HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

A PROJECT REPORT

Submitted by

N. KAVINRAJ

(Reg. No: 22MCR045)

V. KIRUBAKINI

(Reg. No: 22MCR053)

in partial fulfilment of the requirements

for the award of the degree

of

MASTER OF COMPUTER APPLICATIONS

DEPARTMENT OF COMPUTER APPLICATIONS

KONGU ENGINEERING COLLEGE

(Autonomous)

PERUNDURAI, ERODE – 638 052

JUNE 2023

DEPARTMENT OF COMPUTER APPLICATIONS

KONGU ENGINEERING COLLEGE

(Autonomous)

PERUNDURAI, ERODE – 638 052

JUNE 2023

BONAFIDE CERTIFICATE

This is to certify that the project report entitled **"HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS"** is the bonafide record of project work done by **KAVINRAJ N (22MCR045)** and **KIRUBAKINI V (22MCR053)** in partial fulfillment of the requirements for the award of the Degree of Master of Computer Applications of Anna University, Chennai during the year 2022-2023.

SUPERVISOR **HEAD OF THE DEPARTMENT**

Date: **(Signature with seal)**

Submitted for the mini project viva voce examination held on _____

INTERNAL EXAMINER **EXTERNAL EXAMINER**

DECLARATION

We affirm that the project entitled **"HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS"** being submitted in partial fulfilment of the requirements for the award of Master of Computer Applications is the original work carried out by us. It has not formed the part of any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidates.

N. KAVINRAJ (Reg. No: 22MCR045)

V. KIRUBAKINI (Reg. No: 22MCR053)

I certify that the declaration made by the above candidates is true to the best of my knowledge.

Date: **Name and Signature of the Supervisor**

[DR.T. KAVITHA]

ABSTRACT

Heart-related illnesses, often known as cardiovascular diseases (CVDs), are the leading cause of death worldwide during the past several decades and have become the most serious illness in both India and the rest of the globe. Therefore, a trustworthy, accurate, and practical system is required to identify these disorders early enough for effective therapy. In order to automate the examination of massive and complicated data, machine learning methods and techniques have been used on a variety of medical datasets.

To identify the relevant demographic and clinical variables that contribute to the risk of heart disease. To collect and pre-process data from various sources, including electronic medical records and patient surveys. To develop a user-friendly interface for clinicians and patients to input data and obtain a heart disease risk score.

Before doing Feature extraction for heart disease prediction using machine learning involves selecting and extracting relevant information from raw data to create a set of informative features that can be used as input for a machine learning model. Random Forest is an algorithm consisting of decision trees. Random Forest Classifier from the sklearn. ensemble package was used to build the home and all matrices' models. The number of estimators equalled 1000 in both the home and all matrices' models.

ACKNOWLEDGEMENT

We respect and thank our Correspondent **Thiru.A.K.ILANGO BCom., MBA.,** and our Principal **Dr.V.BALUSAMY BE(Hons)., MTech., PhD.,** Kongu Engineering College, Perundurai for providing us with the facilities offered.

We convey our gratitude and heartfelt thanks to our Head of the Department **Dr.R.THAMILSELVAN MCA., ME., PhD.,** Department of Computer Applications, Kongu Engineering College for his perfect guidance and support that made this work to be completed successfully.

We also like to express our gratitude and sincere thanks to our project coordinator **Dr.P.VIJAYKUMAR MCA., M.Phil., Phd., MBA.,** Associate Professor and our beloved **Ms.T KALPANA MCA**., Assistant Professor(s) (Sr.G), Department of Computer Applications, Kongu Engineering College, who have motivated us in all aspects for completing the project in scheduled time.

We would like to express our gratitude and sincere thanks to our project guide **Ms.T.KAVITHA MCA.,** Assistant Professor (Sr.G), Department of Computer Applications, Kongu Engineering College, for giving her valuable guidance and suggestions, which helped us in successful completion of the project.

We owe a great deal of gratitude to our parents for helping us to be overwhelmed in all proceedings. We bow our heart and head with heartfelt thanks to all those who taught us their warm service to succeed and achieve our work

TABLE OF CONTENT

CHAPTER No.	TITLE	PAGE No
	ABSTRACT	4
	ACKNOWLEDGEMENT	5
	LIST OF FIGURES	
	LIST OF ABBREVIATIONS	
1	**INTRODUCTION**	
	1.1 OVERVIEW OF THE PROJECT	9
	1.1.1 Heart Disease	9
	1.1.2 Machine Learning	9
	1.1.3 Heart disease prediction	10
2	**LITERATURE REVIEW**	
	2.1 Study	11
3	**METHODOLOGY**	
	3.1 REQUIREMENT SPECIFICATION	
	3.1.1 Hardware Requirements	18
	3.1.2 Software Requirements	18

3.2 SOFTWARE DESCRIPTION

3.2.1 Google colab 18

3.2.2 Applications of Google colab 19

3.2.3 Python 20

3.2.4 Features of python 20

3.2.5 Tools Description 20

3.3 LIBRARIES 21

3.4 RANDOM FOREST 22

3.5 SUPPORT VECTOR MACHINE 23

3.6 DECISION TREE 26

4 PROJECT IMPLEMENTATION

4.1 Data Collection 26

4.2 Model Traning 27

4.3 Data Preprocessing 28

4.4 Classification Algorithms 28

5 EXPERIMENTAL RESULTS

5.1 RESULTS 32

5.2 PERFORMANCE MEASURES 32

5.2.1 Accuracy Percentage 33

6 CONCLUSION AND FUTURE ENHANCEMENT

6.1 Conclusions 35

6.2 Future Enhancement 35

APPENDIX 36

REFERENCES 43

LIST OF FIGURES

FIGURE No.	TITLE	PAGE No.
1.1	Concept of machine learning	10
3.1	Random Forest	23
3.2	Support Vector Machine	24
4.1	Heart disease prediction dataset	26
4.2	Split the data	27
4.3	Model fitting	27
4.4	Model fitting dataset	29
4.5	Model fitting Support Vector Machine	30
4.6	Model fiiting Rnadom Forest	31
5.1	Accuracy	33
5.2	Accuracy chart	34

LIST OF ABBREVIATIONS

ABBREVIATIONS	EXPANSION
SVM	SUPPORT VECTOR MACHINE
RF	RANDOM FOREST

CHAPTER 1

INTRODUCTIONS

1.1 OVERVIEW OF THE PROJECT

1.1.1 Heart disease

Heart disease, caused by abnormal heart and blood vessel conditions, is widely considered a direct threat to human life and health. It is one of the significant diseases exerting irreversible effects on many middle-aged and older people, in which fatal complications are highly likely to result. Makino states that the absolute risk of cardiovascular heart disease is associated with disability and death among people 65 years or older. The World Health Organization (WHO) declared an estimated 17.7 million people died from cardiovascular disorders in 2015, accounting for one-third of all deaths that year.

1.1.2 Machine learning

Machine learning is a branch of computer science that is committed to examining and interpreting patterns and structures in data to authorize learning, reasoning and decision making outside of manual interaction. In other words, machine learning enables the user to feed a massive amount of data to an algorithm, which then analyses it and potentially generates data-driven recommendations and decisions based on the input data. If there are modifications, the algorithm can take them into account for future decision-making. Machine learning focused on improving computer programmers that can gather data and use it to make their own decisions. Linear regressions, regression, Naive Bayes, KNN, decision tress, random forests, dimensionality reduction and SVM are a few of the frequently used machine learning methods.

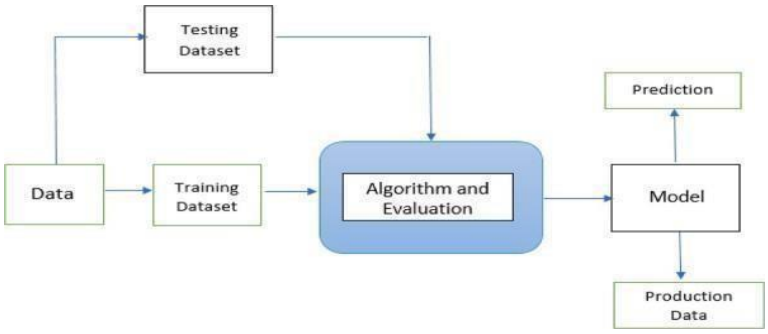

Figure:1.1 Concept of Machine Learning

1.1.3 Heart disease prediction

Computational technology and statistical approach have been popular in discovering the relationship between heart diseases and patients' health conditions. They can help predict the potential risk of heart disease based on the patient's underlying physical condition in advance, thereby reducing the probability of dying from a heart attack. Many statistical methods based on computer calculation have been applied to predict heart attacks. Due to its high accuracy, SVM has been prevalently applied as a classification method to predict heart attacks. Akkaya used Decision tree and the Support Vector Machine (SVM) algorithm to estimate heart failure and accomplished compromising outcomes. With the adoption of Random Forest, the best accuracy of 90.16% has been achieved by modification of feature selection. These algorithms have been proved to predict the risk of heart disease effectively.

CHAPTER 2

LITERATURE REVIEW

INTRODUCTION

In this literature review, the main focus is on the use of machine learning techniques, specifically on hyperparameter techniques for epileptic seizure detection. The purpose of this review is to provide an overview about the epilepsy and the seizure is presented in the human brain or not. To do this, we examine a wide range of literature, including academic papers, conference proceedings and other sources to gather and synthesize information about the hyperparameters.

2.1 STUDY

M A Rizvi, Himanshu Sharma (2017), proposed a Prediction of heart disease using machine learning. Medical professionals working in the field of heart disease have their own limitation, they can predict chance of heart attack up to 67% accuracy [2], with the current epidemic scenario doctors need a support system for more accurate prediction of heart disease. Machine learning algorithm and deep learning opens new door opportunities for precise predication of heart attack. Paper provides lot information about state of art methods in Machine learning and deep learning. An analytical comparison has been provided to help new researches' working in this field.

Abdulwahab Ali Almazroi (2021) suggested utilizing machine learning techniques to predict survival among heart patients. The development of machine learning approaches and algorithms for the early identification, diagnosis, and subsequent treatment of cardiovascular- related disorders was the focus of various works in this area. These studies concentrated on a range of topics, such as identifying crucial characteristics to accurately forecast the emergence of heart- related disorders in order to determine the survival likelihood.

By choosing a standard, well-defined, and well-curated dataset and a set of standard benchmark algorithms, this study adds to the body of literature by allowing the performance of each to be independently verified using a variety of performance evaluation measures. Based on the results of our experimental evaluation, decision trees outperform logistic regression as a performance method. Support vector machines, and artificial neural networks. Decision trees achieved 14% better accuracy than the average performance of the remaining techniques, Decision Tree accuracy 80% predicted.

Machine learning techniques were proposed by Harshit Jindal, Sarthak Agrawal, Rishabh khera, Rachna Jain, and Preeti Nagrath (2021) to predict heart disease. Using the patient's medical history, we developed a system to determine if a heart disease diagnosis is likely or not for the patient. We identified and classified the patient with heart disease with 88.52% accuracy using a variety of machine learning methods, including KNN and logistic regression. The regulation of how the model can be utilized to increase the precision of heart attack prediction in any individual was done in a very useful way. The proposed model's strength was quite pleasing, and it was able to foresee signs of heart disease in a specific person by using KNN and Logistic Regression which showed a good accuracy in comparison to the previously used classifier such as Naive Bayes etc.

Using machine learning, Rishabh Magar, Rohan Memane, and Suraj Raut (2021) suggested predicting heart disease. Based on the parameters provided regarding the patients' health, we will create a system that can effectively identify the rules to estimate the risk level of the patients. The aim is to anticipate the presence of heart illness in patients where the presence is evaluated on a scale and to identify hidden patterns by applying data mining techniques, which are notable to heart disorders. Large amounts of data that are too complex and large to analyse and evaluate using traditional methods are needed for the prediction of cardiac disease. Our goal is to identify the most appropriate machine learning method for heart attack prediction that is both accurate and computationally efficient.

In 2023, Chintan M. Bhatt, Parth Patel, Tarang Ghetia, and Pier Luigi Mazzeo proposed an efficient machine learning technique for heart disease prediction. Cardiologists can better classify patients' cardiovascular conditions by performing accurate diagnoses and prognoses, which enables them to administer the appropriate care to patients. Due to machine learning's ability to identify patterns in data, its applications in the medical sector have grown. Diagnosticians can prevent misdiagnosis by classifying the occurrence of cardiovascular illness using machine learning. In order to lower the fatality rate brought on by cardiovascular disorders, our research creates a model that can accurately forecast these conditions. The approach of k-modes clustering with Huang beginning that is suggested in this paper can increase classification precision. Models include the XGBoost (XGB), multilayer perceptron (MP), decision tree classifier (DT), random forest (RF), and Multilayer perceptron (MLP) 87.23% had a highest accuracy.

Apurv Garg, Bhartendu Sharma, and Rijwan Khan (2021) suggested utilising machine learning approaches to forecast heart disease. In this study, machine learning is utilised to determine whether or not a person has a heart condition. Cardiovascular diseases (CVDs) affect a large number of people and even claim lives globally. By taking into account factors like chest pain, cholesterol levels, a person's age, and other factors, machine learning can be used to determine whether a person has a cardiovascular disease. K-Nearest Neighbor's (K-NN) prediction accuracy is 86.885%.

Mohammed Khalid Hossen (2022) suggested utilising machine learning methods to forecast heart disease. Throughout their growth, machine learning and artificial intelligence have been shown to be beneficial in a variety of fields, particularly with the recent explosion in data. Making quicker and more accurate decisions in terms of disease forecasts may be possible using this method. The heart disease dataset is first converted into the necessary format for machine learning algorithms to run on it. The UCI repository is used to gather patient data, including medical records. The presence or absence of heart disease in the patients is then determined using the heart disease dataset. Second, this essay presents a

wealth of useful findings. In comparison to other algorithms, the Logistic Regression algorithm has a high accuracy rate of 95%.

Deep learning and machine learning were proposed by Pratipal Rai, Shanelle Fernandes, Tenzin Choedon, and Tseten Tashi Bhutia (2023) to detect cardiovascular disease. The largest cause of death worldwide is cardiovascular disease, usually referred to as heart disease. Numerous lives can be saved by early detection of cardiac problems, and it can even help doctors create successful treatment regimens. Cardiovascular disorders can be identified using an electrocardiogram (ECG), a simple and affordable method of measuring the electrical activity of the heart. The application of machine learning algorithms to identify cardiac disease has been the subject of numerous studies, although the bulk of these models do not offer very high accuracy with Nave Bayes 86%.

A Machine Learning-based Prediction and Diagnosis of Heart Disease utilising various models was proposed by Jyoti Maurya and Shiva Prakash (2023). The main factor contributing to illness is thought to be heart disease. Heart disease is becoming a serious issue that affects people of all ages because the majority of people are uninformed of their own kind and degree of heart disease. On the other hand, the manual approach to prediction is difficult and frequently necessitates the capacity to select the appropriate strategy. Different machine-learning models are essential in the automatic disease prediction in the medical industry, which is helping to overcome these problems. In this work, the accuracy of multiple machines learning models, including SVM, KNN, Logistic Regression, Decision Tree, Random Forest, and Gaussian, has been calculated and compared. Naive Bayes, AdaBoost, Extra Tree Classifier and Gradient Boosting for prediction of heart disease using UCI repository dataset for training and testing of models. Among all the models used, the highest accuracy of 95.08% obtained by the Gradient Boosting model The major aim of the paper is to get a reliable, computationally effective machine learning algorithm for heart disease prediction.

A proposal was made by Victor Chang, Meghana Ashok Ganatra, Karl Hall, Lewis Golightly, and Qianwen Ariel Xu (2022). An evaluation of artificial intelligence (AI) algorithms and models for the early detection and diagnosis of diabetes utilizing health markers. Data patterns can be discovered using machine learning models, and predictions can be made based on these patterns. For the diagnosis, prognosis, and treatment of diseases, they are used in healthcare applications. These models are now more efficient than ever at providing patient care thanks to the development of new algorithms and other technical advancements. This study's main goal is to use several machine learning techniques to foretell the diagnosis of diabetes. With an accuracy of 82.26%, Random Forest greatly beat the other models under investigation.

Using hybrid machine learning techniques, Senthilkumar Mohan, Chadrasegar Thirumalai, and Gautam Srivastava (2019) proposed an efficient method for heart disease prediction. With the useof machine learning (ML), it has been demonstrated that it is possible to make predictions and judgements from the vast amount of data generated by the healthcare sector. Additionally, we haveobserved the employment of ML approaches in recent IoT advances across a variety of domains. Only a few research have looked into using ML to predict cardiac disease. In this study, we suggesta unique approach to improve the precision of cardiovascular disease prediction by identifying keyfeatures using machine learning techniques. Different feature combinations and many well-knowncategorization strategies are used to introduce the prediction model.

An Analysis of Heart Disease Prediction Using Different Data Mining Techniques was proposed by Nidhi Bhatla and Kiran Jyoti in 2012. The phrase "heart disease" is used to describe a wide range of heart-related medical problems. These illnesses list the abnormal health disorders that have a direct impact on the heart and all of its components. Today, heart disease is a significant public health issue. The purpose of this research is to examine the various data mining methods that have been developed recently for the prediction of heart disease. The observations show that 15-attribute neural networks beat all other data mining strategies. Another finding from the

investigation is that using a genetic algorithm and feature subset selection, decision trees have also demonstrated good accuracy.Neural networks have the highest accuracy.

2015 saw the proposal of a Heart Disease Prediction using Different Data Mining Techniques by Jaymin Patel and Dr. Samir Patel.Several data mining techniques have been developed by researchers to aid medical practitioners in the identification of cardiac disease. However, fewer tests may be needed if data mining techniques are used. A speedy and effective detection method is required to lower the number of deaths caused by cardiac disorders. One of the efficient data mining techniques is the decision tree. This study evaluates various Decision Tree classification algorithms in an effort to use WEKA to diagnose heart disease more effectively. The J48 algorithm, Logistic model tree algorithm, and Random Forest algorithm are the algorithms that are being tested. The Cleveland database of the UCI repository's current patient files for people with heart illness.

Using machine learning techniques, V.V. Ramalingam, Ayantan Dandapath, and M. Karthik Raja (2018) proposed predicting heart disease. Heart-related illnesses, often known as cardiovascular diseases (CVDs), are the leading cause of death worldwide during the past several decades and have become the most serious illness in both India and the rest of the globe. Therefore, a trustworthy, precise, and workable method is required to identify these disorders early and start the appropriate course of treatment. Various medical datasets have been subjected to machine learning methods and techniques in order to automate the examination of huge and complex data. In recent years, numerous researchers have employed a variety of machine learning techniques to aid the healthcare sector and experts in the detection of heart related.

Jyoti Soni, Uzma Ansari, and Dipesh Sharma (2011), proposed a the healthcare industry is still knowledge-poor but information-rich. Within the health care systems, there is a lot of data. The lack of efficient analysis tools, however, makes it difficult to find hidden links in data. The purpose of this work is to create a GUI-based interface that

allows users to enter patient information and determine using a classifier based on weighted association rules whether or not a patient has heart disease. The forecast is made using historical patient data or another data source. Different weights are assigned to various attributes in the Weighted Associative Classifier (WAC) based on how well they can predict. It has already been established that associative classifiers outperform conventional classifier methods.

Using supervised machine learning algorithms, Md Mamun Ali, Bikash Kumar Paul, Kawsar Ahmed, Francis M.Bui, Julian M.W.Quinn, and Mohammed Ali Moni (2021) suggested a method for predicting heart disease. Analysis and comparison of performance. Although they would be very useful clinically, machine learning and data mining-based methods for the identification and prediction of cardiac disease are extremely difficult to create. The development of precise and effective early-stage heart disease prediction through analytical support of clinical decision-making with digital patient data could address the shortage of cardiovascular competence and the high incidence of cases that are misdiagnosed in the majority of countries. This study sought to find the most accurate machine learning classifiers for these diagnostic uses. For the purpose of predicting cardiac disease, a number of supervised machine-learning algorithms were used and their effectiveness was evaluated. Scores for feature importance

CHAPTER 3

METHODOLOGY

3.1 REQUIREMENTS SPECIFICATION

3.1.1 HARDWARE REQUIREMENTS

Processor : Intel Core i5

RAM : 4 GB

Hard Disk : 1 TB

Keyboard : Standard 104 keys

3.1.2 SOFTWARE REQUIREMENTS

Operating System: Windows 10

Tool : Google Colab

3.2 SOFTWARE DESCRIPTION

3.2.1 Google Colab

In terms of AI research, Google is active. Google spent many years creating the TensorFlow artificial intelligence framework and Collaboratory, an application development platform. TensorFlow is now open-sourced, and Google has made Colaboratory available to everyone for free since 2017. Google Colab or just Colab are the current names for Collaboratory.

A research tool for machine learning instruction and study is Google Colab. The most recent versions of Chrome, Firefox, and Safari were used to test Colaboratory the most completely. A free cloud service called Google Colab now offers free GPU support. Additionally, it supports the Total Processing Unit. Directly imported, mounted, and uploaded files come from discs.

Working with deep learning frameworks like PyTorch, Keras, TensorFlow, and OpenCV is excellent with Colab, as is honing your Python coding skills. You can import most of your favourite directories, mount your Google Drive and use whatever is stored there, create notebooks in Colab, upload personal Jupyter Notebooks, store notebooks directly from GitHub, share notebooks, upload Kaggle data, and download notebooks.

3.2.2 Applications of Google Colab

Google Colab is widely used in different segments like;

3.2.2.1 Business and Financial Analytics

3.2.2.2 Marketing and Trading, Share market

3.2.2.3 Education and Research

3.2.2.4 Financial trend analysis

3.2.2.5 Genetic engineering

3.2.2.6 Space exploration

3.2.2.7 Text Mining

3.2.2.8 Machine Learning

3.2.2.9 Predictive Analytics

3.2.3 Python

python is an interpreted, object-oriented, high-level, dynamically semantic programming language. It is very attractive for Rapid Application Development as well as for use as a scripting or glue language to connect existing components together due to its highlevel built-in data structures, dynamic typing, and dynamic binding. Python's straightforward syntax emphasises readability and makes it simple to learn, which lowers the cost of programme maintenance. Python's support for modules and packages promotes the modularity and reuse of code in programmes. For all popular platforms, the Python interpreter and the comprehensive standard library are freely distributable and available in source or binary form.

3.2.4 Features of python:

3.2.4.1 It is very simple to use.

3.2.4.2 Python is free and open-source.

3.2.4.3 It is easy to maintain.

3.2.4.4 Python is a high-level language

3.2.4.5 It provides a perfect interface to all commercial databases.

3.2.4.6 It is easy to learn and has very simple syntax.

3.2.4.7 Python provides better structure and support.

3.2.5 Tools Description

Python is an object-oriented, high-level software program with dynamic semantics that is interpreted. Python facilitates software flexibility by supporting modules and packages. Enforcing machine learning algorithms can be difficult and time-consuming. Python structures and frameworks are used by programmers to minimize development time.

Colab notebooks let mix executable code and rich textbook, as well as diagrams, HTML, LaTeX, and more, in a single document. Colab notebooks are saved in the Google Drive account. The ability to participate in colab notebooks with associates or familiarity allows users to give commentary or even make changes. Colab is a multi-purpose framework that allows users to visualize and analyze data. It can be used to create and manage Jupyter notebooks, which are hosted by colab.

3.3 LIBRARIES

NumPy

NumPy is a Python library for working with arrays that also includes functions for direct algebra, Fourier conversion, and matrices. Numerical Python is referred to asNumPy.

Matplotlib

Matplotlib is a cross-platform data visualization and graphical contriving library for Python as well as its numerical extension NumPy. Matplotlib has numerous plots, including line, bar, handful, and histogram.Importing matplotlib: import matplotlib.pyplot as plt.

Pandas

Pandas is a powerful library for data manipulation and analysis. It provides datastructures and functions to efficiently handle structured data.
Importing pandas: import pandas as pd.

Seaborn

Seaborn is a visualization library built on top of matplotlib. It provides a high-levelinterface for creating attractive and informative statistical graphics
Importing seaborn: import seaborn as sns

Sklearn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python.

3.4 RANDOM FOREST

A random forest is a machine learning technique that is used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. The forest generated by the random forest algorithm trained through bagging or bootstrap aggregating.

This algorithm establishes the outcome based on the predictions of the output from various trees. Increasing the number of trees increases the precision of the outcome. A random forest eradicated the limitations of a decision tree algorithm. It reduces the overfitting of datasets and increases precision. It generated predictions without requiring configurations in packages.

Random forest is more accurate than the decision tress. It provides an effective way of handling missing data. It can produce a reasonable prediction without hyperparameter tuning. It solves the issue of overfitting in decision trees. In every random forest tree, a subset of features is selected randomly at the node's splitting point.

Random forest takes less training time as compared to other algorithms. It predicts output with high accuracy, even for the large dataset it runs efficiently. It can also maintain accuracy when a large proportion of data is missing. Some of the advantages of random forest algorithm is capable of performing both classification and regression tasks. It is capable of handling large datasets with high dimensionality. It enhances the accuracy of the model and prevents the overfitting issue

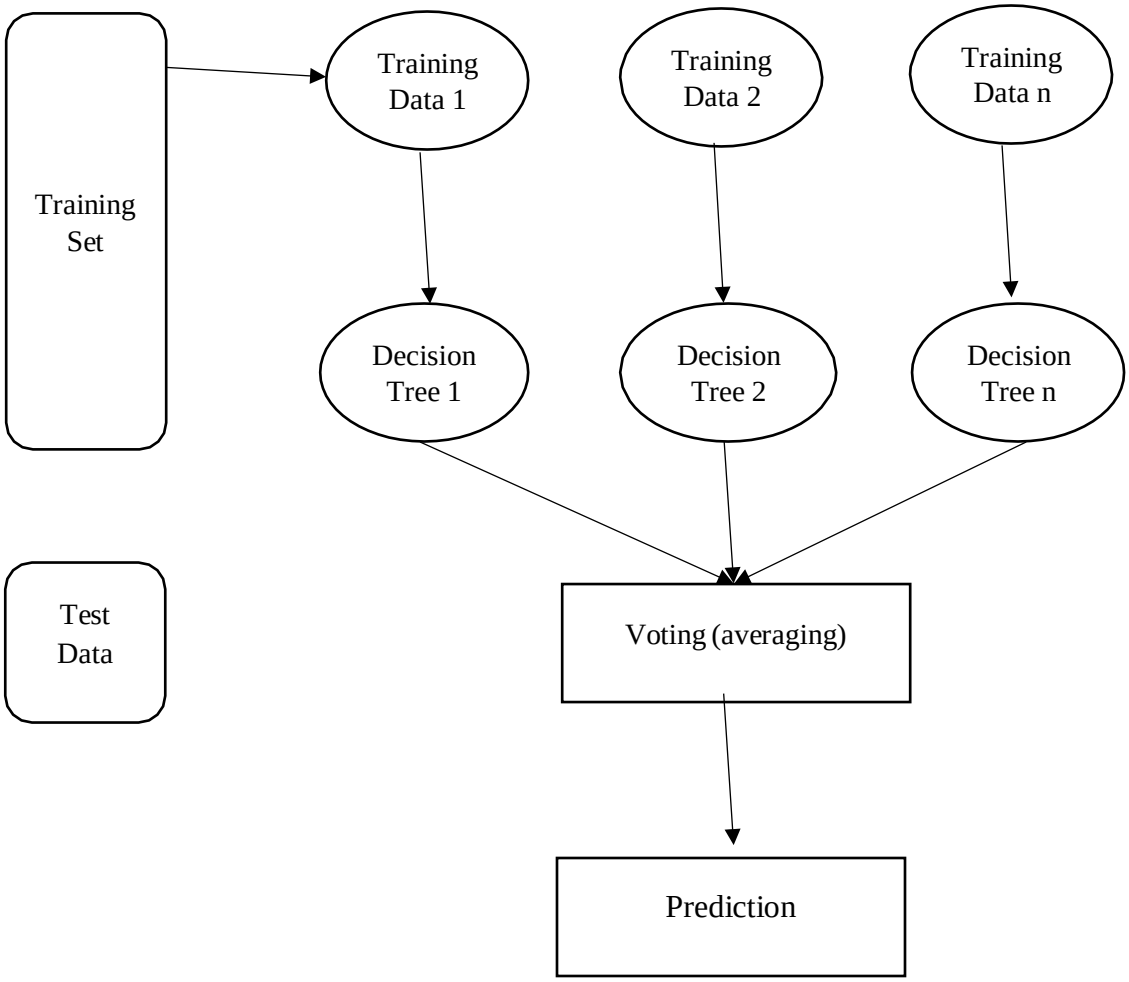

Figure 3.1 Random Forest

3.5 SUPPORT VECTOR MACHINE

Support Vector Machine is one of the most popular supervised learning algorithms, which is used for classification as well as regression problems. However, primarily it is used for classification problems in machine learning. The goal of the support vector machine is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.

Support vector machine chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors and hence algorithm is termed as SVM.

The SVM kernel is a function that takes low-dimensional input space and transforms it into higher-dimensional space it converts non separable problem to separable problems. It is mostly useful in non-linear separation problems. Simply put the kernel, does some extremely complex data transformations and then finds out the process to separate the data based on the labels or output defined.

Some of the advantages of SVM is effective in high-dimension cases. Its memory is efficient as it uses a subset of training points in the decision function called support vectors. Different kernel functions can be specified for the decision functions and it's possible to specify custom kernels.

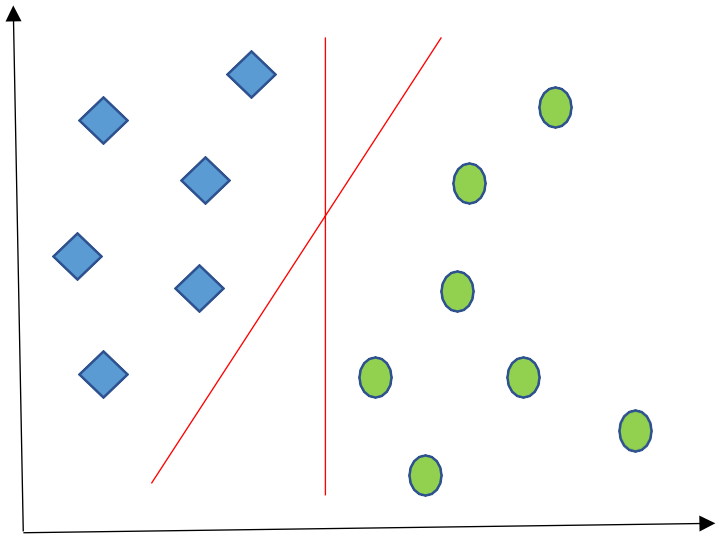

Figure 3.2 Support Vector Machine

3.6 DECISION TREE ALGORITHM

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree- structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.

A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

CHAPTER 4

PROJECT IMPLEMENTATION

4.1 DATA COLLECTION

The methodical procedure of acquiring information about a certain subject is known as data collection. The Kaggle website has provided the dataset. There are 303 data in the dataset. The dataset is made up entirely of numbers. The dataset in question consists of 302 data points or instances. Each data point represents a unique observation or entry in the dataset. The dataset is composed of 14 attributes or features. Attributes are the characteristics or properties of the data points that provide information for analysis and modelling.

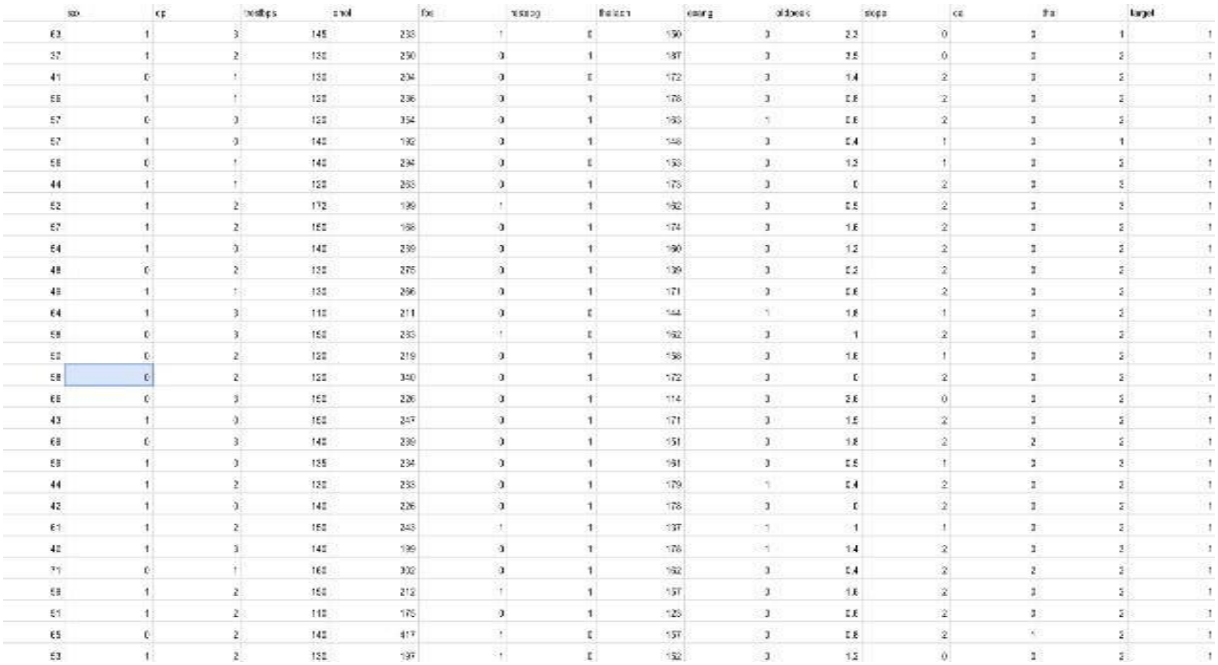

Figure 4.1 heart disease prediction dataset

4.2 MODEL TRAINING

Split the resampled and PCA-transformed dataset into training and testing. Create a random forest classifier object. Train the classifier using the training data. Make the predictions on the data set.

```
[39] from sklearn.model_selection import train_test_split

     predictors = dataset.drop("target",axis=1)
     target = dataset["target"]

     X_train,X_test,Y_train,Y_test = train_test_split(predictors,target,test_size=0.20,random_state=0)

[40] X_train.shape

     (242, 13)

[41] X_test.shape

     (61, 13)

[42] Y_train.shape

     (242,)

     Y_test.shape

     (61,)
```

Figure 4.2 Split the data

```
from sklearn.ensemble import RandomForestClassifier

max_accuracy = 0


for x in range(2000):
    rf = RandomForestClassifier(random_state=x)
    rf.fit(X_train,Y_train)
    Y_pred_rf = rf.predict(X_test)
    current_accuracy = round(accuracy_score(Y_pred_rf,Y_test)*100,2)
    if(current_accuracy>max_accuracy):
        max_accuracy = current_accuracy
        best_x = x

#print(max_accuracy)
#print(best_x)

rf = RandomForestClassifier(random_state=best_x)
rf.fit(X_train,Y_train)
Y_pred_rf = rf.predict(X_test)
```

Figure 4.3 Model fitting

4.3 DATA PREPROCESSING

The describe function and pandas profiling in Python were used in the study's data description to provide an overview of the dataset. For 303 patients, there were 14 factors in the raw data. Data analysis was performed using correlation matrices, extra-tree classifiers, and chi- square values. All variables were chosen for model development since the correlation matrices and Chi-square values revealed that no variables were substantially associated. Additionally, using Standard Scaler, all numerical variables were scaled to normal.

Both home matrices and all matrices were created for the 13 independent variables. Age, sex, resting blood pressure, cholesterol, fasting blood sugar, and thalassemia were the six factors that made up home matrices. All 13 independent variables were present in all matrices. With 80% training data and 20% test data, the research produced the training set and test sets.

The accuracy score, false negative rate, and confusion matrix for each model were displayed using the helper function in Python. The accuracy score was used to calculate the proportion of patients who were correctly identified as either having or not having a risk for heart disease. The resultant score demonstrated how well each model predicted patients' actual risks of developing heart disease. The fraction of patients whose true risk for heart disease was misdiagnosed as being low was determined by the false negative rate. The false negative rate was noteworthy because inaccurate forecasts could cause patients' treatments to be delayed. The accuracy of self-measured home matrices in contrast to all other matrices was determined using those values in the final model comparison.

4.4 CLASSIFICATION ALGORITHMS

The classification of the dataset is done with a different machine learning algorithm. They are Logistic Regression, Random Forest and Decision Tree. Training and testing datasets are separated from the overall dataset.

1. DECISION TREE

Decision trees are nonparametric supervised learning algorithms that are used for classification and regression issues, but especially for classification. By generating a set of rules, a decision tree classifier builds a tree for categorizing input into classes. The root node is where the algorithm begins, predicts the class, equates the record's root value and attributes, and jumps to the next node based on the comparison.

The maximum depth from 1 to 30 was used in the research to create a line graph, and the decision tree classifier score was used as the y axis. For the model building, a maximum depth of 10 was chosen because it gives the best scores.

IMPLEMENTATION RELATED TO THE DATASET:

Step 1: According to dataset, start the tree at the root node (URL), which has the overall dataset.

Step 2: Using the method Attribute Selection Measure, identify the most important attribute in the dataset (ASM).

Step 3: To include potential values for the finest attributes, subset the attribute.

```
[ ] from sklearn.tree import DecisionTreeClassifier

    max_accuracy = 0


    for x in range(200):
        dt = DecisionTreeClassifier(random_state=x)
        dt.fit(X_train,Y_train)
        Y_pred_dt = dt.predict(X_test)
        current_accuracy = round(accuracy_score(Y_pred_dt,Y_test)*100,2)
        if(current_accuracy>max_accuracy):
            max_accuracy = current_accuracy
            best_x = x

    #print(max_accuracy)
    #print(best_x)


    dt = DecisionTreeClassifier(random_state=best_x)
    dt.fit(X_train,Y_train)
    Y_pred_dt = dt.predict(X_test)

[ ] print(Y_pred_dt.shape)

    (61,)

[ ] score_dt = round(accuracy_score(Y_pred_dt,Y_test)*100,2)

    print("The accuracy score achieved using Decision Tree is: "+str(score_dt)+" %")

    The accuracy score achieved using Decision Tree is: 81.97 %
```

Figure 4.4 Model fitting
dataset

2. SUPPORT VECTOR MACHINE

Because Support Vector Machine is a classification and regression algorithm, it was selected as one of the models. Svm from the Python sklearn.svm module was used for the study. For the two machine learning models, the radial basis function kernel was chosen, gamma was set to 0.01, and the regularization parameter was set to 1.

Figure 4.5 model fitting svm

2. RANDOM FOREST

Feature extraction for heart disease prediction using machine learning involves selecting and extracting relevant information from raw data to create a set of informative features that can be used as input for a machine learning model.

Random Forest is an algorithm consisting of decision trees. Random Forest Classifier from the sklearn. ensumble package was used to build the home and all matrices' models. The number of estimators equaled 1000 in both the home and all matrices' models.

The extraction of Data is

- cp
- fbs
- restecg
- exang
- slope
- ca
- thal

```
from sklearn.ensemble import RandomForestClassifier

max_accuracy = 0


for x in range(2000):
    rf = RandomForestClassifier(random_state=x)
    rf.fit(X_train,Y_train)
    Y_pred_rf = rf.predict(X_test)
    current_accuracy = round(accuracy_score(Y_pred_rf,Y_test)*100,2)
    if(current_accuracy>max_accuracy):
        max_accuracy = current_accuracy
        best_x = x

#print(max_accuracy)
#print(best_x)

rf = RandomForestClassifier(random_state=best_x)
rf.fit(X_train,Y_train)
Y_pred_rf = rf.predict(X_test)
```

Figure 4.6 Model fitting random forest

CHAPTER 5

EXPERIMENTAL RESULTS

5.1 RESULT

In this work, the dataset is collected from the kaggle website and applied in python. It is one of the popular programming languages and the language is used in this work to detect the heart disease prediction. In this work, Random Forest, Support Vector Machine and Decision Tree algorithms are used to detect the results. The evaluated results are compared based on accuracy.

In this work, comparing and produce the best result by implementing those three algorithms like Random Forest, Support Vector Machine and Decision Tree. Random Forest produce the highest accuracy of 90.16% among them.

5.2 PERFORMANCE MEASURES

Performance measures also known as performance measures or evaluation metrics are used to access the performance of machine learning model. These metrics provide quantitative measurements that indicate how well the model is performing in terms of accuracy. Estimating the machine learning model performance is the most important step while building a most effective model. Performance measure is used to evaluate the quality of a model. Accuracy can be intended as the number of veracious forecasts to the overall number of forecasts.

Accuracy: It measures the proportion of correctly classified instances out of the total number of instances. Accuracy is a common metric for balanced datasets but may not be suitable for imbalanced datasets.

5.2.1 ACCURACY PERCENTAGE

Accuracy is metric used to measure the performance of a classification model. It represents the percentage of correct predictions made by the model on given dataset. Accuracy measures the overall correctness of the model's predictions. It calculates the percentage of correctly predicted instances out of the total number of instances. Accuracy is a commonly used metric, especially for balanced datasets, as it provides a straightforward measure of the model's success rate.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

However, accuracy can be misleading in certain scenarios, such as imbalanced datasets, where the majority class dominates the predictions. As shown in Figure 4, the following code is used for finding the accuracy of Random Forest algorithm.

I. ACCURACY OF ALGORITHMS

MODEL	ACCURACY SCORE
Random forest	90.16
Support vector machine	81.97
Decision tree	81.97

Figure 5.1. Accuracy score

Figure 5.2. Accuracy chart

```
[ ] scores = [score_svm,score_dt,score_rf]
    algorithms = ["Support Vector Machine","Decision Tree","Random Forest"]

    for i in range(len(algorithms)):
        print("The accuracy score achieved using "+algorithms[i]+" is: "+str(scores[i])+" %")

    The accuracy score achieved using Support Vector Machine is: 81.97 %
    The accuracy score achieved using Decision Tree is: 81.97 %
    The accuracy score achieved using Random Forest is: 90.16 %
```

Figure 5.3. Final Accuracy

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENT

6.1 CONCLUSION

To answer the research question of this study, it is concluded that the machine learning algorithms with only self-measurable physical condition indicators do not predict as accurately as machine learning algorithms with all physical condition indicators. Not only do algorithms with self-measurable physical condition indicators not predict the heart disease outcome as accurately as algorithms with all physical condition indicators, but they are also more likely to falsely predict not having heart disease among patients with heart disease. Thus, machine learning algorithms with only self-measurable physical condition indicators should not be used until more indicators are measurable at home in the future.

The evaluation metrics of accuracy have been demonstrated the models to give the best results. Comparing those three models Random Forest algorithm has exhibited good performance as accuracy of 90.16%. This project highlights the importance of machine learning algorithm for heart disease prediction.

6.2 FUTURE ENHANCEMENT

The limitations of this study have indicated the following areas as recommendations for future work. First, include other health attributes from the original dataset to discover the machine learning algorithm with the highest accuracy and lowest false negative rate. Second, since every patient has different health conditions, it is recommended to group the patients with similar health conditions and ages to investigate each machine learning algorithm’s accuracy rate.

APPENDIX

SAMPLE CODE

```
from google.colab import
drive
drive.mount('/content/drive')
import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline



import os
print(os.listdir())



import warnings
warnings.filterwarnings('ignore')



dataset = pd.read_csv("/content/drive/MyDrive/heart.csv")
```

```
type(dataset)

dataset.shape

dataset.head(5)

dataset.sample(5)


dataset.describe()

dataset.info()


info = ["age","1: male, 0: female","chest pain type, 1: typical angina, 2: atypical angina, 3: non- anginal pain, 4: asymptomatic","resting blood pressure"," serum cholestoral in mg/dl","fasting blood sugar > 120 mg/dl","resting electrocardiographic results (values 0,1,2)"," maximum heart rate achieved","exercise induced angina","oldpeak = ST depression induced by exercise relative to rest","the slope of the peak exercise ST segment","number of major vessels (0-3) colored by flourosopy","thal: 3 = normal; 6 = fixed defect; 7 = reversable defect"]



for i in
  range(len(info)):

    print(dataset.columns[i]+":\t\t\t"+info[i])


y = dataset["target"]


sns.countplot(y)
```

```
target_temp = dataset.target.value_counts()

print(target_temp)




print("Percentage of patience without heart problems:

"+str(round(target_temp[0]*100/303,2))) print("Percentage of patience with heart

problems: "+str(round(target_temp[1]*100/303,2)))



#Alternatively,


# print("Percentage of patience with heart problems:
"+str(y.where(y==1).count()*100/303))

# print("Percentage of patience with heart problems:
"+str(y.where(y==0).count()*100/303))



# #Or,


# countNoDisease = len(df[df.target == 0])


# countHaveDisease = len(df[df.target == 1])



from sklearn.model_selection import train_test_split
```

```
predictors =

dataset.drop("target",axis=1) target =

dataset["target"]



X_train,X_test,Y_train,Y_test =
train_test_split(predictors,target,test_size=0.20,random_state=0)



from sklearn import svm

sv = svm.SVC(kernel='linear')

sv.fit(X_train, Y_train)



Y_pred_svm = sv.predict(X_test)



from sklearn.tree import DecisionTreeClassifier



max_accuracy

= 0 for x in

range(200):

    dt = DecisionTreeClassifier(random_state=x)

    dt.fit(X_train,Y_train)
```

```
    Y_pred_dt = dt.predict(X_test)

    current_accuracy =

    round(accuracy_score(Y_pred_dt,Y_test)*100,2)

    if(current_accuracy>max_accuracy): max_accuracy =

    current_accuracy best_x = x


#print(max_accuracy)

#print(best_x)



dt = DecisionTreeClassifier(random_state=best_x)

dt.fit(X_train,Y_train)

Y_pred_dt = dt.predict(X_test)
from sklearn.ensemble import RandomForestClassifier


max_accuracy = 0



for x in range(2000):

    rf = RandomForestClassifier(random_state=x)

    rf.fit(X_train,Y_train)
```

```
    Y_pred_rf = rf.predict(X_test)

    current_accuracy =

    round(accuracy_score(Y_pred_rf,Y_test)*100,2)

    if(current_accuracy>max_accuracy): max_accuracy =

    current_accuracy best_x = x


#print(max_accuracy)

#print(best_x)


rf = RandomForestClassifier(random_state=best_x)

rf.fit(X_train,Y_train)

Y_pred_rf = rf.predict(X_test)


from sklearn.ensemble import RandomForestClassifier

max_accuracy = 0 for x in range(2000):

rf =

RandomForestClassifier(random_state=x

) rf.fit(X_train,Y_train)

    Y_pred_rf = rf.predict(X_test)

current_accuracy =

round(accuracy_score(Y_pred_rf,Y_test)*100,2)
```

```
if(current_accuracy>max_accuracy): max_accuracy

current_accuracy best_x = x


score_rf = round(accuracy_score(Y_pred_rf,Y_test)*100,2)


print("The accuracy score achieved using Random forest is: "+str(score_rf)+"

%") scores = [score_svm,score_dt,score_rf]

algorithms = ["Support Vector Machine","Decision Tree","Random Forest"]


for i in range(len(algorithms)):

 print("The accuracy score achieved using "+algorithms[i]+" is: "+str(scores[i])+"
                                                                          %")


sns.set(rc={'figure.figsize':(15,8)})

plt.xlabel("Algorithms")

plt.ylabel("Accuracy score")

sns.barplot(x=algorithms,y=scores,alp

ha=0.8)
```

REFERENCES

[1] AHA (2017) American Heart Association. Liu, X., Wang, X.L., Su, Q., Zhang, M., Zhu, Y.H., Wang, Q.G. and Wang, Q. (2017) A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method. Computational and Mathematical Methods in Medicine, 2017, 1-11.

[2] Akkaya, B., Sener, E. and Gursu, C. (2022) A Comparative Study of Heart Disease Prediction Using Machine Learning Techniques. 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, 9-11 June 2022, 1-8.

[3] Desai, F., Chowdhury, D., Kaur, R., Peeters, M., Arya, R.C., Wander, G.S., Gill, S.S. and Buyya, R. (2022) Health Cloud: A System for Monitoring Health Status of Heart Patients Using Machine Learning and Cloud Computing Article ID: 100485.

[4] Fahd Saleh Alotaibi," Implementation of Machine Learning Model to Predict Heart Failure Disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[5] Heron, M. (2012) Deaths: Leading Causes for 2008. National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System, 60, 1-94.

[6] Liu, W., Tang, Q., Jin, J., et al. (2021) Sex Differences in Cardiovascular Risk Factors for Myocardial Infarction. Herz, 46, 115-122.

[7] Lee, H.G., Noh, K.Y. and Ryu, K.H. (2007) Mining Biosignal Data: Coronary Artery Disease Diagnosis Using Linear and Nonlinear Features of HRV. In: Emerging Technologies in Knowledge Discovery and Data Mining, PAKDD 2007, Lecture Notes in Computer Science, Vol. 4819, Springer, Berlin, Heidelberg.

[8] Makino, K., Lee, S., Bae, S., Chiba, I., Harada, K., Katayama, O., Shinkai, Y. and Shimada, H. (2021) Absolute Cardiovascular Disease Risk Assessed in Old Age Predicts Disability

and Mortality: A Retrospective Cohort Study of Community-Dwelling Older Adults. Journal of the American Heart Association, 10, e022004.

[9] Nahar, J., Imam, T., Tickle, K.S. and Chen, Y.P.P. (2013) Computational Intelligence for Heart Disease Diagnosis: A Medical Knowledge Driven Approach. Expert Systems with Applications, 40, 96-104.

[10] Praveen Kumar Reddy, T.Sunil Kumar Reddy, Balakrishnan, Syed Muzamil Basha, Ravi Kumar Poluru, ,August 2019, Heart Disease Prediction Using Machine Learning Algorithm, Blue Eyes Intelligence Engineering & Sciences Publication

[11] WHO (2017) cardiovascular diseases. ABS (2009) Causes of Death, Australia. Australian Bureau of Statistics.

[12] Wilson, P.W.F., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H. and Kannel, W.B. (1998) Prediction of Coronary Heart Disease Using Risk Factor Categories. Circulation, 97, 1837-1847.

[13] Xing, Y.W., Wang, J., Zhao, Z.H. and Gao, Y.H. (2007) Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease. Convergence Information