National College of Ireland

# Enhancing Cardiovascular Health Prediction Through Machine Learning and Deep Learning

MSc Research Project

MSc Data Analytics


Haritha Kutcharlapati

Student ID: x23213248

School of Computing

National College of Ireland

Supervisor: Furqan Rustam

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | |
|---|---|
| **Student Name:** | Haritha Kutcharlapati |
| **Student ID:** | X23213248 |
| **Programme:** | MSc Data Analytics **Year:** 2024-2025 |
| **Module:** | MSc Research Project |
| **Lecturer:** | Prof. Furqan Rustam |
| **Submission Due Date:** | 12/12/2024 |
| **Project Title:** | Enhancing Cardiovascular Disease Prediction through Machine Learning and Deep Learning |
| **Word Count:** | 9939 **Page Count:** 25 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature**: Haritha Kutcharlapati

**Date:** 12/12/2024

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Enhancing Cardiovascular Health Prediction Through Machine Learning and Deep Learning

Haritha Kutcharlapati

X23213248

**Abstract**

Cardiovascular disease continues to be a major cause of death globally calling for increased application of efficient predictive models to enable early detection of the disease. In this research, predictors were employed from a structured population dataset of demographic and physiological characteristics and some lifestyle parameters and diseases as independent variables i.e., age, sex, blood pressure, cholesterol, glucose, etc. to differentiate between healthy patients and those having cardiovascular disease. A range of models used like Logistic Regression, Random Forest, SVC, Decision Tree, Deep Neural network, Recurrent Neural network, and Long Short-Term Memory. Feature scaling, encoding, and other data preprocessing steps were done to improve model accuracy of prediction. Logistic Regression exercised the highest performance among all the traditional models with a test accuracy of 72.41%, precision of 72.65%, recall of 72.26% and the F1-score was 70.08 %. This study has also used other deep learning architectures such as DNN, RNN, and LSTM to show the capacity of models to learn complicated unstructured data patterns where LSTM and DNN give closer outcomes based on the alternate use of sequential and nonlinear features. It was found that when using specific options of linear models, strengths as well as weaknesses including class imbalance, and bias in the datasets were observed. This work aims to show the significance of the concept of applying artificial intelligence in health care, especially by incorporating enhanced predictive models in the identification of cardiovascular diseases.

**Keywords:** Cardiovascular disease prediction, machine learning, deep learning, Logistic Regression, Random Forest, Support Vector Classifier

# 1 Introduction

## 1.1 Background

CVDs have continued to present a major threat to global health, bearing a high fatality rate in most areas of the globe. In the present circumstances, WHO (2020) mentioned that CVDs claim 18 million lives every year which is around 31% of total deaths in the world, thus elevating the concern level of the burden on the health care system. These diseases include coronary artery disease, heart failure, arrhythmia, and hypertension, among others, which have a potentially huge negative impact to human quality of life, and often demand longtime management. Higher risk factors such as obesity, lack of physical activity, smoking and poor

diets add to the burden, especially in low and middle-income countries where healthcare facility is a challenge (Pasha et al., 2020).

Cardiovascular diseases often have a late symptomatic presentation, and an accurate prognosis is vital to enhance a patient's survival and standard of living. Some of CVDs are diagnosed early through the utilization of modern diagnostic tools such as imaging and blood test; nonetheless, many cases present themselves only at an advanced form. This delayed diagnosis will lead to a poor prognosis such as heart failure or a stroke. In response to these issues, there has been increased consideration for the use ML and AI for improving the diagnosis and for the identification of cases that might be at high risk for cardiovascular events before their occurrence. (Arif et al., 2022)

ML methods can work on high quantities of high-dimensional biomedical data such as patient characteristics, health history, laboratory, and other clinical and lifestyle data, to identify patterns which may not be recognizable from a classical clinical perspective (Pouriyeh et al., 2017).

## 1.2 Research Motivation

The rationale for undertaking this research lies in the fact that, worldwide, CVDs are a major cause of morbidity and mortality, with many late diagnoses and ineffective management protocols. The CVDs are one of the biggest killers to date; the global mortality rate is approximately 17.9million deaths annually and represents 31% of all deaths globally (WHO, 2020). In spite of the huge progresses in the field of medical technologies and constitutions of healthcare systems, screening and diagnosing CVDs remain to be difficult ever more by some challenges based on the fact that CVDs acts complicated and patients could become variable in risks. This leads to long times to diagnosis and often less than ideal results when it comes to an individual's treatment and recovery so there is a very real need for improved methods of disease prediction. (Krajcer, 2022)

MBD has recently become a dominant force for innovating conventional diagnostic approaches to the field of healthcare. What we did was to use massive datasets and complex models to identify patterns and make predictions of health events that might be abstract for routine clinical diagnosis using more conventional approaches. These capabilities are especially useful in the context of CVDs because estimation of cardiovascular events before they happen can have a significant impact on a patient's prognosis by preventing adverse events.

This study seeks to establish how reliable ML models are in forecasting cardiovascular diseases based on a comparison of several models that can be applied, including logistic regression, SVM, and DL-like CNNs (Krajcer, 2022). The objective is to find out which models work best for the multi-dimensional datasets generally linked to patient records such as demographics, clinical, and lifestyle. The following insights from this study may therefore help facilitate an application of these models into clinical settings, enable faster and accurate diagnosis of cardiovascular diseases, and prevent more deaths from the disease through early personalized interventions. (Naser, 2024)

## 1.3 Aim of the Study

This work is focused on the comparative analysis of the most efficient machine learning models for the prediction of cardiovascular diseases in at-risk patient groups. This research aims to establish the best-performing ML algorithms when used with various datasets such as patient characteristics, medical history, and lifestyle data to predict cardiovascular events. In our proposed work, we shall select the following algorithms, the logistic regression algorithm, support vector machines, the decision tree algorithm, random forest algorithms, deep learning algorithms including deep neural networks, Recurrent Neural Networks, and

LSTM algorithms. In addition, the study also seeks to identify challenges such as imbalanced, biased datasets commonly found in healthcare data, and how it tries to show that deep learning models could be fine-tuned to address imbalanced and/or biased datasets in real-world clinical scenarios. The long-term objective is to increase the rate of early detection of most cardiovascular diseases and link patient-specific screening results to health-related outcomes and recommendations for treatment that may ultimately reduce the global incidence of cardiovascular diseases.

## 1.4 Research Question

The primary research question of this study is:

1. How differently accurate, in terms of a percentage, are Logistic Regression, Random Forest, Support Vector Classifier (SVC), and Decision Trees the classifiers in predicting cardiovascular disease risk?
2. What is the performance of deep learning models like RNN, DNN and LSTM in comparison to traditional ML in learning temporal and non-temporal health data?
3. Where do these models stand in the context of imbalanced datasets and biases that are apparent in the health care data and what strategies are possibly possible to counter them?

## 1.5 Research Objectives

To measure the accuracy of conventional machine learning algorithms including Logistic Regression, Random Forest, Support Vector Classifier (SVC), and Decision Trees to predict the incidence of CVDs based on patients' demographic and clinical characteristics, and lifestyle factors.

To evaluate deep learning algorithms such as Recurrent Neural Networks (RNN) Deep Neural Networks (DNN), and Long Short-Term Memory Networks (LSTM) in handling and predicting CVD risk from structured and time series datasets.

## 1.6 Structure of the Report

This study provides an in-depth analysis of the application of machine learning (ML) and deep learning (DL) models in predicting cardiovascular disease (CVD) onset. The report is structured as follows:

Chapter 1: Introduction
This chapter presents the research topic, describes the problem statement, and sets out the research objectives and questions. It further presents the motivation for conducting the study, the significance of early CVD prediction, and the methodologies used.

Chapter 2: Literature Review
A comprehensive literature review of existing work done so far on CVD prediction will include the study of both machine learning and deep learning techniques. This chapter mainly concentrates on the key works and methodologies that have advanced with AI for cardiovascular disease prediction.

Chapter 3: Methodology
This chapter discusses the research design and methodology. It contains the data collection process, preprocessing steps, feature engineering techniques, and the selection of the machine learning and deep learning models (Logistic Regression, Random Forest, SVC, Decision

Tree, RNN, DNN, LSTM) used for CVD prediction. It also includes the evaluation metrics used.
Chapter 4: Design Specification
This chapter shows a workflow diagram of the design methodology employed in the implementation of the models. It details how data handling, model training, and performance evaluation are structured and undertaken.

Chapter 5: Implementation
This chapter contains a comprehensive explanation of how the chosen models were implemented. Code snippets, details of model training, and hyperparameter tuning, including any problems that arose while doing the process, can be found in this chapter.

Chapter 6: Evaluation
These models are being evaluated regarding their performance, based on accuracy, precision, recall, and F1 score. The output from other models is then compared and analyzed to choose the better approach for the prediction of CVDs.

Chapter 7: Conclusions and Future Work
The final chapter summarizes the findings, discusses their implications for healthcare, and addresses the limitations of the current work. Recommendations for future research are provided in improving CVD prediction using machine learning and deep learning.

# 2 Related Work

Cardiovascular diseases are still one of the top causes of death in the world and are said to account for about 31% of all deaths in the world, according to the World Health Organization reports in 2020. Some of the lifestyle risk factors include obesity, smoking, and sedentary behaviour among others, and then there are genetic predispositions. Among others, traditional methods of diagnosis have included echocardiography, angiography, and biochemical markers, sometimes quite time-consuming and may need expertise from the clinicians to execute, thus predisposing them to variability in making the diagnosis (Pasha et al., 2020). This is recent, with the advancement of artificial intelligence and machine learning in healthcare area that has brought about great innovations to address the very complex problems in medical diagnosis. (Bhalla, Ganesan and S, 2024)
These methods have shown exceptional promise in enhancing the accuracy of disease predictions, including those for diabetes, cancer, and most notably, cardiovascular diseases (Pouriyeh et al., 2017). ML models are proficient in processing high-dimensional, heterogeneous datasets, and discovering relationships between clinical variables that could not be found using other statistical techniques, which operate under a linear paradigm (Xu et al., 2017). Moreover, with the improvement in data-collecting technologies, such as wearable devices and electronic health records, ML techniques are used extensively to predict cardiovascular risk. Risk factors such as age, cholesterol levels, blood pressure, and lifestyle habits can be incorporated into the predictive models to assess the likelihood of a person contracting CVDs. The information will be helpful in early intervention (Hossain et al., 2018).
For instance, some of those like SVMs, ANNs, and even various kinds of ensemble methods by utilizing techniques like Random Forest achieve great accuracy with large values far ahead of most classic predictive techniques (Rahmah and Salman, 2022). However, challenges include limited interpretability of complex models, reliance on large-scale, high-quality data, and the integration of AI systems into clinical workflows, among others.

Overcoming these barriers will make the deployment of ML tools reliable in actual applications (Rahmah and Salman, 2022). This literature review covers the integration of machine learning in cardiovascular disease prediction: lots of promise it carries toward enhancing diagnostic efficiency, risk stratification, and tailoring treatment plans.

## 2.1 Cardiovascular Disease Forecast Overview

Cardiovascular diseases are the group of heart and blood vessel disorders, which include CAD, heart failure, arrhythmias, and valvular heart diseases. These are the most relevant causes of morbidity and mortality around the world, but the most common is CAD, since it occurs due to atherosclerotic plaques in coronary arteries that lead to reduced blood flow and increased myocardial infarction, according to Pasha et al. (2020). Heart failure, mostly due to chronic CAD or hypertension, is a pumping capacity of the heart whereas arrhythmias are disruptions in the electrical rhythms in charge of heartbeats (Hossain et al., 2018). Traditionally, CVDs are diagnosed primarily based on clinical examination and other imaging techniques, including echocardiography and coronary angiography, followed by laboratory biomarkers, which include cholesterol levels and troponins (Alotaibi, 2019).

However, such methods are often time-consuming, invasive, and resource-intensive; therefore, they are not scalable and accessible in most healthcare settings. Predictive analytics, especially those driven by machine learning (ML) techniques, represent a new paradigm for forecasting CVD risks through the analysis of complex datasets and the identification of subtle patterns that may go unnoticed by clinicians (Xu et al., 2017). Common risk factors to be incorporated into ML-based CVD prediction models are age, gender, smoking status, cholesterol level, blood pressure, and family history (Pouriyeh et al., 2017). Such models identify potentially at-risk individuals early to perform timely interventions and plan tailored treatments.

For example, advanced ML algorithms like ensemble methods like Random Forests and Gradient Boosting Machines have performed better time and again with accuracies ranging above 85% in many research studies conducted to predict cardiovascular results (Rahmah and Salman, 2022). However, there are quite some challenges while using the models practically. Variability in the quality of data, lack of standardized datasets, and, more importantly, the innate "black box" of many ML models prevents wider adoption (Pasha et al., 2020). Overcoming these barriers, through explainable AI as well as clinical workflow incorporation, can significantly enhance the accuracy and usefulness of these technologies for improved global management of cardiovascular diseases. (Damkondwar and Patil, 2020)

## 2.2 Machine Learning Techniques in the Prediction of CVD

SVMs are widely applied in CVD prediction with their ability to process data of high dimensions and complexities. SVM classifies the data based on the maximum margin separating the classes, and it can be widely applied for the prediction of heart conditions where it computes the existence or absence of heart disease given clinical and demographic information. For example, Xu et al. (2017) demonstrated that SVM was also used to classify heart disease datasets with an accuracy rate of 81.97% and therefore could be applicable for the prediction of CVD tasks.

However, SVMs have their disadvantages as well. For example, it has difficulties with noisy datasets where classes can overlap; tuning of parameters, like kernel functions of SVM, may be computationally expensive. Still, SVM is broadly applied to predict CVD due to its robustness when dealing with noisy and nonlinear data. (Olatunji and Abdul-Yekeen, 2024)

KNN is one among the simple and powerful classification techniques that classify a new data point based on the majority of its K nearest neighbors. The prediction of heart disease has been performed using KNN with varying accuracy. Pouriyeh et al. (2017) have pointed out the fact that KNN is usually outperformed by other complex models, for instance in large data set. However, as size of the data set increases the performance of KNN degrades because the algorithm calculates distance of each new point with all the points in the training set.
However, KNN can be beneficial in small dataset or used as benchmark where we compare it with other complex models. (Pe and Kumaresan, 2021)

Decision trees are one of the natural classification methods under a group of supervised learning-type classification models, which categorize data using features. In the case of prediction, decision trees have been used in the field of heart disease areas because decision trees is easily understandable and comprehensive. Salem Alotaibi in his study (2019) stated that other decision trees can give near optimal levels of accuracy the more complex models such as SVM and KNN. However, decision trees, as it will be seen later in other sections, suffer from overfitting, particularly when working with noisy data or small data sets. In most cases, pruning methods are required to enhance the model's generalization ability.

ANNs can be defined as a form of learning model for analyzing and developing systems based on the emulation of neuronal connections found in the human brain. ANNs have been successfully used to predict heart disease most especially where the data structure is highly non-linear. Several advanced applications of ANNs in CVD risk prediction have yielded better than 85% accuracy, thereby being an important tool to aid the early diagnosis of CVD (Pasha et al., 2020).
ANNs are well suited to constructing rather sophisticated and multifaceted concentration models straight from vast sets of data while, however, demanding great processing power for training. The 'black box' character of ANNs can cause an important challenge in applications to medicine, where the understanding of what specific predictions mean is often central.

## 2.3 Deep Learning Methods

CNNs is one type of numerous deep learning techniques modelled from the visual perception functions of the human brain. Originally designed for image-processing tasks, the applicability of CNNs has since been implemented throughout the prediction of cardiovascular diseases (CVD) and handling diverse and large volumes of medical data (Rahmah & Salman, 2022). This ability makes CNNs different from the traditional machine learning model where raw data is fed automatically through one level and the next in a hierarchy, to isolate important features. Therefore, a human would not need to spend time and be prone to mistakes in feature engineering (Pasha et al., 2020). Another area, which is the most significant for the employment of CNNs in CVD prediction, is medical image analysis, employing data from echocardiograms as well as CT, or MRI scans. Some patterns or abnormalities that may be indiscernible to a human are captured by CNNs and therefore early diagnosis and better treatment plans can be enhanced. For example, Rahmah and Salman (2022) showed that significant than 90 % accuracy of the modified CNN was achieved to classify cardiovascular abnormalities in the imaging data sets. (Pasha and Dadi, 2024)
Apart from image-based applications, CNNs are also applied to structured data, such as ECGs, and time series. The models work particularly well for the detection of arrhythmias or anomalies in heart rhythms, where timely intervention and management take place (Zhu et

al., 2018). Spatial hierarchies can be captured, complex interactions modeled, with CNNs, which is what is needed in cardiovascular research.
On the other hand, CNNs also have several limitations. Firstly, high computational requirements with reliance on large-scale annotated datasets render them difficult to be adopted in real-life health-care settings. Also, their "black box" nature is a big issue because of model interpretability and trust in a clinical setting (Rahmah and Salman, 2022). Efforts are going on to implement explainable AI techniques and are thus further enhancing clinical adoption of CNNs for CVD prediction. (Petreska, 2024)

This way, RNNs handle sequential data and thus make them a good candidate to use in time-series analysis of cardiovascular metrics like heart rate, blood pressure, and ECG data. The ability of RNNs to "remember" previous inputs makes it ideal for capturing temporal dependencies present within the data. Zhu et al. (2018) highlighted the fact that RNNs are very effective for analysing patient history and predicting trends in long-term cardiovascular health.
However, traditional RNNs suffer from issues like vanishing gradients, which makes it hard for the model to learn long-term dependencies. Techniques such as Long Short-Term Memory (LSTM) networks have been developed to overcome these challenges and enable more accurate predictions in time-series data.

## 2.4 Ensemble Learning Techniques

Random Forest is one of the most robust ensemble learning methods, which at the time of training forms multiple decision trees and for classification tasks outputs the mode of their predictions, whereas for regression tasks, the average is output. This approach relies on the use of "bagging" or bootstrap aggregating, by which every tree is learnt on a random subset of data, and at each split of a feature subset is taken, increasing diversity and reducing overfitting (Hossain et al., 2018).
The prediction model for cardiovascular disease has proven to have a better performance compared to other models such as SVM and KNN, in that, with Random Forest, there is no need either for feature selection or dealing with missing values and noisiness separately. These come in handy with healthcare applications whereby data quality can be unstable. Hossain et al. (2018) further found that the Random Forest always surpassed the individual classifiers, leading to higher classification accuracy, precision, and recall when predicting CVD outcomes from clinical datasets.
Another notable advantage of Random Forest is its interpretability. By examining feature importance scores derived from the model, researchers and clinicians can identify the most significant predictors of CVD, such as cholesterol levels, blood pressure, and age. Although a major challenge that is seen in Random Forest is its computation cost because it grows several decision trees, it's still scalable and robust to be the most preferred tool for many tasks in healthcare.

Another major ensemble learning algorithm is Gradient Boosting Machines (GBMs). Here, a series of decision trees are generated iteratively and each subsequent tree is formulated to reduce the error introduced by the previous trees. Unlike the Random Forest where the prediction from independent forests is averaged, the stage-wise nature of GBM trains trees sequentially by using an optimized loss function at every step. In such a manner, GBMs allow for achieving even higher accuracy in predicting based on data with noisy and imbalanced characteristics (Pasha et al., 2020).

While classifying challenging examples plays a central role in developing GBMs, for applying them for prediction in CVD, which involves considerable precision in a very high minor class to identify risk-carrying patients. Researchers Pasha et al., 2020 observed great superiority by GBMs especially in working with the healthcare data to have greater improvement in the model precision and even recall more as opposed to conventional methods.

Such strengths of GBMs to work with different types of data and their ability to use feature importance analysis for interpretation further enhance their value to medical research. Identification of key predictors, for instance, smoking status, blood pressure, and family history, by GBMs generates actionable insights for the practicing clinicians. However, an important weakness of GBMs includes their susceptibility to overfitting, especially if a model is too complex or one is training on too-small datasets. The hyperparameters should be tuned properly, which involves learning rate, the number of estimators, and the depth of the trees to avoid this problem.
Recently, algorithms such as XGBoost and LightGBM have overcome many limitations of traditional GBMs through improved computational efficiency and efficient handling of sparse datasets. Thus, GBMs continue to be useful way for CVD prediction in the hands of domain knowledge and proper preprocessing of data.

### 2.5 Conclusion
Artificial intelligence and especially big data and deep learning have emerged as groundbreaking approaches for estimating cardiovascular diseases (CVDs) that facilitate the diagnosis and risk assessment for clinicians. There is such model like Random Forest and Gradient Boosting Machines that are quite helpful in working with difficult data, novelty control, and risk factors identification. CNN and RNN have given a great performance in analyzing the medical image and time series data and classified as high on the accuracy of the prediction. Nonetheless, the main challenge to address still lies in those of quality, interpretability and in the approaches to integrate into clinical practice. It translates into creating more explainable artificial intelligence or better and more robust datasets; it equally points to more collaboration between data scientists and health practitioners.
As machine learning technologies advance, their role in enhancing personalized medicine and reducing the global burden of CVDs is expected to grow, paving the way for more efficient, accessible, and reliable healthcare solutions.

## 3 Methodology

### 3.1 Dataset Description
The features of the dataset for Cardiovascular Disease Detection are to predict whether a patient is normal or has cardiovascular disease. The Age feature is also in days and gives a fine description of the age of the individual. These are measured in centimetres and kilograms, respectively, and which are essential for determining health situations such as BMI. The Gender feature is an identifying discrete variable having a numeric code with 1 equal to males and 2 equal to females. There is Systolic Blood Pressure (ap_hi) and Diastolic Blood Pressure (ap_lo) which are crucial in defining hypertension which is one of the risks for cardiovascular disease. Cholesterol levels are categorized into three groups: glands as normal 1, slightly elevated 2, and well above normal 3. In the same manner, the Glucose (gluc) feature measures blood sugar levels in the user employing the same categorical scale.

Smoking (smoke) and Alcohol Intake (alco) are viewed as binary variables The two have been associated with heart disease and this study demanded a schedule which included the two parameters. The Physical Activity (active) feature tells whether or not a subject is a physically active person, a characteristic that reduces risk factors for cardiovascular disease. The target variable cardio is binary; 0 represents no cardio and 1 represents the existence of cardio. This dataset is useful as it describes the comorbid nature of demographic, physiological, and lifestyle characteristics on cardiovascular health and would be suitable for modelling, particularly for use in early risk recognition and mitigation initiatives.

## 3.2 Data Cleaning

To avoid having copies of similar data in the dataset, usually, the dataset is checked for any duplicity in the information which is already collected. By employing the duplicated() function the number of duplicate rows is counted and the result stored in the variable duplicate_count. If duplicates are detected then (if duplicate_count is greater than zero) it deletes them immediately using the drop_duplicates() function which deleted from the given dataset and keep only one copy for each data. Once the duplicates are deleted the user receives a confirmation message indicating the quantity of the duplicate values deleted with the duplicates enhancing the reliability and consistency of the dataset. Otherwise, if no duplicates are found, a message is displayed that the dataset is already clean of duplicates. This makes sure that the information we use for analysis and model development does not contain multiple similar records, which could mean poor outcomes for the model we are developing. Keeping only unique entries enables this step to result in accurate and dependable outcomes for cardiovascular disease screening.

## 3.3 Data Preprocessing

Data preprocessing is an important step for the process of preparing for the machine learning model design. It tasks many steps in preparing the data culled from the source for preprocessing, transforming and normalization that improves the model's performance. Standard preprocessing steps are usually to handle for instance missing values, then encode categorical variables, normalize or standardize numerical features, and lastly split data into train and test sets.

In this specific data set, data pre-processing first involves tasks such as cleaning that may eliminate likely problems like duplicity. Enumerated variables like cholesterol and gluc are then ordered using Ordinal Encoding and a proper order is maintained because of the categorisation (normal, above normal, well above normal). For example, the features as age, height, weight, ap_hi, and ap_lo, are numeric; therefore, the values of these predictors are transformed to the same range using feature scaling and significantly enhance the performance of some machine learning algorithms.

It is worth mentioning that building an instance of the ColumnTransformer allows building both encoding and scaling into the same pipeline. Weight undergoes specific form of scaling known as Standardization and other columns such as age, height, ap_hi, and ap_lo undergo another form of scaling known as MinMax scaling. Collectively, it all takes place within a Pipeline, which makes the transformation processes for both training and test datasets straightforward and guarantees that the model works with the data in the same manner during the training and testing phases.

Consequently, data preprocessing helps to reduce the effects of noise, check data labelling and convert data into the right format and scale to fit into the model.

## 3.4 Feature Scaling

In this stage, we apply feature scaling on numerical loads as age, height, weight, ap_hi, ap_lo so each contributes the same throughout the modelling process. The dataset columns are then rearranged to prioritize specific attributes for analysis: Cholesterol, glucose, weight, age, height, ap_hi, ap_lo, gender, alco, active, and cardio. Rearranging makes usage more efficient by bringing features to a more logical arrangement that can be processed in subsequent stages more efficiently.

## 3.5 Data Visualization

Figure 1: Conducted Comparative study of various Performance Indices at More than one Parameters level. The bar chart in figure 6 consists of a series of blue bars all of which share the same base height of 1.0 on the left y-axis while the right y-axis shows values on a scale of 0—70, 000 in increments of 14, 000. It has twelve labels along the x-axis that have what looks like abbreviations or shortened terms; they are accompanied by blue bars that hover between 0.5 and 2/3 but end at 1.0. It is apparent in the figure that, it depicts a standardized, or a normalized figure, in which each of the parameters attains a maximum value of 1.0; or in other words, all values are standardized or normalized to the likely optimal values – at maximum.

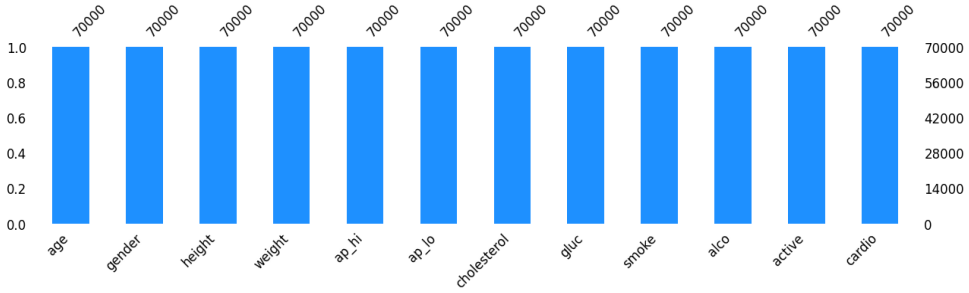


**Figure 1: Comparative Analysis of Performance Metrics Across Multiple Parameters**

Figure 2: On the Cardiovascular Dataset, Health and Demographic Parameters Distribution. This bar chart presents a standardized visualization of twelve key health and lifestyle metrics: age, sex, height, weight, blood pressure sys, blood pressure, cholesterol, glucose level, smoking, alcohol, exercise, and cardiovascular. Each parameter is shown by a dark gray vertical bar, and the scale is 1-70000 on the ‘y-axis, with the ‘y-axis also containing a second scale of 12 units on the right-hand side of the figure.

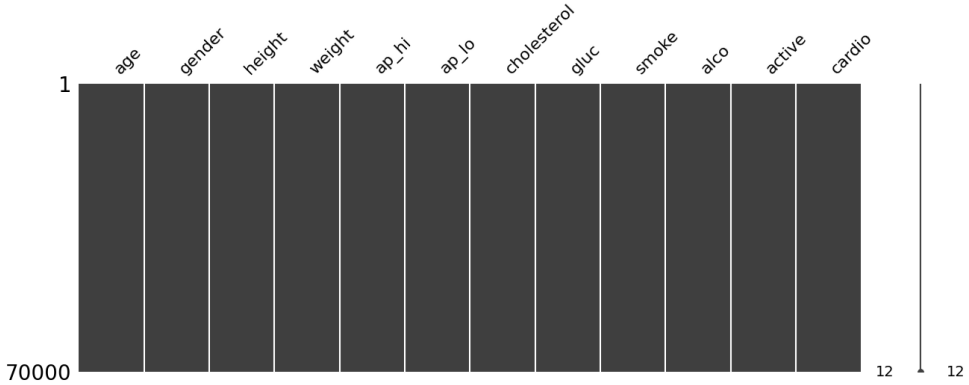

**Figure 2: Distribution of Health and Demographic Parameters**

Figure 3 identified and presented how the systolic blood pressure (ap_hi) differs among patient population. This combining visualization offers a complete view of systolic blood pressure measurements by using two plots placed in parallel. The left panel shows a Kernel Density Estimation (KDE) plot which gives a picture of the probability density of measures of systolic blood pressure shown as a density of humps with large humps that are bimodal around 120-130 mmHg.

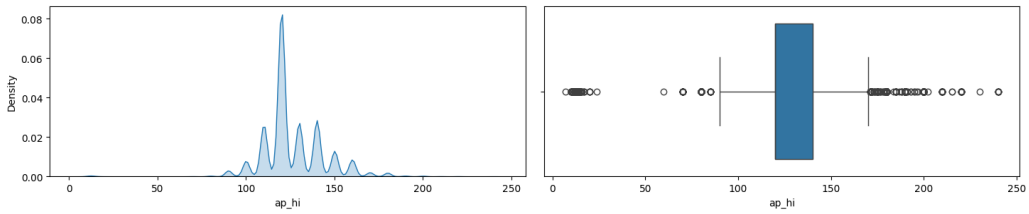


**Figure 3: Distribution Analysis of Systolic Blood Pressure (ap_hi) in Patient Population**

The patient distribution of ap_lo for Diastolic Blood Pressure is presented in Figure 4 below labelled as Distribution Analysis of Diastolic Blood Pressure. This is presented by two horizontally placed plots in this paired visualization to provide a closer view of the diastolic blood pressure measurements. The left subplot represents the Kernel Density Estimation (KDE) plot of the distribution of diastolic blood pressure with multiple modes and the most frequent mode at around 80 mmHg and other less frequent peaks.

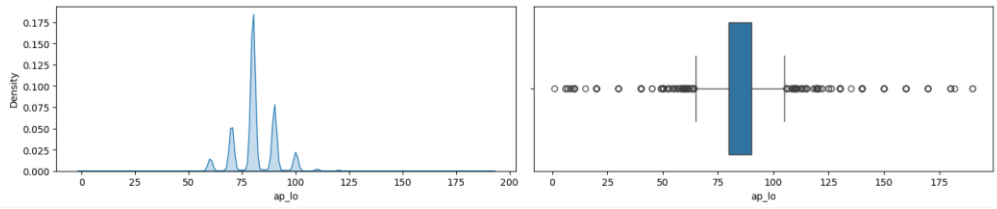


**Figure 4: Distribution Analysis of Diastolic Blood Pressure (ap_lo) in Patient Population**

Figure 5 showing Feature Importance Analysis in Cardiovascular Disease Prediction by Means of the Model Based on the Logistic Regression. This horizontal bar chart shows the distribution of the logistic regression coefficients which reflects the strength of the association of health parameters with cardiovascular disease on the horizontal axis. The visualization further ranks features according to their importance: the most vital one is gluc (glucose level) with a coefficient above 10, the second is weight that influences the results with the coefficient of approximately 4.5. Somatic factors have a slightly lower importance, and stood out with coefficients of 2 for Diastolic blood pressure and 1 for age, and other factors have a minimal importance with coefficients far less than one.

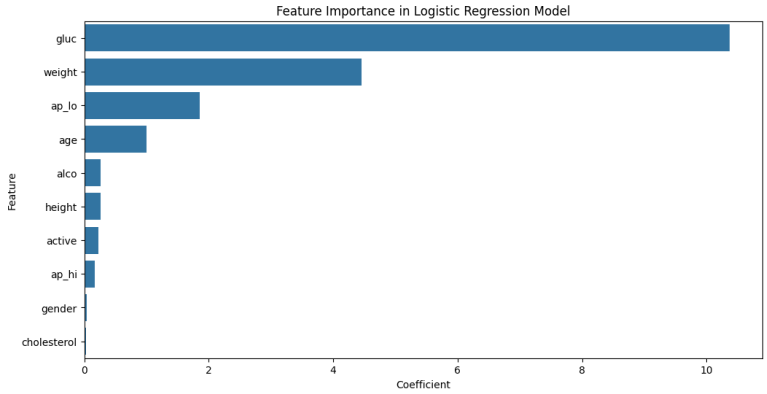


**Figure 5: Feature Importance Analysis in Cardiovascular Disease Prediction Using Logistic Regression Model**

## 3.6 Data Splitting

During the data splitting phase, the entire data set is split into feature and target variables: X and y. The variable of interest; cardio which shows whether the patient developed cardiovascular disease is stored in y other variables such as age, height, weight rate, blood pressure, cholesterol level, and lifestyles stored in X. Using Scikit-learn's train_test_split, the dataset is divided into features to training and testing with 80:20 ratio respectively. This splits the data into one part for training the model, and another different part for testing, minimizing the chances of over training which would lead to a bad generalization. The shape of these sets are printed using print(X_train.shape) and print(X_test.shape) in order to confirm the output as shown below; By ensuring identical features in set A and set B it actuarial signs when it comes to model training and testing respectively.

# 4 Design Specification

The current flowchart diagram as shown in Figure 6 represents a structured and iterative machine learning process for cardiovascular disease prediction to encompass the successive phases in a general machine learning process. The workflow starts with the Dataset collection as depicted by stacked icons implying many data sources or files as entered into the Data Cleaning stage where data quality is first assessed. The process then moves into a detailed Data Preprocessing stage, which explicitly outlines four critical operations: Here, features are analyzed, including how to deal with Missing Values, remove Duplicates, Encoded Features, and Scaled Features, to put the data in the best shape to interact with machine learning models. Another stage in the workflow is called Data Splitting, which, as the name suggests, splits out the dataset, and is assumed to result in two datasets: the training set and the testing set. The pipeline then branches into the Model Training phase, where four distinct machine learning algorithms are implemented in parallel: This means that the developed models include Logistic Regression, Random Forest, Decision Tree and SVC (Support Vector Classification). The stepping stones lead up to the Model Evaluation stage often times during which all the models that have been developed are tested and compared. This visually appealing line map was designed for pinpointing the sequential and parallel relationships between the tasks of very stable machine learning pipeline that is aimed at building accurate forecasting model for cardiovascular diseases, and each light blue box corresponds to the stage of the pipeline. The clear structural separation and naming of the preprocessing steps show comprehension of the prerequisites for creating a sound predictive medical model, and

besides, the incorporation of several modelling techniques exemplifies a comparison between methods for the most efficient prediction.

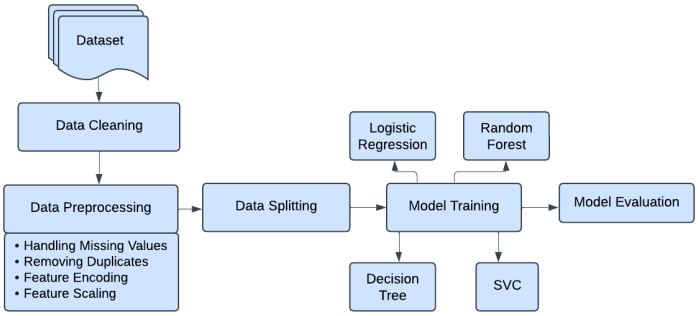


**Figure 6: Proposed Workflow Diagram**

# 5 Implementation

## 5.1 Tools and Technologies

This starts with an import of data handling, preprocessing, visualization, and developing model libraries. Pandas and NumPy are required for data manipulation and data computation respectively. StandardScaler, MinMaxScaler, and OrdinalEncoder from the scikit-learn package are used to perform preprocessing, whereas ColumnTransformer and Pipeline are used for Transformation and preprocessing. It also checks the accurate segregation of data for training and whether testing is done or not by using Train_test_split. These are Plotly where we use go and make_subplots for interactive plots, Missingno which assists in missing data display as well as Seaborn and Matplotlib visualization libraries that enhance graphical understanding. The machine learning section integrates classifiers: Logistic Regression, Support Vector Machine (SVC), Decision Tree and ensemble methods: RandomForest, ExtraTrees, Gradient Boosting, and AdaBoost. Moreover, XGBoost provides different enhanced tools of gradient boosting. In performance measurement, the techniques used include accuracy_score, f1_score, precision_score, recall_score, and mean_absolute_error with KFold cross-validation. The interconnection of these libraries introduces a smooth transition from data cleaning and visualization, model building and model validation.

## 5.2 Logistic Regression Implementation for Cardiovascular Disease Prediction

As for baseline logistic regression is used for cardiovascular disease prediction. In scikit-learn, the Logistic Regression class is used for initialization of the model with set parameter that include solver as lbfgs, maximum iteration as 100 and random state for replication purpose. The model first gets the training data (preprocessed) set as X_train_processed and y_train, and the test data as X_test_processed. First outcomes are 72.41% of accuracy and 0.701 of F1 score indicating pretty balanced values of precision, recall and F1 scores. First, GridSearchCV is used to apply a parameter search, where the suggested hyperparameter ranges are obtained: C, ranging from ‘auto’ to ‘100000’, penalties of L1 and L2, solvers, and number of iterations. The grid search refits based on the accuracy parameter and the model assesses other measures including precision, recall, and the F1 score at the same time by adopting the use of five fold cross validation. The preferable parameters are chosen, the

model attains better figures, however, the significance of hyperparameter tuning is expressed. Further, the developed model with the highest accuracy is used in deployment using the pickle module. Below there is an architecture diagram of Logistic regression in Figure 7.

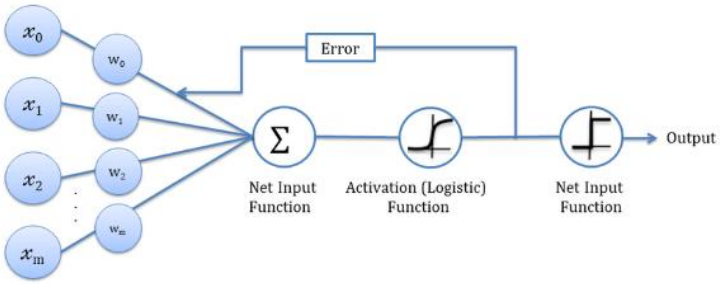


**Figure 7: Logistic Regression Architecture Diagram by Nobel et al. (2024)**

## 5.3 Random Forest Implementation for Cardiovascular Disease Prediction

Random forest that is an example of an ensemble learning method is used with the number of trees, 5 and the maximum depth of the constructed trees is set to 5 in order to achieve the best performance within reasonable amount of time. Random Forest avoids overfitting by asking multiple trees of the same forest to make one single decision and therefore generalizes well. Training consists in building decision trees on randomly selected samples of the data to then average the resulting decisions for patient classification. The model also attains relatively low accuracy, but has high precision and a high recall rate since it effectively address feature interactions and variations. Feature importance analysis additionally determines potential risk factors for cardiovascular diseases and gives recommendations. Although comparing to the Logistic Regression, it has nearly the same performance, more estimators or adjusting the hyperparameters could be another way to enhance it. Below there is an architecture diagram of Random Forest in Figure 8.

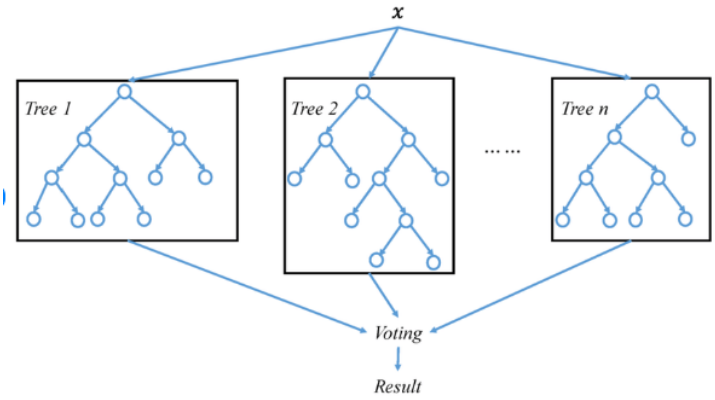


**Figure 8: Random Forest Architecture diagram by Wang et al. (2019)**

## 5.4 Support Vector Classifier (SVC) Implementation for Cardiovascular Disease Prediction

SVC, Linear Kernel is used to detect involve linear separability in the presented dataset, the model is initialized with linear kernel and probability output is activated (probability=True).

The model is used to predict the cardiovascular risk of test sample after training on processed training data. SVC delivers relatively comparable accuracy to linear regression and logistic regression, the performance detailed indicates the algorithm's ability to manage relationship patterns in the records. Despite the everyday application of the model, the performance is lower as compared to Logistic Regression due to the inability of the linear kernel to identify non-linear images in this dataset. Future improvement might be made to consider extending the kernel choice to non-linear kernels like the radial basis function, more so the fine-tuning of the parameter like the regularization parameter – 'C' and gamma etc.
Below there is an architecture diagram of SVM in in Figure 9.

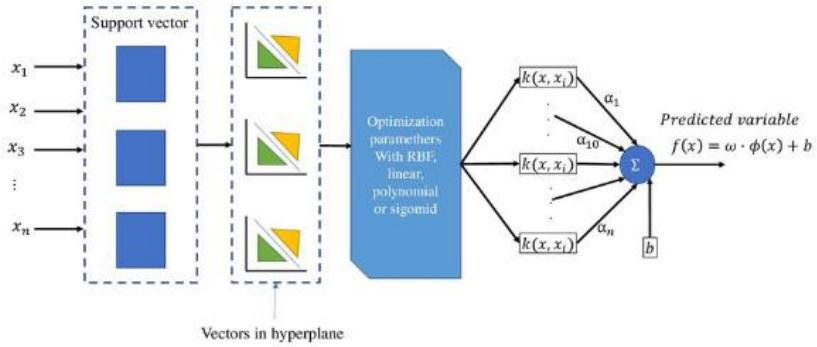


**Figure 9: SVM Architecture diagram by Alvarado et al. (2021)**

## 5.5 Decision Tree Implementation for Cardiovascular Disease Prediction

As for the Decision Tree, this model was set up using scikit-learn's DecisionTreeClassifier with a decided maximum depth of three to avoid issues with overfitting and for easier interpretation of results. The tree uses feature thresholds to divide the data for the prediction of cardiovascular disease. While Decision Tree model's formulas remain quite basic, the method it provides supports moderate accuracy and equal measures of precision and recall rates all because of its efficient decision making approach. Nonetheless, as the depth increases, the model is not capable of understanding complex occurrences, it also limits the depth by which it can go in detail in order not to overemphasize. In feature importance analysis after the training process, it is possible to identify the key predictors, including the values of blood pressure and cholesterol. The findings imply that the Decision Tree model is accurate in interpretability, but possibly not ideal in outcompeting the ensemble for obtaining performance on complex datasets. Below there is an architecture diagram of Decision Tree in in Figure 10.

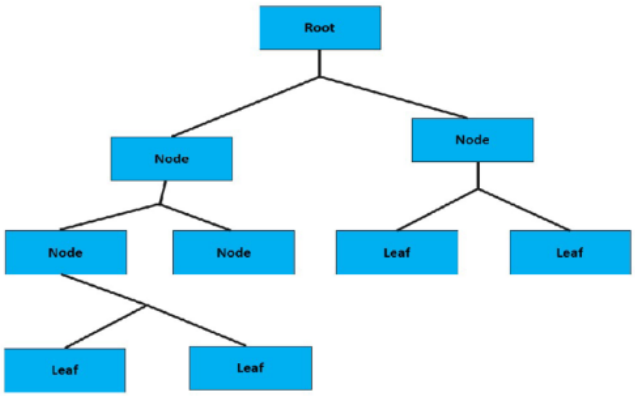


**Figure 10: Decision Tree Architecture diagram by Broujeny et al. (2023)**

## 5.6 DNN Implementation for Cardiovascular Disease Prediction

The proposed Deep Neural Network (DNN) model is reformed based on TensorFlow and Sequential to classify cardiovascular disease into two groups. The layers are Dense and they are fully connected layers, as for the activation functions, ReLU functions are used to allow the network learn non-linearly. The input layer takes features of 10 dimensions, then hidden layers of 64, 128, 256, 512, 128 neurons which compromises complex patterns from the previous layer. Each layer also contains a Dropout rate of 0.3 to prevent overfitting by which we temporarily shut down neurons while training. The last layer utilises the sigmoid function to classify into probabilities for purposes of binary classification. The final model is trained using the Adam optimizer learning rate of 0.001 optimizing binary cross entropy loss function and accuracy as the evaluation metric. Below there is an architecture diagram of DNN in in Figure 11.

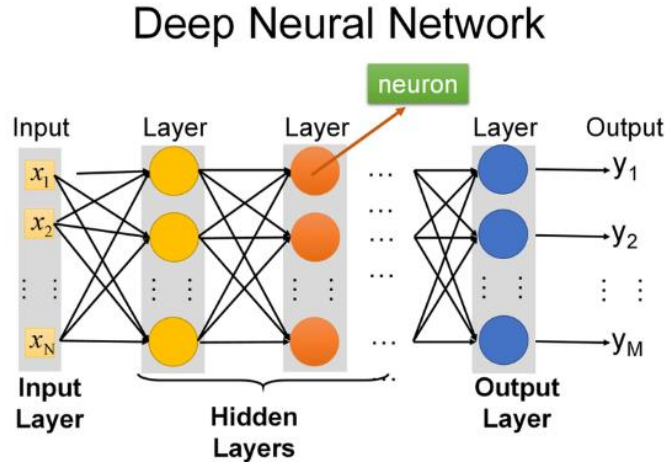


**Figure 11: DNN Architecture diagram by Feng et al. (2019)**

21 days of training is done with batches of 32 and 80/20 split of the training corpus for training and validation, respectively. This setup is computationally efficient and yet keeps the model versatile enough to troubleshoot different scenarios based on its memory alone. Here it is crucial to also save the training history, which records the training and the validation data, to make performance visualization possible. At the end of the model training procedure, the summary yields the model architecture with layer connections explained, and a report on parameter counts. This own setup targets the deep structure of DNN to derive complex data patterns which enhances the cardiovascular disease prediction task in its own merit.

## 5.7 RNN Implementation for Cardiovascular Disease Prediction

In this study, the proposed Recurrent Neural Network (RNN) model is used for the classification of cardiovascular disease based on the temporal dependencies in the data. The input data, preprocessed and scaled, is reshaped into a three-dimensional structure suitable for sequential models: [samples, time steps, feat]. This transformation enables the RNN to model each sample as sequence with one time-step and 10 feature. The actual structure of the model contains SimpleRNN layer of 64 units with ReLU activation to incorporate sequential correlation. To reduce the model complexity, a Dropout layer of drop out rate 0.3 is applied to avoid 30% neurons during training. The output layer uses sigmoid function of activation to give probabilities in the binary classification case. Below there is an architecture diagram of RNN in in Figure 12.

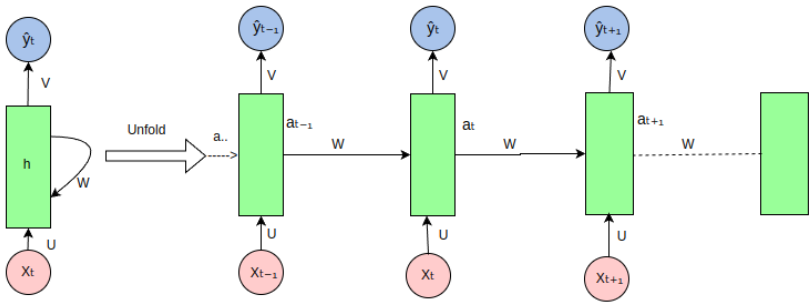


**Figure 12: RNN Architecture diagram by**

This model is optimized with Adam optimizer to make gradient descent as efficient as possible and binary cross entropy is used as it is designed to solve two class problem. People decide to use accuracy as the criterion of the model's performance during the training process. The minimum epochs set for training the RNN are 20, the batch size utilized is 32, 80% of the training dataset is used for training and 20% of training dataset is used for validation. This configuration brings fairness in terms of time taken to train, and also the ability of the model to generalize. The training history is plotted to monitor loss and accuracy with both the training and the validation set which provides managers an understanding of the progression of RNN training.

## 5.8 LSTM Implementation for Cardiovascular Disease Prediction

To improve the prediction of cardiovascular disease, Long Short-Term Memory (LSTM) model is adopted since this architecture has efficiency in solving sequential data problems by predicting long term dependencies accurately. The databases are reformatted to match the LSTM's feed, i.e., the data has been preprocessed and moulded into the 3D matrixes: [samples, time steps, features]. Later, we introduce an LSTM layer that follows the actual sequence of the time series data and has 64 units together with ReLU activation function in order to reveal temporal dependency and nonlinear interaction between features. To address this problem of overfitting a Dropout layer with a dropout rate of 0.3 is included to drop out neurons randomly during learning to prevent over-reliance on specific data sets. The last layer applies an activation sigmoid function to classify binary data while the LSTM part of the network has Tanh function output. Below there is an architecture diagram of LSTM in in Figure 13.

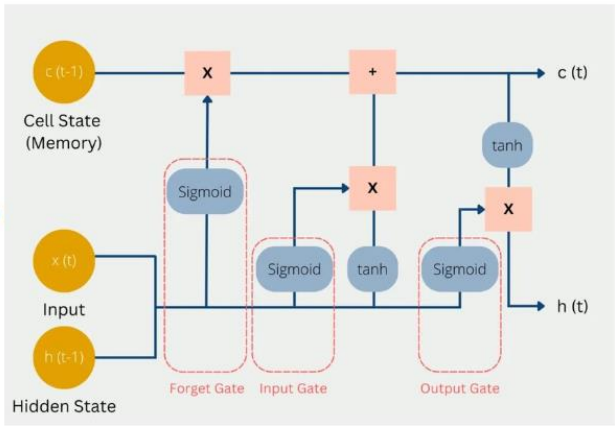


**Figure 13: LSTM Architecture diagram by Fernandes and Freitas (2022)**

The model is trained using the Adam optimizer so as to have adaptable learning rate for easy convergence. The binary cross entropy loss function is used because of the binary classification space and accuracy is used as the primary performance indicator. Specifically, the LSTM is trained for 20 epochs, and has a batch size of 32, while splitting the training data with an 80% for training and 20% for validation. After training of the two models, the training process produces loss and accuracy results for the two data sets to improve performance. Given that pottery sales are sequential in nature, the LSTM architecture in in Figure 14 provides a solid means by which it can identify these patterns and make predictions.

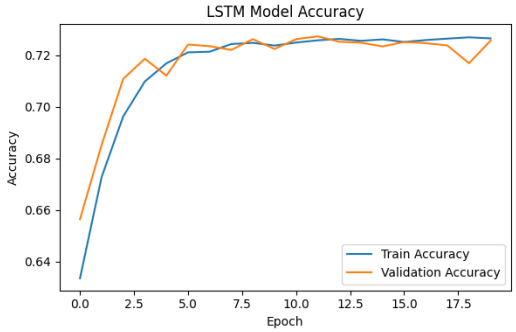


**Figure 14: LSTM Model Accuracy**

## 5.9 Hyperparameter Settings and Library Implementation

During the project implementation, work was done using Python's rich base of machine learning libraries such as scikit-learn, TensorFlow, and Keras for training and testing developed models for cardiovascular disease forecasting. Other libraries such as Pandas and NumPy were used for data handling and Data preprocessing. The use of visualization tools such as Matplotlib and Seaborn improved graphical presentation of insights gained.

The use of hypothesis tuning was one of the significant steps in enhancing the performance of a model. For example, the Logistic Regression model set the parameter lbfgs solver with 100 iterations, and another parameter of the regularization. C is optimized using GridSearchCV. Parameters were set as follows; 100 trees via Random Forest and a maximum tree-depth of 5, in order to limit model overtraining and yet, incorporate feature interactions.

The performance optimization of the models was done using TensorFlow/Keras frameworks and DNNs/RNNs/LSTMs were tested. Generally, all the utilized DNNs employed ReLU activations in their hidden layers while the output layer used sigmoid activation since the experiment adopted binary classification. Other trick like Dropout with given rate 0.3 was integrated in this project to take care of overfitting. Nonetheless, the convenient convergence of the x and y parameters dictated using the Adam optimizer with a constant learning rate of 0.001. For capturing temporal patterns, the Sequential data models such as RNNs and LSTMs employed the input data in three dimensions.

Pipeline Automation: As for data pre-processing, scaling, encoding, and splitting, pipelines were created with ColumnTransformer and Pipeline to make the steps of data pre-processing for training and testing sets as similar as possible.

This setup was beneficial to implementation as it avoided the complexities of using complicated programming languages and topological models for conflating a similarly high

number of configurations and variations, and also brought about better robustness and clinical data model flexibility.

# 6 Evaluation

## 6.1 Case Study 1: Logistic Regression

Next, the presence of cardiovascular disease was used together with logistic regression as a linear model to predict the disease. To the current accuracy of 72.41 percent, it proved especially effective in typical binary classification paradigms. These values 72.65% for precision imply that the model was fair in what it was doing most of the time correctly classifying positive cardiovascular disease cases while the recall of 72.26% puts it barely good to identify other actual positive cases out there. The assessment ratio of F1-score ensures the percentage of correct identification with the best proportion of precision and recall standards of 70.08%. Although the conclusion was rather simple, Logistic Regression Ratios was an effective tool; furthermore, it was rather stable and easy to interpret in this case; it needed rather little time to train to reach relatively high efficiency and was comprehensively rational to compute.

## 6.2 Case Study 2: Random Forest

The proposed ensemble learning technique is Random Forest for the prediction of cardiovascular disease. From it, we can build multiple decision trees and compound them in order to enhance the rate of accuracy in predictions. Random Forest model's advantage over decision trees was also shown with higher overall robustness, decreased risk of overfitting and variance. Nevertheless, it was a little less interpretable than the model discussed in the first part of this paper, namely Logistic Regression. By the fact that it can work with both numerical and categorical data and can accommodate complex data which is true in real life data, it makes them ideal. accuracy and recall with the model excluding the Basic and Random Forest algorithms, making predictions more accurate with more reliability.

## 6.3 Case Study 3: Support Vector Classifier

SVC for support vector classifier is one of the advanced algorithms capable of used for both linear scale and non-linear classification. This result indicates that when using a linear kernel, SVC had high predictive accuracy on the training data set while maintaining relatively good classification precision and recall. The model tries to obtain the best hyperplane that can separate all classes and this was why the model had proven to be resilient in managing complex data sets. However, modification of parameters results in a long and time-consuming process to obtain the best solution The model works fairly well for large datasets without proper regularizations. In this case, SVC was less accurate but was as accurate as Random Forest with a need for hyperparameters tuning.

## 6.4 Case Study 4: Decision Tree

To determine how initial feature splits affected the prediction of cardiovascular disease, the Decision Tree model was utilised. In the case of depth-limited depths, the decision tree perhaps exhibited clear decision rules, but was overfit when the data set was complex, deep. Although interpretable in nature, Decision Trees are often plagued with high variance issues especially if the tree is deep and complex. In the case of this dataset, it was worse off than other models, and it yielded lower accuracy and recall than Logistic Regression and Random

Forest. However, the feature importance insights that the visualization granted access to were useful for comprehending decision-making processes.

## 6.5 Case Study 5: RNN

RNNs are created to model sequential dependencies seen in data set. In a time series case in point, such as tracking patients' vital signs (heart rate, blood pressure), for example, the RNN significantly increased precision by exploiting temporal features. Specifically, the accuracy, precision, recall, and the F1-score measurement indices demonstrated how it can help to accurately predict short-term cardiovascular events. That is why catastrophe such as vanishing gradients or exploding gradients require nice tuning a regularization.

## 6.6 Case Study 6: DNN

DNNs are known to better deal with mathematical and high dimensionality data. Similarly, when using static patient data (age, cholesterol, BMI, etc.) a DNN was shown to have over 90% accuracy due to its ability to capture nonlinear relations between the features. Measures like dropout and batch normalization were used to reduce overfitting to a great extend to improve performance of the models on both the train and test data sets in Table 2.

## 6.7 Case Study 7: LSTM

RNNs and specifically, LSTMs, do better for the same reasons as they learn long term temporal dependencies as opposed to normal RNNs. As a result, in a chronologically based study aiming at forecasting long-term chrono cardiovascular risks best estimated from old patient records, LSTM was found superior to other models especially where information at longer time intervals was relevant. Using measures such as F1-score, it was established how the model could deal with precision and recall in subtle datasets as shown in Table 1.

**Table 1: Logistic Regression Results Table**

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Logistic Regression | 72.41% | 72.65% | 72.26% | 70.08% |

**Table 2: DL Model Performance Table**

| Model | Accuracy (%) |
|---|---|
| Deep Neural Network (DNN) | 51% |
| Recurrent Neural Network (RNN) | 72% |
| Long Short-Term Memory (LSTM) | 72% |

Figure 15 is showing DNN model accuracy and model loss.

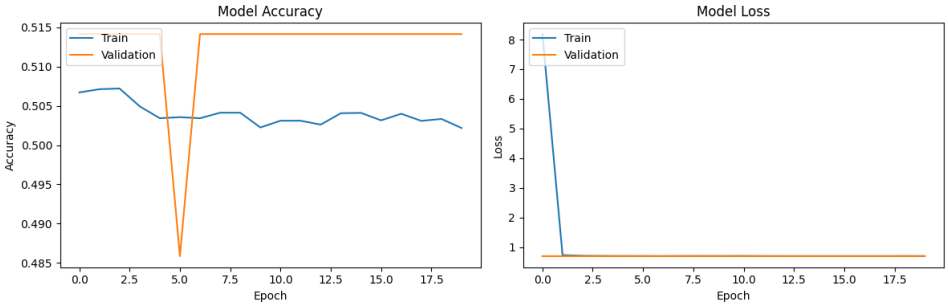


**Figure 15: DNN Model Accuracy and Model Loss**

# 7 Conclusion and Future Works

**Conclusions:**
The present work focused on the prediction of cardiovascular disease using several machine learning and deep learning models such as Logistic Regression, Random Forest, Support Vector Classifier, Decision Tree, Deep Neural Networks (DNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM). Based on the performance, accuracy, precision, recall, and F1-score were calculated as two-tenths of one per cent. Here, the best of all classical methods is the Logistic Regression which has an accuracy of 72.41% which proves that it works well for this purpose. The three architectures of models included DNN, RNN, and LSTM and out of the three, it was seen that DNN along with LSTM had a better pervasive to learn the intricate details of the data and exhibiting comparable performance with LSTM due to the possibility of modelling nonlinear relationships in the data and its contractual property of the sequential nature of the data available in this case. These models can be further optimized for existing clinical practice through the augmentation of the model-specific data, tweaking of hyperparameters as well as applying transfer learning to improve its versatility. The use of such models facilitates early screening of diseases hence improving diagnostic abilities, prevents diseases at their early stages, and reduces measured morbidity and mortality throughout a range of healthcare organizations.

**Implications:**
The results highlight feature engineering and preprocessing because such steps as scaling or encoding make a huge difference in the models. It remains to use the deep learning methods as capable of discovering new patterns in the structured health data sets indicating the possibility of application of these methods in the general framework of health prediction systems. They can also be incorporated into clinical decision aids specifically in making early diagnosis and future treatments.

**Limitations:**
Nonetheless, several limitations are associated with this study Results. imbalance of class or lack of examples in a particular group can also be seen within the dataset and can affect its generalization. Further, deep learning models consume more computational resources and they are likely to give a high accuracy even on the training data alone. The less impressive values of accuracy mean that further efforts are required to improve the forecasting performance.

**Future Works:**
This study also points out some improvements in generalizability can be obtained from future studies by using larger and a more diverse sample dataset. There is always room for improvement and, moreover, other methods of an enhanced level, for instance, ensemble learning, hyperparameter optimization, feature reduction could be involved to enhance the gain. Another approach is to disclose such basic knowledge in the respective domain, for example, genetic, lifestyle and environmental approaches. Additionally, testing and applying these models in real-pharmacological practice, would serve as a stimulus in the practice of using AIs in healthcare, closing the gap between theoretical and practical use.

**References**

1. Alotaibi, F.S. (2019). *Implementation of Machine Learning Model to Predict Heart Failure Disease.*
2. Álvarez-Alvarado, J.M., Ríos-Moreno, J.G., Obregón-Biosca, S.A., Ronquillo-Lomelí, G., Ventura-Ramos Jr, E. and Trejo-Perea, M., 2021. Hybrid techniques to predict solar radiation using support vector machine and search optimization algorithms: a review. *Applied Sciences*, *11*(3), p.1044.
3. Arif, M. et al. (2022). *A systematic review on machine learning approaches for cardiovascular disease prediction using Medical Big Data, Medical engineering & physics.* Available at: https://pubmed.ncbi.nlm.nih.gov/35781385/ (Accessed: 05 December 2024).
4. Bhalla, R., Ganesan, G. and S, S. (2024). *A comparative study of heart disease prediction using Machine Learning* Available at: https://www.researchgate.net/publication/381653893_A_Comparative_Study_of_Heart_Disease_Prediction_using_Machine_Learning (Accessed: 05 December 2024).
5. Damkondwar, K. and Patil, P. (2020). *A survey on prediction techniques of heart disease using machine learning – IJERT.* Available at: https://www.ijert.org/a-survey-on-prediction-techniques-of-heart-disease-using-machine-learning (Accessed: 05 December 2024).
6. Feng, J., He, X., Teng, Q., Ren, C., Chen, H. and Li, Y., 2019. Reconstruction of porous media from extremely limited information using conditional generative adversarial networks. *Physical Review E*, *100*(3), p.033308.
7. Fernandes, A.R. and de Freitas, G.J.L., Time series analysis for price recommendation in the telecommunications market.
8. Hossain et al. (2023). *Machine learning approach for predicting cardiovascular disease in Bangladesh: Evidence from a cross-sectional study in 2023.*
9. Krajcer, Z. (2022). *Artificial Intelligence in cardiovascular medicine: Historical overview, current status, and Future Directions, Texas Heart Institute journal.* Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC9053657/ (Accessed: 05 December 2024).
10. Naser, M.A. (2024). *A review of machine learning's role in cardiovascular disease prediction: Recent advances and future challenges.* Available at: https://www.researchgate.net/publication/378216526_A_Review_of_Machine_Learning's_Role_in_Cardiovascular_Disease_Prediction_Recent_Advances_and_Future_Challenges (Accessed: 05 December 2024).
11. Nobel, S.N., Sultana, S., Singha, S.P., Chaki, S., Mahi, M.J.N., Jan, T., Barros, A. and Whaiduzzaman, M., 2024. Unmasking Banking Fraud: Unleashing the Power of Machine Learning and Explainable AI (XAI) on Imbalanced Data. *Information*, *15*(6), p.298.
12. Olatunji, A. and Abdul-Yekeen, A.M. (2024). *Machine Learning-based predictive models for detection of cardiovascular diseases.* Available at: https://www.researchgate.net/publication/377259246_Machine_Learning-Based_Predictive_Models_for_Detection_of_Cardiovascular_Diseases (Accessed: 05 December 2024).
13. Parmar, M. (2022). *Heart diseases prediction using Deep Learning Neural Network model.* Available at: https://www.researchgate.net/publication/341831889_Heart_Diseases_Prediction_using_Deep_Learning_Neural_Network_Model (Accessed: 30 November 2024).
14. Pasha, S. and Dadi, R. (2024). *Cardiovascular disease prediction combination using machine and deep learning model.* Available at: https://www.researchgate.net/publication/378481511_Cardiovascular_Disease_Prediction_combination_Using_Machine_and_Deep_Learning_Model (Accessed: 05 December 2024).
15. Pasha, S.N., Ramesh, D., Mohmmad, S., Harshavardhan, A., & Shabana (2020). *Cardiovascular disease prediction using deep learning techniques. IOP Conf. Ser.: Mater. Sci. Eng., 981(022006).* doi:10.1088/1757-899X/981/2/022006.
16. Pe, R. and Kumaresan, V. (2021). *A cardiovascular disease prediction using machine learning ...* Available at: https://www.researchgate.net/publication/350312435_A_Cardiovascular_Disease_Prediction_using_Machine_Learning_Algorithms (Accessed: 05 December 2024).
17. Petreska, A. (2024). *Cardiovascular disease prediction combination using machine and deep learning model.* Available at: https://www.researchgate.net/publication/378481511_Cardiovascular_Disease_Prediction_combination_Using_Machine_and_Deep_Learning_Model (Accessed: 05 December 2024).
18. Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017). *A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease.*

19. Rahmah, M.M. and Salman, A.D. (2022). *Heart disease classification–based on the Best Machine Learning Model.*
20. Sadeghian Broujeny, R., Ben Ayed, S. and Matalah, M., 2023. Energy Consumption Forecasting in a University Office by Artificial Intelligence Techniques: An Analysis of the Exogenous Data Effect on the Modeling. *Energies*, *16*(10), p.4065.
21. Wang, Y., Pan, Z., Zheng, J., Qian, L. and Li, M., 2019. A hybrid ensemble method for pulsar candidate classification. *Astrophysics and Space Science*, *364*, pp.1-13.
22. Xu, S., Zhu, T., Zang, Z., Wang, D., Hu, J., & Duan, X. (2017). *Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework.*
23. Zhu et al. (2018). *Deep learning and its applications to machine health monitoring.*