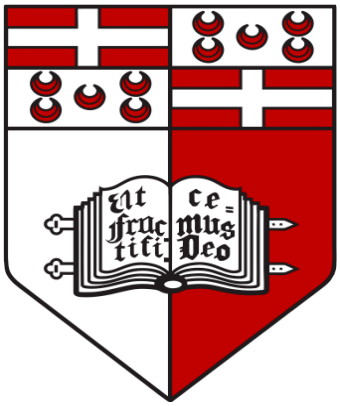
Evaluating Machine Learning models for Cardiac Irregularities (Early Detection, Prediction for Heart Disease)

Adriano Brizi

Supervisor: DR. Godwin Caruana



University of Malta

May 2025

Submitted in partial fulfillment of the requirements for the degree of

B.Sc. (Hons.) in Business and Information Technology

Abstract

The following research's purpose is to tackle one of the most significant global health challenges, responsible for millions of deaths annually. Heart diseases are a serious and common threat to nowadays world and, despite constant medical advances, predicting cardiovascular diseases remains a problem. To address this issue, after a personal experience regarding heart disease, the researcher focused on understanding and creating four machine learning models with limited domain knowledge. Random Forest, Support Vector Machines, Deep Learning and XGBoost are the models leveraged in this study, which, once applied to public available datasets, provided unexpected results.

The datasets used differ in size, quality and feature composition which mirror real-world conditions of clinical practice. Methods such as normalization, encoding, and Synthetic Minority oversampling technique (SMOTE) were implemented directly on the datasets to enhance the model performance when it come to accuracy, recall and precision. The final results showed how XGBoost is the most consistent and reliable model in between datasets. However, Deep learning, while being the second best, it provided unexpected results when working with small datasets. The findings underline how, to improve timing and precision in early detection of cardiovascular diseases, it is important that proper preprocessing and handling of data is performed, together with choosing the right model.

This study's purpose is to show the base of how important the integration of machine learning models with domain experts' judgement is. Future research is certainly needed, and the thesis suggests practical directions to further improve clinical applicability in cardiac care.

Acknowledgments

The current research would have not been possible without the guidance of my supervisor, Godwin Caruana and co-supervisor Joseph Bonello. A special thanks to their invaluable guidance and domain knowledge which made it possible for me to complete this dissertation. Your patience surely contributed to my understanding and appreciation of the treated topic.

I am extremely grateful to my family. To my mother who has always been by my side, to my father who raised me to be ambitious, and to my older brother who is the reference point of my future goals. Without their mental and financial support, I would have never pursued such satisfying academic journey.

Lastly but not of importance, a special thanks to my girlfriend who lifted me up when I was too comfortable laying down. Who pushed me to achieve more when my work felt enough.

Finally, a thanks to myself, who through tough cardiovascular and personal experiences, managed to make the most out of it, enhancing both my knowledge in medicine and machine learning.

Table of Contents

List of Figures

List of tables

1. Introduction

The heart can be defined as a constant perfectly functioning machine. It is the first thing developing when we are conceived and the last thing moving when a life ends. It is the source of life as we know it. However, perfection does not exist and there can be many factors in life which can alter the functioning of a cardiac muscle. The Human being has a fragile essence, and anything we experience does in fact make the heart beating differ from one moment to another. When we train, when happy, when sad or stressed, the heart rhythm changes. Some rhythms are healthy, some others less. A constant unhealthy rhythm, either for genetic reasons or for the wrong lifestyle choices, can bring the heart to not function properly. This is what leads to the development of heart diseases.

Heart diseases are the leading cause of death all around the world. CVD's (Cardiac Vascular Diseases) take an estimated 20 million lives each year. These include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. (World Heart Federation, 2023). Hospitals and researchers are constantly working to avoid premature deaths and increase people's lifespan. Together with governments, who impose taxes on harmful substances such as cigarettes and alcohol. These substances, unhealthy dies and sedentary lives are the main reason of CVD's. The below table provides the numbers and statistics of CVD's deaths in 2021.

	MODIFIABLE RISK FACTOR & ATTRIBUTABLE BURDEN	2021 CVD DEATHS	% OF BURDEN
1	High systolic blood pressure (hypertension)	10.4 million	54%
2	Dietary risks (poor diet)	5.8 million	30%
3	Air pollution (ambient particulate matter pollution)	4.1 million	23%
4	High LDL cholesterol (raised cholesterol)	3.6 million	19%
5	Tobacco (cigarette smoking; second-hand smoke)	2.8 million	15%
6	High fasting plasma glucose (diabetes)	2.2 million	11%
7	Kidney dysfunction (renal failure)	2.1 million	11%
8	High body-mass index (obesity and excess weight)	1.9 million	10%

Table 1.1 - CVD's deaths in 2021 (British Heart Foundation, 2023)

The Early detection of cardiac irregularities is essential for professionals' intervention. It allows to avoid drastic inconveniences and improves patients' lives.
The technological Era in which we live in allows balance and improvement in multiple fields. Disruptive technologies such as Artificial Intelligence, have completely changed healthcare diagnostics. In the early 2000's initial prototypes of AI systems introduced machine learning to analyze electronic health records (EHRs) and medical images (Keragon.com, 2021).

However, true advancement was recognized in the last years, due to an enhancement of computational tools and hardware which before were not powerful enough to solve modern problems.

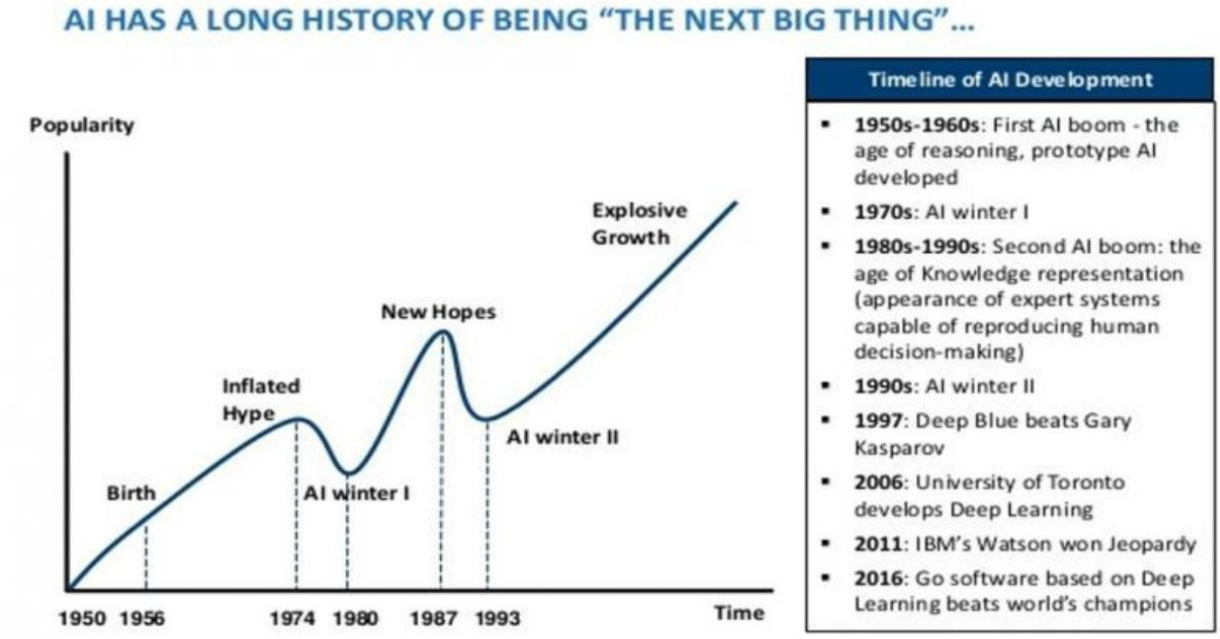


Figure 1.1 - AI timeline (Thedatageneralist.com, 2025)

AI had three "hype" phases. The first two ended up in what so called AI winter, where the expectations for AI did not achieve what was imagined. For this reason, research reduced. The third AI boom represents the present, which after the development of deep learning by the university of Toronto in 2006, the development curve started growing exponentially.

A subset, and important part of AI is machine learning (ML). Machine learning is considered the base of AI. It is used to understand human behavior, create algorithms and self-learning models. (Staff, 2024). A ML model uses data sets to train itself and recognize patterns. It then uses the model to make predictions on new, unseen data. Machine learning has now achieved high accuracy in many fields. Healthcare is one of them.

Machine learning models have shown great promise in identifying patterns in electrocardiogram (ECG) readings, MRI scans, Blood pressure measurements and other diagnostic outputs that human experts might not recognize. Through supervised learning techniques, these models can be trained on historical patient data to detect early warning signs of cardiac irregularities, providing healthcare professionals with additional insights to guide their diagnosis and treatment plans, meaning saving more lives.
This dissertation focuses on building and evaluating machine learning models for the detection and prediction of cardiac irregularities, aiming to contribute to the development of early-warning diagnostic tools in cardiology. By assessing the performance of different machine learning algorithms, such as Support Vector Machines (SVM), Random Forest, Deep Learning Neural Network and XGBoost (supervised machine learning methods), this research will establish the most effective methods for identifying early indicators of cardiac issues. Moreover, the study will incorporate a comprehensive data preprocessing approach,

addressing data preprocessing steps such as data cleaning, normalization, and feature selection, which are critical for improving model accuracy and reliability.

The unicity of this dissertation relies on the researcher's direct experience. Having cardiac irregularities makes an individual understand what type of mental and physical struggle these pathologies can cause. For this reason, this personal insight enhances the commitment of exploring preventative solutions that not only will advance scientific knowledge but delve deeper into real-world implications. By focusing on early detection and preventative strategies, this dissertation aims to address a significant healthcare challenge, potentially reducing the burden of heart disease on individuals and healthcare systems alike.

The objectives of this dissertation are:

- Ensuring that the datasets provided are reliable. Cleaning, handling missing values and standardizing data formats are challenges that will be tackled, together with class imbalance and the application of SMOTE.
- Identify, create and compare machine learning algorithms: This step will consist in evaluating the performance of multiple Machine Learning models (SVM, XGBoost, Deep Learning and random forest) based on datasets obtained from secondary sources. The focus will be on metrics such as accuracy, precision, recall and F1 score.
- Interpret the results by comparing them with existing literature, evaluating whether performance aligns with known clinical patters represents new horizons.
- Based on the results, the dissertation will delve into proposing directions for better research, improving model performance with technologies such as IoT integration and fill in present gaps.

This research can surely give a contribution to the health and technological world. People know how important synergy is when it comes to reaching enhanced results in such complicated field. In fact, through this study the researcher aims to bridge and understand what the gaps between machine learning and cardiology are, giving its help to enhance the healthcare system, hence people's lives.

2. Literature Review

Machine learning has transformed many aspects of the world we live in. When it comes to healthcare, automation has drastically enhanced the diagnosis and prediction of analytics related to various conditions. As explained in the introduction, when it comes to predicting heart diseases, ML offers approaches to analyze complex and vast datasets related to heart health metrics (ECG's, MRI scans, Cholesterol, Blood Pressure, Blood glucose and others). The goal of this literature review consists in examining existing research on ML application and achievements in cardiology, focusing on what are the key models and the best methods to early detect and predict heart diseases.

The review will be divided into different section so to help the reader have an easier understanding and analysis of this vast and complicated topic. It first delves into an overview of cardiac irregularities, together with the importance of its early detection and current limitations. Secondly, the research will discuss how machine learning is deployed in fields such as healthcare and cardiology underlining its applications and recent breakthroughs. Third step will cover some of the best machine learning models for predictive analysis together with the data and processing challenges in medical research. This step is going to summarize how metrics are used to evaluate models' predictive power. Lastly, after summarizing all the literature gathered, the review will identify research gaps and limitations in existing studies which will be the base of the actual dissertation.

2.1 Overview of cardiac irregularities and the importance of early detection:

2.1.1 Understanding cardiac irregularities

Cardiac irregularities, also known as arrythmias can come in multiple forms and types. The word arrythmia signifies that a heartbeat is out of rhythm, either too slow, too fast or just irregular. The most common way to classify arrythmias is based on the rate of conduction as bradyarrhythmia (<60 bpm) and tachyarrhythmia (>100bpm) (Desai and Said Hajouli, 2023). These two categories are composed by multiple subcategories which represents the complexity and variety or heart diseases. These irregularities vary in severity, from benign conditions to life-threatening complications. As stated by (The Texas Heart Institute®, 2020) the main types of arrythmias are Ventricular arrythmias and Atrial Fibrillation (AF). However, there are many more conditions that can cause problems to the cardiac muscle. Cardiovascular diseases (CVDs), as explained by (World, 2024) are the leading cause of death and disability in the WHO European Region. The main cause of CVDs in Europe is related to high blood pressure (hypertension) which can be related to factors such as tobacco and alcohol use, obesity hence unhealthy diets, and sedentarism. In contrast (Eurostat, 2020) shows that in 2016 deaths reported in Europe related to CVD were 4 527 500, almost 300 thousand more than in 2019. This shows how healthcare is focusing on improving prevention of this issue.

2.1.2 Impact of Cardiac Irregularities and importance of early detection.

Since CVDs are one of the most common causes of death globally, they are considered expensive for general healthcare systems. Overall, cardiovascular disease in Europe cost to the economy a total of 282 billion euros early (2021), related on drug expenditure, out patience care, primary and emergency care. (Ox.ac.uk, 2023). For this reason, the UN assembly together with WHO (World Health Organization) assigned methods which governments should follow to minimize and prevent CVDs worldwide. These methods are: reduce tobacco usage (reduce tobacco affordability by increasing taxes and sensitive packaging); reduce alcohol consumption (restrictions on the physical availability such as

reduced hours of sale); reduce unhealthy diets (reduce common salt consumption and saturated fatty acids) and educate people to be physically active (Khaltaev and Axelrod, 2022). This methods in the last years have been considered functional, slightly reducing the death rate.
According to Wang et al. (2023), the early detection of heart failure (HF) is essential for the tackling of guideline-directed medical therapy (GDMT). This plays a key role in lowering mortality rates. However, restricted access to diagnostic tools such as echocardiography can delay diagnosis often resulting in the condition only being identified, when the disease has already progressed. As explained in research conducted by (Grosser, 2021) Atrial Fibrillation (AF) is an irregular heartbeat which causes the heart to beat too fast than it should. Usually, it is not life-threatening however if not taken care of, could lead to blood clots that could cause strokes, which consequently could lead to death. Early detection can lead to management protocols such as the use of beta-blockers and anti-arrhythmic drugs which could reverse completely this issue. In relation to Malta, CVDs are the cause of 40% of the deaths in 2013. Rates for ischemic heart disease are higher in Malta than the EU average. This mortality rate has been constant through the years, for this reason prevention of CVDs is considered a must. (England, 2015)

2.1.3 Current Limitations in Cardiac Diagnostics

Despite the importance of early detection, traditional diagnostic methods still have limitations. Electrocardiograms (ECGs) can provide crucial information; However, they are not fully reliable. Talking from the researcher personal experience, since they require manual interpretation and are often only effective when abnormalities are actively occurring, early-stage heart disease may go undetected. Holter monitors (wearable 24 or 48 hrs) and echocardiograms can provide more continuous data, however the same problem stated above can still occur.
In research by (Muhammad et al., 2020) traditional invasive methods for diagnosing heart disease typically rely on a patient's medical history, and symptom evaluation conducted by healthcare professionals. Among these, angiography is considered one of the most accurate techniques for identifying cardiovascular diseases (CVDs). However, it comes with significant impacts, such as health risks and requirement for advanced technical expertise. Moreover, conventional diagnostic approaches are prone to human error, and can result in less precise outcomes. CVDs are complicated to diagnose due to their clinical silence until serious complication occur. As stated by (Zeynep Altintas, Fakanya and Tothill, 2014) past diagnosis were either expensive or risky invasive techniques, which not everyone can afford. Biosensors, which are devices that convert chemicals information into quantifiable electrical sign, are largely used to diagnose alteration in a patience blood for preventing heart diseases (troponin or CPK values). However, the collection of these data needs to be constant to save a patient life. These limitations need the action of innovative diagnostics approaches. Machine learning models, which can analyze large volumes of patient data such as ECG, MRI and Biosensors offer promising potential in the healthcare field. ML algorithms could identify early warning signs which would otherwise be missed by traditional methods.

2.2 Machine Learning in Healthcare and Cardiology

2.2.1 Overview of Machine Learning in Healthcare

ML has become an invaluable tool in healthcare, offering the ability to analyze large and complex datasets, discovering patterns that could be hidden to the human eye. With the increasing digitalization of healthcare data, such as electronic health records (EHRs), medical imaging and wearable devices, ML algorithms have changed the approach of medical diagnosis. For instance, An et al. (2023) and Javaid et al., (2022) give an overview on machine learning in healthcare. Javaid et al provide an in-depth examination of how ML is deployed across various healthcare domains. Specifically pointing out how ML tools, when integrated into healthcare operations, reduce data handling errors which are often challenges in large healthcare settings. Similarly, An et al talk about the potential of ML's in healthcare, however the research focuses on how these technologies support mobile health and wearable sensors which are becoming extremely important for remote patient monitoring. By analyzing these data in real-time, it contributes to preventive care and might diagnose problems which if not taken promptly could lead to life threatening results. Machine learning and deep learning are used in a vast variety of medical challenges. For example (Liu et al., 2014) give an overview on early diagnosis of Alzheimer' disease with deep learning. To overcome previous invasive based methods for the prevention of general diseases, different predictive machine learning techniques have been used. Some examples can be, Convolutional Neural networks (CNN), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes (NB), and Decision Tree (DT), etc. (Muhammad et al., 2020). These different models are selected base on the type of datasets provided and their performance. For example, CNNs were used to infer a hierarchical representation of low-field knee MRI scans to automatically segment cartilage and predict the risk of osteoarthritis (Miotto et al., 2017). In a recent research study, Liu, Zhang, and Razavian developed a deep learning algorithm using LSTM networks (reinforcement learning) and CNNs (supervised learning) to predict the onset of diseases, such as heart failure, kidney failure, and stroke (Hafsa Habehh and Gohel, 2021). This research shows how well Neural Networks performs when image-based data is involved.

2.2.2 Machine Learning Applications in Cardiology and recent Breakthroughs

An et al. (2023) discuss how continuously monitoring heart rates with wearable devices, is useful to track arrythmias and other irregularities in real time. This allows patients to understand whether their condition requires urgent clinical help and promptly act on it to avoid any further problems. On the other hand, Javaid et al. (2022) explain the use of ML models in analyzing ECG and MRI data. These types of data, being image-related, are well managed by models such as neural networks, which are optimal for prevention and quick intervention.
In a study by Rajpurkar et al. (2017), a CNN model trained on ECG data achieved a diagnostic performance comparable to that of expert cardiologists. This shows how MLM, in certain areas, could help domain experts in providing reliable automated diagnosis. Pu et

al. (2022) on the other hand deployed a CNN to distinguish hypertrophic cardiomyopathy from healthy heart structures with an accuracy exceeding 90%, suggesting that deep learning can facilitate faster and more accurate diagnosis compared to manual image interpretation. However human and domain expert diagnosis is still needed due to an omnipresent lack in accuracy. One main barrier in adopting ML in clinical stings is the so called “black box”. It is the nature of machine learning models where hidden processing layers compute decisions that are difficult for experts to understand. However, one important breakthrough is the concept of explainable AI (XAI). Sadeghi et al. (2024) highlighted the use of XAI to make predictions more interpretable. Techniques such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) help clinicians understand the factors contributing to a model’s predictions, increasing trust in ML’s diagnostic accuracy. For example, Decision tree-based models are extensively used in machine learning due to their interpretability. These, differently than deep learning provide clear pathways on how a certain decision is made. Another important breakthrough is related to privacy concerns in healthcare and how data is used. Since it is important that ML models are trained as well on decentralized data, federated learning allowed these to train without sharing sensitive information. Aggarwal et al. (2023) discuss the application of federated learning in cardiac research, demonstrating that ML models could be trained on patient data from multiple hospitals without transferring the data itself. This innovation not only preserves patient privacy but also enhances the robustness of predictive models by incorporating diverse datasets.

2.3 Machine Learning Models for Predictive Analysis, Data challenges, Model evaluation and metrics

The four models selected for this study: Logistic Regression, Random Forest, XGBoost, and Neural Networks, were chosen based on their extensive use in medical prediction tasks, especially in cardiology. Logistic Regression serves as a reliable baseline for binary classification with interpretability. Random Forest offers balance between performance metrics on structured health data. XGBoost is known for its speed and high predictive accuracy, especially when it comes to imbalanced clinical datasets. Neural Networks, particularly CNNs, are effective in capturing complex patterns in large and high-dimensional data such as ECG signals and imaging. These models represent a diverse range of machine-learning capabilities, which makes them ideal for comparative analysis in heart disease prediction.

2.3.1 Machine Learning Models for Predictive Analysis in Cardiology

Machine Learning has revolutionized predictive analysis in many fields. One of the highest impacts has been seen in the cardiology. Based on a wide range of data, from clinical records to real-time monitoring Machine Learning models are evaluated based on the precision of their predictions. An et al. (2023) describe machine learning as being divided into two main categories: supervised learning and unsupervised learning. Supervised learning algorithms are trained on labeled datasets, where input and outputs features are

known, allowing the model to learn mappings for future predictions. In contrast, unsupervised learning analyzes data without predefined labels which discovers hidden patterns and underlying structures. Supervised learning techniques are used for both classification and regression tasks, whereas unsupervised methods are primarily applied to clustering and dimensionality reduction problems. Each machine learning model is better suited to specific types of prediction scenarios depending on the nature of the data. In this section we will explain and describe multiple models, and which ones are most frequently used in cardiac predictions.

2.3.2 Decision Trees and Random Forests

Decision trees are simple but powerful models. They are among the most interpretable forms of machine learning, that partition the data into homogenous subsets. In cardiology decision trees are useful due to their interpretability, which show a clear decision path when predicting certain conditions. However, even if it offers strong classification, decision trees are more prone to overfitting and might suffer from high variance. For this reason, ensemble methods such as Random forests (RF) have been chosen for medical prediction. Random forest model are an ensemble of multiple decision trees which improve the prediction and accuracy of a certain outcome. Pal and Parija (2021) demonstrated RF effectiveness in heart disease prediction. Their implementation achieved an accuracy of 86.9%, a precision of 90.6%, recall of 82.7%, and an AUC (Area Under the Curve) of 93.3%, indicating the model's strong diagnostic power. An et al. (2023) highlight that decision tree classification techniques have been widely utilized for the detection and prevention of heart disease. In one such application, Pathak and Valan employed a decision tree model using eight patient-related variables and achieved a prediction accuracy of 88%. On the other hand, Baral et al. (2024) show that in a research to predict heart attacks made by Aanthana Krishnan, which used a decision tree model and a Nave Bayes (NB) model, the decision tree had a precision of 91% while NB achieved only 87%. This has shown that decision trees for handling small data sets were the best algorithm. Garg, Sharma and Khan (2021) compared KNN (K-Nearest Neighbor) with random forest algorithm (RFA) based on Kaggle’s datasets to separate people with CVD from normal ones. KNN had a greater accuracy (86.66%) than RFA (81.96%), suggesting that RF may underperform in smaller datasets with low feature diversity. The above studies show how random forest can offer one of the most reliable balances between metrics, when datasets are well preprocessed and trained.

2.3.3 Support Vector Machines (SVMs):

SVM machines are highly used supervised machine learning models when it comes to classification of data. As explained by IBM (2023) SVM classifies data using an optimal Hyperplane maximizing the margin between the closest points of opposite classes. An optimal hyperplane is a line (or boundary) that separates two groups of data by leaving as much space as possible between them. This happens in a 2D plane when it comes to linear classification tasks. However, SVM can handle also nonlinear classification but kernel functions need to be used when the dimensional space has more than two dimensions. In

heart disease prediction, Radial Basis Function (RBF) kernel is preferred due to its ability to handle non-linear decision boundaries, however careful parameter tuning is critical for maximizing classification accuracy. (Keylabs, 2024) This make SVM suitable for complex detection tasks such as heart diseases prediction. However, the dataset needs to be small, structured and well formatted to obtain good results.

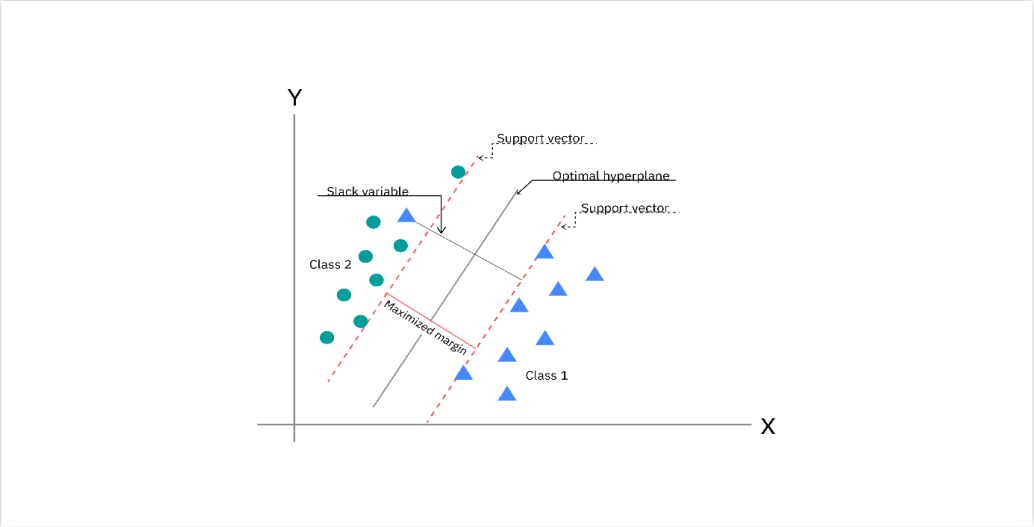


Figure 2.3.3 - Linear classification SVM (IBM, 2023)

Aspect	**Case Study 1**: Heart Disease Prediction Using SVM (Duraisamy et al., 2024)	**Case Study 2**: SVM for Medication Adherence in Heart Failure Patients (Son et al., 2010)
Objective	Predict heart disease using SVM based on patient features collected online.	Predict medication adherence in heart failure (HF) patients using SVM.
Dataset Size & Features	Uses multiple patient attributes (age, gender, BP, cholesterol, symptoms, etc.).	Uses 76 heart failure patients, with features including medication frequency, NYHA functional class, cognitive status, etc.
Modeling Approach	- SVM for classification (binary: heart disease or not). - Compared with KNN, achieving higher accuracy (89% vs. 86%)	- SVM for classification (adherent vs. non-adherent). - Used feature selection to find the best predictive variables.
Evaluation Metrics	Accuracy, precision, recall (SVM: 89% accuracy, outperforming KNN).	Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy (best accuracy: 77.63% using RBF kernel).
Key Findings	- SVM outperformed KNN in heart disease prediction. - Online consultations can provide preliminary heart disease risk assessment	Identified medication knowledge & NYHA functional class as strong predictors. - Found RBF kernel was most effective.

Table 2.3.3 – Comparison of two case studies using SVM (Duraisamy et al., 2024) and (Son et al., 2010)

Duraisamy et al. (2024) use similar dataset features as the ones used in this dissertation, achieving a great accuracy result, emphasizing how SVM tools can be integrated in online medical consultation. On the other hand, Son et al. (2010) used a small dataset to predict heart failure patients, showing how RBF kernel was most effective. However, these study did not address class imbalance, which is a critical limitation in clinical datasets, where negative cases outweigh positive ones. Moreover, SVMs often lack interpretability. In comparison to other models used in this dissertation such as RF, SVMs do not provide feature importance which can limit the usage in medical settings. Hence, their performance highly relies on parameter tuning and kernel choice, which limits their real-time and large-scale adaptability.

2.3.4 Neural Networks and Deep Learning:

Neural networks are a type of machine-learning model which is inspired on how the human brain processes and outputs information. Exactly as the human brain works, these models consist of a number of layers, all interconnected by nodes or neurons that help to learn patterns and make predictions. A basic neural network consists of Input Layer (receives raw data), Hidden Layer (process and transform data using activation functions and adjusting weights) and Output Layer (gives a prediction). This ML model excels to make decisions based on large datasets and complex data format (High dimensional data). Most common learning architectures used in cardiology are Convolutional Neural Networks (CNNs) for image-based diagnosis and Recurrent Neural Networks (RNNs) for ECG waveforms. (Sutanto, 2024). NNs are widely used In cardiology because they help with pattern recognition, early detection and process massive amounts of health records. Murat et al., (2021) explain how Convolutional Neural Networks are used in cardiology for automated analysis and atrial fibrillation (AF) detection. Unlike other machine learning models, CNN eliminate the need for manual feature selection, which improves accuracy and efficiency. Studies reviewed in Murat et al. (2021) show CNN-based models achieving over 98% accuracy, with bidirectional LSTM (long short-term memory) models reaching 99.77% for AF detection. However, these ML models come with challenges and limitations. Large and labeled datasets are needed to obtain good results, but in healthcare it is not easy to obtain such structure. High performance GPUs are required for training, together with the black box dilemma, where it is hard to interpret model decisions. In relation to this issue, explainable AI (XAI) is being developed to fix this.

2.3.5 XGBoost:

XGBoost (Extreme Gradient Boosting) is a machine learning algorithm that uses Gradient boosting decision trees which is optimal for speed and accuracy and interpretation when it comes to predictions using structured data. Gradient Boosting means that the model combines multiple weak decision trees and creates a strong model out of weak ones. A weak decision tree is a simple model that makes only slightly better predictions than random guessing. On its own, it's not very accurate, but when combined with many other weak trees, they form a much stronger and more accurate model. Two research papers were

analysed to understand the importance of this machine learning model in cardiology, and why for certain aspects is overclasses the other models (NNs, SVM, Decision trees). The first research paper “The Prediction and Analysis of Heart Disease Using XGBoost” (Yang, 2024) evaluated this model to predict heart diseases using structured patient data. Yang et al used a csv dataset containing attributes like age, cholesterol, heart rate and blood pressure. Feature selection was performed to identify key factors leading to Heart Diseases. The performance of this model was compared with SVM and Neural networks. The findings show that XGBoost achieved an accuracy of 91%. Influential features present in the datasets used in these studies, such as ST_slope_up, cholesterol and RestingBP are one of the factors improving the accuracy of the models’ predictions. The second paper "XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction" (Moore and Bell, 2022) evaluated XGBoost against logistic regression using the UK Biobank cohort (database with more than half a million participants). 90% of this dataset has been used for training the model and 10% used for testing. To interpret feature importance using explainable AI, SHAP (Shappley Additive Explanations) technique has been applied. XGBoost outperformed logistic regression with a ROC-AUC score of 0.86 vs 0.77, proving better recall for heart attack cases reducing false negatives. Main risk factors identified were related to waist size, blood pressure, cholesterol and sex. Overall, this model is special in cardiology because it can outperform traditional models with better accuracy. This is because it uses feature importance analysis that identifies risk factors which help doctors on relevant biomarkers. Handles well imbalanced data, which recurs in healthcare, where datasets are outnumbered with healthy patients rather than diseased ones. Unlike deep learning, this model provides transparent decision making which makes explainability and interpretability a good tool to adopt in healthcare. Finally, this model is extremely flexible. It works well with both large and small datasets, and does not require much computational power.

2.4 Comparative analysis and research gaps

Numerous machine learning models in various researches have been explored for heart disease prediction. However, most of these studies focused on single datasets, with a lack of consistency in showing preprocessing methods. Subsequently, metrics such as recall and F1 score, were not properly evaluated or implemented. These metrics are crucial for medical applications, since predicting false negative is often more important than the accuracy of a model itself. As explained in the other sections of this literature review, SVMs demonstrate high accuracy in small, structured datasets however their performance decays in larger ones. Decision trees offer good interpretability but are susceptible to overfitting. Random forest, Deep learning and XGBoost models show a strong generalization, mainly on structured datasets. However, the above studies did not evaluate the performance of these models across multiple datasets which is a critical limitation in applying AI models in real world settings. Moreover, the lack of external validation is another issue that can bias model performance evaluation. To address these limitations, this dissertation proposes a multi-model, multi-dataset comparative framework when it comes to predicting heart diseases. Applying four supervised models across four Kaggle datasets of varying complexity, and evaluating performance based on recall, precision, and F1-

score, this study aims to offer a more encompassing and clinically grounded assessment of ML models in cardiology.

3. Methodology

3.1 General overview of methodology

During this study, supervised machine learning approaches are employed to evaluate multiple models for early detection of cardiac irregularities, following a quantitative experimental design using datasets from public sources. Datasets are gathered from Kaggle, an online platform providing the most used datasets for learning and research purposes. Four different datasets are used. Having similar attributes but different sizes will show us which model is more suitable for a certain dataset. The chosen datasets were used in previous researches. Models will be created with RapidMiner, a drag and drop software that enhances machine learning creation. The steps of this methodology are represented in the flowchart below.

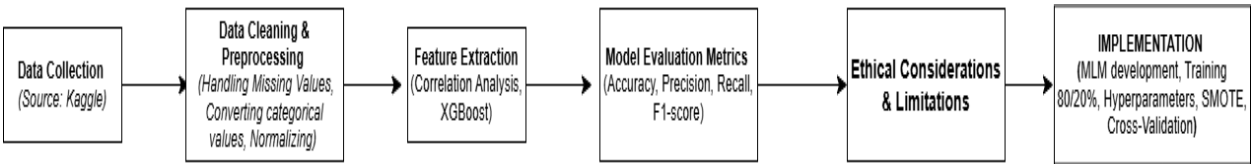


Figure 3.1 Methodology timeline

Hardware used for research:

GPU	NVIDIA GeForce RTX 3060 laptop
CPU	11th Gen Intel(R) Core (TM) i7-11800H @ 2.30GHz
RAM	16 GB
OS	Windows 11 (64-bit OS)

Table 3.1 Hardware used for research

Reasoning behind RapidMiner leverage

During this research, the use of RapidMiner has been the key to achieving good and complete Machine Learning models. RapidMiner is a drag-and-drop, no code data science platform which allows efficient model development. RapidMiner can be used for data collection and preprocessing (Import datasets, Clean missing values, Normalize and Encode), feature selection, model building and training, model evaluation (accuracy, precision, recall, AUC and F1-score), and optimization. Additionally, it offers automated data processing pipelines, hyperparameter tuning capabilities (ensuring optimized

performance) and visualization/reporting tools. Data collection and preprocessing was performed with Python due to better flexibility in comparison to RapidMiner.

3.2 Data collection and explanation of datasets

The different datasets used for this dissertation were sourced form Kaggle. An online platform and repository of datasets containing secondary and anonymous data for research and training purposes. Each dataset collected contains similar records related to heart disease indicators, having a final attribute in common that states the presence or not of cardiovascular disease (0/1). The chosen datasets are 4:

Dataset	Attributes	Classification Label	Usability	Upvote	N. of rows / Instances
Dataset heart Utkarsh Singh	Age, Sex, Chest pain type, resting blood pressure, Serum Cholesterol, Fasting blood sugar, Resting electrocardiographic results, Max heart rate, Exercise-Induced angina, old peak, ST segment, Major vessels, Thal	heart disease 1/0	10/10.	86	271
Heart 2020 Kamil Pytlak	BMI (Body mass index), Smoking, Alcohol Drinking, Stroke, Physical Health, Mental Health, Diff Walking, Sex, Age Category, Race, Diabetic, Physical Activity, Gen Health, Sleep Time, Asthma, Kidney Disease Skin Cancer	Heart Disease 1/0	9.4/10	888	319795
Heart Amirmahdi Aboutalebi Dataset	Age, Sex, Chest Pain Type, Resting BP, Cholesterol, Fasting BS, Resting ECG, Max HR, Exercise Angina, Old Peak, ST Slope.	Heart Disease 1/0	8.8/10	52	919
Heart Disease Oktay Ördekci Dataset	Age, Cholesterol Level, Smoking, Family Heart Disease, Diabetes, BMI, High Blood Pressure, Low HDL Cholesterol, High LDL Cholesterol, Alcohol Consumption, Stress Level, Sleep Hours, Sugar Consumption, Triglyceride Level, Fasting Blood Sugar, CRP Level, Homocysteine Level	Heart Disease Status 1/0	10/10	75	10001

Table 3.2: Datasets overview

The choice was based on 3 factors. Date, Upvote and Usability of the datasets. All four dataset are recent, With historical data of up to 3 years. Upvoting means how many people used and suggested the dataset. As it is possible to see, Pytlak dataset is the one with the most Upvotes, this is due to the vast number of attributes and data samples provided (319

thousand rows and 18 columns). Usability is a score calculated directly by Kaggle that involves Completeness (Subtitle, Tag, Description), Credibility (Provenance, Notebook, Update frequency), and Compatibility (License, File Format, File description, Column Description). Usability is high in all datasets chosen. The choice was made in a way that all datasets are different in values and amount of attributes, but all aspire to the same outcome (cardiovascular disease yes/no). This is to understand how different Machine learning models act in different datasets.

3.3 Cleaning and Preprocessing

The data cleaning and preprocessing phase consists in four different sections. These were tackled with Python programming language (due to its precision and flexibility) and Rapid Miner's software (Simple and efficient).

3.3.1 Handling Missing Values

To identify missing values, python, more specifically the Pandas python library (for manipulation of data) was used. The graphical visualization of these missing values occurred with two more libraries: Seaborn and Matplotlib. When missing values occurred during the analysis of datasets, the researcher used imputation strategies to fill these values. Numerical missing values were filled with Median imputation to reduce the effect of outliers. Categorical feature missing values were filled with mode values (most frequent category). If in a feature (column) in a dataset, a relevant amount of missing values was encountered (e.g. > 40%), the researcher dropped the column to maintain dataset integrity. This decision needed however critical reflection. In healthcare missing values are not always random and might be informative. A missing value could underly clinical judgement which could correlate with disease outcomes. This phenomenon is known as "missing not at random" (MNAR). In section 6.2 it will be explained how future research could explore whether these missing values actually exhibit predictive patterns.

3.3.2 Encoding Categorical variables

Certain machine learning models (e.g. NN and SVM) are highly dependent on numerical inputs for proper functioning. This is because they rely on numerical operations which assume that all features have continuous relationships. For this reason, conversion to categorical values was necessary.
Binary features (e.g. Male/Female, Yes/No) were converted to 0/1 values. Multi-Class features (e.g. overall health: low, medium, high) required one hot encoding (OHE) which creates separate binary columns for each category. For example, if a categorical variable like "Smoking Status" has values [Non-smoker, Occasional Smoker, Heavy Smoker], encoding it as 0, 1, 2 (Ordinal Encoding) might mislead these models into assuming a numerical relationship where Heavy Smoker (2) is twice as "intense" as Occasional Smoker

(1), which is not always true. One-Hot Encoding (OHE) solves this by creating separate binary columns:

Non-smoker → [1, 0, 0]
Occasional Smoker → [0, 1, 0]
Heavy Smoker → [0, 0, 1]

3.3.3 Normalizing and scaling

Normalization and scaling are compulsory steps when it comes to machine learning. Especially when numerical features in datasets have different ranges. It prevents large values from dominating on smaller values, so that a fair weightage is ensured across all features. Scaling consists in adjusting data so that it fits within a certain range. Normalization rescales all values between 0 and 1. Without scaling, machine learning models assign more weight to high-value features. Mix-Max scaling technique was used since the used data does not follow a normal distribution.

$$Xnormalized = \frac{(x - min(x))}{(max(x) - min(x)}$$

Figure 3.3.3: Normalization formula

3.4 Feature selection

Feature selection was conducted through rapid miner operators so to understand the most relevant features in each dataset, enhancing interpretability for each model. Two methods were used and compared. Correlation analysis and XGBoost. They both utilize different methodologies, meaning that the results differed.

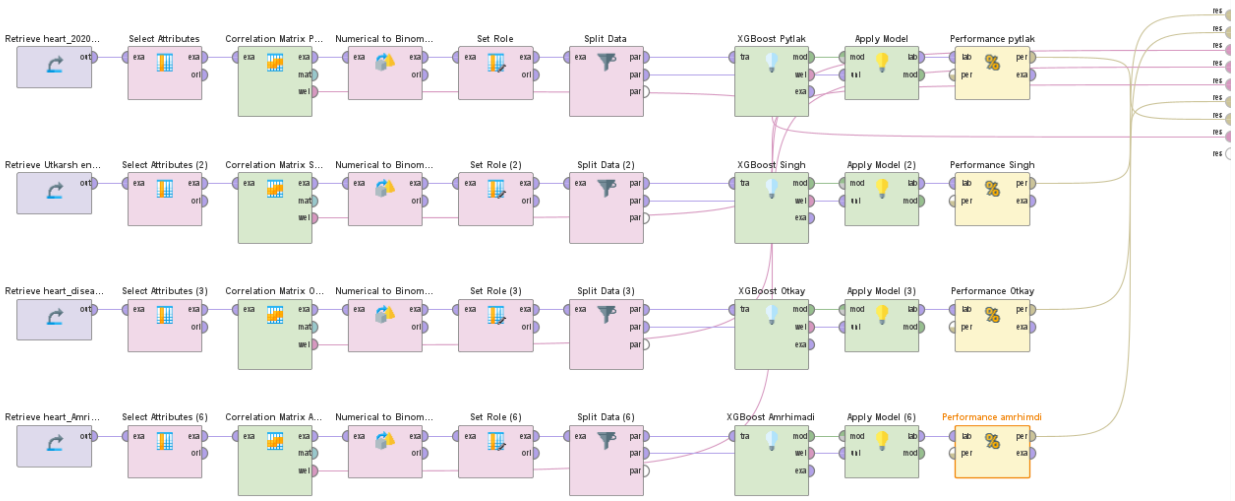


Figure 3.4: Model used to perform Correlation analysis and XGBoost for feature extraction.

3.4.1 Correlation analysis

Correlation analysis is used to establish a relationship or an association between two quantitative variables (Gogtay and Thatte, 2017). This method was used on each dataset. A correlation matrix displays correlation coefficients between variables, which indicates the strength of their linear relationships. Features with high correlation to the prediction label (>0.85) were considered strong predictors and were used for the model development (e.g. Smoking, Alcohol Drinking, Stroke, Sex, Asthma, Kidney Disease). Features with moderate correlation (0.5-0.85) have also been used. Features with little correlation (< 0.5) have been removed so to avoid outliers. When two independent features were found to be strongly related (e.g. Serum Cholesterol and LDL cholesterol both > 0.85) one is likely to be redundant. Keeping both does not add new information and can introduce multicollinearity that can negatively affect performance.

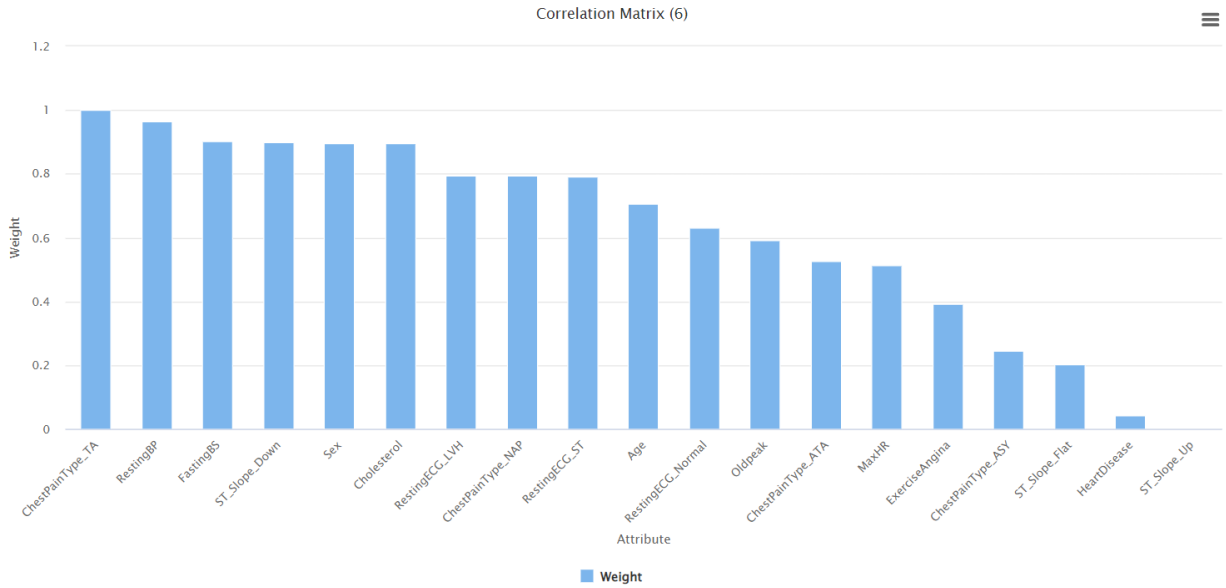


Figure 3.4.1: Correlation Matrix feature extraction based on one dataset

3.4.2 XGBoost method

XGBoost is a powerful gradient-boosting algorithm, used extensively for classification tasks. XGBoost evaluates feature importance through Weight and Gain. Weight counts the number of times a feature is split throughout all trees in the model. The higher the number the more important the feature. Gain shows how a feature enhances accuracy or not. During the training process, this MLM performs automatic feature selection, which allows the researcher to be guided through feature elimination. The researcher eliminated features that had low represented weights. Overall, this model was used to provide feature importance

scores based on how useful each feature is in reducing the model's prediction error. This further refines the choice of relevant features by considering non-linear relationships between features.

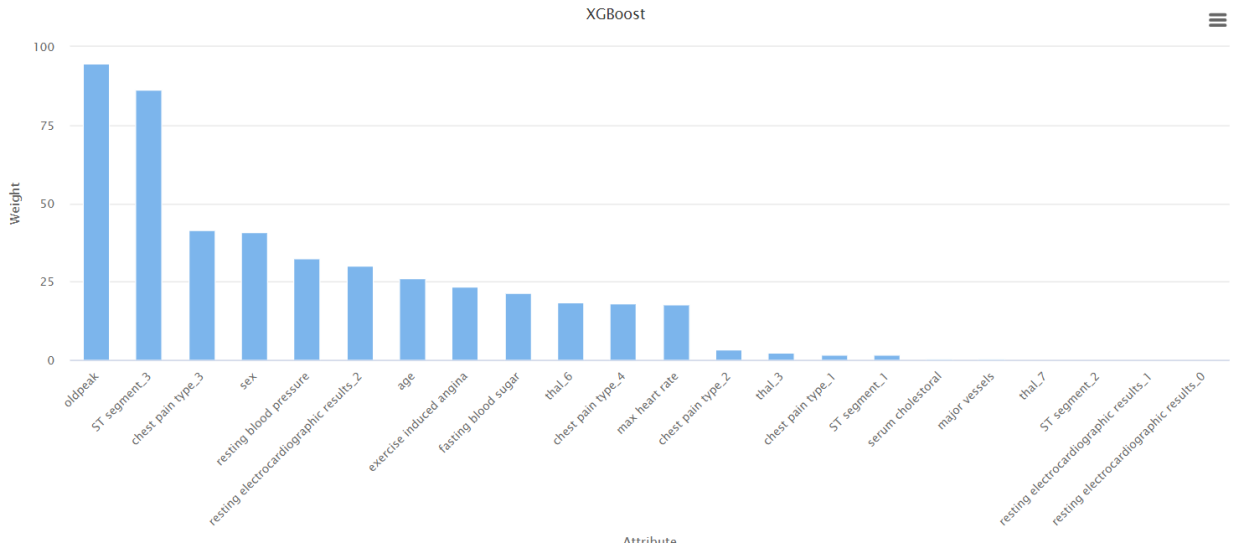


Figure 3.4.2: XGBoost feature weight analysis

3.5 Models evaluation metrics

To evaluate the model's performance, especially in binary classification, four metrics were chosen. Accuracy, Precision, Recall, and F1-score. Given the medical importance of correctly identifying individuals at risk of cardiovascular conditions, Recall was treated as the primary evaluation metric. Recall was prioritized because in medical diagnosis, especially for cardiovascular conditions, it is crucial to identify as many true positive cases as possible to avoid missing individuals who may be at serious risk. The F1 Score was used to balance this with the model's ability to avoid excessive false positives. While Accuracy and Precision were reported for completeness, they were not prioritized due to their susceptibility to skewed interpretation in imbalanced datasets typical of medical diagnosis problems. To understand and calculate the metrics a confusion matrix is used. It is composed by, TP (True Positives), TN (True Negatives), FP (False Positives), FN (False Negatives).

	Actually Positive	Actually Negative
Predicted No	FN	TN
Predicted Yes	TP	FP

Table 3.5: Confusion Matrix

3.5.1 Accuracy

Accuracy is the metric that measures the proportion of correct predictions (true positive and true negatives) out of all predictions made. However, in imbalanced datasets, accuracy is not sufficient since it does not differentiate between false positives and false negatives.

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

Figure 3.5.1: Formula to obtain accuracy

3.5.2 Precision (Positive predicted value)

Precision is used due to its ability to indicate the proportion of positive predictions that are correct. If precision is high, it means that the MLM predicting a positive case, is likely correct.

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

Figure 3.5.2: Formula to obtain precision

3.5.3 Recall (Sensitivity)

This metric measures the proportion of actual positive that are correctly identified by the model. High recall is vital in medical diagnosis. If a positive case is missed (False negative), drastic consequences might occur. For this reason, in this research is important that recall is higher than precision.

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

Figure 3.5.3: Formula to obtain recall

3.5.4 F1-Score

F1-score is the mean balance between precision and recall. It is useful when models deal with imbalanced datasets. It is required in the current research due to the nature of heart

disease datasets. There are usually more non-disease cases than disease cases. If datasets contain 90% healthy patients and 10% with heart disease, accuracy will still achieve 90% when predicting “no disease”, even if it fails to identify sick patients. F1 score prevents this by balancing and avoiding false positives and avoiding false negatives.

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

Figure 3.5.4: Formula to obtain F1-score

3.5.5 Ethical Considerations

The researcher ensured that the current research adheres to ethical standards related to machine learning models and the usage of secondary data. The study involves the processing of heart disease datasets sourced from public repositories such as Kaggle. The above-mentioned ethical standards revolve around data privacy, model fairness, and their explainability. All datasets used are available to the public and fully anonymized. No PII (personally identifiable information) was used or accessed throughout the study, ensuring compliance with GDPR. Subsequently, machine learning models can inadvertently introduce bias, particularly when trained on datasets with imbalanced demographic or clinical characteristics. To mitigate this, techniques such as SMOTE were applied. SMOTE (Synthetic Minority Over-sampling Technique) is a method that creates synthetic examples of the minority class to balance the dataset and improve model performance on underrepresented outcomes. Performance metrics were stratified to examine fairness across subgroups. This model is intended to support, not replace, clinical judgment. Any predictive output should be reviewed and validated by qualified healthcare professionals and machine learning domain experts to ensure the models quality.

3.5.6 Limitations of the Study

This research appeared to have certain limitations that need to be acknowledged. The datasets used are heavily imbalanced, with much more non-disease ("No") cases than disease ("Yes") cases. While SMOTE was used to correct this imbalance, synthetic oversampling may introduce noise or lead to overfitting in some models. Moreover, across all models tested, Recall consistently remained lower than Precision. While high Precision indicates that the model avoids false positives, low Recall is dangerous in the context of heart disease prediction, as it means actual heart disease cases may go undetected (false negatives), posing clinical risks. On the other hand, the chosen datasets may not fully reflect the diversity and complexity of real-world hospital data. As a result, the findings may not generalize well to clinical settings without further validation on institutional datasets. The models were not tested in live clinical environments. Additional validation on external datasets from medical institutions is necessary to confirm real-world performance and usability. Subsequently, deep learning models and SVM machines required significant

computational resources. Some experiments were constrained by hardware limitations, which may have impacted model tuning and scalability.

4. Implementation

4.1 Machine learning models

Machine learning models algorithms used for the scope of this research include Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM) and Deep Learning Neural Networks. A description of the above machine learning models can be found in Section 2.3 of the current research. As already stated, the above-mentioned ML models were created, trained and tested with Rapid Miner Studio. Four Kaggle datasets have been used for training. Section 3.2 shows the attributes of the datasets, the predicting variables used and the number of participants sampled. The model's hypermeters were tested, choosing the ones that provided the best results for each MLM. Predictive features have been reduced in result of correlation and XGBoost analysis. Again, the ones who provided better results have been chosen. Below a comprehensive explanation on how the MLM were created has been conducted.

4.1.1 Handling Missing Values

The process of cleansing the datasets involved the usage of Python programming language. Before cleaning, each dataset was inspected, identifying missing values and detecting categorical and numerical values or general inconsistencies. Tools used in this process are Pandas library and Matplotlib/Seaborn to visualize missing values on a graph. Four scripts were written, one for each dataset. A separate script was created for each dataset to address its unique structure and preprocessing requirements, ensuring accurate data preparation. Only one dataset appeared to have missing values. The other three dataset's N/A values were filled by the owner before publishing. Below is the visualization of the dataset's missing values.

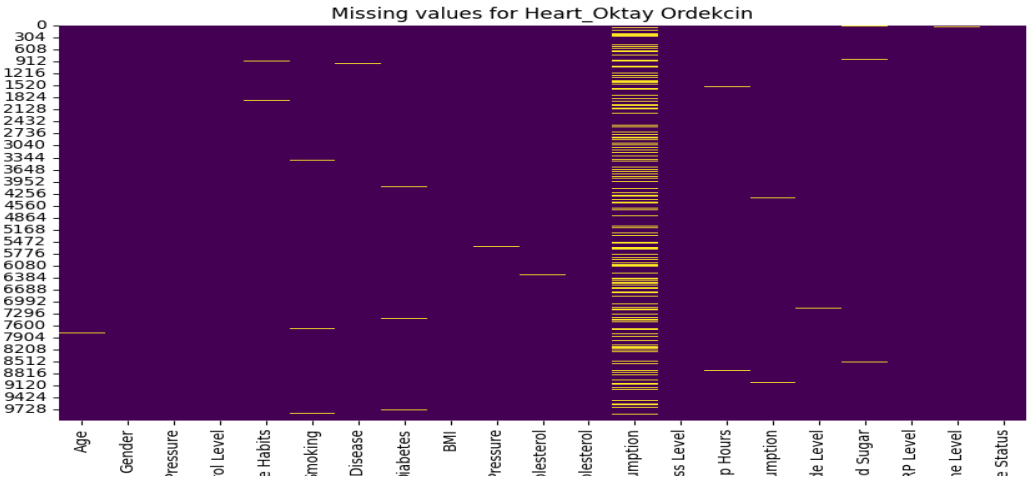


Figure 4.1.1 Handling missing values

To make sure N/A values would not affect the quality of the dataset "Heart Disease (Oktay Ördekçi, 2025)", some techniques have been adopted. The code "print(df.isnull().sum()) showed a large number of missing values for the column "Alcohol Consumption". This column was dropped with "df.drop(columns=("Alcohol Consumption"), inplace=True)" to not affect the Machine Learning Models processing. To handle the rest of the missing data, Numerical missing values were filled using a median approach due to less dependency on outliers, ensuring a more robust dataset. Categorical features were filled with a mode approach, representing the most frequently occurring category, being the most statistically probable value to replace missing data. (Appendix) This code ensured no more N/A values on the dataset.

4.1.2 Converting Categorical Values

Machine learning models struggle to perform when categorical values are present, hence they perform better with numerical values. Next step of cleaning consisted in converting categorical variables to numerical ones. There are two types of categorical features, Binary and Multi-Class.
Binary features consist in features that can have either 1 or 0 as an answer. For example, attributes that would have original values such as Yes or No and Male or Female were converted in binary code (0/1) to make ML reading possible. Multi-class features consist in attributes that have more than two values. For example, "Exercise Habits" could be low, medium or high. Conversion consists in scaling it in numerical values (0,1,2). To better identify categorical columns in each dataset the below code was used. This code identifies all columns in the dataset that contain categorical (non-numeric) data and prints their names.

"categorical_cols = df.select_dtypes(include=["object"]).columns
print("categorial columns", categorical_cols)"

Once categorical values were suggested, for each dataset the researcher split these values between binary and multi-class. Performed manually, only three out of four datasets needed converting categorical values into binary ones. Heart Disease Prediction Dataset (Singh, 2024) was already in numerical values.
On the other hand, multi-class categorical features were handled with what is called One-Hot encoding (OHE). It is a technique used to convert multi-class categorical variables into numerical format, creating a separate binary column for each category in a feature. OHE prevents numerical misinterpretation. Some raw features in categorical multiclass format, such as GenHealth (Very good, good, fair, bad) could not be processed by the MLMs. The researcher used the python code "*df = pd.get_dummies(df, columns=["AgeCategory", "Race", "GenHealth"], drop_first=False"* to apply OHE, also called dummy encoding. Get dummies signifies that the code creates new columns, one for each value of a category, creating Boolean values. (True/False). To convert Boolean to binary the code "*df=df.astype(int)*" was used. The above process was used for all four datasets, where multi-class categorical values were encountered.

4.1.3 Normalization and Scaling

In each of the four chosen datasets, normalization was performed. Normalization and its importance have been discussed in section 3.3.3 of the research paper. Represented below is a non-normalized dataset.

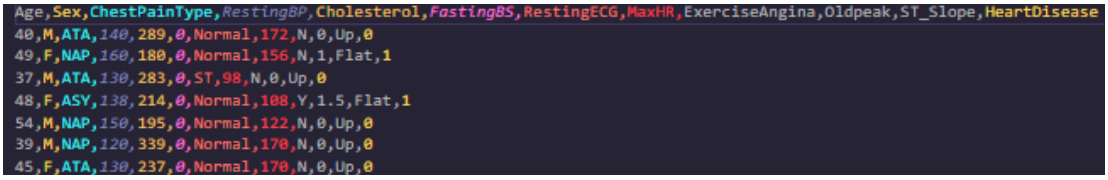


Figure 4.1.3: non-normalized dataset

As it is possible to see, the range for "Age" feature is much smaller than "RestingBP", "Cholesterol" and "MaxHR". This is the case where normalization technique should be applied. As stated previously, normalization ensures that a fair weightage is balanced across all features. If numerical values highly differ from each other, the machine learning model will tend to give more importance to those which have higher ranges. Represented below is the above dataset after normalization.

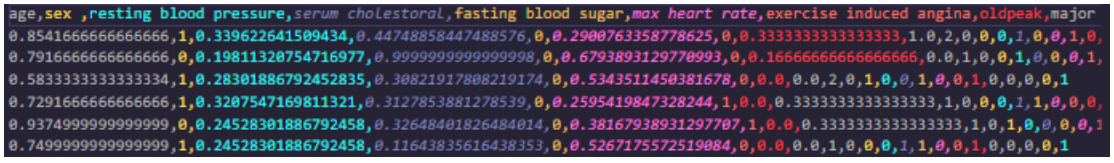


Figure 4.1.3: Normalized dataset

All features are placed in a range between 0-1 which allows non-biased predictions. Normalization was applied in Python using Scikit-learn library. The process consisted in choosing the columns (attributes) that needed normalization, and applied Min-Max scaling method. Finally, the datasets are saved as cleaned and encoded dataset and can be now used

for training and testing the respective MLM. The codes used to normalize each dataset can be found in the appendix.

4.1.4 Feature extraction results

Feature selection was conducted using correlation analysis and XGBoost based-feature importance. Correlation analysis helped assessing the final selection due to its stronger alignment with medical relevant indicators. Features with correlation scores below 0.5 were removed from the datasets, except in cases where their removal negatively affected performance.

In three datasets (Amirmahdi, Otkay, and Pytlak) removing low-correlation features led to a significant drop in model performance across all algorithms. This suggested that these features, although weakly correlated individually, highly contribute to predictive accuracy. As a result, all original features were retained for these datasets.

In contrast, the Singh dataset benefited from feature removal. Features such as chest pain type 4, maximum heart rate, old peak, and certain electrocardiographic indicators were eliminated based on low correlation scores. This improved accuracy for SVM (from 77% to 81%) and Random Forest (from 71% to 77%). On the other hand, however, this pruning negatively affected the performance of XGBoost (dropping from 81% to 67%) and Neural Networks (from 88% to 79%), suggesting that these models were more dependent on complex, non-linear interactions among features.

For clarity, a summary table has been included to show how feature selection decisions were applied across each dataset and their impact on model performance.

Dataset	*Features Removed*	*Feature Selection Used*	*Models Improved*	*Accuracy Impact*
Amirmahdi	No	Correlation	None	Accuracy dropped if features removed
Otkay	No	Correlation	None	Accuracy dropped if features removed
Pytlak	No	Correlation	None	Accuracy dropped if features removed
Singh	Yes	Correlation	SVM, RF	Accuracy ↑ (SVM: +4%, RF: +6%) Accuracy ↓ (XGBoost: -14%, NN: -9%)

Table 4.1.4: Feature extraction outcome

4.2 Machine learning models development in Rapid Miner

4.2.1 Model Development Components

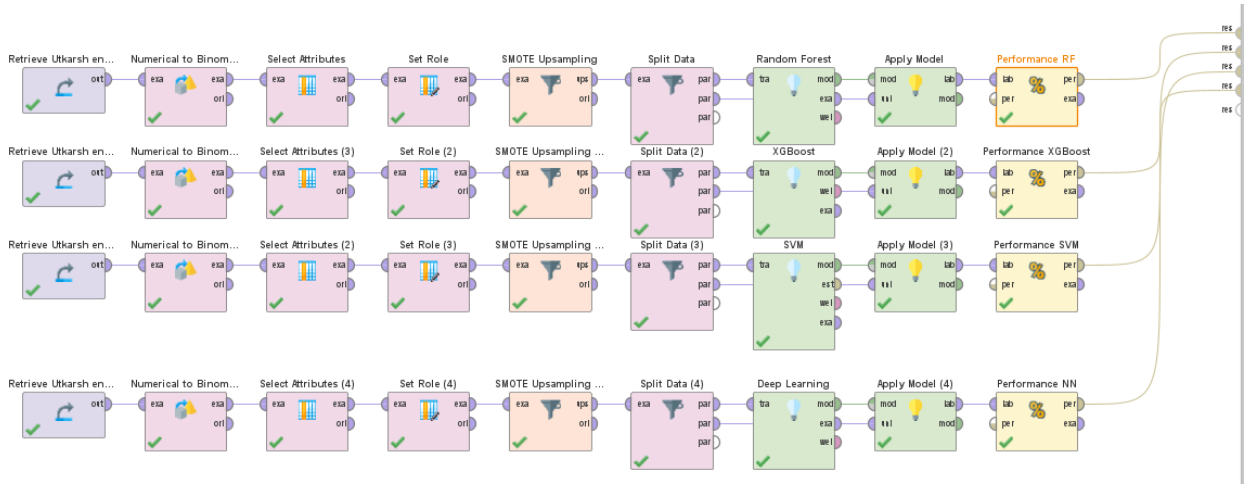


Figure 4.2.1: Process of 4 machine learning models on one dataset.

In supervised learning, the prediction label is the target variable that the model aims to classify or predict—in this case, whether a patient has heart disease (1) or not (0). For each dataset, 4 machine learning models were built. Even though the cleaned datasets had already "Heart Disease" labels as 0/1, Rapid Miner still treated it as continuous numerical variable. The use of "numerical to binomial" operator allowed all models to recognize the prediction label. This is because SVM use mathematical optimization (Hyperplanes) while NN expect a classification task. Rapid miner in this case assumes that the researcher is performing regression and not classification. For this reason, it is important to explicit that the label is set as a categorical label. The "select attribute" operator is used in each model to avoid categories that might influence negatively the training and testing (feature extraction). Set role operator is used to set the prediction label that the ML models use. Heart Disease (1/0) category was chosen as a label. Subsequently Split data Operator (drag and drop operator provided by rapid miner platform) is used to split the dataset in 80% for training and for 20% testing, using a shuffled sampling method for less biased results. Each Machine Learning model is constructed with different parameters. We will discuss this and SMOTE operator in section 4.2.2. The Split Data operator divides the dataset into two parts: 80% for training the model and 20% for testing it. The training portion is sent to the model, while the test portion is connected to the "Apply Model" operator, which uses the trained model to make predictions. Finally, the "Performance" operator evaluates these predictions and provides metrics such as accuracy, precision, recall, and F1-score.

4.2.2 Model parameters

Each hyperparameter used in the MLM was set as default by RapidMiner platform. Only a few changes were made by the researcher. Below the parameters used are shown.

Random Forest	Number of trees: 100
	Criterion: Gain Ratio
	Maximal depth: 10
	Pruning and prepruning – not applied
	Guess subset ratio
	Voting strategy: confidence vote
	Parallel execution enabled
XGBoost	Booster: Tree booster
	Rounds: 25
	Early stopping: none
	Learning rate 0.3
	Min split loss: 0.0
	Max depth: 6
	Min child weight: 1.0
	Subsample: 1.0
	Tree method: auto
	Lambda: 1.0
	Alpha: 0.0
Support Vector Machine	Kernel type: Dot
	Kernel cache: 200
	C: 0.0
	Convergence epsilon: 0.001
	Max iterations:100000
	Scale: Applied
Deep Learning Neural Network	Activation: Rectifier
	Hidden layer sized: 2 layers - 50 / 50
	Epochs: 10.0
	Train samples per iteration: -2
	Adaptive rate: applied
	Epsilon: 1.0E-8
	Rho: 0.99
	Standardize: applied
	L1: 1.0E-5
	L2: 0.0
	Max w2: 10.0
	Loss function: Automatic
	Distribution function: AUTO
	Missing values handling: MeanImputation
	Early stopping: not-applied

Table 4.2.2: model parameters

4.2.3 Synthetic Minority Over-sampling Technique (SMOTE)

As explained in section 3.5.6, one of the main limitations that were encountered in this study was class imbalance in datasets. When it comes to heart disease prediction datasets, "No Heart Disease" cases often severely outnumber "Present heart disease" cases. To address this, SMOTE methodology was employed as a preprocessing step before model training. It generates synthetic samples for the minority class (heart disease present) helping to balance the dataset. This allows the model to reduce its classification focus on the majority class. Different upsampling ratios were tested depending on the model type. 40% for SVM, 50% for Random Forest and XGBoost and Neural networks. Model performance was evaluated with and without SMOTE to assess its impact on recall, precision and F1 score. Recall improved significantly across models post-SMOTE which is critical in the medical field when detecting more heart diseases. A slight drop in precision was noted, however, this led to a more balanced model where minimizing false negatives (improving recall) is more important than reducing false positives (improving precision). While this approach enhanced the representation of the minority class and contributed to improved recall scores, it is important to acknowledge that applying SMOTE before data splitting may lead to data leakage. Specifically, synthetic samples generated from the entire dataset may reflect patterns present in the future test set, potentially inflating performance metrics. Although this design choice was made for workflow simplicity, it introduces a methodological limitation that is discussed further in the section 6.2 (Recommendations for Future Work).

4.2.4 Validation approach and Justification

Due to datasets size and variability, the researcher chose to use a stratified 80/20 train-test split for validation rather than k-fold cross-validation. An 80/20 train-test split was selected to streamline the evaluation process, balancing computational efficiency with sufficient data for both training and testing. Cross-validation, even if recommended for enhancing generalization, it was not possible to implement due to computational limitations given by datasets such as Pytlak's and Oktay's and the use of models such as Deep Learning and XGBoost. Since the research results are based on the evaluation on four machine learning models and their behaviour in different datasets, applying cross-validation and external validation was considered not feasible. This approach, while practical is acknowledged as a methodological limitation that will be discussed in section 5.4.

5. Results and Discussion

This section presents the results provided by the machine learning models across four different-sized heart datasets: Singh dataset (271 rows), Amirmahdi dataset (912 rows), Oktay dataset (10001 rows), and Pytlak dataset (319715 rows). The performance of these models is evaluated on how well these predict the presence of heart disease in patients. Subsequently, these results aim to compare and evaluate clinical feasibility for cardiac risk

detection. Recall results were prioritized due to the metric importance in healthcare scenarios. Random forest, Support vector machine, deep learning and XGBoost MLMs have been leveraged and each model was assessed based on the following key performance metrics: Accuracy, Precision, Recall and F1 Score. The relevance of the said performance metrics can be recalled in section 3.5

5.1 Results

Model	Accuracy	Precision	Recall	F1- Score
XGBoost	92.71%	83.64%	90.20%	86.79%
Random Forest	88.80%	80.21%	76.24%	78.17%
Deep Learning	88.80%	83.91%	71.57%	77.25%
SVM	83.59%	71.91%	62.75%	67.01%

Table 5.1: Model performance using Amirmahdi dataset

In Amirmahdi dataset, which is the second smallest out of the four, XGBoost model performed the best, achieving an F1 score of 86.79% and a Recall of 90.20. This indicates an excellent sensitivity to true positive heart disease cases. This feature is important in a medical context, where failing to detect a heart condition (false negative) could lead to drastic consequences. Deep learning also demonstrated strong performance, having an F1 score of 77.25% and an accuracy of 88.80%. SVM, on the other hand, also if provided a decent precision (71.91%), is the model with the lowest recall. This shows how a noisy and imbalanced dataset, might bias SVM ability to recognize true positives also if SMOTE is applied. Finally, Random Forest achieved balanced results but underperformed in comparison to XGBoost and DL.

Model	Accuracy	Precision	Recall	F1- Score
XGBoost	83.66%	89.80%	75.94%	82.29%
Random Forest	67.61%	62.08%	81.12%	70.33%
Deep Learning	61.61%	53.90%	72.00%	61.65%
SVM	50.66%	50.50%	66.38%	57.36%

Table 5.2: Model performance using Otkay dataset

The Otkay dataset, which is the second largest of the four, introduced challenges such as feature noise, class imbalance, and redundant variables, which persisted even after applying SMOTE and affected overall model performance. XGBoost still maintained a consistent and highest performance out of all models, however random forest achieved the highest recall (81.12%). Despite SMOTE implementation, SVM once again underperformed, producing the lowest accuracy and F1 score. This is because, Otkay's dataset was too large for a model that usually performs well only in small and well-structured datasets. However, deep learning model showed reduced performance compared to its behaviour in other datasets. These results show that class imbalance is not the only factor causing bias in the performance. Factors such as feature noise and redundancy may have contributed to lower effectiveness.

Model	Accuracy	Precision	Recall	F1- Score
XGBoost	79.66%	82.14%	76.67%	79.31%
Random Forest	77.97%	74.19%,	82.14%	77.97%
Deep Learning	84.75%	78.38%	96.67%	86.57%
SVM	81.36%	75.68%	93.33%	83.57%

Table 5.3: Model performance using Singh dataset

The above dataset is the smallest in samples. A consistently high performance was seen across all modes due to its smaller size and a better-balanced class distribution. Deep learning is the model that performed better, which is really uncommon in small datasets. With a recall of 96.67% and a F1 Score of 86.57% it is the most reliable model to detect actual disease cases. Moreover, SVM's performance improved substantially in comparison to all other datasets. This underlines how a smaller dataset is required for the model to be able to process and the limitation it has on large and imbalanced datasets.

Model	Accuracy	Precision	Recall	F1- Score
XGBoost	74.82%	65.94%	43.26%	52.24%
Random Forest	74.17%	64.29%	41.71%	50.60%
Deep Learning	74.54%	57.15%	75.45%	65.04%
SVM	76.82%	69.04%	48.53%	57.00%

Table 5.4: Model performance using Pytlak dataset

The Pytlak dataset, the largest used in this study posed the greatest challenges. As it is possible to state from the results, all models underperformed when it comes to Precision, Recall and F1 score. However deep leaning remained the most effective model, achieving recall of 75.45% and F1 score of 65.04%. The more parameters and samples a dataset has, the better a neural network will train and perform. The low performance of the other models is attributable to the complexity and high dimensionality of the dataset. The machine learning training and development required high processing. The models did not benefit as much from SMOTE up sampling due to the large variety and size of patient profiles.

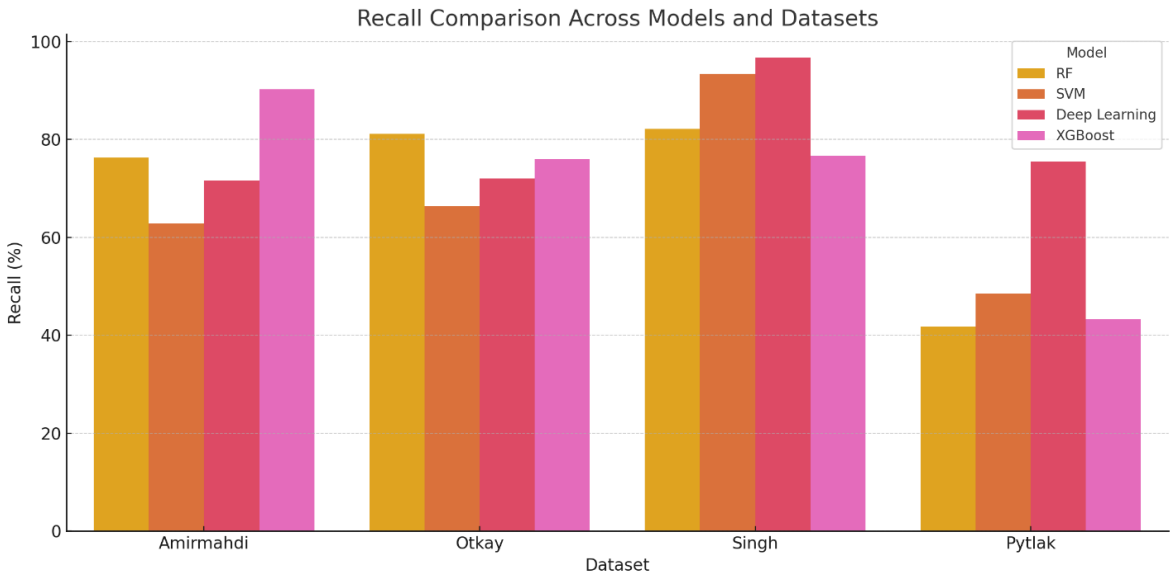


Figure 5.1 Recall comparison across models and datasets

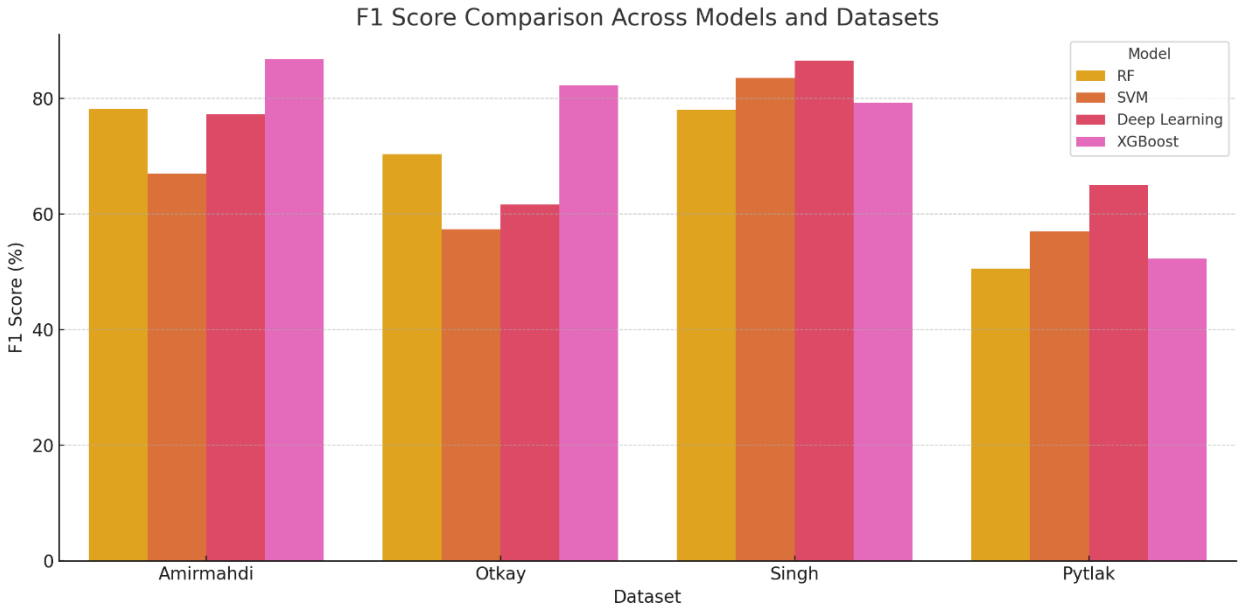


Figure 5.1.2 F1 score comparison across models and datasets

5.2 Best performing model for heart disease prediction.

Among the four machine learning models developed, XGBoost and Deep learning have appeared to be the top performers. Their performance using different data condition shows how model choice should be based on size, quality and complexity of dataset, together with clinical requirements.

5.2.1 XGBoost

XGBoost, as expected when compared to other researches, demonstrated a constant and meticulous performance, mainly in the prediction of heart disease in Amirmahdi dataset. Its achievement of an F1 score of 86.79% and Recall of 90.20% in the above-mentioned dataset shows that the model is inclined to perform well in discrete-sized and well-structured datasets. Its performance remained stable on larger datasets like Otkay where F1 scored 82.29%, however the performance dropped drastically when applied to highly complex datasets (Pytlak) (F1 52.24% ; recall 43.26%). The above performance trend mirrors the XGBoost algorithms strengths. Building decision trees, focusing on correcting errors and fixing weights between iterations, makes this model highly effective in structured data. In addition, XGBoost handles missing values and automatically selects feature, reducing overfitting and making it easily interpretable, which is an essential trait for real-world deployment.

5.2.2 Deep Learning

On the other hand, deep learning model showed good performance in high recall scenarios and unexpected contexts. As expected, in a high multidimensional dataset (Pytlak) the model performed better than other models with a recall of 75.45% and an F1 score of 65.04%. This is because NNs thrive on large volumes of data where linear or non-linear relationships are present. In relation to Singh dataset, the model achieved an F1 score of 86.57% and a Recall of 96.67% which represent the highest results out of all the experiment. Making it well suited for early screening where a missing positive case can be riskier than generating false positives. However, DL performed almost perfectly on the Singh dataset, despite its small size. This is counterintuitive due to the fact that neural networks typically require large datasets to be effective. This anomaly can be explained by the structure of Singh dataset. Well-labeled, clean and strong signal features allow even neural networks to extract patterns efficiently.

5.3 Models performance comparison with existing studies

The following section demonstrates the relevance of our study by comparing the performance of the obtained results with other researches in the same subject matter. The compared researches have been chosen on the datasets used and goal similarities. However, an exact comparison resulted complex and not feasible due to the difference in datasets, data handling, selected predictive features and model development of this research and others found. Moreover, many researches are based on training MLMs on one dataset, while the current research uses four different datasets to understand ML performance in different settings. Only two researches were found using Pytlak and Singh dataset.

Two major studies were considered for comparison:

1- **(Soni, Gupta and Uppal, 2024) – "Optimizing Heart Disease Prediction with Random Forest: Insights from the Kaggle Dataset"**

In this research a Kaggle dataset is used. Very similar to Amirmahdi dataset, it contains 920 entries and 16 attributes, where features correspond to what used in our research. (cholesterol, blood pressure, thalassemia, etc.). Models used are Random Forest, Support Vector Machine, AdaBoost, K-Nearest Neighbors. 70/30 train test split without detailed discussion of normalization or class imbalance handling was recognized by the researcher. The best result showed Random Forest as the best performing algorithm (Accuracy: 67.89%). No precision, recall or F1 score were reported, which limits the depth of the performance evaluation. SVM performance was the lowest with an accuracy of 58.69%. In comparison, the SVM and Random Forest performance of the current research really outperformed Soni's. SVM model achieved an accuracy of 83.58% while random forest an accuracy of 88.80%. The XGBoost model, which was not evaluated in their study, delivered the highest performance with 92.71% accuracy and an F1-score of 86.79%.

2- **(Wang, 2024) – "Predicting heart disease risk using machine learning: A comparative study of multiple algorithms"**

Both dissertations use Kamil Pytlak's 2020 Kaggle dataset comprising 319,795 rows. However, they differ in the processing. While the paper applies under-sampling to balance the dataset, removing 90% of negative cases (decreasing processing power needed) our dissertation uses up-sampling. Up-sampling preserves all positive cases and avoids discarding potentially valuable data from the majority class, which is crucial in healthcare settings. Whereas under-sampling risks losing information by removing a large portion of the majority class. On one hand wang tested traditional models (Logistic regression, Random Forest, Boosted Trees, KNN). Logistic regression reported the best performance (Accuracy: 74%). On the other hand, this study used XGBoost and Deep learning models who had similar results (Accuracy around 74.82%). However, recall and F1 score were prioritized being critical metrics in medical diagnosis. Subsequently, in this research all features were retained after feature extraction with correlation analysis and XGBoost. Wang on the other hand reduced the features to eight using RF importance. Wang's simpler models provided competitive accuracy but lacked of deeper evaluation. Future work could benefit from benchmarking against logistic regression and incorporate AUC in evaluations.

5.4 Implementation feasibility and implications

Overall, this research was performed to test how machine learning models act across datasets that vary in quality and scale. This decision was taken to gather introspections to the real-world clinical environment, where data are noisy and imbalanced. The research results offer valid insights into the behaviour of each machine learning model and their adaptability for heart disease detection.

5.4.1 Implications

Durning this research critical trends were revealed. XGBoost appeared to be the most reliable model throughout all datasets. Its robustness and the way it handles features make it suitable for imperfect datasets, which often recurs in clinical settings. The most surprising discovery is seeing a deep learning model performing really well in a small dataset. This contradicts the nature of Neural network models that often require large training sets to well-perform. However, it shows that when features and data are well-processed, the latter model can have good results even in small datasets. Subsequently, during this research no optimal customization of machine learning model towards each dataset was performed. No extreme hyperparameter tuning allowed the study to achieve better generalization, enhancing the possibility of working well on new unseen datasets. On the other hand, by generalizing, overfitting was avoided, making the model more applicable in different real-world settings. The use of RapidMiner allowed fast model development and ensured accessibility for healthcare stakeholders with less coding experience, which would have not been possible with Python scripting language. Moreover, the results underline the importance of handling class imbalance. SMOTE was useful in enhancing recall and F1 score, that are critical measures in medical diagnosis.

5.4.2 Real-world settings

The results gathered show that the best performing models (XGBoost and Deep Learning) could be integrated into real world settings such as decision support systems within the EHR (Electronic health records) platforms. The fast prediction and relatively high precision and recall allows to be implemented in real time. Recall-oriented models are useful when it comes to initial patient assessing, ensuring that at risk individuals are not missed. Precision oriented models are optimal to reduce unnecessary diagnostics. Integrating these ML models with IoT and wearable technologies would create the possibility of a continuous cardiovascular monitoring.

5.4.3 Challenges

The research has achieved strong performance indicators, however when comparing the current study with similar past ones, some limitations were considered and acknowledged:

- Absence of domain expertise (machine learning and cardiologist) limited the clinical validation of the developed models.
- Rapid miner as a platform is less flexible than python language when it comes to model development. Rapid miner is less performing when it come to hyperparameter tuning, and model stacking.

- Cross-Validation was not conducted during training. Meaning that internal generalizability was based on the dataset variation performance rather than statistical folds.
- The study assumed the reliability of the studied datasets, without a domain expert validation which can affect the model trustworthiness.

5.4.4 Justification for not using cross-validation and External Validation

The dissertation did not implement k-fold cross-validation or external validation due to computational and data-related constraints, given the significant size difference in between the datasets chosen. Ranging from small (e.g. Singh 271 sample) to large and high-dimensional (e.g. Pytlak, 319,000 samples) did not allow the researcher to execute repeated training iterations across all four MLM. This is because it would have introduced considerable time and resource overhead, mainly when using XGBoost and Deep leanring techniques. To balance this, a single stratified 80/20 train-test split was used to simulate something close to a real-world deployment environment. Furthermore, external validation using clinical or real-world hospital data was not pursued due to lack of local institutional datasets and regulatory constraints associated with health data sharing. As a result, the models were only evaluated using Kaggle datasets. While this limits the ability to fully generalize the results to diverse patient populations, the use of four distinct datasets—varying in size, attribute types, and class balance—offers a degree of generalizability and robustness. The above limitations are acknowledged. Future research should incorporate k-fold cross-validation and external validation using real-world, local datasets to better assess model performance, stability, and feasibility for deployment in healthcare environments.

6. Conclusion

6.1 Conclusion

The research that was pursued in this study could possibly help researchers and healthcare system to broaden their perspectives of machine learning models and their performance in different settings. Specifically, Random Forest, Support Vector Machines, Deep Learning, and XGBoost were used to predict heart disease. As previously stated, four different datasets were sourced from Kaggle. These represented real data complexities that can be encountered in clinical environments. The main aim of the above research was not exactly to find the most accurate model, as other research focus on, but rather to understand how each of these predictive models act in realistic, varied contexts.

Among all the developed models, as stated in section 5.2, XGBoost demonstrated to be the most reliable with a strong performance across nearly all datasets. Its ability to handle complexity, noise, and diverse data distributions enhanced its possibility of integration into healthcare settings. Subsequently, Deep Learning models also greatly performed, even when small datasets were involved, which has been a surprising discovery. This is because,

neural networks are typically associated with large-scale data for training purpose (the more the data, the more training, the better is the performance). Yet these findings suggest that if preprocessing and balancing of datasets are well performed, conventional data-size limitations can be overcome.

One of the most relevant parts of the research was addressing dataset imbalances through SMOTE technique. This method improved the model's ability to correctly identify patients who might otherwise be overlooked (recall), critical for heart disease prediction. However, the study lacks of certain characteristics that could have enhanced model generalization further. Deeper Hyperparameter tuning and cross-validation. The researcher reached this conclusion thanks to the comparative analysis assessed in section 5.3. Both the strengths and areas for improvement in this project were evaluated.

The previous existing studies, similar to this dissertation, leveraged public and akin datasets. However, the latter often employed advanced tuning and validation methods, resulting in proofed results and enhanced performance metrics. However, the current project was made to prioritize transparency and reproducibility, showing how intricate methods may not always translate to effective real-world applications.

To conclude, the research demonstrated the possibility to build powerful predictive ML models even with little domain knowledge and public datasets. The researcher considered it relevant to show how each machine learning models behave in different environments. Models such as XGBoost and Deep Learning appeared to be the most flexible when it comes to early heart disease detection. Acknowledging certain methodological constraints allowed the researcher to be aware of the research improvement, even if the results substantively contribute to ongoing efforts to integrate machine learning into medical diagnostics.

6.2 Recommendations for Future Work

Future possibilities were gathered from the results of this research. Thanks to the comparison of the above research to others, where domain knowledge is present, it was concluded that future studies need to consider a deeper extensive tuning and applicable cross-validation. Methods such as Grid search or 10-fold cross-validation methods could help improve model performance and make the models more reliable. Subsequently, External validation of the best working model with a realistic Maltese dataset would ensure that the models created would achieve good results even when applied to unknown datasets. Hybrid modelling is again another possibility that should be not overseen in future works. The combination of traditional MLM such as XGBoost together with NN's could provide robust diagnostic tools. Subsequently, cardiologists and ML specialists need to be implemented in similar researches. Another notable limitation of this research lies in the application of SMOTE before the train-test split. While the technique was effective in improving recall, its implementation on the full dataset may have resulted in data leakage, where synthetic data gets incorporated knowledge from the test set into the training process. This could lead to an extreme enhancement of model performance. Future studies should

apply SMOTE after partitioning the data, ensuring that only the training set is oversampled, to provide a more realistic assessment of model performance. Additionally, future research should tackle whether missing values in specific features carry hidden predictive significance, rather than treating them purely as noise or loss, especially in cases where missingness itself may correlate with disease presence. This would enhance clinical relevance and help the researcher achieve better model interpretability, ensuring that the predictions can actually be used in a real-world context. Moreover, if real-time data would be incorporated in the research (wearable devices insights and IoT) it could enhance predictive capabilities, ensuring faster response in medical emergencies. Finally, this research surely represents a strong base in machine model developing. However, these two fields, Machine learning and cardiology, are extremely complex realities that need to be scraped deeper to achieve greater certainties on results.

7. References:

World Heart Federation (2023). *World heart report 2023*. [online] *World Heart Federation*. Available at: https://world-heart-federation.org/wp-content/uploads/World-Heart-Report-2023.pdf.

Keragon.com. (2021). *When Was AI First Used in Healthcare? The History of AI in Healthcare*. [online] Available at: https://www.keragon.com/blog/history-of-ai-in-healthcare#:~:text=Early%202000s%3A%20Foundation%20and%20Prototypes,(EHRs)%20and%20medical%20images. [Accessed 5 Nov. 2024].

Staff, C. (2024). *What Is Machine Learning? Definition, Types, and Examples*. [online] Coursera. Available at: https://www.coursera.org/articles/what-is-machine-learning?utm_medium=sem&utm_source=gg&utm_campaign=B2C_EMEA__coursera_FTCOF_career-academy_pmax-multiple-audiences-country-multi&campaignid=20858198824&adgroupid=&device=c&keyword=&matchtype=&network=x&devicemodel=&adposition=&creativeid=&hide_mobile_promo&gad_source=1&gclid=Cj0KCQiAoae5BhCNARIsADVLzZcqLN4RrkvKKaFwQH_C3iI9gHPxOfEUi-123QrW22topTw1G7uf6jkaAgZoEALw_wcB [Accessed 5 Nov. 2024].

Desai, D.S. and Said Hajouli (2023). *Arrhythmias*. [online] Nih.gov. Available at: https://www.ncbi.nlm.nih.gov/books/NBK558923/ [Accessed 8 Nov. 2024].

World (2024). *Cardiovascular diseases*. [online] Who.int. Available at: https://www.who.int/europe/news-room/fact-sheets/item/cardiovascular-diseases [Accessed 8 Nov. 2024].

Eurostat (2020). *Deaths due to coronary heart diseases in the EU*. [online] Europa.eu. Available at: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20200928-1 [Accessed 8 Nov. 2024].

The Texas Heart Institute®. (2020). *Categories of Arrhythmias*. [online] Available at: https://www.texasheart.org/heart-health/heart-information-center/topics/categories-of-arrhythmias/ [Accessed 8 Nov. 2024]

Wang, H., Gao, C., M Guignard-Duff, Cole, C., Hall, C., Larman, M., R Baruah, Gao, H., Mamza, J.B., Lang, C.C. and Mordi, I. (2023). Importance of early diagnosis and treatment of heart failure across the spectrum of ejection fraction. *European Heart Journal*, [online] 44(Supplement_2). doi:https://doi.org/10.1093/eurheartj/ehad655.892.

Ox.ac.uk. (2023). *Cardiovascular disease cost the European Union economy €282bn in 2021 — Nuffield Department of Population Health*. [online] Available at: https://www.ndph.ox.ac.uk/news/cardiovascular-disease-cost-the-european-union-economy-20ac282bn-in-2021#:~:text=Overall%2C%20cardiovascular%20disease%20cost%20the,%E2%82%AC620%20per%20EU%20citizen. [Accessed 10 Nov. 2024].

Grosser, D.K. (2021). *Atrial Fibrillation: The Benefits of Early Detection - Kent Cardio*. [online] Kent Cardio. Available at: https://kentcardio.com/early-detection-of-atrial-fibrillation/ [Accessed 10 Nov. 2024].

England, K. (2015). Epidemiology of cardiovascular mortality in the Maltese Islands. *Um.edu.mt*. [online] doi:https://www.um.edu.mt/library/oar//handle/123456789/14072.

Muhammad, Y., Tahir, M., Hayat, M. and Chong, K.T. (2020). Early and accurate detection and diagnosis of heart disease using intelligent computational model. Scientific Reports, [online] 10(1). doi:https://doi.org/10.1038/s41598-020-76635-9.

Altintas, Z., Fakanya, W.M. and Tothill, I.E. (2014). Cardiovascular disease detection using bio-sensing techniques. *Talanta*, [online] 128, pp.177–186. doi:https://doi.org/10.1016/j.talanta.2014.04.060.

An, Q., Rahman, S., Zhou, J. and Kang, J.J. (2023). A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges. *Sensors*, [online] 23(9), p.4178. doi:https://doi.org/10.3390/s23094178.

Javaid, M., Haleem, A., Singh, R.P., Suman, R. and Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, [online] 3, pp.58–73. doi:https://doi.org/10.1016/j.ijin.2022.05.002.

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R. and Feng, D. (2014). Early diagnosis of Alzheimer's disease with deep learning. [online] doi: https://doi.org/10.1109/isbi.2014.6868045.

Miotto, R., Wang, F., Wang, S., Jiang, X. and Dudley, J.T. (2017). Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics, [online] 19(6), pp.1236–1246. doi: https://doi.org/10.1093/bib/bbx044.

Habehh, H. and Gohel, S. (2021). Machine Learning in Healthcare. Current Genomics, [online] 22(4), pp.291–300. doi: https://doi.org/10.2174/1389202922666210705124359.

Rajpurkar, P., et al. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks.

Pu, C., Hu, X., Sangying Lv, Wu, Y., Yu, F., Zhu, W., Zhang, L., Fei, J., He, C., Ling, X., Wang, F. and Hu, H. (2022). Identification of fibrosis in hypertrophic cardiomyopathy: a radiomic study on cardiac magnetic resonance cine imaging. European Radiology, [online] 33(4), pp.2301–2311. doi: https://doi.org/10.1007/s00330-022-09217-0.

Sadeghi, Z., Alizadehsani, R., CIFCI, M.A., Kausar, S., Rehman, R., Mahanta, P., Bora, P.K., Almasri, A., Alkhawaldeh, R.S., Hussain, S., Alatas, B., Shoeibi, A., Moosaei, H., Hladík, M., Nahavandi, S. and Pardalos, P.M. (2024). A review of

Explainable Artificial Intelligence in healthcare. Computers & Electrical Engineering, [online] 118, p.109370. doi: https://doi.org/10.1016/j.compeleceng.2024.109370.

Pal, M. and Parija, S. (2021). Prediction of Heart Diseases using Random Forest. *Journal of Physics: Conference Series*, 1817(1), p.012009. doi:https://doi.org/10.1088/1742-6596/1817/1/012009.

Baral, S., Satpathy, S., Pati, D.P., Mishra, P. and Pattnaik, L. (2024). A Literature Review for Detection and Projection of Cardiovascular Disease Using Machine Learning. EAI Endorsed Transactions on Internet of Things, [online] 10. doi: https://doi.org/10.4108/eetiot.5326.

Garg, A., Sharma, B. and Khan, R. (2021). Heart disease prediction using machine learning techniques. IOP Conference Series: Materials Science and Engineering, [online] 1022(1), p.012046. doi: https://doi.org/10.1088/1757-899x/1022/1/012046.

IBM (2023). Support Vector Machine. [online] Available at: https://www.ibm.com/think/topics/support-vector-machine [Accessed 3 Feb. 2025].

Keylabs (2024). *Support Vector Machines (SVM): Fundamentals and Applications | Keylabs*. [online] Keylabs: latest news and updates. Available at: https://keylabs.ai/blog/support-vector-machines-svm-fundamentals-and-applications/ [Accessed 11 Apr. 2025].

Son, Y.-J., Kim, H.-G., Kim, E.-H., Choi, S. and Lee, S.-K. (2010). Application of Support Vector Machine for Prediction of Medication Adherence in Heart Failure Patients. Healthcare Informatics Research, [online] 16(4), p.253. doi: https://doi.org/10.4258/hir.2010.16.4.253.

Duraisamy, B., Sunku, R., Selvaraj, K., Pilla, V.V.R. and Sanikala, M. (2024). Heart disease prediction using support vector machine. Multidisciplinary Science Journal, [online] 6, pp.2024ss0104. doi: https://doi.org/10.31893/multiscience.2024ss0104.

Sutanto, H. (2024). Transforming clinical cardiology through neural networks and deep learning: A guide for clinicians. Current Problems in Cardiology, [online] 49(4), p.102454. doi: https://doi.org/10.1016/j.cpcardiol.2024.102454.

Yang, J.C. (2024). The prediction and analysis of heart disease using XGBoost algorithm. Applied and Computational Engineering, [online] 41(1), pp.61–68. doi: https://doi.org/10.54254/2755-2721/41/20230711.

Moore, A. and Bell, M. (2022). XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study. Clinical Medicine Insights: Cardiology, [online] 16. doi: https://doi.org/10.1177/11795468221133611.

Ördekçi, O. (2025). Heart Disease. Kaggle.com. [online] doi: https://doi.org/10326308/95bfd024a1dc7d059e7b6f3632ed37c9.

Pytlak, K. (2022). Indicators of Heart Disease (2022 UPDATE). Kaggle.com. [online] Available at: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease [Accessed 19 Feb. 2025].

Aboutalebi, A. (2023). Heart Disease. Kaggle.com. [online] Available at: https://www.kaggle.com/datasets/amirmahdiabbootalebi/heart-disease [Accessed 19 Feb. 2025].

Singh, U. (2024). Heart Disease Prediction Dataset. Kaggle.com. [online] Available at: https://www.kaggle.com/datasets/utkarshx27/heart-disease-diagnosis-dataset [Accessed 21 Feb. 2025].

Gogtay, N. and Thatte, U. (2017). Principles of sample size calculation. Indian Journal of Ophthalmology, [online] 58(6), p.517. doi: https://doi.org/10.4103/0301-4738.71692.

Soni, T., Gupta, D. and Uppal, M. (2024). Optimizing Heart Disease Prediction with Random Forest: Insights from the Kaggle Dataset. [online] pp.741–744. doi: https://doi.org/10.1109/aece62803.2024.10911595.

Prakash, J., Singh, S. and Balamurugan, G. (2024). A Comprehensive Evaluation for Coronary Heart Disease Prediction Using Deep Learning Models. [online] pp.153–159. doi: https://doi.org/10.1109/cicn63059.2024.10847529.

Wang, T. (2024). Predicting heart disease risk using machine learning: A comparative study of multiple algorithms. Theoretical and Natural Science, 35(1), pp.112–118. doi: https://doi.org/10.54254/2753-8818/35/20240925.

Tyagi, P. (2023). Heart Disease Detection using Machine Learning. Ijarsct. [online] Available at: https://ijarsct.co.in/Paper8596.pdf [Accessed 31 Mar. 2025

8. Appendix

Appendix A: Preprocessing and Modeling Code

A.1 Python scripts for data cleaning (missing values, encoding, normalization)

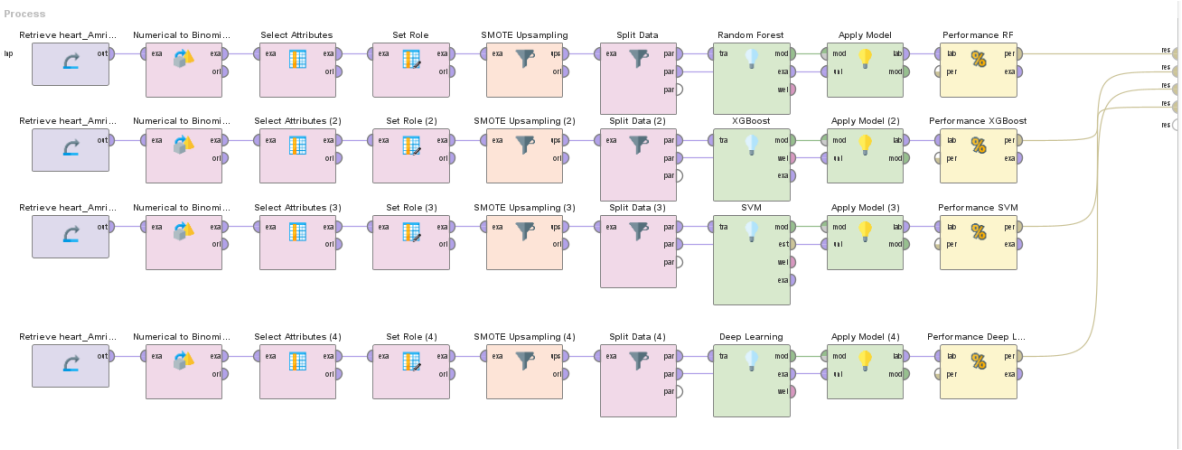
```python
import pandas as pd # Use pandas for datasets
import seaborn as sns #Visualize if any missing values
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler

df = pd.read_csv("heart_Amirmahdi_Aboutalebi.csv") # Load the dataset

#display first few rows
print(df.head())

#check datatypes and missing values
print(df.dtypes)
print(df.isnull().sum()) #Check if any Missing values after fix

#visualization of missing values
plt.figure(figsize=(10,6))
sns.heatmap(df.isnull(), cbar=False, cmap="viridis")
plt.title("Missing values Amirmahdi Dataset")
plt.show()

# Convert Binary Categorical Values (Yes/No, Male/Female) → 0/1
binary_map = {"Y": 1, "N": 0, "M": 1, "F": 0}
df.replace(binary_map, inplace=True)

#One hot encoding
#df = pd.get_dummies(df, columns=["ChestPainType", "RestingECG","ST_Slope"], drop_first=False)


#Convert True/False to 1/0
#df=df.astype(int)

# Define numerical columns To normalize
#num_cols = ["Age", "RestingBP", "Cholesterol","MaxHR"]

# Apply Min-Max Scaling
#scaler = MinMaxScaler()
#df[num_cols] = scaler.fit_transform(df[num_cols])

df.to_csv("heart_Amrimahdi_Not_encoded_or_normalized.csv", index=False)
```

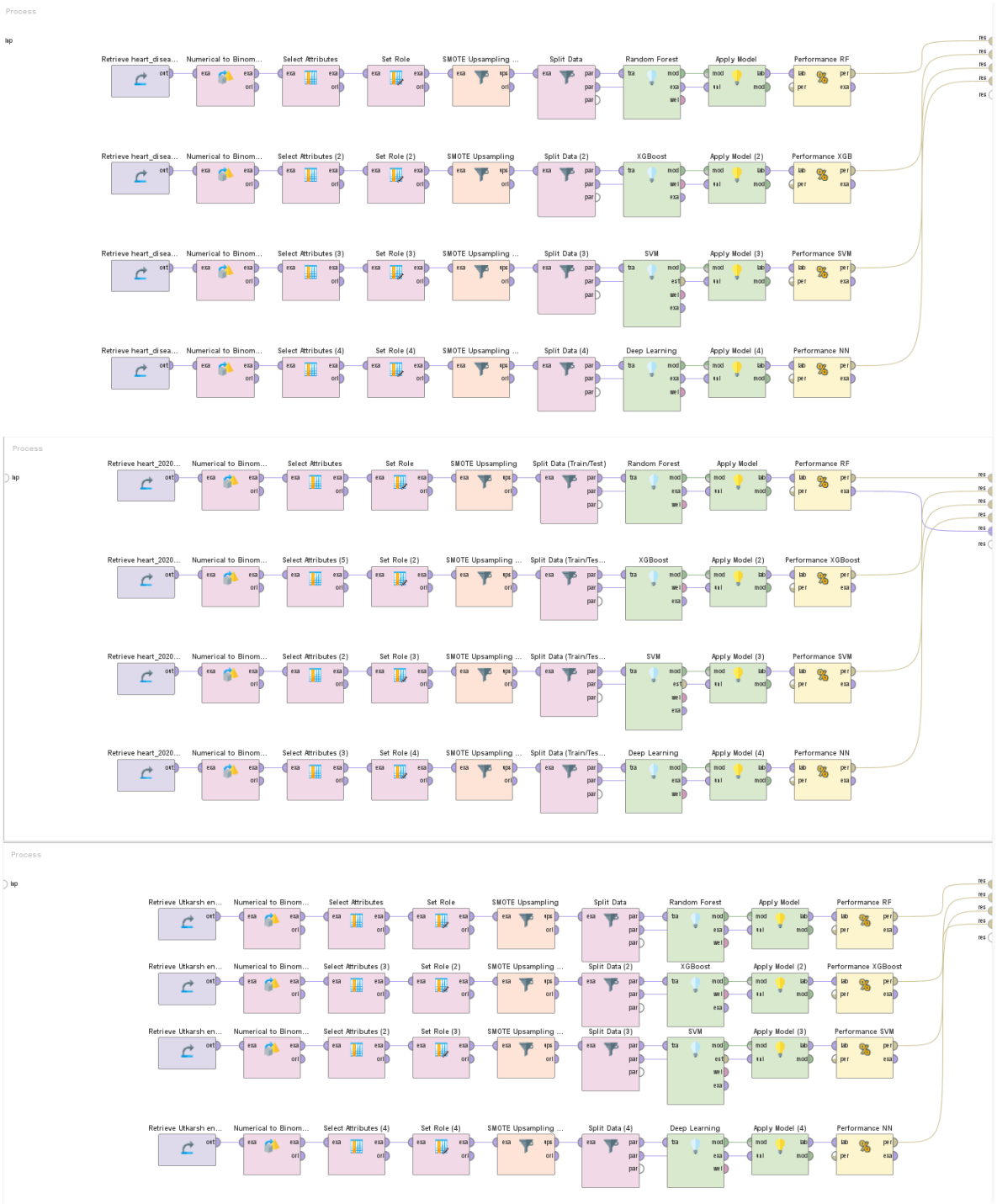
```python
import pandas as pd # Use pandas for datasets
import seaborn as sns #Visualize if any missing values
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler

df = pd.read_csv("dataset_heart_Utkarsh Singh.csv") # Load the dataset

#display first few rows
print(df.head())

#check datatypes and missing values
print(df.dtypes)
print(df.isnull().sum()) #Check if any Missing values after fix

#visualization of missing values
plt.figure(figsize=(10,6))
sns.heatmap(df.isnull(), cbar=False, cmap="viridis")
plt.title("Missing values for Utkarsh Singh Dataset")
plt.show()

#One hot encoding
df = pd.get_dummies(df, columns=["chest pain type", "resting electrocardiographic results","ST segment", "tha

df["heart disease"] = df["heart disease"].replace({1: 0, 2: 1})

#Convert True/False to 1/0
df=df.astype(int)

# Define numerical columns
num_cols = ["age", "resting blood pressure", "serum cholestoral", "max heart rate", "oldpeak", "major vessels

# Apply Min-Max Scaling
scaler = MinMaxScaler()
df[num_cols] = scaler.fit_transform(df[num_cols])

df.to_csv("Utkarsh encoded normalized.csv", index=False)
```

```python
import pandas as pd # Use pandas for datasets
import seaborn as sns #Visualize if any missing values
import matplotlib.pyplot as plt

df = pd.read_csv("heart_2020_cleaned_Kamil Pytlak.csv") # Load the dataset

#display first few rows
print(df.head())

#check datatypes and missing values
print(df.dtypes)
print(df.isnull().sum()) #Check if any Missing values

#identify categorical columns
categorical_cols = df.select_dtypes(include=["object"]).columns
print("categorial columns", categorical_cols)

#visualization of missing values
plt.figure(figsize=(10,6))
sns.heatmap(df.isnull(), cbar=False, cmap="viridis")
plt.title("Missing values for Heart_2020_cleaned_Kamil Pytlak")
plt.show()

#convert categorical to binary values
#binary_map =  {"Yes": 1, "No": 0,"Yes (during pregnancy)": 1, "No, borderline diabetes": 0, "Male": 1, "Fema
#df.replace(binary_map, inplace = True)

#One hot encoding
#df = pd.get_dummies(df, columns=["AgeCategory", "Race", "GenHealth" ], drop_first=False)


#Convert True/False to 1/0
#df=df.astype(int)


df.to_csv("heart_2020_Not_Encoded_Kamil Pytlak", index=False)
```

```python
import pandas as pd # Use pandas for datasets
import seaborn as sns #Visualize if any missing values
import matplotlib.pyplot as plt
#from sklearn.preprocessing import MinMaxScaler

df = pd.read_csv("heart_disease_Oktay Ördekçi.csv") # Load the dataset

#display first few rows
print(df.head())

#check datatypes and missing values
print(df.dtypes)
print(df.isnull().sum())

df.drop(columns=["Alcohol Consumption"], inplace=True)

# Convert Binary Categorical Values (Yes/No, Male/Female) → 0/1
binary_map = {"Yes": 1, "No": 0, "Male": 1, "Female": 0}
df.replace(binary_map, inplace=True)


#One hot encoding
#df = pd.get_dummies(df, columns=["Exercise Habits", "Stress Level", "Sugar Consumption"], drop_first=False)


# Separate categorical and numerical columns
categorical_cols = df.select_dtypes(include=['object']).columns  # Non-numeric
numerical_cols = df.select_dtypes(include=['number']).columns  # Numeric

# Fill missing numerical values with median (Safe Fix ✅)
df[numerical_cols] = df[numerical_cols].apply(lambda x: x.fillna(x.median()))

# Fill missing categorical values with mode (Safe Fix ✅)
df[categorical_cols] = df[categorical_cols].apply(lambda x: x.fillna(x.mode()[0]))

print(df.isnull().sum()) #Check if any Missing values after fix

#visualization of missing values
plt.figure(figsize=(10,6))
sns.heatmap(df.isnull(), cbar=False, cmap="viridis")
plt.title("Missing values for Heart_Oktay Ordekcin")
plt.show()

# Define numerical columns to normalize
#num_cols = ["Age", "Blood Pressure", "Cholesterol Level", "BMI", "Triglyceride Level", "Fasting Blood Sugar", "CRP Level", "Homocysteine Level"]

#Convert True/False to 1/0
#df=df.astype(int)

# Apply Min-Max Scaling
#scaler = MinMaxScaler()
#df[num_cols] = scaler.fit_transform(df[num_cols])

df.to_csv("heart_disease_Oktay_not_encoded_or_normalized.csv", index=False)
```

A.2 RapidMiner workflow, model development and pipeline screenshots for each dataset.




Process
Retrieve heart_disea...
Numerical to Binom...
Select Attributes
Set Role
SMOTE Upsampling ...
Split Data
Random Forest
Apply Model
Performance RF
Retrieve heart_disea...
Numerical to Binom...
Select Attributes (2)
Set Role (2)
SMOTE Upsampling
Split Data (2)
XGBoost
Apply Model (2)
Performance XGB
Retrieve heart_disea...
Numerical to Binom...
Select Attributes (3)
Set Role (3)
SMOTE Upsampling ...
Split Data (3)
SVM
Apply Model (3)
Performance SVM
Retrieve heart_disea...
Numerical to Binom...
Select Attributes (4)
Set Role (4)
SMOTE Upsampling ...
Split Data (4)
Deep Learning
Apply Model (4)
Performance NN
Process
Retrieve heart_2020...
Numerical to Binom...
Select Attributes
Set Role
SMOTE Upsampling
Split Data (Train/Test)
Random Forest
Apply Model
Performance RF
Retrieve heart_2020...
Numerical to Binom...
Select Attributes (5)
Set Role (2)
SMOTE Upsampling ...
Split Data (Train/Tes...
XGBoost
Apply Model (2)
Performance XGBoost
Retrieve heart_2020...
Numerical to Binom...
Select Attributes (2)
Set Role (3)
SMOTE Upsampling ...
Split Data (Train/Tes...
SVM
Apply Model (3)
Performance SVM
Retrieve heart_2020...
Numerical to Binom...
Select Attributes (3)
Set Role (4)
SMOTE Upsampling ...
Split Data (Train/Tes...
Deep Learning
Apply Model (4)
Performance NN
Process
Retrieve Utkarsh en...
Numerical to Binom...
Select Attributes
Set Role
SMOTE Upsampling
Split Data
Random Forest
Apply Model
Performance RF
Retrieve Utkarsh en...
Numerical to Binom...
Select Attributes (3)
Set Role (2)
SMOTE Upsampling ...
Split Data (2)
XGBoost
Apply Model (2)
Performance XGBoost
Retrieve Utkarsh en...
Numerical to Binom...
Select Attributes (2)
Set Role (3)
SMOTE Upsampling ...
Split Data (3)
SVM
Apply Model (3)
Performance SVM
Retrieve Utkarsh en...
Numerical to Binom...
Select Attributes (4)
Set Role (4)
SMOTE Upsampling ...
Split Data (4)
Deep Learning
Apply Model (4)
Performance NN

Appendix B: Dataset Descriptions and Sources

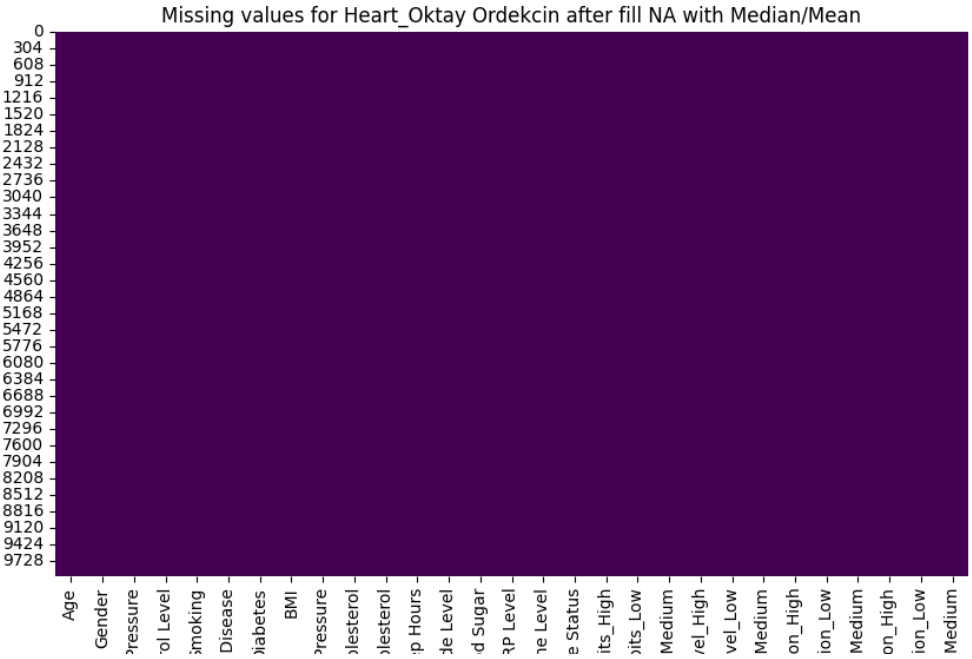
B.1 Overview of datasets (source, number of rows/attribute, type of data, usability, upvote)

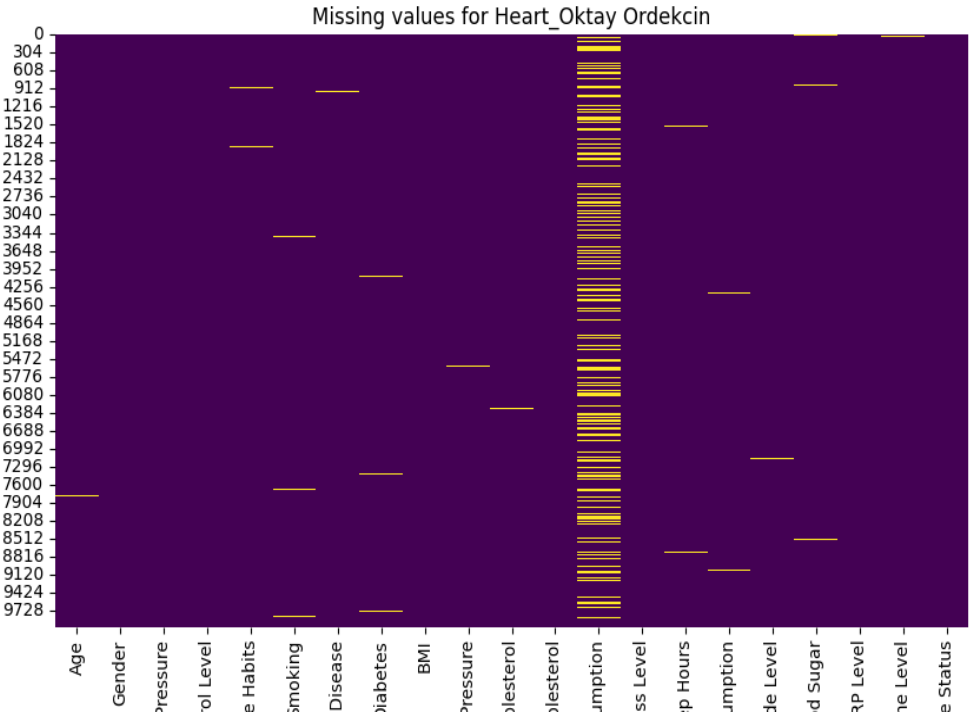
Dataset	Attributes	Classification Label	Usability	Upvote	N. of rows
Dataset heart Utkarsh Singh	Age, Sex, Chest pain type, resting blood pressure, Serum Cholesterol, Fasting blood sugar, Resting electrocardiographic results, Max heart rate, Exercise-Induced angina, old peak, ST segment, Major vessels, Thal	heart disease 1/0	10/10.	86	271
Heart 2020 Kamil Pytlak	BMI (Body mass index), Smoking, Alcohol Drinking, Stroke, Physical Health, Mental Health, Diff Walking, Sex, Age Category, Race, Diabetic, Physical Activity, Gen Health, Sleep Time, Asthma, Kidney Disease Skin Cancer	Heart Disease 1/0	9.4/10	888	319795
Heart Amirmahdi Aboutalebi Dataset	Age, Sex, Chest Pain Type, Resting BP, Cholesterol, Fasting BS, Resting ECG, Max HR, Exercise Angina, Old Peak, ST Slope.	Heart Disease 1/0	8.8/10	52	919
Heart Disease Oktay Ördekci Dataset	Age, Cholesterol Level, Smoking, Family Heart Disease, Diabetes, BMI, High Blood Pressure, Low HDL Cholesterol, High LDL Cholesterol, Alcohol Consumption, Stress Level, Sleep Hours, Sugar Consumption, Triglyceride Level, Fasting Blood Sugar, CRP Level, Homocysteine Level	Heart Disease Status 1/0	10/10	75	10001

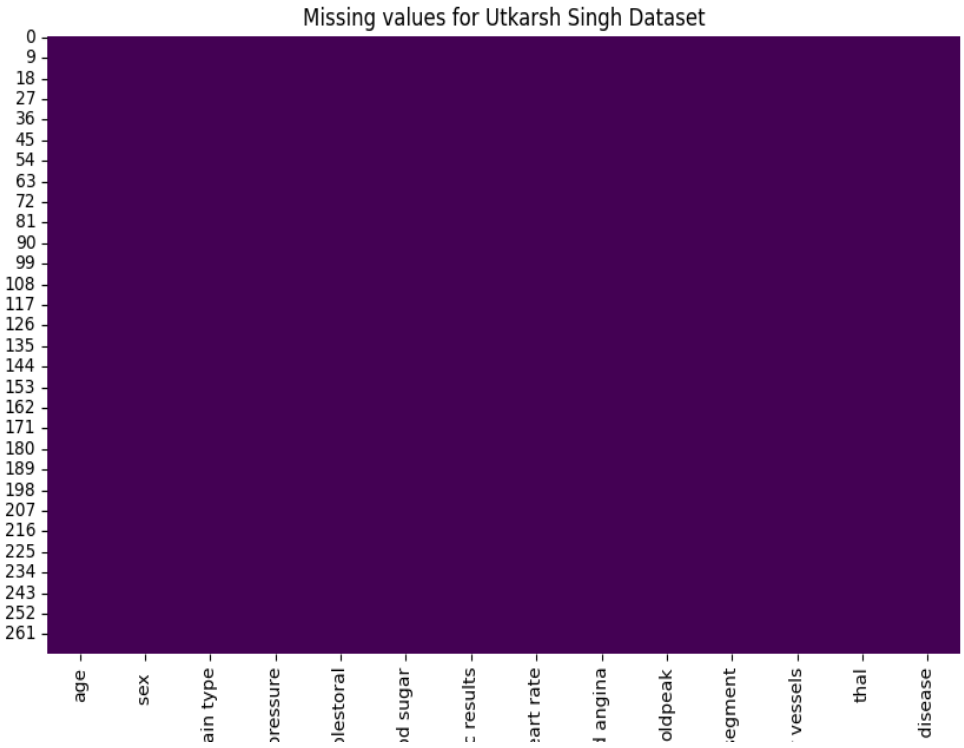
B.2 Links to Kaggle datasets used

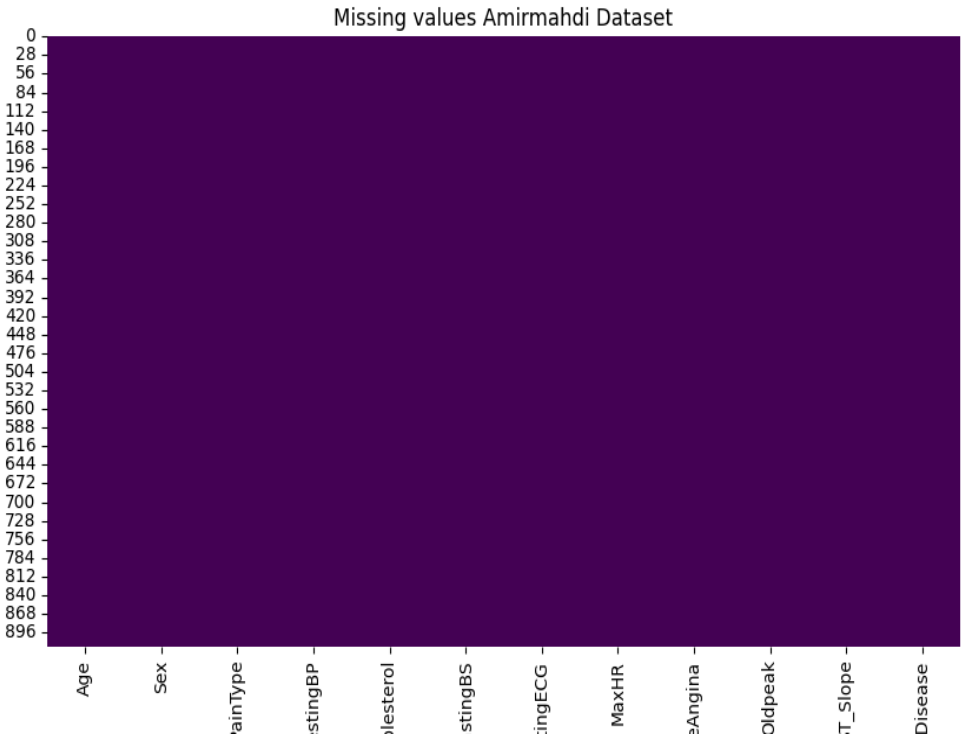
https://www.kaggle.com/datasets/oktayrdeki/heart-disease
https://www.kaggle.com/datasets/amirmahdiabbootalebi/heart-disease
https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
https://www.kaggle.com/datasets/utkarshx27/heart-disease-diagnosis-dataset

B.3 Missing values (per dataset)



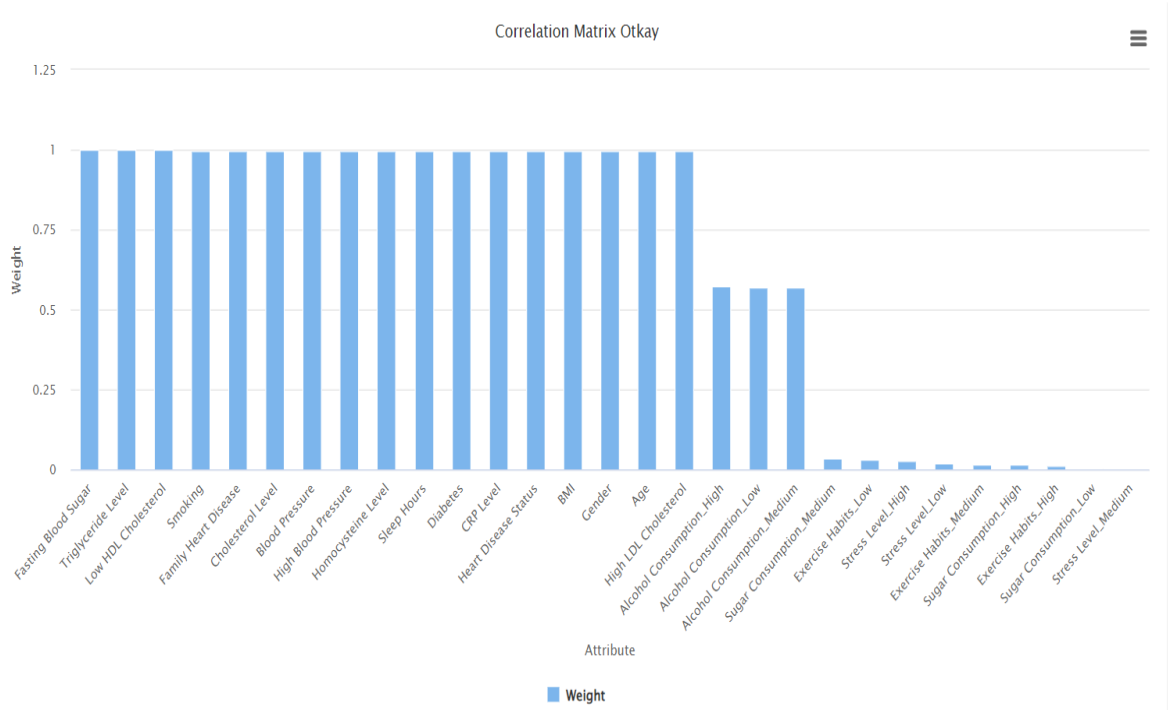


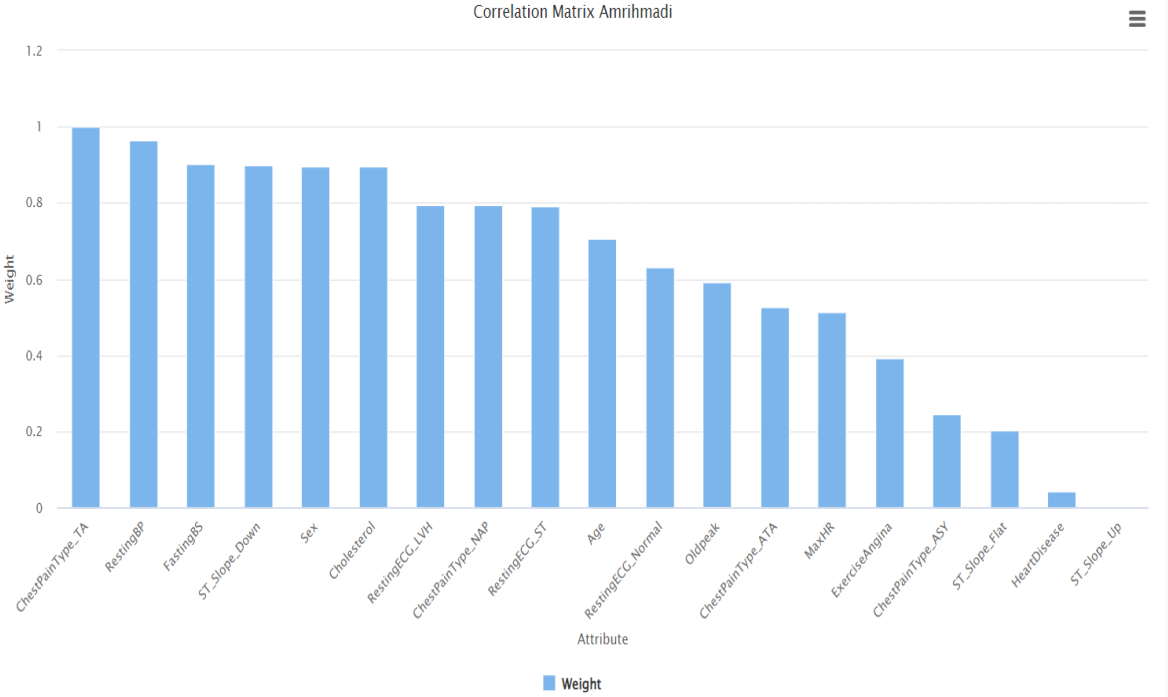

Missing values for Utkarsh Singh Dataset
0
9
18
27
36
45
54
63
72
81
90
99
108
117
126
135
144
153
162
171
180
189
198
207
216
225
234
243
252
261
age
sex
ain type
ressure
olestoral
d sugar
results
art rate
angina
oldpeak
egment
vessels
thal
disease

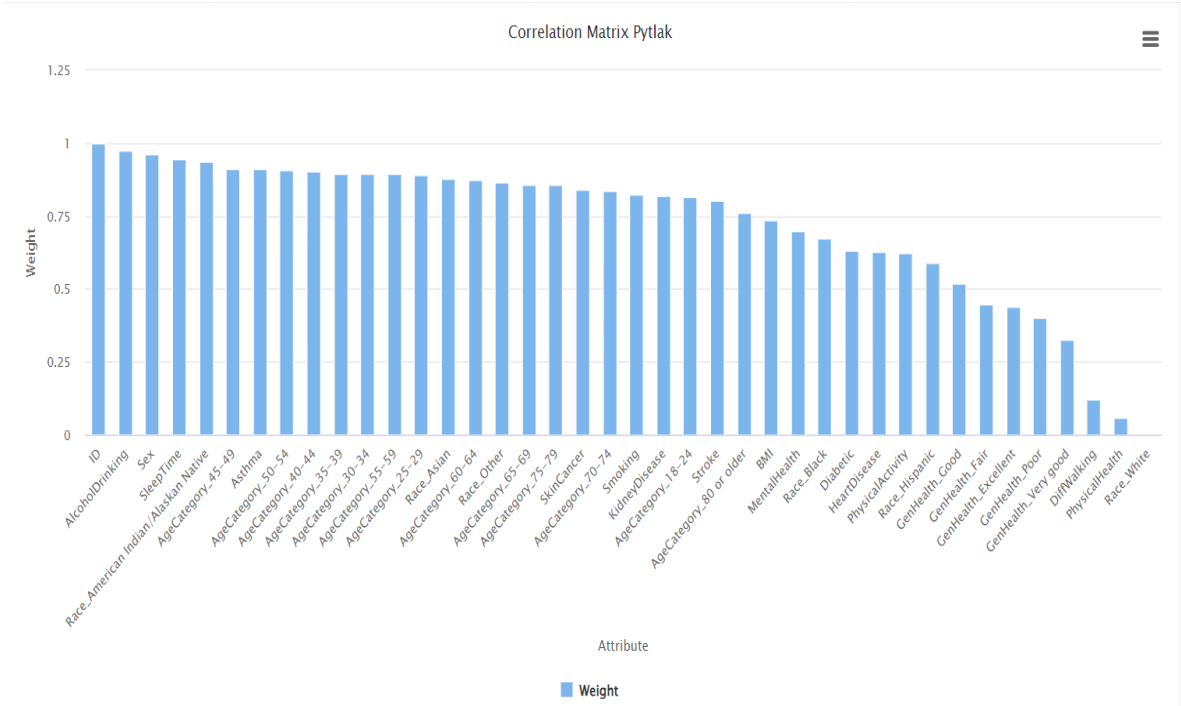

Missing values Amirmahdi Dataset
0
28
56
84
112
140
168
196
224
252
280
308
336
364
392
420
448
476
504
532
560
588
616
644
672
700
728
756
784
812
840
868
896
Age
Sex
ainType
stingBP
olesterol
stingBS
ingECG
MaxHR
eAngina
Oldpeak
T_Slope
Disease

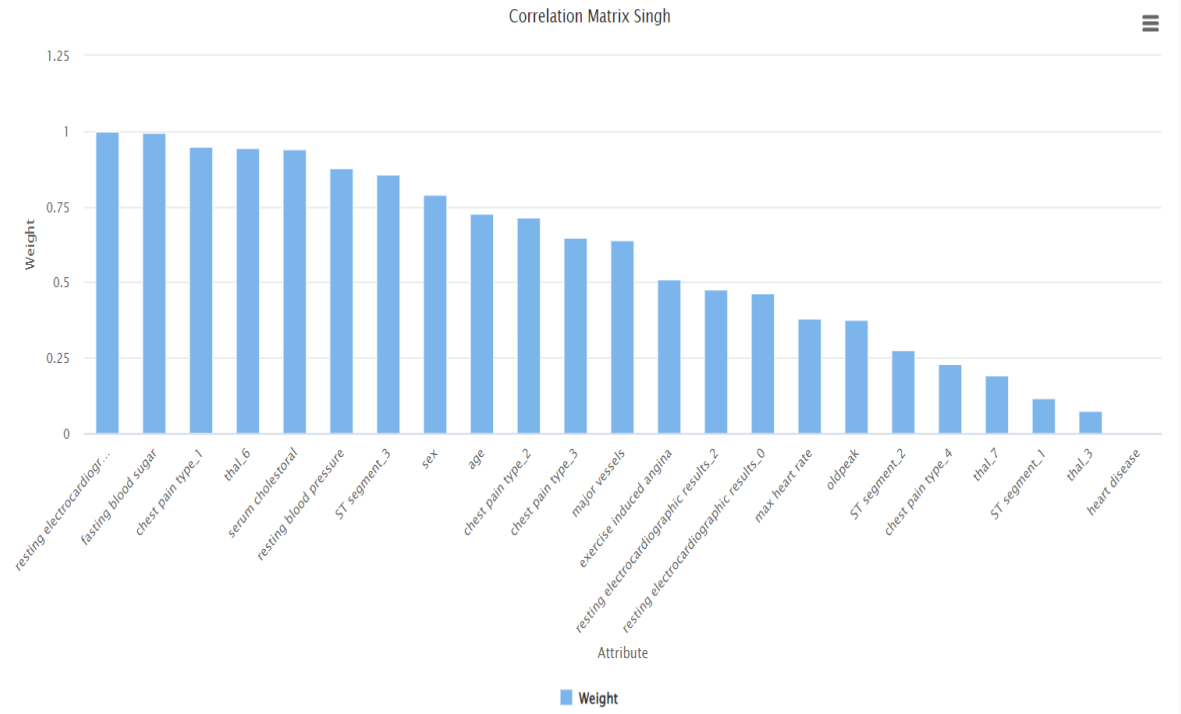
Appendix C: Feature Selection Results

C.1 Correlation matrices for each dataset

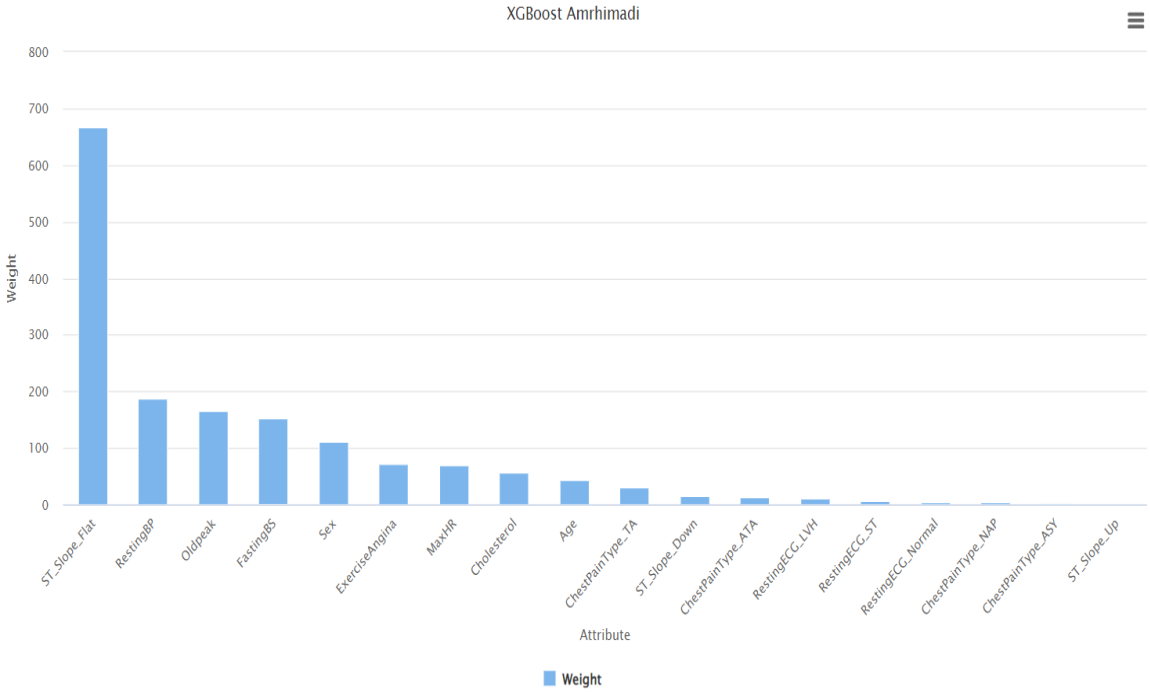


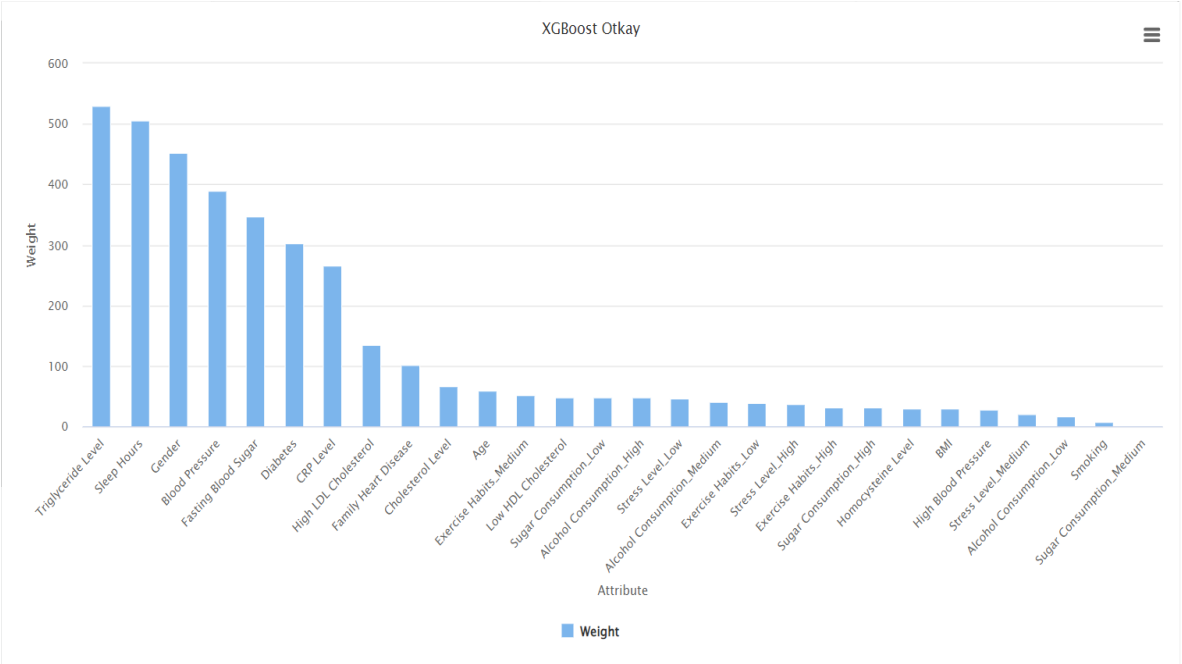


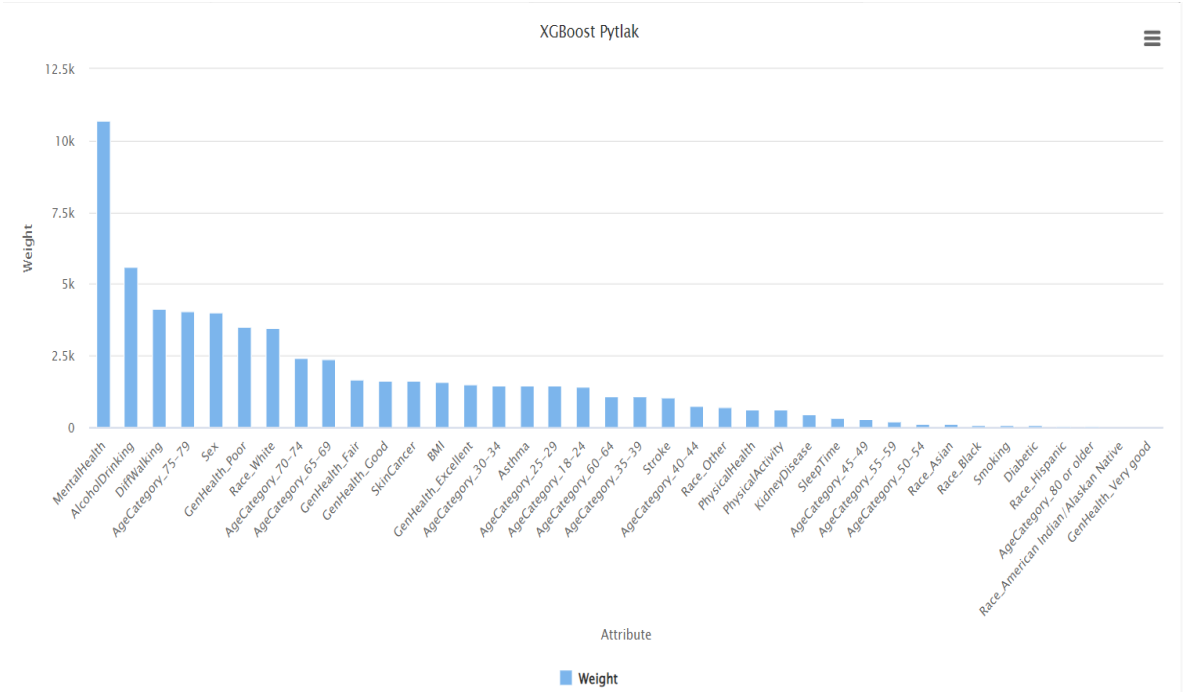

Correlation Matrix Pytlak
1.25
1
0.75
0.5
0.25
0
Weight
ID
AlcoholDrinking
Sex
SleepTime
Race_American Indian/Alaskan Native
AgeCategory_45-49
Asthma
AgeCategory_50-54
AgeCategory_40-44
AgeCategory_35-39
AgeCategory_30-34
AgeCategory_55-59
AgeCategory_25-29
Race_Asian
AgeCategory_60-64
Race_Other
AgeCategory_65-69
AgeCategory_75-79
SkinCancer
AgeCategory_70-74
Smoking
KidneyDisease
AgeCategory_18-24
Stroke
AgeCategory_80 or older
BMI
MentalHealth
Race_Black
Diabetic
HeartDisease
PhysicalActivity
Race_Hispanic
GenHealth_Good
GenHealth_Fair
GenHealth_Excellent
GenHealth_Poor
GenHealth_Very good
DiffWalking
PhysicalHealth
Race_White
Attribute
Weight

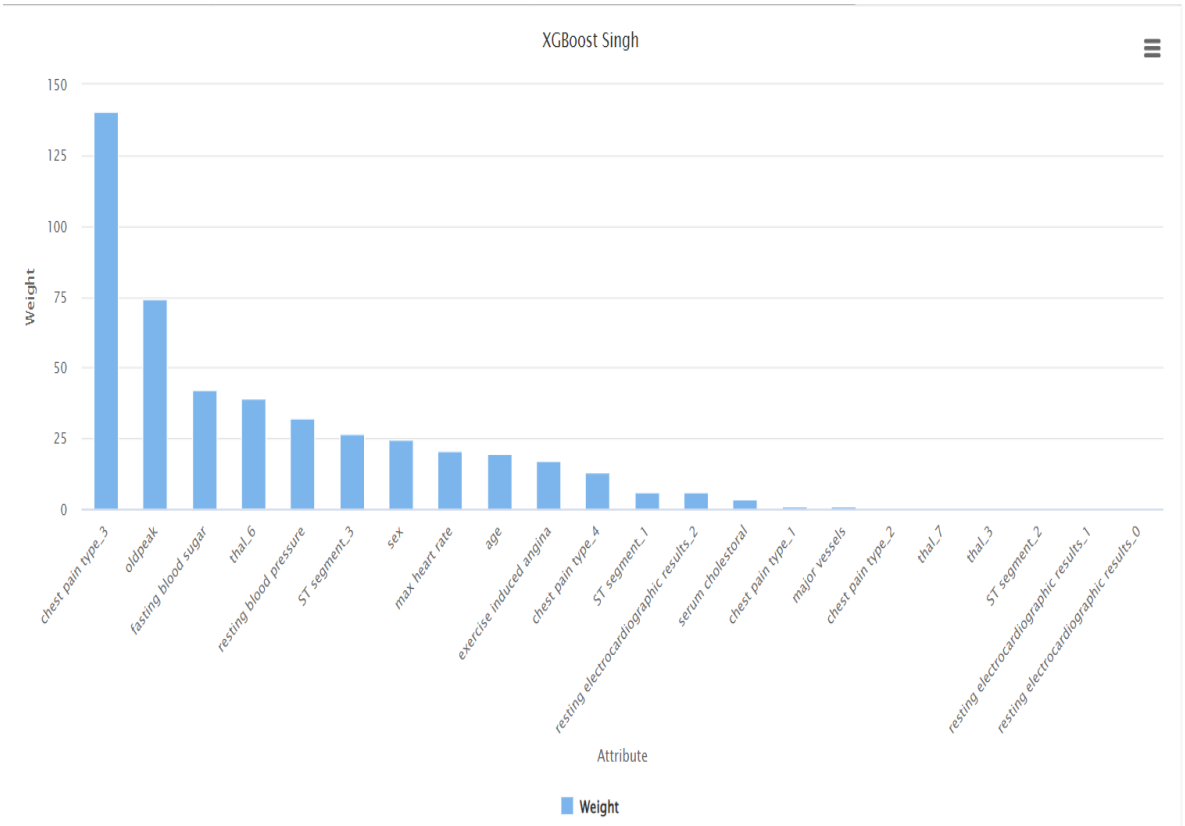

Correlation Matrix Singh
1.25
1
0.75
0.5
0.25
0
Weight
resting electrocardiogr...
fasting blood sugar
chest pain type_1
thal_6
serum cholestoral
resting blood pressure
ST segment_3
sex
age
chest pain type_2
chest pain type_3
major vessels
exercise induced angina
resting electrocardiographic results_2
resting electrocardiographic results_0
max heart rate
oldpeak
ST segment_2
chest pain type_4
thal_7
ST segment_1
thal_3
heart disease
Attribute
Weight

C.2 XGBoost feature importance rankings






XGBoost Pytlak
12.5k
10k
7.5k
5k
2.5k
0
Weight
MentalHealth
AlcoholDrinking
DiffWalking
AgeCategory_75-79
Sex
GenHealth_Poor
Race_White
AgeCategory_70-74
AgeCategory_65-69
GenHealth_Fair
GenHealth_Good
SkinCancer
BMI
GenHealth_Excellent
AgeCategory_30-34
Asthma
AgeCategory_25-29
AgeCategory_18-24
AgeCategory_60-64
AgeCategory_35-39
Stroke
AgeCategory_40-44
Race_Other
PhysicalHealth
PhysicalActivity
KidneyDisease
SleepTime
AgeCategory_45-49
AgeCategory_55-59
AgeCategory_50-54
Race_Asian
Race_Black
Smoking
Diabetic
Race_Hispanic
AgeCategory_80 or older
Race_American Indian/Alaskan Native
GenHealth_Very good
Attribute
Weight

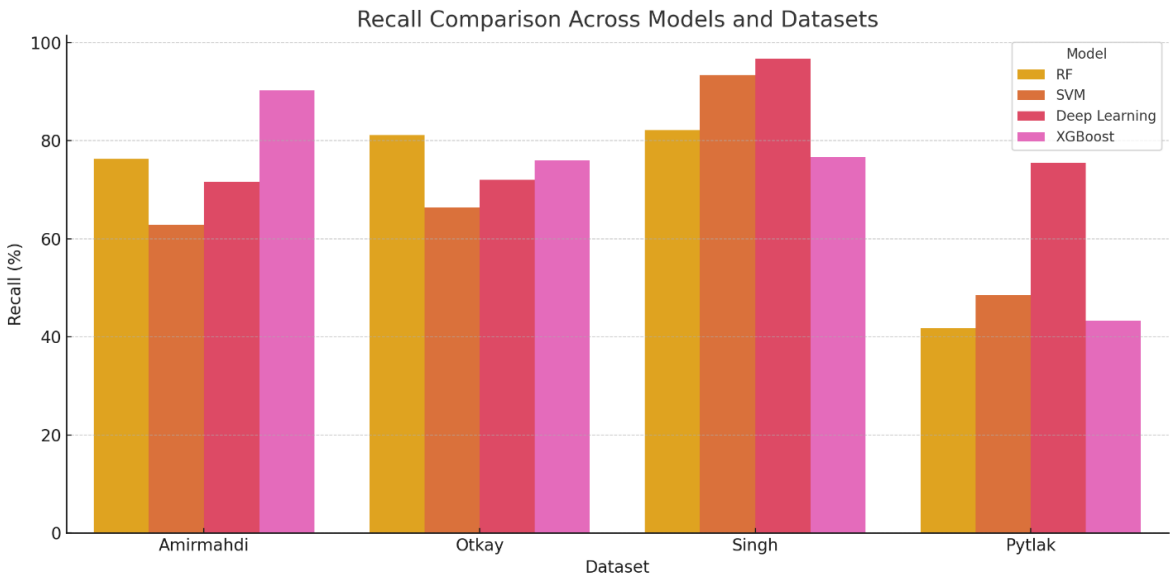

XGBoost Singh
150
125
100
75
50
25
0
Weight
chest pain type_3
oldpeak
fasting blood sugar
thal_6
resting blood pressure
ST segment_3
sex
max heart rate
age
exercise induced angina
chest pain type_4
ST segment_1
resting electrocardiographic results_2
serum cholestoral
chest pain type_1
major vessels
chest pain type_2
thal_7
thal_3
ST segment_2
resting electrocardiographic results_1
resting electrocardiographic results_0
Attribute
Weight

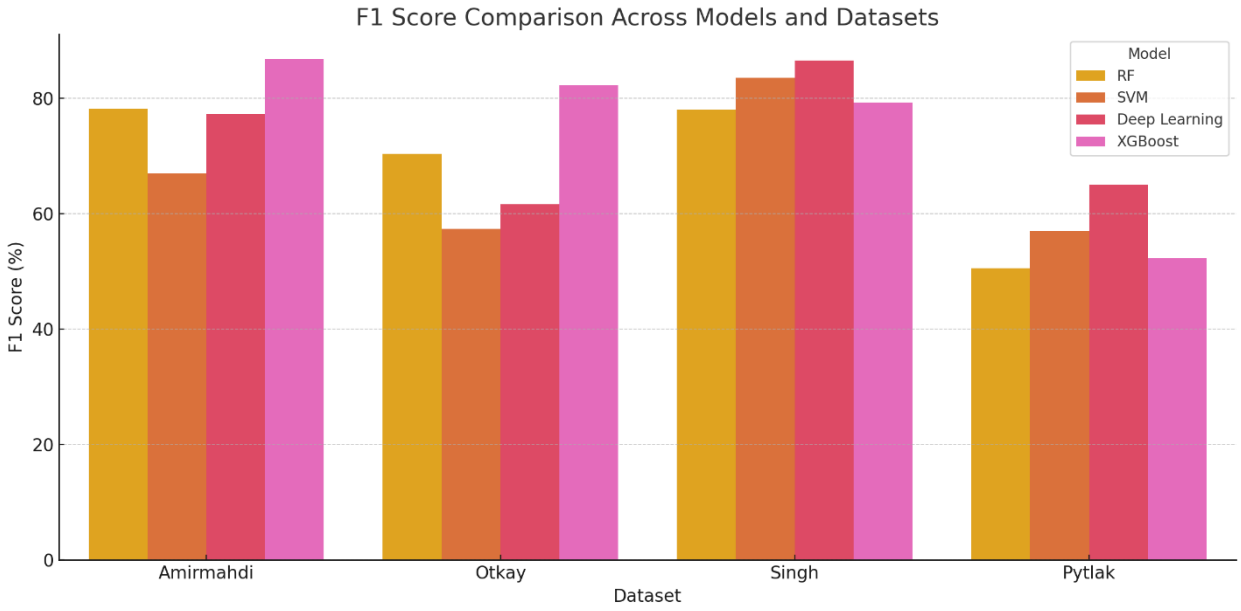
C.3 Final features selected per dataset

Dataset	*Features Removed*	*Feature Selection Used*	*Models Improved*	*Accuracy Impact*
Amirmahdi	No	Correlation	None	Accuracy dropped if features removed
Otkay	No	Correlation	None	Accuracy dropped if features removed
Pytlak	No	Correlation	None	Accuracy dropped if features removed
Singh	Yes (chest pain type 4, maximum heart rate, old peak)	Correlation	SVM, RF	Accuracy ↑ (SVM: +4%, RF: +6%) Accuracy ↓ (XGBoost: -14%, NN: -9%)

Appendix D: Model Performance Metrics

D.1 Recall, F1-score breakdown per class and Dataset





Appendix E: Hyperparameters Used

Random Forest	Number of trees: 100
	Criterion: Gain Ratio
	Maximal depth: 10
	Pruning and prepruning – not applied
	Guess subset ratio
	Voting strategy: confidence vote
	Parallel execution enabled
XGBoost	Booster: Tree booster
	Rounds: 25
	Early stopping: none
	Learning rate 0.3
	Min split loss: 0.0
	Max depth: 6
	Min child weight: 1.0
	Subsample: 1.0
	Tree method: auto
	Lambda: 1.0
	Alpha: 0.0
Support Vector Machine	Kernel type: Dot
	Kernel cache: 200
	C: 0.0
	Convergence epsilon: 0.001
	Max iterations:100000

	Scale: Applied
Deep Learning Neural Network	Activation: Rectifier
	Hidden layer sized: 2 layers - 50 / 50
	Epochs: 10.0
	Train samples per iteration: -2
	Adaptive rate: applied
	Epsilon: 1.0E-8
	Rho: 0.99
	Standardize: applied
	L1: 1.0E-5
	L2: 0.0
	Max w2: 10.0
	Loss function: Automatic
	Distribution function: AUTO
	Missing values handling: MeanImputation
	Early stopping: not-applied

Appendix F: Ethical Considerations

F.1 Public data licensing (e.g., Kaggle terms)

https://www.kaggle.com/discussions/general/116302