UNIVERSITE JOSEPH KI-ZERBO

ECOLE DOCTORALE
INFORMATIQUE ET
CHANGEMENTS CLIMATIQUES

BURKINA FASO

Unité-Progrès-Justice

MASTER RESEARCH PROGRAM

SPECIALITY: INFORMATICS FOR CLIMATE CHANGE (ICC)

MASTER THESIS

Subject:

Development of Cardiovascular Disease Prediction and Patient Management System Using AI in The Gambia. A Comparative Study of Machine Learning Models

Defended on 18th, July, 2023 by:

MUHAMMED FATTY

fatty.m@edu.wascal.org

Major Supervisor

Prof Sidat Yaffa, UTG/WASCAL

Director of Doctoral Research Program

on Climate Change and Education,

University of The Gambia.

Co-Supervisor

Dr. Belko Abdoul Aziz DIALLO

Head of Data Management

(WASCAL Competence

Center, Burkina Faso)

Academic year 2022-2023

DEDICATION

This master thesis is dedicated to my dad and siblings most especially my beloved late mother, Sona, who tragically passed away from a heart disease due to delayed diagnosis. In her honor, I present the SONART A.I Cardiovascular Disease Prediction System developed for this thesis. The name "SONART" combines "Sona" and "rt," from heart symbolizing the significance of timely diagnosis and prevention

May this dedication stand as a lasting tribute to my mother, inspiring my unwavering commitment to advancing medical knowledge and enhancing patient care.

ACKNOWLEDGEMENT

- I am deeply grateful to WASCAL, for their invaluable financial support, which enabled me to pursue my degree and complete this thesis successfully. The opportunity provided by WASCAL has had a profound impact on my personal journey and the lives of many others.
- I extend my heartfelt appreciation to the retiered director, Prof. Tanga Pierre ZOUNGRANA, Dr. Ousmane COULIBALY the interim dicrector, and Dr. Benewinde Jean-Bosco ZOUNGRANA for their unwavering support and encouragement as the esteemed leaders of Ecole Doctorale Informatique et Changement Climatique (ED-ICC).
- Special recognition goes to my supervisors, Prof. Sidat Yaffa and Dr. Belko Abdoul Aziz DIALLO, as well as the jury members, for their guidance and constructive feedback that significantly influenced the direction and quality of my research.
- I am grateful to Dr. Lamin E.S Jaiteh, the Consultant Cardiologist at The Edward Francis Small Teaching Hospital, for his invaluable assistance in facilitating the research process.
- I express sincere gratitude to siblings namely, Awa Fatty, Wuday Fatty, Lamin Fatty and Abdoukarim Fatty and Binta Baldeh, as well as friends like, Kebba Ceesay, Ebou and Muhammed Jaw, for their unwavering support and motivation throughout my academic journey.
- Moreover, I acknowledge the support of WASCAL alumni, Haddy Jawla, Adamou Issaka Moussa and, Boubakar Zourkalaini whose physical and moral support proved crucial during this project.
- Furthermore I extend my thanks to my colleagues, the staff of ED-ICC, students, and the WASCAL Competence Center (CoC) staff for their support. Moreover, special thanks goes to Mame Diarra Diouf, Abdul Wahab and Serign Mbacke Coly for their support and guidance.
- Finally, I express gratitude to the Almighty Allah for granting me the strength and well-being to accomplish this task.

ABSTRACT

Cardiovascular diseases (CVDs) pose a global health challenge, particularly in low-income countries like The Gambia. This study investigates the efficacy of machine learning algorithms in predicting CVDs and explores IoT technology and an AI web application to enhance disease prediction and patient management.

Leveraging historical patient data from the Edward Francis Small Teaching Hospital in Banjul (EFSTH), The Gambia, the dataset comprises 915 rows and 9 columns, encompassing diverse demographics and medical profiles. The combined RF, LR, and SVM algorithms achieve 94.5% and 95% accuracy during validation and testing, respectively.

Results include the secure SONART AI Cardiovascular Prediction System, offering precise predictions, patient monitoring, and secure data storage. The system upholds confidentiality through encryption and access controls, serving as a cutting-edge healthcare solution in resource-constrained settings.

This research contributes valuable insights into predicting CVDs in The Gambia, enhancing patient outcomes and healthcare technologies. IoT integration, utilizing Arduino medical sensors, further empowers disease prediction and patient care.

In conclusion, this thesis addresses accurate CVD prediction and early detection needs in The Gambia. By leveraging machine learning and IoT, it advances health informatics, fostering innovative approaches in healthcare delivery. The findings exemplify a commitment to technology and data security, elevating healthcare standards in resource-limited regions.

Keywords: Cardiovascular diseases (CVDs), Machine learning algorithms, The Gambia, IoT technology, Disease prediction

RÉSUMÉ

Les maladies cardiovasculaires (MCV) représentent un défi majeur pour la santé mondiale, en particulier dans les pays à faible revenu comme la Gambie. Cette étude examine l'efficacité des algorithmes d'apprentissage automatique dans la prédiction des MCV et explore la technologie IoT et une application web d'intelligence artificielle pour améliorer la prédiction des maladies et la gestion des patients.

En exploitant des données historiques de patients provenant de l'Hôpital d'Enseignement Edward Francis Small à Banjul (EFSTH), en Gambie, l'ensemble de données comprend 915 lignes et 9 colonnes, englobant divers profils démographiques et médicaux. Les algorithmes combinés RF, LR et SVM atteignent respectivement des taux d'exactitude de 94,5 % et 95 % lors des phases de validation et de test.

Les résultats incluent le système sécurisé de prédiction cardiovasculaire SONART AI, offrant des prédictions précises, une surveillance des patients et un stockage sécurisé des données. Le système garantit la confidentialité grâce au chiffrement et aux contrôles d'accès, constituant une solution de pointe en matière de soins de santé dans les environnements aux ressources limitées.

Cette recherche apporte des connaissances précieuses dans la prédiction des MCV en Gambie, améliorant les résultats des patients et les technologies de santé. L'intégration de l'IoT, en utilisant des capteurs médicaux Arduino, renforce encore la prédiction des maladies et les soins aux patients.

En conclusion, cette thèse aborde la nécessité d'une prédiction précise des MCV et d'une détection précoce en Gambie. En tirant parti de l'apprentissage automatique et de l'IoT, elle fait progresser l'informatique de la santé, favorisant des approches innovantes dans la prestation des soins de santé. Les résultats témoignent d'un engagement envers la technologie et la sécurité des données, élevant les normes de soins de santé dans les régions aux ressources limitées.

Mots-clés : Maladies cardiovasculaires (MCV), Algorithmes d'apprentissage automatique, Gambie, Technologie IdO, Prédiction de maladies

ACRONYMS AND ABBREVIATIONS

CVD: Cardiovascular Disease

WASCAL: West African Science Service Centre on Climate Change and Adapted Land Use.

EFSTH: Edward Francis Small Teaching Hospital

CoC: Wascal Competence Center

WHO: Who Health Organization

IoT: Internet of Things

SP02: Peripheral Capillary Oxygen Saturation (estimate of the amount of oxygen in the blood)

Resting ECG: Electrocardiogram

Temp: Temperature

Thal: Thalassemia

EDA: Exploratory Data Analysis

ML: Machine Learning

DL: Deep Learning

KNN: K Nearest Neighbour

RF, LR, and GB: Random Forest, Logistic Regression, and Gradient Boosting

RF, LR, and SVM: Random Forest, Logistic Regression, and Support Vector Machine

RF, DT, and GB: Random Forest, Decision Tree, and Gradient Boosting

RF, DT, and SVM: Random Forest, Decision Tree, and Support Vector Machine

RF, GB, and SVM: Random Forest, Gradient Boosting and Support Vector Machine

HGBDTLR: Hybrid Gradient Boosting Decision Tree with Logistic Regression

IoMT: Internet of Medical Things

EDCNN: Enhanced Deep Learning Assisted Convolutional Neural Network

ANN: Artificial Neural Network

DNN: Deep Neural Network

NNE: Neural network ensemble method

NB: Naïve Bayes

RFE: Recursive Feature Elimination

LSTM: Long Short-Term Memory

GRU: Gated Recurrent Unit

CNN: Convolutional Neural Network

Table of Contents

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Cardiovascular diseases (CVD) rank among the primary factors responsible for worldwide mortality and morbidity. According to the WHO (World Health Organization), 17.9 million global deaths were attributed to cardiovascular diseases in 2017 (Hazra et al., 2017a). Climate change can have various impacts on the environment and human health, including contributing to the development and exacerbation of heart diseases. Some of the key causes of heart diseases associated with climate change include extreme temperatures air pollution, changes in infectious disease patterns, changes in precipitation patterns and natural disasters (Pacheco et al., 2021). In addition to climate-related factors, individual risk factors such as unhealthy lifestyle choices (e.g., poor diet, lack of physical activity, smoking) and genetic predisposition also play significant roles. Early detection and accurate prediction of these diseases can significantly reduce the associated health and economic burden (Bhatt et al., 2023).

As there is a recent improvement in medical health care, the healthcare system has collected a massive amount of data about heart disease, and datasets consisting of different medical parameters or features such as age, sex, blood pressure, cholesterol, chest type, and so on are now available for analysis (Patel & Patel, 2016). By applying machine learning algorithms to this massive amount of data, crucial information can be extracted to predict heart disease at an early stage (Garg et al., 2021). Various machine learning techniques such as logistic regression, naïve Bayes, decision tree classifier, k nearest neighbor (knn), etc., can be used for predicting heart disease (Nikhar & Karandikar, 2016).

Cardiovascular diseases are a leading cause of mortality and morbidity globally, and The Gambia is no exception. It is a significant public health concern in The Gambia, where it is one of the leading causes of morbidity and mortality (World Health Organization, 2022). According to the World Health Organization, 17.9 million global deaths were attributed to cardiovascular diseases in 2017 (Hazra et al., 2017b). Early detection and accurate prediction of CVD risk can play a key role in preventing adverse outcomes and improving patient outcomes. Machine learning (ML) algorithms have shown promise in predicting CVD risk (Pal et al., 2022), but there is a need to evaluate and compare the performance of different algorithms in the context of The Gambia.

Cardiovascular disease is a major health challenge in The Gambia however, there is no machine learning tool that can quickly diagnose and detect a cure for the disease. Despite the availability of data on various medical parameters or features such as age, sex, blood pressure, cholesterol, chest type, and others, The Gambia, a low-income country with limited resources and a high burden of cardiovascular diseases, still lacks an accurate and efficient predictive model for cardiovascular diseases. Since the rapid advancement of AI has paved the way for its potential to forecast the future. It is therefore imperative to explore its application in the health sector. Therefore, by leveraging advanced machine learning techniques and healthcare data, this research seeks to explore the development of an accurate and reliable cardiovascular disease prediction system in The Gambia. Additionally, exploring the utilization of IoT technology to address the scarcity of medical equipment in CVD detection and facilitate the collection of patient data for improved predictive accuracy is crucial. This investigation aims to explore the feasibility and benefits of these technologies in the context of CVD management.

Research Questions

This work seeks to address this main question: Can machine learning algorithms be used to improve the accuracy of cardiovascular disease prediction?

From the above main question, four (4) specific questions are derived:

Specific 1: What are the contributing factors to the occurrence of cardiovascular disease and the most effective way of diagnosing it?

Specific 2: How well can machine learning algorithms effectively predict cardiovascular disease?

Specific 3: How can an AI web application with an embedded ML model better support cardiovascular disease diagnosis?

Specific 4: How can IoT be used to mitigate the lack of medical equipment in detecting CVD and facilitate the gathering of patient data for improved predictive accuracy.

Research Hypothesis

The motivation for this work is based on the research question.

Main: Machine learning algorithms can be used to predict cardiovascular diseases.

Specific 1: Age, sex, blood pressure and resting ECG have some influence on a person's heart rate and using machine learning algorithms could be the most effective way of cardiovascular diagnosis.

Specific 2: Machine learning algorithms could most effectively diagnose and predict cardiovascular disease in an effective manner.

Specific 3: An AI web application with an embedded ML model could better support cardiovascular disease diagnosis.

Specific 4: An IoT could be used to mitigate the lack of medical equipment in detecting CVD and facilitate the gathering of patient data for improved predictive accuracy Research Objectives.

Research Objectives
Main Objectives.

The main objective of this research is to improve cardiovascular disease prediction using machine learning algorithms.

Specific 1: To identify the key factors influencing the occurrences of cardiovascular disease and the most effective way of cardiovascular disease diagnosis.

Specific 2: To assess and identify the most effective machine learning algorithm for diagnosing and predicting cardiovascular disease.

Specific 3: To develop an AI web application with an embedded ML model to enhance cardiovascular disease diagnosis support.

Specific 4: To utilize IoT to mitigate medical equipment scarcity in CVD detection and enhance predictive accuracy through improved patient data gathering.

The common cardiovascular diseases in The Gambia include Valvular Heart Disease, Ischemic Heart Disease (Coronary HD), Cardiomyopathy Heart Disease, and Hypertensive Heart Disease (Dr Lamin E.S Jaiteh). The effectiveness of machine learning algorithms in predicting cardiovascular diseases in The Gambia is the focus of this research thesis. Specifically, this study aims to compare the performance of different combined machine learning algorithms in predicting the occurrence of cardiovascular diseases and identifying the key factors contributing to these diseases' development. The study also aims to investigate the potential of IoT (arduino medical sensors) connected to Arduino Nano 33 BLE device in contributing to the patient data gathering to facilitate the prediction of cardiovascular disease.

This study is significant as it will provide valuable insights into the effectiveness of machine learning algorithms in predicting cardiovascular diseases in The Gambia, a low-income country with limited resources and a high burden of cardiovascular diseases. The findings of this research will contribute to the development of a more accurate and efficient predictive model for cardiovascular diseases, which can improve patient outcomes, reduce healthcare costs, and ultimately save lives.

To achieve the research objectives, a comparative study design will be used, involving collecting and analyzing data from various machine-learning algorithms and traditional statistical models. The data will be obtained from medical records of patients who have been diagnosed with cardiovascular diseases at the Edward Francis Small Teaching Hospital in Banjul, The Gambia.

The study will utilize diverse machine learning algorithms, including Random Forest, Support Vector Machine, and Logistic Regression, known for their high accuracy in predicting cardiovascular diseases. IoT technology will be integrated to develop a Smart CVD diagnosis IoT device, enhancing the accuracy and accessibility of healthcare services.

In conclusion, this research thesis aims to investigate the effectiveness of combined machine learning algorithms in predicting cardiovascular diseases in The Gambia, aiming to mitigate the impact of climate change on health and improve patient outcomes. The study will provide insights into the performance of different machine learning algorithms and identify the key factors contributing to the development of these diseases. The findings will significantly contribute to the development of a more accurate and efficient predictive model enabling informed decision-making by healthcare providers. Additionally, the study advances health informatics in The Gambia, fostering innovative approaches in healthcare service delivery and reducing healthcare costs.

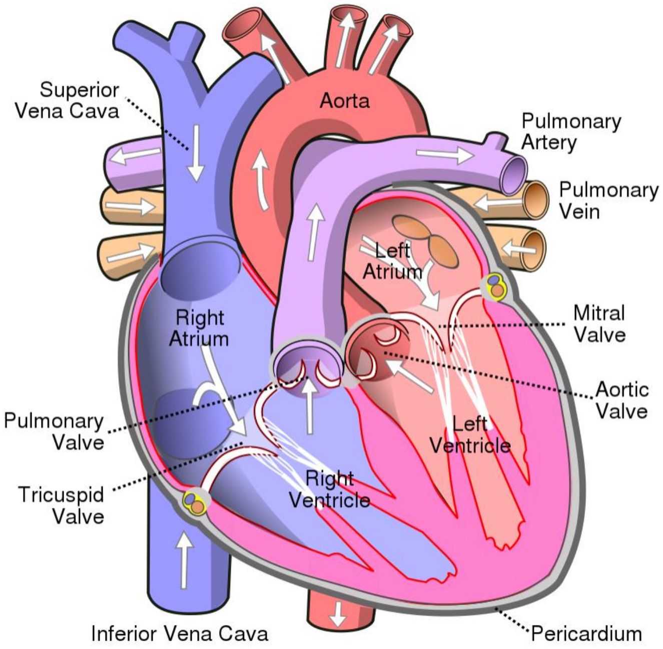


Figure 1: Anatomy of the heart. source:(Prasanth Ganesan, 2015)

CHAPTER ONE: LITERATURE REVIEW

1.1 Introduction

Cardiovascular disease (CVD) is a major public health concern worldwide, with millions of deaths reported each year. The prevalence of CVD has been increasing in recent years due to factors such as unhealthy lifestyles, smoking, and high-fat diets. Early detection and prevention of CVD are crucial to reducing the morbidity and mortality associated with this disease. In recent years, there has been growing interest in the use of Internet of Things (IoT) and machine learning techniques for CVD prediction and prevention.

This chapter presents a review of literature pertinent to the research topic of applying IoT and machine learning-based approaches to predict cardiovascular diseases. This literature review provides an overview of CVD and its risk factors, followed by a discussion of the role of IoT and machine learning in CVD prediction and prevention. It also reviews existing studies on CVD prediction using machine learning techniques, highlighting their efficacy in predicting CVD risk, and comparative analysis of combined algorithms' performance in predicting cardiovascular disease. The chapter concludes with limitations and challenges of using Machine Learning for CVD prediction and future directions for research in CVD prediction using IoT, and Machine Learning.

1.2 Overview of Cardiovascular Disease (CVD) and its Risk Factors

CVD is a broad term that encompasses several diseases of the heart and blood vessels, including coronary artery disease, stroke, and heart failure (World Health Organization, 2021). The human heart is the principal part of the human body. Essentially, it governs the circulation of blood throughout our entire body. Any irregularity in the heart can cause distress in other parts of the body. Any sort of disturbance to the normal functioning of the heart can be classified as Heart disease. In the modern era, cardiovascular disease stands as a leading cause of the majority of fatalities. Heart disease may occur due to an unhealthy lifestyle, smoking, alcohol, and high intake of fat which may cause hypertension (Nagamani et al., 2019), data mining with the MapReduce algorithm can be used for heart disease prediction. According to the World Health Organization more than 10 million die due to Heart disease every single year around the world. A healthy lifestyle and the earliest detection are the only ways to prevent the heart-related diseases.

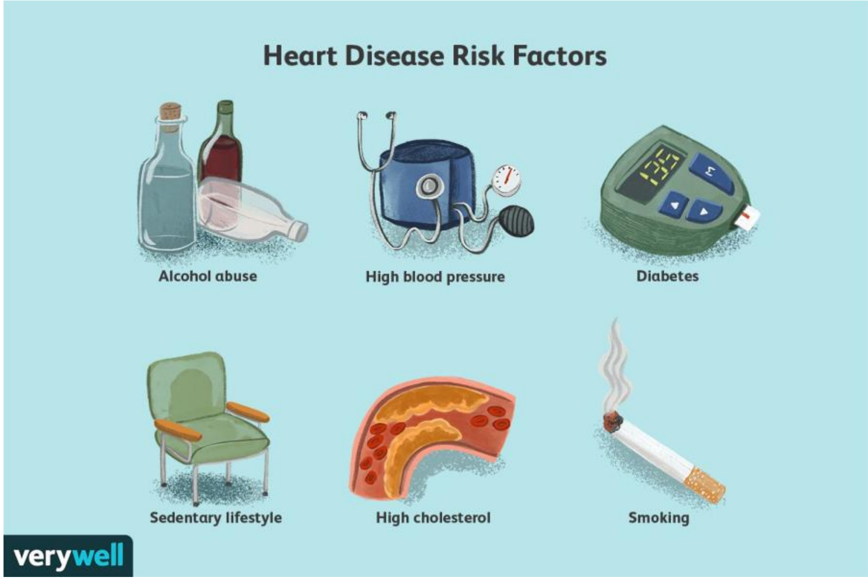


Figure 2: Heart Disease Risk Factors. source:(Jessica Olah, 2021)

1.3 The Role of IoT and Machine Learning in CVD Prediction and Prevention

The use of IoT and Machine Learning for CVD prediction and prevention has shown promising results in recent studies. IoT devices such as wearables and sensors can collect real-time data on physiological parameters such as heart rate, blood pressure, and physical activity, providing a wealth of information for CVD prediction. Machine Learning techniques can be used to analyze this data and develop predictive models for CVD risk. Additionally, Machine Learning can be used to identify patterns and associations between risk factors and CVD, leading to a better understanding of the disease (Brites et al., 2021).

In a study titled " M4CVD: Mobile Machine Learning Model for Monitoring Cardiovascular Disease (Boursalie et al., 2015)," the authors provided a comprehensive review of M4CVD, a Mobile Machine Learning Model for Monitoring Cardiovascular Disease. The system is specifically designed for mobile devices and facilitates the monitoring of cardiovascular disease using wearable sensors to collect observable trends of vital signs combined with data from clinical databases. Instead of sending raw data to healthcare professionals, the system performs analysis on the local device by feeding a hybrid of collected data to a support vector machine (SVM) to classify a patient as either "continued risk" or "no longer at risk" for CVD. The viability of an M4CVD prototype is examined through the evaluation of a synthetic clinical database comprising records of 200 patients. The results of the experiment demonstrate the system's success in classifying a patient's CVD risk with an accuracy of 90.5%.

A review by (Ahamed et al., 2022), proposed a Cardiovascular Disease Prediction System based on IoT using machine learning, which aims to predict heart diseases at an earlier stage to prevent them from occurring. This literature review discusses the potential of using IoT, machine learning, and cloud computing to analyze data related to cardiovascular diseases for early prediction and prevention. The study used a dataset of heart disease patients from Jammu and Kashmir, India, and analyzed it using numerous machines learning techniques like Random Forest, Decision Tree, Naïve based, K-nearest neighbors, and Support Vector Machine revealed the performance metrics (F1 Score, Precision and Recall) for all the techniques. The study found that Naive Bayes is better without parameter tuning while the Random Forest algorithm proved as the best technique with hyperparameter tuning. The authors further proposed a systematic model for obtaining clinical data through IoT and medical sensors for real-time prediction of cardiovascular diseases.

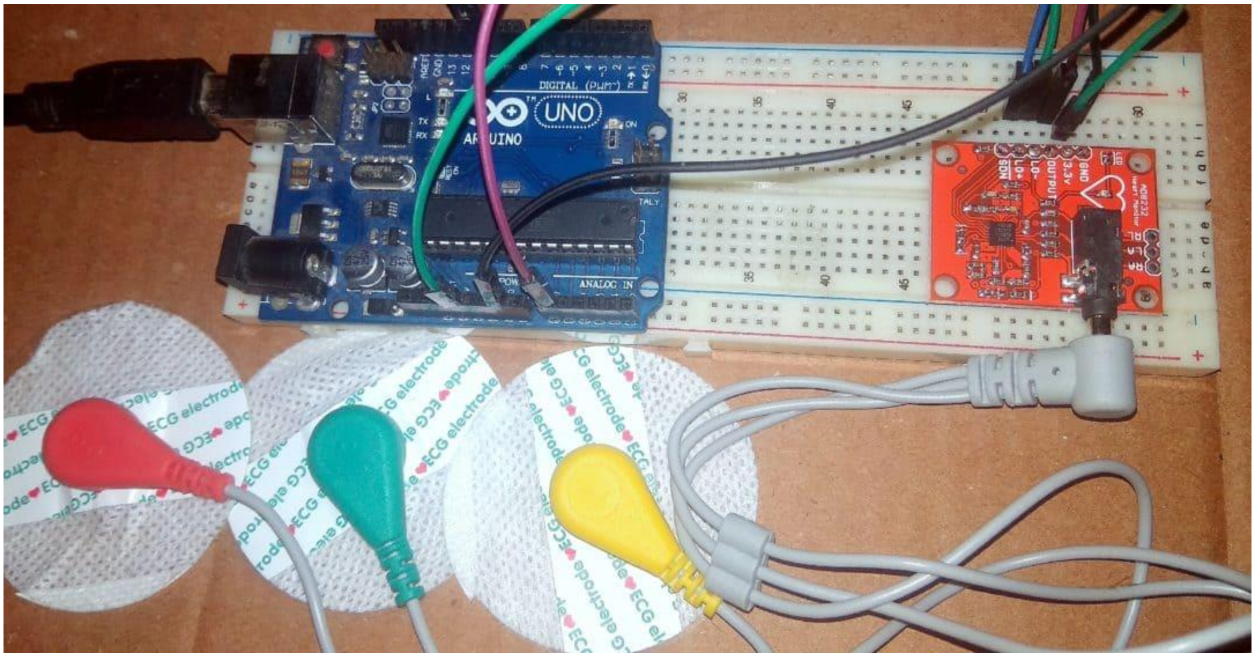


Figure 3: IoT of medical sensors. source:(Arduino projects, 2022)

1.4 Review of Existing Studies on CVD Prediction using Machine Learning Techniques

Several research studies have showcased the effectiveness of Machine Learning in predicting the risk of cardiovascular disease (CVD). For instance, a study conducted by scholars from ABES Institute of Technology in Ghaziabad, Uttar Pradesh, India (Garg et al., 2021), focused on utilizing machine learning techniques for predicting heart disease. This particular study explored the application of Machine Learning (ML) in the detection of CVDs. By considering attributes like chest pain, cholesterol level, and age, supervised ML algorithms, specifically K-Nearest Neighbor

(KNN) and Random Forest, were employed, achieving prediction accuracies of 86.885% and 81.967%, respectively. These findings highlight the efficacy of ML algorithms in diagnosing CVDs and their potential contributions to the medical field.

In another study conducted by Bhatt et al. in 2023, the promise of machine learning techniques in effectively predicting heart disease was demonstrated. The researchers developed a model aimed at accurately predicting cardiovascular diseases to reduce fatality rates associated with these conditions. Their research proposed a method called k-modes clustering with Huang starting, which aimed to improve classification accuracy. Various machine learning models, namely random forest (RF), decision tree classifier (DT), multilayer perceptron (MP), and XGBoost (XGB), were employed. The GridSearchCV technique was utilized to fine-tune the models' parameters. Additionally, the developed model was tested on a real-world dataset consisting of 70,000 instances sourced from Kaggle. The models underwent training using an 80:20 data split and yielded the subsequent accuracy rates: decision tree - 86.37% (with cross-validation) and 86.53% (without cross-validation), XGBoost - 86.87% (with cross-validation) and 87.02% (without cross-validation), random forest - 87.05% (with cross-validation) and 86.92% (without cross-validation), multilayer perceptron - 87.28% (with cross-validation) and 86.94% (without cross-validation). The evaluation of the proposed models yielded impressive AUC (area under the curve) values: decision tree - 0.94, XGBoost - 0.95, random forest - 0.95, and multilayer perceptron - 0.95. The study concluded that the multilayer perceptron with cross-validation outperformed all other algorithms in terms of accuracy, achieving the highest accuracy rate of 87.28%.

1.5 Overview of Machine Learning

Machine learning, a branch of computer science, is currently a rapidly growing and highly relevant field, with further significant growth anticipated in the future. Our world is flooded with data and data is being created rapidly every day all around the world. CSC, a company specializing in Big Data and Analytics Solutions, predicts that by the year 2020, the volume of data will increase by a factor of 44 compared to 2009.

Therefore, it is necessary to understand data and gain insights for a better understanding of the human world. The sheer volume of data in today's world is immense, making it impractical to rely on traditional methods. Analyzing data or building predictive models manually is almost impossible in some scenarios and is also time-consuming and less productive. Machine learning, the on other hand, produces reliable, repeatable results and learns from earlier computation.

1.5.1 Machine Learning Algorithms

To facilitate comparative analysis, this study explores five different Machine Learning algorithms. The different Machine Learning (ML) algorithms Logistic Regression (LR), Decision Tree Classifier (DT). Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM). The reason to choose these algorithms is based on their popularity and accuracy and ability to be applicable on classification tasks.

To test the goodness of each model, we focus on prediction accuracy and also pay attention to efficiency. Among multiple approaches, we prefer ones that provide large numbers of correct predictions with less complexity at the same time.

1.5.2 Logistic Regression

This is the prevailing approach employed by individuals for prediction. Logistic regression, also called logit regression or logit model, is a parametric regression method for predictive analysis. It calculates the likelihood of a categorical outcome happening by considering one or more predictors.

1.5.3 Decision Tree

Decision trees are a commonly used supervised learning approach for solving classification and regression problems. A tree-like structure is created where inner nodes represent dataset characteristics, branches represent decision rules, and leaf nodes represent results. The decision-making process takes place through decision nodes with multiple branches, while leaf nodes represent the final result with no further branches. Decisions and tests are made to determine the appropriate path through the tree by evaluating the characteristics of the dataset. This graphic representation, like the branches of a tree, offers different possible solutions or decisions based on given conditions. The tree is built using the CART algorithm, which stands for Classification and Regression Tree Algorithm. A decision tree starts at a root node, goes through subsequent branches, and produces subtrees based on the answers to the questions.

1.5.4 Random Forest

Classification and regression tree (CART) model is a decision tree algorithm used for regression or classification prediction. It pertains to the iterative division of the input space and the creation of a localized model within each resulting region of the input space. Each CART unit can be represented by a binary tree with one leaf per region. Each leaf node corresponds to a segment of values for the independent variable and each leaf gives a prediction for the dependent variable value. CART algorithm is different from the logistic regression method. Logistic regression algorithm makes classifications based on the relationship between response and predictors, whereas, for decision trees, the idea is to divide the dataset into smaller sections according to certain rules, until a small enough set is reached that data points in it fall under the same label.

1.5.5 Gradient Boosting Machine (GBM)

Boosting method is similar to the random forest algorithm as both of these two serves stronger versions of the CART algorithm. In distinguishing between the two methods, it is notable that the random forest approach enhances model performance through data resampling, whereas the boosting method accomplishes the same objective by reweighing the data.

1.5.6 Support Vector Machine

Support Vector Machine (SVM) is a widely used supervised learning algorithm renowned for its effectiveness in both classification and regression tasks. However, its primary application lies in solving classification problems within the field of machine learning.

The main objective of the SVM algorithm is to construct an optimal line or decision boundary, known as a hyperplane, that effectively separates data points in an n-dimensional space into distinct classes. This process enables the accurate categorization of new data points in the future.

SVM achieves this by selecting the most critical points or vectors, known as support vectors, which play a crucial role in defining the hyperplane. The algorithm derives its name from these support vectors.

1.6 Introduction to Combined Algorithmic Performance and its Significance in CVD Prediction

Combined algorithmic performance refers to the utilization of multiple Machine Learning algorithms to enhance predictive accuracy. The diagnosis of heart disease involves numerous factors, which can present challenges for physicians. To aid physicians in making prompt and accurate diagnoses, classification systems have been developed. These systems employ models that classify medical data based on sample information, allowing for thorough examination of patient records (Anbarasi & Anupriya, 2010). Multiple classification algorithms have been developed and utilized as aids for doctors in the diagnosis of patients with heart disease (Li et al., 2022). Traditional diagnostic methods prove inadequate in dealing with such a complex disease, emphasizing the need for a medical diagnostic system that incorporates feature selection approaches for disease prediction and analysis (Tülay Karayilan et al., 2017)

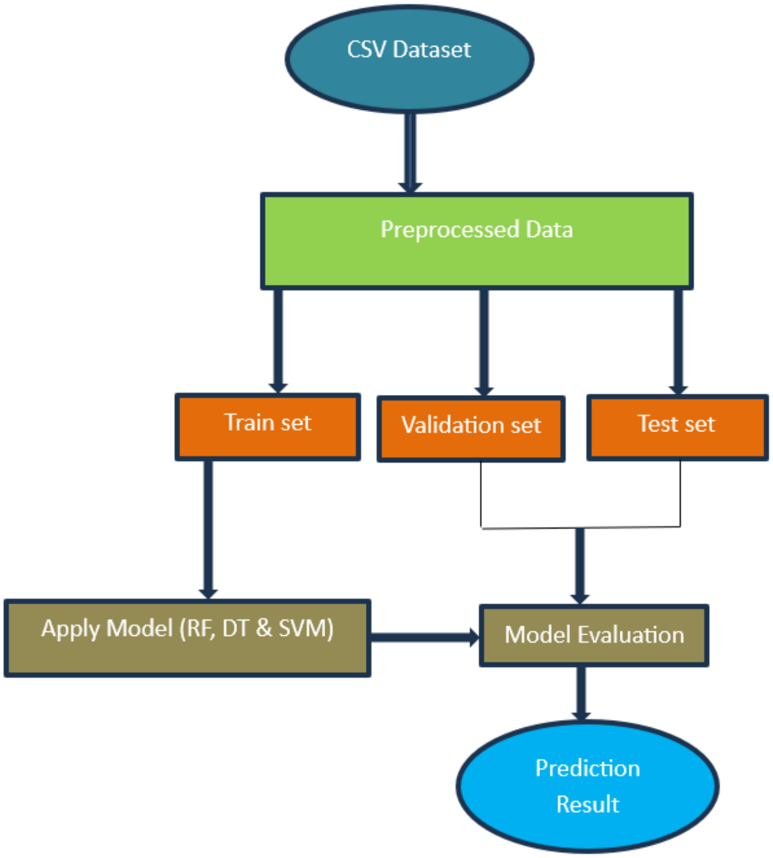


Figure 4: Model Roadmap

1.7 Comparison of Various Combined Machine Learning Algorithms for CVD Prediction

A systematic review by (Asif et al., 2021), compared the performance of various ML algorithms with combined algorithms, in predicting CVDs. Both hard and soft voting combined classifiers (EVCH and EVCS) have demonstrated an accuracy of 92%, surpassing other algorithms in terms of performance. Moreover, these two methods have also displayed better results in terms of other performance metrics. Therefore, these classifiers can be implemented practically to predict patients with cardiovascular disease to ease the diagnosis process and reduce human-made errors. The review also suggested that combining different types of ML algorithms can further improve prediction accuracy.

1.8 Future directions for research in CVD prediction using IoT and Machine Learning

Despite the promising results from previous studies, there is still a need for further research in CVD prediction using IoT and ML. (Moshawrab et al., 2023), conducted a systematic literature review on smart wearables for the detection of cardiovascular diseases. The use of smart wearables for cardiovascular disease detection is a reality, but there is a need for more research and development to improve the process and take proactive and preventive measures. Future perspectives in this field include preserving data privacy and confidentiality through federated learning, removing artifacts and improving data readiness through automation of noise reduction, analyzing heterogeneous and diverse data through multimodal machine learning, and enhancing accuracy, privacy, and explainability to raise trust and wider adoption by users.

In addition, it is of the believe that future directions for cardiovascular disease (CVD) prediction include the investigation of wearable devices, which can monitor vital signs and provide real-time feedback to patients and healthcare providers. Additionally, big data analytics can be explored to analyze large volumes of data from various sources, such as electronic health records and lifestyle factors, to identify new risk factors for CVDs. Lastly, research on the ethical and legal implications of using IoT and ML for CVD prediction is necessary, including privacy concerns, data ownership, and data sharing.

CHAPTER 2: MATERIALS AND METHODS

2.1 Introduction

In order to achieve the study's goals of confirming the hypothesis, fulfilling the research questions, and attaining the objectives, a range of materials and methods were employed.

This chapter is made up of 4 sections: study area, data, tools, and methods. First of all, there will be a presentation on the geography of the study area and its climate. Secondly, the data used in this research would be mentioned. After that, all the tools: programming language, interfaces, environments and libraries, and other software used would be pointed out. The study culminated in the creation of an AI Web Application for predicting cardiovascular disease, data storage, and data analytics, including data visualization. Additionally, the study explores the creation of a Smart CVD diagnosis IoT device to facilitate patient data gathering. This chapter would provide an overview of a detailed description of the steps followed to reach the objectives of this thesis. Figure 5 depicts the conceptual framework of the study.

Initially, the research focused on selecting The Gambia as the study area, followed by data collection and analysis.

2.2 Conceptual Framework

The aim of this research is to accurately determine the presence of heart disease in patients as well as the use of Smart CVD diagnosis IoT device for patient data collection. Healthcare providers input the patient's health report data into a model that calculates the probability of cardiovascular disease occurrence.. Figure 5 shows an in-depth of the entire process involved.

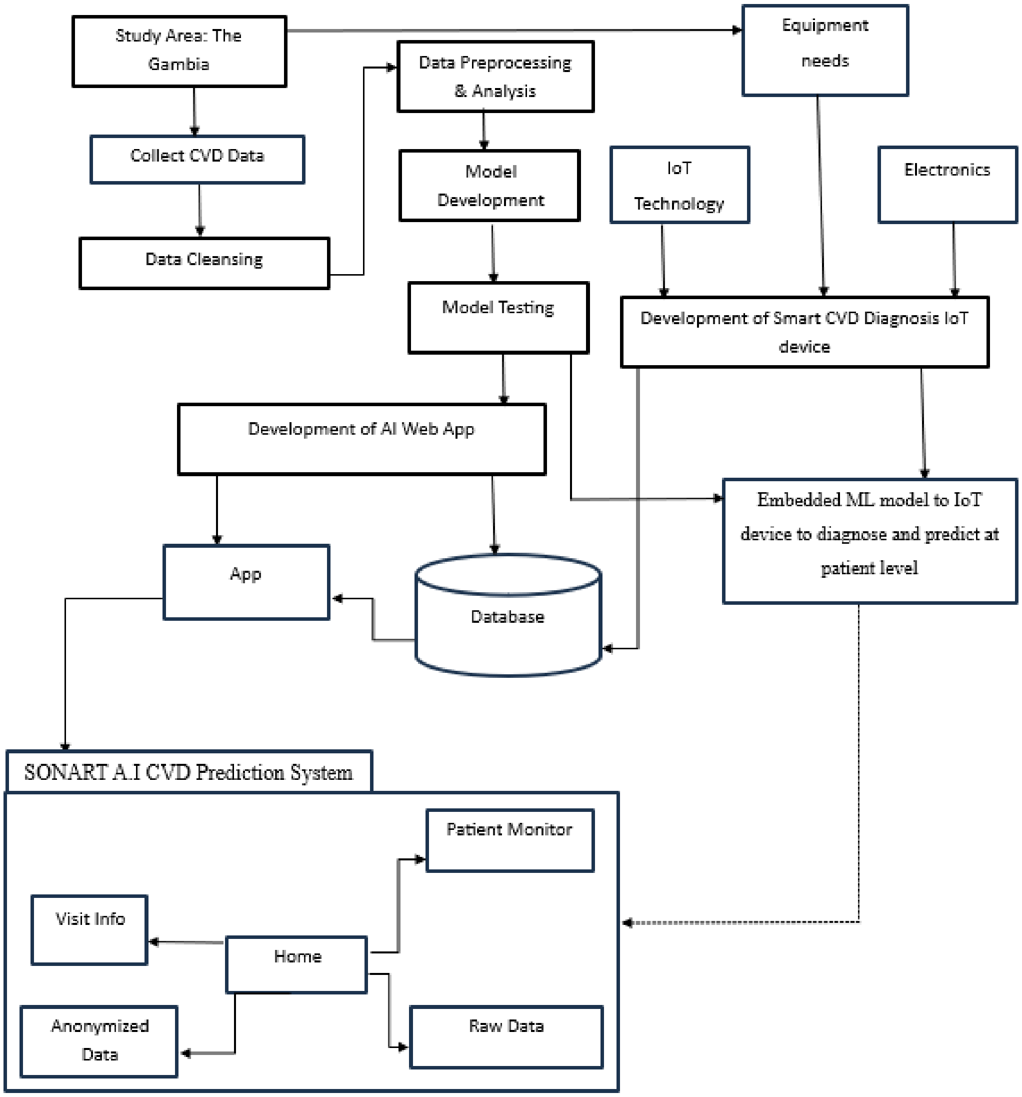

Study Area: The Gambia
Equipment needs
Data Preprocessing & Analysis
Collect CVD Data
Model Development
IoT Technology
Electronics
Data Cleansing
Model Testing
Development of Smart CVD Diagnosis IoT device
Development of AI Web App
Embedded ML model to IoT device to diagnose and predict at patient level
App
Database
SONART A.I CVD Prediction System
Patient Monitor
Visit Info
Home
Anonymized Data
Raw Data

Figure 5: Conceptual Framework

2.3 Study Area

2.3.1 Geography

The Gambia, a small country located in West Africa, is surrounded by Senegal on all sides, except for its western coast, which borders the Atlantic Ocean. However, it has its own unique character and identity, with its capital city of Banjul serving as the political and economic center of the country. One of the most important pieces of information regarding the location of The Gambia is its latitude and longitude coordinates. The country is located at 13.4432° N and 15.3101° W, while Banjul specifically is located at 13.4549° N and 16.5790° W. The country occupies an area of approximately 11,300 square kilometers. As of 2023, it has a population of about 2.7 million inhabitants. It is characterized by a diversity of environments, including savannah lands, tropical rainforests, and savannah forests. Savannah lands make up about 23% of the total area, while rainforests and savannah forests cover around 16%. The country's highest point reaches only 53 meters above sea level, and the coastal region is characterized by sandy beaches, while the inland areas feature grasslands and low hills.

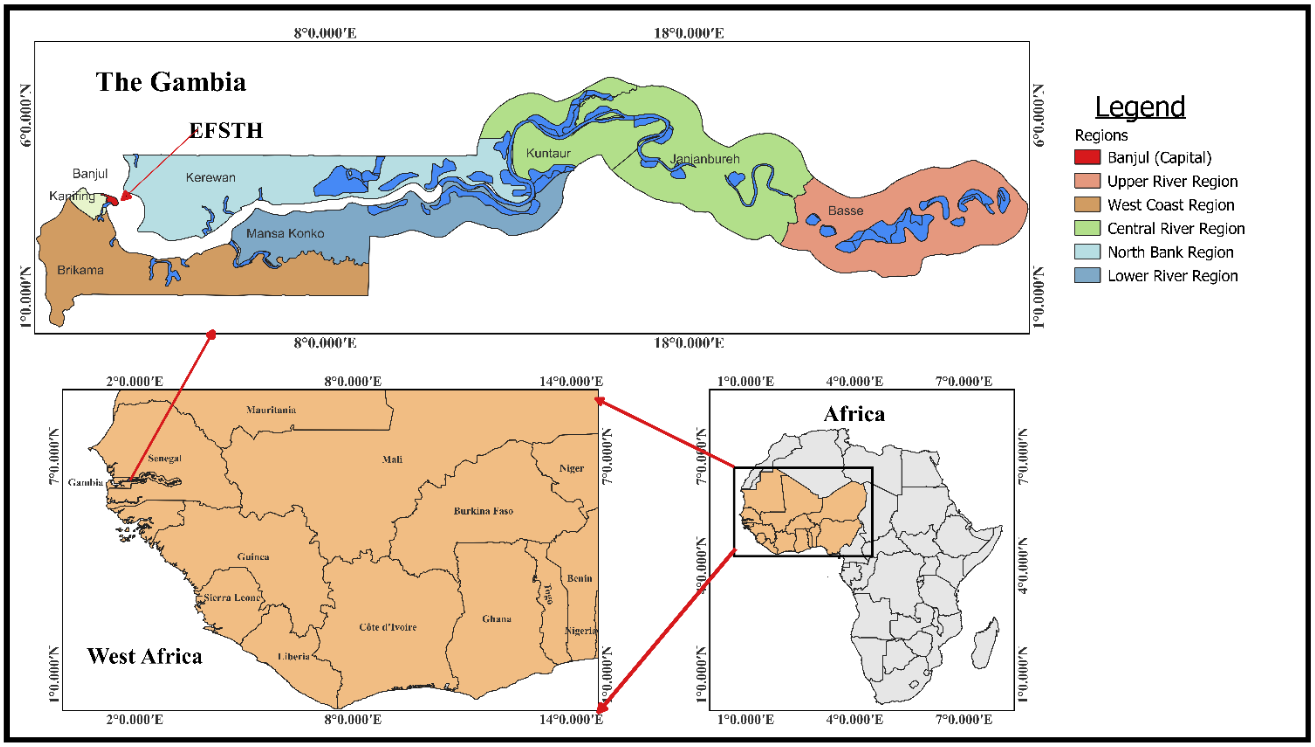
2.3.2 Climate

The Gambia has a tropical climate that is characterized by distinct wet and dry seasons. The country's location on the coast and its proximity to the equator greatly influence its climate, making it relatively consistent throughout the year.

In The Gambia, the rainy season typically runs from June to October, with an average rainfall of around 1,200 millimeters per year in the southern part of the country and around 900 millimeters in the north. During this time, the country experiences heavy rains and thunderstorms, often resulting in flooding in some areas. The wet season is characterized by high humidity levels, with temperatures ranging from around 25°C to 30°C.

The dry season in The Gambia runs from November to May, with almost no rainfall. During this time, the country experiences lower humidity levels and temperatures that range from around 20°C to 34°C. The Harmattan, a dry and dusty wind that blows from the Sahara Desert, is common during this period, often causing haze and reduced visibility.

The Gambia's temperature remains relatively consistent throughout the year, with an average temperature of around 27°C. The country's proximity to the equator means that it experiences only small variations in temperature, with daytime temperatures ranging from around 28°C to 32°C and nighttime temperatures ranging from around 18°C to 24°C.



Author: Muhammed Fatty

Figure 6: Map of the study area

2.4 Dataset and Data Source

This research will utilize historical patient data exclusively from Edward Francis Small Teaching Hospital (EFSTH), located in Banjul, without incorporating data from any private clinics or hospitals. Due to the absence of a structured database for storing medical records digitally, all patient files are maintained individually and housed on shelves at the hospital. This has led to a cumbersome and stressful data collection process since it requires manually searching through patient files.

The research utilizes a dataset that includes 9 cardiovascular disease attributes with 915 rows as the dataset size. Table 1 provides a summary of the dataset and its descriptions.

Sl. No.	Attribute Description	Distinct Values of Attribute
1.	*Age*- represents the age of a person	Multiple values between 4 & 104
2.	*Temperature*- represents the body temperature of the patient	32 - 39
3.	*Sex*- describe the gender of the person (0-Female, 1-Male)	0- female, 1- male
4.	*Pulse*- shows the max heartbeat of the patient in a minute	Multiple values from 31 to 168
5.	*Chest Pain*- represents chest pain of the patient	0- No chest pain, 1- for chest pain
6.	*Blood Pressure*- It represents the patient's BP.	categorized into 4 stages 0-NORMAL, 1-ELEVATED, 2-HYPERTENSION STAGE 1 AND 3-HYPERTENSION STAGE 2
7.	*Sp02 (Oxygen saturation)*- an indirect measurement of oxygen saturation/concentration in the blood	60 to 100
8.	*Resting ECG*-It shows the result of ECG	0- Normal ECG, 1- Abnormal ECG
9.	*Target*-It is the final column of the dataset. It is a class or label, or Column. It represents the number of classes in the dataset. This dataset has binary classification i.e. two classes (0,1). In class "0"	0,1

	represent there is less possibility of heart disease whereas “1” represent a high chance of heart disease. The value “0” Or “1” depends on the other 8 attributes.	

Table 1: Dataset

2.5 Data Processing And Analysis

2.5.1 Data Preparation

The patient data from the cardiology unit was initially stored in an Excel Open XML Spreadsheet (XLSX) format. To facilitate further analysis, the raw data underwent a data cleaning procedure and were subsequently transformed into Common Separated Values (CSV) format. During the data cleaning and preprocessing stage, any missing or incomplete records that could not be utilized in the analysis were discarded. This step ensured the integrity and quality of the dataset.

To gain a comprehensive understanding of the dataset, exploratory data analysis (EDA) was performed. EDA is a method of analyzing and summarizing data sets to comprehend their main characteristics. It involves employing statistical and visualization techniques to uncover patterns, relationships, anomalies, validate assumptions, and develop hypotheses. The primary objective of EDA is to extract meaningful insights from the data and acquire a deep understanding of its underlying structure and distribution. This understanding enables the construction of appropriate statistical models that can make accurate predictions or decisions.

2.6 Exploratory Data Analysis

2.6.1 Descriptive Statistics of the Dataset

These descriptive statistics offer valuable insights into the dataset, enabling a comprehensive understanding of the variables under investigation. They form the foundation for further analysis and interpretation, guiding the formulation of hypotheses and the identification of patterns or relationships among the variables.

Table 2 presents descriptive statistics for the variables included in the dataset, providing valuable insights into the characteristics and distribution of the data. A comprehensive understanding of these statistics is crucial for interpreting the results and drawing meaningful conclusions. The following is a detailed explanation of each column:

Age: The dataset consists of 915 observations of individuals with a wide range of ages, from a minimum of 4 years to a maximum of 104 years. It has an average age of approximately 53.54 years (SD = 18.96). The age distribution is fairly spread out, with 25% of the participants being below 40 years and 25% being above 68 years.

The mean age of the individuals is approximately 53.54 years, with a standard deviation of 18.96, indicating a moderate level of dispersion around the mean.

The dataset captures a broad age range, allowing for the exploration of age-related patterns and their impact on the target variable.

Sex: The dataset includes information about the gender of the participants. The variable takes on binary values, with 0 representing females and 1 representing males. The data indicate that approximately 49.18% of the participants are females.

Temp: This variable represents the body temperature of the participants. The average temperature recorded is approximately 35.95 (SD = 0.60), with the minimum and maximum values being 32 and 39, respectively. The temperature data exhibit limited variability, suggesting a relatively stable body temperature among the participants.

Pulse: The pulse rate of the participants ranges from 31 to 168 beats per minute, with an average pulse rate of approximately 89.93 (SD = 20.06). The distribution of pulse rates indicates considerable variability among the participants, with some individuals having significantly higher or lower rates.

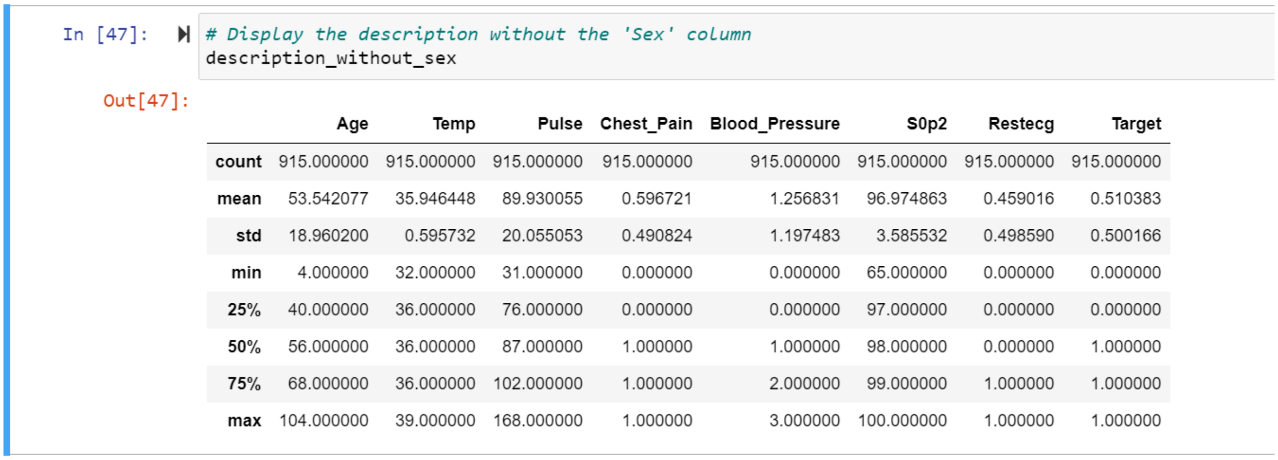
Chest_Pain: This variable indicates the presence or absence of chest pain among the participants. A value of 0 denotes the absence of chest pain, while a value of 1 signifies its presence. The data show that approximately 59.67% of the participants experience chest pain.

Blood_Pressure: The participants' blood pressure readings have an average value of approximately 1.2568. The minimum and maximum blood pressure values recorded are 0 and 3, respectively. The data indicate variability in blood pressure levels, reflecting differences among the participants.

Sp02: The variable denoted as Sp02 in this study represents the oxygen saturation level in the blood. Its maximum value is 100%, while the mean value is calculated to be 65%, with a corresponding standard deviation of 3.585. The dataset comprises 915 observations, and the average oxygen saturation level is recorded as 96.97%.

Restecg: The variable represents the results of resting electrocardiograms for the participants. A value of 0 corresponds to a normal result, while values of 1 indicate abnormal results. The data reveal that approximately 45.90% of the participants have abnormal resting electrocardiographic results.

Target: This variable signifies the presence or absence of heart disease among the participants. A value of 0 denotes the absence of heart disease, while a value of 1 indicates its presence. The data show that approximately 51.04% of the participants in the dataset have heart disease.



```
In [47]:  # Display the description without the 'Sex' column
          description_without_sex
```

Out[47]:

	Age	Temp	Pulse	Chest_Pain	Blood_Pressure	S0p2	Restecg	Target
count	915.000000	915.000000	915.000000	915.000000	915.000000	915.000000	915.000000	915.000000
mean	53.542077	35.946448	89.930055	0.596721	1.256831	96.974863	0.459016	0.510383
std	18.960200	0.595732	20.055053	0.490824	1.197483	3.585532	0.498590	0.500166
min	4.000000	32.000000	31.000000	0.000000	0.000000	65.000000	0.000000	0.000000
25%	40.000000	36.000000	76.000000	0.000000	0.000000	97.000000	0.000000	0.000000
50%	56.000000	36.000000	87.000000	1.000000	1.000000	98.000000	0.000000	1.000000
75%	68.000000	36.000000	102.000000	1.000000	2.000000	99.000000	1.000000	1.000000
max	104.000000	39.000000	168.000000	1.000000	3.000000	100.000000	1.000000	1.000000

Table 2: Descriptive Statistics

2.6.2 Percentage of Cardiovascular Disease Patients in The Dataset

The provided visualization presents the percentage distribution of cardiovascular and normal patients within the dataset. Additionally, it provides the exact count of cardiovascular and normal patients, offering a quantitative understanding of the dataset composition.

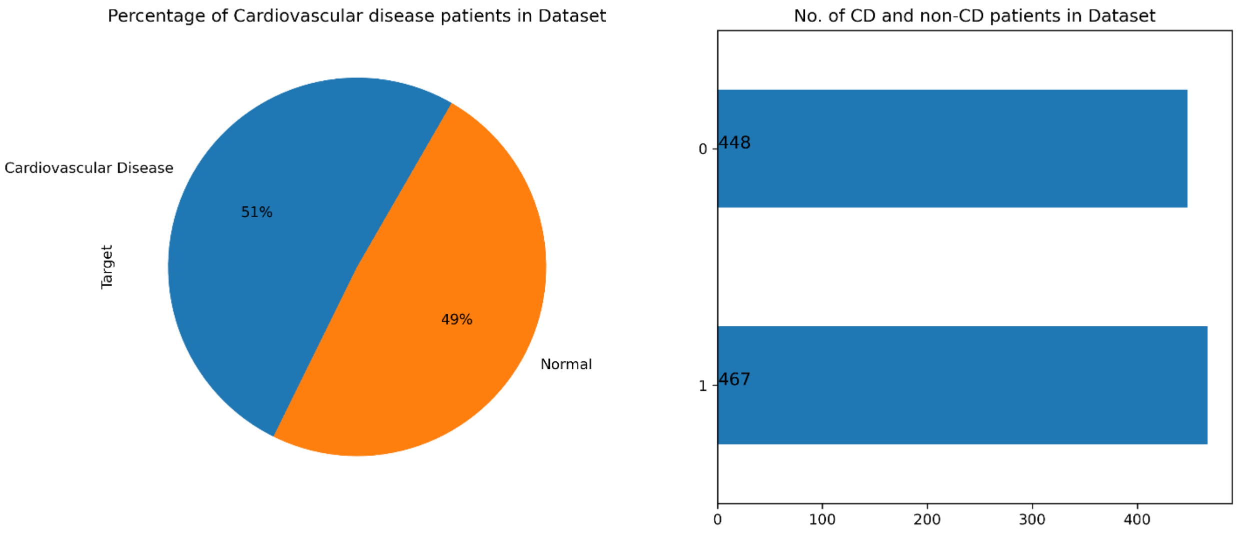


Figure 7: Percentage of CVD Patients in the dataset

2.6.3 Checking Gender and Age Wise Distribution

The depicted plot illustrates the percentage distribution of gender within the dataset, revealing a significantly higher representation of females compared to males. Additionally, the plot showcases the age-wise distribution of patients, indicating an average age of approximately 65 years.

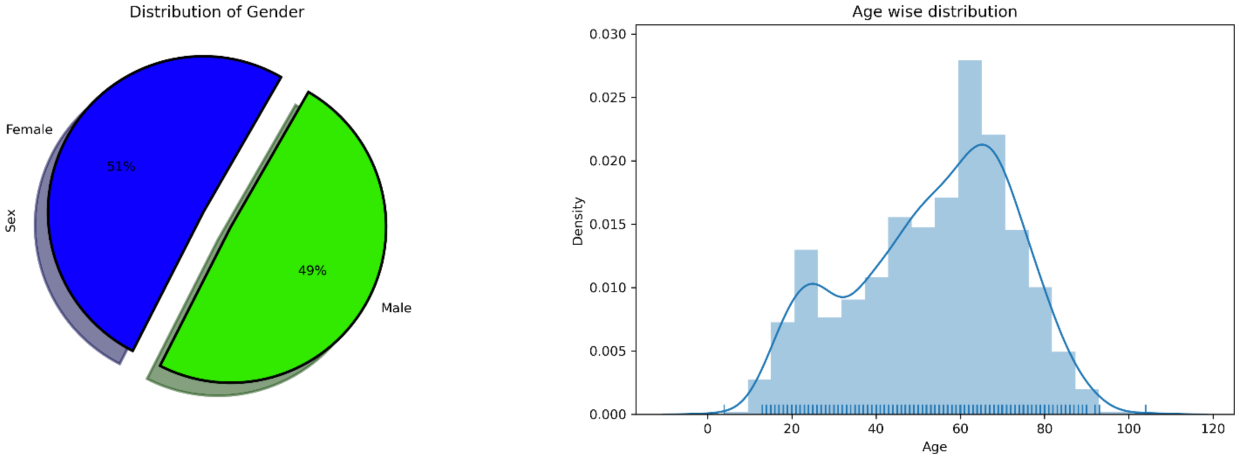


Figure 8: Gender and Age-Wise Distribution

2.6.4 Gender distribution of normal and cardiovascular disease patients

The presented figure displays the gender distribution of both normal and cardiovascular disease patients within the dataset. Among both groups, females exhibit a higher representation compared to males.

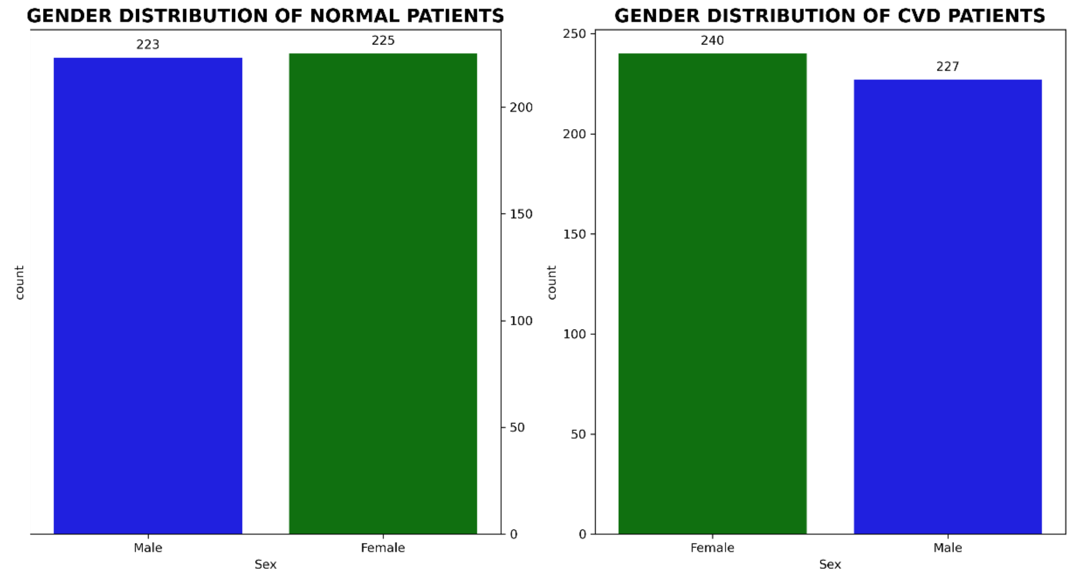


Figure 9:: Gender Distribution of both normal and cardiovascular disease patients

2.6.5 Age Distributions of Patients

Figure 18 visually compares the age distributions of normal patients and those with cardiovascular disease. It enables us to observe whether there are differences in the age profiles between these two groups. If the orange curve is shifted to the right of the blue curve, it indicates that cardiovascular patients tend to be older, while a shift to the left suggests they are younger. This visualization helps identify potential age-related risk factors for cardiovascular disease and enhances our understanding of its relationship with age.

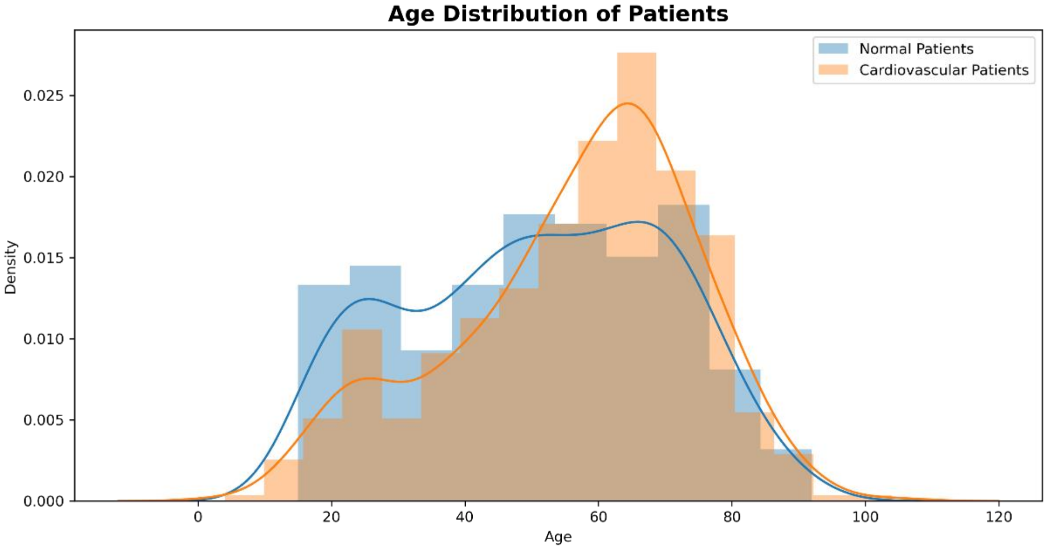


Figure 10: Age Distribution

2.6.6 Cardiovascular Disease Frequency for Ages

This visual representation displays the frequency of cardiovascular disease across different age groups. The ages are depicted on the horizontal axis, while the frequencies are represented on the vertical axis. Each age group is further divided into male and female categories, indicating the number of individuals affected by cardiovascular disease.

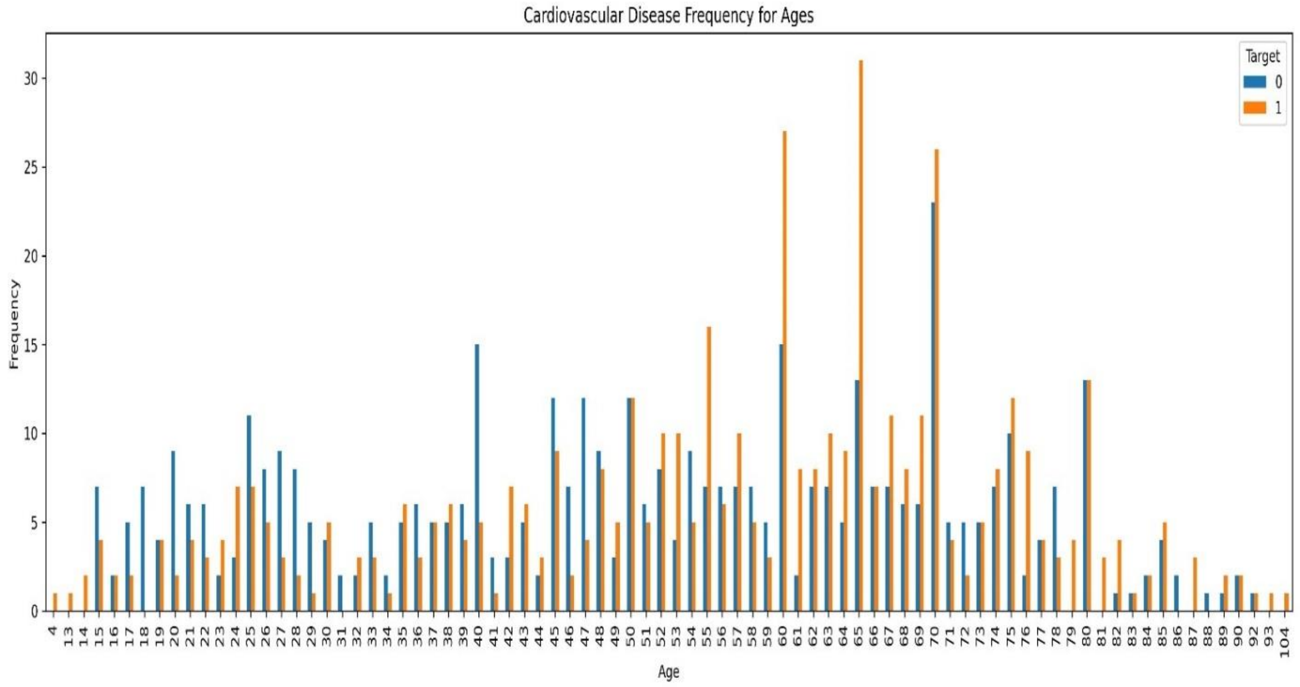


Figure 11: CVD Frequencies for Ages

2.7 Data Analysis and Modeling

In this section, a variety of combined machine learning algorithms, including Random Forest, Decision Tree, Logistic Regression, support vector machine, and Gradient Boost, were employed to analyze the input attributes (x) as listed in Table 1. The objective was to determine whether or not the heart is defective by examining the target (y) output. To assess the performance of each algorithm, the input dataset was divided into two subsets, with 80% of the data used for training the models, the remaining 20% for validation testing and a completely different dataset for testing which represents 20% of the training dataset. The data analysis and modeling tasks were carried out using the Python programming language, a widely-used tool for machine learning and data analysis.

2.8 Tools

Many tools/materials were used for the purpose of this applied research. Especially open-source tools and materials such as electronic components were used for the design and implementation of the SONART AI Cardiovascular Disease Prediction system and the smart CVD diagnosis IoT device.

Different tools are used for this study. All these options are available at no cost and are based on open-source principles.

1. Python 3.10-64 – programming language

2. NumPy 1.22.4 - for managing multidimensional arrays

3. Matplotlib 3.7.1 is a versatile tool used to generate visual representations, including static, animated, and interactive visualizations.

4. Pandas 1.5.3 - for managing data structures

5. Seaborn 0.11.2 is a widely used Python data visualization library that is developed on the foundation of matplotlib. It offers a user-friendly interface for generating visually appealing and informative statistical graphics. This includes a wide range of visualizations such as heatmaps, scatterplots, bar plots, and various other types of plots.

6. SciPy and Scikit-learn 1.0.2 - for normalizing data, and Seaborn for creating informative and attractive statistical graphics.

7. Flask 2.2.3 - a web framework that provides you with tools, libraries, and technologies that allow you to build a web application.

8. Joblib 1.2.0 - build & save an combined algorithmic model

9. VS Code Version 1.77.2: a source code editor developed by Microsoft, providing a range of features for writing and debugging code.

10. Jupyter Notebook 6.4.8: an open-source web application that allows for interactive computing and data visualization.

11. XAMPP v3.3.0: a cross-platform web server solution that includes Apache, MySQL, PHP, and Perl, used for hosting web applications.

12. Google Chrome Version 112.0.5615.49 (Official Build) (64-bit)

13. Arduino medical sensors

14. Arduino: an open-source electronics platform that allows for the creation of interactive electronic projects, including medical sensors.

15. Zotero: a reference management tool and academic social network used for organizing references and staying updated on research.

16. Microsoft Word: a word processing program used for writing the thesis report.

17. Laptop computer: The specification of the system is as follows: model Dell 11th Gen Intel(R) Core(TM) i7-1185G7 @ 3.00GHz, RAM 16GB. System type (64-bit operating system, x64-based processor), Windows 11 Pro

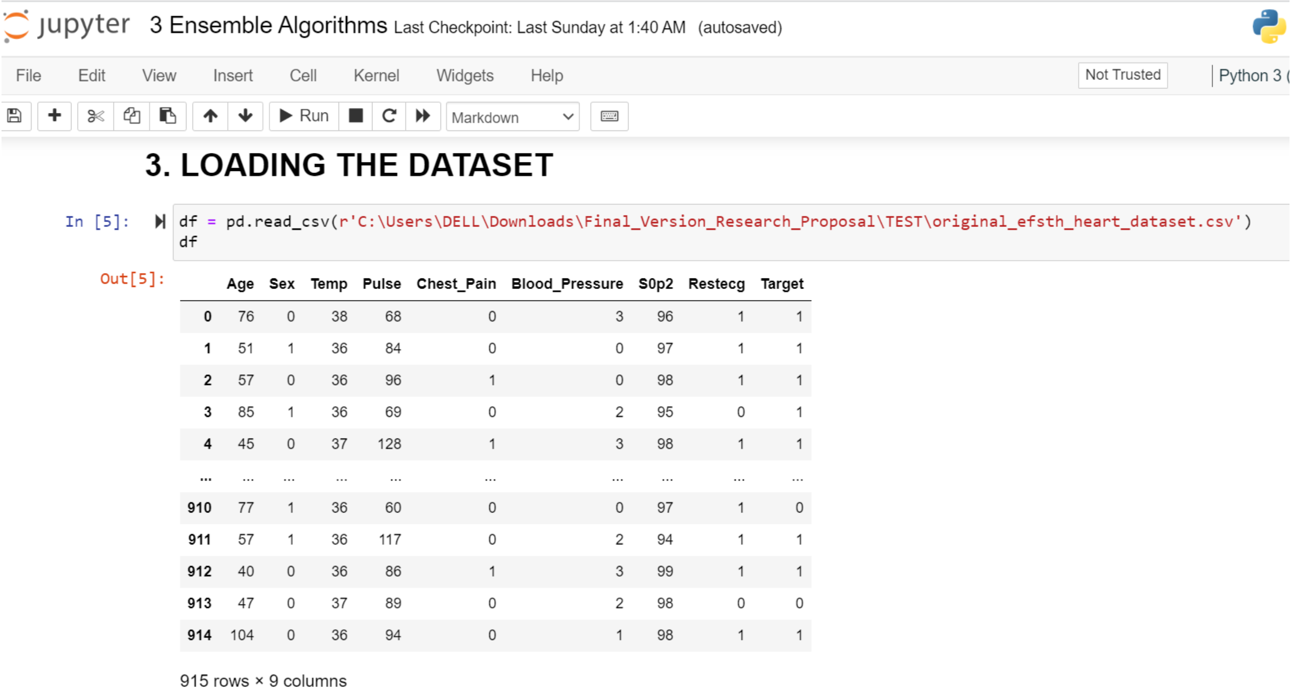
This study primarily relied on open-source tools, with Python serving as the chosen interpreted, high-level programming language. Jupyter Notebook, an open document format rooted in JSON, facilitated interactive computing. Anaconda, a Python distribution focused on scientific computing, was utilized for streamlined package management and deployment.

2.9 Implementation steps

The implementation of the thesis project was conducted using the Python programming language within the Anaconda Navigator's Jupyter Notebook environment. Python was chosen as the language due to its versatility and extensive libraries support for machine learning tasks. Jupyter Notebook was selected as the development platform because of its faster execution compared to other Python IDE tools such as PyCharm or Visual Studio, especially for implementing machine learning algorithms. Additionally, Jupyter Notebook provides convenient features for data visualization and graph plotting, including histogram and heatmap visualization for correlated matrices.

The implementation steps of the project were as follows:

a) Dataset Collection: The required datasets were collected from Edward Francis Small Teaching Hospital (EFSTH) to carry out the analysis and modeling tasks.



3. LOADING THE DATASET

```
In [5]: df = pd.read_csv(r'C:\Users\DELL\Downloads\Final_Version_Research_Proposal\TEST\original_efsth_heart_dataset.csv')
        df
```

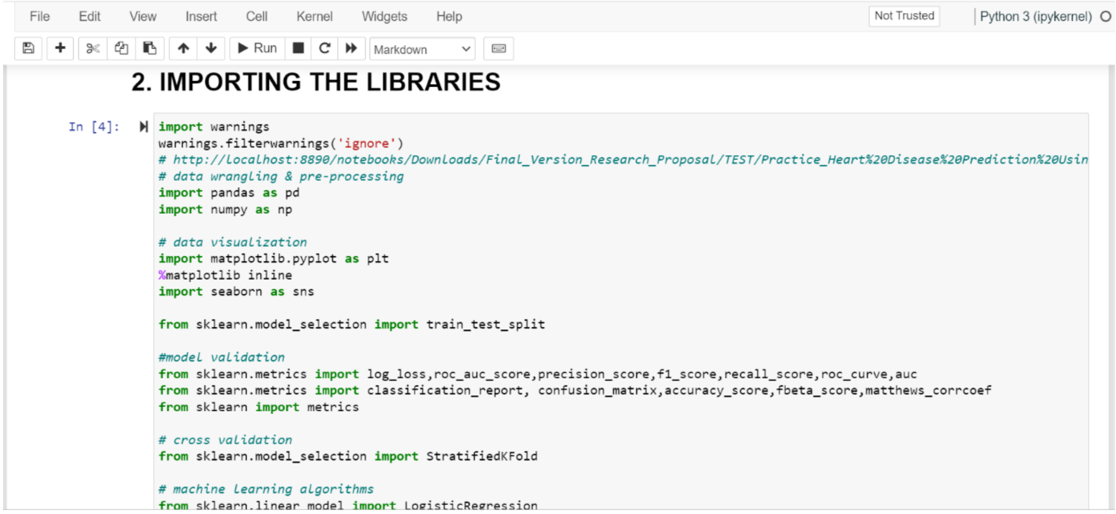
Out[5]:

	Age	Sex	Temp	Pulse	Chest_Pain	Blood_Pressure	S0p2	Restecg	Target
0	76	0	38	68	0	3	96	1	1
1	51	1	36	84	0	0	97	1	1
2	57	0	36	96	1	0	98	1	1
3	85	1	36	69	0	2	95	0	1
4	45	0	37	128	1	3	98	1	1
...	...	...	...	...	...	...	...	...	...
910	77	1	36	60	0	0	97	1	0
911	57	1	36	117	0	2	94	1	1
912	40	0	36	86	1	3	99	1	1
913	47	0	37	89	0	2	98	0	0
914	104	0	36	94	0	1	98	1	1

915 rows × 9 columns

Figure 12: Loading the dataset

b) Importing Libraries: The necessary libraries, including Numpy, Pandas, Scikit-learn, Matplotlib, and Seaborn, were imported to facilitate various operations throughout the project.



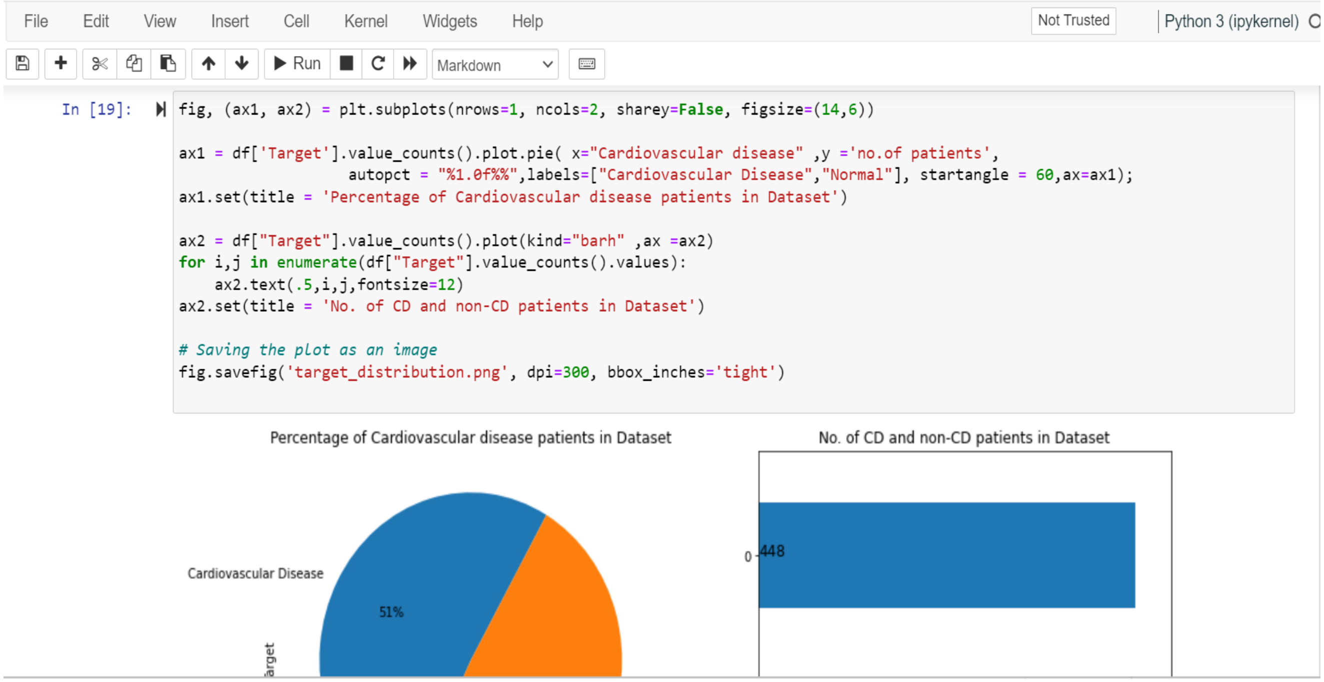
2. IMPORTING THE LIBRARIES

```
In [4]: import warnings
        warnings.filterwarnings('ignore')
        # http://localhost:8890/notebooks/Downloads/Final_Version_Research_Proposal/TEST/Practice_Heart%20Disease%20Prediction%20Usin
        # data wrangling & pre-processing
        import pandas as pd
        import numpy as np

        # data visualization
        import matplotlib.pyplot as plt
        %matplotlib inline
        import seaborn as sns

        from sklearn.model_selection import train_test_split

        #model validation
        from sklearn.metrics import log_loss,roc_auc_score,precision_score,f1_score,recall_score,roc_curve,auc
        from sklearn.metrics import classification_report, confusion_matrix,accuracy_score,fbeta_score,matthews_corrcoef
        from sklearn import metrics

        # cross validation
        from sklearn.model_selection import StratifiedKFold

        # machine learning algorithms
        from sklearn.linear_model import LogisticRegression
```

Figure 13: Importing Librairies

c) Exploratory Data Analysis: Exploratory data analysis techniques were employed to gain deeper insights into the data and understand its characteristics better. This involves using either VS Code

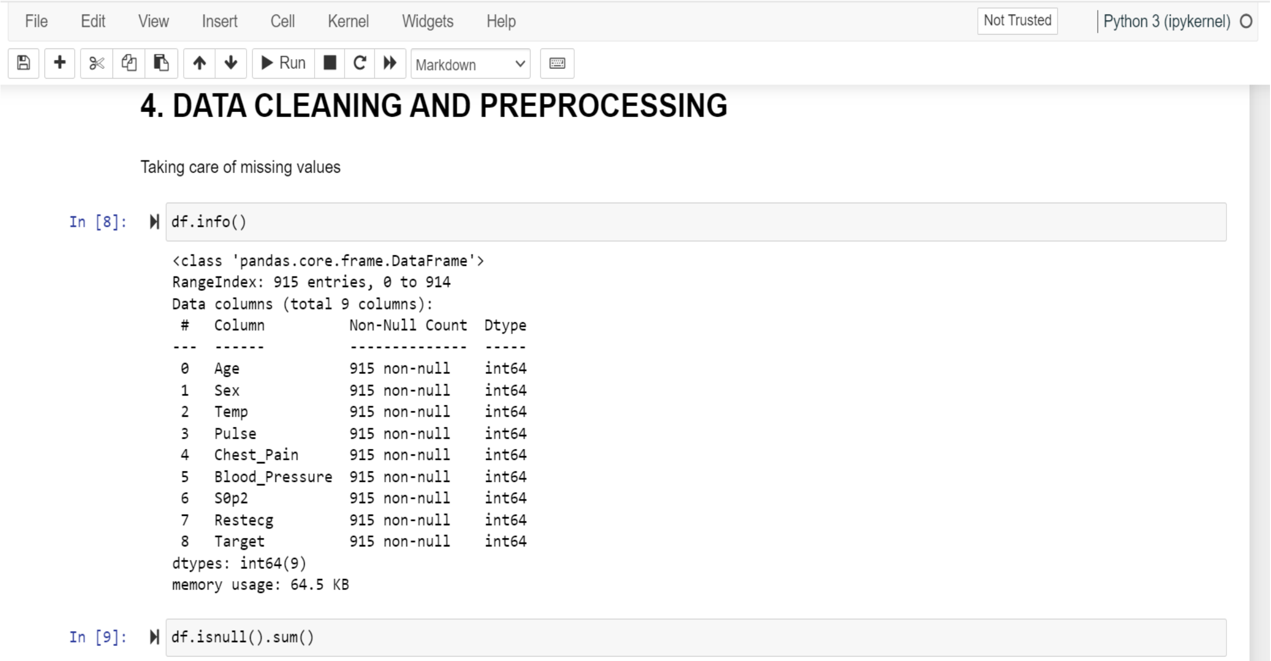
Version or Jupyter Notebook, and visualization libraries such as SciPy and Scikit-learn, Seaborn and Matplotlib.



```
fig, (ax1, ax2) = plt.subplots(nrows=1, ncols=2, sharey=False, figsize=(14,6))

ax1 = df['Target'].value_counts().plot.pie( x="Cardiovascular disease" ,y ='no.of patients',
                   autopct = "%1.0f%%",labels=["Cardiovascular Disease","Normal"], startangle = 60,ax=ax1);
ax1.set(title = 'Percentage of Cardiovascular disease patients in Dataset')

ax2 = df["Target"].value_counts().plot(kind="barh" ,ax =ax2)
for i,j in enumerate(df["Target"].value_counts().values):
    ax2.text(.5,i,j,fontsize=12)
ax2.set(title = 'No. of CD and non-CD patients in Dataset')

# Saving the plot as an image
fig.savefig('target_distribution.png', dpi=300, bbox_inches='tight')
```

Figure 14: Exploratory Data Analysis

d) Data Cleaning and Preprocessing: To ensure data quality, a data cleaning process was conducted. Null and junk values were identified using the isnull() and isna().sum() functions in Python and were either removed or imputed using appropriate techniques. In the preprocessing phase, feature engineering was conducted on the dataset. To include categorical variables in the analysis, they were converted into numerical variables using the get_dummies() function from the Pandas library. This process allows for the inclusion of categorical information in the machine learning models.



4. DATA CLEANING AND PREPROCESSING

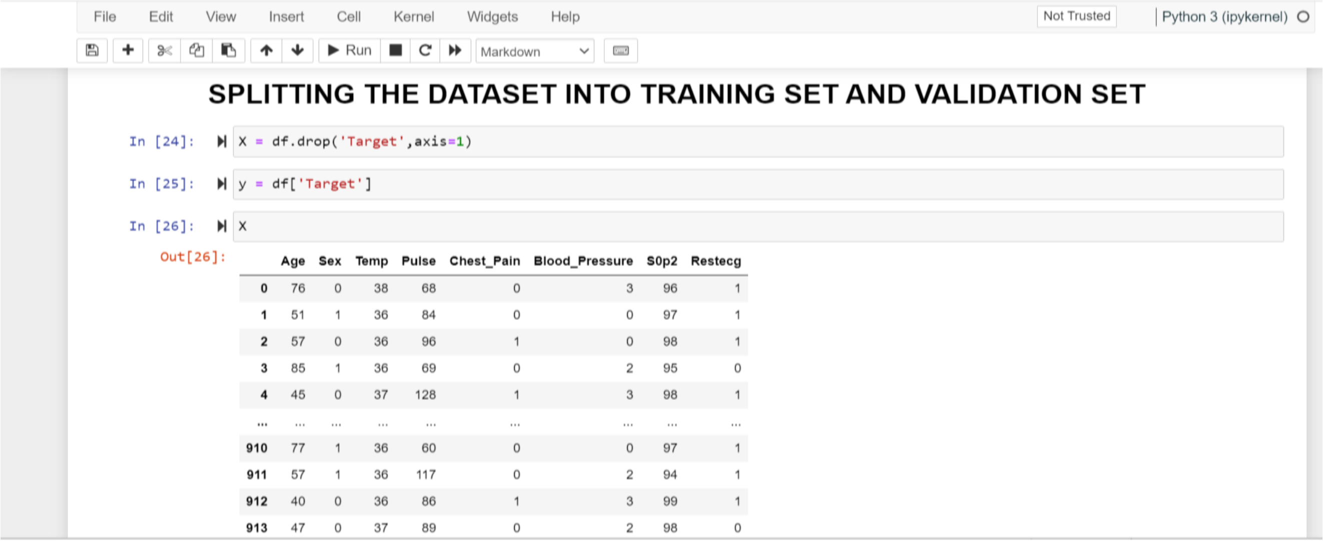
Taking care of missing values

```
In [8]:  df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 915 entries, 0 to 914
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             915 non-null    int64
 1   Sex             915 non-null    int64
 2   Temp            915 non-null    int64
 3   Pulse           915 non-null    int64
 4   Chest_Pain      915 non-null    int64
 5   Blood_Pressure  915 non-null    int64
 6   S0p2            915 non-null    int64
 7   Restecg         915 non-null    int64
 8   Target          915 non-null    int64
dtypes: int64(9)
memory usage: 64.5 KB

In [9]:  df.isnull().sum()
```

Figure 15: Data Cleansing and Preprocessing

e) Model Selection:

The X (input variables) and Y (dependent or target variables) were separated. The sklearn library's train_test_split() function was utilized to split the data into training and testing subsets, with 80% allocated for training and 20% for validation. In addition, a completely separate dataset is used for testing.



SPLITTING THE DATASET INTO TRAINING SET AND VALIDATION SET

```
In [24]:  X = df.drop('Target',axis=1)

In [25]:  y = df['Target']

In [26]:  X
Out[26]:
```

	Age	Sex	Temp	Pulse	Chest_Pain	Blood_Pressure	S0p2	Restecg
0	76	0	38	68	0	3	96	1
1	51	1	36	84	0	0	97	1
2	57	0	36	96	1	0	98	1
3	85	1	36	69	0	2	95	0
4	45	0	37	128	1	3	98	1
...	...	...	...	...	...	...	...	...
910	77	1	36	60	0	0	97	1
911	57	1	36	117	0	2	94	1
912	40	0	36	86	1	3	99	1
913	47	0	37	89	0	2	98	0

Figure 16: Model selection

f) Applied ML Models: To assess the performance of different machine learning algorithms, a range of combined models were applied to the dataset. These included Random Forest, Logistic Regression, and Gradient Boosting (RF, LR, and GB), Random Forest, Logistic Regression, and Support Vector Machine (RF, LR, and SVM), Random Forest, Decision Tree, and Gradient Boosting (RF, DT, and GB), Random Forest, Decision Tree, and Support Vector Machine (RF, DT, and SVM), and Random Forest, Gradient Boosting, and Support Vector Machine (RF, GB, and SVM).

It involves training and evaluating three individual classifiers (Random Forest, Logistic Regression, and Support Vector Machine) and combining their predictions using a Voting Classifier to form the ensemble model. Here's a detailed explanation of each part.

Dataset Loading and Splitting:

The dataset is assumed to be loaded into the variables 'X' (features) and 'y' (target) beforehand. The data consists of features (X) and corresponding target labels (y).

The dataset is split into training and validation sets using the train_test_split function. It takes 'X' and 'y' as input and returns four subsets: 'X_train', 'X_test', 'y_train', and 'y_test'. The 'test_size' parameter is set to 0.2, which means the validation set will be 20% of the entire dataset.

Model Initialization:

Three classifiers, namely Random Forest, Logistic Regression, and Support Vector Machine (SVM), are initialized. Each of these classifiers will be trained individually.

Ensemble Model Creation:

A combined model is created using the VotingClassifier from scikit-learn. The ensemble combines the predictions of the three individual classifiers using the 'hard' voting strategy, where the majority prediction wins.

Model Training:

The three individual classifiers (Random Forest, Logistic Regression, and SVM) and the ensemble model are trained using the fit method. Each classifier is trained on the training set, which is 'X_train' and 'y_train'.

Model Prediction and Accuracy Calculation:

After training, each classifier (Random Forest, Logistic Regression, and SVM) is used to make predictions on the validation set ('X_test').

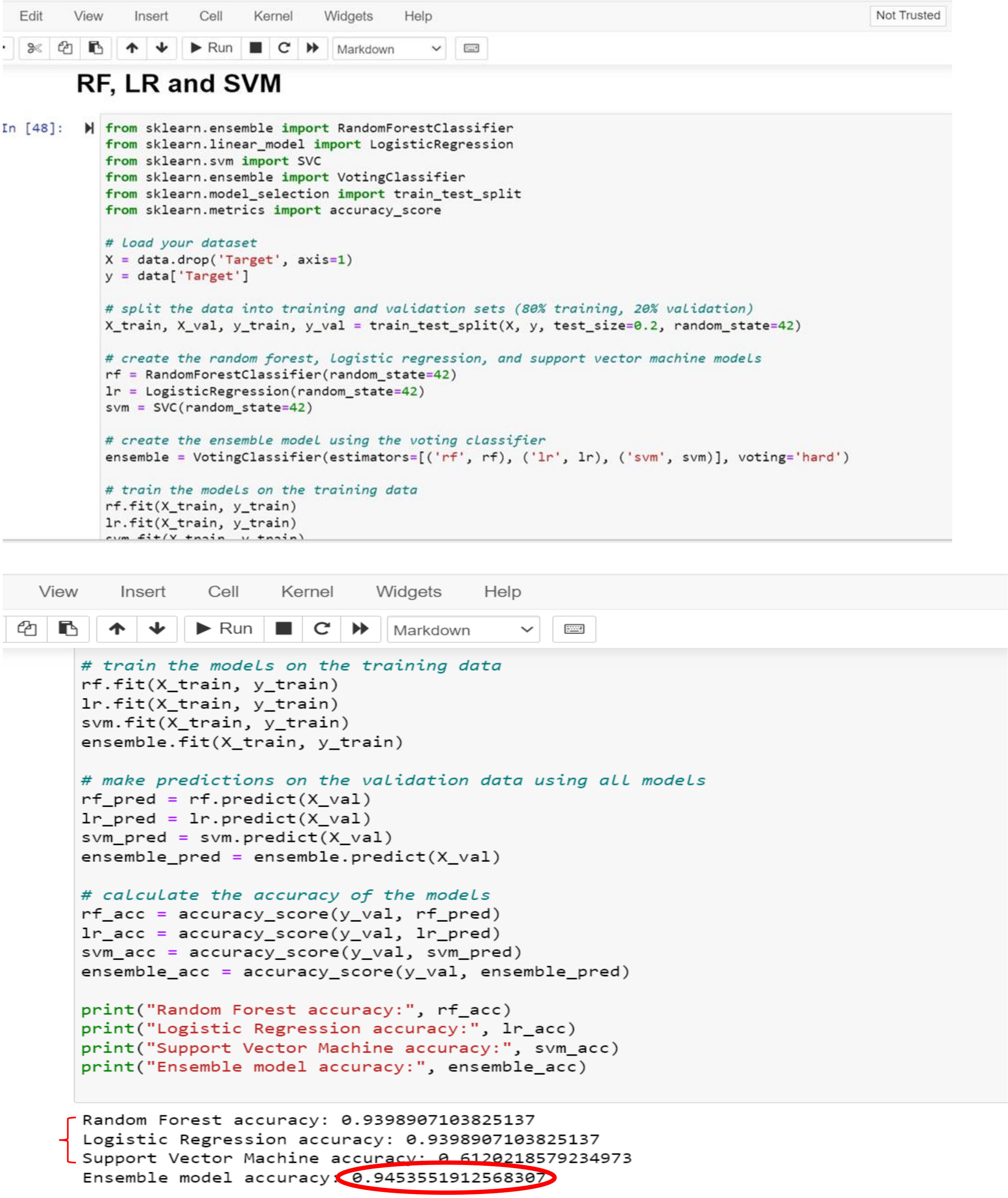
The predictions of each classifier are compared against the true labels in 'y_test' to calculate their accuracy using the accuracy_score function from scikit-learn. The accuracy is a measure of how well the model predictions match the actual target labels.

Results Display:

The accuracy of each individual model (Random Forest, Logistic Regression, and SVM) and the ensemble model is displayed using print.

Model Saving:

The trained ensemble model is saved to a file named 'RF_LR_SVM_ensembled_model.joblib' using the joblib.dump function. Saving the model allows you to use it later for making predictions on new data without retraining.



RF, LR and SVM

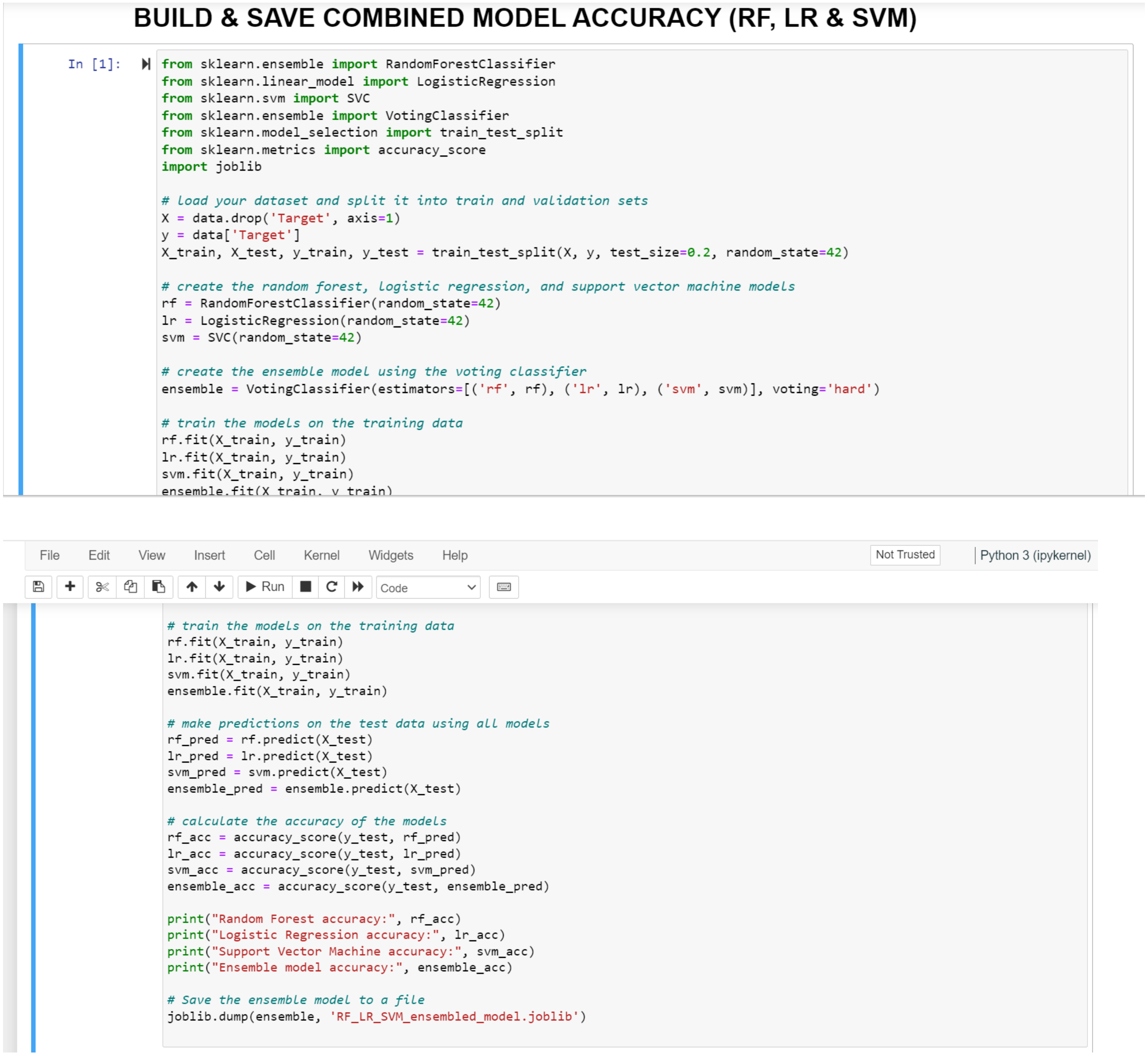
```
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.ensemble import VotingClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

# load your dataset
X = data.drop('Target', axis=1)
y = data['Target']

# split the data into training and validation sets (80% training, 20% validation)
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2, random_state=42)

# create the random forest, logistic regression, and support vector machine models
rf = RandomForestClassifier(random_state=42)
lr = LogisticRegression(random_state=42)
svm = SVC(random_state=42)

# create the ensemble model using the voting classifier
ensemble = VotingClassifier(estimators=[('rf', rf), ('lr', lr), ('svm', svm)], voting='hard')

# train the models on the training data
rf.fit(X_train, y_train)
lr.fit(X_train, y_train)
svm.fit(X_train, y_train)
```

```
# train the models on the training data
rf.fit(X_train, y_train)
lr.fit(X_train, y_train)
svm.fit(X_train, y_train)
ensemble.fit(X_train, y_train)

# make predictions on the validation data using all models
rf_pred = rf.predict(X_val)
lr_pred = lr.predict(X_val)
svm_pred = svm.predict(X_val)
ensemble_pred = ensemble.predict(X_val)

# calculate the accuracy of the models
rf_acc = accuracy_score(y_val, rf_pred)
lr_acc = accuracy_score(y_val, lr_pred)
svm_acc = accuracy_score(y_val, svm_pred)
ensemble_acc = accuracy_score(y_val, ensemble_pred)

print("Random Forest accuracy:", rf_acc)
print("Logistic Regression accuracy:", lr_acc)
print("Support Vector Machine accuracy:", svm_acc)
print("Ensemble model accuracy:", ensemble_acc)
```

```
Random Forest accuracy: 0.9398907103825137
Logistic Regression accuracy: 0.9398907103825137
Support Vector Machine accuracy: 0.6120218579234973
Ensemble model accuracy: 0.9453551912568307
```

Figure 17: RF, LR, and SVM model

g) Deployment of the Best Performing Model: The model with the highest accuracy was selected for deployment. In this section, the performance of various combined machine learning algorithms was assessed using evaluation metrics such as accuracy and precision. The model was then built using joblib library. The objective of the analysis was to determine whether the heart is defective or not based on the input attributes listed in Table 1. The Python programming language, known for its widespread usage in machine learning and data analysis, was employed for data analysis and modeling tasks.



BUILD & SAVE COMBINED MODEL ACCURACY (RF, LR & SVM)

```
In [1]:  from sklearn.ensemble import RandomForestClassifier
         from sklearn.linear_model import LogisticRegression
         from sklearn.svm import SVC
         from sklearn.ensemble import VotingClassifier
         from sklearn.model_selection import train_test_split
         from sklearn.metrics import accuracy_score
         import joblib

         # load your dataset and split it into train and validation sets
         X = data.drop('Target', axis=1)
         y = data['Target']
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

         # create the random forest, logistic regression, and support vector machine models
         rf = RandomForestClassifier(random_state=42)
         lr = LogisticRegression(random_state=42)
         svm = SVC(random_state=42)

         # create the ensemble model using the voting classifier
         ensemble = VotingClassifier(estimators=[('rf', rf), ('lr', lr), ('svm', svm)], voting='hard')

         # train the models on the training data
         rf.fit(X_train, y_train)
         lr.fit(X_train, y_train)
         svm.fit(X_train, y_train)
         ensemble.fit(X_train, y_train)
```

```
# train the models on the training data
rf.fit(X_train, y_train)
lr.fit(X_train, y_train)
svm.fit(X_train, y_train)
ensemble.fit(X_train, y_train)

# make predictions on the test data using all models
rf_pred = rf.predict(X_test)
lr_pred = lr.predict(X_test)
svm_pred = svm.predict(X_test)
ensemble_pred = ensemble.predict(X_test)

# calculate the accuracy of the models
rf_acc = accuracy_score(y_test, rf_pred)
lr_acc = accuracy_score(y_test, lr_pred)
svm_acc = accuracy_score(y_test, svm_pred)
ensemble_acc = accuracy_score(y_test, ensemble_pred)

print("Random Forest accuracy:", rf_acc)
print("Logistic Regression accuracy:", lr_acc)
print("Support Vector Machine accuracy:", svm_acc)
print("Ensemble model accuracy:", ensemble_acc)

# Save the ensemble model to a file
joblib.dump(ensemble, 'RF_LR_SVM_ensembled_model.joblib')
```

Figure 18: Best Performing Model

h) Building the SONART AI web application Cardiovascular Disease Prediction: The SONART AI web application Cardiovascular Disease Prediction System was developed using a systematic implementation approach. To provide a user-friendly interface and deploy the prediction model, Flask, a Python module, was utilized for web application development. Flask is a microframework known for its compact and flexible core, simplifying the development process by offering functionalities such as URL routing and a template engine. Additionally, additional frameworks used for web development include HTML, CSS, and JavaScript.

The coding and implementation of the SONART system took place in the VS Code development environment and XAMPP for database building, which provided a comprehensive platform for web development tasks. Python, as a high-level general programming language, served as the primary language for implementing the SONART AI web application Cardiovascular Disease Prediction system, leveraging its versatility and extensive libraries.

The development of the SONART AI web application Cardiovascular Disease Prediction System represents a significant contribution to the field of healthcare technology. By harnessing the synergistic capabilities of Flask's web framework and Python's flexibility, the system offers users an intuitive and interactive platform to assess the risk of cardiovascular diseases based on input attributes. Within this context, paramount emphasis has been placed on addressing the critical aspect of data security and privacy to safeguard sensitive patient information.

A noteworthy advancement in this system's capabilities arose from the seamless integration of a patient record database utilizing the Flask web framework. This strategic incorporation has revolutionized the storage and management of patient data, ensuring not only the efficiency of access and retrieval but also the preservation of data confidentiality. It is essential to acknowledge that the sensitive nature of patient records necessitates stringent security measures to maintain the highest standard of privacy.

Flask, being a high-level Python web framework of proven reliability, played a pivotal role in managing the database and facilitating the development of our advanced web application. This framework's versatility, empowered by a robust Object-Relational Mapping (ORM) system, has streamlined the interaction between our application and the database, simplifying the development process while upholding the integrity and confidentiality of patient data.

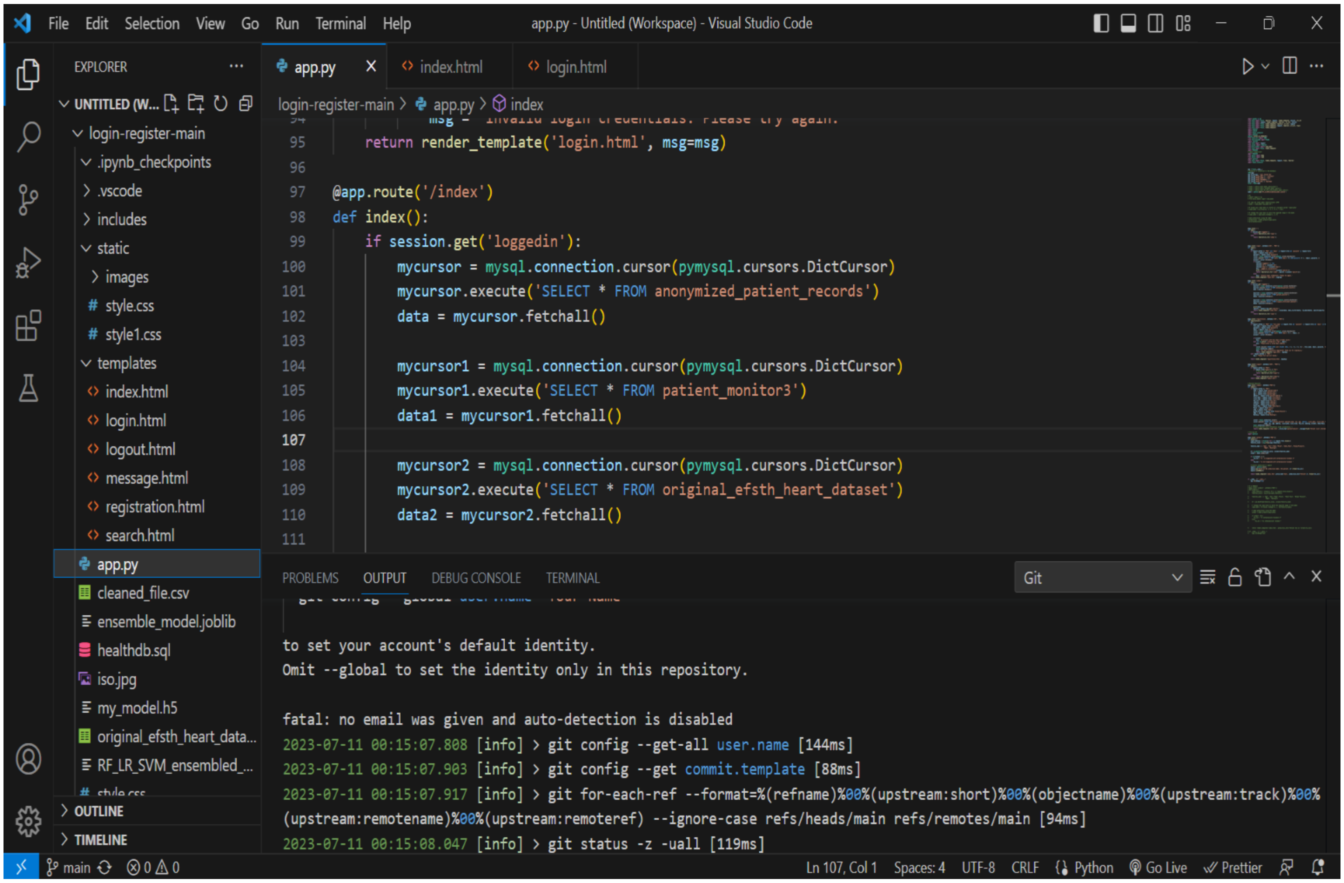
Crucially, the entire design and implementation of the patient record database have been meticulously crafted to prioritize security at every level. The intricacies of data handling, storage, and access have been developed with a comprehensive security-conscious approach to safeguard patient privacy and prevent unauthorized access.

The seamless integration of Flask with the prediction system is a testament to its versatility and adaptability, ensuring a harmonious and cohesive platform. With a steadfast commitment to data security, the patient record database and predictive capabilities have been harmonized to provide a secure, unified environment.

The incorporation of Flask has significantly enhanced the system's data storage, retrieval, and management capabilities, while preserving the confidentiality of patient records. The comprehensive solution now offers a robust and scalable platform for maintaining patient records and predicting cardiovascular disease, maintaining an unwavering focus on ethical and regulatory considerations.

Central to our endeavor is the dedication to leveraging cutting-edge technologies, exemplified by Flask, which manifests our profound commitment to the highest standards of data security and privacy in the domain of cardiovascular health. Security remains an integral pillar, fortifying user trust and confidence in the platform for patient record keeping and disease prediction, ultimately leading to improved patient outcomes and elevated healthcare delivery.

In essence, the integration of Flask within the web development process has engendered a transformative enhancement, characterized by an elevated standard of data security and privacy. This profound advancement not only underscores superior predictive capabilities but also underscores our steadfast commitment to providing a secure and ethically conscientious healthcare technology platform, thereby elevating data protection and healthcare practices to new heights.



```python
            msg = 'Invalid login credentials. Please try again.'
    return render_template('login.html', msg=msg)

@app.route('/index')
def index():
    if session.get('loggedin'):
        mycursor = mysql.connection.cursor(pymysql.cursors.DictCursor)
        mycursor.execute('SELECT * FROM anonymized_patient_records')
        data = mycursor.fetchall()

        mycursor1 = mysql.connection.cursor(pymysql.cursors.DictCursor)
        mycursor1.execute('SELECT * FROM patient_monitor3')
        data1 = mycursor1.fetchall()

        mycursor2 = mysql.connection.cursor(pymysql.cursors.DictCursor)
        mycursor2.execute('SELECT * FROM original_efsth_heart_dataset')
        data2 = mycursor2.fetchall()

```

```
  git config --global user.name "Your Name"

to set your account's default identity.
Omit --global to set the identity only in this repository.

fatal: no email was given and auto-detection is disabled
2023-07-11 00:15:07.808 [info] > git config --get-all user.name [144ms]
2023-07-11 00:15:07.903 [info] > git config --get commit.template [88ms]
2023-07-11 00:15:07.917 [info] > git for-each-ref --format=%(refname)%00%(upstream:short)%00%(objectname)%00%(upstream:track)%00%
(upstream:remotename)%00%(upstream:remoteref) --ignore-case refs/heads/main refs/remotes/main [94ms]
2023-07-11 00:15:08.047 [info] > git status -z -uall [119ms]
```

Figure 19: SONART AI Web Building

This is the link to the uploaded source codes on github: https://github.com/muhafatty/Master_Thesis.git

CHAPTER 3: RESULTS AND DISCUSSIONS

Following a discussion and investigations with a cardiologist at the Edward Francis Small Teaching Hospital (EFSTH), the following contributing factors to the prediction of cardiovascular disease were identified (Dr Lamin E.S. Jaiteh).

3.1 Key Factors Influencing the Occurrence of Cardiovascular Disease

3.1.1 Age

Age is a significant independent risk factor for cardiovascular disease (CVD), even after adjusting for traditional risk factors in a multivariable CVD prediction model. The contribution of age in these models may reflect the intensity and duration of exposure to other traditional CVD risk factors. Studies have shown that the absence of traditional risk factors is associated with a reduction in CVD risk, even at older ages. Factors such as lower midlife blood pressure and cholesterol levels, absence of glucose intolerance, smoking abstinence, higher education, and female gender have been found to predict increased survival up to 85 years of age. The contribution of age to CVD risk prediction declines at older ages due to the limited time available for acquiring other modifiable risk factors (Dhingra & Vasan, 2012).

3.1.2 Sex

Gender contributes to the cardiovascular (CV) health of both women and men, directly and indirectly influenced by various risk factors. Understanding the role of gender domains (identity, roles, relations, institutions) and their interaction with biological sex in the manifestation, progression, and outcomes of cardiovascular disease (CVD) is essential. Detrimental characteristics associated with women in many cultures, such as poverty, low-level jobs, and lower pay, contribute to the multifaceted nature of CVD risk. Gender identity, encompassing diverse identities (cisgender, transgender, gender-neutral), plays a significant but poorly understood role in CVD risk, likely mediated through other gender domains. Personality traits, stress levels at work and home, emotional intelligence, depression, anxiety, and childhood trauma are dimensions within this context (Connelly et al., 2021).

3.1.3 High blood pressure (BP)

Cardiovascular disease (CVD) is strongly influenced by several controllable factors that can be modified to reduce the risk. These include high blood pressure (BP), smoking cigarettes, diabetes mellitus, and abnormal lipid levels. Among these factors, high BP is strongly associated with the development of CVD and is highly prevalent. However, there is evidence suggesting that the biologically normal level of BP in humans is lower than what has traditionally been considered normal in clinical practice and research. This underrepresentation of the role of BP as a risk factor for CVD has led to the proposal of an integrated theory for CVD causation.

The theory suggests that CVD in humans is primarily caused by a right-sided shift in the distribution of BP. This theory is supported by a robust body of coherent and consistent evidence and fulfills the criteria for causality proposed by Bradford Hill.

In the past, high BP was not recognized as a risk factor for CVD because there were no practical noninvasive methods to measure BP. It was only after the development and dissemination of these methods that physicians and actuaries began to associate high BP with disease, particularly CVD events. Early studies comparing the risk of CVD in individuals with high BP to those with lower but still high BP had limitations, as they did not account for the fact that most humans have a BP level above what is biologically normal and desirable (Fuchs & Whelton, 2020).

3.1.4 Resting ECG

Electrocardiography (ECG or EKG) is a non-invasive diagnostic modality that holds significant clinical relevance in investigating the severity of cardiovascular diseases (CVD). Originally developed by Dutch physician William Einthoven in 1902, ECG has emerged as a fundamental technique for assessing heart disorders, earning Einthoven the title of 'father of electrocardiography,' and a Nobel Prize in Medicine in 1924. Over time, ECG has gained recognition as a robust screening and diagnostic tool, finding application in diverse healthcare settings globally (Yasar Sattar & Lovely Chhabra, 2023).

3.1.5 Chest pain

Chest pain serves as a strong indicator of cardiovascular disease (CVD), primarily coronary artery disease (CAD). Recurrent angina pectoris, characterized by chronic chest pain, is the most prevalent form of CAD. Notably, specific medications often alleviate the pain, offering a diagnostic clue. Rapid relief of anginal pain within 3 minutes of sublingual nitroglycerin administration strongly suggests underlying coronary artery disease (Hickam, n.d.).

Nevertheless, it is crucial to acknowledge that chronic chest pain associated with CVD can have other causes. Esophageal disease, such as acid reflux-induced esophagitis, commonly presents as chest pain accompanied by a burning sensation. Pain relief through the use of antacids, topical lidocaine, or reflux-reducing maneuvers can indicate esophageal involvement.

Esophageal motor disorders, involving contraction and spasm of the esophageal muscles, represent another potential cause of chest pain. These disorders may arise as secondary manifestations of reflux esophagitis or exist independently as conditions like achalasia or diffuse esophageal spasm. Complicating the diagnostic process, medications like nitrates and calcium channel blockers, commonly used for CAD treatment, can also provide relief for esophageal motor disorders (Hickam, n.d.).

3.2 Evaluation Metrics

In this section, we present a discussion on the results obtained from employing various machine learning (ML) models in this study to simulate and analyze the outcomes. Python has been used as a language for the implementation purpose. The major packages used were Sklearn, Seaborn, Matplotlib, Pandas, Numpy, etc. In order to evaluate the effectiveness of each machine learning (ML) model, several performance metrics are calculated and analyzed. To evaluate the performance of the predictive models applied in this work, performance measures including accuracy and precision were performed.

– Accuracy: The percentage of the total number of instances that are correctly classified relative to the number of all tested instances.

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

– Precision: The ratio between the number of positive instances that are correctly classified and all instances predicted as positive. The precision presents how confident an instance predicted with a positive target actually has a positive target level.

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

3.3 Performance Results Of The Predictive Models

To enhance the performance of the cardiovascular disease (CVD) prediction model, a combination of multiple machine learning (ML) algorithms was employed. This approach yielded significantly improved performance metrics of accuracy and precision crucial for accurate CVD prediction. By leveraging the strengths of different algorithms, the combined model demonstrated enhanced predictive capabilities, capturing intricate patterns and relationships in the data. This ensemble approach resulted in a more robust and accurate predictive model, reducing false positives and false negatives. Overall, the integration of multiple ML algorithms proved to be a successful strategy for improving CVD prediction, contributing to more effective disease management and improved patient care.

A total of five combined algorithms were applied to the data, and the combined algorithm with the highest scoring accuracy (Random Forest, Logistic Regression, and Support Vector Machine (RF, LR, and SVM)) chosen to build the model. The combined algorithms are as follows.

1. Random Forest, Logistic Regression, and Gradient Boosting (RF, LR, and GB)
2. Random Forest, Logistic Regression, and Support Vector Machine (RF, LR, and SVM)
3. Random Forest, Decision Tree, and Gradient Boosting (RF, DT, and GB)
4. Random Forest, Decision Tree, and Support Vector Machine (RF, DT, and SVM)
5. Random Forest, Gradient Boosting and Support Vector Machine (RF, GB, and SVM)

Accuracy

Figure 20 illustrates the performance results of the five combined machine learning algorithms in terms of validation and testing accuracy. The experiments conducted reveal valuable insights into the accuracy scores achieved by each combined algorithm.

The combined algorithm of Random Forest, Logistic Regression, and Gradient Boosting (RF, LR & GB) demonstrated exceptional performance, achieving an accuracy score of 93.99% during the validation phase and 92.34% during testing. In comparison, the combined algorithm of Random Forest, Decision Trees, and Support Vector Machine (RF, LR & SVM) scored 91.40% during validation and 94.0% during testing, showcasing its robust performance.

Similarly, the combination of Random Forest, Decision Trees, and Gradient Boosting (RF, DT & GB) yielded a validation accuracy score of 93.99% and a testing accuracy score of 92.8%. These results highlight the effectiveness of this combined algorithm in maintaining a high level of accuracy across both validation and testing stages.

Furthermore, the combined algorithm of Random Forest, Gradient Boosting, and Support Vector Machine (RF, GB & SVM) achieved notable scores of 93.44% during validation and 94.3% during testing. These results indicate the strong performance and reliability of this algorithm in accurately predicting outcomes.

Notably, the highest accuracy score of 95.0% during testing and 94.5% during validation testing was achieved by the combined algorithm of Random Forest, Logistic Regression, and Support Vector Machine (RF, LR & SVM). This finding underscores the superior performance of this particular combination, suggesting that it outperforms the other combined models evaluated in this study.

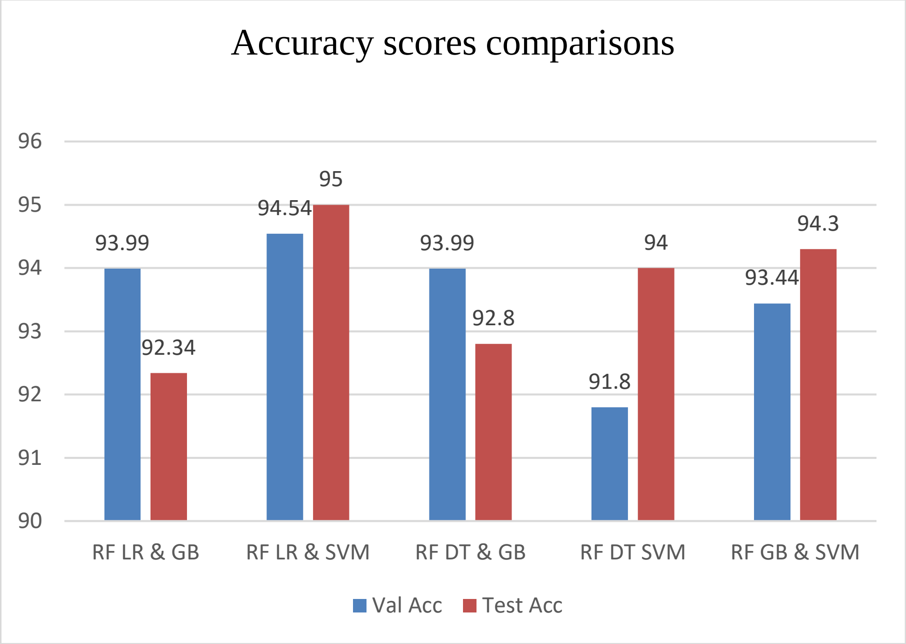


Figure 20: Combined Accuracy Scores

Precision

Figure 21 presents the precision scores obtained during the validation and testing phases for the combined models. These precision scores provide insights into the models' performance in terms of the accuracy of positive predictions.

The combined model consisting of Random Forest, Decision Trees, and Support Vector Machine (RF, DT & SVM) exhibited the lowest precision score of 91.49% during the validation phase. However, during the testing process, this combination achieved a higher precision score of 95.16%. This suggests that the model's precision improved when applied to unseen data.

On the other hand, the combined model of Random Forest, Logistic Regression, and Support Vector Machine (RF, LR & SVM) demonstrated the highest precision score of 96.63% during the validation phase and a slightly lower precision score of 94.73% during testing. This indicates that the model consistently provided accurate positive predictions, both during the validation and testing stages.

Similarly, the combination of Random Forest, Logistic Regression, and Gradient Boosting (RF, LR & GB) achieved a high precision score of 96.59% during the validation phase and 92.55% during testing. These results suggest that the model maintained a high level of precision in predicting positive outcomes during the validation phase, but slightly lower precision during the testing phase.

Furthermore, the combination of Random Forest, Decision Trees, and Gradient Boosting (RF, DT & GB) obtained a precision score of 95.56% during the validation phase and 93.54% during testing. This indicates that the model performed well in accurately predicting positive outcomes during both the validation and testing phases.

Lastly, the combination of Random Forest, Gradient Boosting, and Support Vector Machine (RF, GB & SVM) exhibited a precision score of 94.51% during the validation phase and 95.45% during the testing phase. These results demonstrate that the model consistently provided accurate positive predictions throughout both validation and testing.

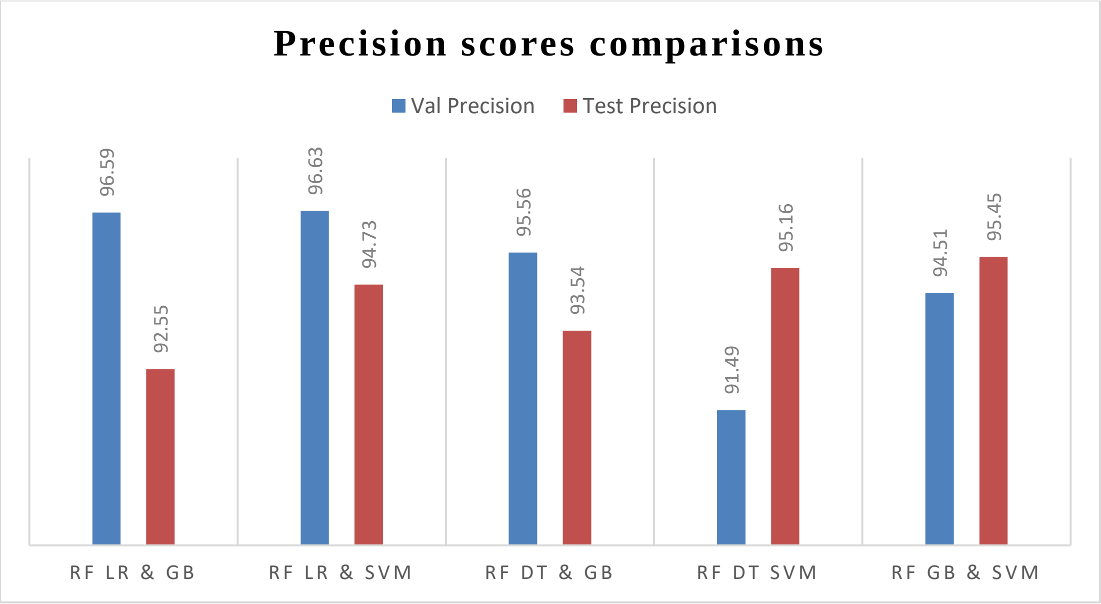


Figure 21: Combined Precision Scores

3.4 Combined Algorithm Evaluation Metrics

This section presents a comprehensive evaluation of the combined algorithms' performance during the testing phase, encompassing the two key metrics accuracy and precision. These metrics provide valuable insights into the effectiveness and discriminative capabilities of the models.

Accuracy serves as a measure of overall correctness in predictions. Precision focuses on the accuracy of positive predictions.

The results highlight variations in the performance of the combined algorithms. Notably, Random Forest, Logistic Regression, and Support Vector Machine (RF LR & SVM) combination stands out as it achieves the highest accuracy score (95.0). This finding suggests superior performance of this over the other models.

On the other hand, the combination of Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM) exhibited the highest precision score of 95.45%. In second place, the combinations of Random Forest with Decision Tree (RF DT) and Random Forest with Logistic Regression and SVM (RF LR & SVM) achieved commendable precision scores as well.

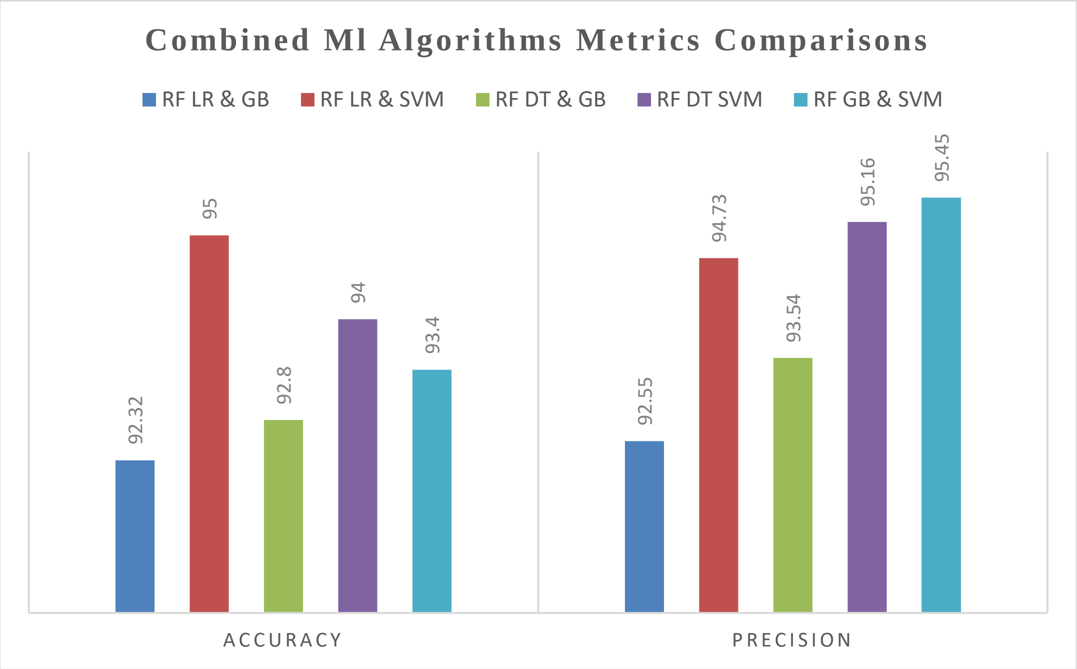
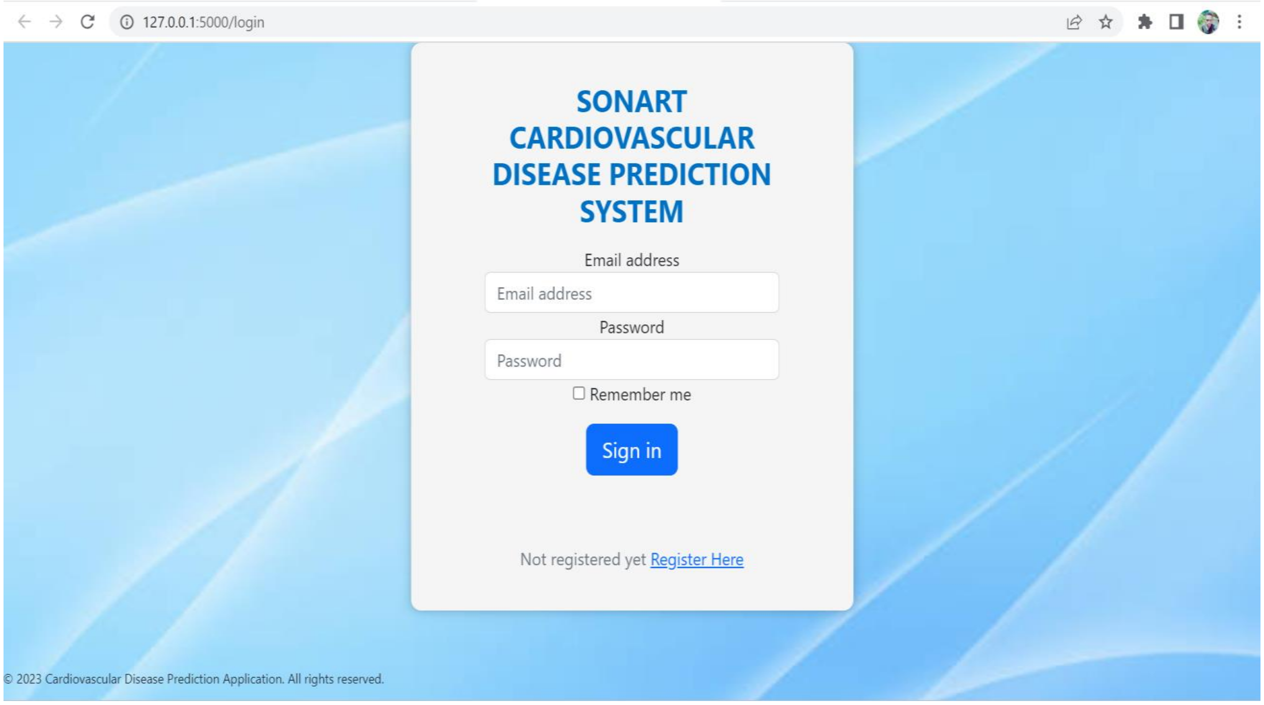


Figure 22: combined Algorithms Evaluation Metrics

3.5 SONART AI Cardiovascular Disease Prediction System (Web Application)

The SONART AI web application is a cutting-edge platform that incorporates a powerful combined machine learning model for accurate cardiovascular disease prediction.

Figure 23 showcases the login interface, allowing users to authenticate themselves as administrators or doctors. Upon entering their credentials, users gain authorized access to the system, enabling them to leverage its advanced features and functionalities. The SONART application prioritizes user authentication to safeguard patient data and provide a secure environment for healthcare professionals.



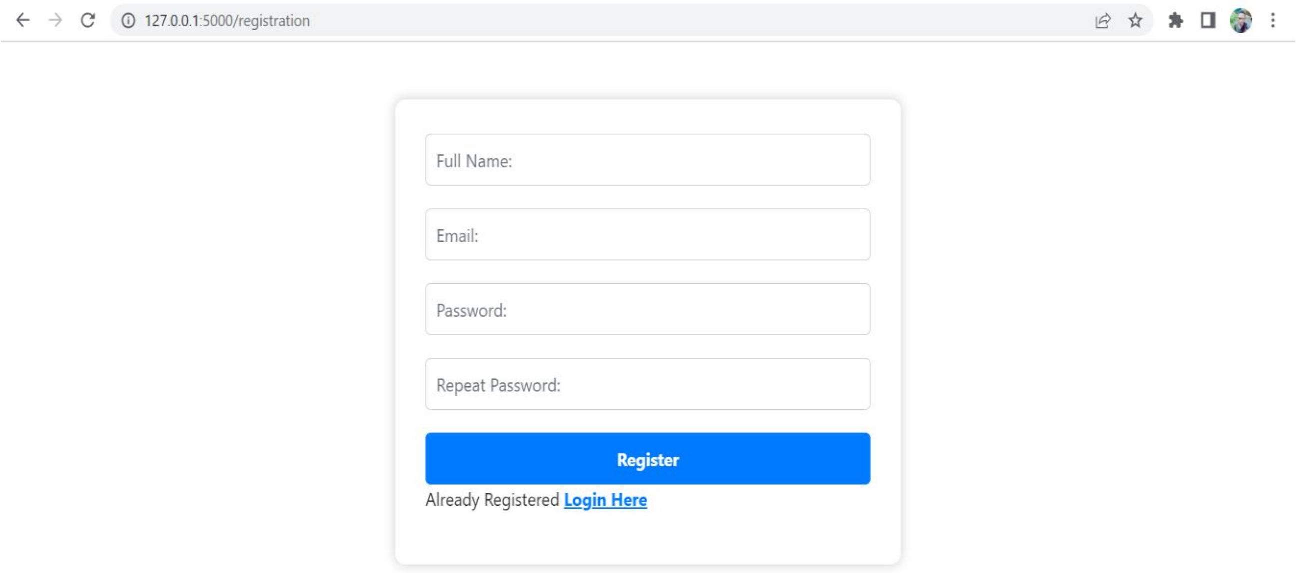
SONART CARDIOVASCULAR DISEASE PREDICTION SYSTEM

Not registered yet Register Here

© 2023 Cardiovascular Disease Prediction Application. All rights reserved.

Figure 23: SONART Login

The depicted figure 24 portrays the registration page, designed to enable users (admin, doctors or patients) to create an account by providing their personal information, including full name, email address, and password.



127.0.0.1:5000/registration

Full Name:

Email:

Password:

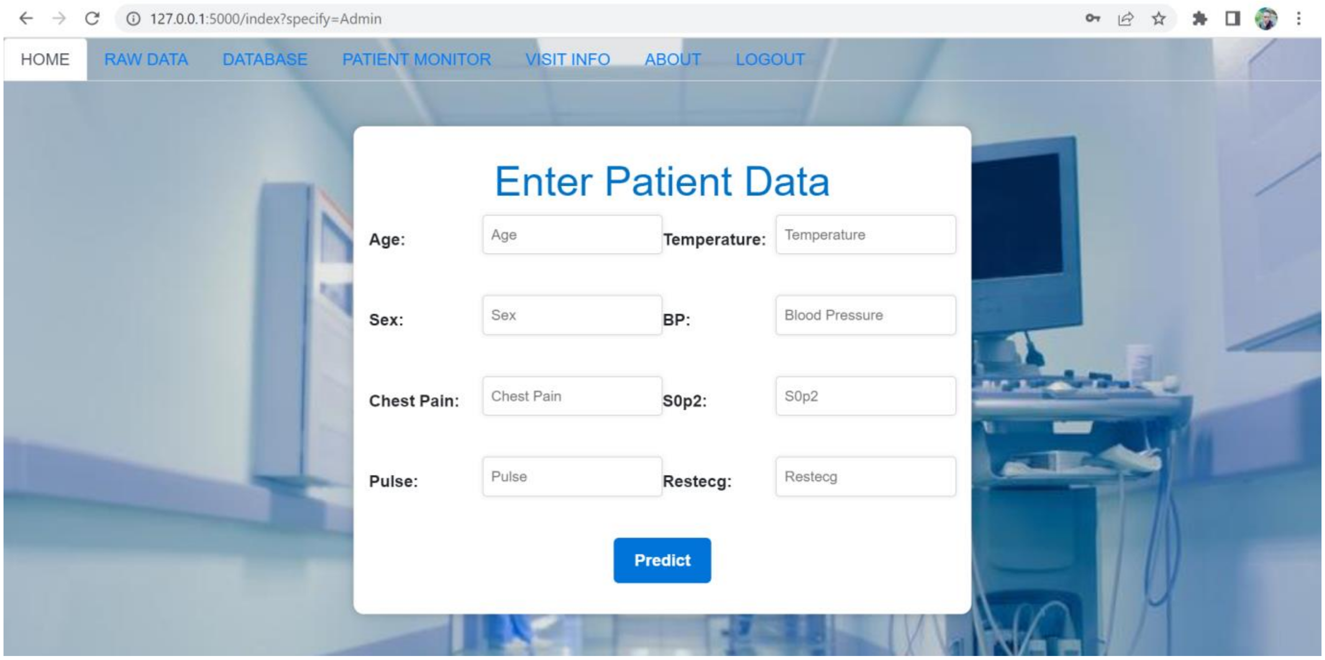
Repeat Password:

Register

Already Registered **Login Here**

Figure 24: SONART Registration Page

The presented figure showcases the SONART AI prediction page, which empowers doctors to assess the cardiac condition of a patient by inputting relevant variables and initiating the prediction process by selecting the "predict" option. The resulting predictions are delivered in a comprehensive manner, encompassing both an automated speech output and textual format

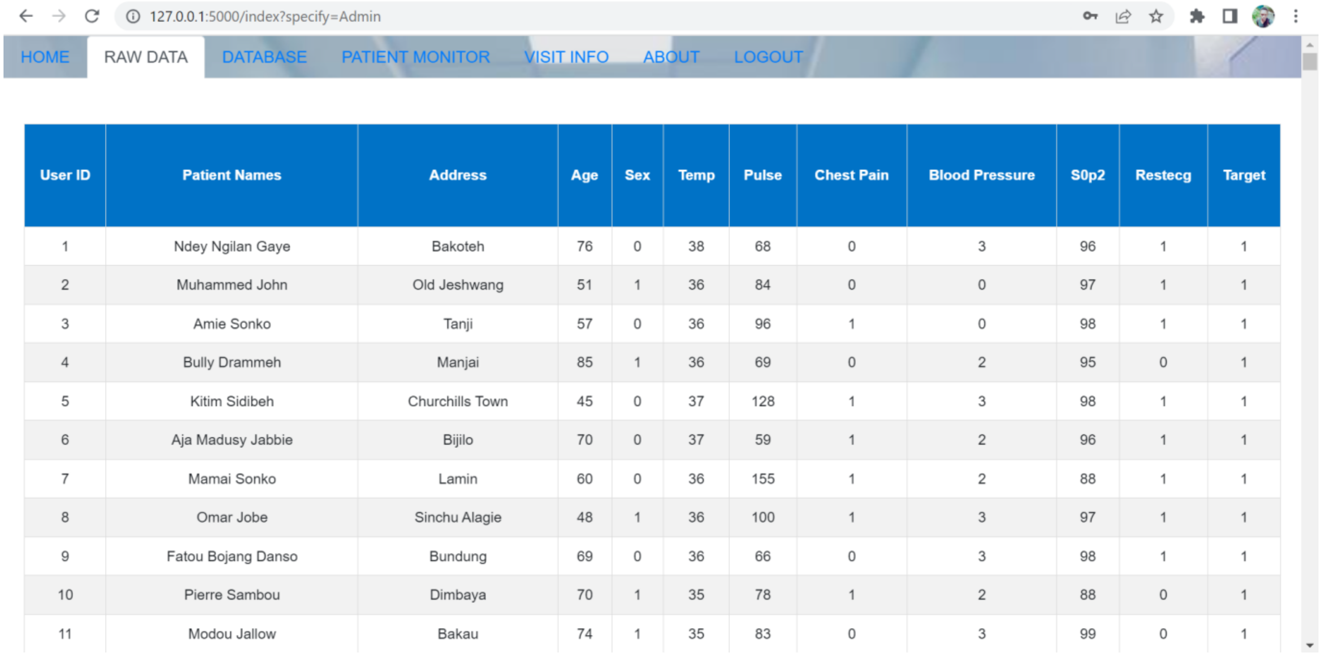


Enter Patient Data

Age: Age
Temperature: Temperature
Sex: Sex
BP: Blood Pressure
Chest Pain: Chest Pain
S0p2: S0p2
Pulse: Pulse
Restecg: Restecg

Figure 25: SONART AI Prediction Page

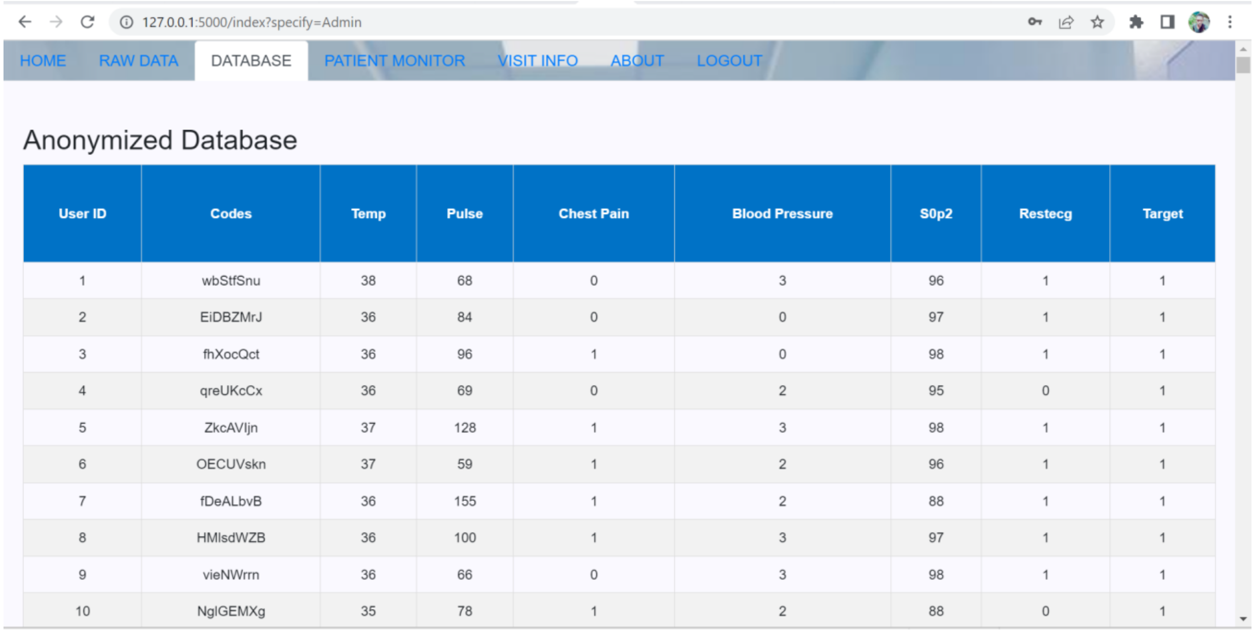
Figure 26 exhibits the compilation of original patient data obtained from the Edward Francis Small Teaching Hospital. To ensure the confidentiality and privacy of the patient information, access to this data is restricted exclusively to the administrator and the doctor.



User ID	Patient Names	Address	Age	Sex	Temp	Pulse	Chest Pain	Blood Pressure	S0p2	Restecg	Target
1	Ndey Ngilan Gaye	Bakoteh	76	0	38	68	0	3	96	1	1
2	Muhammed John	Old Jeshwang	51	1	36	84	0	0	97	1	1
3	Amie Sonko	Tanji	57	0	36	96	1	0	98	1	1
4	Bully Drammeh	Manjai	85	1	36	69	0	2	95	0	1
5	Kitim Sidibeh	Churchills Town	45	0	37	128	1	3	98	1	1
6	Aja Madusy Jabbie	Bijilo	70	0	37	59	1	2	96	1	1
7	Mamai Sonko	Lamin	60	0	36	155	1	2	88	1	1
8	Omar Jobe	Sinchu Alagie	48	1	36	100	1	3	97	1	1
9	Fatou Bojang Danso	Bundung	69	0	36	66	0	3	98	1	1
10	Pierre Sambou	Dimbaya	70	1	35	78	1	2	88	0	1
11	Modou Jallow	Bakau	74	1	35	83	0	3	99	0	1

Figure 26: Raw Data

The depicted figure 27 displays the anonymized names of the patients, employing a Python code to ensure confidentiality. To enhance the preservation of patient privacy, the displayed list intentionally excludes information such as age and sex.



Anonymized Database

User ID	Codes	Temp	Pulse	Chest Pain	Blood Pressure	S0p2	Restecg	Target
1	wbStfSnu	38	68	0	3	96	1	1
2	EiDBZMrJ	36	84	0	0	97	1	1
3	fhXocQct	36	96	1	0	98	1	1
4	qreUKcCx	36	69	0	2	95	0	1
5	ZkcAVljn	37	128	1	3	98	1	1
6	OECUVskn	37	59	1	2	96	1	1
7	fDeALbvB	36	155	1	2	88	1	1
8	HMlsdWZB	36	100	1	3	97	1	1
9	vieNWrrn	36	66	0	3	98	1	1
10	NglGEMXg	35	78	1	2	88	0	1

Figure 27: Anonymized Patient Dataset

The presented figure 28 illustrates the patient monitor page, which serves as a platform for capturing and recording patient details during their visits. Subsequently, this recorded information is securely transmitted and stored within an established database, ensuring effective data management and retention.

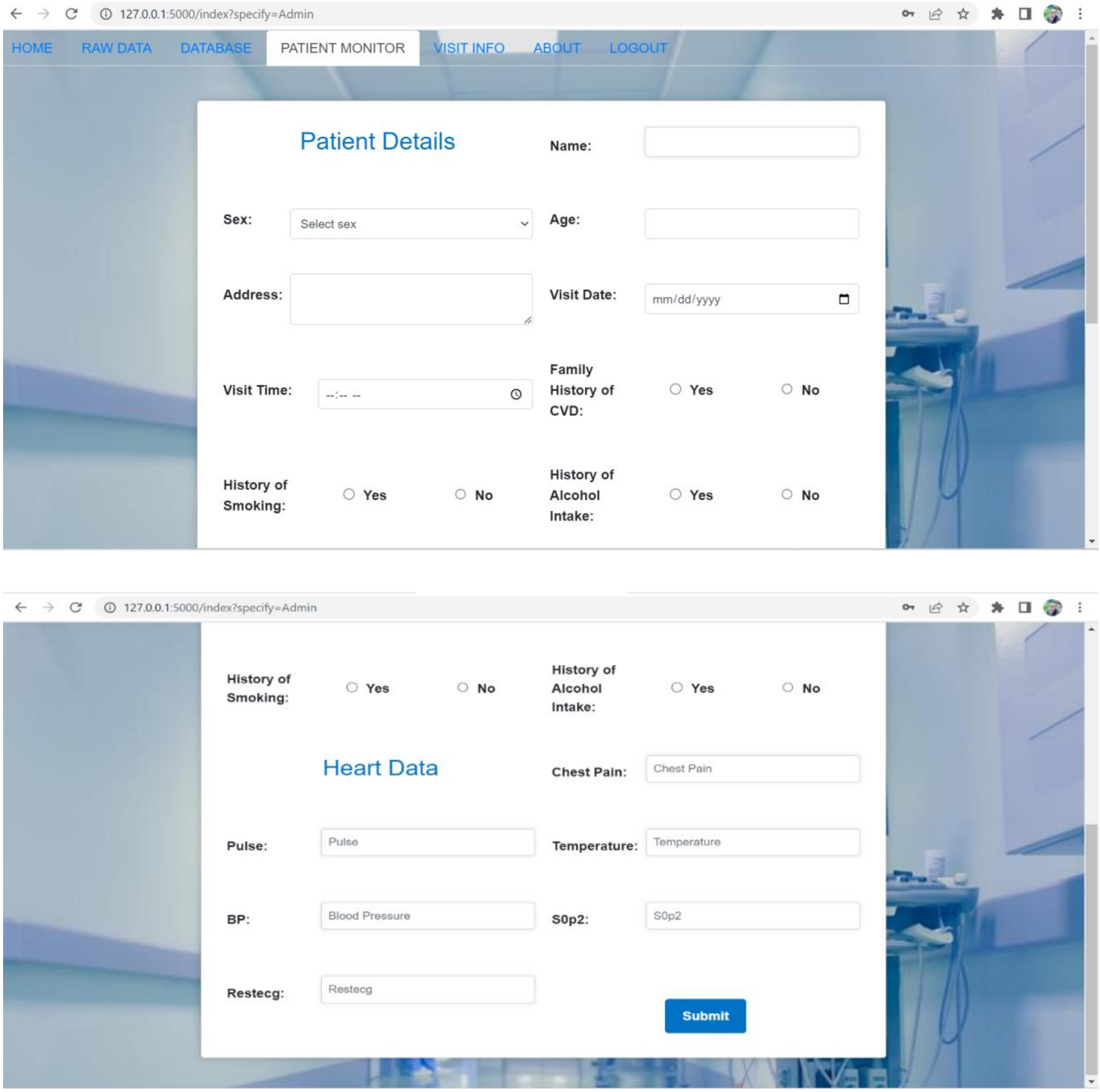
In this anonymization process, a Python script was used to anonymize patient records stored in a CSV file. The script utilizes the Pandas library for data handling and manipulation. The primary goal of anonymization is to protect the privacy and confidentiality of individuals by replacing sensitive information, such as patient names, with pseudonyms or random strings.

Here's a step-by-step explanation of the anonymization process:

- Loading Patient Records: The script starts by importing the required libraries, including Pandas, and loading the patient records from the CSV file located at the given file path C:\Users\DELL\Downloads\Final_Version_Research_Proposal\TEST\original_efsth _heart_dataset.csv.

- Generating Random Strings: To replace patient names with random strings, the script defines a function called generate_random_string(length). This function generates a random string of specified length using a combination of letters from the standard ASCII character set (both lowercase and uppercase).
- Anonymizing Patient Names: The script creates a new column named "Codes" in the DataFrame, where each row is populated with a randomly generated string of 8 characters. The length of the random string can be adjusted according to the desired level of anonymity.
- Saving Anonymized Data : Once the anonymization process is complete, the script saves the modified DataFrame containing anonymized patient records to a new CSV file named "anonymized_patient_records.csv". The index=False parameter ensures that the index of the DataFrame is not included in the output CSV file.
- Recovery of Anonymized Names: To recover the original patient names using the user ID numbers, one would need access to the original (raw) dataset, where the "User ID" column still contains the original IDs linked to the respective patient names. By using this tab, one can look up the corresponding patient names based on the "User ID" and match them with the anonymized records in this tab.

By executing this script, the patient records' original names are replaced with unique and random strings, effectively anonymizing the data. This process helps to protect patient privacy and comply with data protection regulations, making it suitable for use in research and analysis while safeguarding sensitive information.


127.0.0.1:5000/index?specify=Admin
HOME
RAW DATA
DATABASE
PATIENT MONITOR
VISIT INFO
ABOUT
LOGOUT

Patient Details

Name:

Sex: Select sex

Age:

Address:

Visit Date: mm/dd/yyyy

Visit Time: --:-- --

Family History of CVD: ○ Yes ○ No

History of Smoking: ○ Yes ○ No

History of Alcohol Intake: ○ Yes ○ No

127.0.0.1:5000/index?specify=Admin

History of Smoking: ○ Yes ○ No

History of Alcohol Intake: ○ Yes ○ No

Heart Data

Chest Pain: Chest Pain

Pulse: Pulse

Temperature: Temperature

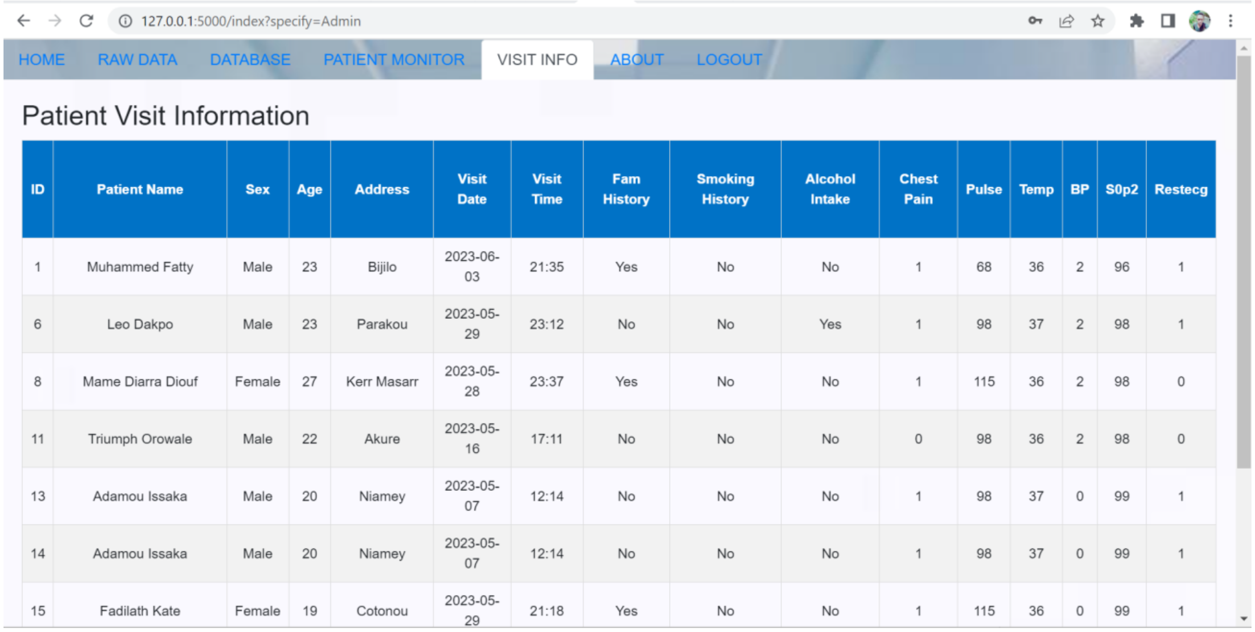
BP: Blood Pressure

S0p2: S0p2

Restecg: Restecg

Figure 28: Patient Monitor Page

The displayed figure 29 showcases the comprehensive record of patient details, encompassing essential information such as patient names, sex, age, address, visit time, visit date, and relevant medical parameters. This compilation provides a comprehensive overview of the patients' demographic and clinical data for effective analysis and documentation.

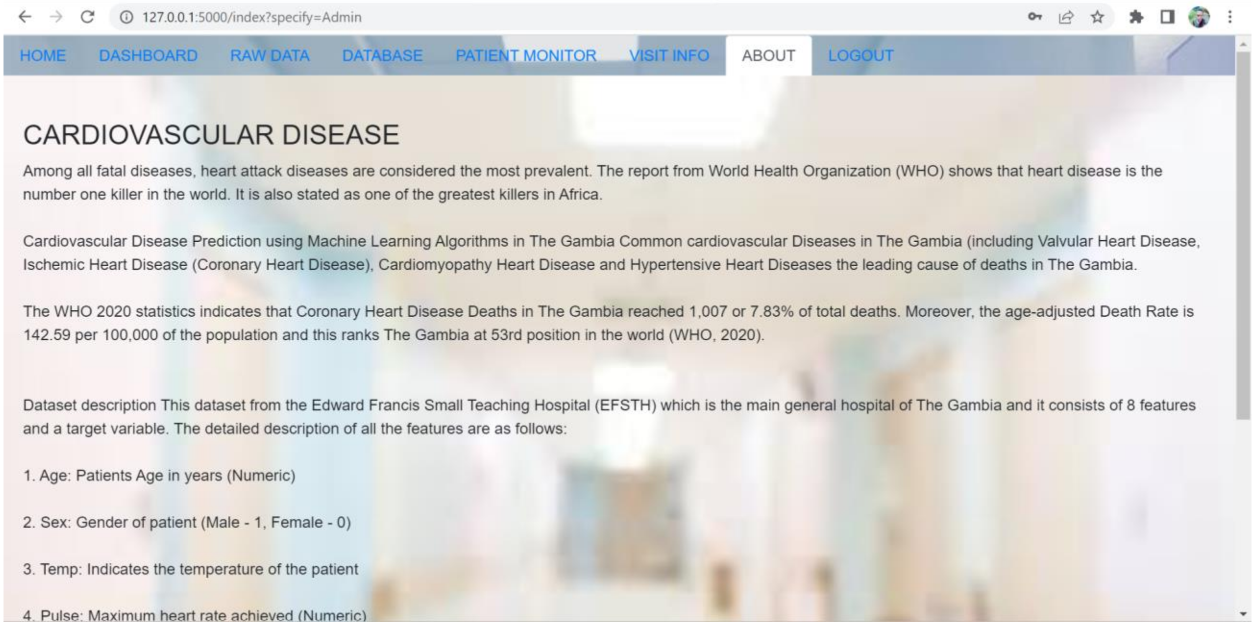


Patient Visit Information

ID	Patient Name	Sex	Age	Address	Visit Date	Visit Time	Fam History	Smoking History	Alcohol Intake	Chest Pain	Pulse	Temp	BP	S0p2	Restecg
1	Muhammed Fatty	Male	23	Bijilo	2023-06-03	21:35	Yes	No	No	1	68	36	2	96	1
6	Leo Dakpo	Male	23	Parakou	2023-05-29	23:12	No	No	Yes	1	98	37	2	98	1
8	Mame Diarra Diouf	Female	27	Kerr Masarr	2023-05-28	23:37	Yes	No	No	1	115	36	2	98	0
11	Triumph Orowale	Male	22	Akure	2023-05-16	17:11	No	No	No	0	98	36	2	98	0
13	Adamou Issaka	Male	20	Niamey	2023-05-07	12:14	No	No	No	1	98	37	0	99	1
14	Adamou Issaka	Male	20	Niamey	2023-05-07	12:14	No	No	No	1	98	37	0	99	1
15	Fadilath Kate	Female	19	Cotonou	2023-05-29	21:18	Yes	No	No	1	115	36	0	99	1

Figure 29: Patient Visit Info Page

Figure 30 depicts the description of cardiovascular disease along with the dataset utilized for the study. It offers valuable insights into the characteristics and composition of the dataset, aiding in a deeper understanding and comprehension of the research context.



CARDIOVASCULAR DISEASE

Among all fatal diseases, heart attack diseases are considered the most prevalent. The report from World Health Organization (WHO) shows that heart disease is the number one killer in the world. It is also stated as one of the greatest killers in Africa.

Cardiovascular Disease Prediction using Machine Learning Algorithms in The Gambia Common cardiovascular Diseases in The Gambia (including Valvular Heart Disease, Ischemic Heart Disease (Coronary Heart Disease), Cardiomyopathy Heart Disease and Hypertensive Heart Diseases the leading cause of deaths in The Gambia.

The WHO 2020 statistics indicates that Coronary Heart Disease Deaths in The Gambia reached 1,007 or 7.83% of total deaths. Moreover, the age-adjusted Death Rate is 142.59 per 100,000 of the population and this ranks The Gambia at 53rd position in the world (WHO, 2020).

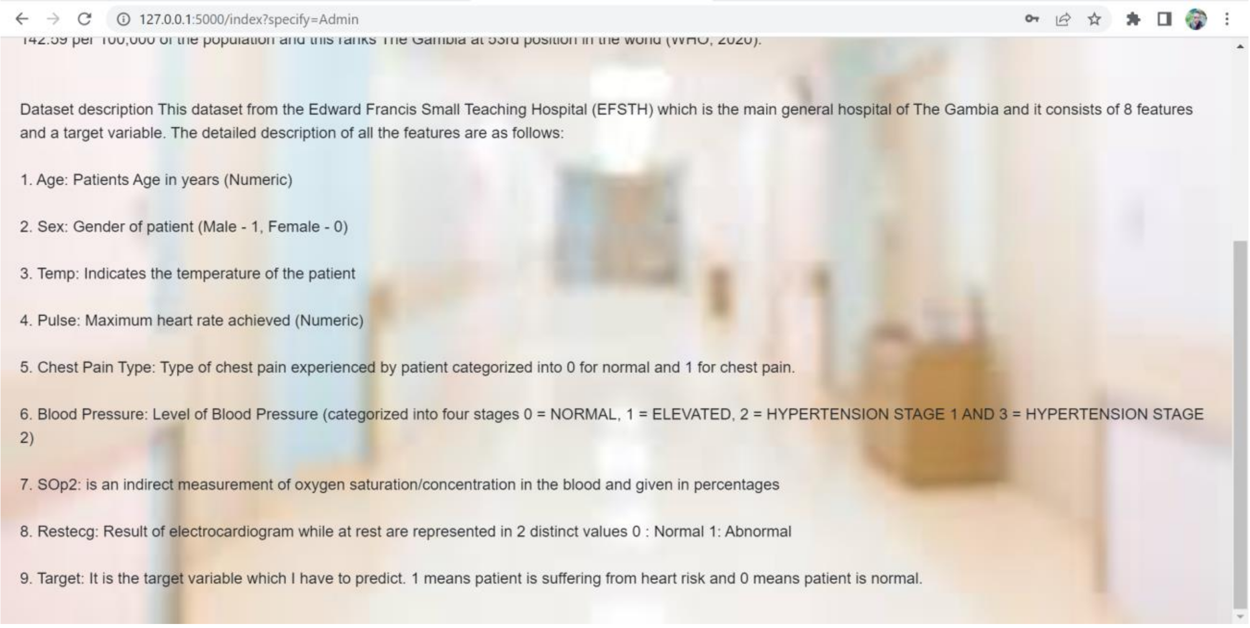
Dataset description This dataset from the Edward Francis Small Teaching Hospital (EFSTH) which is the main general hospital of The Gambia and it consists of 8 features and a target variable. The detailed description of all the features are as follows:

1. Age: Patients Age in years (Numeric)

2. Sex: Gender of patient (Male - 1, Female - 0)

3. Temp: Indicates the temperature of the patient

4. Pulse: Maximum heart rate achieved (Numeric)



142.59 per 100,000 of the population and this ranks The Gambia at 53rd position in the world (WHO, 2020).

Dataset description This dataset from the Edward Francis Small Teaching Hospital (EFSTH) which is the main general hospital of The Gambia and it consists of 8 features and a target variable. The detailed description of all the features are as follows:

1. Age: Patients Age in years (Numeric)

2. Sex: Gender of patient (Male - 1, Female - 0)

3. Temp: Indicates the temperature of the patient

4. Pulse: Maximum heart rate achieved (Numeric)

5. Chest Pain Type: Type of chest pain experienced by patient categorized into 0 for normal and 1 for chest pain.

6. Blood Pressure: Level of Blood Pressure (categorized into four stages 0 = NORMAL, 1 = ELEVATED, 2 = HYPERTENSION STAGE 1 AND 3 = HYPERTENSION STAGE 2)

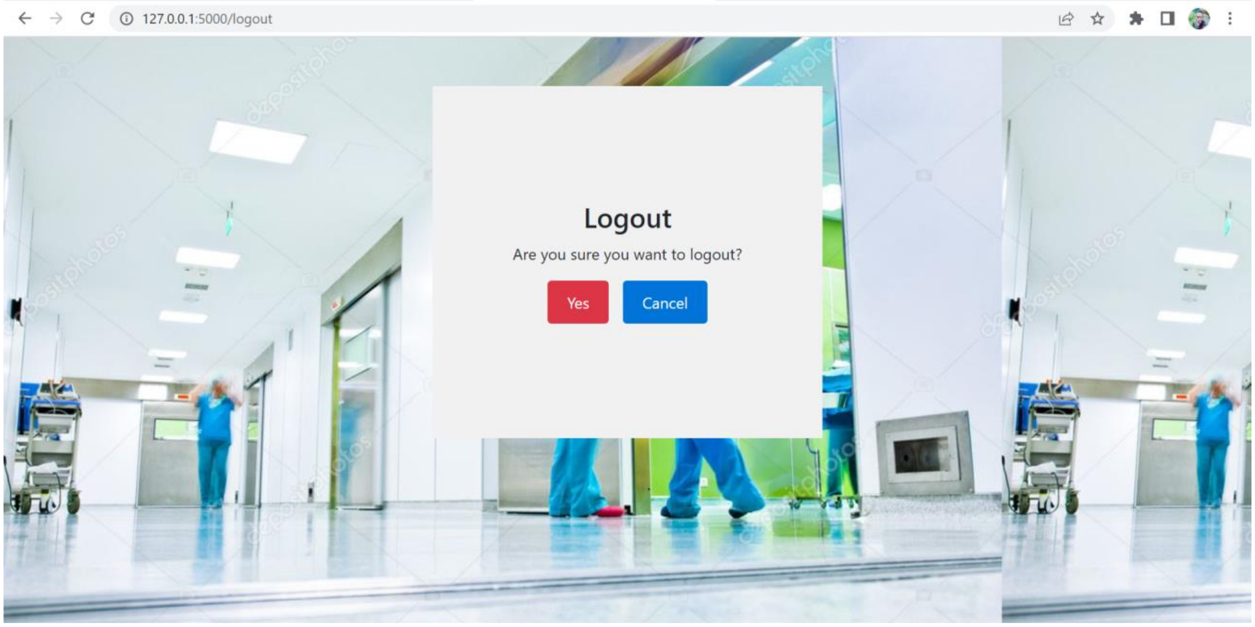
7. SOp2: is an indirect measurement of oxygen saturation/concentration in the blood and given in percentages

8. Restecg: Result of electrocardiogram while at rest are represented in 2 distinct values 0 : Normal 1: Abnormal

9. Target: It is the target variable which I have to predict. 1 means patient is suffering from heart risk and 0 means patient is normal.

Figure 30: About Page

The depicted figure 31 showcases the logout page, providing users with a straightforward means to conclude their session on the web application. By selecting the logout option, users can effectively log out and conclude their usage of the application.



Logout

Are you sure you want to logout?

Figure 31: Logout Page

3.6 Smart CVD diagonis IoT device System

The central focus of this thesis revolves around the creation of an innovative ML model that combines the prowess of Random Forest, Logistic Regression, and Support Vector Machine algorithms (RF, LR & SVM). This sophisticated model is destined to be embedded into the cutting-edge Arduino Nano 33 BLE processor, thereby offering a robust and compact solution to host the ML capabilities. Notably, this processor provides seamless compatibility with medical sensors, which play a crucial role in acquiring essential data for analysis.

The Arduino Nano 33 BLE processor stands as an exceptional platform for integrating our ML model, as it offers a multitude of capabilities while maintaining a compact form factor. By leveraging its computational power, we are able to accommodate and execute our combined ML model efficiently. Moreover, the processor's support for medical sensors ensures the reliable collection of crucial healthcare data, which is vital for the predictive aspects of our research.

To further enhance the user experience, we have implemented a speaker interface as well as an LCD display that will be connected to the Arduino Nano 33 BLE. This feature enables the display of the text on the LCD display and the transformation of the predictive text into a speech format, providing an accessible and user-friendly means of conveying critical medical insights and predictions.

The blood pressure monitor, an essential part of the system, consists of an inflatable cuff that's wrapped around the patient's arm, roughly at the level of the heart, and a monitoring device that measures the cuff's pressure. The monitor measures two pressures: systolic and diastolic. Additionally, the system incorporates an infrared temperature sensor for taking a patient's body temperature, an ECG monitoring sensor for checking for abnormal heart rhythms and signs of potential heart disease, a heart rate sensor for measuring the patient's heartbeat per minute, and an SpO2 sensor to indicate the amount of oxygen being carried by red blood cells. The SpO2 reading is a critical measurement indicating how effectively a patient is breathing and how well blood is being transported throughout the body. This value is represented as a percentage to indicate the measurement.

To ensure seamless transmission of data, the system includes a WiFi & Bluetooth module, which facilitates sending data from the Smart CVD diagnosis IoT device to the SONART AI Cardiovascular Disease Prediction System database.

Through this remarkable integration of advanced ML algorithms and state-of-the-art hardware, we aim to create a self-contained and versatile system that not only possesses the ability to predict cardiovascular diseases with high accuracy but also offers a seamless and interactive user experience. The amalgamation of hardware and software solutions paves the way for an innovative and reliable approach to healthcare data analysis and medical prediction.

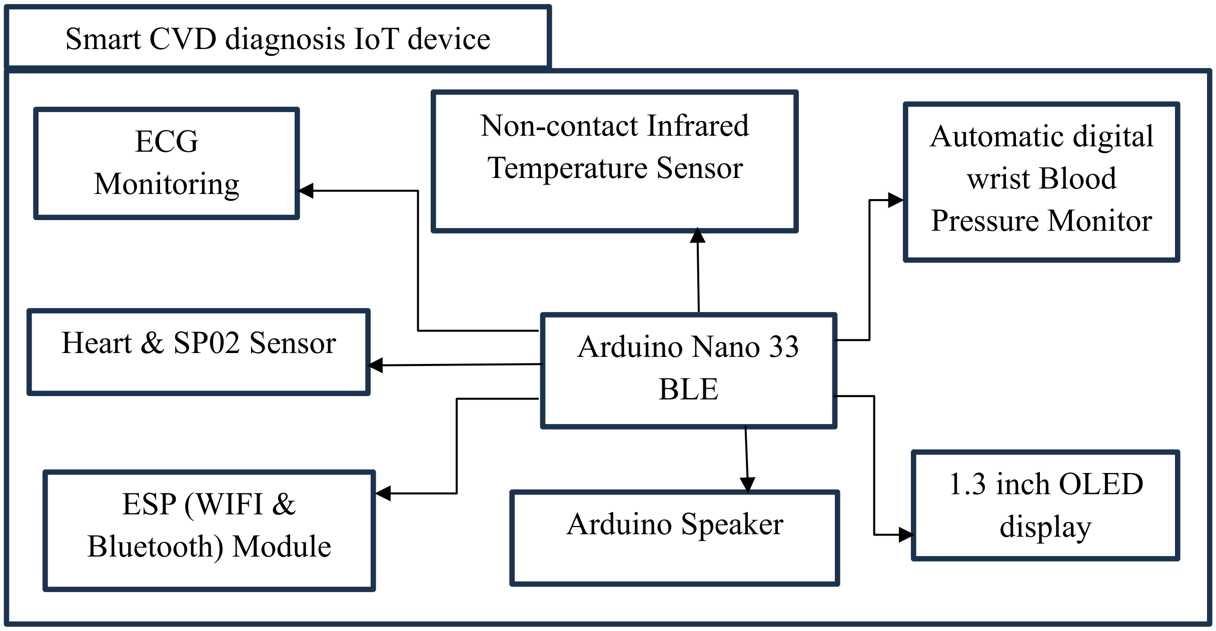

Smart CVD diagnosis IoT device
ECG Monitoring
Non-contact Infrared Temperature Sensor
Automatic digital wrist Blood Pressure Monitor
Heart & SP02 Sensor
Arduino Nano 33 BLE
ESP (WIFI & Bluetooth) Module
Arduino Speaker
1.3 inch OLED display

Figure 32: Work flow

3.6.1 Electronics Components Used to Design the Smart CVD diagnosis IoT device System

The Smart CVD diagonis IoT device system designed for the purpose of this work is a multifunctional device that required a lot of materials. The model is integrated into the Arduino Nano 33 BLE along with additional connected sensors. Subsequently, once all the sensors transmit the data, the predicted outcome is presented on the LCD display. The name, description, and reason for the choice of each piece of equipment are listed below:

Arduino Nano 33 BLE

The Arduino Nano 33 BLE is an advanced version of the conventional Arduino Nano, featuring a powerful nRF52840 processor with a 32-bit ARM® Cortex®-M4 CPU operating at 64 MHz. It offers significant improvements, including larger program memory (1MB) and increased RAM capacity, enabling the development of more extensive and feature-rich programs. The Nano 33 BLE also supports Bluetooth® pairing through NFC and boasts ultra-low power consumption modes.

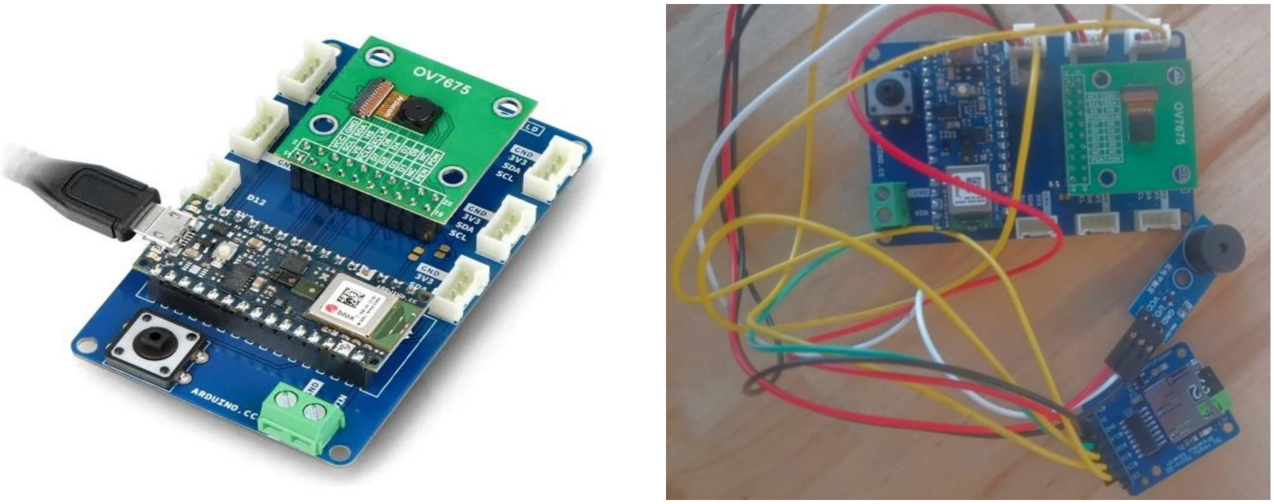


Figure 33: Arduino Nano 33 BLE. source: (aliexpress, 2023)

OLED display module

The 1.3-inch OLED display module is a compact and high-resolution display that offers excellent visual quality. With its organic light-emitting diode (OLED) technology, it provides vibrant colors, high contrast, and wide viewing angles. The 1.3-inch size is ideal for displaying essential information and graphics in a clear and concise manner. It is widely used in various applications, including wearable devices, medical monitoring systems, and IoT projects. The OLED display module offers ease of integration and programming, making it a versatile choice for displaying critical data and enhancing user interaction in cardiovascular disease monitoring and management solutions.

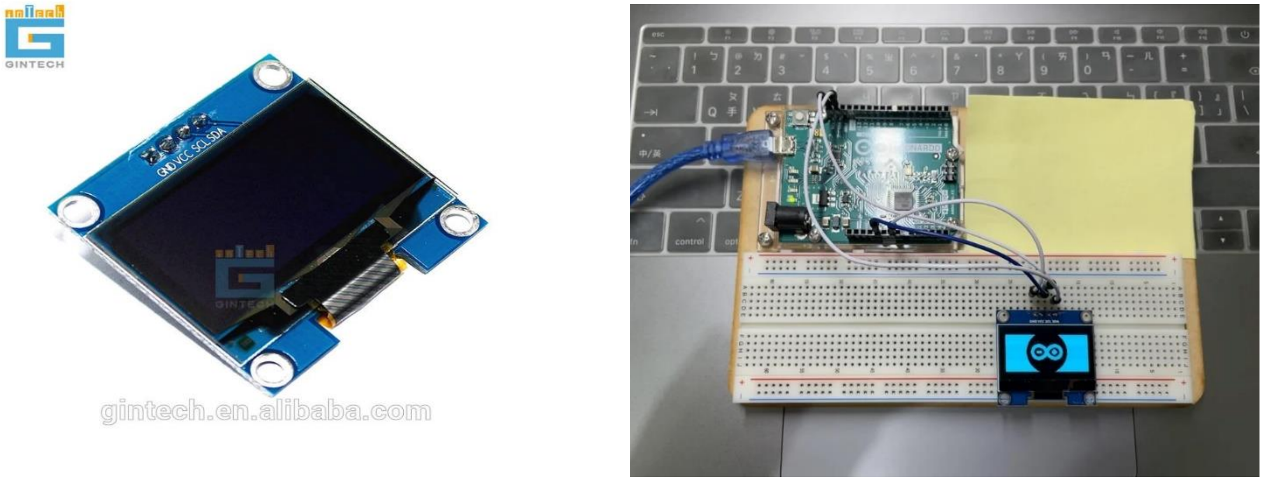


Figure 34: 1.3 inch OLED display module. source: (aliexpress, 2023)

Automatic Digital Wrist Blood Pressure Monitor

The Automatic Digital Wrist Blood Pressure Monitor will be hacked and integrated with the Arduino Nano 33 BLE, enabling seamless data transfer and enhanced functionality. Once connected, the blood pressure monitoring device will be wrapped around the patient's wrist, approximately at the level of the heart. The monitoring device, utilizing the Arduino Nano 33 BLE's capabilities, will measure the cuff's pressure and provide accurate readings for systolic and diastolic blood pressure, as well as pulse rate. With automatic inflation and deflation, the device ensures quick and effortless readings, while the digital display offers clear and easy-to-read

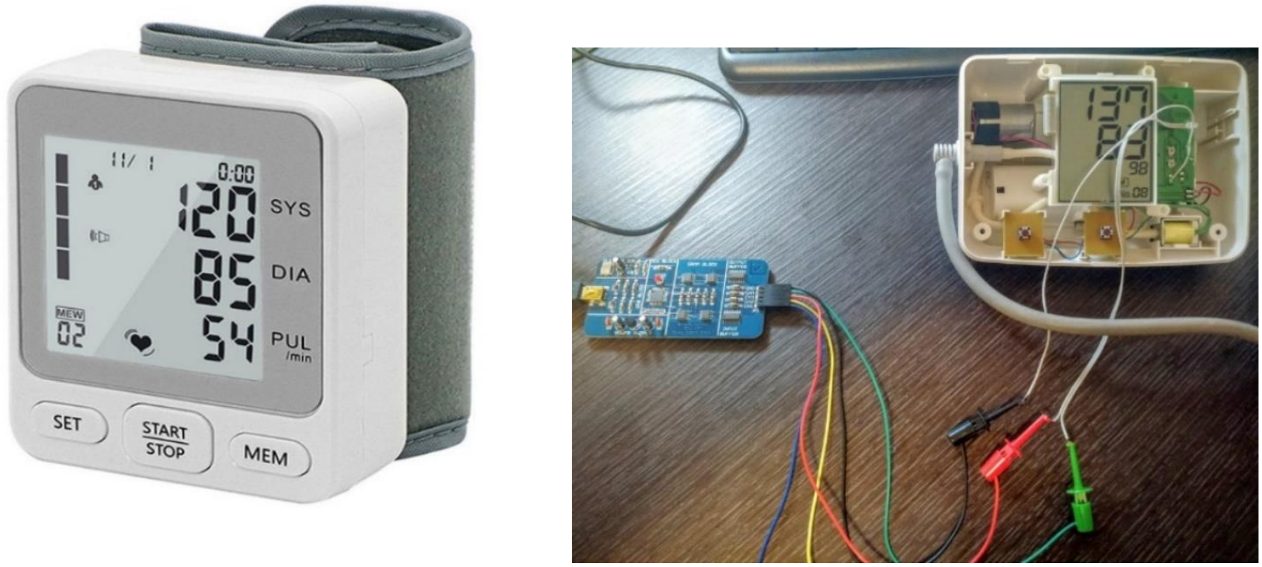


Figure 35: Automatic Digital Wrist Blood Pressure Monitor. source:(aliexpress, 2023)

MLX90614 GY-906 Non-contact Infrared Temperature Sensor IIC Interface Module IR Sensor for Arduino

The MLX90614 GY-906 non-contact infrared temperature sensor module has significant implications for cardiovascular disease prediction. By leveraging its capability to measure body temperature without physical contact, the sensor can be integrated into cardiovascular disease prediction systems to collect relevant physiological data and enhance the accuracy of prediction models.

To take a patient's temperature using the MLX90614 sensor, position the sensor at an appropriate distance, typically a few centimeters, from the patient's forehead. This non-contact approach ensures that the patient's temperature can be measured without causing discomfort or interrupting their activities. It is important to aim the sensor at the center of the forehead to obtain the most accurate temperature reading.

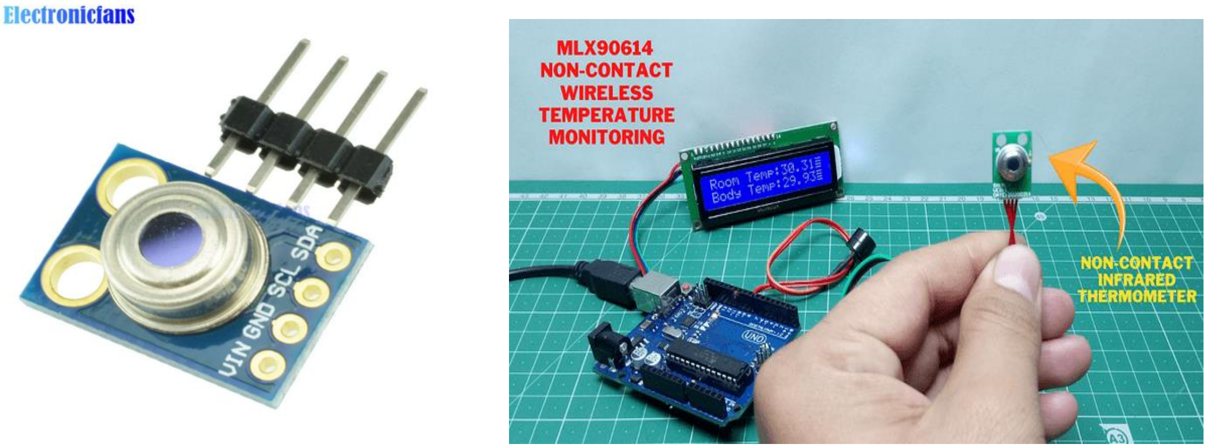

Electronicfans
VIN GND SCL SDA
MLX90614 NON-CONTACT WIRELESS TEMPERATURE MONITORING
Room Temp:30.31
Body Temp:29.93
NON-CONTACT INFRARED THERMOMETER

Figure 36: MLX90614 GY-906 Non-contact Infrared Temperature Sensor IIC Interface Module IR Sensor for Arduino. source:(aliexpress, 2023)

ECG Monitoring Sensor

The ECG Monitoring Sensor DIY Kit with AD8232 technology facilitates the measurement and recording of electrocardiogram (ECG) signals from the human body. To utilize the kit, it must be meticulously assembled in adherence to the provided instructions, and three electrodes should be strategically placed on the patient's chest area following the application of conductive electrode gel to ensure optimal contact. The leads from the AD8232 module are then connected to the electrodes, and the system is powered up. Subsequently, the data acquisition process commences, enabling the acquisition of ECG readings, which are presented and analyzed on a connected device. It is of utmost importance to acknowledge that while the kit holds substantial potential for acquiring ECG data, it does not possess the qualifications of a medical-grade device and, consequently, should not be employed for diagnostic or therapeutic purposes. The interpretation of any anomalous ECG findings should be referred to a qualified healthcare professional to ensure accurate evaluation and appropriate care.

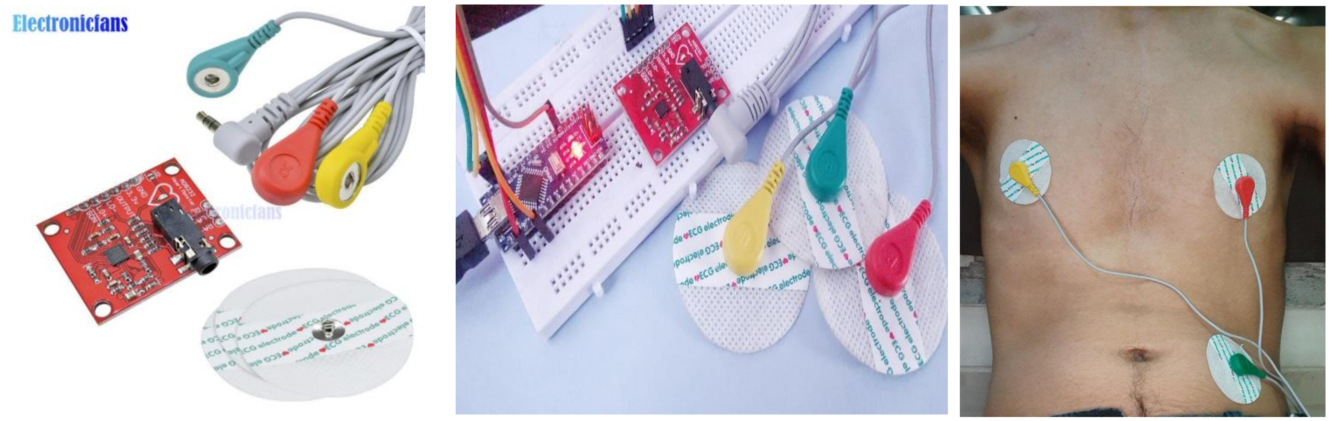


Figure 37: ECG Monitoring Sensor. source:(aliexpress, 2023)

Heartbeat & SOP2 sensor

The 30102 Tech Pulse Prevention Heart Rate Sensor can be seamlessly connected to the Arduino Nano 33 BLE board, resulting in a comprehensive cardiovascular monitoring system. The Arduino Nano 33 BLE's capabilities, such as wireless communication and data processing, enhance the sensor's functionality and user-friendliness. The board can receive heart rate and blood oxygen level data from the sensor, process it, and transmit it to other devices or a central monitoring system. Similarly, the Heart Rate Sensor Prevention 30102 Tech Pulse and Blood Oxygen Detection can be linked to the Arduino Nano 33 BLE board to measure heart rate and blood oxygen levels. By placing the sensor on the patient's chest and fingertip, the Arduino Nano 33 BLE board collects and processes the data, enabling real-time cardiovascular monitoring for remote patient care, personalized healthcare solutions, fitness tracking, and cardiovascular disease prevention.

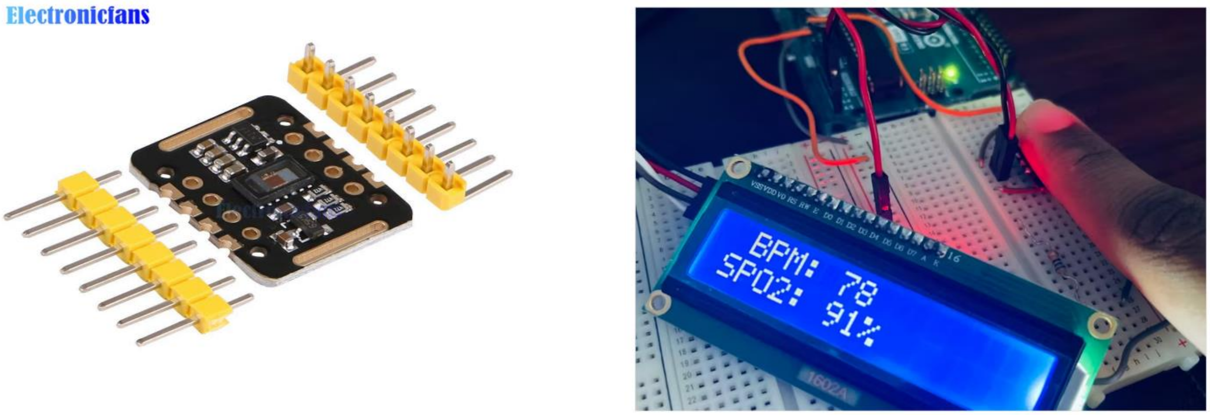


Figure 38: Heartbeat & SOP2 sensor. source:(aliexpress, 2023)

ESP (WiFi & Bluetooth) module

The ESP (WiFi & Bluetooth) module is a versatile component that can be integrated with Arduino boards, such as the Arduino Nano 33 BLE, to provide wireless connectivity capabilities. With this module, the Arduino board can establish WiFi or Bluetooth connections, enabling communication and data exchange with other devices or networks.

By incorporating the ESP module into the Arduino Nano 33 BLE, medical personels can leverage its WiFi and Bluetooth functionalities to create a wide range of applications. It will enable sending of patient data into the database of the SONART AI Cardiovascular Prediction System. This opens up possibilities for remote monitoring, control, and data sharing.

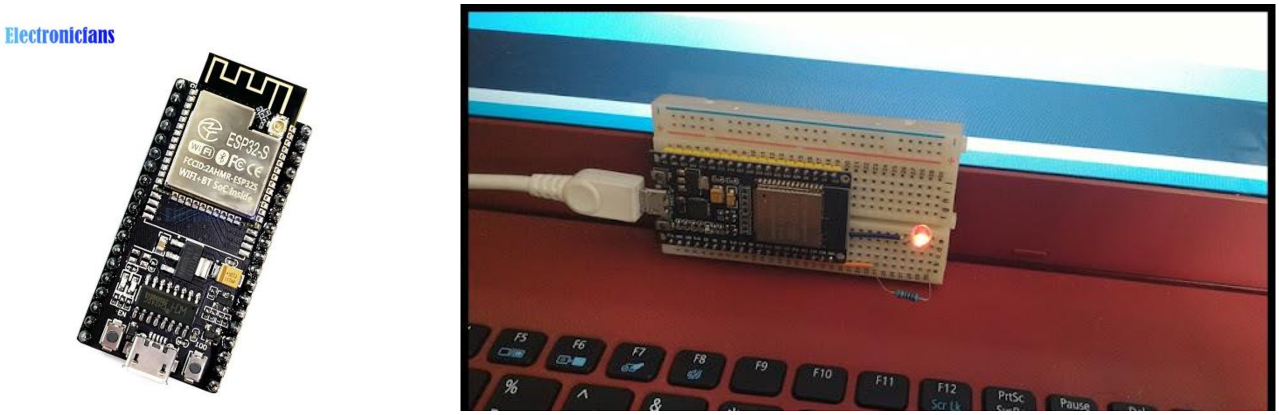


Figure 39: ESP (WiFi & Bluetooth) module. source: (aliexpress, 2023)

Arduino speaker

The Ultrathin Mini Horn Speaker, designed for use with the Arduino Nano 33 BLE and other compatible boards, serves as a compact and powerful audio component, catering to projects requiring audio playback or sound generation. By connecting the speaker to the Arduino Nano 33 BLE with a simple wiring setup, one terminal to a digital pin and the other to the ground (GND) pin, it allows for easy integration. After installing necessary libraries and writing the code, the speaker can trigger speech or audio playback through functions in the code. This functionality enables predictive text to be shown on the LCD display while simultaneously being read out in speech format through the powerful 2-watt speaker. The speaker's 8-ohm impedance guarantees clear and efficient audio delivery.

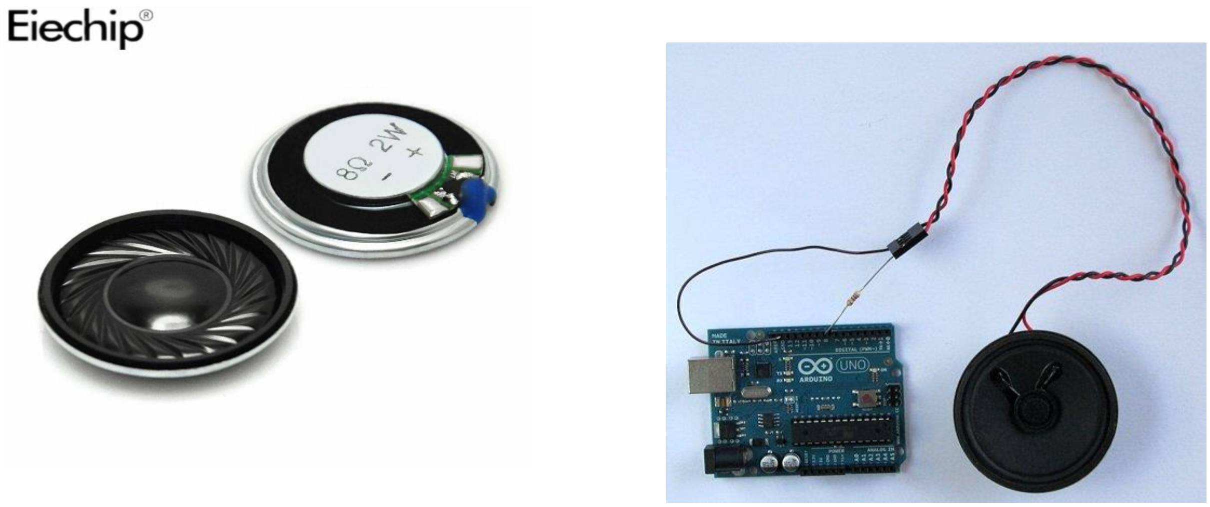


Figure 40: Arduino speaker. source: (aliexpress, 2023)

3.5 Discussions

The aim of this study was to evaluate the performance of various combined machine learning algorithms for cardiovascular disease prediction. Figure 14 presents the performance results, showing the accuracy scores of different combined models. The combined algorithm of Random Forest, Logistic Regression, and Gradient Boosting achieved an accuracy score of 93.99%, outperforming the combined model of Random Forest, Decision Trees, and Support Vector Machine, which scored 91.80%. Similarly, the combined models of Random Forest, Decision Trees, and Gradient Boosting obtained an accuracy score of 93.99%, while Random Forest, Gradient Boosting, and Support Vector Machine achieved a score of 93.44%. The highest accuracy score of 94.54% was achieved by the combined algorithm of Random Forest, Logistic Regression, and Support Vector Machine, indicating its superior performance compared to the other models.

A study by (Yuan et al., 2020) focuses on heart disease prediction utilizing machine learning techniques. Given that heart disease ranks among the leading causes of death globally, the application of machine learning has emerged as a promising approach for prediction and prevention. Nevertheless, the widespread implementation of machine learning in disease prediction on a large scale has yet to be realized.

To address this gap, the authors propose a novel algorithm called Hybrid Gradient Boosting Decision Tree with Logistic Regression (HGBDTLR) based on ensemble learning, aiming to enhance the accuracy of machine learning in heart disease prediction. This algorithm combines the strengths of gradient boosting decision trees and logistic regression, leveraging their respective advantages.

The article emphasizes the significance of big data analysis in machine learning and underscores the importance of precise prediction for heart disease. To evaluate the effectiveness of the HGBDTLR algorithm, the authors conducted experiments using the Cleveland heart disease dataset, yielding a promising prediction accuracy of 91.8%.

In summary, this article makes a valuable contribution to the field of heart disease prediction by introducing a novel algorithm that integrates ensemble learning techniques to improve the accuracy of machine learning models. The findings highlight the potential of machine learning and big data analysis in enhancing the prediction and treatment of heart disease, addressing a critical area of research and application in healthcare.

In another study by (Pan et al., 2020), the article presents a novel approach for heart disease prediction using an Enhanced Deep Learning Assisted Convolutional Neural Network (EDCNN) on the Internet of Medical Things (IoMT) platform. The diagnosis of heart disease is a complex task that requires detailed analysis of clinical test data and health history. The EDCNN model, which incorporates multi-layer perceptron models with regularization learning techniques, aims to improve patient prognostics.

The performance of the EDCNN model is evaluated using full features and reduced features, considering the impact of feature reduction on classification efficiency and accuracy. The article discusses the implementation of the EDCNN system on the IoMT platform, enabling doctors to access and diagnose patient information from anywhere in the world.

Comparative analysis with other conventional approaches, such as Artificial Neural Network (ANN), Deep Neural Network (DNN), Ensemble Deep Learning-based smart healthcare system (EDL-SHS), Recurrent neural network (RNN), and Neural network ensemble method (NNE), demonstrates the effectiveness of the EDCNN system in determining the risk level of heart disease.

The test results indicate that by fine-tuning the hyperparameters of the EDCNN model, a precision rate of up to 99.1% can be achieved. The article focuses on heart disease prediction, convolutional neural networks, and deep learning as key terms.

Overall, this research contributes to the field of heart disease diagnosis by proposing an advanced deep learning model and demonstrating its effectiveness in accurately predicting the risk level of heart disease. The integration of the model with the IoMT platform enhances the accessibility and usability of the diagnostic system for healthcare professionals.

Moreover, in a study by (Rahim et al., 2021), the authors present a framework named MaLCaDD (Machine Learning based Cardiovascular Disease Diagnosis) for the effective prediction of CVDs with high precision. The proposed framework addresses two key challenges: handling missing values (using mean replacement technique) and dealing with data imbalance (employing Synthetic Minority Over-sampling Technique - SMOTE). Furthermore, a Feature Importance technique is applied for feature selection. The authors introduce an ensemble of Logistic Regression and K-Nearest Neighbor (KNN) classifiers to achieve improved prediction accuracy.

To validate the framework, three benchmark datasets (Framingham, Heart Disease, and Cleveland) are utilized, and the achieved accuracies are reported as 99.1%, 98.0%, and 95.5%, respectively. Comparative analysis demonstrates that the predictions generated by MaLCaDD exhibit higher accuracy and utilize a reduced set of features compared to existing state-of-the-art approaches. Consequently, MaLCaDD proves to be a highly reliable framework suitable for real-world applications in the early diagnosis of cardiovascular diseases.

The article provides a comprehensive overview of the proposed MaLCaDD framework, highlighting its methodology, experimental validation, and comparative analysis, showcasing its potential to significantly improve the prediction accuracy of cardiovascular diseases compared to existing approaches.

In another study by (Pan et al., 2020), the article presents a novel approach for heart disease prediction using an Enhanced Deep Learning Assisted Convolutional Neural Network (EDCNN) on the Internet of Medical Things (IoMT) platform. The diagnosis of heart disease is a complex task that requires detailed analysis of clinical test data and health history. The EDCNN model, which incorporates multi-layer perceptron models with regularization learning techniques, aims to improve patient prognostics.

The performance of the EDCNN model is evaluated using full features and reduced features, considering the impact of feature reduction on classification efficiency and accuracy. The article discusses the implementation of the EDCNN system on the IoMT platform, enabling doctors to access and diagnose patient information from anywhere in the world.

Comparative analysis with other conventional approaches such as Artificial Neural Network (ANN), Deep Neural Network (DNN), Ensemble Deep Learning-based smart healthcare system (EDL-SHS), Recurrent neural network (RNN), and Neural network ensemble method (NNE) demonstrates the effectiveness of the EDCNN system in determining the risk level of heart disease.

The test results indicate that by fine-tuning the hyperparameters of the EDCNN model, a precision rate of up to 99.1% can be achieved. The article focuses on heart disease prediction, convolutional neural networks, and deep learning as key terms.

Overall, this research contributes to the field of heart disease diagnosis by proposing an advanced deep learning model and demonstrating its effectiveness in accurately predicting the risk level of heart disease. The integration of the model with the IoMT platform enhances the accessibility and usability of the diagnostic system for healthcare professionals.

CONCLUSION AND PERSPECTIVES

In conclusion, this research thesis aims to investigate the effectiveness of machine learning algorithms in predicting cardiovascular diseases in The Gambia, a country with a high burden of cardiovascular diseases and limited resources. The study compares the performance of different combined machine learning algorithms and identifies key factors contributing to the development of these diseases. The findings of this research have important implications for improving patient outcomes, reducing healthcare costs, and ultimately saving lives.

The study addresses the gap in the availability of an accurate and efficient predictive model and the lack of medical equipment for cardiovascular diseases diagnosis in The Gambia. By analyzing data from medical records of patients diagnosed with cardiovascular diseases at the Edward Francis Small Teaching Hospital in Banjul, the study provides valuable insights into the performance of various combined machine learning algorithms, including Random Forest, Support Vector Machine, and Logistic Regression. These algorithms have shown high accuracy in previous studies, making them suitable for this research.

Furthermore, the study explores the potential of IoT technology, gathering patient data and facilitating the accurate prediction of cardiovascular diseases. This investigation into the integration of IoT technology in healthcare services can lead to improved accuracy and accessibility of healthcare services in The Gambia.

The research findings contribute to the development of health informatics in The Gambia, a country with a developing healthcare system. By utilizing machine learning algorithms and IoT technology, the study promotes data-driven decision-making and innovative approaches to improving the quality and accessibility of healthcare services.

The comparative study design and collection of data from real patient records enhance the reliability and applicability of the research findings. The study aims to achieve better accuracy and efficiency in diagnosing heart disease patients, providing a foundation for developing more accurate predictive models and improving patient outcomes.

Overall, this research thesis addresses the pressing need for accurate prediction and early detection of cardiovascular diseases in The Gambia. By developing this system, we expect to mitigate the impact of climate change on health. The findings will further contribute to the development of a more accurate and efficient predictive model, supporting healthcare providers in making informed decisions and improving patient outcomes. The study also advances the field of health informatics in The Gambia, paving the way for innovative approaches to healthcare service delivery.

The conclusive determination of whether the hypothesis of this study is confirmed or rejected is presented in the table below.

Specific Hypothesis 1 Age, sex, pulse, blood pressure and resting ECG have some influence on a person's heart rate and using machine learning algorithms could be the most effective way of cardiovascular diagnosis	Confirmed
Specific Hypothesis 2 **Machine learning algorithms could most effectively diagnose and predict cardiovascular disease in an effective manner.**	Confirmed
Specific Hypothesis 3 **An AI web application with an embedded ML model could better support cardiovascular disease diagnosis.**	Confirmed
Specific Hypothesis 4 **An IoT could be used to mitigate the lack of medical equipment in detecting CVD and facilitate the gathering of patient data for improved predictive accuracy.**	Yet to be confirmed

Contributions

- ✓ Quick diagnosis.
- ✓ Provides a safe digital patient data storage.
- ✓ Help avoid human biases.

- ✓ Help mitigate the impact of climate change on human health.

Limitations

- ✓ Additional Variables: Additional variables like Cholesterol, FB Sugar, and OldPeak can be collected with the help of IoT.
- ✓ Implementation Challenges: Future work should promote the integration and application of the model in healthcare practice by focusing on data privacy, security, regulatory compliance.

Recommendations

- ✓ Expand the dataset: Collect data from private hospitals across The Gambia to improve the applicability of the findings.
- ✓ Include additional variables: Consider incorporating variables such as Cholesterol level, Fasting Blood Sugar, Exang, OldPeak, CA, and Thal to enhance the predictive models' accuracy and comprehensiveness.
- ✓ Diversify model evaluation metrics: Evaluate the models using recall, F1 score, and AUC-ROC, sensitivity, specificity.

Perspectives

- ✓ Expand the dataset to private hospitals.
- ✓ Test the SONART AI CVD Prediction system on the ground
- ✓ Test the Smart CVD diagnosis IoT device.
- ✓ Future models should focus on predicting specific types of common heart diseases prevalent in the Gambia. Such a targeted approach can provide healthcare practitioners in the country with more relevant and accurate insights.

BIBLIOGRAPHY REFERENCES

(Ahamed et al., 2022): Ahamed, J., Manan Koli, A., Ahmad, K., Alam Jamal, Mohd., & Gupta, B. B. (2022). CDPS-IoT: Cardiovascular Disease Prediction System Based on IoT using Machine Learning. *International Journal of Interactive Multimedia and Artificial Intelligence*, *7*(4), 78. https://doi.org/10.9781/ijimai.2021.09.002

(Anbarasi & Anupriya, 2010): Anbarasi, M., & Anupriya, E. (2010). Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. *International Journal of Engineering Science and Technology*, *2*.

(Bhatt et al., 2023): Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms*, *16*(2), 88. https://doi.org/10.3390/a16020088

(Boursalie et al., 2015): Boursalie, O., Samavi, R., & Doyle, T. E. (2015). M4CVD: Mobile Machine Learning Model for Monitoring Cardiovascular Disease. *Procedia Computer Science*, *63*, 384–391. https://doi.org/10.1016/j.procs.2015.08.357

(Brites et al., 2021): Brites, I. S. G., da Silva, L. M., Barbosa, J. L. V., Rigo, S. J., Correia, S. D., & Leithardt, V. R. Q. (2021). Machine Learning and IoT Applied to Cardiovascular Diseases Identification through Heart Sounds: A Literature Review. *Informatics*, *8*(4), 73. https://doi.org/10.3390/informatics8040073

(Connelly et al., 2021): Connelly, P. J., Azizi, Z., Alipour, P., Delles, C., Pilote, L., & Raparelli, V. (2021). The Importance of Gender to Understand Sex Differences in Cardiovascular Disease. *Canadian Journal of Cardiology*, *37*(5), 699–710. https://doi.org/10.1016/j.cjca.2021.02.005

(Dhingra & Vasan, 2012): Dhingra, R., & Vasan, R. S. (2012). Age As a Risk Factor. *Medical Clinics of North America*, *96*(1), 87–91. https://doi.org/10.1016/j.mcna.2011.11.003

(Fuchs & Whelton, 2020): Fuchs, F. D., & Whelton, P. K. (2020). High Blood Pressure and Cardiovascular Disease. *Hypertension*, *75*(2), 285–292. https://doi.org/10.1161/HYPERTENSIONAHA.119.14240

(Garg et al., 2021): Garg, A., Sharma, B., & Khan, R. (2021). Heart disease prediction using machine learning techniques. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012046. https://doi.org/10.1088/1757-899X/1022/1/012046

(Hazra et al., 2017a): Hazra, A., Mandal, S. K., Gupta, A., Mukherjee, A., & Mukherjee, A. (2017a). *Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review*.

(Hazra et al., 2017b): Hazra, A., Mandal, S. K., Gupta, A., Mukherjee, A., & Mukherjee, A. (2017b). *Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review*. 24.

(Hickam, n.d.): Hickam, D. H. (n.d.). *9 Chest Pain or Discomfort*.

(Li et al., 2022): Li, J.-X., Li, L., Zhong, X., Fan, S.-J., Cen, T., Wang, J., He, C., Zhang, Z., Luo, Y.-N., Liu, X.-X., Hu, L.-X., Zhang, Y.-D., Qiu, H.-L., Dong, G.-H., Zou, X.-G., & Yang, B.-Y. (2022). Machine learning identifies prominent factors associated with cardiovascular disease: Findings from two million adults in the Kashgar Prospective Cohort Study (KPCS). *Global Health Research and Policy*, *7*(1), 48. https://doi.org/10.1186/s41256-022-00282-y

(Moshawrab et al., 2023): Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H., & Raad, A. (2023). Smart Wearables for the Detection of Cardiovascular Diseases: A Systematic Literature Review. *Sensors*, *23*(2), 828. https://doi.org/10.3390/s23020828

(Nagamani et al., 2019): Nagamani, T., Logeswari, S., & Gomathy, B. (2019). *Heart Disease Prediction using Data Mining with Mapreduce Algorithm*. *8*(3).

(Nikhar & Karandikar, 2016): Nikhar, S., & Karandikar, A. M. (2016). *Prediction of Heart Disease Using Machine Learning Algorithms*. *2*(6).

(Pacheco et al., 2021): Pacheco, S. E., Guidos-Fogelbach, G., Annesi-Maesano, I., Pawankar, R., D' Amato, G., Latour-Staffeld, P., Urrutia-Pereira, M., Kesic, M. J., & Hernandez, M. L. (2021). Climate change and global issues in allergy and immunology. *Journal of Allergy and Clinical Immunology*, *148*(6), 1366–1377. https://doi.org/10.1016/j.jaci.2021.10.011

(Pal et al., 2022): Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022). Risk prediction of cardiovascular disease using machine learning classifiers. *Open Medicine*, *17*(1), 1100–1113. https://doi.org/10.1515/med-2022-0508

(Pan et al., 2020): Pan, Y., Fu, M., Cheng, B., Tao, X., & Guo, J. (2020). Enhanced Deep Learning Assisted Convolutional Neural Network for Heart Disease Prediction on the Internet of Medical Things Platform. *IEEE Access*, *8*, 189503–189512. https://doi.org/10.1109/ACCESS.2020.3026214

(Patel & Patel, 2016): Patel, J., & Patel, D. S. (2016). *Heart Disease Prediction Using Machine learning and Data Mining Technique*.

(Rahim et al., 2021): Rahim, A., Rasheed, Y., Azam, F., Anwar, M. W., Rahim, M. A., & Muzaffar, A. W. (2021). An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases. *IEEE Access*, *9*, 106575–106588. https://doi.org/10.1109/ACCESS.2021.3098688

(Tülay Karayilan et al., 2017): Tülay Karayilan, Z., Özkan Kiliç, M., Jones, R. W., & Alberti, T. (2017). Prediction of Heart Disease Using Neural Network. *Proceedings of the 15th*

Annual International Conference of the IEEE Engineering in Medicine and Biology Societ, 277–278. https://doi.org/10.1109/IEMBS.1993.978541

(Yuan et al., 2020): Yuan, K., Yang, L., Huang, Y., & Li, Z. (2020). Heart Disease Prediction Algorithm Based on Ensemble Learning. *2020 7th International Conference on Dependable Systems and Their Applications (DSA)*, 293–298. https://doi.org/10.1109/DSA51864.2020.00052

Visited websites

https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm 13/03/2023

https://docs.arduino.cc/hardware/nano-33-ble-sense 16/04/23

https://www.malicksarr.com/type-of-machine-learning-algorithms-the-complete-overview/ 27/04/23

https://www.britannica.com/place/The-Gambia / 25/04/2023

https://en.wikipedia.org/wiki/The_Gambia/ 25/04/2023

https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm/ 19/04/2023

https://www.hackster.io/DKARDU/how-to-make-blood-oxygen-body-temperature-measurement-583c31 12/05/23

https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm/ 20/04/23

https://botland.store/arduino-nano-boards/19352-arduino-tiny-machine-learning-kit-with-arduino-nano-33-ble-sense-lite-akx00028-7630049202771.html 21/05/23