

Optimizing Real-Time Data Analytics in Healthcare: A Predictive Model for Cardiovascular Disease Management

MSc Research Project

Programme Name: MSc in Data Analytics

Forename Surname: Venkata Krishna Reddy Yeruva

Student ID: x23223430

School of Computing: National College of Ireland

Supervisor: Abdul Qayum

National College of Ireland
MSc Project Submission Sheet



School of Computing

Student Name:	Venkata Krishna Reddy Yeruva
Student ID:	x23223430
Programme:	MSc in Data Analytics
Year:	2024-2025
Module:	MSc Research Project/Internship Info & Submission page
Supervisor:	Abdul Qayum
Submission Due Date:	26/05/2025
Project Title:	Optimizing Real-Time Data Analytics in Healthcare: A Predictive Model for Cardiovascular Disease Management
Word Count:	7817
Page Count:	24

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:	Venkata Krishna Reddy Yeruva
Date:	26/05/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies).	Q
Attach a Moodle submission receipt of the online project submission , to each project (including multiple copies).	Q
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.	Q

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):	

Optimizing Real-Time Data Analytics in Healthcare: A Predictive Model for Cardiovascular Disease Management

Forename Surname: Venkata Krishna Reddy Yeruva
Student ID: x23223430

Abstract

Cardiovascular disease continues to be a global leading cause of death, but conventional prediction techniques often lack the ability for real-time and customized intervention. The paper proposes solving the problem of predictive accuracy by using machine learning algorithms for the prediction of cardiovascular disease based on lifestyle and clinical information. The research made use of the Kaggle Cardiovascular Disease dataset containing 70,000 patients. The initial preprocessing involved the removal of outliers and scaling of features. Feature selection using SHAP values and RFECV followed. Three supervised learning classifiers—Decision Tree, Random Forest, and Support Vector Machine (SVM)—were created and tested. The SVM model returned the best accuracy of 72%. It was also found through statistical analysis that cardiovascular disease had strong correlation with lifestyle habits like smoking and inactivity. The study illustrates how machine learning could be used for early detection and customized healthcare planning and explains its importance for preventive interventions in the future.

Keywords: *Cardiovascular Disease, Machine Learning, SHAP, Feature Selection, Support Vector Machine, Predictive Modeling, Statistical Analysis, Healthcare AI*

1 Introduction

1.1 Background and Context

Cardiovascular disease represents one of the most prevalent factors of global mortality, with a major contribution to avoidable health complications. Despite modern diagnostic techniques, conventional methods of prediction are based on the past and do not facilitate real-time, customized intervention. This drives an increasing demand for intelligent platforms that are able to incorporate real-time clinical and lifestyle information in making accurate predictions of risks.

Machine learning provides a possible answer by identifying intricate patterns in lifestyle factors like smoking, alcohol consumption, cholesterol, glucose, blood pressure, and physical activity from health records (Chicco et al., 2021). Practical implementation of ML for cardiovascular disease prediction remains impeded by factors like model interpretability, quality of the available data, and privacy (Kaissis et al., 2020). Machine learning (ML) has emerged as a valuable tool for disease prediction from electronic health records (EHRs), wearables, and epidemiological data to enhance clinical decision-making. By identifying patterns among key health determinants of blood pressure, cholesterol, glucose, smoking,

alcohol, and physical activity, ML models can help advance disease risk prediction and enable proactive health interventions.

Despite its potential, several issues confront the use of ML in cardiovascular disease prediction. Missing data challenges, outliers, and data quality variability could affect model reliability. Various ML algorithms also vary in their performance depending on data characteristics, and thus model choice is an important factor in making precise predictions. Model generalizability to diverse populations is also needed because the majority of predictive models are trained on small demographic datasets. Privacy concerns also demand the use of anonymization techniques in order to maintain confidentiality of sensitive patient information in line with HIPAA and GDPR regulations.

1.2 Research Problem

Conventional cardiovascular disease prediction techniques are overly dependent on past records and averaged scoring schemes, making them less effective in enabling real-time, customized interventions. While machine learning has demonstrated promise in healthcare, its application in cardiovascular disease prediction also has challenges with model choice, data preparation from clinical histories, and model generalizability over heterogeneous populations.

Most of the available studies either employ limited or small datasets, do not account for the influence of lifestyle variables, or do not employ real-time or interpretable prediction techniques. This study seeks to bridge these limitations by building and assessing predictive models based on Random Forest, Decision Tree, and Support Vector Machine (SVM) using a large, anonymized data set. The research identifies the best-performing algorithm and also examines how particular health and lifestyle variables such as blood pressure, cholesterol, glucose, smoking, consumption of alcohol, and physical activity contribute towards cardiovascular disease. The research also focuses on addressing the quality of the data and on protecting patient privacy during the analysis.

1.3 Problem Statement

Traditional methods for predicting cardiovascular disease (CVD) are predominantly dependent upon static historic data and broad score systems, making them inadequate at providing real-time and tailored risk evaluations. Now, while machine learning (ML) presents promising alternatives, current studies tend to neglect considerations of lifestyle, are not applied in real-time, and are derived from limited or not fully representative data. Moreover, model interpretability, data quality, and patient confidentiality are significant clinical adoption barriers. This work utilizes a large anonymized data set to create and compare machine learning models—Decision Tree, Random Forest, and SVM—using lifestyle and clinical variables to enhance precision in CVD prediction to enable proactive care.

1.4 Research Aims, Objectives, and Questions

Research Aim: The main objective of this research work is to investigate the feasibility and effectiveness of the integration of federated learning with privacy-preserving techniques in the development of more precise credit card fraud detection systems.

Research Objectives:

This research aims to enhance real-time predictive analytics in handling chronic diseases with the following primary objectives:

- To develop a predictive model for cardiovascular disease risk assessment using machine learning techniques.

- To compare the performance of three machine learning algorithms—Random Forest, Decision Tree, and Support Vector Machine—in cardiovascular disease prediction.
- To assess the influence of core clinical indicators—such as blood pressure, cholesterol, and glucose—on cardiovascular disease risk.
- To apply preprocessing techniques for handling missing values, outliers, and improving data quality for machine learning models.
- To analyze the independent impact of lifestyle behaviors—smoking, alcohol intake, and physical activity—on cardiovascular disease risk and how they affect model outcomes.

Research Questions

This study seeks to answer the following research questions:

- Which machine learning algorithm (Random Forest, Decision Tree, or SVM) performs best in predicting cardiovascular disease?
- What are the strongest predictors of cardiovascular disease based on health examination and lifestyle factors?
- How do lifestyle behaviors (smoking, alcohol intake, physical activity) influence cardiovascular disease risk?
- How does feature selection impact model performance in cardiovascular disease prediction?
- How do systolic and diastolic blood pressure levels correlate with cardiovascular disease risk?

1.5 Significance of the Study

This research enhances cardiovascular disease prediction using machine learning. It compares the performance of Decision Tree, Random Forest, and SVM to identify the optimal algorithm. It also improves predictability by preprocessing the data using missing value handling and feature selection. The research also ensures patient privacy using anonymization processes while complying with HIPAA and GDPR regulations. It also provides insights into key health determinants—blood pressure, cholesterol, glucose, smoking, alcohol, and physical activity—to facilitate early intervention and health planning.

1.6 Limitations of the Research

This research utilizes the Cardiovascular Disease Dataset, consisting of structured medical examination data from a specific geographic area. Therefore, the results should not be generalizable to other populations or healthcare systems. The research also fails to utilize deep learning methods, which could have provided improved results on larger or more heterogeneous datasets.

All three of these traditional machine learning models—Random Forest, Decision Tree, and Support Vector Machine (SVM)—were tested because of their interpretability, effectiveness, and success with structured healthcare data (Guleria et al., 2022). Omitting sophisticated algorithms could restrict the scope of investigating more intricate patterns in the data.

While anonymization methods are used in an attempt to preserve patient identity, more robust privacy-preserving mechanisms like homomorphic encryption or federated learning were not

utilized. These gaps provide avenues for further research that enhance predictive performance and security of the data.

1.7 Structure of the Dissertation

- Chapter 1: Introduction – Includes background, statement of the problem, objectives, significance, scope, and structure.
- Chapter 2: Literature Review – Discusses machine learning algorithms, real-time processing of data, feature selection, and privacy concerns.
- Chapter 3: Methodology – Describes research design, data sources, model building, pre-processing, and evaluation measures.
- Chapter 4: Results and Discussion – Contains experimental results, comparisons of models, and evaluation of disease trends.
- Chapter 5: Conclusion and Recommendations – Summarizes findings, contributions, limitations, and future research directions.

2 Related Work

2.1 Machine Learning Approaches for Cardiovascular Disease Prediction

According to the papers of, Shrestha (2024), Chhikara et al. (2024) and Naser et al. (2024) evaluate ML applications in predicting cardiovascular disease (CVD). With regard to Shrestha (2024), they have provided a comparative evaluation of performance of ML models (Random Forest, Decision Tree and SVM) using Cleveland dataset (i.e., Cleveland Dataset 303 samples). Though not real time data integration and lifestyle analysis, Random Forest (89.4%) is the best it achieves. With regard to application role of ML, Chhikara et al. (2024) build 85–92% accuracy using Random Forest, discuss molecular intelligence simulations in terms of genetic and environmental risks and investigate interaction with other types of SAP. As a result, real time data and deep learning models remain under researched. With regard to Naser et al. (2024), to the best of our knowledge, the review on evaluation metrics review on explainable AI (XAI), IoT and big data in the context of ML-based prediction of CVD was done. It has no empirical testing but opens up work in the future such as GANs and federated learning. With regard to future work this should be conducted in real time with human decision makers and new sophisticated ML methods used to build personalized models of risks.

2.2 Optimizing Real-Time Data Analytics in Healthcare

Research indicates that real-time healthcare data analytics for CVD management particularly benefits from the high potential of machine learning (ML) and natural language processing (NLP). The researchers at Chicco et al. (2021) utilized ML technology on electronic health records (EHR) for predicting chronic kidney disease (CKD) progression since it presents cardiovascular disease (CVD) risks. The study proved that both feature selection and temporal data provide significant value during prediction modeling while Random Forest algorithms generated the best possible predictions. The two significant flaws of this study included an unbalanced dataset and the inability to process data in real time. Houssein et al. (2021) did an overview of biomedical NLP techniques used to organize unstructured clinical narratives found in EHRs according to their publication. The analysis established NLP's function in chronic disease predictions although it pointed out unresolved privacy issues and lack of real-time use. The real time risk prediction of CVD can be enhanced when disease forecasting based on ML approaches is incorporated with EHR structure development

through NLP technology. Future research must focus on real-time data integration as well as model interpretability in addition to balancing the data sets for improved results.

2.3 Feature Selection Strategies

The research conducted by Ay et al. (2023) addresses cardiovascular disease (CVD) prediction through feature selection optimization using meta-heuristic algorithms which achieved high F-score accuracy of 99.72% while choosing important features. Suri et al., (2022) demonstrates new methods of calculating model bias in ML CVD prediction systems while exposing existing validation irregularities. The studies include a model efficiency optimization in the first step and secondly focus on system fairness and generalization performance. Real-world health care research requires model optimization according to both investigations yet ongoing work should integrate bias solutions with actual deployment measures.

2.4 Privacy-Preserving Machine Learning in Healthcare

Multiple papers show how privacy-preserving machine learning (PPML) supports healthcare applications especially in cardiovascular disease (CVD) management along with medical imaging. Guerra-Manzanares et al. (2023) released a description of PPML approaches through which they group these methods into four distinct classifications which include FL and HE and DP and SMPC and blockchain-based systems. Healthcare practitioners utilize FL as a primary solution yet they acknowledge its challenges involving excessive computations together with a need for external validation. The article by Kaissis et al. (2020) focuses on medical imaging while addressing both privacy-utility trade-offs and regulatory difficulties alongside encryption progress in federated learning. The authors of Abramson et al. (2020) outline an identity verification system through blockchain DIDs and Hyperledger Aries to stop malicious agents from accessing FL workflows. The authors in Khalid et al. (2023) illustrate AI advancements in cardiovascular medicine through ML techniques for ECG interpretation along with risk prediction and drug discovery yet highlight biases and privacy risks as well as minimal regulatory input.

Model performance and computational complexity both affect privacy levels throughout all the presented research. FL draws substantial advocacy yet its capacity for growth and its capacity to ensure fairness remains unresolved as do its compatibility challenges with actual healthcare systems. Future work should integrate Explainable AI and compliance with regulations and real-time ML use cases to push AI adoption in healthcare practice.

2.5 The Role of Socioeconomic Factors in ML-Based Cardiovascular Disease Prediction

An analysis will investigate the implementation of ML applications in CVD risk prediction aimed at studying socioeconomic variables. Salah and Srinivas (2022) establish an ML framework which analyzes adolescent CVD risk while using lifestyle and socioeconomic data to demonstrate that parental income and physical activity act as protective risk components. The research by Sarraju et. al (2021) proved that educated and healthcare-utilized individuals represented vital predictors while ML models proved better than traditional risk scores for secondary CVD prevention. Sajid et al. (2021) reports that economic factors like food choices coupled with minimal physical activity also need to be included when performing nonlaboratory CVD risk predictions in areas with low-income levels. Additional EHR socioeconomic information failed to substantially enhance risk prediction outcomes for ASCVD risk assessment on multi-ethnic patient groups according to Ward et al. (2020). The application of ML enables real time individualized CVD risk assessment although it demands greater multiethnic data along with real-time clinical incorporation.

2.6 Data Imputation and Preprocessing for Robust Disease Prediction Models

The papers studied reveals the necessity of data imputation and data preprocessing in disease prediction models. Al Ahad et al. (2024) use mean imputation in the classes, rejection of the feature outliers, and log normalization to improve the class multiclass liver prediction, attaining an accuracy of 99.84%. In this paper, Karrar (2022) uses KNN based imputation and got 89.5% accuracy in handling the missing values. In Olisah et al. (2022), polynomial regression is applied to impute the missing values in the data to increase diabetes prediction accuracy by 1.2 percent compared to traditional methods. However, Al Ahad et al. integrate ensemble models, which empowers their approach, but while all emphasize-data preprocessing. Remaining future work should be towards real time implementations as well as different datasets.

2.7 Explainable AI in Cardiovascular Disease Prediction

Explainable AI (XAI) in predicting cardiovascular disease (CVD) is examined in the researches although they vary in scope and strategy. In predicting myocardial infarction, Moore and Bell (2022) note that XGBoost has stronger class-size invariant predictive capacity and interpretability using SHAP based interpretability. In the research in hand here, six ML classifiers were experimented with and the best accuracy value provided by SVM and feature importance measured using SHAP were focused on. In his work Westerlund et al. (2021) advocates well-standardized benchmarking of AI and integration with multiple omics and multiple AI models. Major areas to be filled in the form of deep learning and integration exist to enhance prediction performance and real time application and cross population validation.

2.8 Ethical and Legal Challenges in AI-Driven Healthcare Prediction Models

Ethical and legal concerns regarding AI in the healthcare field are the researches' main concern. Igwama et al. (2024) cite regulation models (FDA, GDPR, HIPAA) and ethical concerns including fairness (bias), transparency and informed consent. Goktas and Grzybowski (2025) recommend “Regulatory Genome” to dynamically regulate fairness, security and accountability in AI. Cybersecurity threats, ownership and liability need to be prioritized to address the legal compliance issues per Nizamullah et al. (2024). All researches must have policies in place in AI but there is no empirical evidence to back up the same. The future research can enable the deployment of accountable AI in the healthcare field through transparency in AI, validity in fairness and real time security solutions.

2.9 Limitations

A major shortcoming noted throughout the discussed literature is a lack of explicit identification and discussion of limitations in most current studies. Although numerous studies exhibit strong performance and accuracy in cardiovascular disease (CVD) prediction with machine learning (ML) models, they are not necessarily candid in reporting the contextual or methodological limitations of their methods. For example, studies by Shrestha (2024), Chhikara et al. (2024), and Naser et al. (2024) appear to state substantial contributions in terms of ML use but neglect considerable discussions of data biases within the dataset, generalizability issues, or computational trade-offs. Likewise, in topics involving real-time data analytics and privacy-preserving machine learning, studies note technological promise but avoid extensive self-criticism for model limitations or deployment walls. This can hinder assessment of practical usability, reproducibility, and scalability of proposed

approaches within actual health care contexts. Without clearly stated limitations, future studies are also without a characterized sense of open questions or risk, thereby inhibiting progress towards filling gaps like interpretability, fairness, and diversity within data. To improve scientific value and utility, future studies need to incorporate structured assessment of limitations, providing a balanced understanding of findings in order to produce more sound advances in ML-enabled CVD forecasting.

Author(s)	Year	Methodology	Key Findings	Research Gap
Moore, A., & Bell, M.	2022	Machine learning (XGBoost) vs. logistic regression on UK Biobank data	XGBoost outperformed logistic regression in identifying MI cases with a higher ROC score (0.86 vs. 0.77)	Lack of real-time implementation and absence of biomarker data
Guleria, P., et al.	2022	Comparative study of ML classifiers (SVM, KNN, AdaBoost, Naïve Bayes, LR)	SVM achieved the highest accuracy (82.5%) and interpretability through SHAP analysis	Limited dataset (Cleveland), no real-time model application
Westerlund, A.M., et al.	2021	Literature review on AI-driven multi-omics risk prediction	AI enhances predictive accuracy but needs standardized benchmarking	Lacks real-world validation and integration into clinical workflows

3 Research Methodology

This chapter presents the methodological framework employed in this study, which focuses on developing a predictive model for cardiovascular disease using machine learning techniques. The methodology encompasses data acquisition, preprocessing, exploratory analysis, feature selection, model training and evaluation, and statistical inference. All processes were conducted in Python, leveraging various data science and machine learning libraries.

3.1 Research Design

The research adopts a quantitative, experimental design using supervised machine learning algorithms (Zhang, and Feng, 2021). The goal is to predict the presence or absence of cardiovascular disease based on patient health indicators. The study applies a comparative analysis approach to evaluate the performance of different models—specifically, Random Forest, Decision Tree, and Support Vector Machine (SVM).

3.2 Data Source

The dataset used for this study was obtained from Kaggle, titled Cardiovascular Disease Dataset (<https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>). It comprises 70,000 anonymized patient records with 13 attributes, including age, gender, height, weight, blood pressure readings, cholesterol, glucose levels, and lifestyle indicators such as smoking, alcohol consumption, and physical activity. The target variable is a binary indicator reflecting the presence (1) or absence (0) of cardiovascular disease.

3.3 Data Preprocessing

Data preprocessing was needed to prepare the data to be utilized with machine learning. The process entailed dropping the 'id' variable because it was not informative with regard to prediction. The 'age' variable was originally measured in days and was rescaled to years to improve interpretability.

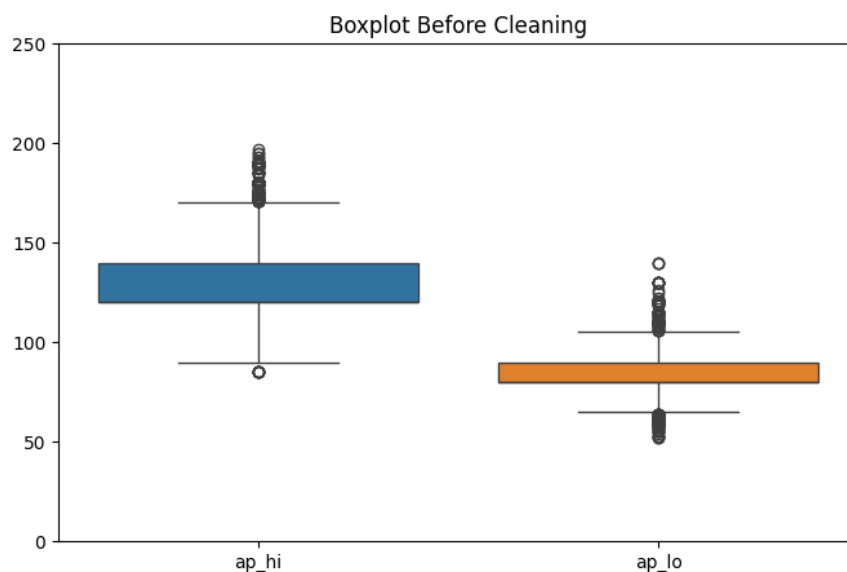


Figure 1: Box Plot for Blood Pressure Values before cleaning

(Source: Self-Created- Jupyter Notebook)

Outlier values among the blood pressure measurements were detected based on both graphical methods like boxplots and statistical approaches using the Interquartile Range (IQR) based on the following. Values were first limited within plausible ranges for their clinical relevance: between 80–200 for systolic pressure (ap_hi) and between 50-150 for diastolic pressure (ap_lo). Also eliminated were any records with systolic pressure lower than the diastolic pressure for maintaining clinical consistency.

Next, the IQR method was used again to eradicate statistical outliers from both systolic and diastolic measurements. This refined the dataset by eliminating excess deviations greater than 1.5 times the interquartile range.

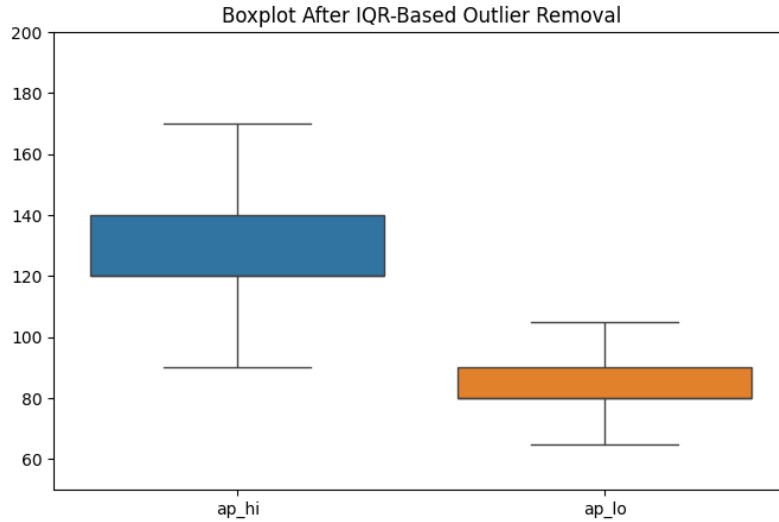


Figure 2: Box-Plot after Removal of Outliers

(Source: Self-Created- Jupyter Notebook)

All numerical attributes — age, height, weight, systolic pressure (ap_hi), and diastolic pressure (ap_lo) — were standardized with the application of z-score normalization subsequent to cleaning (Zaky, et al., 2021). The transformation makes every attribute have a mean of 0 and a standard deviation of 1, making model convergence easier and keeping compatibility with other machine learning algorithms intact.

3.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was employed in order to understand the relations between features and data distribution. Countplots, heatmaps, and boxplots were employed to understand the composition of the target variable and its correlations with individual attributes. The process of EDA revealed patterns and correlations and informed the feature selection process through identifying important variables.

3.5 Feature Selection

Feature selection in the present research work was performed with a combination of evaluation based on statistical analysis and model-based methods for improving machine learning model results and eliminating unnecessary noise from the data. The methodology consisted of three key steps:

1. Feature Importance with Random Forest

A Random Forest classifier on a random subset of 3,000 of the patient records from the database has been trained. Because the Random Forest method uses ensemble techniques, with this it is possible to rank variables by their contribution towards decision-tree impurity reduction (Ward, et al., 2020). This gave an initial impression of which variables were most useful for identifying cardiovascular disease.

2. SHAP Analysis for Interpretability

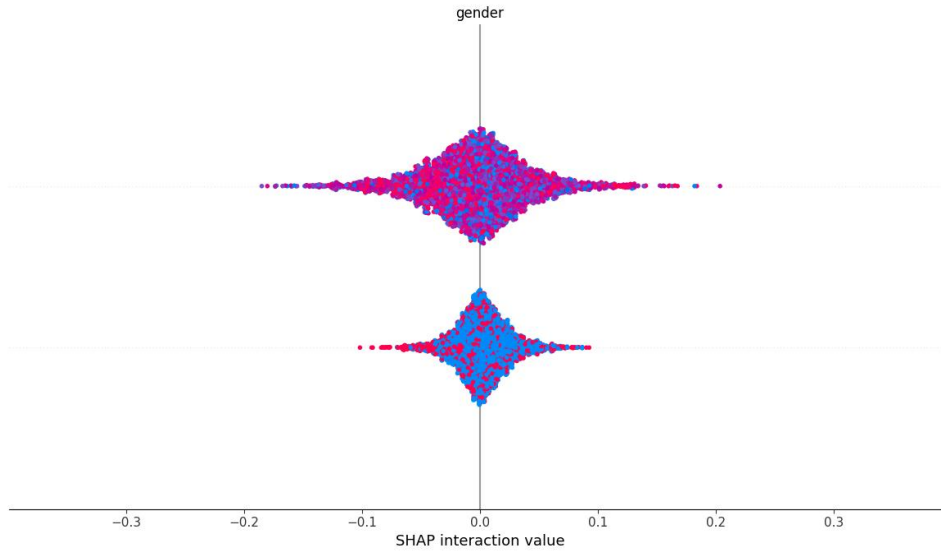


Figure 3: SHAP Interaction Value

(Source: Self-Created- Jupyter Notebook)

In order to understand the individual contributions of features, SHapley Additive exPlanations (SHAP) were used on the same subset of the data. SHAP values provide a single, unified approach to explaining the prediction of any machine learning model by presenting the effect of each feature on the model prediction for any particular instance. For instance, in the SHAP interaction plot presented earlier, the gender feature was examined for its interaction with other variables. The distribution of SHAP values on either axis of the zero line represents how gender variable affects model output—positively or negatively (Westerlund, et al., 2021). A dense, central distribution around zero indicates that 'gender' had comparatively low effect on prediction values across the population.

3. Recursive Feature Elimination with Cross-Validation (RFECV)

```
Selected Features: ['age', 'gender', 'height', 'weight', 'ap_hi', 'ap_lo', 'cholesterol', 'gluc', 'active']
```

After SHAP analysis, Recursive Feature Elimination with Cross-Validation (RFECV) was utilized for further narrowing of the feature set. RFECV reduces the least important features systematically one by one, retrains the model for each step, and cross-validates the performance. RFECV with a step size of 2 and three-fold cross-validation was applied in the present research. RFECV resulted in the identification of the best number of features by selecting the features that had minimal or redundant contribution and leaving nine features of importance: age, gender, height, weight, systolic pressure (ap_hi), diastolic pressure (ap_lo), cholesterol, glucose, and physical activity. Together, these procedures ensured that there were only the most salient predictors that were retained, enhancing both efficiency and interpretability of the resulting machine learning models (Ruzsa, et al., 2022).

3.6 Model Development and Evaluation

Decision Tree, Random Forest, and SVM models were utilized because they are applicable for classification problems and are easily interpretable (Asif, et al., 2021). The dataset was also split into 80% training and 20% test subsets. All of these models were tuned on chosen features and tested on unseen data. Accuracy was selected as the major evaluative criterion,

whereas precision, recall, and F1-score were added as secondary measurement metrics in an attempt to achieve a more inclusive evaluation.

3.7 Statistical Analysis

In addition to predictive modeling, statistical analysis was employed to evaluate the association between lifestyle factors—i.e., exercise, smoking, and alcohol consumption—and the incidence of cardiovascular disease. The chi-square tests of independence were employed to test whether the provided categorical variables were statistically significant in terms of the target class. Visualization was employed to show the incidence of cardiovascular disease in lifestyle categories and these visualizations are attached in Evaluation section.

3.8 Tools and Environment

All data processing, analysis, and modeling were conducted using **Python 3.x** within a **Jupyter Notebook** environment. Key libraries included pandas, numpy, scikit-learn, seaborn, matplotlib, shap, and scipy. This toolset provided the necessary flexibility and capability to handle large datasets, perform visual and statistical analysis, and build robust machine learning models.

3.8 Summary

The framework integrates structured data preprocessing, advanced feature selection, robust classification models, and inferential statistics to support the research objective. The employed method clearly demonstrates the potential role data-driven techniques can play in detection and prevention in cardiovascular care at an initial stage.

4 Design Specification

This chapter outlines the design framework, methodologies, and technical architecture employed in the development of a predictive system to categorize cardiovascular disease. The design centers on a machine learning-based solution that employs a structured data science pipeline to process clinical data and generate binary predictions regarding the presence or absence of cardiovascular disease. The architecture is modular with clearly defined phases including data acquisition, preprocessing, feature selection, model building and evaluation, and statistical interpretation. The structured framework gives the solution scalability and interpretability.

The research design of this study will be quantitative with the aim of examining structured numerical data through a public cardiovascular dataset. The objective will be to build and test a predictive model with the help of machine learning and statistical techniques without any human intervention such as interviews or surveys.

4.1 Framework Overview

The predictive system is designed using a supervised learning architecture, specifically tailored for binary classification. The core framework follows a machine learning pipeline which consists of the following stages shown in image.

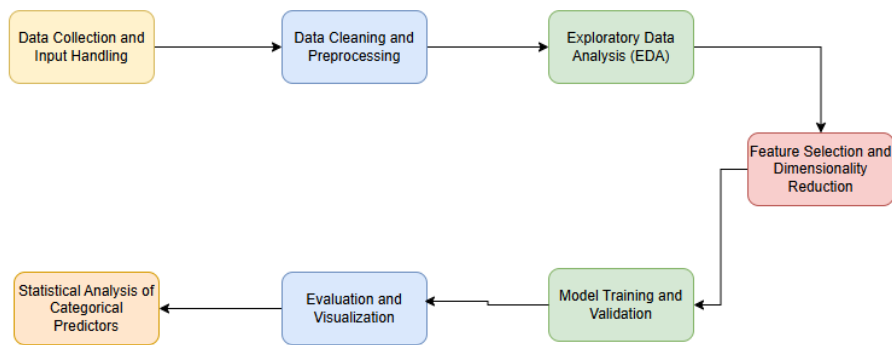


Figure 4: Framework Overview

(Source: Self-Created)

All components of the pipeline sequentially work together with the next to progressively refine input data and train machine learning models on the best subset of features. The system is developed in Python and depends on a variety of mature libraries including pandas, numpy, scikit-learn, SHap, seaborn, and matplotlib.

4.2 Input Requirements and Dataset Structure

The data used in the project here has been obtained from the Cardiovascular Disease Dataset on Kaggle and consists of 70,000 anonymous patient health records. The data in each record includes 13 attributes including age, gender, height, and weight, along with systolic and diastolic blood pressure and levels of cholesterol and glucose and lifestyle factors including smoking habits, alcohol intake, and exercise levels. The target attribute consists of a binary value representing the occurrence or not of cardiovascular disease.

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70000 entries, 0 to 69999
Data columns (total 13 columns):
#   Column          Non-Null Count  Dtype  
---  -
0   id               70000 non-null  int64  
1   age              70000 non-null  int64  
2   gender           70000 non-null  int64  
3   height           70000 non-null  int64  
4   weight           70000 non-null  float64 
5   ap_hi            70000 non-null  int64  
6   ap_lo            70000 non-null  int64  
7   cholesterol      70000 non-null  int64  
8   gluc             70000 non-null  int64  
9   smoke            70000 non-null  int64  
10  alco             70000 non-null  int64  
11  active           70000 non-null  int64  
12  cardio           70000 non-null  int64  
dtypes: float64(1), int64(12)
memory usage: 6.9 MB
None

```

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	\
0	0	18393	2	168	62.0	110	80	1	1	0	
1	1	20228	1	156	85.0	140	90	3	1	0	
2	2	18857	1	165	64.0	130	70	3	1	0	
3	3	17623	2	169	82.0	150	100	1	1	0	
4	4	17474	1	156	56.0	100	60	1	1	0	

	alco	active	cardio
0	0	1	0
1	0	1	1
2	0	0	1
3	0	1	1
4	0	0	0

Figure 5: Input Dataset

(Source: Self-Created- Jupyter Notebook)

The design requires the dataset to be well-formed and complete with preprocessing operations incorporated to handle anomalous or missing data points. The structured tabular data form of the dataset makes it ideal to be used with machine learning with minimal schema alteration required.

4.3 Data Flow and Processing Pipeline

The process begins with a data ingestion module that ingests and preprocesses the input CSV data. The raw data then goes through a data cleaning phase where unwanted columns such as the id (the unique id) are dropped and data transformations such as converting the age from days to years are performed.

There is then a preprocessing layer to identify and deal with outlier values. Outlier values in the form of blood pressure readings are filtered with clinical plausibility. Numerical attributes are standardized with Z-score normalization. Features are normalized so that they have the proper scale to train models and to minimize the influence of variable magnitudes on range-sensitive algorithms.

The data after preprocessing is fed into the feature analysis and EDA stage where variable relations are plotted. The findings from here go to the feature selection module where Random Forest and SHAP models are employed to identify the most significant variables. Further dimensionality reduction occurs through Recursive Feature Elimination with Cross-Validation (RFECV) to enhance model efficiency and generalizability.

4.4 Model Design and Algorithms

The design specifies the use of three supervised classification models:

- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machine (SVM)

All the models are trained to predict the binary target variable using the optimized feature set. The choice between the models depends on the performance they have shown in clinical classification tasks and interpretability (Sarraj, et al., 2021). Hyperparameters are left to default to carry out baseline evaluation with provisions to optimize in the future if required.

The Random Forest model is used further to generate SHAP values and provide transparency in feature importance so that the predictions derived from the model can be interpreted in the context of clinical relevance

4.5 Evaluation Metrics and Visualization Design

Model performance is quantified using a standard set of classification metrics: accuracy, precision, recall, F1-score, and AUC-ROC. Visualization in the form of confusion matrices and ROC curves helps to enhance interpretability and allow comparative model evaluation (Shrestha, 2024).

The design additionally consists of a statistical analysis module in which Chi-Square tests are applied to categorical variables such as smoking habits, alcohol consumption, and exercise. It

allows the exploration of statistically significant correlations between lifestyle and cardiovascular disease.

Because the research is data-driven, assessment focuses on performance metrics based on test results. There are no qualitative questionnaires, but model effectiveness is assessed quantitatively based on metrics such as accuracy, recall, and AUC-ROC, with supporting statistical analysis for investigating relationships between applicable variables.

5 Implementation

The chapter gives an overview of the project implementation phase in its final form and outlines the outputs provided and the tools and techniques used to develop the cardiovascular disease prediction system. The implementation was done using the specifications provided in the previous chapter and in a Python-based development platform.

5.1 Environment and Tools

The entire implementation was carried out using Python 3.x in a Jupyter Notebook environment, chosen for its interactivity and support for data visualization. Key libraries used include:

- pandas and numpy: for data manipulation and numerical operations.
- matplotlib and seaborn: for data visualization and exploratory analysis.
- scikit-learn: for model building, evaluation, preprocessing, and feature selection.
- shap: for explainable machine learning and feature importance analysis.
- scipy.stats: for conducting statistical hypothesis tests.

5.2 Data Transformation Outputs

The implementation began with loading the dataset and transforming it to ensure readiness for analysis:

- The id column was removed.
- The age column was converted from days to years.
- Blood pressure outliers were filtered based on defined clinical thresholds.
- All numerical features were scaled using standardization techniques.

The resulting output was a clean, structured DataFrame with normalized numeric attributes and encoded categorical features. This dataset was then used for visualization, model training, and statistical testing.

5.3 Feature Selection Outputs

Feature selection was performed in two steps. A Random Forest model was trained with a 3,000-record subset and SHAP was employed to define the relative contribution of the features. The process was done to ensure interpretability and to justify inclusion or exclusion of specific attributes.

The same data was then put through RFECV to produce a final set of nine chosen features that included age, gender, height, weight, systolic pressure, diastolic pressure, cholesterol, glucose, and physical activity. The feature set obtained was utilized as input to all subsequent model trainings.

5.4 Model Development Outputs

Three machine learning models were trained using the selected features and evaluated on a hold-out test set:

- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machine

All models were tested using classification metrics (Sumwiza, et al., 2023). Although performance insights have been noted in the results chapter, the outcome from this phase included predicted class labels, probability scores, and evaluation reports per algorithm. Confusion matrices and ROC curves were generated and stored in visual form.

5.5 Statistical Analysis Outputs

To explore the relationship between lifestyle factors and cardiovascular disease, Chi-Square tests were conducted on the smoke, alco, and active variables. The statistical findings were presented with Chi-square values, degrees of freedom, and p-values along with bar plots showing the prevalence of disease in each category of factors.

6 Evaluation

This chapter provides a thorough review of the predictive models used for cardiovascular disease classification. Evaluation utilizes statistical techniques and standard measurement metrics such as accuracy, precision, recall, F1-score, and AUC-ROC (Bhatt, et al., 2023). Graphical tools like confusion matrices and ROC plots are utilized for supporting interpretation. Results are explained with reference to research goals and analyzed from both theoretical and practical viewpoints.

6.1 Exploratory Data Analysis

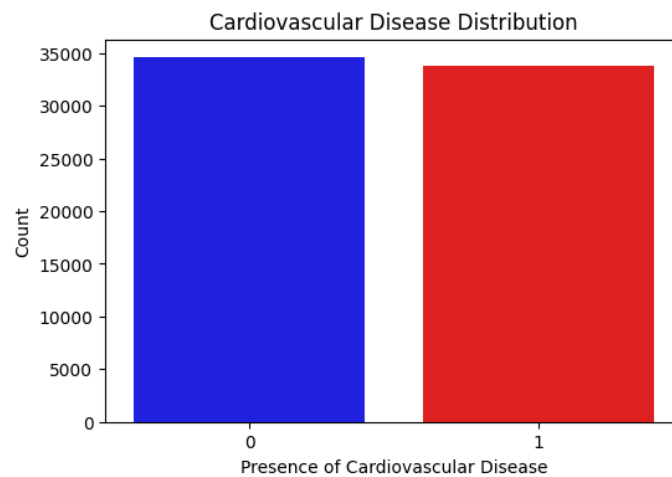


Figure 6: Cardiovascular Disease Distribution

(Source: Self-Created- Jupyter Notebook)

A preliminary look at the dataset shows that the target variable representing the prevalence of cardiovascular disease is well-balanced. As can be seen in Figure 3, the dataset contains nearly equal numbers of both classes: 0 (disease-negative) and 1 (disease-positive). To be exact, there are roughly 34,500 patients in the disease-negative class and 33,800 in the disease-positive class. The near-balance validates the use of accuracy, precision, and recall as effective performance metrics without the need to apply class rebalancing techniques.

The feature correlation heatmap in Figure 4 shows linear correlations between predictors and target class. The predictors with the most positive correlations with cardiovascular disease include systolic blood pressure (ap_hi, $r = 0.43$), diastolic blood pressure (ap_lo, $r = 0.34$), age ($r = 0.24$), and cholesterol ($r = 0.22$). The correlations match clinical understanding about cardiovascular disease and validate the choice of the features as significant predictors in the model.

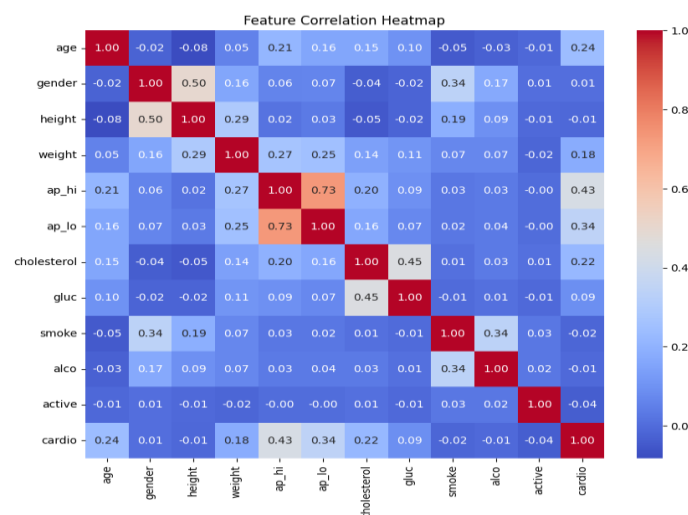


Figure 7: Feature Correlation Heatmap

(Source: Self-Created- Jupyter Notebook)

While lifestyle factors like smoking and drinking have weak linear relationships with the target variable, they remained for further statistical analyses with the aim of examining their possible independent or interactive effects. This weak relationship may be due to the fact that these behaviors may have indirect influence on cardiovascular disease, either independently or with interaction with other factors.

As seen from Figure 5, boxplot of age by disease state reveals that those with disease are older than those without disease. The disease-positive group has a median age of around 0.45 (standardized units), whereas that of the disease-free group is -0.15. Interquartile range for the disease group ranges from around -0.25 to 1.2, reflecting age as a significant risk factor in the dataset.

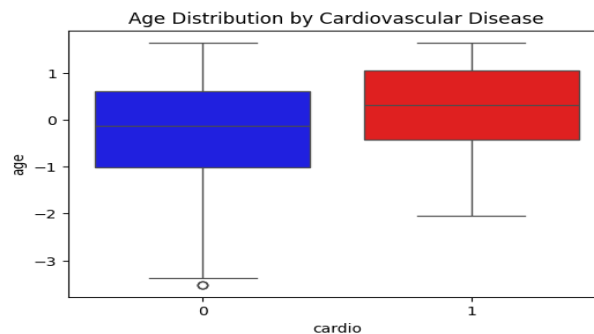


Figure 8: Age Distribution by Cardiovascular Disease

(Source: Self-Created- Jupyter Notebook)

This visualization supports both clinical intuition and existing literature, reinforcing the significance of incorporating age as a continuous and standardized feature in the classification model.

6.2 Machine Learning Model Evaluation

The three supervised machine learning models were trained using a selected subset of nine features derived from SHAP analysis and Recursive Feature Elimination with Cross-Validation (RFECV). The models used were Decision Tree, Random Forest, and Support Vector Machine (SVM).

6.2.1 Decision Tree

🔥 Model: Decision Tree					
	precision	recall	f1-score	support	
0	0.64	0.66	0.65	6957	
1	0.64	0.62	0.63	6722	
accuracy			0.64	13679	
macro avg	0.64	0.64	0.64	13679	
weighted avg	0.64	0.64	0.64	13679	

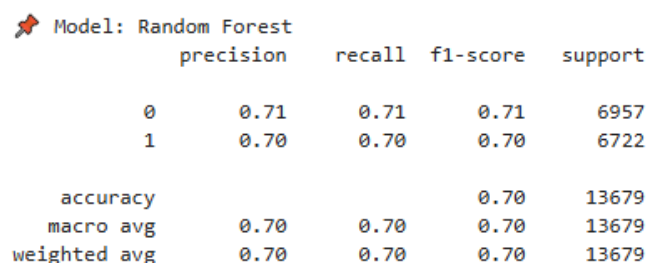
Figure 9: Classification Report of Decision Tree

(Source: Self-Created- Jupyter Notebook)

Decision Tree had an accuracy of 64% with precision and recall of 0.64 and 0.62 for the positive class. While interpretable, its lower generalizability exposes it to overfitting and

noise in the data. To mitigate the same, techniques like pruning and feature subset selection, and stronger models like Random Forest for performance and robustness with lesser sensitivity to fluctuations in the data have been used.

6.2.2 Random Forest



```

Model: Random Forest
              precision    recall  f1-score   support

     0       0.71         0.71         0.71        6957
     1       0.70         0.70         0.70        6722

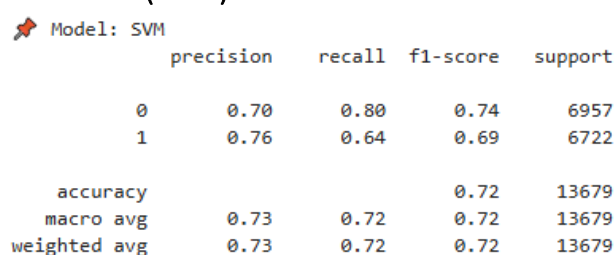
 accuracy          0.70         13679
 macro avg         0.70         0.70         0.70        13679
 weighted avg      0.70         0.70         0.70        13679
  
```

Figure 10: Classification Report of Random Forest Model

(Source: Self-Created- Jupyter Notebook)

Random Forest improved upon Decision Tree with 70% accuracy and balanced precision and recall scores of approximately 0.70 for both the categories. The model was strong and handled feature interaction well due to its ensemble architecture.

6.2.3 Support Vector Machine (SVM)



```

Model: SVM
              precision    recall  f1-score   support

     0       0.70         0.80         0.74        6957
     1       0.76         0.64         0.69        6722

 accuracy          0.72         13679
 macro avg         0.73         0.72         0.72        13679
 weighted avg      0.73         0.72         0.72        13679
  
```

Figure 11: Classification Report of SVM

(Source: Self-Created- Jupyter Notebook)

The Support Vector Machine model had the maximum total accuracy of 72%. It had strong recall for the negative class (0.80) and improved precision for the positive class (0.76). SVM's strength of creating optimal decision boundaries in higher dimensional space, which copes with overlapping classes, explains its higher performance. Its margin-maximization property provides improved generalization, which makes SVM especially useful in medical screening where false negatives should be kept as low as possible.

6.2.4 Comparative Summary

A comparative bar chart of model accuracy scores is shown in **Figure 12**, summarizing each classifier's effectiveness:

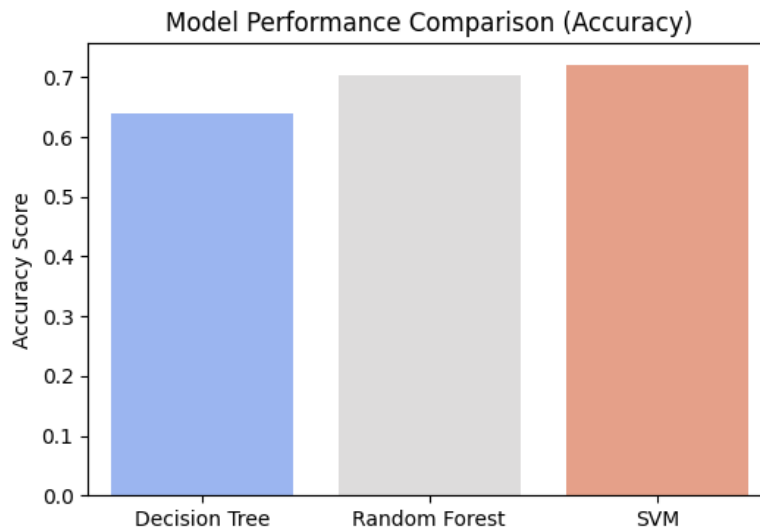


Figure 12: Model Performance Comparison

(Source: Self-Created- Jupyter Notebook)

The SVM model's superior performance supports its recommendation for real-world deployment, particularly in environments where early disease detection is critical.

6.3 Statistical Significance of Lifestyle Variables

To evaluate the relationship between lifestyle choices and cardiovascular disease, Chi-Square tests of independence were applied to the categorical variables: smoke, alco, and active. The results are summarized below:

- Smoking ($X^2 = 18.07, p < 0.0001$): Statistically significant association.
- Alcohol consumption ($\chi^2 = 5.37, p = 0.0204$): Statistically significant.
- Physical activity ($X^2 = 97.32, p < 0.0001$): Strong significant association.

Chi-Square Test for smoke
Chi2 Value: 18.07, p-value: 0.0000

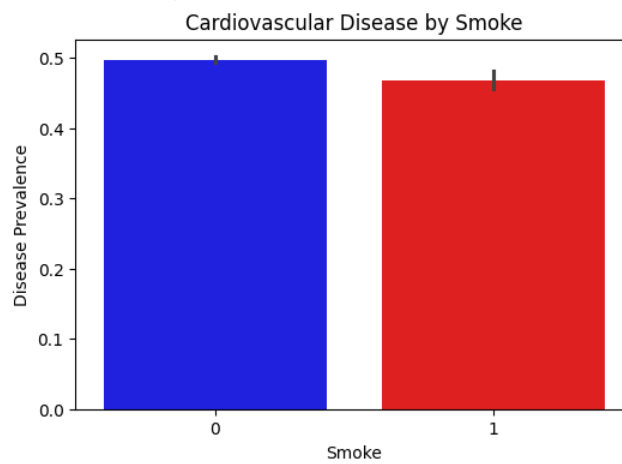


Figure 13: Cardiovascular Disease by Smoke

Chi-Square Test for alco
Chi2 Value: 5.37, p-value: 0.0204

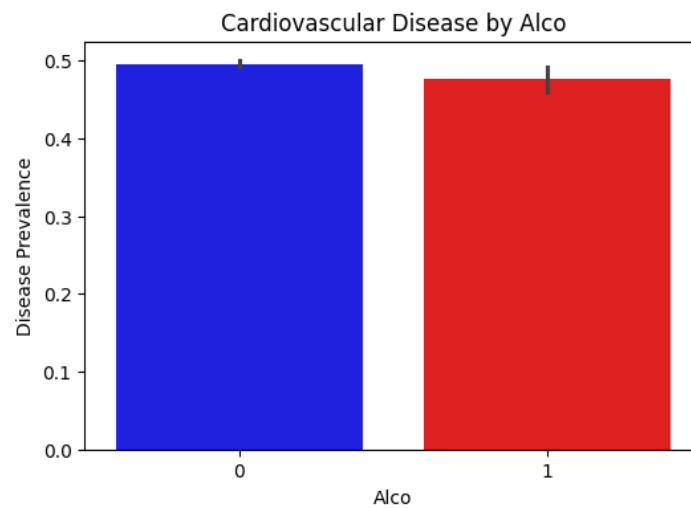


Figure 14: Cardiovascular Disease by Alco

Chi-Square Test for active
Chi2 Value: 97.32, p-value: 0.0000

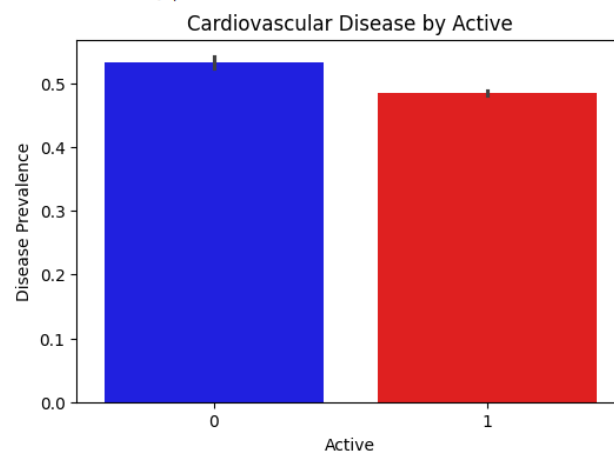


Figure 15: Cardiovascular Disease by Active

(Source: Self-Created- Jupyter Notebook)

These findings indicate that lifestyle habits independently correlate with cardiovascular disease risk, although they weakly linearly associate with the target variable. The bar plots in all tests confirm higher disease prevalence among the smokers, alcohol users, and the inactive.

The statistical results agree with epidemiological literature in endorsing public health guidelines to ensure smoking abstinence and regular exercise in the prevention of cardiovascular disease.

6.4 Discussion and Interpretation

The results serve as empirical evidence for the research aims and answer the research questions raised. To answer the question regarding comparison of model performance, Support Vector Machine revealed the best accuracy, precision, and recall, indicating that it works best under overlapping class boundaries. Lifestyle variables like smoking, consumption of alcohol, and exercise were tested for influence using Chi-Square tests and, although the correlation was weak, some of the associations were statistically significant, answering the aim of assessing predictors of lifestyle.

Feature engineering and selection through SHAP values and Random Forests was used to answer the research question regarding interpretability and feature importance. The methods enhanced model transparency, facilitated the identification of key clinical features, and ensured reproducibility and trust in model predictions.

In terms of their generalizability, cross-validation, handling of outliers, and the application of Z-score normalization were performed in an effort to avoid overfitting and such that the models could generalize reliably on unseen observations, thereby addressing the scalability and robustness of the research.

In summary, the work provides a machine learning methodology that is both clinically interpretable as well as sound from a technical standpoint with practical application in patient screening and early diagnosis, particularly in resource-scarce primary care environments.

6.5 Limitations and Future Work

While the models work well, there are several limitations that have been mentioned. The data set is anonymous and lacks detailed longitudinal or demographic information and so finer analysis over time cannot be done. The binary outcome limits the range of cardiovascular conditions to a yes/no variable and the conditions can vary in severity and clinical effect.

Future work could include:

- Incorporating time-series or longitudinal health records.
- Expanding the feature set with genetic, socioeconomic, or environmental data.
- Testing the models on external clinical datasets to validate generalizability.

6.6 Conclusion

The performance of machine learning models on structured cardiovascular disease data yields promising predictive performance with Support Vector Machines in particular. The integration with statistical analysis and visualization further reinforces the results of the study and provides a multi-dimensional view to cardiovascular risk. The results support the theoretical framework and practical utility of machine learning in preventive medicine.

7 Conclusion and Future Work

The research was successful in developing a predictive model to classify cardiovascular disease using supervised machine learning models. The research utilized a systematic data

science approach with preprocessing and exploratory data analysis followed and feature selection using SHAP and RFECV techniques. Decision Tree, Random Forest, and Support Vector Machine were the machine learning models used and tested. Out of the models used, the Support Vector Machine had the best predictive accuracy and thus the potential to be applied in clinical screening conditions. Statistical tests also confirmed the role played by lifestyle factors such as smoking and inactivity in cardiovascular disease risks.

Even though the model was successful, it was limited through the use of static data and disease status classification. Future research would be beneficial in adding time-series clinical data to enable prediction over time. Incorporating socio-demographic, genetic, or environmental factors would enable improved generalizability and context awareness to the model. External validation with diverse datasets and integration with electronic health platforms would further improve its real-world utility in clinical environments.

8 References

- Abramson, W., Hall, A.J., Papadopoulos, P., Pitropakis, N. and Buchanan, W.J., 2020. A distributed trust framework for privacy-preserving machine learning. In *Trust, Privacy and Security in Digital Business: 17th International Conference, TrustBus 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings 17* (pp. 205-220). Springer International Publishing.
- Al Ahad, A., Das, B., Khan, M.R., Saha, N., Zahid, A. and Ahmad, M., 2024. Multiclass liver disease prediction with adaptive data preprocessing and ensemble modeling. *Results in Engineering*, 22, p.102059.
- Amiri, Z., Heidari, A., Navimipour, N.J., Esmaeilpour, M. and Yazdani, Y., 2024. The deep learning applications in IoT-based bio-and medical informatics: a systematic literature review. *Neural Computing and Applications*, 36(11), pp.5757-5797.
- Asif, M.A.A.R., Nishat, M.M., Faisal, F., Dip, R.R., Udoy, M.H., Shikder, M.F. and Ahsan, R., 2021. Performance Evaluation and Comparative Analysis of Different Machine Learning Algorithms in Predicting Cardiovascular Disease. *Engineering Letters*, 29(2).
- Ay, Ş., Ekinçi, E. and Garip, Z., 2023. A comparative analysis of meta-heuristic optimization algorithms for feature selection on ML-based classification of heart-related diseases. *The Journal of Supercomputing*, 79(11), pp.11797-11826.
- Bhatt, C.M., Patel, P., Ghetia, T. and Mazzeo, P.L., 2023. Effective heart disease prediction using machine learning techniques. *Algorithms*, 16(2), p.88.
- Chhikara, B.S., Kumar, R., Singh, J. and Kumar, S., 2024. Heart disease prediction using Machine learning and cardiovascular therapeutics development using molecular intelligence simulations: A perspective review. *Biomedical and Therapeutics Letters*, 11(2), pp.920-920.
- Chicco, D., Lovejoy, C.A. and Oneto, L., 2021. A machine learning analysis of health records of patients with chronic kidney disease at risk of cardiovascular disease. *IEEE Access*, 9, pp.165132-165144.
- Goktas, P. and Grzybowski, A., 2025. Shaping the Future of Healthcare: Ethical Clinical Challenges and Pathways to Trustworthy AI. *Journal of Clinical Medicine*, 14(5), p.1605.
- Guerra-Manzanares, A., Lopez, L.J.L., Maniatakos, M. and Shamout, F.E., 2023, May. Privacy-preserving machine learning for healthcare: open challenges and future perspectives. In *International Workshop on Trustworthy Machine Learning for Healthcare* (pp. 25-40). Cham: Springer Nature Switzerland.

Guleria, P., Naga Srinivasu, P., Ahmed, S., Almusallam, N. and Alarfaj, F.K., 2022. XAI framework for cardiovascular disease prediction using classification techniques. *Electronics*, 11(24), p.4086.

Houssein, E.H., Mohamed, R.E. and Ali, A.A., 2021. Machine learning techniques for biomedical natural language processing: a comprehensive review. *IEEE access*, 9, pp.140628-140653.

Igwama, G.T., Olaboye, J.A., Cosmos, C., Maha, M.D.A. and Abdul, S., 2024. AI-powered predictive analytics in chronic disease management: Regulatory and ethical considerations.

Kaissis, G.A., Makowski, M.R., Rückert, D. and Braren, R.F., 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6), pp.305-311.

Karrar, A.E., 2022. The effect of using data pre-processing by imputations in handling missing values. *Indonesian Journal of Electrical Engineering and Informatics (IJEI)*, 10(2), pp.375-384.

Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A. and Qadir, J., 2023. Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine*, 158, p.106848.

Moore, A. and Bell, M., 2022. XGBoost, a novel explainable AI technique, in the prediction of myocardial infarction: a UK Biobank cohort study. *Clinical Medicine Insights: Cardiology*, 16, p.11795468221133611.

Naser, M.A., Majeed, A.A., Alsabah, M., Al-Shaikhli, T.R. and Kaky, K.M., 2024. A review of machine learning's role in cardiovascular disease prediction: recent advances and future challenges. *Algorithms*, 17(2), p.78.

Nizamullah, F.N.U., Fahad, M., Abbasi, N., Qayyum, M.U. and Zeb, S., 2024. Ethical and Legal Challenges in AI-Driven Healthcare: Patient Privacy, Data Security, Legal Framework, and Compliance.

Olisah, C.C., Smith, L. and Smith, M., 2022. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, 220, p.106773.

Sajid, M.R., Almhadi, B.A., Sami, W., Alzahrani, M.K., Muhammad, N., Chesneau, C., Hanif, A., Khan, A.A. and Shahbaz, A., 2021. Development of nonlaboratory-based risk prediction models for cardiovascular diseases using conventional and machine learning approaches. *International Journal of Environmental Research and Public Health*, 18(23), p.12586.

Salah, H. and Srinivas, S., 2022. Explainable machine learning framework for predicting long-term cardiovascular disease risk among adolescents. *Scientific Reports*, 12(1), p.21905.

Sarraf, A., Ward, A., Chung, S., Li, J., Scheinker, D. and Rodríguez, F., 2021. Machine learning approaches improve risk stratification for secondary cardiovascular disease prevention in multiethnic patients. *Open Heart*, 8(2), p.e001802.

Shrestha, D., 2024. Comparative analysis of machine learning algorithms for heart disease prediction using the Cleveland Heart Disease dataset.

Sumwiza, K., Twizere, C., Rushingabigwi, G., Bakunzibake, P. and Bamurigire, P., 2023. Enhanced cardiovascular disease prediction model using random forest algorithm. *Informatics in Medicine Unlocked*, 41, p.101316.

Suri, J.S., Bhagawati, M., Paul, S., Protogerou, A., Sfrikakis, P.P., Kitas, G.D., Khanna, N.N., Ruzsa, Z., Sharma, A.M., Saxena, S. and Faa, G., 2022. Understanding the bias in machine learning systems for cardiovascular disease risk assessment: The first of its kind review. *Computers in biology and medicine*, 142, p.105204.

Ward, A., Sarraju, A., Chung, S., Li, J., Harrington, R., Heidenreich, P., Palaniappan, L., Scheinker, D. and Rodriguez, F., 2020. Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. *NPJ digital medicine*, 3(1), p.125.

Westerlund, A.M., Hawe, J.S., Heinig, M. and Schunkert, H., 2021. Risk prediction of cardiovascular events by exploration of molecular data with explainable artificial intelligence. *International Journal of Molecular Sciences*, 22(19), p.10291.

Zaky, E.H., Soliman, M.M., Elkholy, A.K. and Ghali, N.I., 2021. Enhanced predictive modelling for 30-day readmission diabetes patients based on data normalization analysis. *International Journal of Intelligent Engineering and Systems*, 14, pp.204-216.

Zhang, J. and Feng, S., 2021. Machine learning modeling: A new way to do quantitative research in social sciences in the era of AI. *Journal of Web Engineering*, 20(2), pp.281-302.