Addis Ababa University

COLLEGE OF TECHNOLOGY AND BUILT ENVIRONMENT

SCHOOL OF BIOMEDICAL ENGINEERING

PREDICTING HEART-ATTACK RISK USING MACHINE- AND DEEP-LEARNING METHODS IN ETHIOPIA

by

Emirt Worku

In partial fulfillment of the requirements for the degree of Master of Science in Biomedical Engineering

Supervisors

Dr. Melkamu Hunegnaw Asmare (PhD)

Dr. Dejuma Yadeta Goshu (Cardiologist)

Addis Ababa, Ethiopia

November 9, 2025

Declaration

I hereby declare that this study is original and has not been submitted for any other degree award to any other university. It is a true copy of the thesis, including any required final revisions, as accepted by my examiners. Permission is herewith granted to Addis Ababa Institute of Technology, Center of Biomedical Engineering, Addis Ababa University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

Name: Emirt Worku **Signature:** .... **Date:** ....

This MSc thesis has been submitted for examination with my approval as the main advisor.

Name: Dr. Melkamu Hunegnaw **Signature:** .... **Date:**.............................

Approval

We the undersigned committee hereby approved that we have read and recommended to the office of graduate Program for acceptance a dissertation entitled “Predicting Heart-Attack Risk Using Machine- and Deep-Learning Methods in Ethiopia” by Emirt Worku in partial fulfilment for the requirements for the Degree of Master of Science in Biomedical Engineering (Bio-instrumentation and imaging) complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Approved by Board of Examiners:

Supervisor’s Name Dr. Melkamu Hunegnaw *Sign:*___________ *Date:*_________________

Co-supervisor’s Name Dr. Dejuma Yadeta *Sign:*___________ *Date:*_________________

External Examiner’s Name Dr.Fantahun Bogale *Sign:*___________ *Date:*_________________

Internal Examiner’s Name Dr.Masreshaw Demelash *Sign:*___________ *Date:*_________________

Chairman’s Name Mr.Bemnet Zelalem *Sign:*___________ *Date:*_________________

Dedication

To my mother for her love, sacrifices, and dreams that I strive to fulfil.

Acknowledgements

First and foremost, I would like to thank GOD for His boundless grace and unwavering support throughout my life. His blessings have sustained me through every challenge and gifted me more than I could ever deserve.

I am deeply grateful to my advisors, Dr. Melkamu H. Asmare and Dr. Dejuma Y. Goshu, for their exceptional guidance, mentorship, and encouragement at every stage of this work. Their insightful advice and continuous support were instrumental in shaping my academic journey as well as the outcome of this research.

I extend my sincere thanks to all individuals and institutions who contributed to this work in one way or another. Whether through academic discussions, technical support, or moral encouragement, your help has been invaluable.

My heartfelt appreciation also goes to my family, whose love has been my greatest source of strength. I am especially thankful to my mother, whose unconditional love and sacrifices have made this journey possible, and to my husband, for his unwavering support, patience, and belief in me.

Finally, I would like to acknowledge my friends, whose motivation, companionship, and consistent encouragement made even the most difficult moments easier to bear. Your presence has made this journey truly meaningful.

Abstract

Heart attack, one of the most severe forms of cardiovascular disease (CVD), remains a leading cause of mortality worldwide. In 2022, CVD accounted for an estimated 19.8 million deaths, representing 32% of all global deaths, with a growing burden in low- and middle-income countries such as Ethiopia. Major risk factors contributing to heart attack include high blood pressure, high cholesterol, diabetes, smoking, obesity, poor diet, and physical inactivity. The early detection and prediction of heart attack risk are critical in reducing mortality.

The purpose of the thesis is to predict risk of heart attack in the human body and to provide suggestions to individuals to reduce the risk in the future. In this thesis, we develop a model using machine learning (ML) and deep learning (DL) using a public data set obtained from the Behavior Risk Factor Surveillance System (BRFSS), the world's largest continuously conducted health survey system, with the help of the Centers for Disease Control and Prevention (CDC) and secondary data from Tikur Anbessa Specialized Hospital (TASH) that include clinical data and demographic data from the person.

We evaluated a wide range of classical ML models, including logistic regression (LR),K-Nearest Neighbours (KNN), Decision trees (DT), Random Forest (RF), and Gradient Boosting (GB) to assess performance compared to DL models. The performance of each ML algorithm was evaluated using cross-validation techniques and standardized metrics to ensure reliability. We also evaluate several DL models, including the Feedforward Neural Network (FNN), Wide & Deep model, Residual Network, and Attention-based model.The diversity of DL models explored in our study allows one to capture complex, nonlinear relationships within health data.

The public availability of large-scale health data has allowed us to develop computational techniques to improve medical diagnostics and screen the high-risk patient. In this context, our work contributes not only to the development of predictive modeling but also to the development of the graphical user interface (GUI) application, which is designed to be accessible for clinical use and aims to support healthcare professionals.

Finally,We found that most DL models achieved fairly similar performance, with the best results showing balanced accuracy (95%), precision (0.95), recall (0.96), F1 score (0.95), and AUC (0.97) in the FNN. This highlights the promise of DL approaches in advancing early diagnosis and personalized care for CVD. Based on these results, the FNN model was integrated into the development of a user-friendly GUI application for the prediction of CVD risk and decision support in real time.

Keywords: Cardiovascular disease , Machine learning , Deep learning , Models , prediction , Ethiopia

Contents

List of Figures

List of Tables

List of Acronyms

AAiT	Addis Ababa Institute of Technology
AAU	Addis Ababa University
AdaB	AdaBoost
AHA	American Heart Association
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
ASCVD	The Atherosclerotic Cardiovascular Disease
AUC	Area Under the Curve
BME	Biomedical Engineering
BMI	Body Mass Index
BRFSS	Behavioural Risk Factor Surveillance System
CAD	Coronary Artery Disease
CDC	Center of Disease Control and Prevention
CDSS	Clinical Decision Support System
CHD	Congenital Heart Diseases
CNN	Convolutional Neural Network
CT	Computed Tomography
CVDs	Cardiovascular diseases
DL	Deep Learning
DNN	Deep Neural Network
DT	Decision Trees
ECGs	Electrocardiograms
ESC	European Society of Cardiology
ET	ExtraTrees
FN	False Negative
FNN	Feedforward Neural Network
FP	False Positive
FRS	Framingham Risk Score
GB	Gradient Boosting
GBD	Global Burden of Disease
GUI	Graphical User Interface
HF	Heart Failure
IQR	Interquartile Range
ISH	International Society for Hypertension

KNN	K-Nearest Neighbors
LDL	Low-Density Lipoprotein
LMICs	Low- and Middle-Income Countries
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multi-Layer Perceptrons
MRI	Magnetic Resonance Imaging
NCD	Non-Communicable Diseases
NFR	Novel Feature Reduction
NN	Neural Networks
RF	Random Forest
ROC	Receiver Operating Characteristics
SAE	Sparse Autoencoder
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TASH	Tikur Anbessa Specialized Hospital
TIA	Transient Ischemic Attacks
TN	True Negative
TP	True Positive
UCI	University of California, Irvine
UK	United Kingdom
US	United States
WHO	World Health Organization
XGB	XGBoost

Symbols

μ	mean of variables
σ	standard deviation

Chapter 1

Introduction

The human heart is a muscular organ responsible for pumping blood throughout the body, maintaining oxygen circulation, and transporting nutrients to tissues. Figure 1.1 shows that the human circulatory system of the heart has four chambers, namely; the right atrium (receives deoxygenated blood from the body), the right ventricle (pumps deoxygenated blood to the lungs) the left atrium (receives oxygenated blood from the lungs), and the left ventricle (pumps oxygenated blood to the body). All chambers work in a coordinated cycle in the form of contraction and relaxation to ensure continuous and efficient blood flow. The proper functioning of the heart system is vital for maintaining life and supporting the metabolic demands of the body. However, disruptions in the heart functioning process can lead to serious health conditions. Cardiovascular diseases (CVDs) are a group of conditions that occur when the arteries become blocked or narrow, leading to conditions such as coronary artery disease, heart attacks, and heart failure[1]. We discuss the background and motivation study of heart attack risk in Sections 1.1 and 1.3, respectively. We also presented the significance of this thesis study in Section 1.4, and a general overview of machine learning (ML) and deep learning (DL) has been used to improve our ability to detect patterns in large-scale health data in Section 1.5. In Section 1.6, we mentioned the general and specific objectives of the thesis, and the general structure of the thesis is given in Section 1.7.

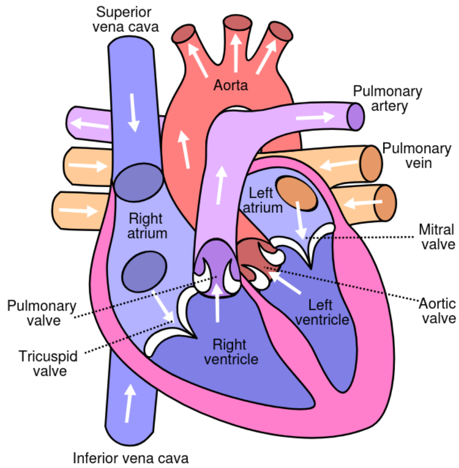

Figure 1.1: Diagram of Human Heart Circulatory System[2].

1.1 Background of the study

Heart attack, one of the most severe forms of cardiovascular disease (CVD),which is one of the non-communicable diseases (NCDs), and the report shows that it is the leading cause of death for millions of people around the globe. Figures 1.2 and 1.3 show CVD deaths reports from "our world data"[3] between 1980 and 2021 in the world and Ethiopia, respectively. As shown in the figures, the number of CVD deaths has increased significantly both globally and in Ethiopia. However, CVD deaths in Ethiopia decreased slightly around 2008, although there is no clear study explaining the reason behind this decline.

[1] https://my.clevelandclinic.org/health/body/21704-heart

[2] https://commons.wikimedia.org/wiki/File:Diagram_of_the_human_heart.svg

[3] https://ourworldindata.org/grapher/deaths-from-cardiovascular-disease?tab=line&time=earliest..2021&country=ETH~OWID_WRL

As shown in Figure 1.3, the number of deaths from CVDs in Ethiopia increased by 5 million between 2015 and 2021, indicating that CVD is a major public health concern.

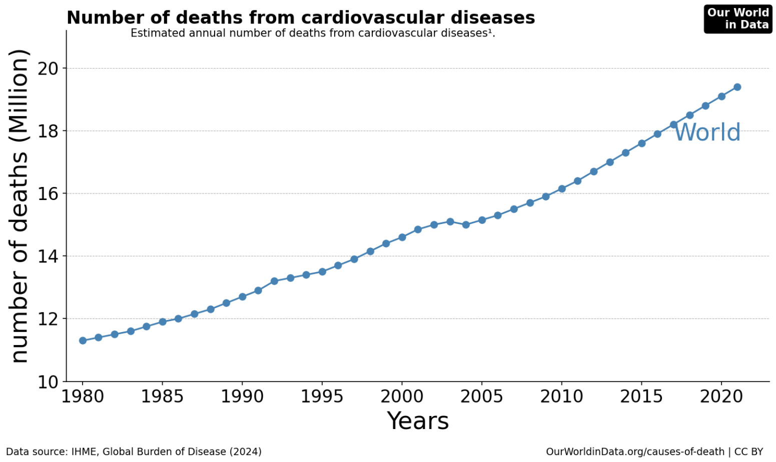

Figure 1.2: The number of deaths from CVD around the globe reported by Global Burden of Disease (GBD).

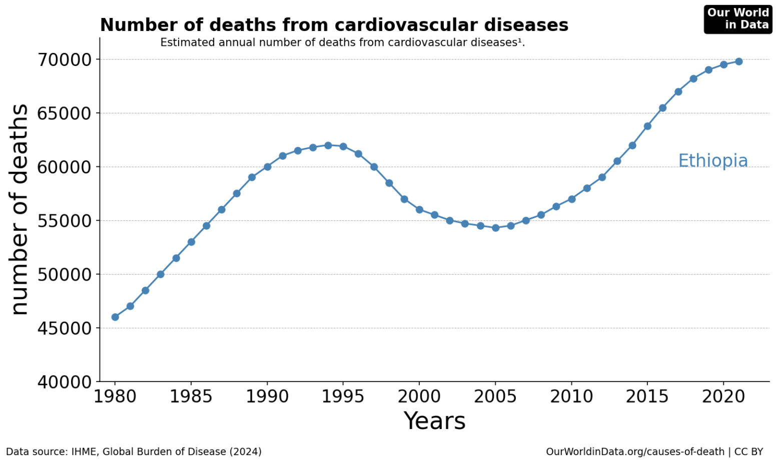

Figure 1.3: The number of deaths from CVD in Ethiopia reported by GBD.

Studies indicate that risk factors for heart attack include sex, cholesterol, high blood pressure, diabetes, smoking, physical inactivity, BIM status, diet, excessive alcohol consumption, and other factors that contribute to heart attack risk [4]. According to the World Health Organization (WHO), the number of CVD deaths is projected to increase to 23.6 million by 2030 [1]. This alarming trend underscores the urgent need for improved strategies for predicting and preventing heart diseases. Predicting heart attack risk is essential for early detection and effective intervention. Identifying high-risk individuals allows timely actions, such as lifestyle changes, medication, and regular monitoring, which can prevent disease progression. Early detection not only improves patient outcomes but also reduces the strain on healthcare systems by mitigating the need for intensive treatments in advanced stages. Fundamentally, improved heart attack risk prediction has the potential to save millions of lives worldwide.

1.2 Statement of the Problem

The leading causes of CVDs include heart attacks; they are one of the major global health challenges and cause millions of deaths annually. In Ethiopia, over the years, the burden of the disease has been increasing due to rapid urbanization and changing dietary patterns, sedentary lifestyles, and an increase

[4] https://ada.com/cardiovascular/cardiovascular-disease-risk-factors/

in non-communicable diseases such as hypertension and diabetes. Cardiovascular conditions usually attract little attention compared with the infectious diseases that have dominated public health programs for years and thus result in late diagnosis and limited preventive care. In Ethiopia, many patients are diagnosed only at a stage when severe symptoms have already manifested, which provides limited and very expensive treatment options. This situation has contributed to an increased number of premature deaths and imposes a heavy burden on the healthcare system. This becomes particularly difficult in the Ethiopian population due to the involvement of several interlinked risk factors, such as age, blood pressure, cholesterol level, diabetes, obesity, smoking, and physical inactivity. Most of the heart attack risk prediction models developed so far have been based on datasets collected from high-income countries with very different population characteristics, genetic profiles, environmental conditions, and health systems.

These models may not perform well or capture local risk patterns well in Ethiopia. A significant limitation to the development of population-specific predictive tools in Ethiopia is the scarcity of well-structured, comprehensive health datasets. This gap in the literature necessitates an urgent need for developing a locally valid, robust system that can apply machine learning and deep learning for the prediction of heart attack risk using Ethiopian data. In this regard, the development of such a model would enable the early identification of high-risk individuals, facilitate clinical decision-making, and hence significantly reduce cardiovascular morbidity and mortality across the country.

1.3 Research motivation

Given the growing challenges posed by Heart Attack in both Ethiopia and the world. We are motivated to contribute to and develop a model that can effectively predict heart attack risk and develop an app graphical user interface (GUI). Our goal is to explore which models can most effectively predict heart attack risk and identify key risk factors that improve the applicability of clinical diagnosis and prevention. Through the availability of a large data source for the public, this work aims to improve early detection and prevention of heart attacks by testing with local data, ultimately improving patient outcomes and healthcare efficiency. Early prediction of heart attack risk is crucial to implement preventive measures and reduce mortality rates. ML and DL techniques have shown promise in the analysis of complex medical data to predict the risk of heart attacks. However, most existing models are developed using data from high-income countries, which are not tested with low-income country local data, limiting their applicability in countries such as Ethiopia. There is a pressing need to develop and validate predictive models based on ML and to validate local data to improve their relevance and effectiveness.

1.4 Significance of the study

Heart attack risk is a growing public health concern in Ethiopia and contributes significantly to morbidity and mortality. Early detection is essential to reduce the progression of these conditions and mitigate the risk of life-threatening complications. Using early diagnosis to allow for timely intervention, early diagnosis can improve patient outcomes and reduce healthcare costs. Predicting heart disease through advanced techniques is therefore crucial to improving preventive care, lowering mortality rates, and ensuring the efficient use of healthcare resources, including the delivery of personalized treatment strategies. Currently, the diagnosis of CVD in Ethiopia is hindered by the limited availability of advanced diagnostic tools such as cardiac catheterization, magnetic resonance imaging (MRI), or computed tomography (CT), which are often only accessible in cities and a few hospitals. Standard diagnostic tests such as blood tests, ECGs, and chest radiographs are used more widely but may not always be sufficient for the early detection or prediction of heart attack risk. With the increasing incidence of heart attack risk, it is critical to develop efficient and affordable methods to predict heart attack risk, tailored to the unique healthcare context of Ethiopia. Factors such as limited access to specialized health professionals, inadequate diagnostic facilities, and a growing prevalence of risk factors such as hypertension, diabetes, and smoking exacerbate the situation.

This study aims to develop a predictive model based on ML and DL, specifically validated with the local Ethiopian population, and to create a GUI for screening and clinical decision support. This model has the potential to support healthcare providers in identifying high-risk individuals early and making informed, data-driven decisions. By integrating international health data and contextual factors, the model can support the optimization of clinical workflows and resource distribution. Ultimately, the successful implementation of this predictive approach can contribute to reducing the national burden of CVD and support the advancement of precision medicine in Ethiopia.

1.5 Machine- and deep-learning in healthcare

Advancements in computational techniques and artificial intelligence enable more accurate diagnoses and assessment plans, enhancing the efficiency and precision of modern healthcare. ML and DL models can process large and complex data sets, uncover hidden patterns, and improve diagnostic accuracy, and the models used optimise predictions, increase reliability, and facilitate early detection of heart attack risk by analysing clinical, demographic, behavioural, and ECG data [2–7]. For more than dadeds, there are different classical ML models that are used for heart attack risk prediction such as logistic regression (LR) and decision trees (DT), naive Bayes (NB), random forest (RF), k-nearest neighbours (KNN), gradient boosting, and others that used for accurate and efficient predictive systems [8, 9]. In addition, different DL models, also used for heart attack risk prediction, for example, Convolutional Neural Network (CNN), Feedforward Neural Network (FNN), Wide-Deep, Residual, and Attention, perform prediction by learning a function that maps input features to outputs through layers of transformations using trained parameters. These systems not only help healthcare professionals diagnose heart attack risk but also enable individualised treatment planning and proactive health management, ultimately reducing the global burden of these life-threatening conditions.

In Ethiopia, where healthcare resources are limited, the use of predictive models based on ML and DL could support early diagnosis and improve clinical decision-making. However, the adoption of ML for heart attack risk prediction remains underexplored in the Ethiopian context. This highlights the need for localised research that considers the unique health determinants and challenges in data availability within the country.

Using clinical and lifestyle-related characteristics from available datasets, the study assesses the feasibility and effectiveness of ML and DL techniques to identify people at high risk of heart attacks and a predictive model tailored to the Ethiopian population [10]. Such kind of findings will offer valuable information to healthcare professionals and policymakers, supporting the integration of AI-driven approaches into the Ethiopian healthcare system for better disease prevention and management.

The WHO reported that NCDs have become a major health concern in Ethiopia; the mortality rate for CVDs, chronic respiratory diseases, cancers, and diabetes is estimated to be 631 per 100,000 men and 549 per 100,000 women. Among these, heart attacks are a leading cause of death[5].A national strategic plan indicates that a 10-year risk of heart disease is $\geq 30\%$ prevalence. of[6]. Angaw studied that heart disease was the leading cause of death and a common reason for hospital admission in Ethiopia based on data from the Tikur Anbessa and St. Paul hospitals[11].

The increasing prevalence of heart disease in Ethiopia, combined with limited healthcare resources, underscores the urgent need for predictive tools to support early detection and intervention. ML and DL offer promising avenues for building accurate data-driven models for early heart attack risk detection and prevention, which is crucial. These models can help clinicians identify high-risk patients and facilitate proactive healthcare strategies, particularly in resource-constrained settings.

[5] https://www.afro.who.int/sites/default/files/2023-08/Ethiopia.pdf

[6] https://extranet.who.int/ncdccs/Data/ETH_B3_s21_National_Strategic_Plan_for_Prevention_and_Control_of_NCDs2021.pdf

In the context of heart attack risk, ML and DL techniques can process clinical and demographic data to detect subtle risk factors and improve diagnostic precision [12–14]. They enable early disease detection, risk assessment, and better clinical decision-making, thus playing a critical role in modern healthcare.

1.6 Objectives of the study

1.6.1 General objective

The primary objective of this research is to develop a predictive model for heart attack risk in the Ethiopian population using both machine learning (ML) and deep learning (DL) techniques.

1.6.2 Specific objectives

- Create and preprocess clinical and demographic data sets for model training.
- Perform feature selection to identify the most relevant predictors.
- Evaluate and compare various classical ML and DL models to determine optimal performance.
- Develop and fine-tune a predictive model tailored to the Ethiopian population.
- Assess model performance using standard evaluation metrics such as accuracy, precision, and recall.
- Validate the model externally.
- Design and integrate a GUI application to Deploy the model.

1.7 Thesis structure

This is the remainder of this thesis structured . Chapter 1 presents the introduction, background, and objective of the study. And also the significance and research gap of the study. Chapter 2 reviews the relevant literature and theoretical frameworks pertinent to the study. Chapter 3 outlines the research methodology used, detailing the techniques and analytical approaches used. Chapter 4 presents and discusses the research findings, while the final chapter 5 summarises the key conclusions, discusses implications, and provides recommendations for future research.

Chapter 2

Literature Review

2.1 Introduction

In this chapter, we discuss the general overview of CVD and its classification of types and risk factors in both the global and the Ethiopian context. We examine the work being done by existing models, significant findings, and theoretical frameworks relevant to our research. We also highlight the results of previous studies and the research gaps that current studies aim to address.

2.2 Types of CVD

We are in the era of artificial intelligence, which is that the ML and DL models significantly help us in the diagnosis and prediction of CVD. However, it requires an understanding of the types of heart disease and large data for model training. CVDs are typically classified according to the anatomical structures affected, the underlying pathology, or the physiological impact on the cardiovascular system [15, 16].

The main types of heart disease include

- ♡ **A heart attack:** It is known as a myocardial infarction, which occurs when blood flow to a part of the heart muscle is blocked, usually by a blood clot in the coronary arteries. This blockage prevents oxygen-rich blood from reaching the heart muscle, leading to heart failure or death [17, 18].
- ♡ **Coronary artery disease (CAD):** Also known as ischemic heart disease, it is the narrowing or blockage of the coronary arteries. CAD often results in angina or myocardial infarction [heart attack] [19–21].
- ♡ **Cerebrovascular disease:** such as stroke and transient ischemic attacks (TIA), which are caused by restricted blood flow or haemorrhage in the brain [22, 23].
- ♡ **Hypertensive heart disease:** Resulting from high blood pressure, it can cause structural and functional changes in the heart [19, 24].
- ♡ **Heart failure (HF):** This condition occurs when the heart cannot pump enough blood to circulate to the body as needed [25].
- ♡ **Arrhythmias:** These involve abnormalities in heart rhythm and can cause serious complications, including stroke or sudden cardiac death [26].
- ♡ **Valvular heart diseases:** These are due to the dysfunction of one or more heart valves [27].
- ♡ **Congenital heart diseases (CHD):** Structural heart abnormalities present at birth.They can significantly affect oxygenated blood flow and often require surgical correction [28, 29].
- ♡ **Inflammatory heart diseases:** These include myocarditis, pericarditis, and endocarditis, which are commonly triggered by infections or autoimmune reactions [30].

This classification highlights the diversity and complexity of heart disease, all types of disease known as CVD. Therefore, accurate diagnosis and prevention are essential for effective treatment planning, risk management, and evaluation.

2.2.1 Risk factors for developing a heart attack

Based on their impact on the heart, the risk factors for developing a heart attack were also broadly classified as high, medium, and low risk factors. High-risk factors significantly increase the likelihood of developing a heart attack and include hypertension (high blood pressure), hyperlipidemia (high cholesterol levels), diabetes mellitus, smoking, obesity, and a sedentary lifestyle [1].

Medium-risk factors have a moderate association with heart attack and include excessive alcohol consumption, psychosocial stress, poor dietary habits (high sodium and fat intake), and a family history of heart disease. Although these factors may not independently cause a heart attack, however, they significantly contribute to its progression, especially when combined with high-risk factors. The Global Burden of Disease Study highlights the interplay between modifiable and non-modifiable risk factors in the development of heart attack risk in diverse populations [24].

Low-risk factors, such as occasional alcohol consumption, mild physical inactivity, and inadequate sleep, have a comparatively lower impact but can still influence long-term heart diseases [2].

2.2.2 Heart attack risk in Ethiopia

Ethiopia is one of the countries where a significant portion of the population suffers from heart attacks. The Global Burden of Disease (GBD) study shows that CVDs were among the leading causes of death in Ethiopia, with 15% of all deaths in 2019. Previous studies indicate that CVDs in Ethiopia include hypertensive heart disease, ischemic heart disease, rheumatic heart disease, heart attacks, and stroke [31, 32]. Studies conducted in urban and rural areas show that high rates of undiagnosed hypertension, unhealthy eating patterns, physical inactivity, and limited awareness are the main modifiable risk factors [32]. Furthermore, behavioural risk factors such as smoking and chewing khat, particularly among men, have shown strong associations with increased heart attack risk[33, 34].

Ethiopia lacks advanced diagnostic tools, such as blood pressure monitors and glucose test kits, particularly in rural primary hospitals and health care. Due to this, Ethiopian hospitals and health centres face significant challenges in providing chronic care and early diagnosis. Furthermore, there is a marked shortage of trained healthcare professionals capable of managing chronic diseases [35].

To address these challenges and provide a good quality health service, the Ethiopian Ministry of Health launched the national strategic action plan for NCD prevention and control (2014-2016), with an emphasis on integrating heart attack detection and health promotion into primary care [36]. However, implementation has been hindered by limited funding, insufficient inter-sectoral collaboration, and competing health priorities, including ongoing efforts to control infectious diseases [36, 37].

In Ethiopia, the heart attack prediction study has relied primarily on traditional statistical methods such as logistic regression, which identified key risk factors such as hypertension, smoking, and khat chewing. Although the health record in general in Ethiopia is poor, there is still a lack of health records. heart attacks prevalence study in Addis Ababa and in the eastern part of Ethiopia, estimated 24% 7. 2%, respectively, [31].

[1] https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[2] https://www.nhlbi.nih.gov/health/heart-attack/causes

As we know, artificial intelligence technologies have a significant impact on our daily lives, including improving early detection and intervention of non-communicable diseases. Heart attack risk is one of them. However, the effectiveness of early detection depends on the availability of high-quality medical data and well-trained models. In the last decade, the prediction of heart attack risk using ML and DL has grown significantly. However, predicting heart attack risk in Ethiopia presents both opportunities and challenges. One of the main challenges is to get a large amount of sample data from both heart attack patients and healthy patients, which is very important to develop effective ML and DL models. Taking into account local risk factors, healthcare infrastructure, and the population in Ethiopia, early prediction and effective intervention are crucial.

2.3 Traditional methods for heart disease diagnosis and irsk prediction

In the world health sector, traditional risk prediction tools have been developed to assess heart disease in individuals based on established risk factors, which contribute significantly to the prevention and management of heart attacks. Some of the traditional methods used for the prediction of risk factors for heart disease. One of the traditional methods is the Framingham risk score that estimates the 10-year risk of coronary heart disease based on factors such as age, sex, blood pressure, cholesterol levels, smoking status, and diabetes [38]. This model focuses mainly on coronary heart disease and may not accurately predict other outcomes of heart disease, such as heart attack, stroke, or heart failure.

The Atherosclerotic Cardiovascular Disease (ASCVD) Risk Estimator, introduced by the American College of Cardiology and the American Heart Association (AHA), estimates the 10-year risk of major ASCVD events, including myocardial infarction and stroke. Incorporating factors such as age, sex, race, cholesterol levels, blood pressure, diabetes, and smoking status [39, 40]. However, this estimator has limitations, particularly in terms of probabilistic risk scores. This could be a miscalculation of risk in certain populations, especially among low-risk individuals or non-US cohorts. Another traditional method, the Systematic Coronary Risk Estimation 2 (SCORE2) and its older population variant, SCORE2-OP, were developed by the European Society of Cardiology (ESC) to estimate the 10-year risk of heart disease of first onset among European populations. SCORE2 is targeted at people aged 40 to 69 years, while SCORE2-OP is designed for people aged 70 and older [41, 42]. The estimator designs were only for the old population, which is age-dependent, and were calibrated using data from predominantly European cohorts, limiting their applicability to the rest of the world. The risk factor result shows non-linear relationship, which indicates that the model might not capture.

Furthermore, the WHO and the International Society for Hypertension (ISH) developed region-specific risk charts, including versions for eastern Sub-Saharan Africa, such as Ethiopia. These charts estimate the 10-year risk of a fatal or non-fatal heart attack based on age, sex, smoking status, systolic blood pressure, and total cholesterol [43]. This model is better for which it is targeted in Eastern Sub-Saharan Africa, the kind of life style similarity. However, the graphs excluded important local or socioeconomic risk factors, reducing their completeness.

In general, traditional heart attack risk models face critical limitations that hinder their precision and clinical utility. Most are based on static variables collected during periodic health assessments, which do not account for dynamic and longitudinal changes in individual health status over time.

As we mentioned above, there are shortcomings of traditional methods, and there is a clear growing interest in advancing to develop predictive tools taking into account the importance of ML, DL, and large data. Such models can integrate large continuous clinical data from healthcare, wearable devices, and genetic information to provide more precise and personalised heart attack risk assessments [44].

2.4 Classical machine learning for heart attack risk prediction

More than decades, several studies have been conducted to develop accurate ML and DL models to predict heart attack risk using publicly available data such as the Cleveland and University of California, Irvine (UCI), Hungary, Switzerland, and other databases and patient attributes to help physicians. With this publicly available online data has been used by many researchers based on dependent on the attributes, commonly used attributes are: age, sex, type of chest pain, resting blood pressure, cholesterol, diabetes, ECG, angina, smoking status, family history of heart disease, BMI, physical inactivity, alcohol consumption, stress, and mental health.

A comprehensive comparison of several ML algorithms for the prediction of heart attack risk from Cleveland data, which is 303 samples with 276 characteristics [45]. Some of the models used by the anthers were used: DT, NB, and KNN, and found an accuracy of 77.55%, 83.49%, and 83.16%, respectively. The researcher used the ensemble prediction of classifiers, bagging, boosting, and stacking, which were applied to the dataset, and found significant improvement in some cases after applying those approaches.

Sajja et al.[46] studied a comparison between classical ML approaches and CNN to develop a predictive model of heart attack risk in the Cleveland heart disease of UCI that contains 14 features. However, the researcher used cleaned data by splitting them into 80% for training and 20% to test trained models. ML models; LR achieved 89.91% (86.83%), NB achieved 80.62% (77.04%), and KNN achieved 79.76%(68.86%) training (testing) accuracy, respectively.

Vayadande et al. [47] conducted a comparative study using different ML algorithms on the UCI heart disease data, which contains a total of 303 instances and 14 attributes. The authors applied GB, RF, LR, KNN, and DT classifiers accuracies achieved were 88.25%, 88.52%, 88.5%, 80.33%, 78.69%, and 85.25%, respectively. The researchers highlight the effectiveness of ensemble methods such as GB and RF in achieving higher predictive performance on this data.

RF and LR were used to predict heart attack risk outcomes in a large cohort in the United States [48]. Their model achieved accuracy in excess of 85%, with a specific focus on the importance of feature selection and pre-processing in improving model performance. Similarly, studies highlighted the importance of data quality and suggested combining multiple ML models to improve the validity of prediction [49].

Shah et al. [50] explore the performance of four ML algorithms for predicting heart disease using clinical data from UCL with a total of 300 examples of data with 14 various attributes. The authors evaluated these algorithms, including NB, KNN, DT, and RF, and obtained accuracies of 88.16%, 90.79%, 80.26%, and 86.84%, respectively. The researcher identified the most effective model for accurate models of NB and KNN diagnosis and significantly improved the early detection of heart disease, potentially improving patient outcomes.

Recently, a prospective study was conducted in the UK Biobank (UK) to assess the impact of integrating psychological data into ML models for the prediction of heart attack risk [51]. The researcher contrasted models trained on psychological aspects with those trained on traditional risk factors alone using an ensemble model that included algorithms of DT, RT, XG, SVM, and deep neural networks (DNN). The prediction accuracy increased from 71.31% to 85.13% when psychological data were included, indicating the importance of mental health in determining one of the risk indicators of heart attacks.

2.5 Deep learning models for heart attack risk prediction

In recent years, the advancement of AI and the large clinical and demographic data enabling the use of DL have increased significantly to develop models for the prediction of heart attack risk. Various studies have explored different architectures, preprocessing techniques, and hybrid models to improve diagnostic accuracy and decision support. Several studies used different deep learning methods with good accuracy, reporting different results, regardless of the data and the number of samples. In [52], studied with a wide and deep neural network was studied, the data consists of five different independent datasets that have never been combined before, which in the data become 918 instances and 11 features. From the three DL models, the wide and deep neural network has the highest accuracy of 84.24%.

Samuel et al. proposed an integrated decision support system that combines ANN with a Fuzzy Analytical Hierarchy Process (Fuzzy-AHP) to predict the risks of heart failure [53]. Their system utilised a multi-criteria decision-making approach that enhanced interpretability and incorporated expert knowledge, leading to improved diagnostic precision. The model achieved an accuracy of 91.11% on data comprising 300 patient records. There also developed an automated diagnostic system that uses a statistical model χ^2 to select characteristics and an optimally configured DNN for the prediction of heart disease [54]. Their model achieved a high accuracy of 98.54% in the Cleveland Heart Disease data, which contains 303 instances and 76 attributes. The effectiveness of statistical feature refinement was demonstrated prior to deep learning model training. However, both studies use a small sample of potential data complexity introduced by the fuzzy logic component.

With current access to large samples, heart attack risk prediction benefited significantly from deep learning models, each model having a distinct architecture that captures complex patterns in clinical and demographic data. One of the multilayer perception models is the FNN model used to learn nonlinear relationships between high-risk factor indicators (e.g., cholesterol level, high blood pressure, smoking status, BMI, and age) and heart attack risk output.

With an advanced DL model to extract complex patterns, the CNN model proposed the prediction of heart attack risk using clinical data from UCI data [55]. The researcher achieved the highest precision, 93.33%; however, their work lacks the interpretability of CNN models, which can hinder their adoption in clinical data.

Das et al. develop a neurofuzzy model integrated with post-feature reduction techniques for biomedical data analysis [56]. The hybrid architecture combined the interpretability of fuzzy logic with the learning capacity of neural networks, producing robust predictions for heart disease. The model was evaluated with data of 270 instances, achieving an accuracy of 92.59%. However, the complexity of the neurofuzzy model may pose challenges in terms of computational efficiency and scalability.

Nancy et al. developed a smart healthcare monitoring system that used the cloud Internet of Things(IoT) architecture coupled with deep learning models [57]. Their system ensured real-time monitoring and efficient prediction of heart conditions, indicating the potential of DL in remote healthcare applications. Using a Bi-LSTM model, the system achieved an accuracy of 98. 86% of the data collected from the wearable sensors. This study has ethical limitations, including potential data privacy concerns associated with IoT-based systems.

Vayadande et al. conducted a benchmark study that evaluated various ML and DL algorithms for the prediction of heart disease [47], data from the Cleveland Heart Disease dataset, which comprises 303 instances with 14 attributes. They compared models such as SVM, RF, and MLP. The MLP model achieved the highest accuracy of 94.12%, outperforming traditional ML models.

Garcia et al. proposed a DL approach to enhance with feature enhancement techniques to improve the prediction of heart disease risk [58]. The researcher used combined data from five sources, including Cleveland, Hungary, Switzerland, Long Beach VA, and Statlog, with a total of 918 samples with 11 clinical features each. The researcher used a methodology that involved a Sparse Autoencoder (SAE) for feature augmentation and classifiers using either MLP or CNN. The proposed models achieved an accuracy of 90%, outperforming traditional methods by 4.4%.

Almazroi et al. developed a Clinical Decision Support System (CDSS) using DL techniques for the diagnosis of heart disease [59]. Their system was evaluated using four different heart disease datasets, with consistently high performance results in all. The integration of deep learning into the CDSS framework highlighted substantial improvements in predictive accuracy and emphasised clinical applicability. However, the researcher did not provide detailed information on the specific DL architectures used, and the a lack of discussion on integrating the system into existing clinical workflows.

In summary, all of these studies collectively present the rapid advances and diverse applications of deep learning in cardiovascular diagnostics, highlighting its growing potential in clinical decision support systems.

2.6 Research gaps about heart attack risk prediction

Globally, heart attack is the leading non-communicable cause of death. The use of ML and DL has increased significantly due to the availability of public data for heart attack risk predictions. However, there are many research challenges that continue to hinder the translation of these models into clinical practice. In particular, developing a heart attack risk predictive model is based on high-quality and structured local data, even if the major risk and the contributing risk factors are internationally well recognised. One of the reasons it depends on structured and high-quality local data is that it includes lifestyle, environmental, unique genetic, and educational level (awareness in general, yearly health check-up), and differs from country to country. For example, in Ethiopia, developing an early predictive and prevalence model is very challenging with local data.

For a developing country such as Ethiopia, the lack of health equipment and hospitals is increasing, as well as the significant increase in heart attacks that lead to silent death. The existing models were built based on high-income countries and with better life and awareness population data, and were not tested with local data. The lack of structured and high-quality data with large sample data in Ethiopia, considering all possible concerns using publicly available data from high-income countries to develop, is one of the options to test the model with local data.

In addition, there are limited resources and less national concern in using AI for early prediction and prevalence for the use of health diagnosis. As we now know, our hospital has a patient health history record system that is not well organised, making it very difficult to get a large sample of data.

Chapter 3

Methodology

In this chapter,we present data and preprocessing steps, as well as detailed methods used in developing models that help predict heart attack risk using ML and DL techniques.Our research follows an experimental research method, in which we design, implement, and evaluate predictive models through controlled experiments. We begin by describing the data acquisition and preprocessing steps, including feature engineering and model development in both classical ML and DL techniques. Additionally, a comparative analysis based on standard performance metrics is conducted to determine the most effective model for heart attack risk prediction.In our research,we develop both classical ML and DL algorithms to compare the performance of robust predictive models. Furthermore, we provide a comprehensive guide to the architecture of the system and the experimental framework, illustrating the process of selecting the best predictive model and deploying it in an accessible application. As shown in Figure 3.1, the detailed workflow of our improved method follows a sequential process: from data source and preprocessing, through feature selection and data confirmation, to training, testing, and explaining the improved model. We also discuss the design and development of a user-friendly GUI that integrates the selected model, enabling real-time prediction and ease of use for end users.

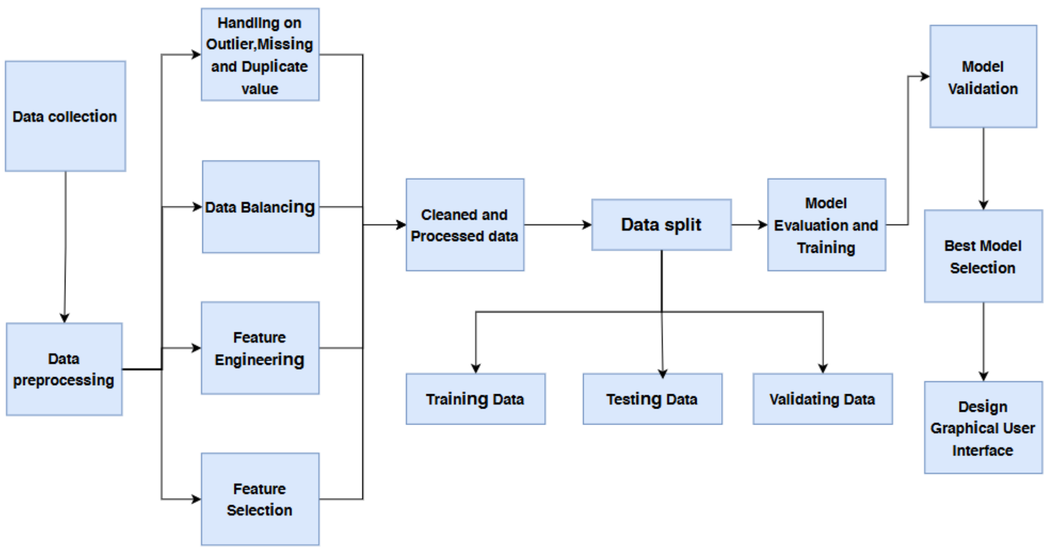

Figure 3.1: Workflow of the developed model

3.1 Data acquisition

For this thesis, we used public data for model development obtained from the Behavioral Risk Factor Surveillance System (BRFSS), the world's largest continuously conducted health survey system. Established in 1984 with participation from 15 states[1],According to the Centre for Disease Control and Prevention (CDC), the BRFSS data, which was collected in 2023 from all U.S. states, includes more than 433,323 adult interviews and 350 variables. [2] The local data for external validation obtained from

[1]https://www.cdc.gov/brfss/annual_data/all_years/states_data.htm
[2]https://www.cdc.gov/brfss/annual_data/annual_2023.html

TASH with 100 entries with 37 variables collected from the data already collected in the department for patients with heart attacks and collected with Google Forms[3] for people who had no heart attacks.

As indicated in the CDC 2023 BRFSS Codebook, the data entries have 23,451 people having heart attacks, while 407,304 people are not having heart attacks. In addition, 2,314, 251, and 3 individuals responded "Don't know/Not sure," "refused", and "Not asked or missing" for the targeted variable (Had heart attack). These values may affect the analysis, so we removed them from the targeted variable because we want only two classes in the target variable. Moreover, 12,407 duplicate records were in the data removed. Based on our research goals, we selected 37 variables based on previous studies' risk contributions for heart attack risk.

3.2 Data preprocessing

In the development of classical ML and DL models, one of the crucial steps is data preprocessing, which transforms raw data into a clean and suitable format for analysis, model training, and testing. And also it enhances the accuracy and performance of predictive models. Here are some preprocessing steps that were performed for this research as follows.

3.2.1 Data visualisation

As we mentioned above, our public data is a large sample. Figure 3.2 shows the distribution of the samples according to the heart attack risk status after removing duplicate data. This visualization is essential to understand the number of heart attack and non-heart attack persons as indicated in the data entries. It clearly shows a significant class imbalance, with 387,936 samples labelled "No Heart Attack" and only 23,413 labelled "Had Heart Attack." This disparity highlights that the data set is heavily biased towards the negative class. This imbalance can bias the ML and DL models towards the majority class, leading to poor predictive performance for the minority class, particularly in the detection of high-risk patients. For this data imbalance, we use undersampling techniques to balance our data, which we discuss in detail in Section 3.2.3.

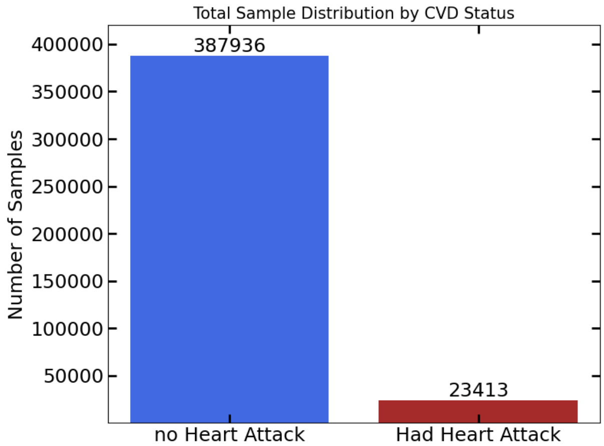

Figure 3.2: Distribution of class

The data we used comprises a total of 37 characteristics, including 3 numerical characteristics and 34 categorical characteristics. A detailed description of the 37 variables and their corresponding questions for numerical and categorical attributes is provided in Tables A.1 and A.2, respectively. Figure 3.3 shows the distribution of numerical attributes, and Figure C.1 illustrates the distribution of categorical attributes.

[3] https://docs.google.com/forms/d/1kKBbDRCNJtspChlMVNvqSPUMAXsfLs1UQI3jl1tivSQ/edit

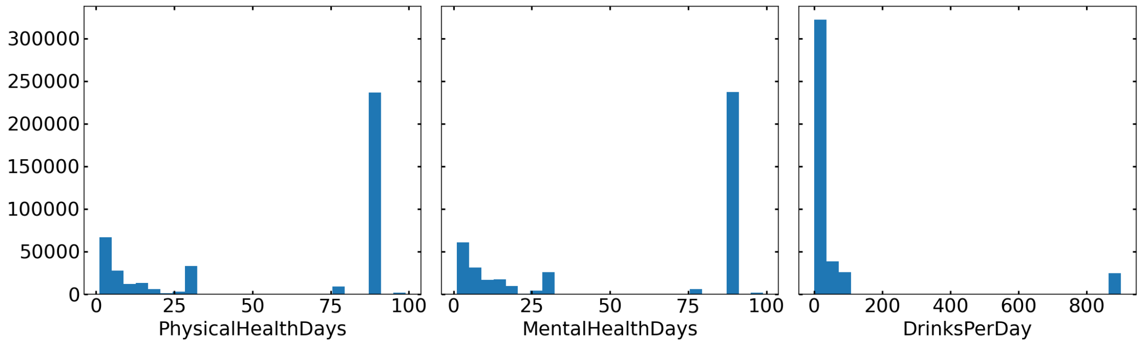

Figure 3.3: Histograms of key Numerical features

3.2.2 Data cleaning

For our study, there were also responses such as "Don't know/Not sure," "Refused," and "Not asked or missing" for non-targeted variables. However, unlike the target variable, we did not remove these responses from non-targeted variables. Instead, we handled them using imputation techniques. Since our dataset contains both numerical and categorical variables, we applied regression imputation for numerical data, which predicts missing values using a regression model based on other observed variables. For categorical data, we used simple imputation. To ensure data completeness and uniqueness, we addressed missing entries and inconsistent responses that could negatively affect model performance during data preprocessing.

For numerical variables, the general form of the regression model is:

$$X_j = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \cdots + \beta_p X_p + \varepsilon \tag{3.1}$$

Where $\mathrm{X_j}$ is the variable with missing values, $X_k$is other observed variables, β_k is the regression coefficients, and ε is an error term.

For categorical variables, missing values were imputed using the most frequent category.

$$X_i^{\text{imputed}} = \text{mode}(X) \tag{3.2}$$

Where is X_i^{imputed} is the imputed value and mode(X) is the most frequent category in x

3.2.3 Data balancing

As we visualized in our previous step there is a data imbalance. It is essential to apply data balancing techniques such as oversampling, undersampling, or synthetic methods such as the synthetic minority oversampling technique (SMOTE) to ensure that the models learn effectively from both classes and provide accurate and fair predictions.

For such imbalanced data, a similar data partitioning technique was applied for heart disease prediction and found that undersampling improved recall of minority classes without significantly sacrificing precision [60]. Furthermore, other techniques such as SMOTE, cost-sensitive learning, and ensemble-based resampling have been extensively explored in the literature to handle class imbalance[61, 62].

We address this data imbalance; a 17:1 undersampling strategy was used. The approach involves dividing the majority class (Class 0) into 17 equal subsets and pairing each with the full set of samples from the minority class (Class 1). This produces 17 balanced datasets that preserve all positive (heart attack) cases and a fraction of negative (no heart attack) cases, thus improving model training and validation reliability. After testing the 17 balanced data sets, we found similar results across them and decided to use one of the 17 balanced data sets for the rest of our study. This approach not only ensures class balance but also enables robust cross-validation and efficient experimentation. The steps involved are outlined below:

- **Step 1: Splitting the data set:** To have an equal number of hearts attak and no heart attack, we divided the original data into Class 0 (No Heart Attack) and Class 1 (Had Heart Attack). Then we divided the larger target variable (Class 0) into 17 sub-samples of No Heart Attack.
- **Step 2: Merging and splitting:** After having 17 sub-samples of Class 0 then we combine Class 1 data with each sub sample to have 17 balanced data.
- **Step 3: Data Preprocessing:** For categorical variables, we encoded using label encoding to convert them into a numerical format suitable for ML models.
- **Step 4: Training and testing:** To check the balance of the data, we use a logistic regression model and split the data into training and testing with a ratio of 80:20. We found that all 17 samples have a quite similar accuracy as shown in Table 3.1.

Table 3.1: Classification metrics for each of the 17 balanced data samples

Data	**Accuracy**	**Precision**	**Recall**	**F1 Score**
1	0.983454	0.968362	1.0	0.983927
2	0.987131	0.975219	1.0	0.987454
3	0.983238	0.967962	1.0	0.983720
4	0.983779	0.968963	1.0	0.984237
5	0.984752	0.970771	1.0	0.985169
6	0.980210	0.962392	1.0	0.980836
7	0.982913	0.967362	1.0	0.983410
8	0.986050	0.973192	1.0	0.986414
9	0.985833	0.972788	1.0	0.986206
10	0.983887	0.969164	1.0	0.984341
11	0.985509	0.972182	1.0	0.985895
12	0.982048	0.965766	1.0	0.982585
13	0.983995	0.969365	1.0	0.984444
14	0.985401	0.971980	1.0	0.985791
15	0.982156	0.965965	1.0	0.982688
16	0.988645	0.978070	1.0	0.988914
17	0.979890	0.961799	1.0	0.980528

We also tried to use SMOTE which helps balance imbalanced datasets by creating new, synthetic samples for the minority class instead of simply duplicating existing ones. However, it has some disadvantages compared to undersampling. One major drawback is the risk of overfitting, since SMOTE generates artificial data that may not perfectly represent real-world patterns, especially when the dataset is small or noisy. It can also increase training time and memory usage, as it expands the dataset size. In addition, SMOTE might create unrealistic or noisy samples, particularly when dealing with categorical or highly non-linear features, and may even blur class boundaries by generating points between overlapping regions. On the other hand, undersampling avoids creating artificial data and is computationally faster, but it can lead to loss of valuable information from the majority class. In summary, SMOTE preserves all data but may introduce artificial noise and complexity, while undersampling keeps data realistic but sacrifices some information from the majority class. Figure3.4 shows the class distribution after applying smote by taking 200000 from majority class and use smote for minority class.

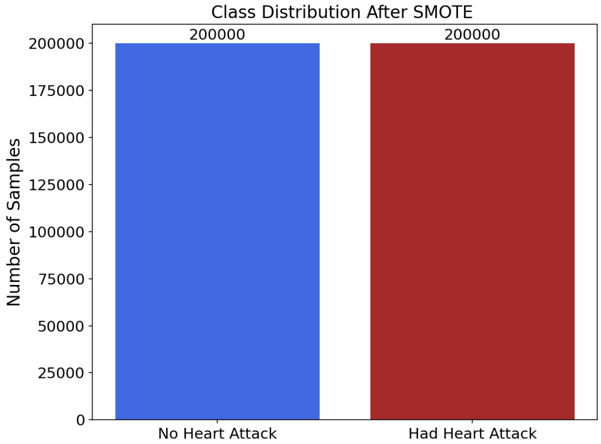

Figure 3.4: Balanced distribution of heart attack and no heart attack data after applying SMOTE.

Table 3.2: Classification Report of SMOTE used balanced data

Class	**Precision**	**Recall**	**F1-Score**
0 (No Heart Attack)	0.75	0.72	0.73
1 (Heart Attack)	0.73	0.77	0.75
Accuracy		0.74	
Macro Avg	0.74	0.74	0.74
Weighted Avg	0.74	0.74	0.74

By comparing classification performance report of SMOTE and undersampling techniques we decide to use Undersampling technique. The distribution of the balanced data we used in this study is shown in Figure 3.5, with almost an equal number of samples for both the "Heart Attack" and "No Heart Attack" categories.

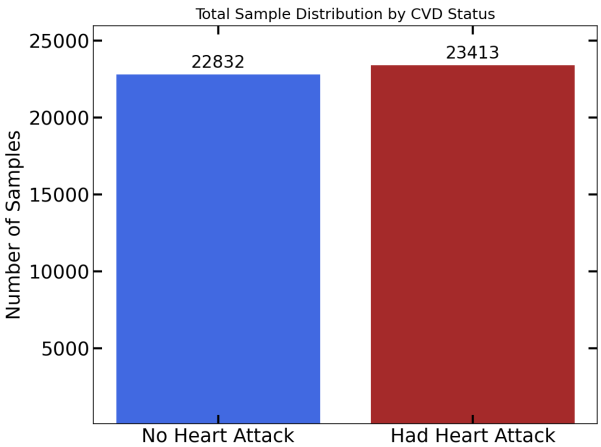

Figure 3.5: Balanced distribution of heart attack and no heart attack data after applying undersampling.

For the development of medical prediction models, maintaining balanced data is crucial. For example, in our study for heart attack risk prediction, where false negatives (that is, predicting no heart attack when it is likely) can have serious consequences. Thus, the adjustment shown in Figure 3.5 is a critical step toward building a more reliable and fair predictive model.

3.3 Outlier detection and treatment

For our data, we handle outliers, which is a critical step in preparing data for statistical analysis, ML, and DL. Outliers are data points that deviate significantly from the mean in the data. These anomalies can arise due to rare events, measurement errors, or inherent variability in the data. If not addressed properly, outliers can distort statistical estimates and degrade the performance of the ML and DL algorithms [63]. Several methods have been used to understand the existence of outliers in the data, including both statistical and visual approaches. Visualising outliers helps to understand the structure and distribution of the data, identify anomalies, and inform preprocessing decisions, thus ensuring the robustness of the downstream analysis and modelling. Some of the visual techniques such as box plots, scatter plots, and histograms, are commonly used for this purpose. For our study to detect outliers, we used interquartile range (IQR) outlier detection methods. The IQR method identifies outliers based on the spread of the middle 50% of the data. The IQR is defined as:

$$\text{IQR} = Q_3 - Q_1 \tag{3.3}$$

A data point is classified as an outlier if it satisfies the following condition:

$$\text{Value} < Q_1 - 1.5 \times \text{IQR} \quad \text{or} \quad \text{Value} > Q_3 + 1.5 \times \text{IQR} \tag{3.4}$$

where Q_1 is the first quartile (25th percentile), and Q_3 is the third quartile (75th percentile).

As can be seen in Figure 3.6, the box plot provides a concise summary of the distribution of numerical variables, the range between the first quartile (25th percentile) and the third quartile (75th percentile). The line within the box denotes the median, whereas the "whiskers" typically extend to data points within 1.5 times the IQR from the quartiles. Any data points outside this range are considered outliers and are plotted as individual dots outside the whiskers [64].

Once outliers have been detected, we apply capping outliers, which is replacing outlier values with the nearest acceptable boundary (e.g., the 1st or 99th percentile) as shown in Figure 3.6.

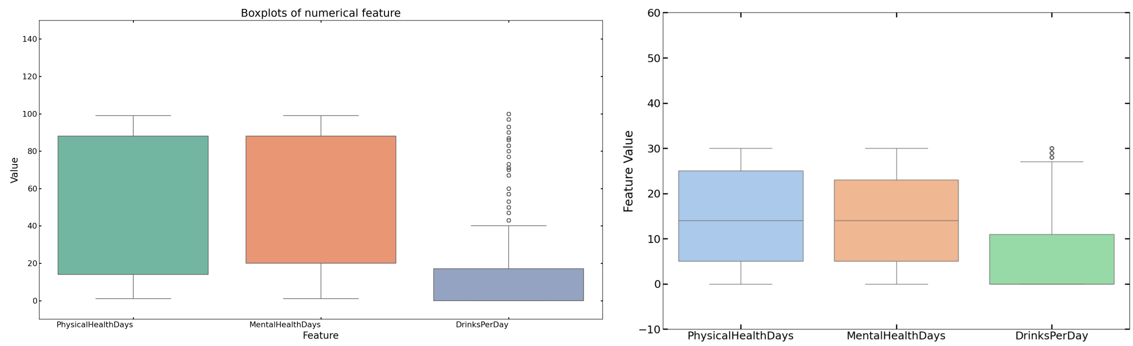

Figure 3.6: Boxplot output before outlier treatment (right) and after outlier treatment (left).

3.4 Feature engineering

Feature engineering and feature selection are crucial steps in improving the performance of ML models. Feature engineering involves creating new features or transforming existing ones to better capture patterns in the data. This process can be applied in different ways, as described below.

3.4.1 Handling categorical variables

In our data, we have different categorical variables that represent different groups or categories (e.g., "Yes"/"No," "Male"/"Female"). For ML and DL models, they generally perform optimally with numerical data; therefore, categorical variables must be transformed into a numerical format before model training.

- **Nominal variables**: Categories without inherent order (e.g., "Race," "Blood Type").
- **Ordinal variables**: Categories with a meaningful order (e.g. "Age Group," "Education level").

To handle categorical features, we used encoding, which is the process of converting categorical (textual) data into numerical representations that can be processed by ML and DL algorithms. There are common encoding methods that include the following:

- **One-hot encoding**: Suitable for non-ordinal categorical variables.
- **Ordinal encoding**: Appropriate for ordinal variables where the order is meaningful.

3.4.2 Scaling and normalization

ML algorithms perform better when all features are on a similar scale. Scaling and normalisation ensure that feature values are adjusted without distorting the relationships among data points.

- **Standardization (Z-score scaling)**: Transforms features to have zero mean and unit variance. It is useful for algorithms that assume normally distributed input features, such as LR and KNN.

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma} \tag{3.5}$$

$$\begin{aligned} X &= \text{Feature value,} \\ \mu &= \text{Mean of the feature,} \\ \sigma &= \text{Standard deviation of the feature.} \end{aligned}$$

- **Normalization (Min-Max Scaling)**: Rescales features to a fixed range, typically [0, 1], to avoid large differences in magnitude.

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{3.6}$$

where:

$$\begin{aligned} \text{X} &= \text{Feature value,} \\ X_{\min} &= \text{Minimum value of the feature,} \\ X_{\max} &= \text{Maximum value of the feature.} \end{aligned}$$

3.5 Feature selection

In this study, a total of 37 features were selected from the BRFSS 2020 dataset. The selection was guided by a combination of literature review and domain knowledge on cardiovascular risk factors, including demographic, behavioral, and clinical variables known to influence heart disease. Although the features were drawn from a large pool of variables in the BRFSS dataset, they were chosen to represent the most relevant predictors identified in previous studies, such as age, sex, diabetes, hypertension, physical activity, and general health status.

And to compare the correlation and select out of 37 variables We applied three different feature selections to better understand the most important features for predicting heart attack risk. The first method we used is the filter, which uses the ANOVA F-test to score each feature based on its individual relationship with the target. The second method that we used is the embedded method, which uses a random forest model to measure the importance of each feature during the model training process. The third method that we used was the wrapper method, which uses Recursive Feature Elimination (RFE) combined with Logistic Regression to iteratively select the best subset of features Figure C.1 describes the overall feature importance table for all 36 features' correlation with targeted features. We tested the performance of the classical model using top 10,20, 30, and all 36 features. Tables 3.3describe the average performance metric with different numbers of features; we found the values of the performance metrics allmost similar..

Therefore, to ensure that the models benefit from all available information without losing predictive power, all 37 features were kept for the final model training and evaluation. This approach helps maintain both the accuracy and reliability of the predictions.

Table 3.3: Overall average Evaluation metrics for the top 10,20,30 and all features

Model	**Accuracy**	**Balanced Accuracy**	**Precision**	**Recall**
Logistic Regression	0.79	0.79	0.83	0.75
KNN	0.85	0.84	0.84	0.86
Decision Tree	0.81	0.81	0.82	0.80
Gradient Boosting	0.83	0.83	0.82	0.84
Random Forest	0.83	0.83	0.82	0.84

Figure3.7explain all performance are allready the same when we select different numbers of features based on their importance rank described in TableC.1 so we prefer to use all features to make the model more learned within the features

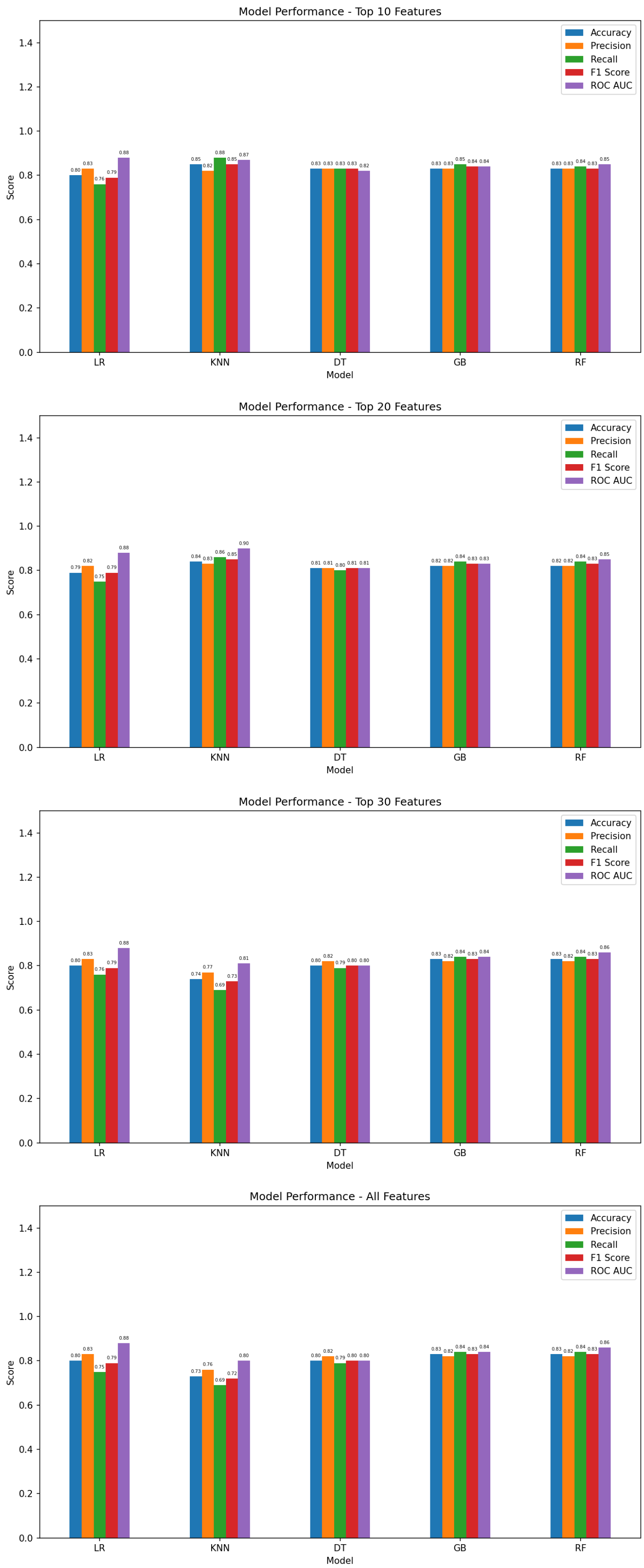

Figure 3.7: Performance plot over ten, twenty, thirty, and all 36 features

3.6 Model selection

Model selection is a crucial step in the ML development process, which involves the identification of the most suitable algorithm or model from a set of candidates based on their performance on a given dataset. This process ensures that the chosen model is generalised well to unseen data, thus enhancing predictive performance and reliability.

3.6.1 Classical machine learning model

In this research, we used five classical ML models to predict heart attack risk. Those models are LR, KNN, DT, GB, and RF. We choose those models based on their performance from the literature. Those models we used to predict heart attack risk are particularly effective when working with tabular data. Here are the classical models that we used for our study:

- LR model algorithm is a statistical model used for binary classification tasks, estimating the probability that a given input belongs to a particular category by applying the logistic (sigmoid) function to a linear combination of input features.For this study it uses hyperparameters like C=1, penalty like L1 (Lasso) or L2 (Ridge) regularization to prevent overfitting and solver like saga suitable for large dataset.

  $$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} \tag{3.7}$$

  where:

  - $P(y = 1 \mid \mathbf{x})$ is the probability that the outcome is class 1 given the input features $\mathbf{x}$,
  - β_0 is the intercept (bias),
  - $\beta_1, \beta_2, \ldots, \beta_n$ are the model coefficients,
  - $x_1, x_2, \ldots, x_n$ are the input features.

  This function maps any real-valued input to a value between 0 and 1, making it suitable for binary classification.

- KNN model algorithm is a non-parametric method that classifies data points based on the majority class among their 'k' closest neighbours in the feature space, relying on distance metrics like Euclidean distance to determine proximity. KNN classifies data by looking at the closest training samples and uses parameters like n_neighbors, weights, and metric. A higher number of neighbors (k) helps reduce overfitting.

- DT model algorithm is a flowchart-like structure where the internal nodes represent feature tests, the branches represent the results, and the leaf nodes represent class labels, splitting the data based on the value of the features to make predictions and controlled by parameters such as max_depth, min_samples_split, and criterion. Regularization in decision trees is typically done by limiting the depth of the tree.

- GB model algorithm is an ensemble technique that builds models sequentially, with each new model attempting to correct the errors of the previous ones, combining weak learners to form a strong predictor using parameters like n_estimators, learning_rate, and max_depth. It avoids overfitting by using techniques like shrinkage, limiting tree complexity.

- Finally, RF combines multiple decision trees, each trained on different random subsets of the data and features. Its key parameters include n_estimators, max_depth, and max_features, and it reduces overfitting through averaging. All of these classical ML models are integral to many applications due to their balance of performance and interpretability.

Table 3.4 shows that the details of the hyperparameters for each model were set to optimal values or by default: We used regularisation of L2 with C = 1 and max_iter = 1000 to ensure convergence for the

Table 3.4: Classical ML models: hyperparameters and justifications

Model	Hyperparameters / Default Values	Justification
LR	C=1, penalty=L2, solver=saga, max_iter=1000	Linear baseline, interpretable coefficients, suitable for probability outputs. Higher max_iter ensures convergence for complex datasets.
KNN	k=7, weights=distance, metric=euclidean	Non-parametric, captures local patterns.
DT	max_depth=10, min_samples_split=2, criterion=gini	Interpretable, captures non-linear relationships. Can overfit, especially on small datasets, but robust to feature scaling.
RF	n_estimators=200, max_depth=None, max_features=sqrt	Ensemble of decision trees with bagging; reduces overfitting compared to a single tree.
GB	n_estimators=200, learning_rate=0.05, max_depth=3	Ensemble boosting method that builds trees sequentially; handles complex non-linear patterns; regularization controlled via learning_rate and n_estimators.

complex dataset in the classical model of LR. In the KNN model, we used 7 neighbours with distance weighting which captures local patterns and non-parametric; For DT and RF models, we had controlled depth and splitting criteria to reduce overfitting; for GB model, we used 200 estimators, a learning rate of 0.05, and a maximum depth of 3. We chose these configurations to balance the interpretability of the model and the ability to capture nonlinear relationships in the dataset, ensuring reliable and comparable performance across all models.

3.6.2 Deep learning model

We implemented four deep learning architectures to further improve prediction performance and capture more complex patterns in the data. These models are: FNN, a Wide & Deep model, a Residual model, and an Attention-based model. DL automatically learns features through its layered structure, enabling it to detect complex patterns and temporal dependencies. DL models are more prone to overfitting, although techniques such as dropout, early stopping, and batch normalisation help. These models often rely on manually engineered features and are simpler, faster to train, and easier to interpret, making them suitable for small to medium-sized datasets with static tabular information, such as feedforward neural networks, residual networks, attention-based models, and wide-and-deep architectures, designed to handle more complex data types. For our study, we used the following models:

- FNN is the simplest form of neural network, where information moves in only one direction from input to output without any cycles or loops. It is suitable for tasks where data relationships are straightforward.
- The Wide and Deep models integrate linear models (wide) with deep neural networks to capture both memorisation and generalisation aspects of data.
- Residual networks introduce skip connections that allow gradients to flow through the network more effectively, facilitating the training of deeper networks.
- Attention mechanisms enable the network to focus on the most relevant parts of the input data, enhancing performance in tasks like language translation and image captioning.

Table 3.5 expresses the detail regularisation and justification of the methods for all DL models. They were trained using Adam Optimiser with binary cross-entropy loss, early stoppage (patience = 10) to prevent overfitting. Across all models, dropout at 0.5 is used for three models and dropout 0.6 used for

Table 3.5: DL models: architecture, regularization, and key advantages

Model	**Architecture**	**Regularization**	**Justification**
FNN	Dense(16, ReLU) → Dropout(0.5) → Dense(1, Sigmoid)	Dropout	Simple baseline; interpretable; fast training
Wide & Deep	Input → Wide(Dense16) + Deep (Dense16 → Dropout0.6) → Concatenate → Dropout0.6 → Dense1	Dropout	Captures linear and non-linear feature interactions; robust to overfitting
Residual	Dense16 → Dropout0.5 → Residual(Dense16) → Add → Dropout0.5 → Dense1	Dropout, Residual connections	Improved gradient flow; better stability
Attention	Input → Dense16 (query, key, value) → Dot → Softmax → Multiply → Dropout0.5 → Dense16 → Dense1	Dropout, attention mechanism	Dynamically weights important features; enhanced focus on critical risk factors

wide& deep model serves as the key regularizer to prevent overfitting.The detailed architecture of all DL models is explained in Figure 4.3 and Figure 4.4.

3.7 Performance evaluation metrics

After building a model and training the network, its performance must be evaluated to know the actual result. There are different ways of evaluating a model. The first is to use a confusion matrix and get the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) rates of the predicted values. The second is to calculate and obtain the precision, recall, F1 score, and overall accuracy values of a model.

- TP value indicates that what is predicted is true.
- TN value indicates that the predicted class is truly negative.
- FP value indicates that something is predicted as if it is part of the class, while it is not.
- FN the prediction indicates that it is not part of the class, while it is.

Once the confusion matrix is ready, the classification report can be done which contains the precision, recall, and F1 score. So, given a class prediction from the classifier, precision is the one that answers the question 'how likely is it to be correct?' Recall, or sensitivity, will indicate the answer to 'will the classifier detect it? The F1 score is the harmonic mean of precision and recall. The model is said to perform well if we have a high F1 score, and the specificity is that it determines the proportion of actual negatives that are correctly identified. They are calculated using the following equations.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{3.8}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{3.9}$$

$$\text{F1} - \text{score} = \frac{2 * \text{Precision} * \text{ Recall}}{(\text{Precision} + \text{Recall})} \tag{3.10}$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \tag{3.11}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FN} + \text{FP})} \tag{3.12}$$

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right) \tag{3.13}$$

Figure 3.8 shows that given a binary classifier of class A and class B, a model can be evaluated using a confusion matrix. The table indicates the TP, TN, FP and FN components for class A. From the table, we can see that the green shaded part indicates the TP and TN parts.

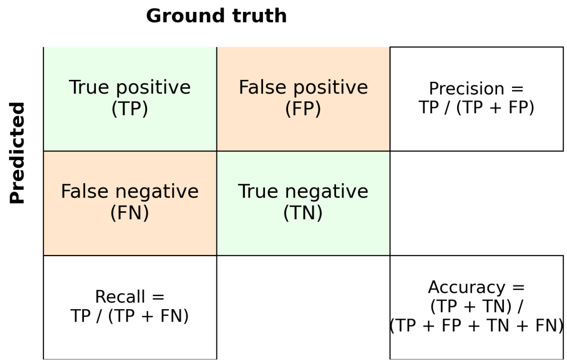

Figure 3.8: Confusion Metric

For our models, we used the above performance matrix measurement methods.

And also, the area under the curve (AUC)-receiver operating characteristics (ROC) plot visualises how well the model can distinguish between classes. The ROC curve can be plotted with two known methods for binary and multiclass classification, which include one versus one class and one versus the rest classes, respectively. In addition, the AUC score can be calculated for each class.

3.8 Model validation

One of the crucial cross-checks of the performance of the model is model validation to evaluate the generalisability and robustness of models, particularly in healthcare applications it is critical. In this study it performed using both internal and external approaches to assess generalization and robustness. Internally, the deep learning models were trained using a validation split during each training run. Early stopping with a patience of 10 epochs monitored the validation loss, halting training when no improvement was observed, which prevents overfitting and ensures that the best weights are restored.

Externally, model performance was evaluated on a separate test set (20% of the original dataset) that was not seen during training. This evaluation included computation of multiple metrics such as accuracy, balanced accuracy, precision, recall, F1 score, and ROC AUC.This approach effectively serves as external validation, providing a realistic estimate of how the models would perform on unseen patient data. By combining early stopping, validation splits, and testing on held-out data, the models' predictive reliability and generalization ability were rigorously assessed.

We evaluated our model based on training and testing accuracy and loss, which helps assess both its learning performance and generalisation ability. We also evaluated our results with our local data for external validation and discussed in 4.1.5.

3.9 Model deployment using graphical user interface (GUI)

Model deployment using a graphical user interface (GUI) is used for making the developed heart attack prediction model accessible and easy to use by non-technical users, such as healthcare professionals or patients. The GUI provides a visual and interactive platform where users can input health-related information through simple elements like text boxes, sliders, buttons, and drop-down menus. It eliminates the need for coding knowledge and allows users to obtain prediction results instantly through a well-designed web interface. We developed a web-based GUI to build the heart attack risk prediction model using Streamlit, a Python framework designed to build interactive data applications. We discussed in 4.2 in detail.

3.10 Material used in the research

We listed the main materials used to carry out the study in the Table 3.6.

Table 3.6: Materials used for this study

Material Name	**Applications**
HP EliteBook 840 G5 computer	To process and organize all thesis work
Streamlit Cloud and GitHub	To develop the GUI application
TeXstudio software and Overleaf website	Used to write the thesis
Python 3.7.3 with various modules	For writing and training the entire system
PyCharm and Jupyter Notebook	For writing and executing the code

3.11 Limitation of the study

This study has several limitations that should be acknowledged; we mentioned two broad points. First, the data used for model development was publicly available data from the US, while the local data for external validation was from Ethiopia. There is a significant lifestyle difference between these two countries, which may affect the model's performance. For example, generally, access to healthcare services varies, which can influence the diagnosis, treatment, and reporting of health conditions. These differences should be taken into account when interpreting the model performance by comparing the model accuracy and the external validation context. Secondly, the local clinical data that we used for the external validation are relatively small and collected from only one hospital, which may limit the generalisability of the results to other populations or regions. In addition, some clinical records may contain missing or inconsistent information, which can affect the accuracy of the model. The study also relied on cross-sectional data, which does not capture the progression of health conditions over time, an important factor in the prediction of chronic diseases. Furthermore, the classification was limited to binary outcomes (risk or no risk), which can oversimplify the complexity of heart attack risk levels. Lastly, while ML models can offer high predictive performance, some lack transparency, making them less interpretable for clinical decision making.

Chapter 4

Results and Discussion

In this chapter, we present and discuss the details of our model findings in both the classical ML and the DL models. We also discuss performance evaluation metrics, which include accuracy, balanced accuracy, precision, recall, and the F1 score, and discuss their implications in the context of heart attack risk. Furthermore, we explain the importance of the feature and the GUI performance accordingly.

4.1 Model training and evaluation

In this section, we describe the procedures used to train the model and evaluate its performance. To train and evaluate the predictive performance of various ML and DL models in our tabular data based on clinical and demographic characteristics, we followed a systematic and structured process. We divide the data into three subdata: 80% for training, 10% for validation, and 10% for testing, using stratified sampling to preserve the distribution of the target variable ('HadHeartAttack').

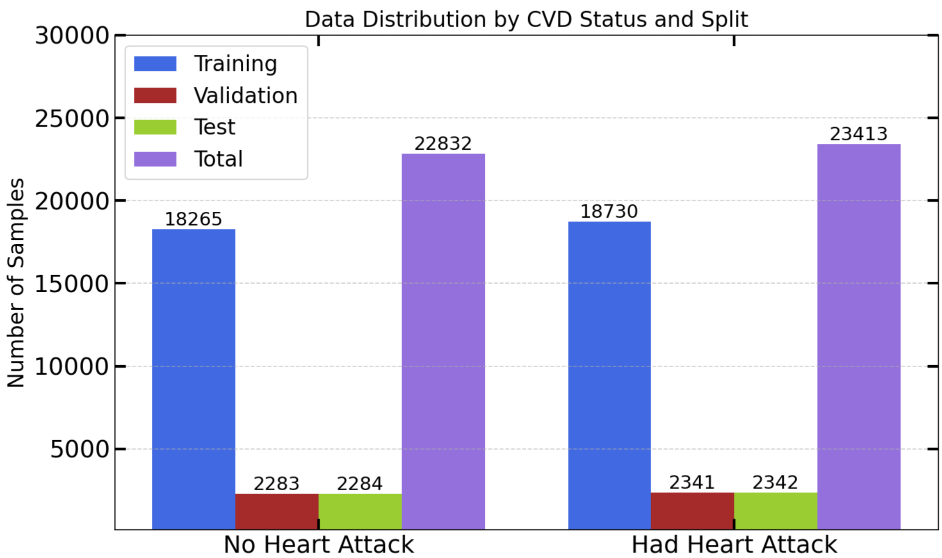

Figure 4.1: Data split for Model Training

Figure 4.1 presents the distribution of samples in the training, validation, and test data. The training data contains 18,265 'no heart attack' and 18,730 'had heart attack' cases; the validation data includes 2,283 'no heart attack' and 2,341 'had heart attack' cases; and the test data consist of 2,284 'no heart attack' and 2,342 'had heart attack' cases. In total, the data comprises 22,832 'no heart attack' and 23,413 'had heart attack' samples. This stratified split approach ensures that the distributions of 'had heart attack' and 'no heart attack' cases are representative across all sub-data, which is critical for preventing bias during model training and enabling reliable performance evaluation.

4.1.1 Classical machine learning models results

As mentioned above, the data used for model training and evaluation was applied to several ML algorithms, including LR, KNN, DT, GB, and RF. Table 4.1 shows that using these models, we found that RF achieved the highest overall performance, with a precision of 0.83, a balanced precision of 0.82, and an F1 score of 0.83. GB also performed competitively, with slightly lower metrics. In contrast, KNN showed the lowest performance in most evaluation criteria, indicating potential difficulty in capturing minority class instances. In general, ensemble models such as GB and RF outperformed simpler models in handling complex heart attack risk prediction tasks.

Table 4.1: Performance of classical ML models for heart attack risk prediction

Model	Accuracy	Balanced Accuracy	Precision	Recall	F1score	ROC AUC
LR.	0.80	0.80	0.82	0.76	0.79	0.88
KNN	0.74	0.74	0.76	0.70	0.73	0.80
DT	0.80	0.80	0.81	0.78	0.80	0.80
GB	0.82	0.82	0.82	0.84	0.83	0.83
RF	0.83	0.82	0.82	0.84	0.83	0.85

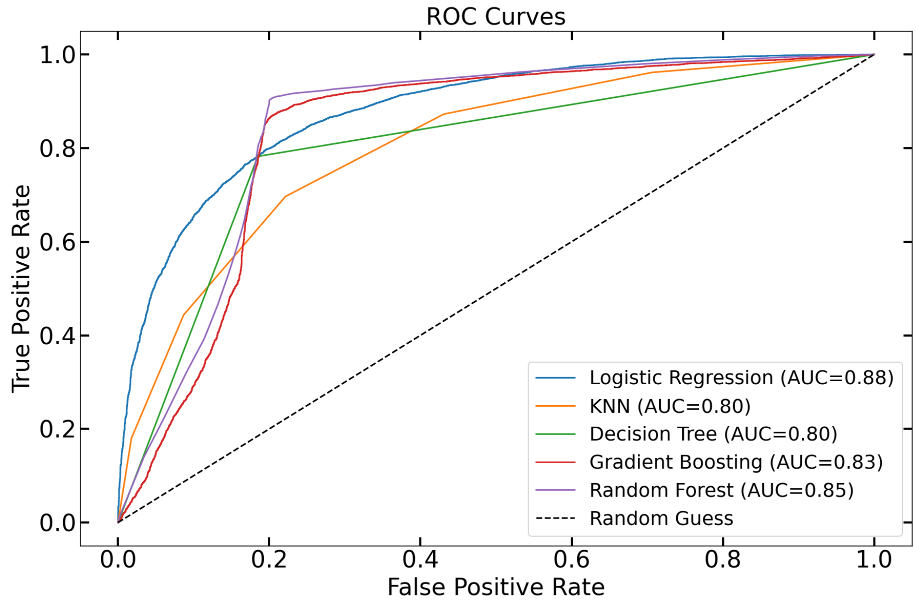

Figure 4.2: ROC curves of classical ML models for heart attack risk.

Figure 4.2 shows the ROC curve and provides additional information on the discriminatory power of each heart attack risk prediction model. Models with the largest AUC, for example, LR, confirm their ability to distinguish between heart attack risk and non-heart attack risk cases with high sensitivity. Together, the ROC analysis complements the tabular metrics and confirms the reliability of the ensemble-based models for predicting heart attack risk. In particular, the model with high recall and AUC could serve as an effective screening tool to identify high-risk individuals for further medical evaluation.

Based on the results of the classical ML models, it is not sufficient to definitively select a single best-performing model. Although these models offer useful benchmarks, their performance alone may not fully capture the complexity. Therefore, we also explored DL approaches, which are capable of learning more complex, non-linear patterns from the data. We aim to identify the most suitable model for our prediction task. By comparing the evaluation results of both the classical ML and the DL models.

4.1.2 Deep learning models results

DL model architectures are multilayered neural networks to model complex patterns in data. These architectures are particularly adept at handling large-scale and unstructured data, such as images, text, and time-series data. To address the diversity of patterns in tabular medical data, we implemented and compared several deep learning architectures. FNN is a baseline model that learns hierarchical patterns by stacking dense layers. It is effective for tabular data. For advanced pattern learning, we also implemented a Wide and Deep model, which simultaneously learns linear relationships through a wide component and nonlinear interactions through deep layers, making it suitable for combining categorical and continuous health-related data. The residual network uses skip connections to mitigate vanishing gradient issues and enables deeper learning, which is advantageous when capturing subtle interactions in high-dimensional tabular data. Finally, the attention-based model dynamically weighs the importance and interactions of features by allowing the network to focus on the most relevant input features, enhancing performance in heterogeneous medical datasets. And we also try to describe the architecture of the DL models as listed below:

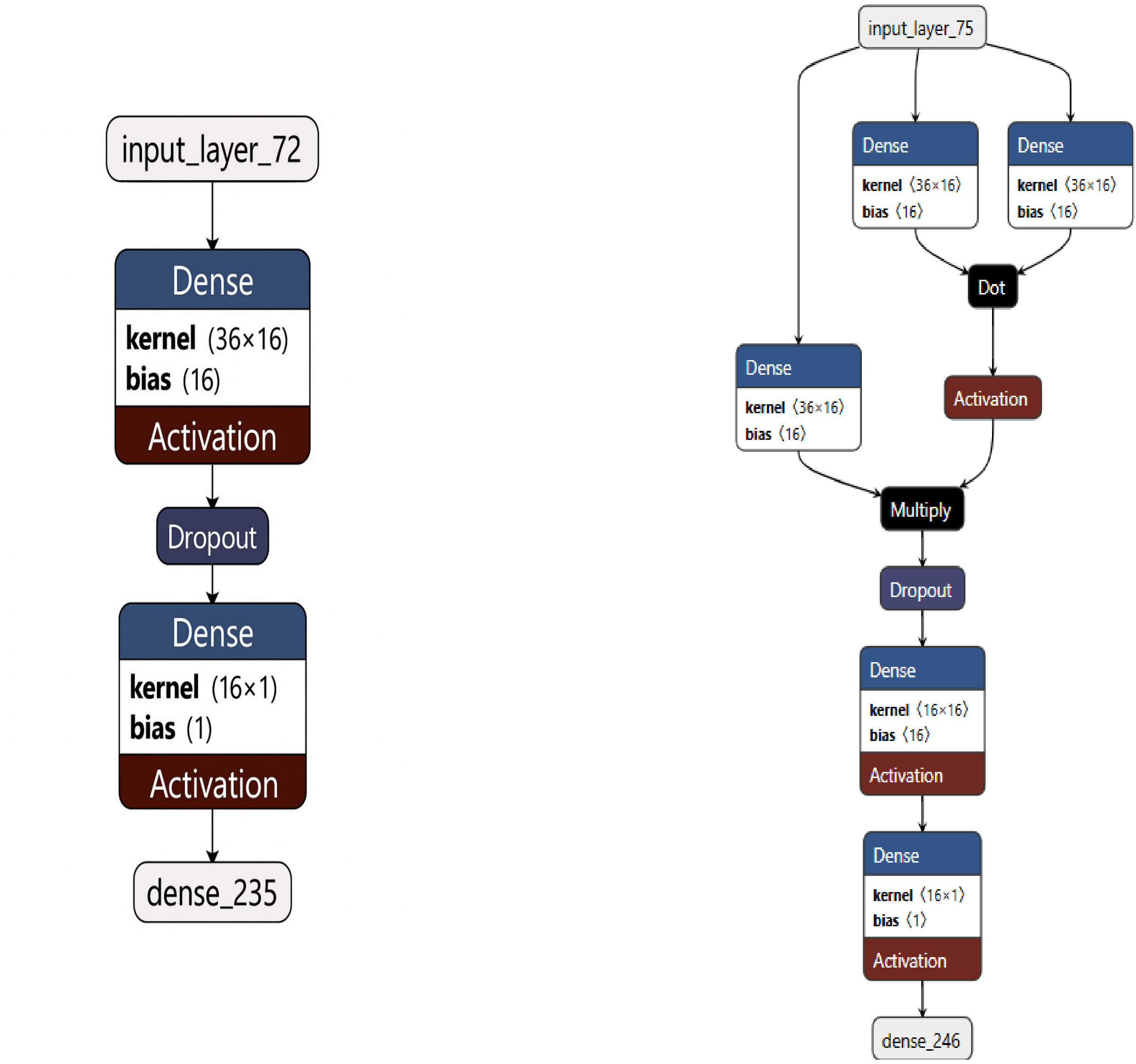

Figure 4.3: FNN (left) and Attention (right) model architectures.

In Figure 4.3 indicates that the FNN model (left) takes 36 input features and passes them through two fully connected layers (Dense). The first Dense layer has 16 neurones and uses an activation function to introduce non-linearity. After that, a dropout layer helps prevent overfitting by randomly turning off some neurones during training. The second dense layer has one neurone, which produces the final output useful for predicting a single value like yes / no (e.g., heart attack risk). This model is simple and commonly used for binary classification tasks.

In Figure 4.3 describes the architecture of the attention-based model (right), takes a sequence input and applies self-attention using dot-product attention. It projects the input into two branches (queries and keys) using dense layers, computes attention scores via a dot product, and applies an activation function. These scores are then multiplied by a third density-transformed version of the input (values). The result is passed through Dropout and two additional dense layers with activations, ending with a single output neurone. It is useful for capturing relationships between different positions in a sequence.

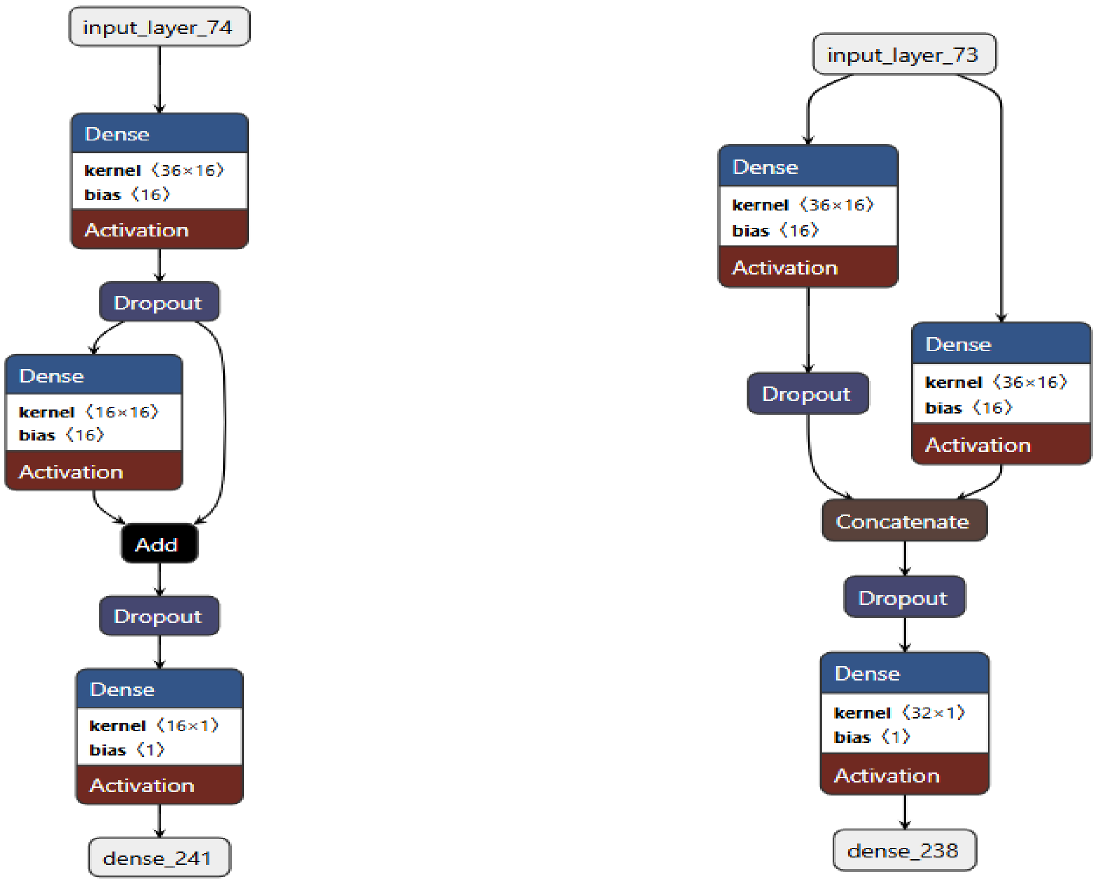

Figure 4.4: Residual (left) and wide deep (right) model architectures.

In Figure 4.4 indicates that the model (left) uses a residual connection, which means that it adds the input of a layer to the output of a deeper layer. It starts with a dense layer of 16 units, followed by dropout. Then it passes through another dense layer (also 16 units), and its output is added back to the previous layer's input (a skip connection). This helps the model learn better by preserving earlier information. A final density layer with 1 unit gives the output. This architecture is useful for training deeper models by reducing vanishing gradient problems.

In Figure 4.4 indicates that a wide deep model (right) combines two parallel dense layers that both receive the same 36 input features. Each path processes the input independently, and their outputs are concatenated (joined together), forming a wider feature vector. After a dropout layer, the combined features are passed to a final dense layer with 1 output unit. This architecture captures both general patterns (deep) and specific rules (wide), making it powerful for the classification task.

Each model was evaluated using accuracy, balanced accuracy, precision, recall, F1-score and ROC AUC metrics. In addition, ROC curves were plotted to visualise and compare the diagnostic capacity of each model.

Table 4.2: Performance comparison of selected DL models for heart attack risk prediction

Model	**Accuracy**	**Balanced Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC AUC**
FNN	0.95	0.95	0.95	0.96	0.95	0.97
Wide_Deep	0.93	0.93	0.95	0.91	0.93	0.97
Residual	0.91	0.91	0.96	0.86	0.91	0.97
Attention	0.94	0.94	0.94	0.95	0.94	0.97

Based on the results reported in Table4.2, the FNN model stands out with the strongest overall performance: achieving an accuracy of 95%, a balanced accuracy of 95%, precision of 95%, recall of 96%, a F1 score of 95% and an ROC AUC of 97%, clearly demonstrating its excellent classification capabilities. Close behind, the WideDeep architecture maintains similarly strong performance: 93% in accuracy and balanced accuracy, with precision of 95%, recall of 91%, F1 score of 93% and AUC of 97%.

The residual model also performs solidly, reaching 91% for precision and balanced accuracy, precision of 96%, recall of 86%, F1 of 91%, and AUC of 97%. Finally, the Attention-based model delivers particularly robust results, registering 94% for accuracy and balanced accuracy, precision, and recall of 94 to 95%, an F1 score of 94%, and AUC reaching 97%. In summary, all four deep models achieve high ROC AUCs (97%), with FNN marginally leading in overall predictive performance.

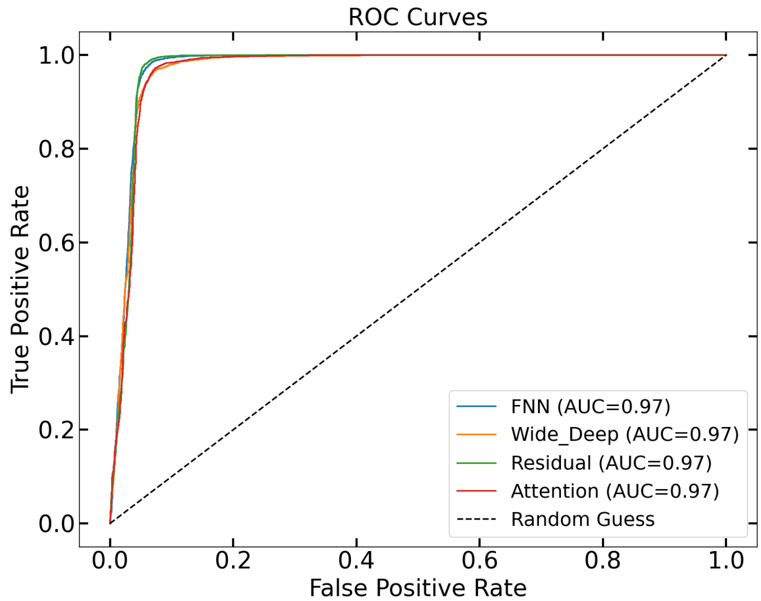

Figure 4.5: ROC curves for different DL model

4.1.3 Model accuracy and loss

Training accuracy and loss show how well a model performs on the data on which it was trained. Training accuracy is the percentage of correct predictions during training, and training loss measures how far the model predictions are from the actual values. The validation accuracy and loss do the same, but on a separate set of data the model has not seen before, called the validation set. These help to check how well the model can generalise to new data. If the training accuracy is high but the validation accuracy is low, it usually means that the model is overfitting. It learnt the training data too well, but it cannot perform well on new data. So, we use all four metrics to track learning progress and make sure that the model is not just memorising, but actually learning. Figure 4.6 (left panel) shows the progression

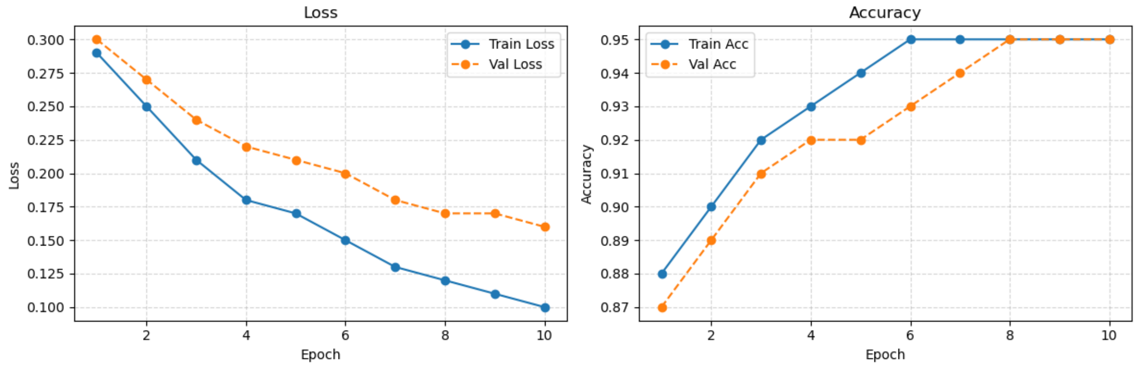

Figure 4.6: Training and validation loss and accuracy curves for the FNN model

of the loss for the FNN model. Training loss consistently decreased from 0.29 to 0.19, while validation loss decreased from 0.30 to 0.16, demonstrating steady convergence. In particular, the validation loss stopped improving and stabilized at around 0.17. The gap between the two curves remains minimal and stable throughout all epochs, never exceeding 0.025 in the final stages. This parallel convergence indicates the model is learning generalizable patterns without memorizing the training data, representing an ideal scenario with no concerning overfitting.

Figure 4.6 (right panel) also illustrates the accuracy of the FNN model. The FNN exhibited strong and consistent performance, starting with a training accuracy of 88% and a validation accuracy of 87% in the first epoch. Both metrics improved steadily, reaching a training accuracy of 95% and a validation accuracy of 95% at the tenth epoch. The model demonstrated rapid convergence and stability, with validation accuracy slightly surpassing training accuracy, which is a sign of effective generalisation and minimal overfitting.

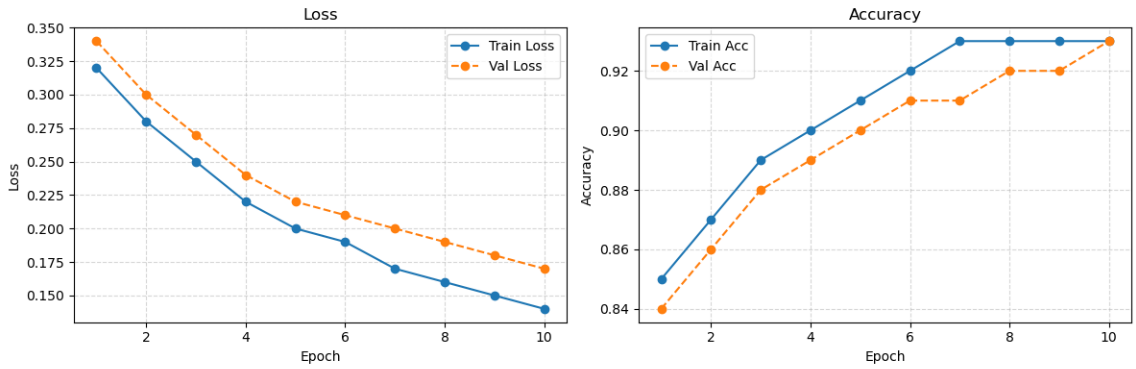

Figure 4.7: Training and validation loss and accuracy curves for the wide & deep model

Figure 4.7 (left panel) presents the progression of the loss of training and validation for the Wide & Deep model. The training loss decreased steadily from approximately 0.32 to 0.14, while the validation loss followed a similar downward trend from around 0.34 to 0.17,creating a persistent gap of approximately 0.025 that remains consistent throughout training. While both curves show downward trends, the maintained separation suggests the model is beginning to overfit, learning patterns that don't fully generalize to the validation data, though not yet at a critical level.

Figure 4.7 (right panel) illustrates the corresponding accuracy metrics. Training accuracy increased from 85% to 93% in 10 epochs, while validation accuracy increased from 84% to 93%. The parallel growth in both metrics, with a consistently small gap between them, confirms that the model achieved strong and stable performance with effective generalisation.

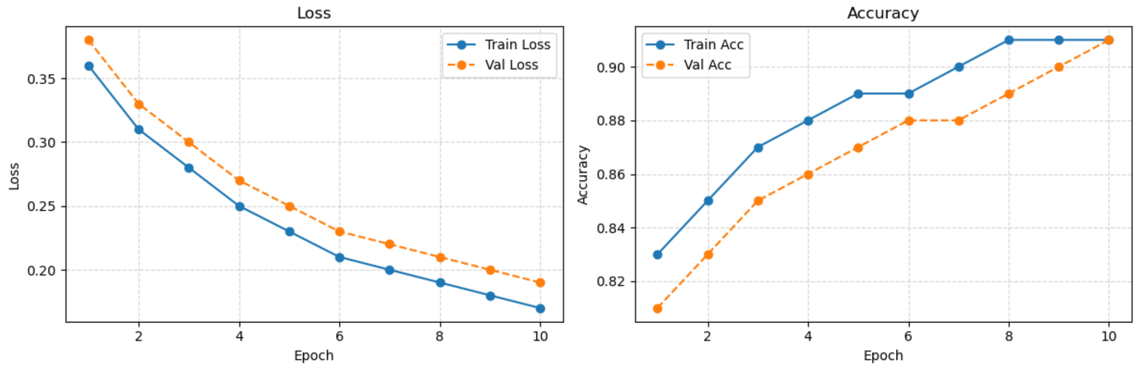

Figure 4.8: Training and validation loss and accuracy curves for the residual network model

Figure 4.8 shows the progression of the loss of the residual network model. The training loss improved from 0.36 to 0.17, while the validation loss decreased from 0.38 to 0.19. The consistent improvement and minimal gap between training and validation losses demonstrate the exceptional stability and learning efficiency of the model. Figure 4.8 (right panel) presents the accuracy trends of the residual network model. Starting with 83% (training) and 81% (validation) accuracy, the model rapidly improved, reaching 91% in both training and validation accuracy by the final epoch.

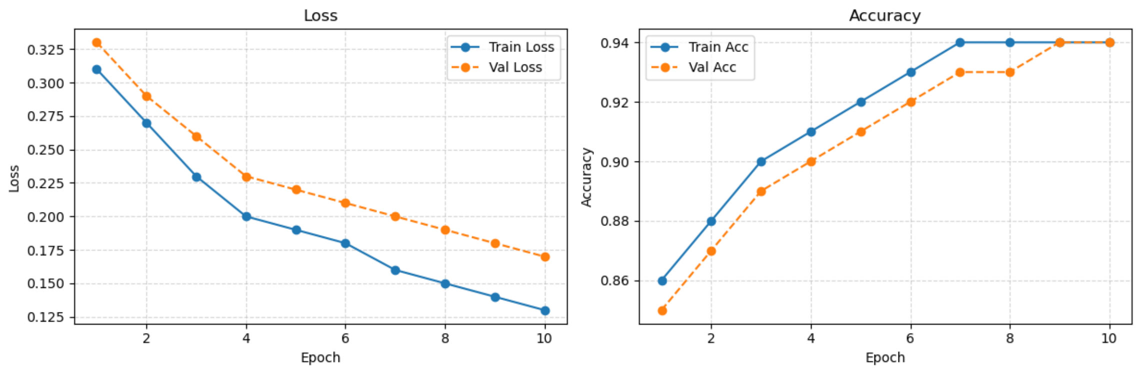

Figure 4.9: Training and validation loss and accuracy curves for the Attention-based model

Figure 4.9 presents the performance of the attention-based model, with data available for ten epochs. The training loss increased from 0.31 to 0.13, while the validation loss of 0.33 to 0.17 remained stable around 0.48. creating a significant and growing gap of 0.05 by the final epochs. This divergence indicates the model is increasingly memorizing the training data rather than learning generalizable features.

Figure 4.9 (right panel) also shows the accuracy metrics of the attention-based model. The training accuracy saw minimal improvement, increasing from 86% to 94%, while the validation accuracy 85% to 94%.

4.1.4 Confusion matrix

In the development of heart attack risk prediction models, we used several architectures of models and used the values of the confusion matrix and the training validation accuracy plots to assess their performance.

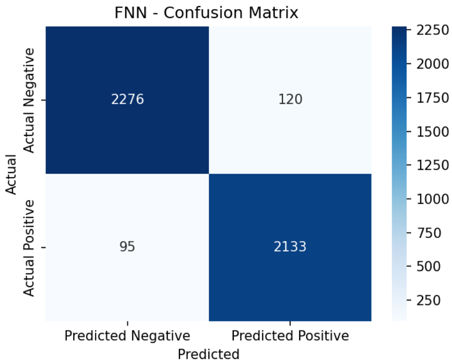

Figure 4.10: Confusion matrix for the FNN model.

.

Figure 4.10 shows the performance of the FNN, which achieved TP = 2276, FP = 120, FN = 95 and TN = 2133. These results indicate strong classification performance with a low misclassification rate, further supported by accuracy graphs, which demonstrate consistently high accuracy in both training and validation datasets.

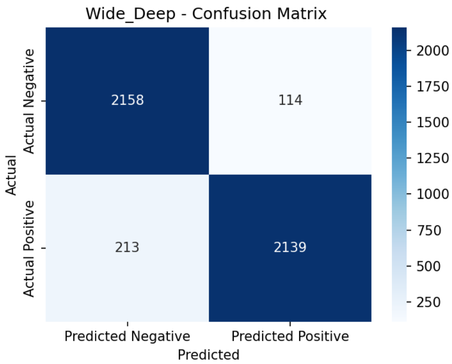

Figure 4.11: Confusion matrix for the wide & deep model.

Figure4.11 shows the performance of the wide & deep model, which achieved TP = 2158, FP = 114, FN = ,213 and TN = 2139. These results indicate strong classification performance with a low misclassification rate, further supported by accuracy graphs, which demonstrate consistently high accuracy in both training and validation datasets. Figure 4.12 shows the residual network performance, with the best results in all models: TP = 2039, FP = 85, FN = 332, and TN = 2168. This minimal error rate highlights the strength of residual connections in mitigating vanishing gradients and promoting deeper model effectiveness. Finally, 4.13 shows the Attention-based model, which scored TP = 2252, FP = 144, FN =119, and TN =2109 . Despite its advanced architecture for focussing on relevant features, the model performed similarly to simpler architectures such as FNN. The confusion matrix suggests difficulties in discriminating between classes and the validation curve reveals a noticeable drop in training performance, indicating overfitting or ineffective attention weighting.

4.1.5 External validation uaing local heart attack risk data

External validation refers to the evaluation of a trained model on entirely independent data that were not involved in the training, validation, or tuning process. Unlike internal validation, which relies on subsets of the same dataset (e.g., using cross-validation or train-test splits), external validation tests the model's generalisability to new populations or settings. This is particularly important in health applications, where differences in the demographic, environmental, and healthcare systems can significantly affect the performance of the model.In this study, we consider local health data from Ethiopia and compare

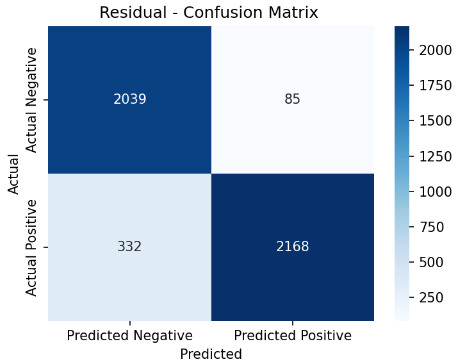

Figure 4.12: Confusion matrix for the residual network model.

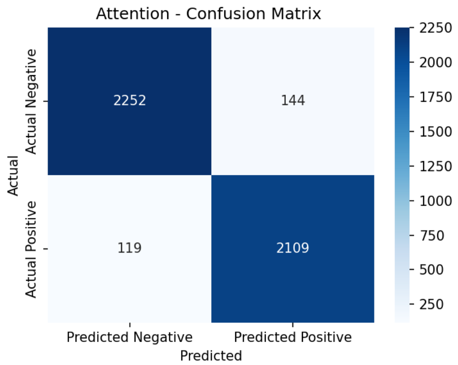

Figure 4.13: Confusion matrix for the Attention-based model.

it with BRFSS 2023 data from the United States. These data sets differ in several aspects, including socioeconomic context, access to healthcare, lifestyle risk factors, and prevalence rates of heart attack. As a result, a model trained solely on BRFSS data may not generalise well to Ethiopian data due to domain shift.

Given the small size of the Ethiopian dataset (100 samples), we applied a few-shot domain adaptation strategy. Specifically, we fine-tuned a DL model pre-trained on BRFSS data using only 20% of the Ethiopian data and validated it on the remaining 80%. And also IsotonicRegression Calibration to fit on calibration set to transform validation probabilities and Thresholding to Converts probabilities to binary predictions using a fixed threshold (0.6).This approach allows the model to adapt to domain-specific features (e.g., unique risk patterns in Ethiopia) while avoiding overfitting. Early stopping was used to maintain generalisation. This few-shot fine-tuning proved effective in improving cross-population performance.For our model, the FNN achieved an accuracy of 75% and a ROC AUC of 0.97 with out calbration and achieved an accuracy of 91% and a ROC AUC of 0.93, with calibration.which is the highest among the models and indicated some level of predictive capacity in the local data set. It showed a balanced performance across both classes with calibration.

Figure4.14 show the performance of the classification model in identifying positive and negative cases. The model correctly predicted 32 true positive cases (TP), which means that it accurately detected 32 instances where the condition was present. It also correctly identified 28 true negative cases (TN), accurately recognising 28 instances in which the condition was absent. However, the model made 12 false positive errors (FP), where it incorrectly predicted the condition as present when it was actually absent, and 8 false negative errors (FN), where it failed to detect the condition despite its presence. These results demonstrate the overall effectiveness of the model while also highlighting specific types of errors that could be addressed to improve accuracy using more local data, equal to public data.

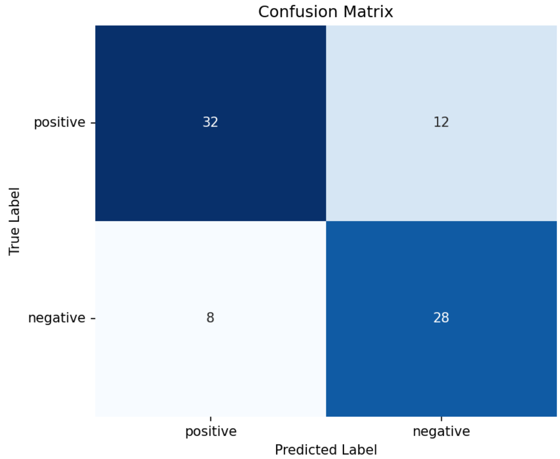

Figure 4.14: Confusion matrix for external validation of FNN model without calibration.

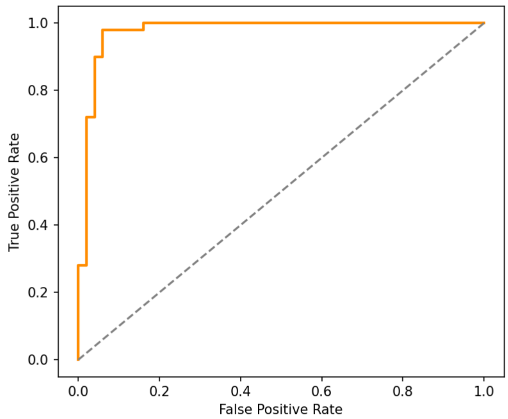

Figure 4.15: ROC for external validation for FNN model without Calibration.

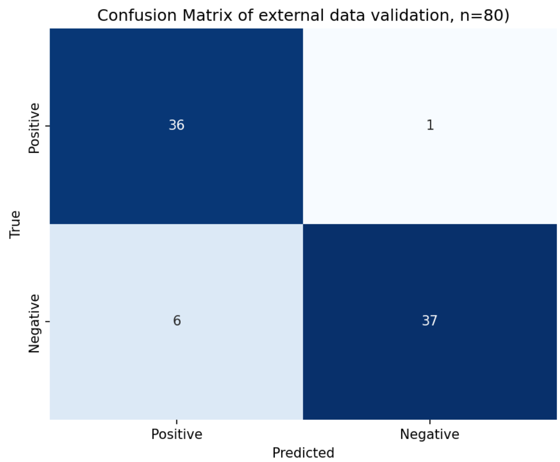

Figure 4.16: Confusion matrix for external validation of FNN model with calibration.

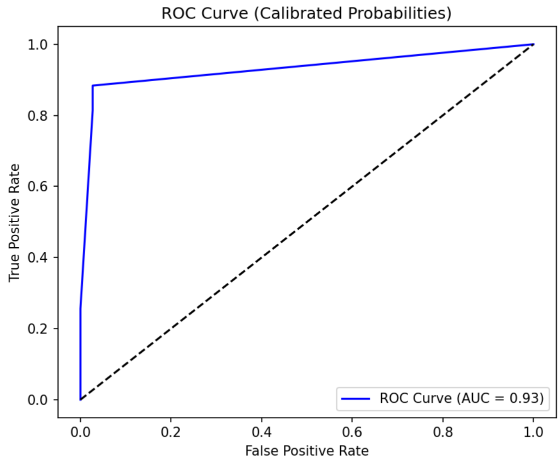

Figure 4.17: ROC for external validation for FNN model with Calibration.

Figure4.16 show the performance of the classification model in identifying positive and negative cases. The model correctly predicted 36 true positive cases (TP), which means that it accurately detected 32 instances where the condition was present. It also correctly identified 37 true negative cases (TN), accurately recognising 37 instances in which the condition was absent. However, the model made 1 false positive errors (FP), where it incorrectly predicted the condition as present when it was actually absent, and 6 false negative errors (FN), where it failed to detect the condition despite its presence. These results demonstrate the overall effectiveness of the model with isotonic regression calibration.

4.2 Graphical user interface

Taking into account the relatively better balance of classification and simplicity in architecture, the FNN model is the most suitable model for integration into a GUI application. For these reasons, the FNN model is selected as the core predictive engine for the proposed heart attack risk assessment GUI application. We developed a Web-Based GUI that makes the heart attack risk prediction model using Streamlit, a Python framework designed to build interactive data applications. Our new GUI will allow healthcare professionals to input clinical and lifestyle information through a carefully selected set of characteristics to examine the risk of heart attack in the new patient.

We organised the GUI into a clean column layout that supports a mix of numeric fields, dropdown menus, and categorical selectors. Upon user submission, the collected input data are sent to the ML model hosted on the back end, which processes the information and returns a heart attack risk prediction, presented as a risk category (Low, Medium, High) and a probability score.

To streamline development and version control, the entire application, including Python scripts, model files, and interface logic, was hosted on GitHub. This allows for collaboration, change tracking, and continuous integration. The GUI was then deployed using the Streamlit Community Cloud, which connects directly to the GitHub repository. Any updates pushed to the repository automatically reflect in the live application, ensuring a seamless development and deployment pipeline.

This integration of GitHub and Streamlit not only enhances transparency and reproducibility but also ensures that the system can be easily maintained and updated as new features or models are added.

[0] https://knsywh55ysigzbzx8zfnpr.streamlit.app

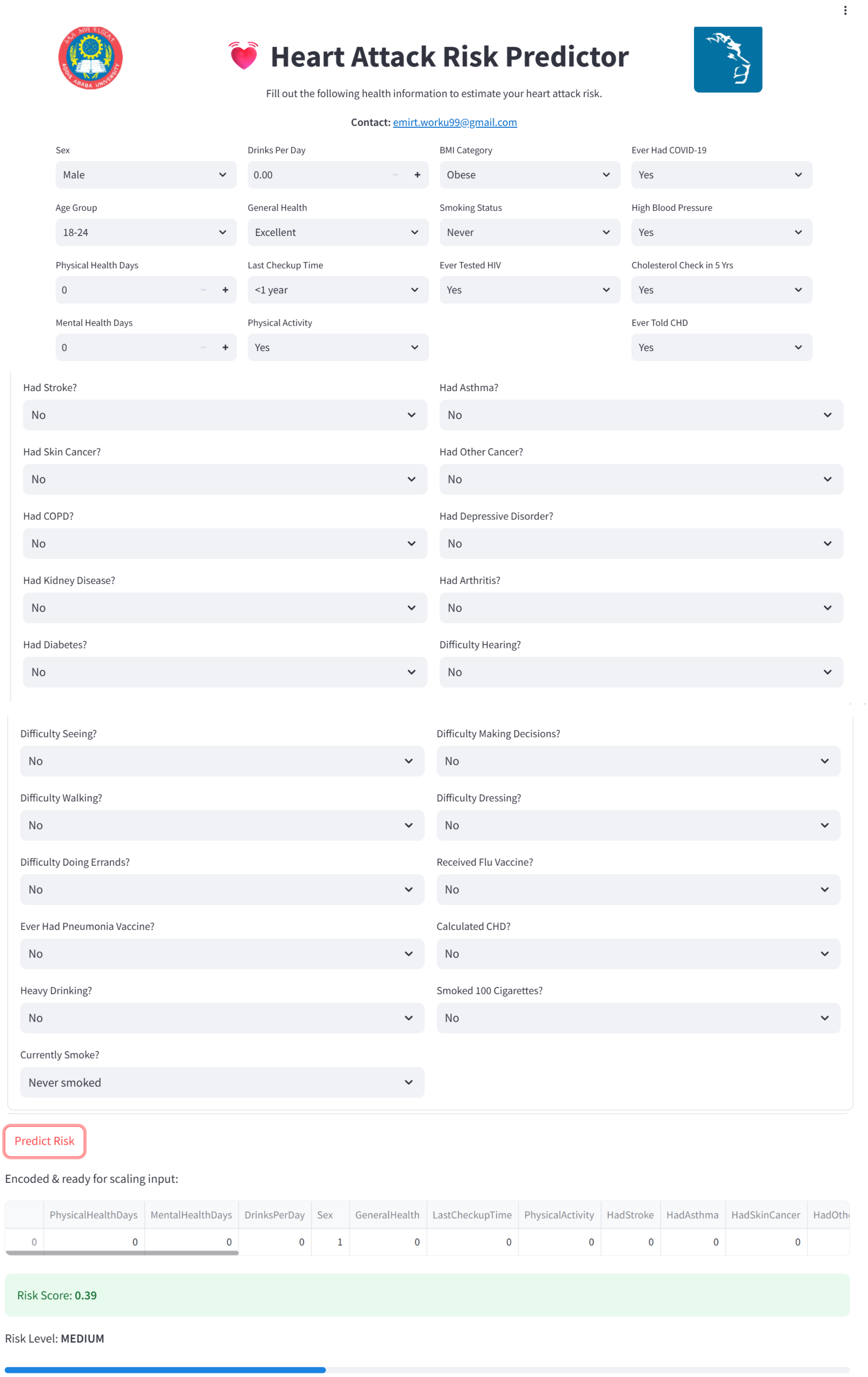

Figure 4.18: Graphical user interface for heart attack risk prediction

4.3 Discussion

In this study, we used classical ML models and DL models to predict heart attack risk using publicly available data from the CDC and externally validated with local data. We found that the DL models perform better than the classical ML models, which is expected. DL models learn these complex feature interactions and handle non-linearity and high-dimensional input, making them better suited for heart attack risk prediction.

The confusion matrix analysis provided a granular understanding of the strengths and weaknesses of the model. The FNN model showed high true positive (TP) and true negative (TN) counts, with relatively low false negatives (FN) and false positives (FP). This balance is vital in medical contexts; minimising FN reduces missed diagnoses and minimising FP avoids unnecessary anxiety and interventions.

The lower FP of the residual model suggests that it is more conservative in positive predictions, potentially reducing false alarms, but risking more missed cases (higher FN). The confusion matrix of the Attention model revealed a more balanced but less optimal separation, consistent with its training behaviour.

Among the models evaluated, RF and GB achieved the highest overall performance with a balanced accuracy of 0.82, while LR and DT also showed strong and comparable results; KNN demonstrated the lowest performance in all metrics. The performance of our ML models is similar to that of previous studies on heart attack risk prediction using similar models [45, 46, 48]. For example, studies used the Cleveland Heart Disease dataset from the UCI repository and achieved 0.78 accuracy using RL, which is comparable to our LR model. Similarly, studies found that ensemble methods such as RF and GB outperformed simpler models, achieving an accuracy of 0.80, which closely matches our results. In contrast, KNN tends to perform poorly in our studies with previous studies, which found a precision of 0.80 [47, 49, 50]. Although our data and features differ from those used in previous studies, our data is larger and incorporates more features. In general, our findings confirm that the ensemble models provide robust and reliable performance for heart attack risk prediction, in agreement with previous machine learning studies on clinical data.

LR and DT showed moderate performance, with comparable accuracy but slightly lower ROC AUCs, indicating that while these models can capture some predictive signals, their discriminative ability is limited compared to ensemble methods. The notably lower recall and F1 scores of KNN suggest that it can be difficult to identify all positive cases. These classical ML results highlight the trade-offs between model complexity, interpretability, and predictive power, suggesting that this model could be a good clinical deployment regardless of model accuracy.

DL models performed significantly better than classical ML approaches in all metrics, highlighting their ability to learn more.FNN emerged as the model that performed the best, achieving the highest balanced accuracy (95%), recall (96%), and ROC AUC (97%). This exceptional performance suggests that the FNN effectively discriminates between heart attack risk-positive patients and non-heart attack risk-negative patients, making it a suitable tool for clinical decision support.

The Wide & Deep model also delivered strong results, with slightly lower recall but comparable precision and ROC AUC. The ability of this architecture to integrate memorisation (wide component) and generalisation (deep component) may provide a balanced approach, especially useful when categorical and continuous features coexist.

The residual network, although deeper and architecturally more complex, showed slightly reduced recall (86%) compared to the FNN and Wide & Deep models. This indicates that increased depth and skip connections improve learning stability, but do not guarantee superior sensitivity to positive cases. The Attention-based model, designed to focus on important features dynamically, performed on par with simpler architectures but showed signs of potential overfitting and less stable training behaviour. This highlights the challenge of effectively tuning more sophisticated models and suggests that model complexity must be carefully balanced against available data and training strategies. The Residual and Attention models showed slightly wider gaps or less consistent improvements, indicating possible overfitting.

Analysis of training curves revealed that all deep models converged well, with training and validation losses steadily decreasing and accuracies rising in tandem. The close alignment between training and validation metrics, particularly in the FNN and Wide & Deep models, reflects good generalisation capability and minimal overfitting, a crucial consideration for real-world applications where models must perform reliably on unseen data. We suggest that these models are a good candidate model for clinical deployment in terms of both predictive power and accuracy.

External validation in a local Ethiopian heart attack risk dataset, distinct from the BRFSS training data, was a critical test of model generalisability. Despite domain shifts due to differing demographics, healthcare infrastructure, and lifestyle factors, the FNN model retained strong predictive power with an ROC AUC of 0.93 and balanced precision around 91%. This promising transferability indicates that the model captures fundamental heart attack risk patterns, but the somewhat reduced accuracy highlights the need for local data integration.

The few-shot domain adaptation approach, fine-tuning the pre-trained model with a small fraction of local data, proved effective in bridging population differences and improving performance without extensive retraining. This strategy is particularly important for the deployment of DL in various healthcare settings where large annotated data sets are scarce, such as in our country, Ethiopia.

In general, these findings underscore the superiority of DL, especially the FNN architecture, in predicting heart attack risk compared to classical ML models. The consistent high ROC AUC across deep models confirms their strong discriminatory ability, critical for clinical risk stratification.

However, the observed variation in recall and false negative rates between models points to the need for context-aware model selection based on clinical priorities. Although the FNN model offers the best overall balance, other architectures might be preferred in settings that emphasise sensitivity or specificity.

Training and validation loss patterns reinforce the need for careful model development to avoid overfitting and maintain generalisability. Successful external validation demonstrates the value of domain adaptation techniques in enhancing the applicability of the model between populations, addressing health equity concerns.

Future research should focus on increasing the size and diversity of local datasets to further validate and refine these models. In addition, exploring the interpretability and expandability of the model will be essential to building trust among healthcare providers.

Chapter 5

Conclusion

5.1 Summary of the study

We conducted a comprehensive evaluation of heart attack risk predictive models of a wide range of classical ML models, including LR, KNN, DT, RF, and GB. In addition, we implemented DL models, such as FNN, Wide & Deep models, Residual Networks, and attention-based architectures. Our results demonstrated that the DL models, the FNN model, achieved the highest overall performance, with an accuracy of 95%, an F1 score of 0.96, and an AUC of 0.97. Two distinct data sets were used: a large publicly available data set from the BRFSS survey, supported by the CDC, for model training and development, and clinical data collected locally from TASH in Ethiopia for external validation. This combination allowed us to build predictive models for heart attack risk that capture both broad population-level trends and local healthcare characteristics.

Our findings indicate that deep learning models outperform classical approaches, with the Feedforward Neural Network demonstrating the best overall performance, achieving an accuracy of 95%, an F1 score of 0.96, and an ROC AUC of 0.97. These results highlight the strong capability of DL models to effectively capture complex and nonlinear patterns inherent in the data, making them highly suitable for heart attack risk prediction tasks. Furthermore, external validation in the Ethiopian dataset underscores the importance of domain adaptation to maintain predictive accuracy in diverse populations. Finally, the integration of the best-performing model into a graphical user interface demonstrates the practical potential for clinical decision support, allowing healthcare professionals to efficiently assess patient risk and facilitate timely interventions.

5.2 Contributions of the study

This study makes several important contributions.First, we developed a user-friendly graphical user interface (GUI) application that enables real-time prediction of heart attack risk. This tool can support both healthcare professionals and patients by providing quick, data-driven risk assessments.Second, the study demonstrates the feasibility and effectiveness of combining machine learning (ML) and deep learning (DL) techniques for heart attack prediction using structured clinical data.Third, we conducted a comparative analysis of multiple model architectures on the same dataset, providing valuable insights into which algorithms perform best for this type of medical data.Fourth, we validated the models externally using real hospital data collected in Ethiopia, ensuring the results are grounded in real-world clinical conditions.Finally, the integration of these models into a GUI-based application provides a practical and accessible tool that could be incorporated into clinical workflows, supporting evidence-based decision-making in healthcare settings.

5.3 Recomendation

We recommend that health facilities invest in adequate data collection and digitisation systems to improve the quality of clinical records, which are critical to reliable predictions. Healthcare increasingly relies on structured electronic patient record data to support research and facilitate the use of ML and DL tools for early detection and risk prediction of heart attacks.

5.4 Future work

Future research should focus on expanding the data set by including multicenter and longitudinal data to enhance the robustness of the model and capture the temporal dynamics of patient health. Incorporating more detailed patient histories and lifestyle information could also improve predictive accuracy. In addition, interpretable models are needed to ensure transparency and trust in medical decision-making. Finally, future studies should consider integrating the developed models into real-time hospital systems and evaluating their impact on clinical outcomes through prospective trials.

References

[1] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, “Effective heart disease prediction using machine learning techniques,” *Algorithms*, vol. 16, no. 2, p. 88, 2023.

[2] M. B. Abubaker and B. Babayiğit, “Detection of cardiovascular diseases in ecg images using machine learning and deep learning methods,” *IEEE transactions on artificial intelligence*, vol. 4, no. 2, pp. 373–382, 2022.

[3] L. Chen, Z. Han, J. Wang, and C. Yang, “The emerging roles of machine learning in cardiovascular diseases: a narrative review,” *Annals of Translational Medicine*, vol. 10, no. 10, p. 611, 2022.

[4] M. M. Ahsan and Z. Siddique, “Machine learning-based heart disease diagnosis: A systematic literature review,” *Artificial Intelligence in Medicine*, vol. 128, p. 102289, 2022.

[5] P. Mathur, S. Srivastava, X. Xu, and J. L. Mehta, “Artificial intelligence, machine learning, and cardiovascular disease,” *Clinical Medicine Insights: Cardiology*, vol. 14, p. 1179546820927404, 2020.

[6] K. K. Wong, G. Fortino, and D. Abbott, “Deep learning-based cardiovascular image diagnosis: a promising challenge,” *Future Generation Computer Systems*, vol. 110, pp. 802–811, 2020.

[7] H. Sadr, A. Salari, M. T. Ashoobi, and M. Nazari, “Cardiovascular disease diagnosis: a holistic approach using the integration of machine learning and deep learning models,” *European Journal of Medical Research*, vol. 29, no. 1, p. 455, 2024.

[8] R. Alizadehsani, M. Abdar, M. Roshanzamir, A. Khosravi, P. M. Kebria, F. Khozeimeh, S. Nahavandi, N. Sarrafzadegan, and U. R. Acharya, “Machine learning-based coronary artery disease diagnosis: A comprehensive review,” *Computers in biology and medicine*, vol. 111, p. 103346, 2019.

[9] F. Girlanda, O. Demler, B. Menze, and N. Davoudi, “Enhancing cardiovascular disease prediction through multi-modal self-supervised learning,” *arXiv preprint arXiv:2411.05900*, 2024.

[10] H. T. Mekonnen and M. M. Woldeyohannis, “Towards predicting the risk of cardiovascular disease using machine learning approach,” in *International Conference on Advances of Science and Technology*. Springer, 2021, pp. 489–499.

[11] D. A. Angaw, R. Ali, A. Tadele, and S. Shumet, “The prevalence of cardiovascular disease in ethiopia: a systematic review and meta-analysis of institutional and community-based studies,” *BMC cardiovascular disorders*, vol. 21, no. 1, pp. 1–9, 2021.

[12] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, “Can machine-learning improve cardiovascular risk prediction using routine clinical data?” *PloS one*, vol. 12, no. 4, p. e0174944, 2017.

[13] K. Shameer, K. W. Johnson, B. S. Glicksberg, J. T. Dudley, and P. P. Sengupta, “Machine learning in cardiovascular medicine: are we there yet?” *Heart*, vol. 104, no. 14, pp. 1156–1164, 2018.

[14] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, “Risk prediction of cardiovascular disease using machine learning classifiers,” *Open Medicine*, vol. 17, no. 1, pp. 1100–1113, 2022.

[15] L. I. Katzel and S. R. Waldstein, “Classification of cardiovascular disease,” in *Neuropsychology of cardiovascular disease*. Psychology Press, 2001, pp. 19–30.

[16] P. Podolec, G. Kopeć, P. Rubiś, J. Stkepniewski, J. Podolec, M. Komar, L. Tomkiewicz-Pajkak, A. Leśniak-Sobelga, A. Kabłak-Ziembicka, and P. T. Matusik, “Clinical classification of rare cardiovascular diseases and disorders: 2018 update,” *J Rare Cardiovasc Dis*, vol. 3, no. 7, pp. 233–239, 2018.

[17] E. Antman, J.-P. Bassand, W. Klein, M. Ohman, J. L. Lopez Sendon, L. Rydén, M. Simoons, and M. Tendera, “Myocardial infarction redefined—a consensus document of the joint european society of cardiology/american college of cardiology committee for the redefinition of myocardial infarction: the joint european society of cardiology/american college of cardiology committee,” *Journal of the American College of Cardiology*, vol. 36, no. 3, pp. 959–969, 2000.

[18] K. Thygesen, J. S. Alpert, A. S. Jaffe, B. R. Chaitman, J. J. Bax, D. A. Morrow, H. D. White, and E. G. on behalf of the Joint European Society of Cardiology (ESC)/American College of Cardiology (ACC)/American Heart Association (AHA)/World Heart Federation (WHF) Task Force for the Universal Definition of Myocardial Infarction, "Fourth universal definition of myocardial infarction (2018)," *Journal of the American college of cardiology*, vol. 72, no. 18, pp. 2231–2264, 2018.

[19] E. Escobar, "Hypertension and coronary heart disease," *Journal of human hypertension*, vol. 16, no. 1, pp. S61–S63, 2002.

[20] C. W. Tsao, A. W. Aday, Z. I. Almarzooq, C. A. Anderson, P. Arora, C. L. Avery, C. M. Baker-Smith, A. Z. Beaton, A. K. Boehme, A. E. Buxton *et al.*, "Heart disease and stroke statistics—2023 update: a report from the american heart association," *Circulation*, vol. 147, no. 8, pp. e93–e621, 2023.

[21] R. Hajar, "Risk factors for coronary artery disease: historical perspectives," *Heart views*, vol. 18, no. 3, pp. 109–114, 2017.

[22] R. L. Sacco and T. Rundek, "Cerebrovascular disease," *Current opinion in neurology*, vol. 25, no. 1, pp. 1–4, 2012.

[23] H. M. Rifin, M. A. Omar, K. S. Wan, and W. S. Rodzlan Hasani, "10-year risk for cardiovascular diseases according to the who prediction chart: findings from the national health and morbidity survey (nhms) 2019," *BMC Public Health*, vol. 24, no. 1, p. 2513, 2024.

[24] G. A. Roth, G. A. Mensah, C. O. Johnson, G. Addolorato, E. Ammirati, L. M. Baddour, N. C. Barengo, A. Z. Beaton, E. J. Benjamin, C. P. Benziger *et al.*, "Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the gbd 2019 study," *Journal of the American college of cardiology*, vol. 76, no. 25, pp. 2982–3021, 2020.

[25] C. W. Tsao, A. W. Aday, Z. I. Almarzooq, A. Alonso, A. Z. Beaton, M. S. Bittencourt, A. K. Boehme, A. E. Buxton, A. P. Carson, Y. Commodore-Mensah *et al.*, "Heart disease and stroke statistics—2022 update: a report from the american heart association," *Circulation*, vol. 145, no. 8, pp. e153–e639, 2022.

[26] C. X. Wong, A. Brown, D. H. Lau, S. S. Chugh, C. M. Albert, J. M. Kalman, and P. Sanders, "Epidemiology of sudden cardiac death: global and regional perspectives," *Heart, Lung and Circulation*, vol. 28, no. 1, pp. 6–14, 2019.

[27] R. A. Nishimura, C. M. Otto, R. O. Bonow, B. A. Carabello, J. P. Erwin III, L. A. Fleisher, H. Jneid, M. J. Mack, C. J. McLeod, P. T. O'Gara *et al.*, "2017 aha/acc focused update of the 2014 aha/acc guideline for the management of patients with valvular heart disease: a report of the american college of cardiology/american heart association task force on clinical practice guidelines," *Circulation*, vol. 135, no. 25, pp. e1159–e1195, 2017.

[28] J. I. Hoffman, "The global burden of congenital heart disease," *Cardiovascular journal of Africa*, vol. 24, no. 4, pp. 141–145, 2013.

[29] J. W. Rossano, "Congenital heart disease: a global public health concern," *The Lancet Child & Adolescent Health*, vol. 4, no. 3, pp. 168–169, 2020.

[30] S. Sagar, P. P. Liu, and L. T. Cooper, "Myocarditis," *The Lancet*, vol. 379, no. 9817, pp. 738–747, 2012.

[31] A. Misganaw, D. H. Mariam, A. Ali, and T. Araya, "Epidemiology of major non-communicable diseases in ethiopia: a systematic review," *Journal of health, population, and nutrition*, vol. 32, no. 1, p. 1, 2014.

[32] A. Adem, D. Bacha, and A. M. Argaw, "Pattern of cardiovascular diseases at a teaching hospital in addis ababa, ethiopia: An echocardiographic study of 1500 patients," *Medicine*, vol. 102, no. 34, p. e34795, 2023.

[33] F. Tesfaye, P. Byass, and S. Wall, "Population based prevalence of high blood pressure among adults in addis ababa: uncovering a silent epidemic," *BMC cardiovascular disorders*, vol. 9, pp. 1–10, 2009.

[34] T. A. Mega and N. E. Dabe, "Khat (catha edulis) as a risk factor for cardiovascular disorders: systematic review and meta-analysis," *The Open Cardiovascular Medicine Journal*, vol. 11, p. 146, 2017.

[35] Y. G. Tefera, T. B. Abebe, A. B. Mekuria, M. S. Kelkay, and T. M. Abegaz, "Prescribing trend in cardiovascular patients at ethiopian university hospital: The number of medications and implication on the clinical improvement," *Pharmacology Research & Perspectives*, vol. 7, no. 3, p. e00474, 2019.

[36] O. Non-Communicable, "National strategic action plan (nsap) for prevention & control of non-communicable diseases in ethiopia," *Google Scholar*, 2016.

[37] Y. Mamo, E. Seid, S. Adams, A. Gardiner, and E. Parry, "A primary healthcare approach to the management of chronic disease in ethiopia: an example for other countries," *Clinical Medicine*, vol. 7, no. 3, pp. 228–231, 2007.

[38] G. C. R. Consortium, "Global effect of modifiable risk factors on cardiovascular disease and mortality," *New England Journal of Medicine*, vol. 389, no. 14, pp. 1273–1285, 2023.

[39] E. Goldsborough III, E. Tasdighi, and M. J. Blaha, "Assessment of cardiovascular disease risk: a 2023 update," *Current opinion in lipidology*, vol. 34, no. 4, pp. 162–173, 2023.

[40] R. Johnson and M. Lee, "Atherosclerotic cardiovascular disease risk prediction: current state and future directions," *Circulation*, vol. 147, no. 12, pp. 1045–1056, 2023. [Online]. Available: https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.122.056993

[41] D. Chipayo-Gonzales, H. Ramakrishna, and I. J. Nuñez-Gil, "Score2: A new updated algorithm to predict cardiovascular disease risk in europe," *Journal of Cardiothoracic and Vascular Anesthesia*, vol. 36, no. 1, pp. 18–21, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053077021008478

[42] S. W. Group and E. C. R. Collaboration, "Score2: A new updated algorithm to predict cardiovascular disease risk in europe," *European Heart Journal*, vol. 42, no. 25, pp. 2439–2453, 2021. [Online]. Available: https://academic.oup.com/eurheartj/article/42/25/2439/6297709

[43] A. Vusirikala, F. Wekesah, C. Kyobutungi, and O. Oyebode, "Assessment of cardiovascular risk in a slum population in kenya: use of world health organisation/international society of hypertension (who/ish) risk prediction charts-secondary analyses of a household survey," *BMJ open*, vol. 9, no. 9, p. e029304, 2019.

[44] T. Shah, J. Engmann, C. Dale, S. Shah, J. White, C. Giambartolomei, S. McLachlan, D. Zabaneh, A. Cavadino, C. Finan *et al.*, "Population genomics of cardiometabolic traits: design of the university college london-london school of hygiene and tropical medicine-edinburgh-bristol (ucleb) consortium," *PloS one*, vol. 8, no. 8, p. e71345, 2013.

[45] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in *2017 IEEE symposium on computers and communications (ISCC)*. IEEE, 2017, pp. 204–207.

[46] T. K. Sajja and H. K. Kalluri, "A deep learning method for prediction of cardiovascular disease using convolutional neural network." *Rev. d'Intelligence Artif.*, vol. 34, no. 5, pp. 601–606, 2020.

[47] K. Vayadande *et al.*, "Heart disease prediction using machine learning and deep learning algorithms," in *2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*. IEEE, 2022, pp. 393–401.

[48] E. J. Benjamin, P. Muntner, A. Alonso, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, A. R. Chang, S. Cheng, S. R. Das, F. N. Delling, L. Djousse, M. S. V. Elkind, J. F. Ferguson, M. Fornage, L. C. Jordan, S. S. Khan, B. M. Kissela, K. L. Knutson, T. W. Kwan, D. T. Lackland, T. T. Lewis, J. H. Lichtman, C. T. Longenecker, M. S. Loop, P. L. Lutsey, S. S. Martin, K. Matsushita, A. E. Moran, M. E. Mussolino, M. O'Flaherty, A. Pandey, A. M. Perak, W. D. Rosamond, G. A. Roth, U. K. A. Sampson, G. M. Satou, E. B. Schroeder, S. H. Shah, N. L. Spartano, A. Stokes, D. L. Tirschwell, C. W. Tsao, M. P. Turakhia, L. B. VanWagner, J. T. Wilkins, S. S. Wong, and S. S. Virani, "Heart disease and stroke statistics—2019 update: A report from the american heart association," *Circulation*, vol. 139, no. 10, pp. e56–e528, 2019. [Online]. Available: https://doi.org/10.1161/CIR.0000000000000659

[49] P. A. Heidenreich *et al.*, “Heart disease and stroke statistics—2022 update: A report from the american heart association,” *Circulation*, vol. 145, no. 8, pp. e153–e639, 2022.

[50] D. Shah, S. Patel, and S. K. Bharti, “Heart disease prediction using machine learning techniques,” *SN Computer Science*, vol. 1, no. 6, p. 345, 2020.

[51] M. Dorraki, Z. Liao, D. Abbott, P. J. Psaltis, E. Baker, N. Bidargaddi, H. R. Wardill, A. van den Hengel, J. Narula, and J. W. Verjans, “Improving cardiovascular disease prediction with machine learning using mental health data: a prospective uk biobank study,” *JACC: Advances*, vol. 3, no. 9_Part_2, p. 101180, 2024.

[52] B. Pan, “Predicting heart disease based on wide and deep neural network,” *Applied and Computational Engineering*, vol. 2, pp. 174–179, 03 2023.

[53] O. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, “An integrated decision support system based on ann and fuzzy-ahp for heart failure risk prediction,” *Expert Systems with Applications*, vol. 68, pp. 163–172, 2017.

[54] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, “An automated diagnostic system for heart disease prediction based on χ^2 statistical model and optimally configured deep neural network,” *IEEE Access*, vol. 7, pp. 34 938–34 945, 2019.

[55] A. Mehmood *et al.*, “Prediction of heart disease using deep convolutional neural networks,” *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 3409–3422, 2021.

[56] H. Das, B. Naik, H. S. Behera, S. Jaiswal, P. Mahato, and M. Rout, “Biomedical data analysis using neuro-fuzzy model with post-feature reduction,” *Journal of King Saud University-Computer and Information Sciences*, 2020.

[57] A. Nancy, D. Ravindran, P. Raj Vincent, K. Srinivasan, and D. Gutierrez Reina, “Iot-cloud-based smart healthcare monitoring system for heart disease prediction via deep learning,” *Electronics*, vol. 11, no. 15, p. 2292, 2022.

[58] M. T. García-Ordás, M. Bayón-Gutiérrez, C. Benavides, J. Aveleira-Mata, and J. A. Benítez-Andrades, “Heart disease risk prediction using deep learning techniques with feature augmentation,” *Multimedia Tools and Applications*, vol. 82, no. 20, pp. 31 759–31 773, 2023.

[59] A. A. Almazroi, E. A. Aldhahri, S. Bashir, and S. Ashfaq, “A clinical decision support system for heart disease prediction using deep learning,” *IEEE Access*, vol. 11, pp. 61 646–61 659, 2023.

[60] A. Pandey, B. A. Shivaji, M. Acharya, and K. K. Mohbey, “Mitigating class imbalance in heart disease detection with machine learning,” *Multimedia Tools and Applications*, pp. 1–26, 2024.

[61] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *arXiv e-prints*, p. arXiv:1106.1813, Jun. 2011.

[62] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, “Learning from class-imbalanced data: Review of methods and applications,” *Expert systems with applications*, vol. 73, pp. 220–239, 2017.

[63] C. C. Aggarwal and C. C. Aggarwal, *An introduction to outlier analysis*. Springer, 2017.

[64] J. W. Tukey *et al.*, *Exploratory data analysis*. Springer, 1977, vol. 2.

Chapter A

Features in the Dataset from BRFSS 2023 Core Variable Definitions

Table A.1: Numerical Features in the Dataset

Feature Name	Description
PhysicalHealthDays	Number of days in the past month with poor physical health
MentalHealthDays	Number of days in the past month with poor mental health
DrinksperDays	Drink-occasions-per-day

Table A.2: Categorical Features in the Dataset

Features	Description
Sex	Sex of respondent (male/female)
GeneralHealth	Self-rated general health
LastCheckupTime	Time since last routine health checkup
PhysicalActivity	Whether engaged in any physical activity in the past 30 days
HadHeartAttack	Ever told by doctor had heart attack
HadStroke	Ever told by doctor had stroke
HadAsthma	Ever told by doctor had asthma
HadSkinCancer	Ever had skin cancer diagnosis
HadOtherCancer	Ever had other cancer diagnosis (non-skin)
HadCOPD	Ever told doctor had chronic obstructive pulmonary disease
HadDepressiveDisorder	Ever told by doctor had depressive disorder
HadKidneyDisease	Ever told by doctor had kidney disease
HadArthritis	Ever told by doctor had arthritis
HadDiabetes	Ever told by doctor had diabetes
DifficultyHearing	Self-reported hearing difficulty
DifficultySeeing	Self-reported vision difficulty
DifficultyMakingDecisions	Cognitive difficulty of making decisions
DifficultyWalking	Mobility difficulty
DifficultyDressing	Self-reported difficulty dressing or bathing
DifficultyErrands	Self-reported difficulty doing errands alone
AgeGroup	Age grouped into categories
BMICategory	Body mass index category
EverTestedHIV	Whether respondent ever tested for HIV
ReceivedFluVax	Whether respondent received influenza vaccination in past 12 months
EverHadPneumoniaVax	Ever had pneumococcal pneumonia vaccination
EverHadCOVID	Ever had COVID-19 (self-reported)
HighBloodPressure	Ever told by doctor had high blood pressure (hypertension)
CholesterolCheck5yrs	Had cholesterol checked in past 5 years
EverToldCHD	Ever told had coronary heart disease
CalculatedCHD	Calculated indicator combining heart diagnosis variables
HeavyDrinking	Binge or heavy drinking
Smoked100Cigarettes	Ever smoked at least 100 cigarettes in lifetime
CurrentlySmoke	Current smoking status (yes/no)

Chapter B

python code

B.1 code Listing

```
# Import necessary libraries

# Basic libraries for numerical operations, data manipulation, and visualization
import numpy as np  # For numerical computations
import pandas as pd  # For data manipulation and analysis
import seaborn as sns  # For statistical data visualization
import matplotlib.pyplot as plt  # For creating plots and figures
import warnings  # To handle warnings in the code
warnings.filterwarnings('ignore')  # Ignore warnings to reduce clutter

# Data Preprocessing and Feature Scaling
from sklearn.preprocessing import StandardScaler, MinMaxScaler
# For feature scaling (Standard and Min-Max scaling)

# Data Splitting and Model Evaluation
from sklearn.model_selection import train_test_split, cross_val_score,
GridSearchCV, RandomizedSearchCV, cross_validate
# For splitting data, cross-validation, and hyperparameter tuning

from sklearn.metrics import classification_report, accuracy_score, confusion_matrix
# For model evaluation and performance metrics

# Handling Imbalanced Datasets
from imblearn.over_sampling import SMOTE  # For handling imbalanced datasets

# Machine Learning Models (for classification)
from sklearn.linear_model import LogisticRegression
# Logistic Regression for binary classification
from sklearn.neural_network import MLPClassifier
# Multi-layer Perceptron (Neural Network) for classification
from sklearn.svm import SVC  # Support Vector Machine classifier

from sklearn.ensemble import (
    RandomForestClassifier,  # Random Forest classifier
    GradientBoostingClassifier,  # Gradient Boosting classifier
    AdaBoostClassifier,  # AdaBoost classifier
    ExtraTreesClassifier,
    StackingClassifier,  # Stacking classifier
    VotingClassifier  # Voting classifier
)

from xgboost import XGBClassifier  # XGBoost classifier

from sklearn.metrics import (
    accuracy_score, precision_score, recall_score,
    f1_score, roc_auc_score, classification_report, roc_curve
```

```
)

# Deep Learning Libraries
import tensorflow as tf
from tensorflow.keras.models import Sequential, Model
from tensorflow.keras.layers import Dense, Dropout, BatchNormalization,
Conv1D, MaxPooling1D, Flatten, Input, LSTM, Concatenate, Add, Attention

# Model Interpretability (SHAP)
import shap
```

All the code used in this study is available on GitHub at: https://github.com/emirt-worku/cvd/tree/main. This includes scripts for CVD data preprocessing, data balancing, machine learning models, deep learning models, a Streamlit-based GUI application and how the data extract from ascii file data and converted to csv file.

Chapter C

Categorical attribute

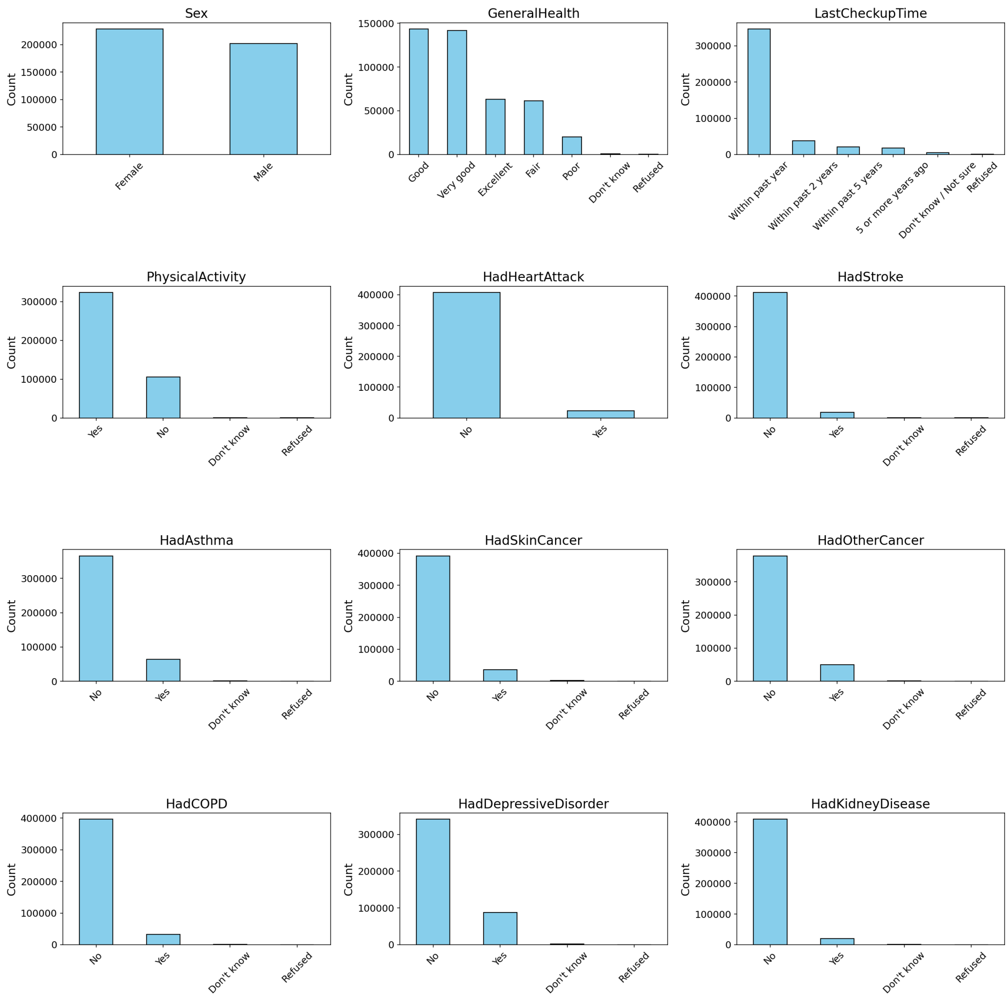

Figure C.1: Categorical attribute distribution of publicly available data from the CDC 2023 BRFSS.

article booktabs multirow array siunitx

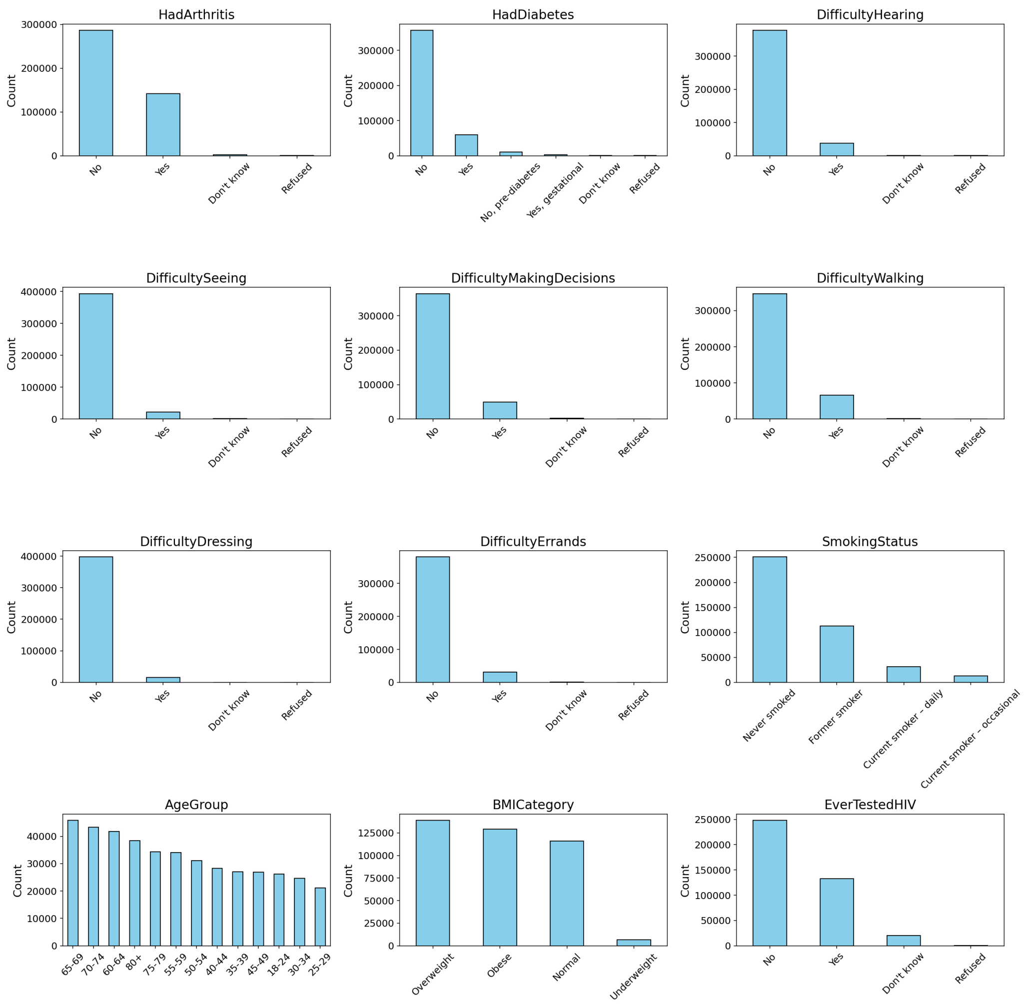

Figure C.2: The same description as figure C.1

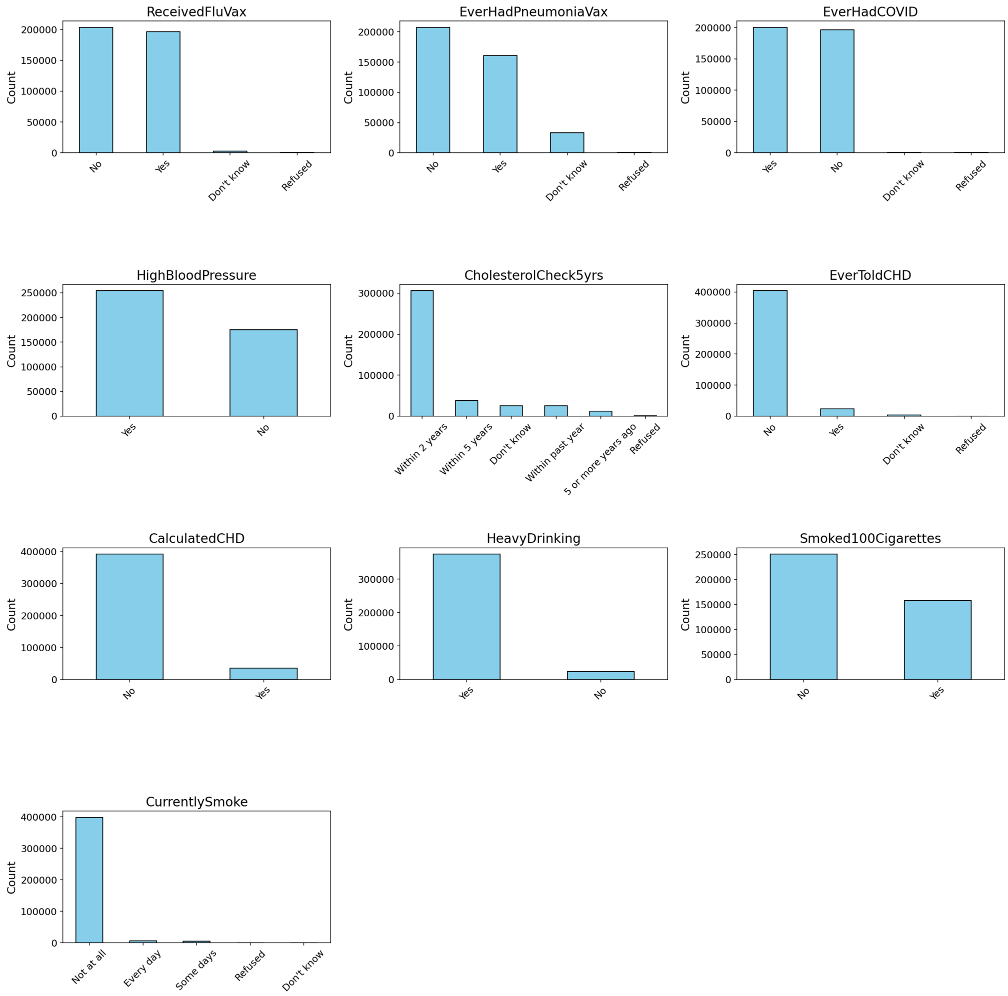

Figure C.3: The same description as figure C.1

Table C.1: Feature Importance Rankings Across Different Methods

Feature	**F-score**	**RF Importance**	**RFE Selected**	**RFE Ranking**
CalculatedCHD	595988.356885	0.692768	True	1
EverToldCHD	11007.555663	0.092429	True	1
AgeGroup	7031.411198	0.037222	False	6
HighBloodPressure	4879.398291	0.022659	False	19
EverHadPneumoniaVax	2998.560743	0.010342	True	1
HadStroke	2866.317791	0.015558	True	1
DifficultyWalking	2713.564242	0.010762	False	16
HadArthritis	2594.770293	0.010467	False	13
Smoked100Cigarettes	2287.790705	0.005250	False	2
HadCOPD	2187.768232	0.006831	True	1
HadDiabetes	2183.114820	0.005399	True	1
HadKidneyDisease	1443.647894	0.002993	True	1
SmokingStatus	1380.546324	0.005920	True	1
DifficultyHearing	1366.081402	0.002901	False	11
Sex	1342.708813	0.006677	True	1
LastCheckupTime	819.898937	0.002209	True	1
DifficultyErrands	810.815176	0.001746	False	9
HadOtherCancer	788.953602	0.002521	False	15
HadSkinCancer	669.537604	0.002223	False	12
ReceivedFluVax	633.129102	0.002578	True	1
DifficultyDressing	606.483903	0.001386	False	10
PhysicalHealthDays	496.093817	0.010527	False	21
DifficultyMakingDecisions	363.471107	0.001885	False	5
DrinksPerDay	271.627732	0.006730	False	22
PhysicalActivity	234.173296	0.002502	True	1
DifficultySeeing	205.067038	0.001765	False	4
HadDepressiveDisorder	199.053592	0.002213	False	17
GeneralHealth	177.175260	0.006488	False	8
EverTestedHIV	144.016813	0.003044	False	18
MentalHealthDays	123.617260	0.011022	False	20
EverHadCOVID	100.679967	0.002778	False	3
HeavyDrinking	89.772569	0.000767	True	1
HadAsthma	77.862670	0.002006	True	1
CholesterolCheck5yrs	52.155501	0.002303	True	1
BMICategory	16.689454	0.004333	False	7
CurrentlySmoke	0.150693	0.000796	False	14