Cardiovascular Disease Prediction with Advanced Machine Learning Techniques: A Comprehensive Analysis

Rifath Ahmed

BSc Computer Science

22 April 2024

Supervised by

Razia Sulthana Abdul Kareem

Senior Lecturer in Computer Science

Second Supervisor

Ik Soo Lim

Senior Lecturer in Computer Science

Abstract

Cardiovascular diseases (CVD) continue to be the primary cause of death around the world, and it is necessary to develop more sophisticated methods for the early identification and efficient management of these conditions. Using substantial datasets from Hungary, Switzerland, and Cleveland, this study investigates the application of machine learning (ML) approaches to increase the accuracy of cardiovascular disease (CVD) prediction and diagnostics. The paper utilised two complex machine learning models, 1D-CNN and TabNet, to handle class imbalances. These models were optimised by intensive preprocessing, which included data augmentation with CTGAN. Its technique's training, validation, and testing phases were highly rigorous. During these phases, the paper carefully compared the performance of different models based on accuracy, precision, recall, and F1 scores. The findings proved that both models can accurately predict cardiovascular events with high reliability over a wide range of demographic variables. Specifically, TabNet generally provided slightly better results across most metrics, particularly in precision and F1 scores, suggesting its effectiveness in reducing false positives and balancing precision and recall. For instance, TabNet achieved an F1 score of 0.83 on the Cleveland dataset and 0.88 on the Hungarian dataset, slightly outperforming the 1D-CNN. This study highlights the potential of machine learning to revolutionise cardiovascular disease (CVD) treatment by providing substantial insights into the optimisation of models and their use in clinical settings to improve patient care. The study's results propose using machine learning to enhance preventative cardiovascular healthcare, which could potentially lead to a reduction in the worldwide illness burden.

Acknowledgements

I extend my heartfelt gratitude to my supervisors for their invaluable guidance and support throughout this journey. I am immensely thankful to Dr. Razia Sultana Abdul Karim for her enduring patience, insightful critiques, and encouraging approach, greatly enhancing my research capabilities. Similarly, my appreciation goes to Dr. Ik Soo Lim for his expert advice and continuous inspiration, which were crucial in refining my work.

I am also grateful to my fellow students and the entire academic staff at the Faculty of Engineering and Science. Their collaborative spirit, thoughtful suggestions, and scholarly environment have been instrumental in personal and educational growth.

Thank you all for your contributions, without which this endeavour would not have been possible.

TABLE OF CONTENTS

List of Tables

List of Figures

1. Introduction

1.1 Motivation

Cardiovascular diseases (CVD) represent the principal cause of mortality worldwide, accounting for approximately one-third of all deaths globally (Anon., 2021). This stark statistic highlights the profound impact of these diseases on public health and underscores the critical necessity for enhanced methods in their early detection and management. The motivation for this study is the persistent challenges that hamper the efficacy of current diagnostic and therapeutic strategies, particularly in the context of the evolving global health landscape.

Despite significant medical advancements and increased public awareness, the burden of CVD remains unacceptably high. This is complicated by several factors, including the ageing population, the prevalence of lifestyle-related risk factors, and the complexities associated with managing chronic health conditions. Current diagnostic tools and predictive models often fail to detect cardiovascular issues early, primarily due to their inadequate sensitivity and specificity, especially in heterogeneous populations. (Morgenstern, 2020).

The integration of diverse datasets presents a formidable challenge, as data variability across different demographics and geographic regions can significantly affect the performance of predictive models. Additionally, there is a notable deficiency in models capable of effectively handling data imbalances (Raniya R. Sarra, 2022)—a common issue in medical datasets where healthy cases significantly outnumber instances of disease. This imbalance can skew the predictive accuracy, leading to high rates of false negatives or false positives, which are clinically undesirable and lead to inefficient use of healthcare resources. (Rahman & Davis, 2013).

This research aims to bridge these gaps by harnessing the power of advanced machine learning techniques to develop an accurate, equitable, and adaptable predictive framework for diverse healthcare settings. By ensuring the accuracy and reliability of CVD predictions, this study seeks to facilitate a shift towards more personalised and preventive cardiovascular care, ultimately reducing the global burden of these deadly diseases.

1.2 Problem Statement

Cardiovascular disease prediction has historically been challenged by the heterogeneity of patient data, which reflects diverse ethnicities, age groups, and socio-economic backgrounds. Each group may exhibit unique disease manifestation patterns, complicating the task of developing universally effective predictive models. This study seeks to create machine learning models capable of assimilating and analyzing such multifaceted data to offer accurate, timely predictions of cardiovascular risks. The goal is to leverage advanced algorithms that can handle high-dimensional, disparate data sources to unearth patterns that traditional analytical techniques might overlook.

Several factors, including the data's quantity and quality, directly impact the effectiveness of machine learning (ML) models. The creation of precise and reliable models is made possible by the availability of high-quality big datasets. On the other hand, the acquisition of such datasets is frequently hampered by practical and ethical reasons. These considerations include patient permission, privacy rules, and the logistical obstacles of aggregating massive amounts of health database information. In addition, the quality of the data, which includes its accuracy, completeness, and relevance, continues to be a significant problem that has implications for the dependability of the predictions that machine learning models generate.

Beyond mere prediction, the reliability and precision of these models in clinical applications are paramount. In real-world healthcare settings, the consequences of false predictions can be severe, ranging from unnecessary treatments (based on false positives) to overlooked conditions (due to false negatives). This research aims to refine these machine learning models, optimising them to provide dependable results that healthcare professionals can trust. Efforts will be focused on validating and testing the models across various populations and conditions to ensure high accuracy and reliability, regardless of the demographic or geographic variables involved. This includes precise evaluation protocols that test these models against established benchmarks and simulated real-world scenarios to ensure their robustness and readiness for deployment in diverse clinical environments.

1.3 Objectives

Cardiovascular diseases (CVD) continue to pose a significant public health challenge across the globe. As the leading cause of mortality, these diseases are responsible for nearly one-third of all deaths worldwide. This alarming statistic underscores the critical need for effective strategies that can significantly improve early detection and management of the condition. In response to this urgent demand, the present research initiative seeks to harness the sophisticated capabilities of machine learning (ML) technologies. Machine learning offers transformative potential for the medical field, particularly in enhancing diagnostic and predictive methodologies. By integrating advanced ML techniques, this study aims to:

1. Investigate the applicability of various machine learning models, including deep learning architectures like 1D-CNN and TabNet, for predicting cardiovascular diseases using diverse and augmented CTGAN datasets from multiple demographics (Cleveland, Hungarian and Switzerland).

2. Compare the effectiveness of these models in terms of accuracy, precision, recall, and F1-score to determine the most efficient in clinical settings.

3. Examine the influence of different hyperparameters, such as data splits, optimisers, and input shapes, on model performance to optimise prediction capabilities.

This research employs a systematic approach to achieve the objectives, starting with collecting extensive datasets from three key geographical locations (Andras Janosi, 1988): Hungary, Switzerland, and Cleveland. These datasets include vital patient metrics such as age, gender, blood pressure, cholesterol levels, and other relevant cardiovascular health indicators. Following the data augmentation via CTGAN to address class imbalance and small sample sizes, the study applies 1D-CNN and TabNet models, critically evaluating their performance across various configurations. The models will be fine-tuned using different train-test splits, optimisers, and batch sizes to identify optimal settings that maximise predictive accuracy and reliability. The anticipated outcome is a comprehensive understanding of how different machine learning models perform across various demographic datasets and configurations, leading to potential recommendations for their implementation in clinical practice. These models are designed to predict the occurrence of such events and assess their potential severity, which can vary widely among patients. The application of these refined models holds promise for revolutionising cardiovascular care. With more precise predictions, healthcare providers can adopt a more proactive approach to treatment and intervention. This proactive strategy could lead to earlier interventions, tailored treatment plans, and improved clinical outcomes for patients at risk of cardiovascular diseases.

2. Literature Review:

2.1 Introduction

The emergence of machine learning (ML) and artificial intelligence (AI) is revolutionising the fight against cardiovascular disease (CVD), which is the leading cause of death in the world. This happens when we are on the verge of entering a new age in the healthcare field. With the help of this literature review, it will shed light on the revolutionary impact that machine learning has had on predicting, comprehending, and treating cardiovascular illnesses. The paper investigates the journey from the foundational understanding of cardiovascular disease (CVD), including its cause of disease and epidemiology, to the cutting-edge developments in machine learning (ML) that offer a leap forward in precision medicine for heart-related illnesses.

The review systematically reveals the fundamental nature of cardiovascular illness, shedding light on the numerous manifestations that it might take and the burden it places on the entire world. While we are moving away from the conventional clinical understanding of cardiovascular disease and into the area of computational healthcare, this review will investigate the role that machine learning plays in going beyond the standard diagnostic and predictive methods. This showcases the potential of machine learning to transform this approach to evaluating the risk of cardiovascular disease (CVD), diagnosing it at an early stage, and creating customised treatment options. Machine learning has the ability to process complex and multidimensional data, making it a powerful tool in this field.

This review, which summarises the current state of research, not only highlights the accomplishments made to this point but also draws attention to problems that need to be addressed and the potential for further investigation.

2.2 Understanding of Heart Disease

Various illnesses that affect the heart and blood vessels are referred to as cardiovascular diseases. The most common cause of cardiovascular disease is atherosclerosis, which is a condition that happens when fat and cholesterol accumulate in the walls of blood vessels (arteries). Plaque is the name given to this accumulation, which contributes to the narrowing of arteries and may result in cardiovascular events such as heart attacks or strokes. The most common form of coronary heart disease (CHD) is characterised by the accumulation of plaque in the arteries that supply blood and oxygen to the heart. This condition restricts the flow of blood and oxygen to the heart, which can lead to a heart attack, heart failure, or arrhythmias. Heart failure is a condition that can be caused by coronary heart disease (CHD) or excessive blood pressure, and it can affect one or both sides of the heart. Heart failure is characterised by the heart becoming too weak or stiff to pump blood efficiently. Arrhythmias, also known as irregular heartbeats, are caused by disturbances in the heart's electrical system. On the other hand, heart valve illnesses can cause disruptions in blood flow due to malfunctioning valves. In contrast to high blood pressure, which raises the risk of heart failure, stroke, and heart attacks, peripheral artery disease limits the amount of blood that flows to the legs and feet, which in turn causes damage to the tissues in those areas. Stroke, caused by blood arteries in the brain becoming clogged or bleeding, is associated with many of the same risk factors as heart disease. In addition, congenital heart disease, which is present from birth, is the most frequent type of birth abnormality. A variety of problems in the structure and function of the heart characterise it. (Metkus, et al., 2022).

The prevalence of cardiovascular disease is a significant public health concern, and individuals all over the world are affected by this condition. In 2019, cardiovascular diseases were responsible for the deaths of around 17.9 million individuals, which accounts for 32 per cent of all deaths worldwide. Heart attacks and strokes were responsible for 85 per cent of these fatalities. More than three-quarters of deaths caused by cardiovascular diseases occur in nations with poor and moderate incomes. In 2019, cardiovascular disorders were responsible for 38 per cent of the 17 million premature deaths that occurred among people under the age of 70 owing to non-communicable diseases. (Anon., 2021). Modifiable risk variables contributed considerably to worldwide CVD mortality in 2021, emphasising the need for focused public health initiatives. Dietary considerations affect heart health, as high LDL cholesterol kills 3.8 million people. High fasting plasma glucose caused 2.3 million deaths, making diabetes management harder. Air pollution killed 4.8 million people, while high BMI, a measure of obesity, killed 2.0 million, illustrating the global obesity epidemic. Tobacco use caused 3.0 million CVD deaths. Exercise is essential since 397,000 deaths were linked to inactivity. High blood pressure caused 10.8 million deaths, emphasising the necessity for hypertension therapy and control. These figures underscore the need for comprehensive measures to address modifiable risk factors to minimise CVD fatalities worldwide (Mariachiara Di Cesare, 2023).

Doctors start with personal and family medical histories, symptoms, blood testing, and electrocardiogram. Based on these first findings, more testing may be needed. Blood tests for cardiovascular health and risk variables like cholesterol, CRP, and homocysteine are essential for heart disease risk assessment. These tests can also evaluate kidney, liver, electrolyte, and thyroid health, which affect the heart. ECG, echocardiography, stress tests (EKG or nuclear), carotid and abdominal ultrasounds, Holter monitoring, and event recording are non-invasive examinations. These tests don't require instrument insertion and can reveal the heart's electrical activity, anatomy, function, and artery health. Heart diseases are diagnosed and treated via invasive testing like cardiac catheterisation, coronary angiography, and electrophysiology. These entail putting catheters into blood arteries to check heart function, blockages, pressures, and arrhythmias. These operations may include balloon angioplasty, stent insertion, or ablation. These diagnostic technologies let clinicians assess heart health, diagnose disease, and create treatment options. (Anon., n.d.).

2.3 Machine Learning in Healthcare

Machine learning (ML), an integral element of artificial intelligence (AI), is progressively employed in healthcare. A substantial proportion of organisations are adopting ML solutions and formulating AI plans. Supervised learning employs various algorithms, including linear regression, support vector machine regression, decision tree regression, and LASSO regression, to make quantitative predictions based on labelled data. Supervised learning uses logistic regression, K-nearest neighbours, Naive Bayes, and decision tree classification for categorical outcomes (Alanazi, 2022). In contrast, unsupervised learning is a method that aims to find patterns in data without the presence of labelled outcomes. This is achieved through techniques such as clustering and anomaly detection. Reinforcement learning is a cognitive process that entails acquiring decision-making skills by receiving feedback on the activities performed.

The significance of machine learning (ML) is underscored in its ability to improve decision-making processes, forecast health outcomes, and help with clinical and public health interventions. The efficacy of machine learning models is strongly dependent on the calibre of data, hence requiring broad, precise, and impartial datasets to guarantee the ability to generalise and maintain reliability. One notable challenge is the interpretability of machine learning models, specifically for healthcare personnel lacking proficiency in data science. The idea that models are mysterious entities that make predictions without specific reasons is created due to the complicated nature of models and the intense learning that takes place. This perception impacts the trustworthiness of models and their adoption in medical applications. In addition, the legal implications of machine learning predictions in the healthcare industry, including the liability for decisions generated from these predictions, provide a complex and complicated issue. Errors or inaccuracies in machine learning (ML) predictions can have significant implications for patient care, giving rise to legal considerations for healthcare practitioners and organisations operating within a legal framework that may not be comprehensively ready for the complexities associated with artificial intelligence (AI) and ML applications (Adlung, 2021).

2.4 Evolution of Machine Learning Techniques in Cardiovascular Disease Prediction

Traditional machine learning models, such as decision trees (DT), random forests (RF), and support vector machines (SVM), were utilised in the early stages of the cardiovascular disease prediction process. The importance of these conventional models in the initial phases of cardiovascular disease (CVD) prediction research cannot be exaggerated. They have demonstrated the feasibility of machine learning in medical prognosis and established a solid basis for investigating more sophisticated and intricate models. Their initial achievement in identifying persons at risk through pattern recognition in medical data has stimulated more investigation into deep learning and neural network models. These methods were intended to provide a foundational understanding of the potential for machine learning to analyse complicated medical datasets and uncover patterns that predict illness risk.

The study described in (Bárbara Martins, 2021) Concentrates on applying Data Mining Techniques (DMTs) to forecast cardiovascular diseases (CVDs) by examining clinical data obtained during medical examinations. The objective of this study is to identify individuals who are at a heightened risk of acquiring cardiovascular diseases (CVDs) in the early stages to prevent untimely deaths. The study utilised the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to apply five classifiers (DT, Optimised DT, RI, RF, and DL) on a dataset obtained from the Kaggle repository. The dataset comprises 70,000 patient records with 12 relevant parameters for diagnosing cardiovascular diseases (CVDs). These studies have laid the framework for further investigation into more complicated models.

The development of deep learning, which provides the capability to learn hierarchical representations of data automatically, was a big step forward in machine learning. In the study, by (Oguz Akbilgic, 2021), the predictive accuracy of electrocardiogram (ECG) data using artificial intelligence processing highlighted the capability of deep learning models to attain comparable accuracy in forecasting future heart failure. The study employed a deep residual convolutional neural network (CNN) as the methodology, utilising the dataset from the Atherosclerosis Risk in Communities (ARIC) study. The dataset consisted of electrocardiogram (ECG) data from 14,613 people collected between 1987 and 1989. The CNN model, specifically developed to forecast heart failure (HF) only based on ECG data, contrasted with the HF risk calculators currently utilised in the ARIC and Framingham Heart projects (FHS). The CNN model attained an area under the curve (AUC) of 0.756, like the risk calculators' performance. When the output of the CNN model was integrated with clinical predictors in a Light Gradient Boosting Machine (LGBM) model, it achieved the most excellent Area Under the Curve (AUC) value of 0.818, suggesting a high level of prediction accuracy for Heart Failure (HF). The output of the ECG-AI model was identified as the most significant predictor. These studies marked a substantial change towards utilising more complex algorithms that could handle the complexities of cardiovascular data more efficiently.

2.5 Challenges of Data Handling and Feature Selection

One of the most persistent issues when predicting cardiovascular illness using machine learning is managing imbalanced and incomplete datasets and selecting features pertinent to the problem. Principal Component Analysis (PCA), the Relief algorithm, and the Least Absolute Shrinkage and Selection Operator (LASSO) are some methods to address these challenges. These methods have been employed to find the most significant predictors of cardiovascular disease.

A novel strategy is introduced by (AZAM MEHMOOD QADRI, 2023) The Principal Component Heart Failure (PCHF) method improves the performance of machine learning models in predicting heart illness by optimising the feature set by selecting the eight most significant characteristics. The study utilised nine distinct machine learning algorithms to assess their efficacy, with the decision tree model attaining the best accuracy on data from the Kaggle repository, which included 1025 patient records.

To develop an effective model that could predict heart disease with high accuracy, efficient data collection, pre-processing, and transformation methods, integrating datasets from various sources and using new hybrid classifiers for training (PRONAB GHOSH, 2021) have been utilised. These algorithms, enhanced by Relief and LASSO feature selection techniques, have been used to predict cardiovascular diseases (CVD).

In addition, hybrid genetic algorithm (GA) and particle swarm optimisation (PSO) approaches with a random forest (RF) classifier, referred to as GAPSO-RF, were proposed by (Mohamed G. El-Shafiey, 2022), have demonstrated potential in terms of optimising feature selection and model performance. The study employed two datasets for evaluation: the Cleveland and Statlog datasets from the University of California Irvine (UCI) machine learning repository. This suggests a trend towards more sophisticated methods for managing dataset challenges.

2.6 Integration of Advanced Algorithms for Improved Prediction Accuracy

The incorporation of sophisticated algorithms and ensemble approaches has emerged as a central focus of research that aims to achieve the highest possible level of accuracy in projections. It has been demonstrated that hybrid models, which capitalise on the advantages of various machine learning algorithms, have shown excellent performance and, as an illustration, (Rout, 2023) They have developed an effective prediction model for heart disease using an improved Particle Swarm Optimization (IPSO) algorithm and an ensemble classification technique. Given the growing size and complexity of medical datasets, this research addressed the challenges in accurately predicting heart disease. Two primary datasets were used for this purpose: the Shahid Rajaei Hospital dataset and the UCI Cleveland dataset. The methodology involved several key steps: data collection and preprocessing using min-max normalisation, feature extraction and selection through Recursive Feature Elimination and the IPSO algorithm, and classification using an ensemble classifier incorporating multiple machine learning models. This comprehensive approach aimed to optimise feature selection and improve the accuracy of heart disease prediction, demonstrating significant improvements over traditional machine learning techniques.

The utilisation of generative adversarial networks (GAN) for data augmentation, as utilised by (Raniya R. Sarra, 2022), bi-directional long short-term memory (Bi-LSTM) networks are examples of creative tactics being developed to improve the robustness and accuracy of models. The dataset utilised was the publicly available Cleveland dataset from the University of California, Irvine (UCI) Machine Learning Repository. Not only do these methods overcome the shortcomings of conventional and singular machine learning models, but they also pave the path for the achievement of greater precision in disease prediction.

2.7 Critical Evaluation

Paper	Dataset	Proposed	Techniques
(Bárbara Martins, 2021)	Kaggle Data Repository	To satisfy the urgent need to extract valuable knowledge hidden in clinical data, particularly to develop a solution to predict the presence/absence of CVDs through Data Mining.	K-NN, Naïve Bayes, DT, RF, Gradient Boosted Tree, Rule Induction, Deep Learning, Generalized Linear Model and Logistic Regression.
(Oguz Akbilgic, 2021)	Atherosclerosis Risk in Communities (ARIC)	This study assesses the utility of electrocardiograms (ECGs) in HF risk prediction.	ECG-AI (CNN)
(AZAM MEHMOOD QADRI, 2023)	Kaggle Repository	PCHF feature engineering technique is proposed to select the most prominent features to improve performance.	PCHF features engineering on DT, LR, RF, SVM, K-NN, Naïve Bayes Multilayer Perceptron
(PRONAB GHOSH, 2021)	Combined Dataset (UCI and Stat log)	Relief and LASSO extract the most relevant features for overfitting and underfitting problems.	Relief and LASSO feature selection on DT, RF, K-NN, AdaBoost and Gradient Boosting
(Mohamed G. El-Shafiey, 2022)	Cleveland and Statlog	Select the best features to improve heart disease prediction.	Genetic Algorithm and Particle Swarm Optimization on Random Forest (GAPSO-RF)

(Rout, 2023)	Shahid Rajaei Hospital and Cleveland	Effective feature optimisation and classification methodologies for heart disease prediction.	Improved Particle Swarm Optimization (IPSO) algorithm and Ensemble Classification (KNN, SVM, DT, LT and Naïve Bayes)
(Raniya R. Sarra, 2022)	Cleveland and Stat log	Augment the imbalanced and limited data to have balanced and more extensive data; improve the performance.	GAN-1D-CNN and GAN-Bi-LSTM

Table 1: Recent works on CVD

Much of the research, including highly accurate studies such as those conducted by (In Table 1), strongly depends on datasets. Determining whether these models can be generalised to other populations or datasets is challenging. Because cardiovascular disease risk factors and manifestations can vary significantly across different demographic groups, the difficulty of dataset diversity and representativeness is a noticeable gap.

One of the most important intersections between clinical medicine and data science is highlighted by evaluating heterogeneous datasets using various models to achieve worldwide clinical acceptance in predicting cardiovascular disease (HD). Because cardiovascular health is very complicated and patient populations worldwide are very different, it is essential to investigate how machine learning (ML) models can be used successfully in various demographic and clinical settings. One of the goals of this critical evaluation is to fully understand the consequences of these factors on global health systems. This study considers the interaction between data diversity, model selection, and potential clinical advantages.

A promising new area that has the potential to revolutionise cardiovascular treatment is the implementation of machine learning algorithms for predicting cardiovascular disease in clinical settings around the world. The dedication to increasing the diversity of datasets, selecting models with care, and resolving ethical and legal challenges is essential to the success of this endeavour. To achieve global clinical acceptance, it will be necessary to innovate technologically and make a concentrated effort to guarantee that these models are interpretable and seamlessly incorporated into healthcare systems worldwide. To maximise the potential of machine learning to improve cardiovascular health outcomes across a wide range of people, it will be necessary to conduct continuing evaluations and adjust as the field advances.

2.8 Conclusion

Every single theoretical technique utilised within the realm of cardiovascular disease prediction through machine learning has contributed to the overall comprehension and development of the area. Pattern identification in complicated datasets was initially addressed by traditional machine learning algorithms, which created the framework for the field. As a result of the development of hybrid models and improved algorithms, the unique difficulties of data imbalance and feature selection have been targeted, leading to increased prediction accuracy. The introduction of deep learning has provided the power to extract deeper insights from medical data simultaneously. In the future, research should concentrate on improving these advanced methodologies, ensuring that models are generalised across a wide range of populations, and investigating the possibility of integrating multimodal data sources to further enhance the accuracy and reliability of predictions regarding cardiovascular disease.

By incorporating machine learning models into clinical practice for cardiovascular disease prediction, the ultimate objective is to improve patient care by enabling early identification, personalised therapy, and effective resource allocation. For these models to gain widespread adoption, it is necessary to demonstrate their capacity to predict outcomes accurately, flexibility in various clinical contexts, simplicity of integration into preexisting workflows in the healthcare industry, and adherence to ethical standards. In addition, models need to be verified in various populations to guarantee their dependability and performance in different socioeconomic and geographical settings.

These outcomes inspire the current research, which leads to a comparative analysis of deep-learning models, specifically 1D-CNN and TabNet, across varied demographic datasets, including those from Hungary, Switzerland, and Cleveland, utilising augmented data.

3 Requirements Specification

3.1 Functional Requirements

System Capabilities

Heart Disease Prediction: The system should accurately predict the risk of heart disease using demographic and medical data inputs. This involves processing and analysing data through 1D-CNN and TabNet models to generate risk assessments.

Data Processing and Analysis

Data Augmentation: Implement data augmentation techniques to enhance the size and quality of the dataset, ensuring robust model training and evaluation.

Model Training: The system must facilitate training 1D-CNN and TabNet models using the specified datasets, incorporating a validation process to monitor performance and overfitting.

Model Evaluation: Using the augmented datasets, perform a comparative analysis of the models, focusing on each model's efficacy and accuracy in predicting heart disease.

3.2 Non-Functional Requirements

Performance Metrics

Accuracy, Precision, and Recall: Define and measure each model's key performance indicators, such as accuracy, precision, and recall, to assess and compare their effectiveness in predicting heart disease.

Response Time: The system should provide predictions promptly, ensuring quick feedback to healthcare professionals.

Scalability

Data and Model Scalability: Ensure the system can handle increasing data and additional models without compromising performance.

3.3 Data Requirements

Data Types and Sources

Demographic Information: Age, gender, and ethnicity data from the specified geographic regions (Hungary, Switzerland, and Cleveland).

Medical Data: Historical medical data, including but not limited to blood pressure, cholesterol levels, and other relevant health indicators.

4 Implementation

4.1 Programming Languages and Libraries:

Python: The primary programming language used for data processing, model training, and evaluation. Python's extensive ecosystem of libraries makes it a popular choice for machine learning and data science projects.

Pandas: A library used for data manipulation and analysis. This project loads datasets, handles data frames, and performs data-cleaning operations (Anon., 2024).

NumPy: Essential for numerical operations on arrays. Although not explicitly mentioned, it's commonly used alongside Pandas and for handling data structures for machine learning models (Anon., 2023).

CTGAN: A Python library for generating synthetic tabular data using Conditional Generative Adversarial Networks. It's used here to augment data by synthesising new records (Anon., 2020).

Table Evaluator: This library evaluates and compares the statistical properties of the original dataset against the synthetic dataset generated by CTGAN (Anon., 2023).

Scikit-learn: A machine learning library for Python. It's used for data preprocessing (e.g., train-test split, standard scaling) and dimensionality reduction (PCA) and could potentially be used for model evaluation metrics (Anon., 2024).

TensorFlow (Keras): An open-source machine learning framework. Keras, TensorFlow's high-level API, is used to construct, compile, and train the 1D Convolutional Neural Network model (Anon., 2024).

PyTorch (TabNet): Another open-source machine learning library, used here specifically for TabNet, a deep learning model for tabular data. It's noted for its use of sparse feature learning and self-attention mechanisms (Anon., 2024).

4.2 Tools:

Google Colab: Although not explicitly mentioned, the code snippet provided is compatible with what you would typically run in a Google Colab environment. This environment is also commonly used for Python scripting, data analysis, and prototyping machine learning models. Google Colab provides an interactive coding environment in the cloud, facilitating easy access to powerful computing resources and collaboration features.

4.3 Data Preparation

4.3.1 Data Found

The datasets are collected from the UCI Machine Learning Repository. (Andras Janosi, 1988). The three datasets are from Cleveland, Hungary and Switzerland. They contain 76 attributes, but all published experiments use a subset of 14 of them. The 14 features are mentioned below. (Andras Janosi, 1988):

1. Age: The patient's age in years (continuous).

2. Sex: The patient's sex (1 = male; 0 = female).

3. Chest Pain Type: Type of chest pain experienced (categorical: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic).

4. Resting Blood Pressure: Resting blood pressure (in mm Hg on admission to the hospital; continuous).

5. Serum Cholestoral: Serum cholesterol in mg/dl (continuous).

6. Fasting Blood Sugar: Fasting blood sugar > 120 mg/dl (1 = true; 0 = false).

7. Resting Electrocardiographic Results: Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria).

8. Maximum Heart Rate Achieved: Maximum heart rate achieved (continuous).

9. Exercise-Induced Angina: Exercise-induced angina (1 = yes; 0 = no).

10. ST Depression Induced by Exercise Relative to Rest: ST depression induced by exercise relative to rest (continuous).

11. Slope of the Peak Exercise ST Segment: The slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping).

12. Number of Major Vessels (0-3) Colored by Flourosopy: Number of major vessels (0-3) colored by fluoroscopy (categorical but represented as objects which may indicate some missing or placeholder values).

13. Thal: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect; represented as objects, which again may indicate non-numeric data entries).

14. Diagnosis of Heart Disease (Angiographic Disease Status): Diagnosis of heart disease (angiographic disease status; value 0: < 50% diameter narrowing, value 1-4: > 50% diameter narrowing, with higher values indicating more severe disease).

Cleveland

The Cleveland dataset contains 303 instances of clinical records with 14 attributes and a few missing values. (Andras Janosi, 1988). The visualisation illustrates the demographic and correlation of features in the dataset:

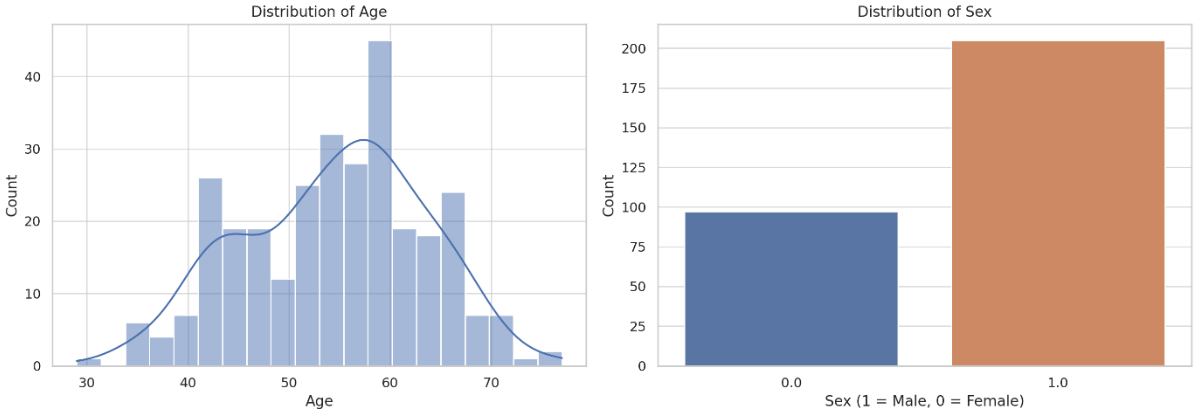
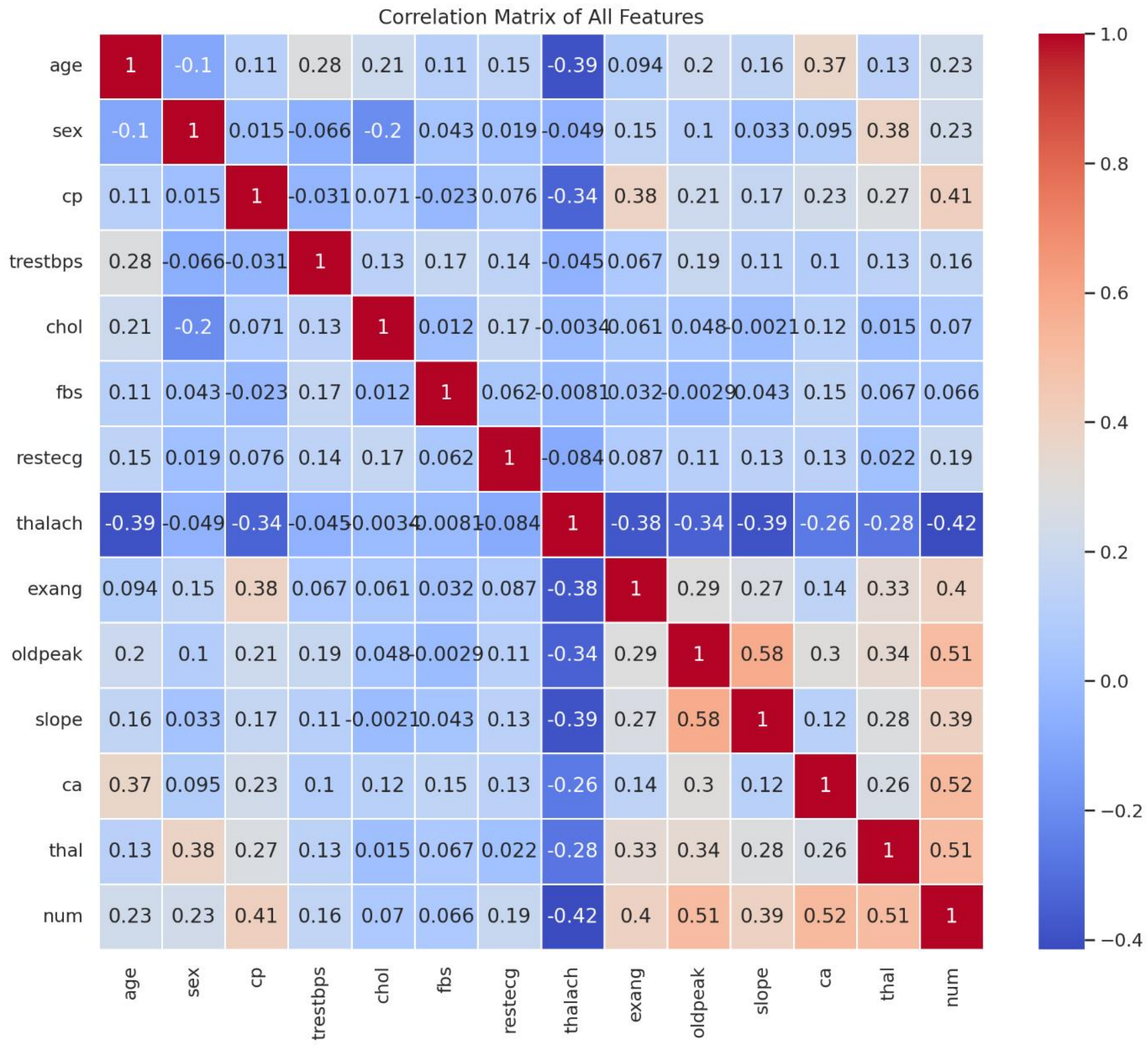


Figure 1: Cleveland Distribution of Age and Gender

Distribution of Age: The histogram (Figure 1) indicates a broad distribution that peaks in the late 50s to early 60s. This suggests a significant representation of older individuals, aligning with the higher risk of cardiovascular disease in older age groups.

Distribution of Sex: The bar chart (Figure 1) shows the distribution between male and female patients, with more male patients in the dataset. This demographic detail is crucial for understanding the dataset's bias towards male patients and considering the impact of sex on cardiovascular disease risk and outcomes.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
age	1	-0.1	0.11	0.28	0.21	0.11	0.15	-0.39	0.094	0.2	0.16	0.37	0.13	0.23
sex	-0.1	1	0.015	-0.066	-0.2	0.043	0.019	-0.049	0.15	0.1	0.033	0.095	0.38	0.23
cp	0.11	0.015	1	-0.031	0.071	-0.023	0.076	-0.34	0.38	0.21	0.17	0.23	0.27	0.41
trestbps	0.28	-0.066	-0.031	1	0.13	0.17	0.14	-0.045	0.067	0.19	0.11	0.1	0.13	0.16
chol	0.21	-0.2	0.071	0.13	1	0.012	0.17	-0.0034	0.061	0.048	-0.0021	0.12	0.015	0.07
fbs	0.11	0.043	-0.023	0.17	0.012	1	0.062	-0.0081	0.032	-0.0029	0.043	0.15	0.067	0.066
restecg	0.15	0.019	0.076	0.14	0.17	0.062	1	-0.084	0.087	0.11	0.13	0.13	0.022	0.19
thalach	-0.39	-0.049	-0.34	-0.045	-0.0034	-0.0081	-0.084	1	-0.38	-0.34	-0.39	-0.26	-0.28	-0.42
exang	0.094	0.15	0.38	0.067	0.061	0.032	0.087	-0.38	1	0.29	0.27	0.14	0.33	0.4
oldpeak	0.2	0.1	0.21	0.19	0.048	-0.0029	0.11	-0.34	0.29	1	0.58	0.3	0.34	0.51
slope	0.16	0.033	0.17	0.11	-0.0021	0.043	0.13	-0.39	0.27	0.58	1	0.12	0.28	0.39
ca	0.37	0.095	0.23	0.1	0.12	0.15	0.13	-0.26	0.14	0.3	0.12	1	0.26	0.52
thal	0.13	0.38	0.27	0.13	0.015	0.067	0.022	-0.28	0.33	0.34	0.28	0.26	1	0.51
num	0.23	0.23	0.41	0.16	0.07	0.066	0.19	-0.42	0.4	0.51	0.39	0.52	0.51	1

Figure 2: Correlation of Cleveland Features

Correlation Matrix of All Features: The heatmap (Figure 2) presents the correlation coefficients between all variables. While many features show low to moderate correlations with each other, a few points stand out:

- 'oldpeak' (ST depression induced by exercise relative to rest) shows some level of positive correlation with 'age', 'sex', and 'exang' (exercise-induced angina), suggesting that older males with exercise-induced angina are more likely to exhibit ST depression.

- 'thalach' (maximum heart rate achieved) is inversely correlated with 'age', which is physiologically expected as maximum heart rate typically decreases.

- The target variable 'num' (diagnosis of heart disease) shows various degrees of correlation with several predictor variables, notably 'cp' (chest pain type), 'thalach', 'exang', and 'oldpeak'. This indicates that these features could significantly predict heart disease in the model.

Hungary

The Hungarian dataset contains 293 clinical records with 14 attributes and many missing values on the ‘slope’, ‘ca’, and ‘thal’ features. (Andras Janosi, 1988). Because of this feature's high number of missing values, these columns were dropped from the data frame. Illustrate how demographic and correlation of features in the dataset is by the visualisation down below:

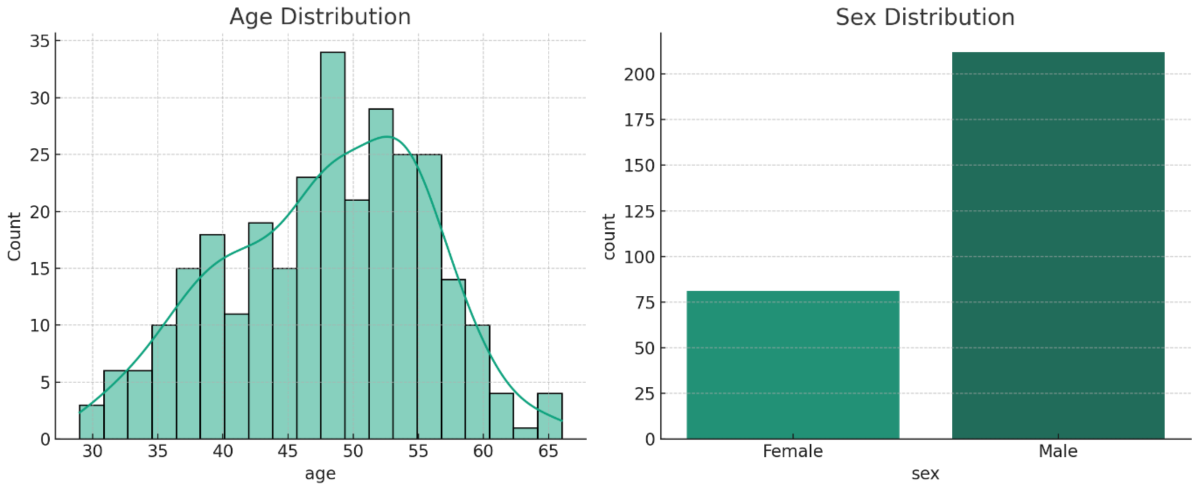


Figure 3: Hungarian distribution of Age and Gender

Age Distribution

The age range of individuals in the dataset is 29 to 66. The age histogram shows a relatively normal distribution with a slight right skew, suggesting a more significant proportion of individuals in the dataset are in mid-to-late middle age. There's a noticeable concentration of individuals in the 50-60-year age bracket (Figure 3).

Sex Distribution

The `sex` feature likely represents biological sex, with 0 for female and 1 for male, based on standard conventions in medical datasets. There's a notable imbalance in the distribution of sexes; approximately 210 individuals are marked as `sex` 1 (male), and more than 75 individuals are marked as `sex` 0 (female). This indicates a higher representation of males in the dataset (Figure 3).

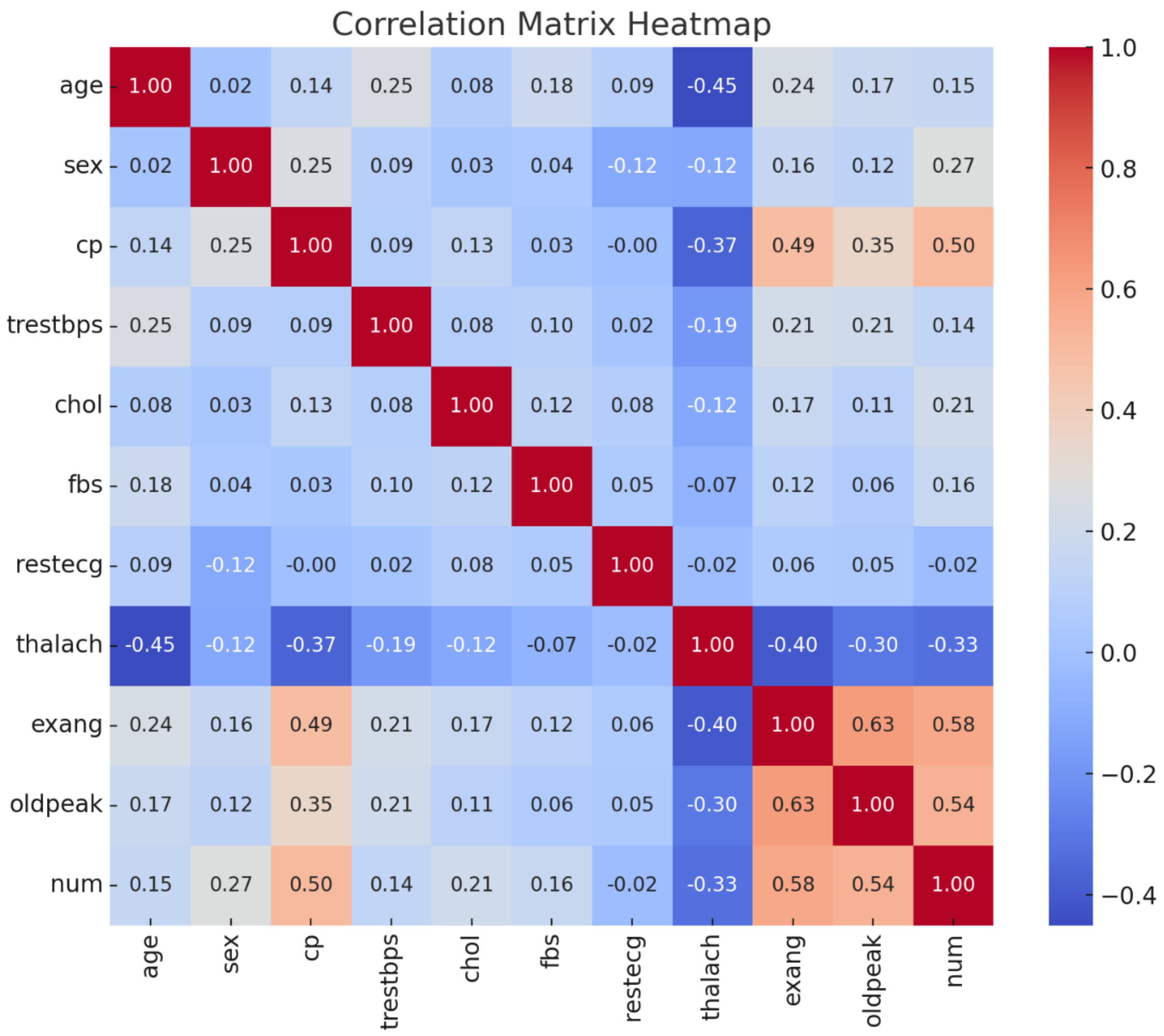


Figure 4: Correlation of Hungarian Features

Key points from the heatmap include:

Certain features might show stronger correlations with others, indicated by the colours; warmer colours (e.g., red) denote a higher positive correlation, whereas cooler colours (e.g., blue) denote a negative correlation. The diagonal, as expected, shows a perfect correlation of 1.00 for each variable with itself (Figure 4). Off-diagonal elements give insights into potential relationships between health indicators and demographic features. This visualisation is handy for identifying features that may significantly impact the outcome (`num`, the diagnosis of heart disease) and could be essential predictors in heart disease risk assessment models.

Switzerland

The Switzerland dataset contains 122 entries with 14 attributes and many missing values on ‘chol’, ‘FBS’, ‘ca’ and ‘thal’ features. (Andras Janosi, 1988). Because of this feature's high number of missing values, these columns were dropped from the data frame. Illustrate how demographic and correlation of features in the dataset is by the visualisation down below:

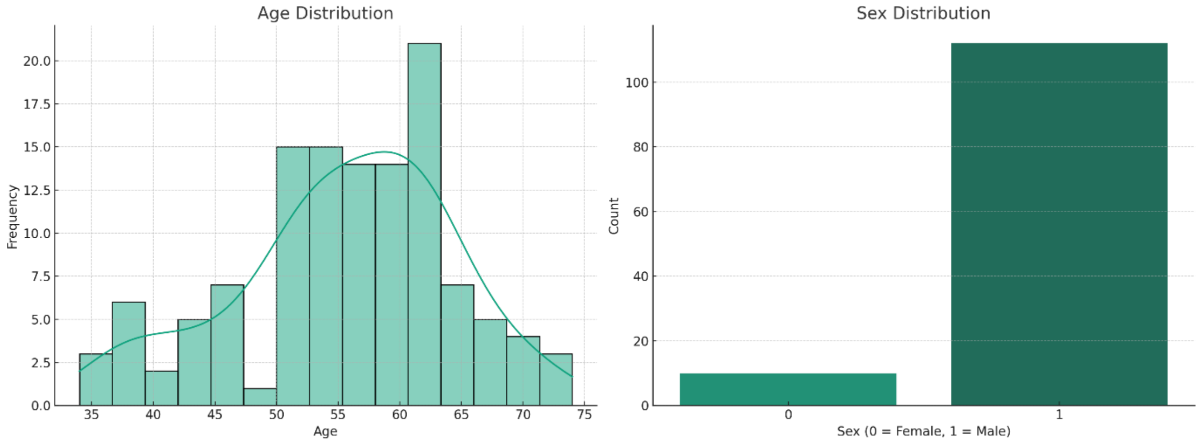


Figure 5: Distribution of Age and Gender (Switzerland)

Age Distribution: The age distribution of patients shows a wide range, with a concentration of patients in the middle-aged to older age range. The histogram reveals a slightly right-skewed distribution, indicating a more significant number of older patients (Figure 5).

Sex Distribution: The sex distribution highlights a significant predominance of male patients (1) over female patients (0) in this dataset (Figure 5).

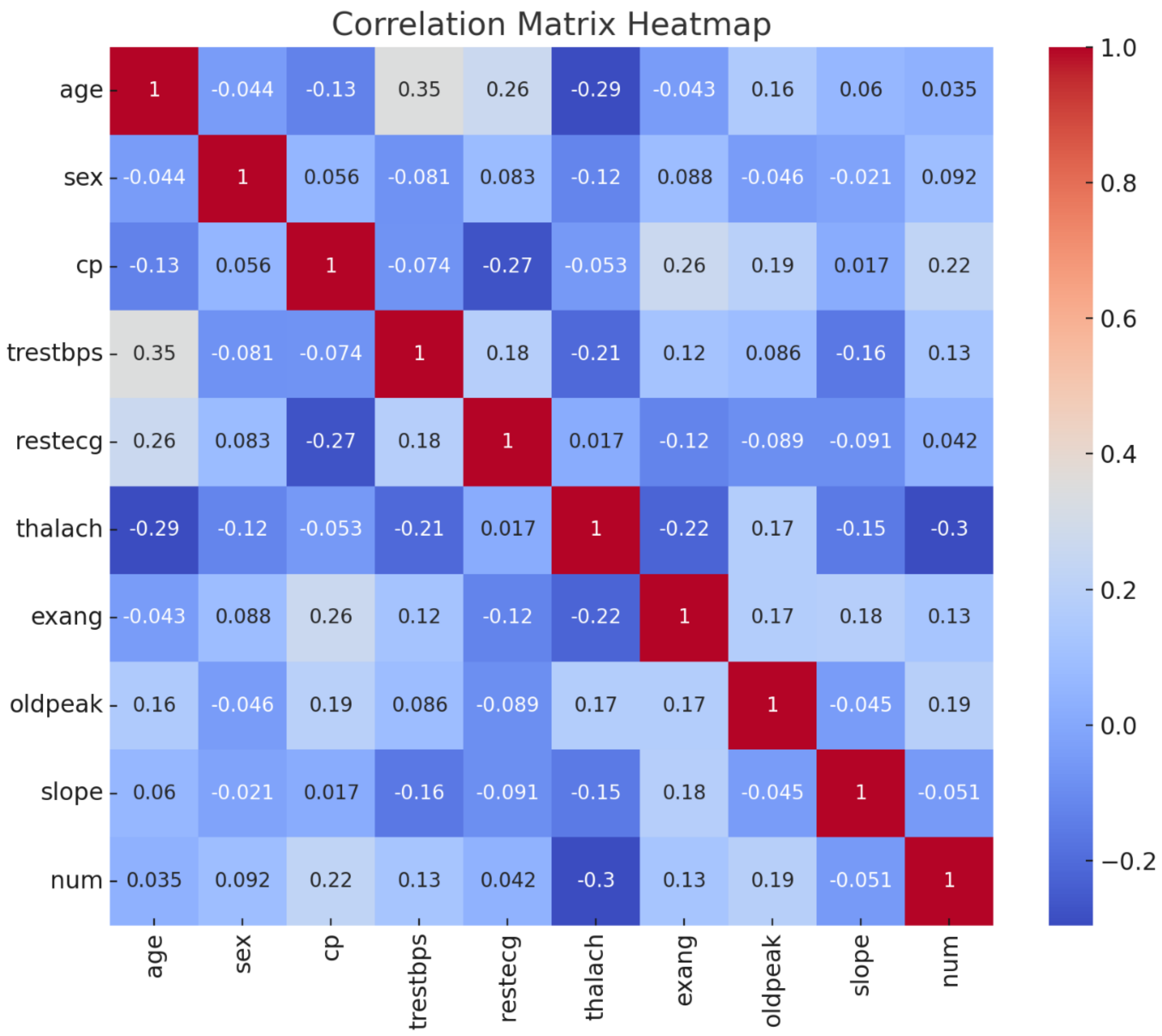


Figure 6: Correlation of Switzerland Features

Correlation Matrix Heatmap:

The heatmap (Figure 6) of the correlation matrix for the ten features in the dataset reveals several key observations:

Age shows a mild positive correlation with the target variable (`num`), suggesting that older individuals might have a slightly higher risk of heart disease. Sex has a very low correlation with the target variable, indicating that sex may not be a strong predictor of heart disease within this dataset. Chest pain type (cp) shows a positive correlation with the target variable, which aligns with clinical expectations that certain types of chest pain are more indicative of heart disease. trestbps (resting blood pressure) and thalach (maximum heart rate achieved) have low to moderate correlations with the target variable, suggesting some relationship to the presence of heart disease, albeit not very strong. Exercise-induced angina (exang), oldpeak (ST depression induced by exercise relative to rest), and the slope of the peak exercise ST segment show moderate to strong correlations with the target variable, emphasising their importance in diagnosing heart disease. The correlations between features and the target variable offer insights into potential risk factors and their relationships with heart disease presence in the Switzerland dataset. However, it's important to remember that correlation does not imply causation and further analysis is required to understand these relationships deeply.

This information covers various factors relevant to diagnosing and assessing cardiovascular disease, including demographic data, symptoms, and results from multiple medical tests. The mixture of continuous, categorical, and binary data types allows for a multifaceted analysis of cardiovascular health, which is essential for developing accurate machine-learning models. The presence of numerical and object data types for some categorical variables suggests that the dataset may require cleaning or preprocessing to handle non-numeric or missing values appropriately.

4.3.2. Handling Missing Values:

For the preprocessing of the datasets, start by dealing with missing values, referred to as '?'. Managing missing data is generally accomplished by finding and substituting placeholders of disappeared values ('?') with NaN (Not a Number), utilising the replace method in the pandas library. This translation process ensures that missing values are uniformly represented throughout the collection. Afterwards, the dropna() technique eliminates rows containing these NaN values from the dataset. This approach eliminates missing values in the dataset, streamlining subsequent data processing and modelling tasks.

4.3.3. Outlier Detection and Removal:

Remove the outliers after dealing with the missing values in the data frame. Outliers in a dataset can significantly impact statistical analyses and machine learning models by causing biased or inaccurate results. For this project, outliers are handled using the Interquartile Range (IQR) method. The Interquartile Range (IQR) is a crucial statistic for identifying and handling outliers in a dataset, representing the middle 50% of the data by calculating the difference between the 75th percentile (Q3) and the 25th percentile (Q1). Bounds are set around this range to determine which data points are outliers. The lower bound is defined as (Q1 - 1.5 × IQR); any values below this are considered outliers. Similarly, the upper bound is set at (Q1 + 1.5 × IQR), with values above this threshold also deemed outliers. Data points outside these bounds are filtered out from the dataset, ensuring extreme values do not skew the analysis or predictive modelling processes. This method effectively standardises the dataset by removing these extreme values, promoting more accurate and reliable statistical and machine-learning outcomes. (Ch. Sanjeev Kumar Dash, 2023).

4.3.4 Data Augmentation (CTGAN):

The datasets found from the UCI repository are considerably small. Therefore, the paper had to produce augmented data to create a large dataset for the models. This research used the CTGAN (Conditional Generative Adversarial Network) model. CTGAN is a sophisticated variation of the traditional Generative Adversarial Network (GAN) designed to generate synthetic tabular data. The architecture consists of two main components: the generator and the discriminator. The generator uses input noise vectors to create artificial data instances that closely mimic the actual data distribution. At the same time, the discriminator evaluates these instances against actual data to determine their authenticity. This adversarial process is enhanced by conditional vectors, where the model conditions the generation process on specific label or feature columns, promoting the creation of diverse and representative data samples. Throughout the training, both the generator and discriminator iteratively enhance their capabilities through adversarial training, employing backpropagation and optimisation techniques such as the Adam optimiser to refine their performance. This continuous interaction helps the generator produce increasingly realistic data and the discriminator to improve its accuracy in distinguishing between real and synthetic instances (AYESHA SIDDIQUA DINA, 2022). CTGAN is particularly useful for datasets containing categorical and continuous features. It handles categorical data by embedding and transforming these features, enabling the network to better model and generate such data.

4.3.5 Scaling:

Data scaling is essential in preparing datasets for analysis and modelling in machine learning and statistical computing. It is different from data augmentation and is a vital preparation step. This process is meant to standardise the features' scale, making it easier for variables to be the same. In this project, 'StandardScaler' is applied to the training data (X_train) using the fit_transform method, which computes the parameters (mean and standard deviation for each feature) and transforms the training data. The transformation is then applied to the test data (X_test) using only the transform method, which uses the same parameters computed from the training data to ensure consistency in how the data is scaled. The `StandardScaler` method standardises a feature by subtracting the mean (μ) and then scaling it to unit variance. The mean is the average of the values in the feature, and the standard deviation (σ) measures how much the values deviate from the mean. The standardisation process transforms each element (x) in the feature according to the formula ($z = \frac{(x-\mu)}{\sigma}$\). In this formula, z represents the standardised value (Danny Hartanto Djarum, 2021). This process effectively adjusts the scale of the feature values to have a mean of zero and a standard deviation of one, which is crucial for models sensitive to the magnitude of input features.

4.3.6 Dimensionality Reduction (PCA):

Dimensionality reduction is one of the most essential steps in machine learning and data science preprocessing. It solves the critical problem of managing large datasets with many dimensions. This project utilised Principal Component Analysis (PCA) to reduce the dimensionality of the dataset. Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction while retaining as much variance (information) as possible. Principal Component Analysis (PCA) begins by standardising the data, ensuring that each feature contributes equally, which is essential when dealing with variables measured in different units. Next, PCA computes the covariance matrix to understand how the variables vary. The covariance matrix is then decomposed into eigenvalues and eigenvectors. The eigenvectors determine the directions of the new feature space, while the eigenvalues indicate their magnitude, representing how much variance there is in the data along these new axes. The eigenvectors are sorted by decreasing eigenvalues, and the top 'N' eigenvectors are selected to form a new feature space. This selection is based on keeping the dimensions that capture the most variance. Finally, the original data is projected onto this new feature space, resulting in a transformed dataset where the principal components with the most significant variance come first, minimising information loss while reducing dimensionality. (Bharadiya, 2023).

This comprehensive data preparation process ensures that the datasets are clean, standardised, and augmented, ready for practical machine learning model training and evaluation.

4.4 Model Implementation

4.4.1 1D-CNN Implementation

The 1D Convolutional Neural Network (1D-CNN) is designed to process sequential or time-series data, making it well-suited for analysing the patterns within the feature set of the heart disease dataset despite it not being temporal. The application of 1D-CNN here is somewhat innovative, leveraging its ability to extract local patterns from the individual features of the dataset.

Model Structure

The architecture of 1D Convolutional Neural Network (1D-CNN) model is specifically designed to process one-dimensional signal data effectively. The model's input layer receives a data sequence with six-time steps, each containing one feature. This setup is particularly tailored to handle time series or sequential data where temporal dynamics are crucial.

The first layer of the model is a convolutional layer equipped with 64 filters, each with a kernel size of two. The filters in this convolutional layer slide across the input data, performing element-wise multiplications by their weights and summing the results, a process defined mathematically as:

$$y(t) = \mathrm{ReLU}\left(\sum_{k=0}^{1} f(\mathrm{k}) \times x(t-k) + b\right)$$

- f(k) represents the weights of the k-th filter.
- x(t−k) represents the input at position t−k.
- b is the bias.
- ReLU activation: ReLU(x) = max(0, x).

The ReLU (Rectified Linear Unit) function introduces non-linearity, allowing the model to learn more complex patterns. (Manali Saini, 2022).

Following the first convolutional layer, a max pooling layer with a pool size of one is applied. Typically, pooling operations help reduce the data's dimensionality and prevent overfitting by taking the maximum value within the window (pool size) as the representative value.

$$y(t) = max(x(t))$$

A dropout layer with a dropout rate of 0.2 is then applied to reduce overfitting by randomly setting 20% of the input units to zero at each update during training time.

Subsequently, a second convolutional layer containing 128 filters, each with a kernel size of two, is employed. This layer uses the ReLU activation function to output values between 0 and 1, which helps normalise the output of neurons.

Another dropout layer with a rate of 0.2 follows to enhance the model's generalisation capabilities further. The output from this layer is then flattened to transform the multi-dimensional feature maps into a one-dimensional feature vector, facilitating the transition to fully connected layers.

The flattened output is fed into a dense layer consisting of 64 neurons, which applies the ReLU activation function to continue the non-linear transformation of features. An additional dropout layer follows at the same rate of 0.2, aiming to minimise overfitting.

After the feature extraction and regularisation steps carried out by the convolutional, pooling, and dropout layers, the model's architecture incorporates a fully connected layer. This layer, often called a dense layer, integrates learned features across the entire model. This is a standard fully connected layer with 64 neurons. Each neuron in this layer receives input from all the neurons of the previous layer (post-flattening), meaning the input is fully connected to these 64 units. The output for each neuron is computed as follows:

$$y = ReLU(Wx + b)$$

- W represents the weight matrix connecting the input from the flattened layer to the neurons in the dense layer.
- x is the input vector to the dense layer.
- b is the bias vector.

Following the dense layer, a dropout layer is applied, randomly setting 20% of the neuron outputs to zero during training. Finally, the output layer comprises a single neuron with a sigmoid activation function to produce a probability score between 0 and 1, making it suitable for predicting the likelihood of one class over another. (Manali Saini, 2022) The mathematical expression for the sigmoid activation is:

$$y = \sigma(Wx + b)$$

Where:

- σ(x) = 1/ (1+ r^{-x}) The sigmoid function squashes the input x to a value between 0 and 1.
- W and b are the weight and bias for this layer, and x is the input from the previous (fully connected or dropout) layer.

The model is compiled using the Adam optimiser, renowned for its efficiency in handling sparse gradients and noisy problems. The loss function selected is the mean squared error (MSE). It effectively measures the average of the squares of the differences between predicted and actual values, making it particularly suitable for regression tasks. The metric used to evaluate the model's performance during training and validation is accuracy, which assesses the proportion of correctly predicted instances over the total number of cases evaluated.

This methodology leverages the inherent capabilities of 1D-CNNs to process sequential data through convolutional, pooling, and fully connected layers, enabling the effective extraction, transformation, and utilisation of temporal features for classification purposes. (Serkan Kiranyaz, 2021).

Challenges and Solutions

Dimensionality: The primary challenge was adapting the 1D-CNN to work effectively with non-temporal data. This was addressed by reshaping the input data to have a 'channel' dimension, allowing the convolutional layers to process the data.

Feature Selection and Reduction: Before applying the 1D-CNN, it was crucial to reduce the dimensionality of the data using PCA. This step was essential to ensure the model could efficiently learn from the most variance-explaining features without being overwhelmed by the curse of dimensionality.

Overfitting: Due to the relatively small dataset size, there was a risk of overfitting. Dropout layers and data augmentation (using CTGAN to generate synthetic data) were strategies to mitigate this issue.

4.4.2 TabNet Implementation

TabNet is a novel deep learning model for tabular data that utilises sparse attention in its architecture to select which features to focus on during training. This approach allows TabNet to learn complex patterns in the data and makes it particularly effective for datasets where relationships between features are not linear or easily discernible.

Model Structure

In this section, the paper describes the architecture and operational mechanics of the TabNet classifier, a novel neural network specifically designed for tabular data. The TabNet model integrates the concepts of deep learning with the interpretability typically associated with decision tree algorithms. It achieves this through the innovative use of an attention mechanism that allows for the selective processing of the most relevant features at each decision step. The mathematical formulation of the TabNet model is rooted in its distinctive components: feature transformation, feature selection via attention, and training methodology. (Sercan O. Arık, 2021).

The TabNet architecture uses a sequential attention mechanism to learn which features to attend to at each decision step. This attention mechanism is guided by the model's learning and past decisions. The model structure comprises primarily two types of transformations: the feature transformer and the attentive transformer. The feature transformer is a feed-forward network with a structure akin to that found in typical neural networks and is designed to capture the interactions between features.

The operation begins with the input features, $f \in R^D$, Where D is the number of features, the model computes a feature mask M[i] using the attentive transformer for each decision step I. This mask is used to perform instance-wise feature selection. The mask is generated through the following transformation:

$$M[i] = sparsemax(P[i-1] \times h(a[i-1]))$$

Where h represents a trainable transformation function that typically comprises fully connected layers followed by batch normalisation, P[i-1] denotes the prior scale that adjusts feature selection based on previous usage to encourage feature reuse across steps, and sparsemax is a normalisation function that promotes sparsity in the attention weights (Martins, 2016).

Following the application of the mask, the feature transformer applies a non-linear transformation to the selected features:

$$[d[i], a[i]] = f\_i(M[i] \times f)$$

Where f_i denotes the layers of the feature transformer specific to step i, d[i] is the decision output for that step, and a[i] is the aggregated output that serves as an input to the next decision step. The feature transformer includes layers shared across steps and layers specific to each step, enhancing the model’s ability to generalise across different types of feature interactions.

Training the TabNet model involves optimising a loss function that combines a predictive performance metric (cross-entropy or mean squared error) and a sparsity regularisation term. This regularisation term is essential for the interpretability of the model as it enforces sparsity in the feature masks:

$$L\_sparse = \sum_{i=1}^{N_{steps}} \sum_{b=1}^{B} \sum_{j=1}^{D} \frac{-M\_b,j[i]\,log(M\_b,j[i] + \epsilon)}{N\_steps\ \times B}$$

N_steps is the number of decision steps, B is the batch size, and D is the number of features, with ∊ a small constant to ensure numerical stability.

The aggregation of decision steps' outputs is achieved through an element-wise non-linear transformation, typically a Rectified Linear Unit (ReLU), followed by a summation across steps:

$$\text{output} = \sum_{i=1}^{N\_steps} \text{ReLU}(\text{d}[\text{i}])$$

This final aggregated output is then used to make predictions or further processed depending on the specific application.

The TabNet model was initialised with an Adam optimiser and a learning rate of 1e-3. Learning rate scheduling was also employed to adjust the learning rate over training epochs, optimising the training process. The model was set to use the 'entmax' function for feature selection, a generalisation of softmax that can yield sparse probabilities, enhancing feature selection capabilities. PyTorch was the underlying framework, with the option to run the model on a GPU if available, accelerating the training process. (Sercan O. Arık, 2021).

Challenges and Solutions

- Feature Learning: One of the primary challenges was ensuring that TabNet could effectively learn from the reduced dimensionality data produced by PCA. This was critical as PCA-transformed features are linear combinations of the original features and might obscure some original relationships. TabNet's attention mechanism, which helps it focus on the most relevant features, was crucial here.

- Computational Resources: Given TabNet's complexity and resource-intensive nature, ensuring the model could be trained efficiently required careful management of computational resources. Running the model on GPU when available and adjusting the batch and virtual batch size parameters were practical solutions to manage memory usage and computational time (Sercan O. Arık, 2021).

5 Product Evaluation

5.1 Introduction

This section discussed four key evaluation metrics used to assess the 1D-CNN and TabNet model's performance. These are Accuracy, Precision, Recall and F1 Score.

It is important to note that the research objectives are to develop machine learning models that accurately detect cardiovascular disease (CVD) in diverse populations using augmented, large-scale data. Therefore, the paper aims to evaluate the model's accuracy across all predictions objectively, ensuring a balance between detecting all relevant cases and minimising false positives. The evaluation metrics—accuracy, precision, recall, and F1-score will help to optimise models for better healthcare outcomes and more efficient patient care.

5.2 Accuracy

In predictive modelling and machine learning, the accuracy metric occupies a foundational role, serving as a primary measure to evaluate the performance of classification algorithms. Defined mathematically as the proportion of true positive and true negative predictions to the total number of instances, accuracy is articulated through the formula:

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

TP, TN, FP, and FN denote the counts of true positives, true negatives, false positives, and false negatives, respectively. This metric encapsulates the overall correctness of a model's predictions, offering an aggregate assessment of its effectiveness in correctly classifying instances across both positive and negative classes. (Hossin, 2015).

5.3 Precision

Precision is a critical metric in evaluating classification models within machine learning and statistics, particularly in contexts where the cost of false positives is high. It quantifies the accuracy of a model's positive predictions, thereby assessing its ability to identify only relevant instances as positive. Mathematically, precision is defined as the ratio of true positive predictions to the total number of optimistic predictions made by the model, encapsulated by the formula:

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

TP represents true positives or correctly identified positive instances, and FP represents false positives or negative instances incorrectly classified as positive. (Hossin, 2015).

5.4 Recall

Recall, also known as sensitivity or actual positive rate, is a pivotal metric in evaluating classification models, especially in contexts where detecting all positive instances is critical. It measures the proportion of actual positives correctly identified by the model, thus assessing its capability to capture relevant instances without missing them. Mathematically, recall is defined by the formula:

$$\text{Recall} = \frac{TP}{TP + FN}$$

TP denotes the true positive, or the instances correctly identified as positive, and FN represents the false negatives or the positive instances the model incorrectly classifies as negative. (Hossin, 2015).

5.5 F1-Score

The F1 score is a crucial metric in classification models, providing a harmonised measure that balances the precision and recall of a model. It is precious when an equitable trade-off is desired between precision (the model's accuracy in predicting positive instances) and recall (the model’s ability to identify all actual positives). Mathematically, the F1-score is defined as the harmonic mean of precision and recall, offering a singular metric that encapsulates both aspects of model performance. The formula for calculating the F1-score is given by:

$$\text{F1 score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Precision is the ratio of true positives to the sum of true and false positives, and Recall is the ratio of true positives to the sum of true positives and false negatives. This formula ensures that the F1 Score only achieves a high value when both precision and recall are high, thus requiring the model to maintain a balance between these metrics. (Hossin, 2015).

6 Testing and Integration

6.1 Introduction

This part shows the outcomes of testing two different machine learning models, the 1D Convolutional Neural Network (1D-CNN) and TabNet, on three data sets: Cleveland, Hungary, and Switzerland. The models were used to identify heart disease. Several measures, such as accuracy, precision, recall, and F1 score, were used to judge how well these models worked. These metrics fully show how well the models can predict heart disease. Accuracy measures how accurate predictions are overall; precision measures how accurate positive predictions are; recall measures how well the model can find positive cases; and the F1 score balances precision and recall in a single metric.

The datasets used in this study—Cleveland, Hungarian, and Switzerland—are freely available. (Andras Janosi, 1988) And have been used extensively in medical computing to predict heart disease. The 1D-CNN model uses convolutional layers to understand the spatial order in structured data, making it a good choice for looking at sequential data. TabNet, a new deep learning model, on the other hand, makes choices by using sequential attention, which lets it focus on essential traits for classification tasks.

There are three data sets, and the tables show how well both models performed on each one. Each table shows the accuracy, precision, recall, and F1 score of the 1D-CNN and TabNet models to compare them.

6.2 Train-Test Split

Dataset	Split Ratio	Accuracy	Precision	Recall	F1-Score
Cleveland	50-50	0.82	0.81	0.82	0.81

	70-30	0.82	0.78	**0.86**	0.82
	80-20	**0.83**	**0.82**	0.81	**0.82**
Hungarian	50-50	0.89	0.90	0.85	0.87
	70-30	0.89	0.90	**0.86**	0.88
	80-20	**0.89**	**0.91**	0.85	**0.88**
Switzerland	50-50	0.92	0.93	0.98	0.96
	70-30	0.92	0.94	0.98	0.96
	80-20	0.92	0.93	0.98	0.96

Table 2: Split Ratio for 1D-CNN

Dataset	Split Ratio	Accuracy	Precision	Recall	F1-Score
Cleveland	50-50	0.82	0.81	0.81	0.81
	70-30	0.83	0.83	0.80	0.82
	80-20	**0.84**	**0.83**	**0.83**	**0.83**
Hungarian	50-50	0.89	0.90	0.85	0.87
	70-30	0.89	0.90	0.85	0.88
	80-20	0.89	0.91	0.85	0.88
Switzerland	50-50	0.92	0.93	0.98	0.96
	70-30	0.92	0.92	0.99	0.96
	80-20	0.92	0.93	0.99	0.96

Table 3: Split Ratio for TabNet

The paper aimed to identify the most effective data split ratio that maximises each model's accuracy, precision, recall, and F1 score. The 80-20 split ratio seems optimal for the 1D-CNN and TabNet models when applied to the Cleveland dataset. This split ratio resulted in an accuracy of 0.83 for the 1D-CNN model and 0.84 for the TabNet model. The TabNet model demonstrates a marginal performance improvement, most likely attributed to its greater capacity to utilise relational and time-series data adequately. Regarding the Hungarian dataset, both models demonstrated their highest level of performance when the data was split into an 80-20 ratio. In terms of precision, TabNet slightly surpassed 1D-CNN. The Switzerland dataset exhibits a distinctive scenario where both models demonstrate identical performance across all split ratios. This implies that the dataset's characteristics may be highly compatible with both model architectures or that the size of the dataset and the distribution of its features restrict the influence of the split ratio on model performance.

After thorough analysis, it is evident that the 80-20 split ratio consistently proves to be the most effective across all datasets and models. This ratio strikes a healthy compromise between the difficulty of training and the model's generalisation. The Cleveland dataset highlights the significance of selecting the suitable model for optimal outcomes; TabNet outperforms 1D-CNN in this split ratio. Furthermore, the consistent performance of both models on the Switzerland dataset, regardless of the divided ratios, suggests a possibility for improving the model parameters or architecture rather than manipulating the data split. This analysis highlights the intricate connection between data split ratios, model architecture, and dataset features, emphasising the need for a customised approach to training and evaluating models for heart disease prediction.

6.3 Optimisers

Dataset	Optimiser	Accuracy	Precision	Recall	F1-Score
Cleveland	Adam	0.84	0.84	0.83	0.83
	SGD	**0.84**	**0.85**	**0.83**	**0.84**
	AdaGrad	0.84	0.84	0.82	0.83
Hungarian	**Adam**	**0.90**	**0.87**	0.85	**0.86**
	SGD	0.90	0.85	**0.87**	0.86
	AdaGrad	0.87	0.90	0.81	0.86
Switzerland	Adam	0.91	0.92	0.98	0.95
	SGD	0.90	0.92	0.97	0.95
	AdaGrad	0.89	0.89	1.00	0.94

Table 4: Different Optimisers for 1D-CNN

Dataset	Optimiser	Accuracy	Precision	Recall	F1-Score
Cleveland	**Adam**	**0.84**	**0.85**	**0.81**	**0.83**
	SGD	0.83	0.84	0.80	0.82
	AdaGrad	0.82	0.85	0.76	0.81
Hungarian	**Adam**	**0.90**	**0.86**	**0.87**	**0.86**
	SGD	0.90	0.86	0.86	0.86
	AdaGrad	0.84	0.93	0.71	0.81
Switzerland	**Adam**	**0.91**	**0.92**	**0.99**	**0.95**
	SGD	0.91	0.91	0.99	0.95
	Adagrad	0.90	0.90	1.00	0.94

Table 5: Different Optimisers for TabNet

When evaluating the efficiency of different optimisers (Adam, SGD, AdaGrad) on the performance of 1D-CNN and TabNet models across different datasets (Cleveland, Hungarian, Switzerland), it is essential to consider accuracy, precision, recall, and F1-score metrics. These metrics help determine the most suitable optimiser for each model and dataset combination. In the Cleveland dataset, the Adam and SGD optimisers demonstrate similar performance in the 1D-CNN model, with a tiny advantage for SGD regarding the F1-score (0.84 vs. 0.83). This indicates that Stochastic Gradient Descent (SGD), despite its straightforwardness, continues to be efficient in effectively exploring the model's parameter space for this specific dataset. Within the Hungarian dataset, the 1D-CNN model demonstrates comparable performance between Adam and SGD. However, AdaGrad lags, notably in terms of accuracy and precision. This suggests that AdaGrad's adaptive learning rate mechanism may not be well-suited to the dataset's properties.

Regarding the Switzerland dataset, the Adam optimiser demonstrates exceptional performance for both models, exhibiting excellent proficiency in all measures, particularly in recall for the TabNet model (0.99). Adam's capacity to handle sparse gradients and dynamically modify learning rates demonstrates its great effectiveness for datasets with intricate patterns and imbalanced classes.

When evaluating the TabNet model, it is observed that the Adam optimiser consistently outperforms or performs equally well compared to the SGD and AdaGrad optimisers across all datasets and metrics. This is particularly evident in the accuracy and recall values for the Cleveland and Hungarian datasets. This highlights Adam's proficiency in managing tabular data and its capacity to rapidly converge towards a more optimal solution in intricate feature spaces.

Adam and SGD optimisers perform excellently across all datasets and models. However, Adam significantly surpasses SGD as the favoured optimiser due to its versatility and solid performance, particularly in complex datasets such as Switzerland. The adaptive learning rate mechanism and effective management of sparse gradients make it highly suitable for models like TabNet, which aim to capture complex patterns in data. Therefore, Adam is suggested as the most suitable optimiser for these models and datasets, effectively managing performance in terms of accuracy, precision, recall, and F1-score while guaranteeing strong capabilities in model training and prediction.

6.4 Input Shape

Dataset	Input Shape	Accuracy	Precision	Recall	F1-Score
Cleveland	**4**	**0.86**	0.85	0.83	**0.84**
	6	0.86	0.82	**0.86**	0.84
	8	0.85	**0.86**	0.79	0.82
Hungarian	4	0.90	0.88	0.84	0.86
	6	0.90	0.87	0.85	0.86
	8	**0.90**	**0.88**	**0.85**	**0.87**
Switzerland	4	0.90	0.92	0.98	0.95
	6	**0.91**	**0.92**	**0.99**	**0.95**
	8	0.90	0.92	0.99	0.95

Table 6: Different Input Shape for 1D-CNN

Dataset	Input Shape	Accuracy	Precision	Recall	F1-Score
Cleveland	4	0.87	0.84	0.85	0.85
	6	0.86	**0.86**	0.82	0.84
	8	**0.87**	0.84	**0.86**	**0.85**
Hungarian	4	0.90	0.86	**0.87**	0.87
	6	0.90	0.88	0.84	0.86
	8	**0.91**	**0.88**	0.86	**0.87**
Switzerland	**4**	**0.90**	**0.91**	0.99	**0.95**
	6	0.90	0.90	**1.00**	0.95
	8	0.90	0.90	0.99	0.95

Table 7: Different Input Shape for TabNet

Examining the influence of various input shapes on the efficacy of 1D-CNN and TabNet models across the Cleveland, Hungarian, and Switzerland datasets provides detailed insights into the most effective setup for predicting heart disease. The input shape, which refers to the number of dimensions in the input data, is essential for the model to learn and make generalisations effectively.

Depending on the datasets, the 1D-CNN model produces inconsistent outcomes when the input shapes are 4, 6, and 8. Both input forms 4 and 6 in the Cleveland dataset exhibit the same high accuracy of 0.86. However, there are modest differences in precision and recall, indicating that the input shape minimally affects the model's performance. The Hungarian dataset exhibits similar accuracy across all input shapes, with a little inclination towards an input shape of 8 in achieving a balance between precision and recall, resulting in an F1-score of 0.87. Similarly, in the case of the Switzerland dataset, an input shape of 6 slightly outperforms the other shapes in terms of recall and F1 score, suggesting a better trade-off between sensitivity and precision.

On the other hand, the TabNet model shows an apparent inclination for an input shape of 8 in both the Cleveland and Hungarian datasets. This choice leads to a slight enhancement in the F1 score and accuracy, specifically in the Hungarian dataset. This implies that the TabNet model could potentially gain advantages from having a higher number of dimensions. This could be because it can more effectively capture and utilise the interconnected information in larger input shapes. Nevertheless, with the dataset from Switzerland, all input forms yield similar performances across the measures, suggesting that the intricacy of the dataset may overshadow the impact of input shape.

Based on these results, the best input shape depends on the context and is influenced by both the dataset's properties and the model's design. Using an input shape of 6 in 1D-CNN provides a more optimal data distribution across datasets, especially when dealing with intricate patterns in the Switzerland dataset. TabNet demonstrates somewhat enhanced effectiveness with an input shape 8, particularly in the Cleveland and Hungarian datasets. This indicates its improved ability to utilise additional input information, resulting in better prediction performance. Hence, the selection of input shape should be guided by the dataset's unique attributes and the model architecture's intrinsic capabilities to achieve an optimal balance between accuracy, precision, recall, and F1-score.

6.5 Batch Size

Dataset	Batch Size	Accuracy	Precision	Recall	F1-Score
Cleveland	10	0.85	**0.86**	0.79	0.82
	15	**0.86**	0.85	0.83	**0.84**
	20	0.86	0.83	**0.84**	0.84
Hungarian	**10**	**0.90**	0.86	**0.87**	**0.87**
	15	0.90	**0.88**	0.85	0.86
	20	0.90	0.88	0.84	0.86
Switzerland	**10**	**0.90**	**0.92**	**0.98**	**0.95**
	15	0.90	0.92	0.98	0.95
	20	0.90	0.92	0.98	0.95

Table 8: Different Batch Sizes for 1D-CNN

Dataset	Batch Size	Accuracy	Precision	Recall	F1-Score
Cleveland	10	0.87	0.84	**0.86**	0.85
	15	0.87	0.84	0.85	0.85
	20	**0.87**	**0.85**	0.85	**0.85**
Hungarian	10	0.90	0.88	0.84	0.86
	15	**0.90**	**0.88**	**0.85**	**0.86**
	20	0.90	0.88	0.85	0.86
Switzerland	**10**	**0.90**	**0.91**	**0.99**	**0.95**
	15	0.90	0.90	0.99	0.95
	20	0.90	0.91	0.99	0.95

Table 9: Different Batch Sizes for TabNet

The examination of the effect of batch size on the performance metrics (accuracy, precision, recall, F1-score) of 1D-CNN and TabNet models on three datasets (Cleveland, Hungarian, Switzerland) reveals a subtle impact of batch size on both the training and evaluation of the models. The batch size is a crucial hyperparameter in training neural networks since it dictates the number of samples the network will process before updating its parameters. It impacts the model's generalisation capacity, learning process speed, and stability.

The 1D-CNN model performs best with a batch size 15 on the Cleveland dataset. This batch size achieves a balanced performance in terms of precision and recall, resulting in an F1 score of 0.84. These findings indicate that using a batch size in the middle range can provide a favourable trade-off between learning stability and the frequency of model updates, resulting in improved model generalisation from the training data. Within the Hungarian dataset, the accuracy remains similar regardless of the batch size. However, a batch size of 10 demonstrates a slight advantage in achieving a balance between recall and precision, as indicated by an F1-score of 0.87. This suggests that a smaller batch size may better manage the dataset's complexities through more frequent updates. The performance metrics for the Switzerland dataset remain consistent regardless of the batch size used. This indicates that the choice of batch size has minimal influence on the model's performance for this dataset. The dataset's size and complexity may overshadow the effects of batch size.

The TabNet model exhibits consistent performance across various batch sizes for all datasets, with low accuracy, precision, recall, and F1-score variation. The stability observed in TabNet's performance across different batch sizes may be attributed to its architectural features, which enable it to effectively gather and analyse relational and time-series data independent of the batch size.

Based on these facts, it can be concluded that no one batch size is optimal for all circumstances. However, a batch size of 15 for the 1D-CNN model in the Cleveland dataset and 10 for the same model in the Hungarian dataset appears to improve performance. Across all datasets, including the Switzerland dataset, the performance metrics for the TabNet model indicate that the batch size has a negligible effect on the model outcomes. This suggests that other factors, such as model architecture, learning rate, and dataset characteristics, may significantly influence these contexts. Hence, it is crucial to consider the batch size selection in combination with other hyperparameters and dataset-specific attributes to optimise model performance efficiently.

6.6 Final Model Implementation

Following the manual optimisation process above, a final network architecture was developed based on the best results observed, as shown in Table 9.

Model		Dataset	Accuracy	Precision	Recall	F1 Score
1D-CNN		Cleveland	0.84	0.81	0.83	0.82
TabNet			0.84	0.82	0.83	0.83
1D-CNN		Hungarian	0.91	0.89	0.85	0.87
TabNet			0.91	0.92	0.83	0.88
1D-CNN		Switzerland	0.90	0.91	0.98	94
TabNet			0.90	0.91	0.99	0.95

Table 10: Result of Final Model

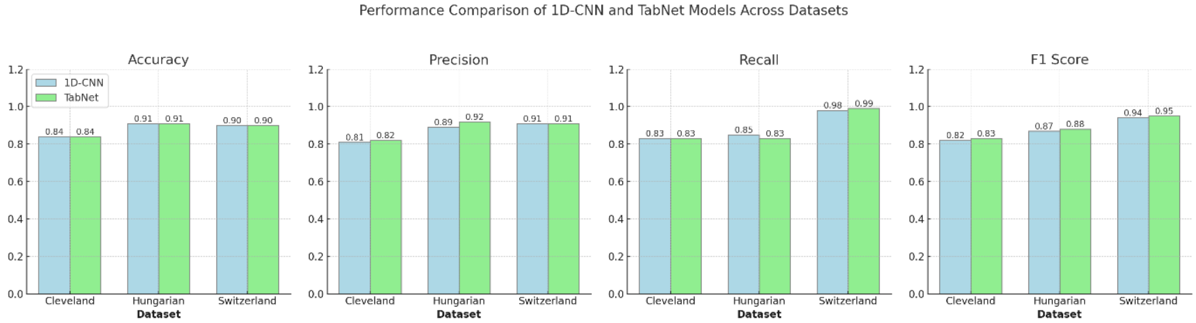


Figure 7: Result of Final Model

From the results presented, TabNet generally demonstrates a slight advantage over 1D-CNN in precision and F1 scores across all datasets, indicating its superior ability to minimise false positives and maintain a balanced approach between precision and recall. Specifically, TabNet achieves an F1 score of 0.83 on the Cleveland dataset compared to 1D-CNN's 0.82 and 0.88 on the Hungarian dataset against 0.87 from 1D-CNN. On the Switzerland dataset, both models show very high recall rates, but TabNet edges out with an F1 score of 0.95 to 1D-CNN's 0.94.

1D-CNN, however, exhibits a strong recall performance, which is particularly useful for ensuring actual cases of heart disease are not missed. This slight edge in recall suggests that while TabNet is better at confirming disease presence with fewer false positives, 1D-CNN is more effective at identifying all possible cases of heart disease.

In summary, TabNet slightly outperforms 1D-CNN overall, especially in minimising false positives and achieving slightly higher F1 scores, making it a marginally more effective model in these tests. Both models, however, provide high F1 scores and show great potential for use in healthcare settings to assist medical personnel in the accurate and reliable diagnosis of cardiac issues.

7 Conclusion

7.1 Recap of the Study

7.1.1 Assessment Against Objectives

The main aim of this work was to tackle the worldwide problem of predicting cardiovascular disease (CVD) by employing machine learning (ML) approaches that cover a wide range of demographics. The study aims to utilise augmented data to construct machine learning models that can accurately and robustly diagnose cardiovascular illness in diverse populations. The models must be flexible enough to accept numerous influencing factors and demographic variations. The comparative examination of the 1D Convolutional Neural Network (1D-CNN) and TabNet machine learning models on three datasets (Cleveland, Hungarian, and Switzerland) fully assess their effectiveness in predicting heart disease. The models were evaluated using accuracy, precision, recall, and F1 score, essential criteria for evaluating prediction performance in medical diagnostics.

The study successfully achieved its aims by proving that 1D-CNN and TabNet can achieve excellent performance across many parameters. This indicates that these models are robust and adaptable to different data representations and demographic factors. Evaluating performance under various situations, including train-test split ratios, optimisers, input shapes, and batch sizes, provided additional confirmation of the models' adaptability and the effectiveness of the optimisation tactics implemented. Furthermore, the ultimate implementation of the model, derived from the combined results of the manual optimisation process, demonstrated that both models exhibit high accuracy, precision, recall, and F1 scores across all datasets. The result is consistent with the study's objectives of creating dependable and comprehensible machine learning models for predicting heart disease in various populations, strengthening their potential usefulness globally in clinical settings.

7.1.2 Key Contributions

This study provides substantial advancements in heart disease prediction through the utilisation of deep learning techniques, contributing to both the methodology and comprehension of the subject. An important finding is that the 1D-CNN and TabNet models effectively handle the complexity and variability seen in heart disease data from various demographic groups. This study emphasises the significance of thorough dataset augmentation and precise hyperparameter tuning to improve the model's performance. The paper is innovative in systematically comparing 1D-CNN and TabNet models on several datasets while considering different operational factors. This comparative method has yielded fresh insights into selecting and optimising models to predict cardiac disease. The research highlights the crucial significance of data split ratios, optimisers, input shapes, and batch sizes in enhancing model performance, providing significant recommendations for future studies in medical machine learning applications.

Moreover, this research adds to the broader discussion on the ethical application of artificial intelligence in healthcare, specifically in guaranteeing that machine learning models are efficient for various populations and adhere to regulatory regulations. Considering these factors, the study establishes a model for future research focused on the moral advancement and use of artificial intelligence in global health settings. Overall, this study greatly enhances the comprehension of machine learning (ML) use in predicting cardiac illness. It introduces novel approaches and perspectives that facilitate future investigations and advancements in this critical healthcare field. Thoroughly analysing and optimising machine learning models actively contributes to continuous endeavours to enhance patient outcomes through early identification and personalised treatment options. This highlights the potential of machine learning to transform healthcare delivery for cardiovascular illnesses significantly.

7.2 Reflection on the Study

7.2.1 Challenges and Limitations

During this work, various problems and constraints arose that highlight the inherent difficulty of using machine learning techniques to make predictions in healthcare, particularly for a widespread and diverse ailment like cardiovascular disease (CVD). A significant obstacle was obtaining and expanding a wide range of datasets that accurately reflect the global population. Although the study employed three datasets (Cleveland, Hungarian, and Switzerland), it is essential to note that they, while commonly utilised, only capture a small portion of the diverse demographic characteristics seen in populations worldwide. This constraint highlights a broader problem in medical data science: the lack of extensive and varied datasets caused by concerns about privacy, ethical considerations, and logistical challenges in collecting and distributing health data.

Furthermore, the study encountered constraints in the applicability of its results. Although the models prioritise generalisation across multiple datasets, it could be done at the expense of specialised optimisations that could yield higher accuracy on specific datasets. These may happen because of the noise data. The outcomes may differ when implemented on alternative datasets or with varying operational parameters, emphasising the need for ongoing validation and adjustment of machine learning models in novel situations.

7.2.2 Learnings

This work has produced crucial insights, particularly in data management, model development and assessment, and the broader incorporation of machine learning in healthcare forecasting.

Data quality, variety, and representation are essential in healthcare predictions. This study emphasised the importance of thorough data preparation, which involves cleaning, normalising, and augmenting the data. This process is crucial for improving model training and ensuring that the resulting models can be applied to various demographic groups. The study emphasised the intricate equilibrium necessary in selecting, training, and evaluating models. It demonstrated the efficacy of both 1D-CNN and TabNet models in accurately predicting heart illness, highlighting the versatility of utilising different model architectures according to the individual needs of the job. Furthermore, it highlighted the importance of using a wide range of assessment metrics (such as accuracy, precision, recall, and F1 score) to evaluate a model's performance from many angles properly.

The most crucial insight gained from this study is the potential influence of Machine Learning in enhancing healthcare results by accurately predicting diseases such as CVD early on. The research demonstrates the practicality of incorporating machine learning models into clinical environments, indicating a potential future where artificial intelligence could significantly enhance human proficiency in diagnosing and treating diseases. Nevertheless, it emphasises the importance of tackling ethical, privacy, and symbolic issues to guarantee that new technologies provide fair advantages to all population sectors. To summarise, although the study encountered various difficulties and restrictions, the insights it yielded have significant ramifications for the future of machine learning in the healthcare sector. The knowledge acquired from overcoming these obstacles provides substantial understanding in the continuous endeavours to utilise AI's potential for improving worldwide health results.

8 Future Work

8.1 Advanced Model Explainability and Fairness Techniques

In healthcare, deploying machine learning (ML) models demands high accuracy, reliability, and a solid commitment to fairness and transparency. This is especially crucial given the potential impact of these models on patient outcomes and treatment decisions. Ensuring that ML models do not inadvertently perpetuate existing biases or introduce new ones is essential for maintaining trust and ethical integrity in medical AI applications.

Advanced explainability techniques should be integrated into the development and deployment phases to enhance the transparency of ML models. Techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) provide valuable insights into the decision-making processes of complex models. SHAP values, derived from game theory, offer a unified measure of feature importance and how each feature contributes to the prediction fairly and consistently (Lundberg, 2017). LIME, on the other hand, approximates the predictions of any classifier in an interpretable and faithful manner by perturbing the input data and observing the changes in forecasts (Ribeiro, 2016). These techniques help demystify the model's operations, making them more accessible and understandable to stakeholders, including clinicians, patients, and regulatory bodies. Alongside explainability, implementing algorithms specifically designed to detect and mitigate biases is critical. These algorithms analyse training data and model outputs to identify patterns of bias—such as those based on race, gender, or socioeconomic status—and adjust the model accordingly to neutralise these biases. This process not only helps enhance the fairness of the models but also improves their accuracy and generalizability by ensuring they perform well across diverse demographic groups. Monitoring model performance and impact is necessary to maintain and enhance model fairness and transparency. This involves regularly updating the model with new data, reassessing its predictions for bias, and refining the explainability methods to keep pace with advancements in AI and ML technologies. Such ongoing evaluations help adapt to population demographics, disease profiles, and healthcare practices, ensuring the model remains relevant and fair.

Achieving advanced model explainability and fairness also requires collaboration across multiple disciplines, including data science, clinical medicine, ethics, and law. Developing and adhering to regulatory frameworks that mandate transparency and fairness in AI applications can guide and standardise efforts in this direction. These frameworks should encourage innovation while safeguarding against AI technologies' misuse or unintended consequences in healthcare. By expanding the use of advanced explainability and fairness techniques in developing and applying machine learning models in healthcare, researchers and practitioners can ensure these models are both practical and equitable. This approach enhances the acceptability and trustworthiness of AI technologies in medical settings and contributes to more personalised and equitable healthcare solutions. As such, prioritising transparency and fairness is not merely a technical requirement but a fundamental ethical imperative in the continued integration of AI into healthcare practices.

8.2 Closing Remarks

This study underscores the transformative potential of machine learning (ML) in healthcare, particularly in the predictive diagnosis of cardiovascular diseases (CVD). By rigorously analysing and comparing the performance of two advanced ML models, the 1D Convolutional Neural Network (1D-CNN) and TabNet, this research contributes to a refined understanding of how these technologies can be tailored to meet diverse global healthcare needs effectively. The findings of this research not only demonstrate the high potential of these models in achieving diagnostic accuracy but highlight the critical importance of model adaptability and robustness across different demographic settings.

The systematic comparison of the 1D-CNN and TabNet across multiple datasets sheds light on the practical challenges and the immense possibilities of using ML to enhance diagnostic processes. This study is a testament to the strength of deep learning in navigating the complexities of medical data, offering insights that could lead to more personalised, efficient, and proactive healthcare interventions. As it advances, it must continue to refine these technologies, ensuring they are accessible and equitable, amplifying their global impact on medical diagnostics.

Considering the substantial advancements made by this study, continued exploration in this area is not just advisable but necessary. It can potentially revolutionise healthcare outcomes by providing more accurate, timely, and democratised predictions of cardiovascular diseases. As such, the research community should remain diligent in pushing the boundaries of what artificial intelligence can achieve in healthcare, striving to turn these advanced predictive capabilities into everyday clinical practices that save lives and enhance the quality of care for patients worldwide.

By championing these technologies and addressing the inherent challenges head-on, we can harness the full power of machine learning to create a future where predictive healthcare is not only a vision but a reality.

References

1. Adlung, L. a. C. Y. a. M. U. a. E. E., 2021. Machine learning in clinical decision making. *Med,* Volume 2, pp. 642-665.

2. Alanazi, A., 2022. Using machine learning for healthcare challenges and opportunities. *ScienceDirect,* Volume 30.

3. Andras Janosi, W. S. M. P. R. D., 1988. *UC Irvine Machine Learning Repository.* [Online]
   Available at: https://archive.ics.uci.edu/dataset/45/heart+disease
   [Accessed 01 04 2024].

4. Anon., 2020. *Pypi.* [Online]
   Available at: https://pypi.org/project/ctgan/0.2.2/
   [Accessed 02 04 2024].

5. Anon., 2021. *World Health Organization.* [Online]
   Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
   [Accessed 20 03 2024].

6. Anon., 2023. *NumPy.* [Online]
   Available at: https://numpy.org/
   [Accessed 02 04 2024].

7. Anon., 2023. *PyPi.* [Online]
   Available at: https://pypi.org/project/table-evaluator/
   [Accessed 02 04 2024].

8. Anon., 2024. *Pandas.* [Online]
   Available at: https://pandas.pydata.org/
   [Accessed 02 04 2024].

9. Anon., 2024. *PyTorch.* [Online]
   Available at: https://pytorch.org/
   [Accessed 02 04 2024].

10. Anon., 2024. *Scikit-learn.* [Online]
    Available at: https://scikit-learn.org/stable/
    [Accessed 02 04 2024].

11. Anon., 2024. *TensorFlow.* [Online]
    Available at: https://www.tensorflow.org/
    [Accessed 02 04 2024].

12. Anon., n.d. *UCSFHealth.* [Online]
    Available at: https://www.ucsfhealth.org/education/diagnosing-heart-disease#
    [Accessed 21 03 2024].

13. AYESHA SIDDIQUA DINA, A. B. S. A. D. M., 2022. Effect of Balancing Data Using Synthetic Data on the Performance of Machine Learning Classifiers for Intrusion Detection in Computer Networks. *IEEE Access,* Volume 10, pp. 96731-96747.

14. AZAM MEHMOOD QADRI, A. R. K. M. A. M. S. A., 2023. Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning. *IEEE Access,* Volume 11, pp. 56214-56224.

15. Bárbara Martins, D. F. C. N. A. A. &. J. M., 2021. Data Mining for Cardiovascular Disease Prediction. *Journal of Medical Systems,* Volume 45.

16. Bharadiya, J. P., 2023. A Tutorial on Principal Component Analysis for Dimensionality Reduction in Machine Learning. *International Journal of Innovative Science and Research Technology ,* 8(5), pp. 2028-2032.

17. Ch. Sanjeev Kumar Dash, A. K. B. S. D. A. G., 2023. An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal,* Volume 6.

18. Danny Hartanto Djarum, Z. A. &. J. Z., 2021. *Comparing Different Pre-processing Techniques and Machine Learning Models to Predict PM10 and PM2.5 Concentration in Malaysia.* Singapore, Springer.

19. Hossin, M. a. S. M., 2015. A REVIEW ON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATION. *International Journal of Data Mining & Knowledge Management Process,* Volume 5.

20. Lundberg, S. M. a. L. S.-I., 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems,* Volume 30.

21. Manali Saini, U. S. M. D. U., 2022. One-dimensional convolutional neural network architecture for classification of mental tasks from electroencephalogram. *Biomedical Signal Processing and Control,* Volume 74, p. 103494.

22. Mariachiara Di Cesare, H. B. T. G. L. H. C. K. D. V. M. J. M. B. P. P. P. D. P. S. T. F. P., 2023. *World Heart Report - 2023,* Geneva: World Heart Federation.

23. Martins, A. a. A. R., 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In: *International conference on machine learning.* s.l.:PMLR, pp. 1614--1623.

24. Metkus, T. S., Hopkins, J., Zieve, D. & Conaway, B., 2022. *MedlinePlus.* [Online] Available at: https://medlineplus.gov/ency/patientinstructions/000759.htm#:~:text=Coronary%20heart%20disease%20(CHD)%20is,can%20cause%20a%20heart%20attack. [Accessed 20 03 2024].

25. Mohamed G. El-Shafiey, A. H. E.-S. A. E.-D. &. M. A. I., 2022. A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest. *Multimedia Tools and Applications,* Volume 81, pp. 18155-18179.

26. Morgenstern, J. D. a. B. E. a. O. M. a. P. T. a. G. V. a. F. D. a. K. K. a. R. L. C., 2020. Predicting population health with machine learning: a scoping review. *BMJ open,* Volume 10, p. e037860.

27. Oguz Akbilgic, L. B. I. K. P. P. C. D. W. K. A. A. L. Y. C. a. E. Z. S., 2021. ECG-AI: electrocardiographic artificial intelligence model for prediction of heart failure. *European Heart Journal - Digital Health,* Volume 2, pp. 626-634.

28. PRONAB GHOSH, S. A. M. J. A. K. F. M. J. M. S. E. I. S. S. A. R. B. A. F. D. B., 2021. Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques. *IEEE Access,* Volume 9, pp. 19304-19326.

29. Rahman, M. M. & Davis, D. N., 2013. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing,* Volume 3, p. 224.

30. Raniya R. Sarra, A. M. D. M. A. M. M. K. A. G. a. M. A. A., 2022. A Robust Framework for Data Generative and Heart Disease Prediction Based on Efficient Deep Learning Models. *Diagnostics,* 12(12).

31. Ribeiro, M. T. a. S. S. a. G. C., 2016. Why should i trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* s.l.:s.n., pp. 1135--1144.

32. Rout, N. V. M. &. R. K., 2023. Effective heart disease prediction using improved particle swarm optimization algorithm and ensemble classification technique. *Soft Computing,* Volume 27, pp. 11027-11040.

33. Sercan O. Arık, T. P., 2021. TabNet: Attentive Interpretable Tabular Learning. *Proceedings of the AAAI Conference on Artificial Intelligence,* Volume 35.

34. Serkan Kiranyaz, O. A. O. A. T. I. M. G. D. J. I., 2021. 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing,* Volume 151.

Appendix A

The appendix presents the code snippets of the main model of 1D-CNN and TabNet.

1D-CNN

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Conv1D, MaxPooling1D, Flatten,
Dense, Dropout


# Define the 1D-CNN model
model = Sequential([
    Conv1D(filters=64, kernel_size=2, activation='relu',
input_shape=(6, 1)),
    MaxPooling1D(pool_size=1),
    Dropout(0.2),
    Conv1D(filters=128, kernel_size=2, activation='relu'),
    Dropout(0.2),
    Flatten(),
    Dense(64, activation='relu'),
    Dropout(0.2),
    Dense(1, activation='sigmoid')
])

# Compile the model
model.compile(optimizer='Adam', loss='binary_crossentropy',
metrics=['accuracy'])

# Train the model
history = model.fit(X1_train, y_train, epochs=20, validation_split=0.2,
batch_size=15)
```

TabNet

```
!pip install pytorch-tabnet
from pytorch_tabnet.tab_model import TabNetClassifier
import torch

# Set device to GPU if available, else CPU
device = 'cuda' if torch.cuda.is_available() else 'cpu'

# Initialize TabNetClassifier
clf = TabNetClassifier(optimizer_fn=torch.optim.Adam,
                       optimizer_params=dict(lr=1e-3),
                       scheduler_params={"step_size":100, # how to
update learning rate
                                         "gamma":0.9},
                       scheduler_fn=torch.optim.lr_scheduler.StepLR,
                       mask_type='entmax', # "sparsemax"
                       device_name=device)

# Fit the model on the training data
clf.fit(X_train_pca, y_train,
        eval_set=[(X_test_pca, y_test)],
        eval_name=['test'],
        eval_metric=['accuracy'],
        max_epochs=25,
        patience=50,
        batch_size=15,
        virtual_batch_size=8,
        num_workers=0,
        drop_last=False)
```