train_test_split(X,y,test_size=0.40,
random_state= 0)
```

For building the program we used the following algorithms: Support vector machine (SVM), decision tree (DT), and random forest (RF).

To evaluate algorithms performing we used: accuracy, precision, recall, F1-score, and confusion matrix (figure 13).

Accuracy: the base metric used for model evaluation is often *Accuracy*, describing the number of correct predictions over all predictions.

Precision: is a measure of how many of the positive predictions made are correct (true positives TP).

Recall / Sensitivity: is a measure of how many of the positive cases the classifier correctly predicted, over all the positive cases in the data.

F1-Score: is a measure combining both precision and recall. It is generally described as the harmonic mean of the two.

Confusion Matrix: is sometimes used to illustrate classifier performance based on the above four values (TP, FP, TN, FN). These are plotted against each other to show a confusion matrix.

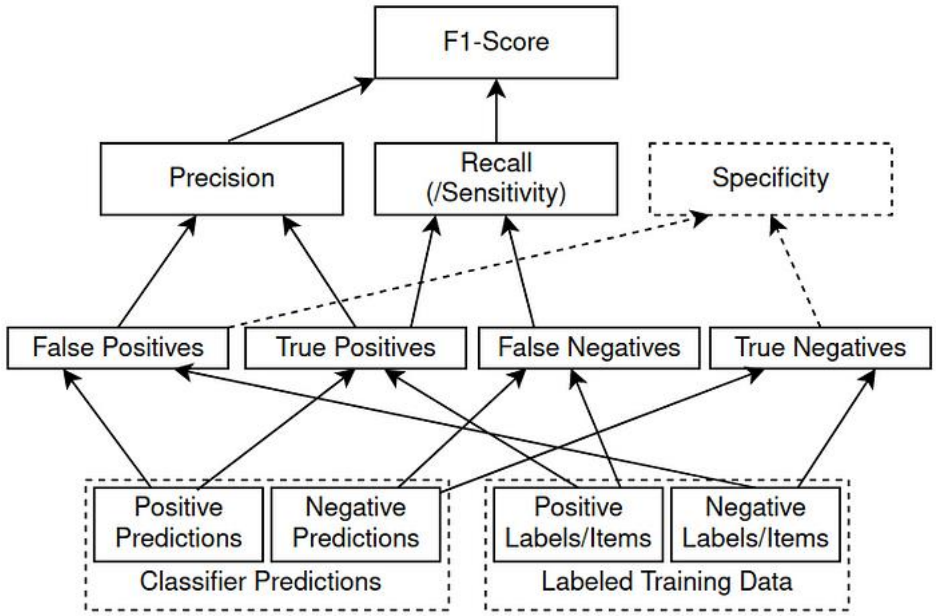


Figure (15): Evaluating criteria used in the study.

2.2.1. Support vector machine (SVM) results:

SVM achieved a very good accuracy of 71.62%.

Analyzing the confusion matrices provides valuable insights into the classification performance of the algorithms and helps identify any specific areas of improvement or errors in predictions.

It was observed that there were a few misclassifications, the algorithm correctly predicted 53 patients correctly (42 patients with hypertension, 7 patients with infarction, 1 patient with congestive heart failure, and 3 patients with arrhythmia). However, it made incorrect predictions for:

4 patients classifying them as hypertension when they are actually with infarction.

8 patients classifying them as with infarction when they are actually with congestive heart failure.

6 patients classifying them as with infarction when they are actually with arrhythmia.

2 patients classifying them as with congestive heart failure when they are actually with arrhythmia.

1 patient classifying him as with infarction when he is actually with hypertension.

```
[[42  1  0  0]
 [ 4  7  0  0]
 [ 0  8  1  0]
 [ 0  6  2  3]]
                 accuracy: {} 0.7162162162162162

                 classification report
----------------------------------------------------------
               precision    recall  f1-score   support

           1       0.91      0.98      0.94        43
           2       0.32      0.64      0.42        11
           3       0.33      0.11      0.17         9
           4       1.00      0.27      0.43        11

    accuracy                           0.72        74
   macro avg       0.64      0.50      0.49        74
weighted avg       0.77      0.72      0.70        74
```

2.2.2. SVM with balanced weight data:

Which is a technique that is commonly used to fine-tune hyperparameters and obtain a more robust evaluation of the model's performance.

Interestingly, the accuracy of SVM increased to reach 91.74%, which reflects the strong performance of this algorithm.

Backing to the confusion matrix, it was observed that there was a fewer misclassification, the algorithm correctly predicted 64 patients correctly (43 patients with hypertension, 11 patients with infarction, 8 patients with congestive heart failure, and 2 patients with arrhythmia). However, it made incorrect predictions for:

1 patient classifying him as with infarction when he is actually with congestive heart failure.

9 patients classifying them as with congestive heart failure, when they are actually with arrhythmia.

```
Best parameters set found on development set:

{'C': 25, 'gamma': 0.01, 'kernel': 'rbf'}

Training accuracy
0.9174174174174174
SVC(C=25, class_weight='balanced', gamma=0.01)


****Results****
====================================================
time cost: 2.440507173538208

confusion matrix
 [[43  0  0  0]
 [ 0 11  0  0]
 [ 0  1  8  0]
 [ 0  0  9  2]]

               accuracy: {} 0.8648648648648649

               classification report
----------------------------------------------------
              precision    recall  f1-score   support

           1       1.00      1.00      1.00        43
           2       0.92      1.00      0.96        11
           3       0.47      0.89      0.62         9
           4       1.00      0.18      0.31        11

    accuracy                           0.86        74
   macro avg       0.85      0.77      0.72        74
weighted avg       0.92      0.86      0.84        74
```

2.2.3 Decision tree (DT) results:

When applying the DT algorithm, we notice that it also achieved a very good accuracy 85.14%.

As for the confusion matrix, the algorithm correctly predicted 63 patients correctly (43 patients with hypertension, 11 patients with infarction, 9 patients with congestive heart failure).

Surprisingly, it made incorrect predictions in only one class, classifying 11 patients as with congestive heart failure, when they are actually with arrhythmia.

```
[[43  0  0  0]
 [ 0 11  0  0]
 [ 0  0  9  0]
 [ 0  0 11  0]]
               accuracy: {} 0.8513513513513513

               classification report
---------------------------------------------------
              precision    recall  f1-score   support

           1       1.00      1.00      1.00        43
           2       1.00      1.00      1.00        11
           3       0.45      1.00      0.62         9
           4       0.00      0.00      0.00        11

    accuracy                           0.85        74
   macro avg       0.61      0.75      0.66        74
weighted avg       0.78      0.85      0.81        74
```

2.2.4. Random forest (RF) results:

Finally, the Random Forest model was applied, where each decision tree learns from a random sample of data points, ensuring that each tree is trained on a different sample.

We used several depths of the decision tree, to see how would the accuracy change.

Max_depth=5:

When determining the depth of the decision tree (5), the RF algorithm achieved a remarkable accuracy of 85.14%.

As for the confusion matrix, the algorithm correctly predicted 63 patients correctly (43 patients with hypertension, 11 patients with infarction, 9 patients with congestive heart failure). However, it made incorrect predictions in only one class, classifying 10 patients as with congestive heart failure, when they are actually with arrhythmia and classifying 1 patient with infraction, when he is actually with arrhythmia.

```
[[43  0  0  0]
 [ 0 11  0  0]
 [ 0  0  9  0]
 [ 0  1 10  0]]
0.8513513513513513
```

<u>Max_depth=10, 15, and 20:</u>

When increasing the depth of the forest to 10, 15, or 20, the accuracy increased to 87.84% and remained steady.

As we can see from the confusion matrix, the algorithm correctly predicted 65 patients correctly (43 patients with hypertension, 11 patients with infarction, 9 patients with congestive heart failure, 2 patients with arrhythmia). However, it made incorrect predictions in only one class too, classifying 9 patients as congestive heart failure, when they are actually with arrhythmia.

```
[[43  0  0  0]
 [ 0 11  0  0]
 [ 0  0  9  0]
 [ 0  0  9  2]]
               accuracy: {} 0.8783783783783784

               classification report
---------------------------------------------------
              precision    recall  f1-score   support

           1       1.00      1.00      1.00        43
           2       1.00      1.00      1.00        11
           3       0.50      1.00      0.67         9
           4       1.00      0.18      0.31        11

    accuracy                           0.88        74
   macro avg       0.88      0.80      0.74        74
weighted avg       0.94      0.88      0.86        74
```

In conclusion, the results highlight the varying performance of classification algorithms in predicting CVD. SVM with balanced weights demonstrated the best accuracy 91.74%, while decision tree and random forest (tree depth=5) achieved the same accuracies of 85.14%. On the other hand, when increasing the depth of the trees to 10, 15, or 20, the accuracy increased to 87.84% and remained steady.

3. Recommendations and Future aspirations:

Based on the results, patients should choose and prepare foods that contain little or no salt. Eating large amounts of salt can increase blood pressure and the risk of stroke and heart failure. The American Heart Association (AHA) recommends limiting salt intake to less than 1,500 mg per day and eat plenty and a variety of fruits and vegetables they provide essential nutrients and phytochemicals that may lower your risk of cardiovascular disease (CVD)and Choose foods made mostly with whole grains rather than refined grains. Whole grains can improve your cardiovascular risk factors by lowering your blood pressure, cholesterol, and blood sugar levels.

We look forward to developing this research with important additions, most notably:

· Training model on a larger set of data.

· Collect data from multiple hospitals to reduce bias.

· The application of this model becomes a complete innovation that can be used in the health sector and hospitals in Syria.

4. References:

1. Sherrel Z, what to know about the cardiovascular system? Medical news today. (2021) July 30.
2. How does the heart work?.InformedHealth.org [Internet].(January 31, 2019).
3. Farley A, McLafferty E, Hendry C. The cardiovascular system. 2012 Oct 31-Nov 6Nurs Stand. 27(9):35-9.
4. Benjamin EJ, Virani SS, Callaway CW, Chamberlain AM, Chang AR, Cheng S, Chiuve SE, Cushman M, Delling FN, Deo R, de Ferranti SD, Ferguson JF, Fornage M, Gillespie C, Isasi CR, Jiménez MC, Jordan LC, Judd SE, Lackland D, Lichtman JH, Lisabeth L, Liu S, Longenecker CT, Lutsey PL, Mackey JS, Matchar DB, Matsushita K, Mussolino ME, Nasir K, O'Flaherty M, Palaniappan LP, Pandey A, Pandey DK, Reeves MJ, Ritchey MD, Rodriguez CJ, Roth GA, Rosamond WD, Sampson UKA, Satou GM, Shah SH, Spartano NL, Tirschwell DL, Tsao CW, Voeks JH, Willey JZ, Wilkins JT, Wu JH, Alger HM, Wong SS, Muntner P., American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart Disease and Stroke Statistics-2018 Update: A Report From the American Heart Association. Circulation. 2018 Mar 20;137(12):e67-e492.
5. Jensen RV, Hjortbak MV, Bøtker HE. Ischemic heart disease: an update. Semin Nucl Med. 2020 May;50(3):195–207.
6. Sanchis-Gomar F, Perez-Quilis C, Leischik R, Lucia A. Epidemiology of coronary heart disease and acute coronary syndrome. Ann Transl Med. 2016 Jul;4(13):256.
7. Atherosclerosis.pennmedicine.org [Internet].
8. Lu L, Liu M, Sun R, Zheng Y, Zhang P. Myocardial infarction: symptoms and treatments. Cell Biochem Biophys. 2015 Jul;72(3):865–7.
9. Puymirat É. [Epidemiology of coronary artery disease]. Rev Prat. 2015 Mar;65(3):317-20.
10. Tsao CW, Aday AW, Almarzooq ZI, Anderson CAM, Arora P, Avery CL, et al. Heart Disease and Stroke Statistics-2023 Update: A Report From the American Heart Association. Circulation. 2023 Feb 21;147(8):e93–621.
11. [anatomy/cardiovascular/blood/classification].training.seer.cancer.gov.
12. Wang JC, Lise S, Normand T, Mauri L, Kuntz RE. Coronary artery spatial distribution of acute myocardial infarction occlusions. *Circulation.* 2004; 110:278–84.
13. Wirtz PH, von Känel R. Psychological stress, inflammation, and coronary heart disease. Curr Cardiol Rep. 2017 Sep 20;19(11):111.
14. Sata M, Saiura A, Kunisato A, Tojo A, Okada S, Tokuhisa T, Hirai H, Makuuchi M, Hirata Y, Nagai R. Hematopoietic stem cells differentiate into vascular cells

that participate in the pathogenesis of atherosclerosis. Nat Med. 2002 Apr;8(4):403-9.

15. Stary HC, Chandler AB, Dinsmore RE, Fuster V, Glagov S, Insull W, Rosenfeld ME, Schwartz CJ, Wagner WD, Wissler RW. A definition of advanced types of atherosclerotic lesions and a histological classification of atherosclerosis. A report from the Committee on Vascular Lesions of the Council on Arteriosclerosis, American Heart Association. Circulation. 1995 Sep 01;92(5):1355-74.
16. Atherosclerosis.britannica.com [Internet].(Nov 23, 2023).
17. Allport SA, Kikah N, Abu Saif N, Ekokobe F, Atem FD. Parental Age of Onset of Cardiovascular Disease as a Predictor for Offspring Age of Onset of Cardiovascular Disease, PLOS ONE. 2016; Dec;11(12).
18. Lloyd-Jones DM, Hong Y, Labarthe D, Mozaffarian D, Appel LJ, Van Horn L, Greenlund K, Daniels S, Nichol G, Tomaselli GF, Arnett DK, Fonarow GC, Ho PM, Lauer MS, Masoudi FA, Robertson RM, Roger V, Schwamm LH, Sorlie P, Yancy CW, Rosamond WD., American Heart Association Strategic Planning Task Force and Statistics Committee. Defining and setting national goals for cardiovascular health promotion and disease reduction: the American Heart Association's strategic Impact Goal through 2020 and beyond. Circulation. 2010 Feb 02;121(4):586-613.
19. Malakar A , Choudhury D, Halder B, Paul P, Uddin A ,Chakraborty S, A review on coronary artery disease, its risk factors, and therapeutics, J Cell Physiol. 2019;1–12.
20. Magriplis, E., Panagiotakos, D. B., Kyrou, I., Tsioufis, C., Mitsopoulou, A., Karageorgou, D., Dimakopoulos, I., Bakogianni, I., Chourdakis, M., Micha, R., Michas, G., Ntouroupi, T., Tsaniklidou, S., Argyri, K., & Zampelas, A.Presence of Hypertension Is Reduced by Mediterranean Diet Adherence in All Individuals with a More Pronounced Effect in the Obese: The Hellenic National Nutrition and Health Survey (HNNHS). *Nutrients*, 2020 ;*12*(3), 853
21. Micha, R., Peñalvo, J. L., Cudhea, F., Imamura, F., Rehm, J., & Mozaffarian, D. Association between dietary factors and mortality from heart disease, stroke, and Type 2 diabetes in the United States. *JAMA*, 2017;*317*(9), 912.
22. Wei, T., Liu, J., Zhang, D., Wang, X., Li, G., Ma, R., Chen, G., Lin, X., & Guo, X. (2021). The relationship between nutrition and atherosclerosis. *Frontiers in Bioengineering and Biotechnology*, *9*.
23. Carime, N. A., Cottenet, J., Clerfond, G., Eschalier, R., Quilliot, D., Eicher, J., Joly, B., & Quantin, C. (2022). Impact of nutritional status on heart failure mortality: a retrospective cohort study. *Nutrition Journal*, *21*(1).
24. Moyer, M. W. (2010). Carbs against Cardio. *Scientific American*, *302*(5), 19–21.
25. Kelly, R. E., Tong, T. Y., Watling, C. Z., Reynolds, A., Piernas, C., Schmidt, J. A., Papier, K., Carter, J. L., Key, T. J., & Pérez-Cornago, A. (2023). Associations

between types and sources of dietary carbohydrates and cardiovascular disease risk: a prospective cohort study of UK Biobank participants. *BMC Medicine*, *21*(1).
26. Micha, R., Shulkin, M., Peñalvo, J. L., Khatibzadeh, S., Singh, G., Rao, M., Fahimi, S., Powles, J., & Mozaffarian, D. (2017). Etiologic effects and optimal intakes of foods and nutrients for risk of cardiovascular diseases and diabetes: Systematic reviews and meta-analyses from the Nutrition and Chronic Diseases Expert Group (NutriCoDE). *PLOS ONE*, *12*(4), e0175149.
27. Mount WD. Bioinformatics: sequence and genome analysis. 2nd ed. New York: Cold Spring Harbor Laboratory Press; 2004. 692 p. doi:10.1086/431054
28. 2.Bioinformatics [Internet]. 2016. https://en.wikipedia.org/wiki/Bioinformatics. Accessed: 2016-03-10
29. Abdurakhmonov, I. Y. (2016). Bioinformatics: Basics, development, and future. In *InTech eBooks*.
30. Khomtchouk, B. B., Tran, D., Vand, K. A., Might, M., Gozani, O., & Assimes, T. L. (2019). Cardioinformatics: the nexus of bioinformatics and precision cardiology. *Briefings in Bioinformatics*, *21*(6), 2031–2051.
31. Uzun Ozsahin D, Ozgocmen C, Balcioglu O, Ozsahin I, Uzun B. Diagnostic AI and cardiac diseases. Diagnostics (Basel). 2022 Nov 22;12(12).
32. Zhang C, Lu Y. Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration. 2021 Sep;23:100224.
33. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol. 2020 Feb 27;9(2):14
34. Karatzia L, Aung N, Aksentijevic D. Artificial intelligence in cardiology: Hope for the future and power for the present. Front Cardiovasc Med. 2022 Oct 13;9:945726.
35. Karatzia L, Aung N, Aksentijevic D. Artificial intelligence in cardiology: Hope for the future and power for the present. Front Cardiovasc Med. 2022 Oct 13;9:945726.
36. Ahsan MM, Luna SA, Siddique Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare (Basel). 2022 Mar 15;10(3).
37. Pal M, Parija S. Prediction of Heart Diseases using Random Forest. J Phys: Conf Ser. 2021 Mar 1;1817(1):012009.