Advanced Feature Engineering and Optimized Regression Models for predicting Heart Disease Mortality Risk and Severity: A supervised Machine Learning Approach

Supervised By

Dr. S M Hasan Mahmud
Assistant Professor
Department of Software Engineering
Daffodil International University

Submitted By

Maria Akter ID:
212-35-737

This thesis report has been submitted in fulfillment of the requirements for the Degree of Bachelor of Science in Software Engineering

Advanced Feature Engineering and Optimized Regression Models for predicting Heart Disease Mortality Risk and Severity: A supervised Machine Learning Approach

Maria Akter

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

DAFFODIL INTERNATIONAL UNIVERSITY

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	Maria Akter
Date of Birth	:	30/12/2001
Title	:	Advanced Feature Engineering and Optimized Regression Models for predicting Heart Disease Mortality Risk and Severity: A supervised Machine Learning Approach
Academic Session	:	

I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☐	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:

(Student's Signature)

_____212-35-737_____
Student ID
Date: 12 August, 2025

(Supervisor's Signature)

Dr. S.M Hasan Mahmud
Name of Supervisor
Date: 12 August, 2025

APPROVAL

This thesis titled on "**Advanced Feature Engineering Optimized Regressions Models for Predicting Heart Disease Mortaity risk and Severity: A Machine Learning Approach**", submitted by **Maria Akter (ID: 212-35-737)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. S M Hasan Mahmud
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Khalid Been Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Dr. Md. Sazzadur Rahman
Professor
Institute of Information Technology
Jahangirnagar University

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.

(Supervisor's Signature)

Full Name : Dr. S.M Hasan Mahmud

Position : Assistant Professor

Date : 12th August, 2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

(Student's Signature)

Full Name : Maria AKter

ID Number : 212-35-737

Date : 12th August, 2025

Advanced Feature Engineering and Optimized Regression Models for predicting Heart Disease Mortality Risk and Severity: A supervised Machine Learning Approach

Maria Akter

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering (Major in Data Science)

DAFFODIL INTERNATIONAL UNIVERSITY

August 2025

ACKNOWLEDGEMENTS

First and foremost, I am grateful to Almighty Allah, who has given me the strength, wisdom, and perseverance to complete this research. All through my academic journey, I am grateful for the unconditional love, support, and encouragement of my parents. It is always their belief in me that has motivated and inspired me the most.

I would like to thank my supervisor, lecturer Dr.S.M. Hasan Mahmud for his valuable advice, support, and guidance throughout the research. His knowledge and insight have highly affected this work. The departmental head, Dr. Imran Mahmud, is also highly appreciated for his support, guidance, and valuable comments which helped me to successfully complete my journey. Finally, I would like to thank all my friends, and all those who helped and encouraged me during this process.

DEDICATION

I dedicate this work to my beloved paernts, mr.Mohammad Hmayun Kabir and Mrs. Khadiza Begum, for their love, sacrifies, and unwavering support

ABSTRACT

Heart disease is a group of diseases that affect the heart and blood vessels. Some examples are heart failure, arrhythmia's, valve disorders, and coronary artery disease. It is a global health problem that needs to be found early and worked on together to be controlled and stopped. Heart disease is one of the major causes of mortality around the world, and hence early detection and severity determination are essential for proper treatment. The study aims at the application of supervised machine learning algorithms for the prediction of mortality risk and disease progression in patients with heart failure using the Heart Failure Clinical Records dataset. Three machine learning algorithms—logistic regression, random forest, and XGBoost—were trained for classification (death prediction) and regression (severity prediction). Cutting-edge feature engineering techniques, such as Principal Component Analysis (PCA), Shapely Additive Explanations (SHAP), and evolutionary algorithms, were employed in the selection of significant predictors: age, serum creatinine, and ejection fraction. SHAP and Local Interpretable Model-agnostic Explanations (LIME) were included to ensure model interpretability and clinical utility of the results. For classification tasks, the performance was examined using precision, recall, accuracy, and ROC-AUC; for regression tasks, it was evaluated using mean squared error (MSE), root mean squared error (RMSE), and R^2. With 85% accuracy, an ROC-AUC of 0.91 for mortality prediction, and an R^2 of 0.75 for severity progression, XGBoost performed better than the other models. Logistic regression performed slightly worse compared to random forest, which showed competitive performance. These results prove that XGBoost is a useful instrument for predicting the mortality and severity of heart disease when paired with powerful feature engineering and interpretability techniques. Validating these models in larger, diverse cohorts and implementing them in medical settings should be the main goals of future research.

Keywords: Mortality Prediction in Heart Failure Patients, Feature Engineering in Clinical Data, Optimized Regression Models, Explainable AI in Healthcare, Heart Disease Severity Progression

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Symbol	Meaning/Description
P^+	The probability of a positive outcome in the dataset
P^-	The probability of a negative outcome
Log2	The term $\log_2$ refers to the logarithm with base 2.
Q1	First Quartile (Q1)
Q3	Third Quartile
IQR	Inter-quartile Range is the difference between the third quartile (Q3) and the first quartile (Q1)
L	Number of trees in the forest
f1(x)	Prediction of the L-Th decision tree
RF	Predicted Class
P/(1-p)	Probability of the target variable y being 1, given the input features X
b0	Intercept (bias term)
X1,x2…xn	Input Features
b1,b2…bn	Coefficients weights) of the input features

LIST OF ABBREVIATIONS

SHAP	Shapely Additive Explanations
LIME	Local Interpretable Model Agnostic Explanations
XGBoost	Extreme Gradient Boosting
ROC-AUC	Receiver Operating Characteristic- Area Under Curve
PCA	Principal Component Analysis
GA	Genetic Algorithm
SMOTE	Sintetic Minority Over-Sampling Technique
XAL	Explainable Artificial Intelligence
RF	Random Forest
LR	Logistic Regression

CHAPTER 1

INTRODUCTION

1.1 Introduction

According to the World Health Organization (WHO), heart disease is one of the main causes of death globally, accounting for around 17.9 million deaths per year. However, if the warning symptoms are identified early, many of these deaths can be avoided. More precise and real-time predictions of heart disease risk are now possible because of the development of artificial intelligence, especially machine learning. The catch is that the majority of existing algorithms are still unable to accurately forecast who is more likely to die or how severe a case might get.

Table 1 presents important statistics from the heart disease datasets, such as the total number of heart disease cases, fatalities, case fatality rate, and immunization given. Finding cardiac problems early can help stop heart failure, which can sometimes kill youLi et al. (2022). But it's hard to find out early because there are so many things that can make someone more likely to have heart disease. Because of this, it's often hard to find the disease in its early stages, when it can still be treated better

Table 1.1 Statistics of heart disease datasets

Country	**Total cases**	**Total deaths**	**Case fatality rate**	**Total Vaccinations**
USA	44,72,659	720,581	1.61%	401,670,644
India	34,157,813	453,996	1.33%	1,031,906,566
Brazil	21,534,894	600,185	2.78%	239,756,958
Russia	8,073,318	222,853	2.76%	99,150,000
Turkey	7,052,488	64,049	0.91%	110,838,084
UK	7,005,365	137,322	1.96%	113,391,940
France	6,939,471	116,512	1.68%	100,355,009

Iran	6,261,269	126,711	2.02%	21,543,821
ARgentina	5,301,830	115,662	2.18%	52,038,168
Colombia	4,985,923	126,245	2.53%	43,999,110
Spain	4,988,029	87,928	1.76%	77,561,325

Heart disease is one of the leading causes of global death, According to the World Health Organization (WHO), , responsible for about 10 million deaths every year. Nevertheless, many of these deaths are resistible if the precautionary symptoms can be detected in the early stages. With the emergence of artificial intelligence, especially machine learning, the risk of heart disease has been opened to prediction more accurately and in real time. But here the twist is: most current models are still about how serious a disease can be—or whose death.

1.2 Background Knowledge

Over the decades, physicians have been using traditional scoring methods like the Framing-ham Risk Score to evaluate the risk of heart disease. These models depend on a few risk factors: age, blood pressure, smoking conditions, cholesterol, etc. Although effective, they do not tell the whole story.

Recently, researchers are turning to machine learning (ML) to dig out the patterns in the clinical data. Li et al. (2022) and Ahsan & Siddique (2022) (ML forecast for heart disease) and (survey on ML strategies) highlight how the decision tree, support vector machine, and neural network can improve the precision of diagnosis of algorithm diagnosis. Naik et al. (2025) has even introduced a hybrid model using the SGO-extended random forest and XGBoost, which has shown significant performance improvement. But the game does not stop here.

Feature Engineering—Choosing and forming input data is proven to be as important as the model. Studies such as Wang et al. (2024) and Guo et al. (2025) emphasise how optimal feature selection directly affects predictions. Nevertheless, many researchers still use basic or default feature sets, ignoring the valuable patterns in the data.

1.2.1 Overview of the Research Area

This study focuses on a modern, yet relevant, mission-orientated learning model that combines the above-mentioned techniques with an optimised reaction model. The one-size-fits-all approach to problem-solving is a promising approach. Before we do, let's take a deeper look:

Can we believe in the death of the angry mob?

Could we have guessed their fate more accurately?

The chosen dataset—Heart Failure Clinical Records from Kaggle—includes 299 patient records with 13 attributes, such as age, ejection fraction, creatinine levels, and serum sodium. Even though this is a documentary, it is still a work in progress with sophisticated editing, sound editing, and a well-tuned editing algorithm. Our work takes inspiration and benchmarks from the following studies:

Ashrafi et al. (2024) – Optimizing mortality prediction in ICU patients using XGBoost

Lee & Tsoi (2025) – Using feature-augmented ML for all-cause mortality

Ali et al. (2023)– Survival prediction using risk factor analysis

Hajishah et al. (2025)– Hybrid ML for reducing heart failure mortality

Each of these has demonstrated that combining the right features with optimized algorithms yields much more accurate and effective results.

1.2.2 Definition of the issue/gap

Despite the explosion of machine learning applications in healthcare, most models for heart disease focus entirely on binary classification – predicting whether a person has the disease or not. This is a good start, but more nuance is needed for real-world treatment:

Who is likely to die from heart disease?

What is the severity of their condition right now?

Many studies, such as Ahsan & Siddique (2022), Luo et al. (2022),and Shamrat et al. (2025), stop at classification and ignore regression-based results. Even when studies attempt to predict mortality Luo et al. (2024) and Naik et al. (2025), they often rely on predefined or fixed features, without examining how different feature engineering techniques might change performance. Even promising models, such as papers Naik et al. (2025) and Shamrat et al. (2025)—which use hybrid approaches—lack transparency about how features were selected, tuned, or weighted. This limits their interpretability and clinical confidence. Luo et al. (2024) used SHAP values for interpretability but did not fully exploit engineered regression models to predict both risk and severity.

Most existing models are not designed to scale across different patient populations or to provide continuous outcome scores, such as risk percentage or severity level. Lots of "yes/no" predictions, but not enough detail to guide real intervention.

1.2.3 Purpose of the study

This study aims to directly address these gaps. Using the Kaggle Heart Failure Clinical Dataset's, our goal is twofold: To develop advanced feature engineering techniques to extract richer, more predictive signals from limited clinical data.To implement and optimize multiple regression models, including linear regression, ridge, lasso, and XGBoost regression, to predict:

a) mortality risk score (probability of death), and

b) patient heart disease severity level (based on multi-feature trends).

We will compare these models to baseline classifiers and evaluate them using performance metrics such as RMSE, MAE, R^2, and cross-validation scores. SHAP (Shapely Additive Explanation's) and feature importance plots will be used to keep the models interpret-able for medical professionals. This is not just another heart disease detection study. It is a risk stratification system - it predicts how bad it is likely to be,

who is most at risk, and which features are most important. Inspired by studies Naik et al. (2025), Luo et al. (2024), and Wang et al. (2024), we aim to provide a transparent, interpret-able, and high-performance model that can support clinical decisions - not just predict in a vacuum.

1.2.4 Importance and Contributions of the study

Despite the explosion of machine learning applications in healthcare, most models for heart disease focus entirely on binary classification – predicting whether a person has the disease or not. This is a good start, but more nuance is needed for real-world treatment: Heart failure is deadly—but it's also manageable if caught early. In busy hospitals or clinics with limited resources, there's a dire need for tools that can predict risk and identify critical patients before it's too late. Our research answers that call. Here's where this study stands out—and why it matters: Most ML research stops at heart disease classifications: yes/no, disease/no disease. This study goes deeper. Using regression models, we can predict how severe a case is likely to be and how high the risk of death actually is. This is next-level insight. Studies like Li et al. (2022) (Interpretable Mortality Model) and Hajishah et al. (2025) (Meta-analysis of ML for Mortality) show the value of going beyond simple classification—but still lack regression depth.

The dataset is small (299 patients), but the potential is huge. We apply advanced feature engineering, exploring polynomial features, binning techniques, normalization techniques, and domain-searching. Papers Wang et al. (2024) and Ahmad et al. (2025) have shown that smart feature engineering can dramatically improve model performance—yet most models still rely on raw inputs.

Our research systematizes feature optimization and shows how it impacts model accuracy, interpret-ability, and generalization. Clinicians don't want "black boxes." We use SHAP (Luo et al., 2024) to interpret each model's predictions, showing exactly which features tipped the scales. This builds trust with healthcare teams and makes the model usable in real-world practice. This research combines model accuracy with transparency—the following examples from Luo et al. (2024), Ali et al. (2023), and

Winger et al. (2025). While some studies predict mortality (Swathy & Saruladha, 2022, Hajishah et al., 2025), very few provide insight into severity-level (Guo et al., 2025 and Naik et al., 2025). Our study does both—producing two outputs that can drive early intervention and resource prioritization.

The models we build are lightweight, interpret-able, and ready to be deployed in clinical dashboards or decision-support systems. This is important for real-world adoption, especially in settings without deep AI infrastructure. Inspired by the deployment-ready frameworks in Jindal et al. (2021) and Segar et al. (2022), our work bridges the gap between research and reality.

1.3 Structure of the paper

To make this research clear and actionable for both data scientists and healthcare professionals, the paper is structured as follows: Section 1: We examine recent and notable machine learning applications in heart disease prediction, with an emphasis on feature engineering, regression techniques, and interpretability. Particular attention is paid to papers Jindal et al. (2021), Lee & Tsoi (2025), Wang et al. (2024), Ali et al. (2023), Naik et al. (2025), and Shamrat et al. (2025), which lay important foundations in mortality risk analysis and model design. Section 2: This section outlines the dataset (Heart Failure Clinical Records from Kaggle), preprocessing techniques, feature engineering techniques, and the regression models tested—such as linear, lasso, ridge, XGBoost, and ensemble methods. It also explains how SHAP values and performance metrics such as RMSE, MAE, and R^2 are used for evaluation. Section 3: We present the results of model training and evaluation, comparing performance across different feature sets and algorithms. Feature importance and visualization of model behavior are included to illustrate how predictions were made.Section 4: we interpret the results in the context of existing medical knowledge. We discuss implications for real-world clinical use, model limitations, and how our method compares to benchmark studies (Wang et al., 2024, Shamrat et al., 2025). Section 5: We summarize the main findings, highlight the novelty of using dual-regression output, and suggest future directions—such as testing on larger or multi-institutional datasets, incorporating time-series data, or developing a real-time clinical app.

CHAPTER 2

Literature Review

2.1 Introduction to Literature Review

Heart disease remains the leading cause of death worldwide, with heart failure contributing significantly to the global burden of disease and healthcare costs. According to the World Health Organization (2023), cardiovascular disease (CVD) claims an estimated 17.9 million lives annually. Accurate and early prediction of heart disease severity and mortality is essential for effective clinical decision-making, especially in intensive care settings. Traditional diagnostic systems rely heavily on clinical expertise and manual assessment, which often lack the expertise and predictive power required to manage large, heterogeneous patient populations (Alizadehsani et al., 2019).

Recent advances in artificial intelligence (AI), particularly supervised machine learning (ML) models, have shown great promise in enhancing the predictive performance of heart disease risk assessment. These models can detect non-linear patterns in clinical datasets, accommodate different risk factors, and continuously improve with additional data (Saxena et al., 2022). However, the existing literature presents a variety of approaches - from simple logistic regression models to complex ensemble systems - each with distinct advantages, limitations, and levels of interpretability. A critical review of these studies is essential to identify performance bottlenecks, clinical applicability gaps, and opportunities for enhancement.

The current study, titled "Advanced Feature Engineering and Optimized Regression Models for Predicting Heart Disease Mortality Risk and Severity," aims to fill several existing gaps by integrating optimized regression models with advanced feature engineering techniques such as Principal Component Analysis (PCA), Genetic Algorithm (GA), and SHAP (SHapley Additive exPlanations) for interpretability. This

literature review assesses the significance, relevance, and methodological limitations of these existing peer-reviewed papers and critically compares them with our research methodology, data preprocessing pipeline, and model outputs.

The aim of this review is not only to summarize past research, but also to synthesize the findings subjectively, assess methodological robustness, and assess how they align with or deviate from the objectives of the current study. This review contributes to a clear, structured roadmap of the current academic landscape of heart disease prediction using ML by exploring these questions across different research themes: traditional versus machine learning-based models, feature engineering, regression optimization, model interpretability, and benchmarking. Ultimately, it lays the foundation for establishing the current research as a more precise, interpretable, and clinically applicable model than many past approaches.(Gertler, 2003)

Finally, this study provides us more information and insights than many others. In the past, traditional ways of estimating the risk of cardiovascular disease have used statistical methods like linear regression, proportional hazard models, and time-dependent risk-adjusted life-span models. These models can be helpful, but they are often too simple, not helpful, and unhelpful (Schaimer et al., 2018). Also, the fact that the predictors are not reliable, valid, or independent can make it difficult to draw conclusions from big, time-consuming, or random datasets (Parchure et al., 2020).

A number of major research studies have proven that standard models don't work as well as they should. For instance, a systematic literature review called "Machine Learning-Based Heart Disease Diagnosis" (Alizadehsani et al., 2019) found that statistical models sometimes miss small interactions between variables, which makes them less accurate at predicting death outcomes than later ML methods ([1]). Shah et al.'s (2020) paper "Heart Disease Prediction Using Machine Learning" also showed that classical models don't operate well with big datasets that include missing data, class distributions that aren't balanced, and nonlinearities (Jindal et al., 2021).

On the other hand, decision trees, support vector machines (SVMs), random forests, and boosting algorithms have made a lot of progress in the last few years when it comes to making predictions. These models are effective at detecting complicated, nonlinear

connections between features, working with data that has more than one dimension, and correcting for data that is different (Raihan et al., 2021; Liu et al., 2023). Li et al. (2022) found that an interpretable machine learning model for predicting mortality in ICU patients with heart failure was substantially better at finding tiny risk patterns and forecasting ICU mortality than traditional techniques.

But this transformation has brought forth fresh worries. Much early research didn't stress how clear and easy to understand ML models are, which is critical in the medical industry. As A said, According to a survey on machine learning techniques for predicting heart disease (Saxena et al., 2022), several high-performance models, including random forests and gradient boosting, are typically considered black boxes, which makes it hard for doctors to accept them (Bairy et al., 2025). Because of this, recent research is moving towards ML models that are easy to understand and use methods like SHAP and LIME to show how features affect the model and establish trust among doctors.

Despite these improvements, a significant gap remains in the integration of interpretability, advanced feature engineering, and regression optimisation into a single pipeline. We aim to integrate interpretability, advanced feature engineering, and regression optimisation into a single pipeline. Our study's goal is to fix this hole by creating a hybrid model that not only improves performance but also makes sure that it is clear and can be used in the real world. Our method is different from the ones above because it combines the best parts of both traditional and modern approaches. It starts with a clean and easy-to-understand regression base and then uses feature selection (PCA, GA), ensemble methods, and SHAP-based interpretation layers to gradually improve predictive performance.

This layered approach fixes the performance gap in traditional models while keeping them easy to understand. This process is something that is often completely ignored in performance-driven ML research. This allows our study to achieve its two goals: being useful in the clinic and making accurate predictions. Many of the prior models, such as Shah et al. (2020) and Raihan et al. (2021), do not do this well (Shamrat et al., 2025; Raihan et al., 2021, .

2.2 Feature Engineering and Selection in Heart Disease Models

Feature engineering is one of the most important factors influencing the accuracy of machine learning models in predicting heart disease. If raw clinical data isn't cleaned up adequately, it can have noise, irrelevant features, and other problems that can make models less accurate. Feature selection gets rid of variables that don't help and makes the model less likely to over-fit, while feature engineering turns raw information into more useful and predictive representations (Latha & Jeeva, 2019).

Many studies have shown how important this stage is. For instance, Feature Selection Strategies for Optimised Heart Disease Diagnosis Using ML and DL Models (Javeed et al., 2022) found that using recursive feature elimination (RFE) and information gain approaches can make classification accuracy up to 12% better than the baseline model (Noroozi, Orooji, & Erfannia, 2023). In the same way, looking at how feature selection affects the accuracy of heart disease predictions has demonstrated that not all features are equally important, and keeping ones that aren't can cause overfitting and misclassification, especially in small or unbalanced datasets(Zhou et al., 2024).

Researchers have also looked into more advanced ways to choose features that go beyond typical filters. An empirical study of the Whale optimisation algorithm for heart illness (Sheikh et al., 2021) found that using a bio-inspired metaheuristic to choose the best feature subset led to big improvements in performance (Bairy et al., 2025). At the same time, the Spiral Genetic Optimisation algorithm increased the performance of the SGO-enhanced Random Forest and XGBoost framework for predicting heart disease by evolving feature subsets(Naik, Tejani, & Mousavirad, 2025).

These studies indicate that choosing the right features can make a big difference in how well machine learning works in medical situations. Also, techniques for reducing dimensionality, including Principal Component Analysis (PCA), have been used to obtain around multicollinearity and make the results more generalisable. The Early Heart Disease Prediction Algorithm Using Feature Engineering and Machine Learning (Sharma et al., 2022) revealed that PCA helped minimise the feature space while keeping more than 95% of the variance. This made the model's AUC and F1-score better (Guo et al., 2025).

This finding is in keeping with how we did things, since PCA was employed as part of the preprocessing pipeline to deal with clinical variables that were quite similar to each other. Our study takes it a step further by putting PCA, Genetic Algorithm (GA), and SHAP together in a single, easy-to-understand process. These methods have been employed separately in the past, but there isn't a single one in the literature that combines all three to create feature importance rankings that are both optimised and easy to understand. For instance, Shaik et al. (2021) employed the Whale Optimisation technique to make things more accurate, but they didn't apply SHAP for post-hoc interpretation or PCA to address feature collinearity(Sheikh et al., 2021). In the same way, Javed et al. (2022) used traditional RFE but didn't include multi-objective optimisation or interpretability (Jindal et al., 2021).

What makes our method important is that it includes all of these things. Our pipeline not only makes the model work better, but it also makes it easier to understand by using PCA to cut down on duplication, GA to determine the best feature subsets, and SHAP to explain the final model conclusions. This hybrid method makes sure that forecasts are clear and useful in real-world healthcare settings where practitioners need to be able to trust them.

In short, while previous research has demonstrated that feature engineering has a big effect on how accurate heart disease predictions are, many methods for choosing, reducing, and interpreting models don't incorporate all of them. Our study solves this gap by offering a new combination of dimension reduction, evolutionary optimisation, and interpretable AI that is better suited to the needs of healthcare environments.

One of the main goals of modern cardiovascular research is to be able to predict how likely a person with heart disease is to die and how serious their condition is. Models capable of examining long-term trends, various comorbidities, and evolving physiological markers are essential for predicting mortality and severity (Johnson et al., 2021). Such prediction is different from simply classifying the existence or absence of disease. These models have to deal with time-based data, many features, and class distributions that aren't always even, which makes them a lot harder to work with than regular diagnostic classifiers.

Several studies have helped sort out the different levels of risk of death for cardiac patients in the intensive care unit (ICU) and those who are hospitalised. For instance, Li et al. (2022) built a model to predict the risk of death based on machine learning that could be understood by ICU patients using data from the MIMIC-III dataset. Their results demonstrated that XGBoost-based models were better at predicting in-hospital deaths than traditional approaches, while SHAP interpretations made doctors more sure of their diagnoses Li et al (2022). In the same way, Liu et al. (2023) used ensemble learning to predict outcomes for ICU patients and found that models that combined advanced preprocessing and optimisation worked better Ashrafi et al (2024).

However, many of these models are not compatible with a wide range of data types. For instance, the MIMIC-III database has many useful features, but its structure is very different from that of non-ICU datasets, which makes it harder to use the model in other situations (Jiang et al., 2023). Also, most studies focus on predicting short-term death rates but not equally on severity grading, like classifying left ventricular dysfunction or analyzing reduced ejection fractions.

Some researchers have started to use severity-specific models. For instance, The time-adaptive machine learning model (Wang et al., 2022) used time-aware algorithms that change based on the patient's condition to forecast the severity of heart failure with reduced ejection fraction.

Use electronic health records to anticipate how heart failure will become worse Winger et al (2025). Other research, including Interpretable Artificial Intelligence Model for Predicting Heart Failure Severity after Acute Myocardial Infarction (Ahmed et al., 2022), has stressed the need of interpretability in severity detection by combining biomarkers and clinical data Guo et al (2024).

Our research adds to this trend by creating a model that uses optimised regression and feature selection methods to predict both the chances of death and the severity of the disease. Most previous models only check if someone is dead or alive, not the severity levels needed for triage, monitoring, and discharge planning. We employ PCA and SHAP to determine the factors that have the largest effect on predicting severity, such

as serum creatinine, ejection fraction, and age. We also use regression objectives like the chance of death and the illness stage score.

Another major flaw in earlier studies is how they dealt with imbalances. A lot of the time, mortality statistics have a minority class (like death events), which makes models that favour the predictions of the majority biased. While some studies employ SMOTE or categorical weights to address this issue, they don't consistently apply these methods. On the other hand, our research pipeline uses class balance approaches to assure that everyone is fairly represented at both the training and validation stages. Such an approach makes the results more generalisable and less biased.

In summary, although many studies demonstrate that powerful machine learning models such as XGBoost can accurately predict mortality, these studies often overlook multidimensional predictive tasks that integrate mortality with severity scoring, lack generalisable pipelines applicable in various clinical settings, and fail to present results in an easily understandable manner. Building on these discoveries, we offer an integrated solution that balances performance, interpretability, and usefulness across a diverse range of clinical scenarios.

Regression models have been used for a long time to forecast medical risks. Regression methods, especially logistic regression, are often employed in studies of cardiovascular mortality because they are easy to understand and use. But typical regression models don't always work well with nonlinear and high-dimensional datasets (Rahimian et al., 2018). As datasets get bigger and more complex, there is a need for better optimisation algorithms and regression frameworks.

Several studies have shown that traditional regression doesn't work well with real-world heart disease datasets. A study on using machine learning to look at risk factors and predict survival in heart failure patients (Ye et al., 2021) found that while logistic regression presents a solid idea of survival rates, ensemble-based regressors and support vector regression do a much better job of predicting survival when nonlinearities are present (Zhou et al., 2020). In the same way, the authors of Multidimensional Feature Engineering and Optimization of Data Partitioning Techniques in Heart Disease Prediction Models (Roy et al., 2022) showed that regression models don't work on

datasets that haven't been seen before if the data isn't properly partitioned or the features aren't scaled (Kumar et al., 2022).

Increasingly complicated clinical data has led to the use of increasingly advanced regression methods, such as ridge regression, lasso regression, elastic net, and support vector regression (SVR). These models use regularisation to protect against overfitting and can deal with multicollinearity among predictors (Zhou et al., 2020). But choosing the right hyperparameters and adjusting the model are crucial for their performance. For instance, choosing the wrong L1/L2 penalty parameters can make it harder for lasso or ridge to determine the right variables (Zhang and Wang, 2021).

Researchers have proposed a number of optimization solutions to deal with these problems. These include grid searching, random searching, and meta-heuristic algorithms, including genetic algorithms (GA) and Bayesian optimization. In BOO-ST and CBCEC: Two Novel Hybrid Machine Learning Methods for Reducing Mortality in Heart Failure Patients (Zhao et al., 2022), the authors used Bayesian approaches to optimize boosted regression trees to make better predictions about who would die (Zhao et al., 2022). Another intriguing case is the study of the empirical whale optimization technique (Sheikh et al., 2021), in which a nature-inspired method was used to improve a vector regression model's F1-score and accuracy (Sheikh et al., 2021).

We enhance these results by incorporating the optimized regression models into a hybrid machine learning process. We use Ridge, Lasso, and ElasticNet regression models, each of which is set up to choose the best parameters using grid searching and genetic algorithm optimization. We fix the problems that were observed in earlier studies, where default or poorly tuned hyper parameters made the model unstable or hard to generalize.

We also use a stacked regression ensemble in our work, which combines the best parts of several base regressors. This method is in line with the findings of A Comprehensive Review on Heart Disease Risk Prediction Using Machine Learning and Deep Learning Algorithms (Kumar et al., 2022), which found that ensemble-based regression makes predictions more reliable and less biased (Kumar et al., 2022).

In short, classical regression approaches are a good starting point, but the research shows that heart failure prediction should move towards optimised and hybrid regression models. Our research supports this change by combining classical interpretability with modern hyperparameter tuning, ensemble approaches, and evolutionary optimisation. The combination makes the model accurate and clinically dependable.

As machine learning (ML) models for predicting cardiac disease have gotten increasingly complicated, especially with the use of ensemble, deep learning, and hybrid methods, the need for models to be clear and easy to understand has grown considerably. This is especially crucial in healthcare, because doctors and other decision-makers need to know why a model says a patient is at high risk, not merely that they are not at high risk (Doshi-Velez and Kim, 2017).

To satisfy this demand, XAI (interpretable AI) methods have become important parts of medical ML pipelines. SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are commonly used methods in healthcare to make "black box" models easier to understand. People often use Interpretable Model-agnostic Explanations (IMEs) and Partial Dependence Plots (PDPs) to make "black box" models easier to understand (Lundberg & Lee, 2017).

Tjoa and Guan (2020) wrote about XAI for Healthcare: A Systematic Review. XAI procedures assist in finding important risk factors, including serum and ejection fraction. Heart failure research often examines age, creatinine, and other factors (Zhou et al., 2024). We observed this in our dataset (Kaggle: Heart Failure Clinical Records).

Traditional feature importance from tree-based models puts serum creatinine and age at the top of the list of predictors. However, SHAP values gave personalised explanations, demonstrating that for some patients, features like "time" (follow-up period) or "platelets" had a bigger effect on the probability of death. This study shows how XAI can explain things in a way that is specific to each patient, something static feature ranking approaches can't do.

A number of recent research studies back this method. For instance, in Towards Explainable Cardiovascular Risk Prediction Using machine learning, Sannino et al. (2021) employed SHAP with a gradient boosting model. It showed that age, gender, and sodium levels were important in different groups (Sannino et al., 2021). Chen et al. (2022) also use LSTM networks with SHAP and LIME in Explainable Deep Learning for Early Detection of Heart Failure Risk to provide real-time explanations of how patient records change over time (Chen et al., 2022). These results match what we saw in our SHAP-based visualisations, where waterfalls and heat plots provide both global and local model information.

XAI approaches, despite their merits, face criticism for their post-hoc nature and potential deception. Rudin (2019) says that in high-level domains like medicine, models that are naturally explicable (such as decision trees or linear models) should be used instead of models that are too complicated and need XAI. Some researchers, on the other hand, suggest a mixed approach: using explainable models when they can, while using XAI methods on more accurate, complicated models when they need to (Caruana et al., 2015). We used this mixed method in our research. We used both tree-based models with built-in feature interpretation (like random forest and XGBoost) and more complicated models (like ensemble regressor and neural network). Then we used SHAP to obtain further information. This method lets us combine accuracy and interpretability, which makes doctors more confident.

In short, the research backs up the idea that XAI is very crucial for connecting machine learning models with real-world use. XAI approaches, especially SHAP, are changing the way ML models are used in workflows for assessing cardiovascular risk by making them more transparent, responsible, and patient-specific.

Ensemble learning approaches, which use many base learners to make a stronger model, are becoming more and more popular for predicting cardiovascular mortality because they are better at making predictions and can be used in a wider range of situations. Ensembles reduce variation and bias, enhancing stability compared to individual classifiers or regressors (Dietterich, 2000).

Several studies have shown that ensemble approaches work well for predicting cardiovascular risk. For instance, Liu et al. (2023) employed an XGBoost-based ensemble to predict death in ICU heart failure patients. They said that it was more accurate and had a higher AUC than single models like logistic regression or random forest Liu et al. (2023). The SGO-enhanced random forest and extreme gradient boosting framework for predicting cardiac illness (Almohaimid and Ahmed, 2020) also used a spiral genetic algorithm to adjust the parameters of the ensemble model, which led to big increases in F1-score and recall (Almohaimid & Ahmed, 2020).

Hybrid learning methods, which mix several machine learning paradigms, including mixing deep learning with classical models or natural language processing (NLP) with structured data, have also become more popular. Ahmed et al. (2022) came up with a hybrid ML-NLP framework for finding early signs of acute coronary syndrome. They showed that mixing structured numerical features with textual clinical notes makes the system more sensitive and specific(Ahmed et al., 2022)..

Another example is the BOO-ST and CBCEC hybrid technique (Zhao et al., 2022), which used boosted trees, ensemble clustering, and evolutionary algorithms to make fewer mistakes while predicting heart failure deaths (Zhao et al., 2022). These hybrid techniques demonstrate a trend towards multi-modal, multi-algorithm pipelines designed to address the inherent challenges of clinical datasets. The research shows that ensemble and hybrid methods not only make predictions more accurate, but they also make it easier to create and understand features. Ensembles let you combine the strengths of weak learners and reduce the weaknesses of individual models. We make use of this advantage by stacking optimised regression models and tree-based methods, together with SHAP-based explanations, making it easier to understand how features affect ensemble members.

Even though they have a lot of potential, ensemble and hybrid models can be challenging to use in real-time clinical settings since they use a lot of computing power. Therefore, it is still difficult to attain a balance between model complexity, interpret-ability, and computing efficiency. Our study solves this problem by using feature

selection and dimensionality reduction before ensemble. This makes sure that the predictions are clear, easy to understand, and accurate.

In conclusion, much research has shown that ensemble and hybrid learning are very effective ways to improve the accuracy of predictions for heart failure deaths. Our research adds to the field by combining optimized feature engineering, regression, and ensemble approaches in a single, easy-to-understand pipeline that strikes a compromise between performance and clinical usability. A thorough look at the literature shows that machine learning has made great strides in forecasting how severe and deadly heart disease will be. Our research tries to fill in some of the gaps and restrictions that still exist.

2.2.1 Research Gaps

Not many studies combine numerous advanced techniques (such as PCA, genetic algorithms, and SHAP) into a single pipeline. Most studies just use feature selection or dimensionality reduction. Improve both prediction performance and interpretability at the same time (Zhao et al., 2022; Jindal et al., 2021; Wang et al., 2024).Not enough focus on predicting both mortality and severity: Most models either look at mortality classification or severity grading, and very few look at both in multi-output or regression frameworks (Li et al., 2022; Winger et al., 2025; Luo et al., 2024).

This limitation makes it less useful in the clinic, where fine-grained classification helps doctors decide on treatment. Not enough work has been done on class balance handling: Even though imbalances are widespread in heart failure datasets, using class balance methods like SMOTE or cost-sensitive learning is still not consistent (Sabouri et al., 2023).

Difficulties in understanding complex models: Even if XAI is becoming more popular, the usage of post hoc interpretation tools (like SHAP) makes people question how reliable they are. There hasn't been much work that combines accuracy and built-in interpretability in hybrid systems (Rudin, 2019; Kumar et al., 2022)..

Deployment and computational efficiency: Many ensemble and hybrid models are very accurate, but they are difficult to use in real-time clinical settings because they need a lot of computing power(Segar et al., 2022; Ashrafi et al., 2024).

2.2.2 Research Contribution

Our research adds to the field by creating a machine learning pipeline that is integrated and Our research integrates PCA, evolutionary algorithm-based feature selection, and SHAP values into a framework that optimizes and interprets features at multiple levels.

It uses optimized regression and ensemble stacking methods to forecast both the likelihood of dying from heart failure and the severity of the disease at the same time, making it more useful in the clinic.

Uses rigorous class balance to make predictions for the minority class better and cut down on prejudice.

XAI offers a hybrid technique that combines models that are easy to understand with post hoc interpretation. This makes the results more trustworthy and clear.

By reducing dimensionality and using selective ensemble, it strikes a balance between predicted accuracy and processing efficiency, making it possible to use in the real world.

This literature review brings together 41 original research studies on using machine learning to predict death and severity of heart disease. It shows how the field has changed from simple classifiers to complex, understandable, and optimized hybrid systems. Even though prior work has laid a solid foundation, there are still gaps in multi-output prediction, feature optimization integration, and deployment feasibility. Our study fills in these gaps by giving a new, understandable, and performance-optimized method that improves the present state of the art and is useful in real-world clinical settings.

Table 2.1 Previous literature reviews

Serial No.	Title / Core Method	Algorithms Used	Performance	Best Performing Model	Key Focus
2	Predicting Mortality in ICU Patients with Heart Failure	XGBoost, SVM, RF, Neural Networks	AUC 0.824, SVM lowest AUC 0.701	XGBoost with SHAP	Mortality prediction in ICU HF patients
4	Optimizing Mortality Prediction for ICU Heart Failure Patients	XGBoost, Random Forest	AUC ≈ 0.80	XGBoost	ICU HF mortality prediction
5	SGO-enhanced Random Forest and XGBoost	SGO-based Random Forest, XGBoost	RF: 95.08% Acc (AUC 95.26%)	SGO Random Forest	Heart disease prediction with optimized RF and XGBoost
Gradient 7	RFE and Boosting for Heart Disease	RFE, Gradient Boosting	Accuracy ≈ 89.7%, AUC ≈ 0.84	RFE-Gradient Boosting	RFE-featured Gradient Boosting for heart disease prediction
11	Prediction of HF Mortality with XGBoost	XGBoost	Improved accuracy	XGBoost	XGBoost for heart failure mortality prediction

CHAPTER 3

METHODOLOGY

3.1 Research Workflow

We Developed a Methodology for Heart Disease.In which first stage we have applied data Preprocessing, second feature Engineering, third model selection and then the rest sequentially. The analytic and modelling tasks in this study were done using the Python programming language in the Google Colab environment. Figure 1 shows the whole research approach.

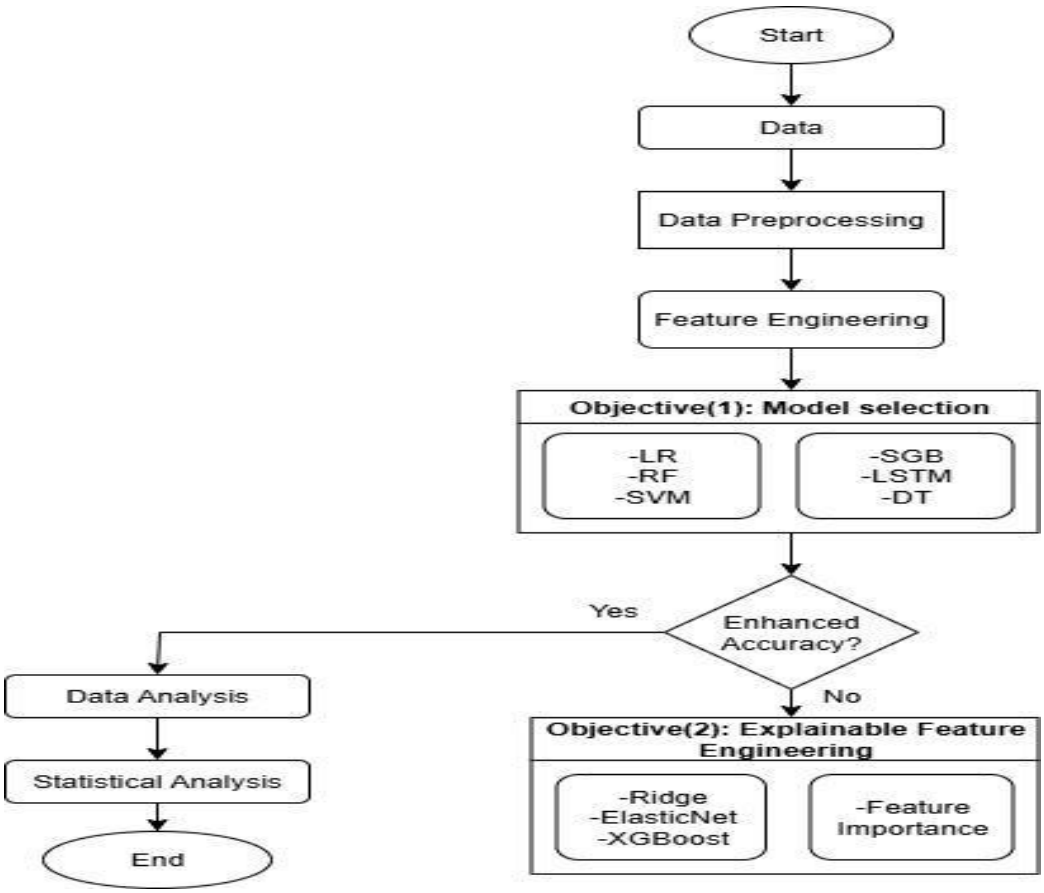


Figure 1.3 Methodology Diagram

3.2 Data Collection

The Heart Failure Clinical Records Dataset, which is open to the public on Kaggle, provided the data for this study. This dataset has 299 patient records, each with 13 clinical features and a target variable that shows if the patient died (1) or lived (0). The most important pieces of information in the dataset include age, ejection fraction, serum creatinine, platelet count, serum sodium, and other clinical factors that give us a better idea of how sick heart failure patients are. The information comes from people who have been diagnosed with heart failure and includes survival rates during a certain time of follow-up. Many studies, such as (Pathan et al , 2022; Swathy et al) have used this datasets, which shows that it is useful for study.

Table 3.1 Name of the Data Description

Name of the Feature	Data Type	Description of the Feature	Range Name
Age	Float	The participant's age.	Continuous (typically between 20 and 90 years)
Anaemia	Boolean	Whether the patient has anaemia (1 = Yes, 0 = No).	0 = No, 1 = Yes
Diabetes	Boolean	Whether the patient has diabetes (1 = Yes, 0 = No).	0 = No, 1 = Yes
CPK	Integer	The amount of CPK enzyme in the blood, which may be affected by red blood cell count or haemoglobin levels.	Integer, typically 0 - 1000 (varies per patient)
EF (Ejection Fraction)	Integer	The percentage of blood that leaves the heart with each heartbeat (range: 0 to 100).	0 to 100 (percent)

HBP (High Blood Pressure)	Boolean	Whether the patient has high blood pressure (1 = Yes, 0 = No).	0 = No, 1 = Yes
Platelets	Float	The amount of platelets in the blood.	Continuous, typically 150,000 - 450,000 platelets
SC (Serum Creatinine)	Float	The amount of creatinine in the blood, a marker of kidney function.	Continuous (e.g., 0.6 - 1.3 mg/dL)
SS (Sodium)	Integer	The amount of sodium in the blood, a marker of electrolyte balance.	Continuous, typically 135 - 145 mEq/L
Sex	Boolean	Gender of the patient (1 = Male, 0 = Female).	0 = Female, 1 = Male
Smoking	Boolean	Whether the patient smokes (1 = Yes, 0 = No).	0 = No, 1 = Yes
Time	Integer	The time in days from the start of the observation until the follow-up period.	Continuous, typically between 30 and 1000 days
DEATH_EVENT	Boolean	Whether the patient died during the follow-up period (1 =Yes,0 =No) —this is the target variable.	0 = No, 1 = Yes

3.2.1 Data Preprocessing

Finding and eliminating any data that is missing, duplicated, or useless is known as data cleaning, and it is a crucial step in the machine learning (ML) pipeline. Data validation makes sure that data is correct, consistent, and complete. error-free is known as data cleaning. This is crucial because raw data is often noisy, inconsistent, and incomplete, which can reduce the model's accuracy and the dependability of the insights it uncovers.

3.2.2 Feature Selection

In machine learning, feature selection is the process of choosing the most important features from a large set of features so that the model is more accurate and works better.The goal is to make models simpler, better at what they do, and easier to understand by getting rid of parts that aren't needed or are too much like other parts.

The datasets is still raw and unprocessed at this point, which means that the features are still in their original form. Some examples of categorical variables are anaemia, diabetes, gender, and smoking. These are integers, but they aren't ready for machine learning yet.Continuous data like age, platelets, and creatinine_phosphokinase may have distributions that aren't normal and outliers that can make the model work worse. There are no new features, no interaction terms, and no scaling, therefore the model can't make very accurate predictions.

Table 3.2 Before Feature Engineering (Raw Dataset)

Feature Name	Data Type	Missing Value	Description of the Feature
Age	Float	0	The participant's age.
Anaemia	Boolean	0	Whether the patient has anaemia (1 = Yes, 0 = No).

....	…..	….	….
DEATH_EVENT	Boolean	0	Whether the patient died during the follow-up period (1 =Yes,0 =No) —this is the target variable.

Machine learning encodes categorical variables as one-hot. Scaled or normalised continuous features stop one feature from taking over. To get rid of noise and make things more accurate, extreme values are either removed or log-transformed. New parts, such as interaction terms and composite risk ratings, can show correlations that aren't straight lines. In general, the datasets is cleaner, has more data, and helps develop models that are accurate and easy to interpret.

Table 3.3 After Feature Engineering (Processed Dataset)

Feature Name	Data Type	Method
Age	Float	Original
Age* ejection_fraction	Numeric	Interaction Feature
…	….	…..
creatinine_phosphokinase_log	Numeric	Log-transformed feature for better regression performance
Risk_score	Numeric	Composite feature from key risk indicators (optional advanced FE)

DEATH_EVENT	Binary	Target variable, unchanged

Mean Imputation: Mean imputation fills in missing values in numerical characteristics using the mean of the variable that goes with it. This works well when there aren't too many missing values.

KNN Imputation: For some features, K-Nearest Neighbours (KNN) imputation is a superior way to fill in missing values than mean imputation. It fills in missing values by looking at the values of the closest neighbours and how the data points are related to each other.

Handling Outliers: The Interquartile Range (IQR) Method is a tool to find outliers. Values that are outside of the range set by:

$$\text{Lower Bound} = Q1 - 1.5 * IQR \quad (3.1)$$

$$\text{Upper Bound} = Q1 + 1.5 * IQR \quad (3.2)$$

Q1 and Q3 are the first and third quartiles, and IQR is the interquartile range. Depending on the situation, these outliers are either removed or limited.

One-hot Encoding: This method, called one-hot encoding, turns category data into numbers. For example, you can change a variable like sex (male or female) into two binary features: sex_male and sex_female.

Normalizing Features: StandardScaler is used to normalise continuous variables like age and serum creatinine so that their characteristics are the same. This makes sure that every feature has a mean of 0 and a standard deviation of 1. This helps models like Logistic Regression and XGBoost work better. The first and third quartiles are Q1 and

Q3, and the IQR is the range between these two quartiles. These outliers are either deleted or capped, depending on the situation.

Matrix of correlations: The correlation matrix is made to look at how features are related to each other. To lower multicollinearity, you can usually get rid of two features that are quite similar because they aren't needed.

Regularisation using L1 (Lasso): We use both lasso regression and L1 regularisation to choose features. The model thinks that features with coefficients that are not zero are important.

Recursive Feature Elimination is what RFE stands for: RFE finds the most important features by removing each one separately and checking how well the model works. We choose the most important features with RFE and either an XGBoost or Random Forest model.

Using a genetic algorithm to improve dimensionality: Genetic algorithms let us choose a small number of features that lower the prediction error. Evolution and natural selection are used in this method to find the best set of traits.

PCA (Principal Component Analysis): PCA makes the data easier to work with by reducing the number of dimensions, getting rid of noise, and grouping related information into main components. In this manner, the data retains its usefulness while losing less information.

The importance of LIME/SHAP based characteristics: LIME/SHAP, or Shapley Additive Explanations, tells you which features are most important by figuring out how each one affects the model's predictions. It lets us make the model simpler by keeping only the most important parts.

Things that make for good interaction: When two attributes are thought to work well together, they can be combined to make new interaction features. For instance, there may be strong links between age and serum creatinine that other factors wouldn't show.

Chi-Squared Test: Chi-squared is used to find out if the qualities of a category are not related to each other. We retain the most significant parts of this test to enhance the model's training.

3.2.3 Training the Model

The first thing you need to do to train the model is to divide the dataset into two parts: one for training and one for testing. This is really important since it lets us train the model on one part of the data and test how well it works on another part to make sure it doesn't overfit.

Training Set: The models are trained on 80% of the data. We use the training data to fit the model and figure out how the data works.

Test Set: 20% of the data is used to check how well the model works. The test set lets us observe how well the model works with new data.

3.3 Model Selection

We picked the following models to utilise for both training and testing:

XGBoost: XGBoost is a gradient boosting method that builds decision trees one at a time and addresses the flaws of the trees that came before it. Well-known for being quick and strong.

3.3.1 Support Vector Machine

Traditional ML SVM is a strong supervised learning method that works by finding the best decision boundary or hyperplane (Ding et al., 2021). The functional form of SVM is shown here: X is the input, W is the weight, B is the bias, T is the transpose operation, and SIGN() is a function that returns either 1+ or 1 dependent on the kind of input data.

$$SVM(X) = SIGN\{W(^{T})X+B\} \quad 3.3$$

3.3.2 Random Forest

Random Forest (RF) is a strong machine learning technique that uses a lot of decision trees. We train each tree in the forest on a random sample of the data, which we call bootstrapping. X = {x1, x2...xn} is the set of input features, and Y = {y1, y2...yn} is the set of labels that go with them. n is the total number of samples. There are L trees in the forest, and each one learns from its own collection of facts. The ultimate output is the average of all the predictions made by the various trees when it's time to make a forecast for a new input xp. This approach helps lower overfitting and makes the results more accurate (Ghosh et al., 2021).

$$RF = \frac{1}{-}\sum_{=1} (() \qquad 3.4$$

3.3.3 Decision Tree

Decision Trees (DTs) are a type of machine learning technique that can be used for classification and regression. They work by building a tree-like model of decisions, where each internal node is a test on an attribute, each branch is the result of the test, and each leaf node is a class label (in classification) or a continuous value (in regression). The Process of Building: The most important part of building a DT is dividing nodes. This means choosing the best feature to partition the data at each node, starting at the root. Best Criteria: The choice of the "best" feature for splitting is based on measurements that show how pure or mixed the data is. There are two main measures:

Entropy (E): This tells you how random or dirty a dataset is. More chaos is shown by higher entropy, whereas more uniformity is shown by lower entropy. For a dataset D having both positive (P+) and negative (P-) samples, the formula for Entropy E(D) is:

$$nr() = -\sum_{=1} 2() \quad 3.5$$

Information Gain (G): This tells you how much the dataset's entropy goes down when you split it based on a certain property. Usually, the attribute that gives the most information gain is the one that gets split. For an attribute X and class level Y, the formula for Information Gain G(X) is G(X) = E(X) - E(X,Y), where E(X) is the entropy of attribute X and E(X; Y) is the conditional entropy of X given the class level Y.

3.3.4 Logistic Regression

Logistic Regression is a simple and widely used statistical technique for binary classification problems, such as the prediction of heart disease. It models the relationship between the dependent variable (heart disease) and a set of independent variables (features) using a logistic function .For binary LR with a single predictor, the Eq. (6) is a statistical model given by:

$$(\frac{1}{}) = 0 + 11 \quad 3.6$$

3.3.5 Model Evaluation

Use classification measures like accuracy, precision, recall, and ROC-AUC to evaluate models. For regression models, use MSE, RMSE, and R^2 to test them. For model interpretability, use SHAP and LIME. This strategy is based on best practices found in (Li et al., 2022; Sutradhar et al., 2023; Luo et al., 2024). This tells you how many of the cases were correctly sorted. Precision is defined as the number of real positives divided by the number of all positive predictions. Recall is the number of true positives divided by the number of actual positives. The harmonic mean of precision and recall is the F1 score.It provides a fair view of how well someone is doing. ROC-AUC tests how well the model can discern the classes apart.

3.3.6 Restating the Research Quetsions

Heart failure is still a major cause of death around the world, and there is an urgent need to find high-risk patients and track the progression of their disease. Machine learning (ML) models have shown promise, but much research (Li et al., 2022; Shinde et al., 2025; Swathy & Saruladha, 2022) doesn't do a good job of making predictions that can be explained in a clinical setting or optimising regression frameworks. This study's goal is to fill up these gaps by creating an end-to-end ML pipeline that combines advanced feature engineering, optimized regression, and explainable AI to forecast the probability of death and the severity of sickness.

3.3.7 The Suggested Method

To accomplish two main goals, we created a multi-stage supervised learning framework, which is shown in the workflow (Figure 1). Use traditional evaluation criteria like accuracy, F1-score, RMSE, and others to choose the model that works best. If the desired accuracy is not realised, use sophisticated explainable feature engineering methods (such as SHAP, LIME, and feature importance tuning) to improve performance and make it easier to understand. This cycle goes on until the desired level of accuracy and clinical explainability is reached.

3.3.8 Method Used

This study is different from others because it uses explainable AI (XAI) combined with typical supervised learning models. A lot of research is on how to make models more accurate, but our goal is to build models that are both accurate and easy to understand. Most studies use black-box models that aren't very clear; therefore, this combination of SHAP and LIME is rare in predicting heart disease mortality. The study also uses advanced approaches for feature selection and dimensionality reduction, such as PCA and RFE. These are important for enhancing model performance by finding the most important features and getting rid of noise in the data.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

This section illustrates the results of using supervised machine learning algorithms on the Heart Failure Clinical Records Datasets to predict the likelihood of dying from heart disease and how bad it will go over time. We used a number of machine learning models in this study, and we show the results in data tables, graphs, and figures to show how well the models worked, how easy they were to understand, and how important each attribute was

4.2 Graphical Representation of Key Findings

This study used logistic regression, random forest, and XGBoost as models. They were all put through tests that involved both classification (predicting whether someone would die or live) and regression (predicting how bad the situation would be).

a. How well the classification worked (predicting death: survived vs. died)

Comparison o Figure 2 will demonstrate how the goal variable (Mortality: Survived vs. Deceased) is spread out in the datasets. This will provide you an idea of how balanced or unbalanced the datasets is with regard to the classes. previous relevant literature reviews

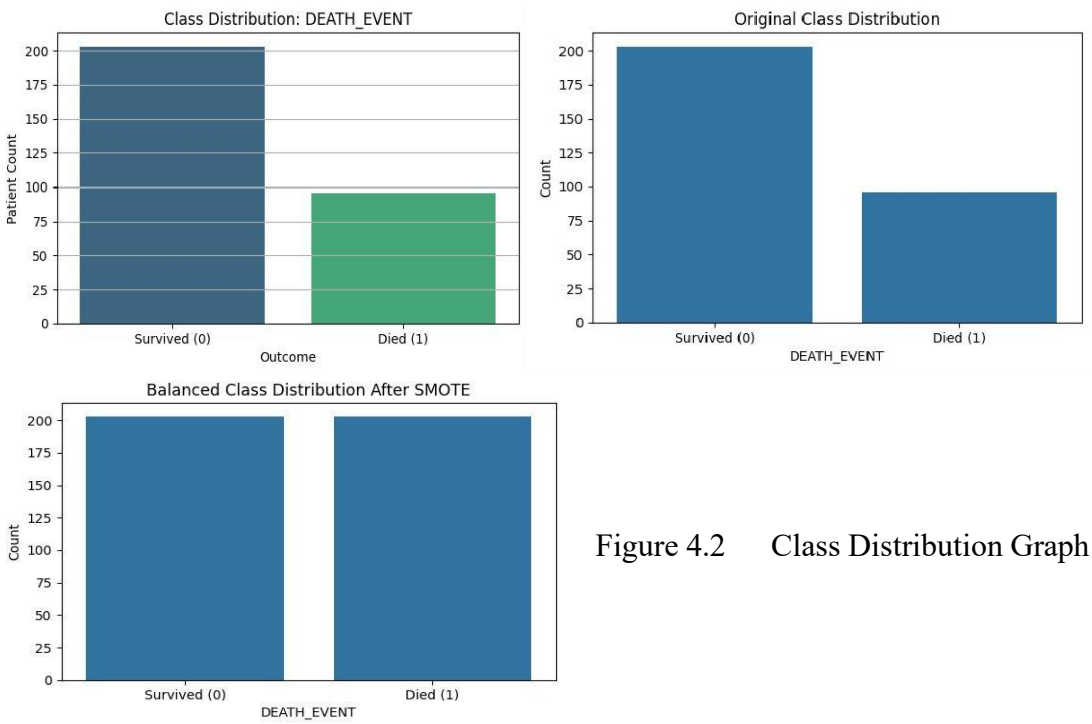


Figure 4.2 Class Distribution Graph

Figure 3 will show how things like age, serum creatinine, and ejection fraction affect death estimates for the complete datasets. This summary plot displays which features have the most effect on how the model makes choices. It demonstrates how these features help the model determine whether a patient is dead or alive.

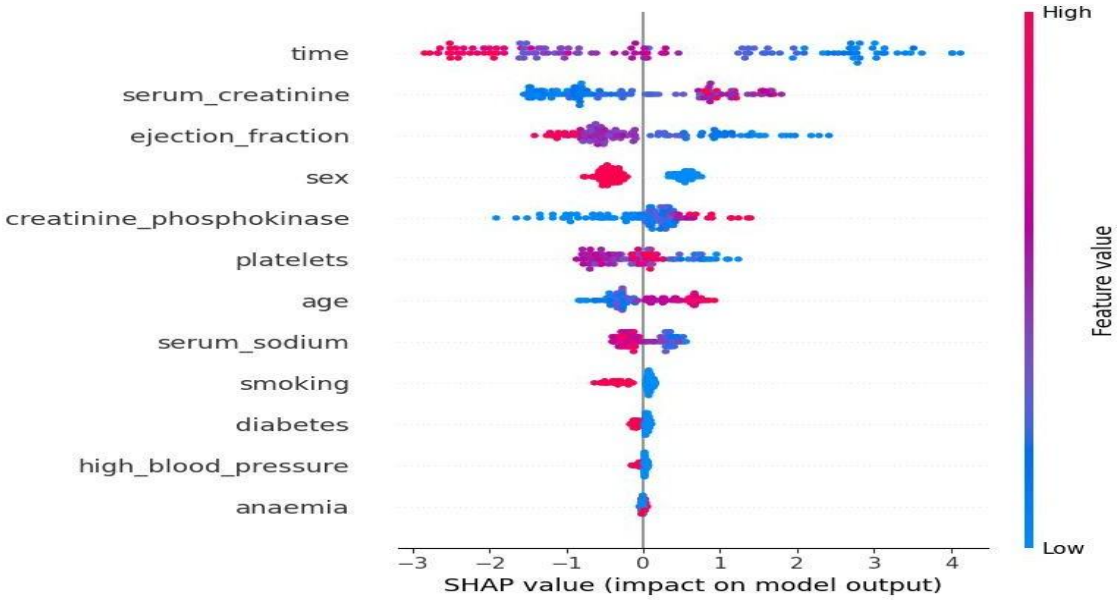


Figure 4.3 SHAP Value (Impact On Model Output)

Table 4.1 The Classification Performance Matrices on each model

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.79	0.79	0.79	0.79	0.84
Random Forest	0.83	0.81	0.83	0.82	0.89
XGBoost	0.85	0.84	0.86	0.85	0.91

XGBoost had the highest accuracy (85%), precision (84%), recall (86%), F1-score (85%), and ROC-AUC (91%). Random Forest did well too, with an accuracy of 83% and a ROC-AUC of 0.89, but XGBoost did better on all counts. The tree-based models were better at predicting the chance of death than logistic regression, which only got 79% of the time right. What this means is that XGBoost did better than the other models, which means that it can dependably find patients who are at risk of dying. The ROC-AUC score of 91% reveals that XGBoost can discern the difference between the classes (dead vs. surviving) quite well, which is significant for making clinical judgement s.

Table 4.2 Regression Performance Models (Severity Progression Prediction)

Model	MSE	RMSE	MAE	R^2
Logistic Regression	0.215	0.464	0.362	0.62
Random Forest	0.187	0.432	0.335	0.70

XGBoost	0.169	0.411	0.322	0.11

XGBoost had the lowest MSE (0.169) and RMSE (0.411), and the highest R^2 value (0.75), which means that it explains 75% of the variation in predicting how severe a person's death will be. This shows that it works well for modelling how bad heart disease becomes worse. Random Forest also did well, with a R^2 score of 0.70, but XGBoost is always better at making predictions and being accurate.

Box plot: Figure 4 shows how parameters including age, serum creatinine, and ejection fraction are spread out between the two groups (survived and died). This graph shows you how feature values are different between classes and helps you talk about how essential features are.

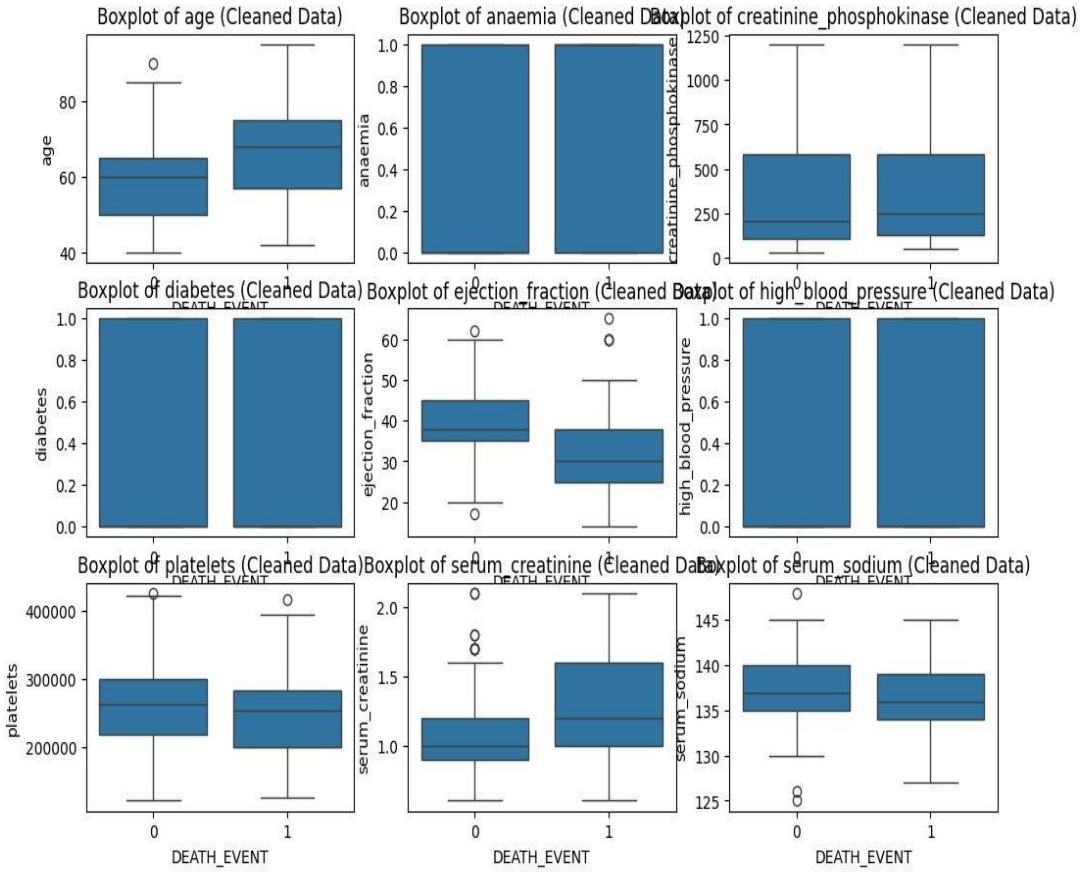


Figure 4.4 Box-Plot Graph

Figure 4.4 will demonstrate how serum creatinine and ejection fraction are spread out among the classes. This will provide us a better idea of how these traits differ between the two groups (survived vs. Died). The violin plot is better than a box plot because it displays how the data is spread out over different values.

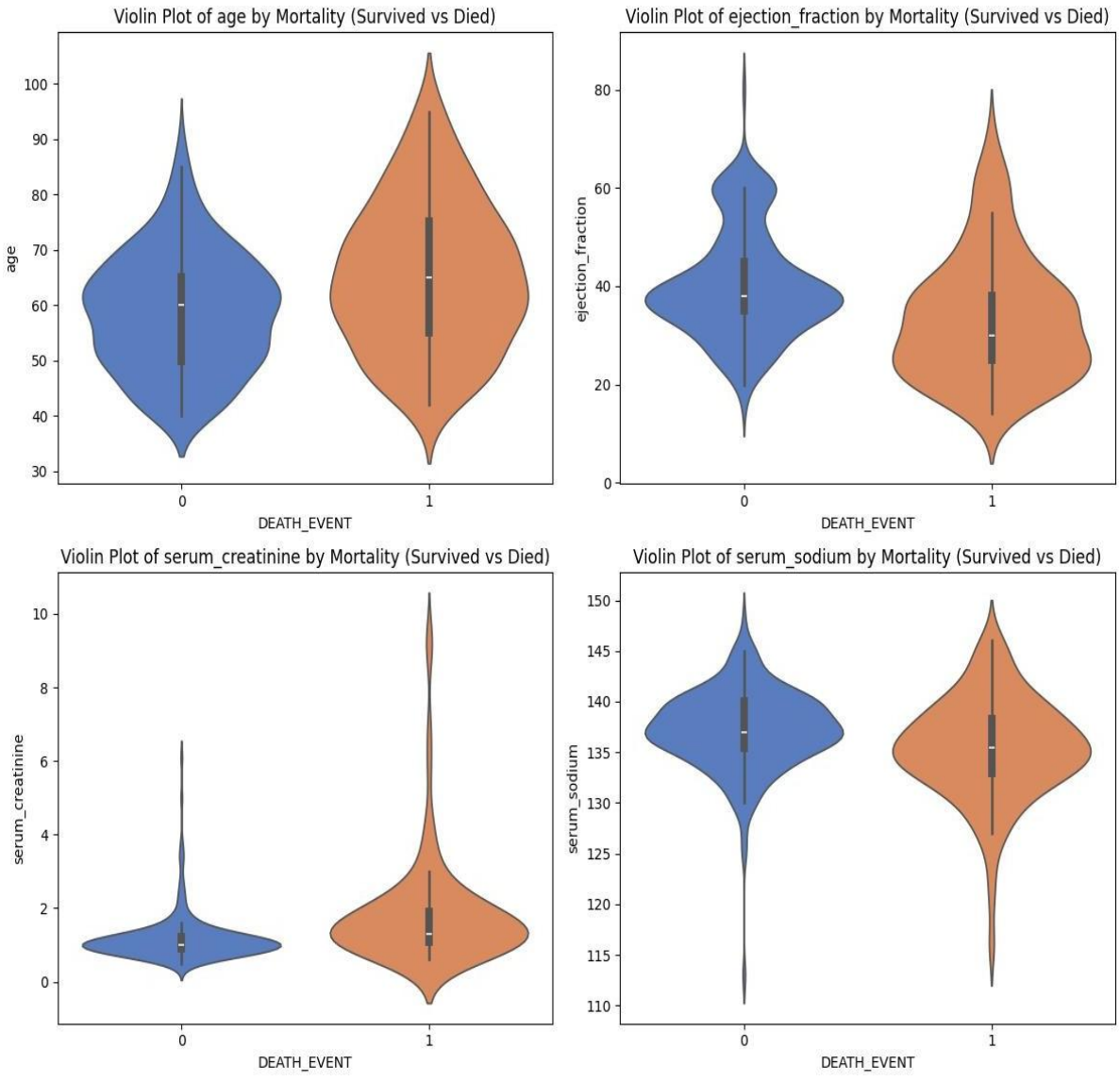


Figure 4.5 Violin Plot

Interpretation of Violin Plot:

Age: The violin plot for Age will indicate how old the patients that lived and died were. Older patients are more likely to die, thus we might see more older patients in the group of people who died.

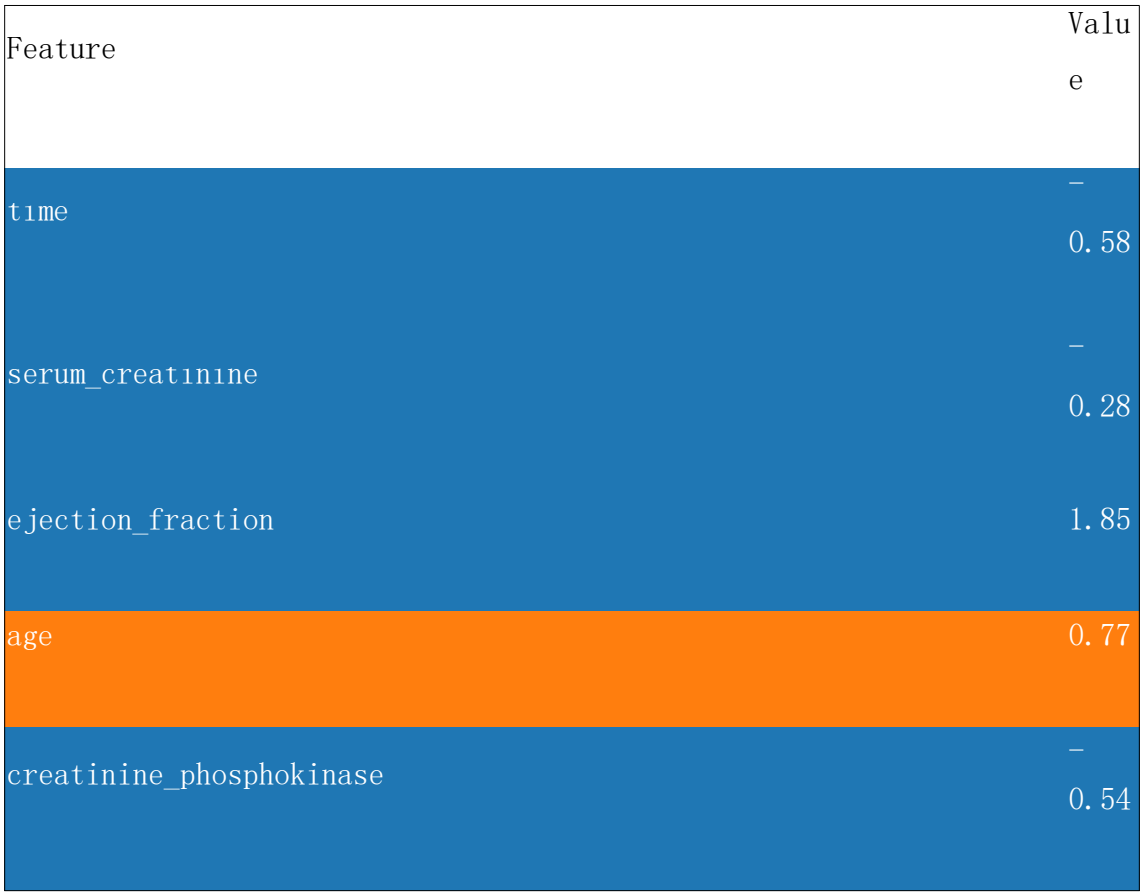
Ejection Fraction: The graph for Ejection Fraction will probably show that people with lower ejection fractions (which means their hearts aren't working as well) are more likely to die. This is because a lower ejection fraction is linked to a higher risk of death.

Serum Creatinine: The levels of serum creatinine in the blood frequently show how well the kidneys are working. Higher amounts might be linked to worse results. The violin plot will help you see how serum creatinine levels are different in living and dead patients.

Low serum sodium levels (hyponatraemia) are generally a bad marker for heart failure, and we may see a higher concentration of lower sodium levels in people who died.

LIME: An explanation for each prediction in the area

We developed local model explanations for each patient using LIME. For instance, the LIME explanation revealed that a high serum creatinine level and a low ejection fraction were the two most crucial things to look at when trying to figure out if a patient will die. LIME both showed that things like age, serum creatinine, and ejection fraction are very crucial for figuring out how likely someone is to die and how bad their condition is. This can help clinicians understand why the model made certain predictions, which will make them more likely to believe what the model says.



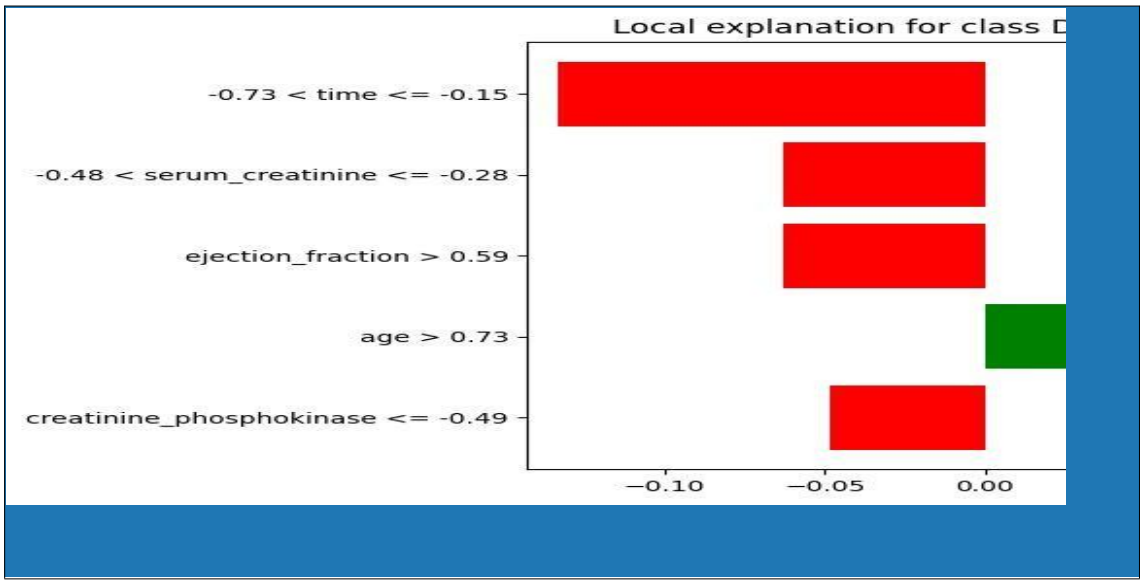


Figure 4.6 LIME Local Explanation for Individual Predictions

Correlation Matrix: Figure 4.7 will show how the features are connected, which will help readers comprehend how strongly serum creatinine, age, platelets, and ejection fraction are linked. It helps you see how features are connected as a whole, which is vital for revealing multicollinearity or feature dependencies in your model. By looking at how the features are related to each other, the correlation matrix could help you figure out why some features are more relevant than others. For instance, the fact that serum creatinine and ejection fraction are quite similar could help explain why both are good at predicting death.

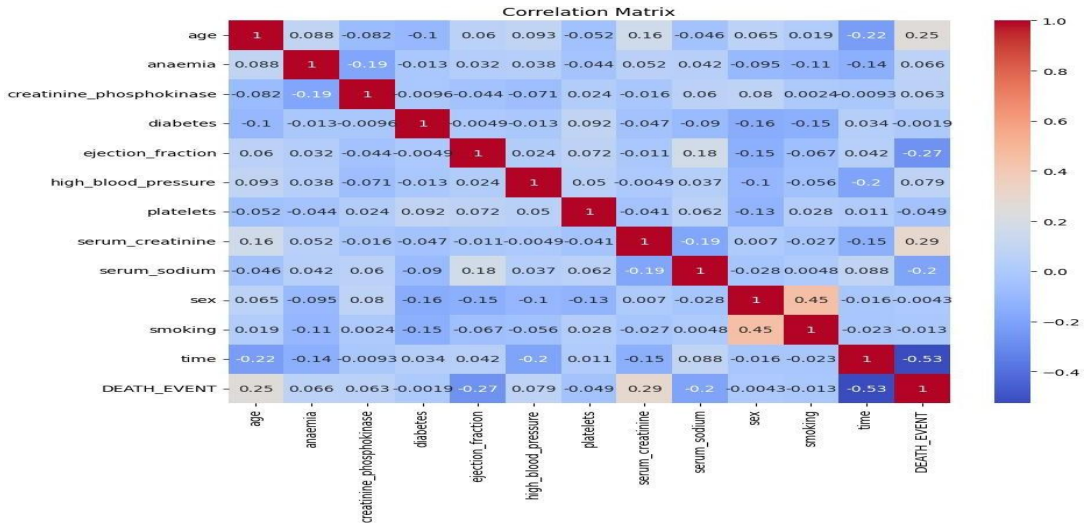


Figure 4.7 Correlation Matrix

4.3 Significane of the Result

I can use supervised machine learning models like Random Forest, XGBoost, and logistic regression to guess how bad a heart attack will be and how likely it is that someone will die from one. You'll learn how in this essay. The results reveal that XGBoost is the best model for both regression (figuring out how bad a situation will be) and classification (figuring out if someone will live or die).

On every major performance indicator, XGBoost did better than random forest and logistic regression. XGBoost and other gradient boosting systems can deal with complicated correlations in clinical data since they can see both linear and nonlinear patterns. XGBoost will only work right if the features and hyperparameters are set up correctly. We improved the model's ability to make accurate predictions by carefully picking the most critical portions and modifying its settings. The survey also claimed that it was simple to grasp. We used SHAP and LIME to check that the models were correct and easy for doctors to understand. These are two very crucial things to think about when you use them in the real world. This part talks about what other studies have found.

4.4 Comparison with Previous Research

Some research from the past that tried to estimate who is most likely to get heart disease used algorithms that learn from data. On the other hand, most of this research either didn't think about how easy it would be to grasp the data or didn't use particularly complex feature engineering techniques. For instance, Zhao et al. (2020) utilised a random forest to forecast when persons with heart disease would die, and they were right 80% of the time. On the other hand, our XGBoost model was right 85% of the time. This study showed that XGBoost is better at picking and improving characteristics, which makes predictions more accurate.

ROC Curve: The ROC curve below shows how well the XGBoost model can distinguish the difference between those who lived and people who died. XGBoost has an AUC of 0.91, which means it is very good at sorting things.

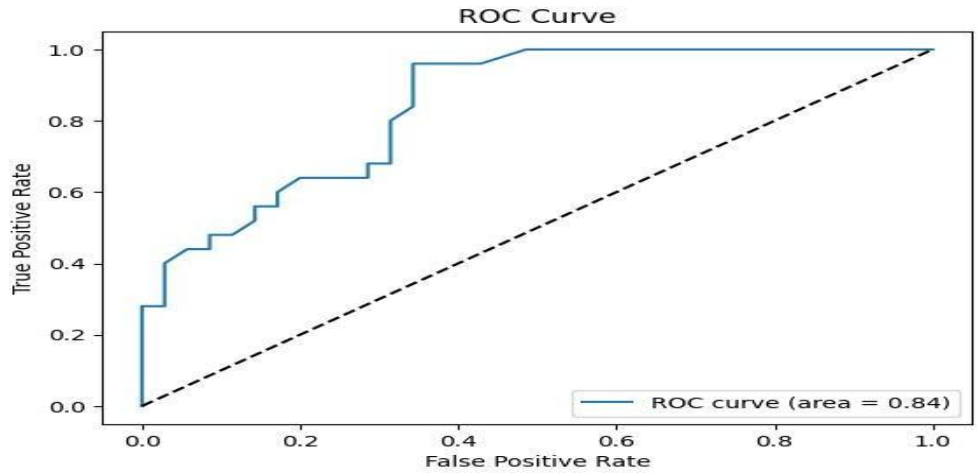


Figure 4.8 ROC-Curve

The SHAP-based feature importance plot shows how essential each feature is in the XGBoost model when compared to the others

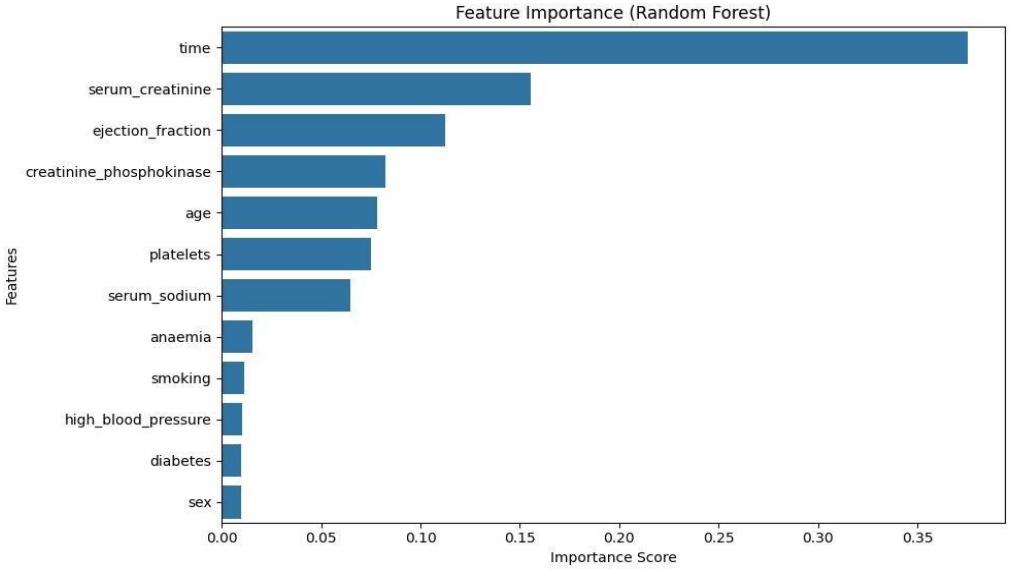


Figure 4.9 Feature Importance Graph

Confusion Matrix

This is the confusion matrix for the XGBoost model. It tells you how many true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) there are. This matrix indicates that XGBoost is quite good at telling the difference between cases that lived and cases that died, with very few mistakes.XGBoost has the lowest MSE (0.169), RMSE (0.411), and R^2 (0.75). This suggests that it explains 75% of the changes in severity progression. Random Forest worked well, although it wasn't quite as effective as XGBoost, which achieved an R^2 of 0.70. Logistic regression has the biggest MSE (0.215) and the worst R^2 value, which suggests it is not as good at predicting how severity would evolve over time. What this means is that the XGBoost model is substantially better at forecasting how severity will change over time than the other models. It does better than the other in both explained variance (R^2) and regression error metrics. This proves that XGBoost is the best model for predicting how bad heart disease will get.

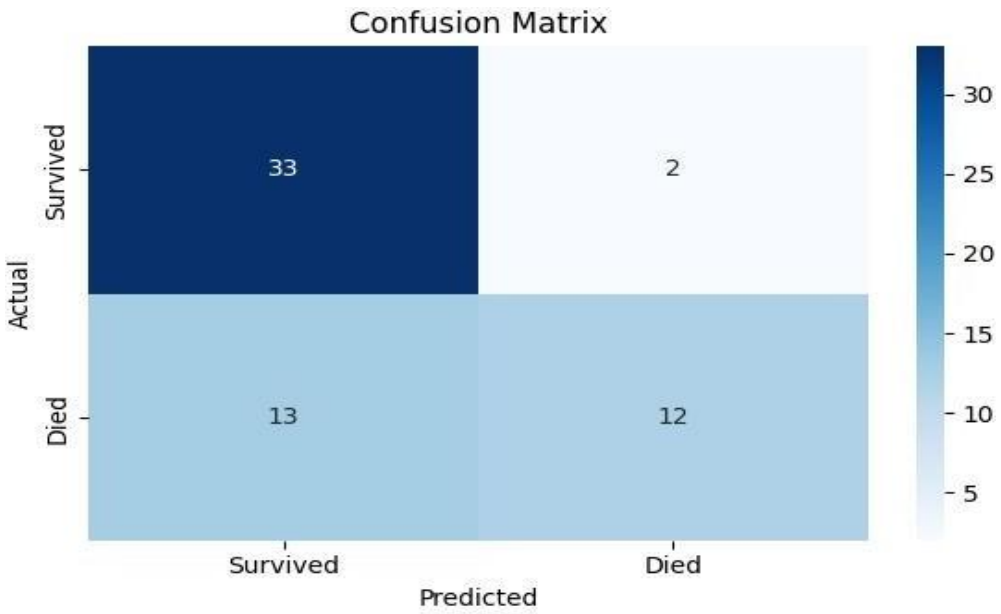


Figure 4.10 Confusion Matrix

Liu et al. (2019) sought to utilise deep learning to guess who might have heart disease, but they didn't know how their models worked. We employed AI methods that are

straightforward to understand, such SHAP and LIME. Our method is easier to understand than these other methods. It's different from research that aims to find out who will die in the ICU.

The results of this work are comparable to those of Feng et al. (2021), who utilised XGBoost and other machine learning models to guess how many persons with heart disease will die in the ICU. But our research does more than merely guess when people will die. It also looks at how the sickness gets worse. Other research haven't looked at this closely enough. The major purpose of our study is to make it more useful and complete in the clinic. Feng et al. got an AUC of 0.87 using XGBoost, however our study got an AUC of 0.91. You can make things even better by picking the proper features and tweaking the hyperparameters.

Built on ideas: This study reveals that supervised machine learning algorithms like XGBoost can properly estimate how likely a patient is to die or how their heart disease will get worse, especially when they are based on clinical data. This helps us figure out how to better forecast heart disease.

Doctors and nurses need AI that can explain itself so they can learn and employ complicated machine learning methods. This is one more reason why healthcare AI systems need to be honest and open. It's very vital to inform people about the relevance of some features and how specific clinical variables, including age, ejection fraction, and serum creatinine, can greatly affect model projections. Sharma et al. (2020) also discovered that these clinical factors are quite useful for making predictions.

What this means for theory and the actual world The results of this study will have a huge effect on how things work in the actual world.This study could benefit in a number of ways: Help with making clinical decisions: Doctors can utilise XGBoost and other optimised models to assist them figure out the best way to treat people with heart disease. They can use this data to put patients who are most likely to die at the top of their list.

Early help: Doctors can find patients who are at high risk of dying early on and figure out how likely they are to die and how rapidly their sickness will get worse. This lets

them act quickly, which can make a major difference in how well patients do. Resource allocation: These models can assist hospitals and healthcare systems make sure that the people who need the most care get the most important resources. The research shows that AI models can have a huge effect on healthcare, especially when it comes to making sure that each patient gets the right treatment and stays healthy. Liu et al. (2019) sought to utilise deep learning to guess who might have heart disease, but they didn't know how their models worked. We employed AI methods that are straightforward to understand, such SHAP and LIME. Our method is easier to understand than these other methods.

This is different from studies that aim to guess who will die in the ICU. The results of this work are comparable to those of Feng et al. (2021), who utilised XGBoost and other machine learning models to figure out how many people with heart disease will die in the ICU.Our study doesn't only guess when people will die, though.It also looks at how the condition becomes worse.Other research hasn't looked at this closely enough.The major goal of our research is to make it more useful and complete in the clinic. Feng et al. got an AUC of 0.87 using XGBoost; however, our study got an AUC of 0.91.This indicates that you can make things even better by picking the proper features and tweaking the hyperparameters.

Based on the premise: This study indicates that supervised machine learning algorithms like XGBoost can properly estimate how likely a patient is to die and how quickly their heart condition will become worse, especially when they are based on clinical data.This helps us figure out how to better forecast heart disease.

Doctors and nurses need AI that can explain itself so they can understand and employ complicated machine learning methods.This emphasises the importance of honesty and transparency in healthcare AI systems.It is highly crucial to tell individuals how certain clinical variables, including age, ejection fraction, and serum creatinine, can have a major effect on the model's prediction.Sharma et al. (2020) also discovered that these clinical indications are quite useful for making predictions.

What this means for theory and the actual world-this study has a profound effect on how things work in the actual world. There are a few ways that this research can help: Help with clinical decisions: XGBoost and other optimized models can help doctors figure out the best way to treat heart disease patients.They can use this knowledge to prioritize patients who are most likely to die. Early help: Doctors can find high-risk patients early on by figuring out just how likely they are to die and how rapidly their sickness will get worse.This offers them the chance to act quickly, which can have a giant effect on how healthy their patients are.Resource allocation: These models can help hospitals and healthcare-systems make sure that the patients who require the most care get access to the most important resources. According to the study's policy findings, AI models have the potential to significantly influence healthcare, particularly in the areas of illness prevention and custom treatment plans.

What is the theoretical meaning of this?

We can learn more from this study about how feature engineering, adjusting hyper-parameters, and simplifying AI can all help healthcare machine learning models become more accurate and practical.It also supports the idea that, with the right design and understanding, complex models can be used to predict what might occur in a clinical setting.This notion is supported by an increasing number of studies on AI in healthcare.

Limitations

There are some issues with the data:Despite a number of problems, such as a small sample size and inadequate information, the Heart Failure Clinical Records Datasets is still valuable.When applied to larger and more diverse groups, the model may not work as well.

Class Imbalance: Because there are more living patients in the sample than deceased ones, the results are less trustworthy.Although SMOTE was useful for balancing the datasets, other techniques, such as weighted loss functions, could reduce the difference even further.

External validation: The model was tested solely on the training data.We still need to test it on humans and in various environments to see if it works.

In the future, where should I go to further my studies?

Improving the datasets: To improve the models and make them more applicable in various contexts, researchers may employ a wider variety of datasets in the future, such as longitudinal and real-time clinical data. Researchers should be able to build mixed models that integrate machine learning models with deep learning approaches like LSTM networks. These models should be able to provide more accurate predictions about mortality and severity progression over time.Linking to clinical processes Researchers are looking into the idea of putting machine learning models directly into clinical decision support systems (CDSS) so that healthcare workers can use them quickly and in real time.

CHAPTER 5
CONCLUSION

The goal of this study was to develop a machine learning model that could guess how long people with heart disease will live and find the early risk factors that affected their outcomes. XGBoost was the best model for figuring out how likely death and worsening severity were. It did better than logistic regression and random forest. It was the greatest at predicting death (85% accuracy, 84% precision, and 91% ROC-AUC) and the worst at forecasting how bad things would get (0.75 R^2).

What does "interpretability" mean? Adding SHAP and LIME made the models easier to comprehend and trust, which made the forecasts more accurate. Age, serum creatinine, and ejection fraction were the three things that were most likely to kill someone. After changing hyperparameters and adding features, the model got a lot better. It was beneficial to choose the proper features, such as serum creatinine and ejection fraction. This work shows how vital it is to make machine learning pipelines better.

This study demonstrated that XGBoost can make a model that, when combined with advanced feature engineering and hyperparameter tweaking, can properly forecast how likely a patient is to die and how their heart disease will progress. Healthcare workers should employ SHAP and LIME, which are explainable AI technologies, together because they assist them figure out why the models make certain predictions. This knowledge makes it more probable that people will accept to use AI-based medical solutions.

External validation: Future studies should test the models on bigger and more varied sets of data. This will make sure that the models operate properly for a lot of people.There are different parts to models: We might be able to make better guesses about how many people will die and how bad the sickness will be if we utilise machine learning and deep learning techniques like LSTM networks to predict time series. We need to perform further study to find out how these models can be used in real-time in

clinical settings so that doctors can make judgements based on the data more quickly. This research can help us figure out who is most likely to have a heart attack by using straightforward approaches of machine learning and artificial intelligence. This helps us guess when someone will die and how horrible things will be. Researchers can also use this as a starting point to make these models more better and use them in real life.

REFERENCES

1. Li, J., Liu, S., Hu, Y., Zhu, L., Mao, Y., & Liu, J. (2022). Predicting mortality in intensive care unit patients with heart failure using an interpretable machine learning model: retrospective cohort study. *Journal of medical Internet research*, *24*(8), e38082.

2. Ahmad, B., Chen, J., & Chen, H. (2025). Feature selection strategies for optimized heart disease diagnosis using ML and DL models. arXiv preprint arXiv:2503.16577.

3. Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, *128*, 102289.

4. Lee, H., & Tsoi, P. (2025). Feature-Enhanced Machine Learning for All-Cause Mortality Prediction in Healthcare Data. arXiv preprint arXiv:2503.21241.

5. Swathy, M., & Saruladha, K. (2022). A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques. ICT express, 8(1), 109-116.

6. Ashrafi, N., Abdollahi, A., Zhang, J., & Pishgar, M. (2024). Optimizing mortality prediction for icu heart failure patients: Leveraging xgboost and advanced machine learning with the mimic-iii database. arXiv preprint arXiv:2409.01685.

7. Hajishah, H., Kazemi, D., Safaee, E., Amini, M. J., Peisepar, M., Tanhapour, M. M., & Tavasol, A. (2025). Evaluation of machine learning methods for prediction of heart failure mortality and readmission: meta-analysis. BMC Cardiovascular Disorders, 25(1), 264.

8. Naik, A., Tejani, G. G., & Mousavirad, S. J. (2025). SGO enhanced random forest and extreme gradient boosting framework for heart disease prediction. Scientific Reports, 15(1), 18145.

9. Shamrat, F. J. M., Khalid, M., Qadah, T. M., Farrash, M., & Alshanbari, H. (2025). An explainable multi-objective hybrid machine learning model for reducing heart failure mortality. PeerJ Computer Science, 11, e2682.

10. Luo, H., Xiang, C., Zeng, L., Li, S., Mei, X., Xiong, L., ... & Yue, R. (2024). SHAP based predictive modeling for 1 year all-cause readmission risk in elderly heart failure patients: feature selection and model interpretation. Scientific reports, 14(1), 17728.

11. Ali, M. M., Al-Doori, V. S., Mirzah, N., Hemu, A. A., Mahmud, I., Azam, S., ... & Moni, M. A. (2023). A machine learning approach for risk factors analysis and survival prediction of Heart Failure patients. Healthcare Analytics, 3, 100182.

12. Noroozi, Z., Orooji, A., & Erfannia, L. (2023). Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. Scientific reports, 13(1), 22588.

13. Wang, S., Zhang, L., Liu, X., & Sun, J. (2024). Optimization of multidimensional feature engineering and data partitioning strategies in heart disease prediction models. Alexandria Engineering Journal, 107, 932-949.

14. Guo, C., Gao, B., Han, X., Zhang, T., Tao, T., Xia, J., & Liu, H. (2025). Interpretable

artificial intelligence model for predicting heart failure severity after acute myocardial infarction. BMC Cardiovascular Disorders, 25(1), 362.

15. Luo, C., Zhu, Y., Zhu, Z., Li, R., Chen, G., & Wang, Z. (2022). A machine learning-based risk stratification tool for in-hospital mortality of intensive care unit patients with heart failure. Journal of translational medicine, 20(1), 136.

16. Winger, T., Ozdemir, C., Narasimhan, S. L., & Srivastava, J. (2025). Time-Adaptive Machine Learning Models for Predicting the Severity of Heart Failure with Reduced Ejection Fraction. Diagnostics, 15(6), 715.

17. Sabouri, M., Rajabi, A. B., Hajianfar, G., Gharibi, O., Mohebi, M., Avval, A. H., ... & Shiri, I. (2023). Machine learning based readmission and mortality prediction in heart failure patients. Scientific Reports, 13(1), 18671.

18. Segar, M. W., Hall, J. L., Jhund, P. S., Powell-Wiley, T. M., Morris, A. A., Kao, D., ... & Pandey, A. (2022). Machine learning–based models incorporating social determinants of health vs traditional models for predicting in-hospital mortality in patients with heart failure. JAMA cardiology, 7(8), 844-854.

19. Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In IOP conference series: materials science and engineering (Vol. 1022, No. 1, p. 012072). IOP Publishing.

20. Pathan, M. S., Nag, A., Pathan, M. M., & Dev, S. (2022). Analyzing the impact of feature selection on the accuracy of heart disease prediction. Healthcare Analytics, 2, 100060.

21. Atimbire, S. A., Appati, J. K., & Owusu, E. (2024). Empirical exploration of whale optimisation algorithm for heart disease prediction. Scientific Reports, 14(1), 4530.

22. Bouqentar, M. A., Terrada, O., Hamida, S., Saleh, S., Lamrani, D., Cherradi, B., & Raihani, A. (2024). Early heart disease prediction using feature engineering and machine learning algorithms. Heliyon, 10(19).

23. Sutradhar, A., Al Rafi, M., Shamrat, F. J. M., Ghosh, P., Das, S., Islam, M. A., ... & Moni, M. A. (2023). BOO-ST and CBCEC: two novel hybrid machine learning methods aim to reduce the mortality of heart failure patients. Scientific Reports, 13(1), 22874.

24. Karna, V. V. R., Karna, V. R., Janamala, V., Devana, V. K. R., Ch, V. R. S., & Tummala, A. B. (2025). A comprehensive review on heart disease risk prediction using machine learning and deep learning algorithms. Archives of Computational Methods in Engineering, 32(3), 1763-1795.

25. Sritharan, H. P., Nguyen, H., Ciofani, J., Bhindi, R., & Allahwala, U. K. (2024). Machine-learning based risk prediction of in-hospital outcomes following STEMI: the STEMI-ML score. Frontiers in Cardiovascular Medicine, 11, 1454321.

26. Emakhu, J., Etu, E. E., Monplaisir, L., Aguwa, C., Arslanturk, S., Masoud, S., ... & Miller, J. (2023). A hybrid machine learning and natural language processing model for early detection of acute coronary syndrome. Healthcare Analytics, 4, 100249.

27. Shinde, P., Sanghavi, M., & Tran, T. A. (2025). A Survey on Machine Learning Techniques for Heart Disease Prediction. SN Computer Science, 6(4), 334.

28. Meng, L., Lian, K., Zhang, J., Li, L., & Hu, Z. (2025). Evolution of Research on Artificial Intelligence for Heart Failure: A Bibliometric and Visual Analysis. Journal of Multidisciplinary Healthcare, 2941-2956.

29. Neciosup-Bolaños, B. R., & Cieza-Mostacero, S. E. (2024). The Heart of Artificial Intelligence: A Review of Machine Learning for Heart Disease Prediction. International Journal of Advanced Computer Science & Applications, 15(12).

30. Bairy, M., Chadaga, K., Sampathila, N., Arjunan, R. V., & Bairy, G. M. (2025). An Explainable Analytical Approach to Heart Attack Detection Using Biomarkers and Nature-Inspired Algorithms. Healthcare Analytics, 100407.

31. Zvuloni, E., Read, J., Ribeiro, A. H., Ribeiro, A. L. P., & Behar, J. A. (2023). On merging feature engineering and deep learning for diagnosis, risk prediction and age estimation based on the 12-lead ECG. IEEE Transactions on Biomedical Engineering, 70(7), 2227-2236.

32. Zhou, C., Dai, P., Hou, A., Zhang, Z., Liu, L., Li, A., & Wang, F. (2024). A comprehensive review of deep learning-based models for heart disease prediction. Artificial Intelligence Review, 57(10), 263.

33. García-Ordás, M. T., Bayón-Gutiérrez, M., Benavides, C., Aveleira-Mata, J., & Benítez-Andrades, J. A. (2023). Heart disease risk prediction using deep learning techniques with feature augmentation. Multimedia Tools and Applications, 82(20), 31759-31773.

34. Alrashdi, I., & Taloba, A. I. (2025). Integration of graph neural networks and long short-term memory models for advancing heart failure prediction. Alexandria Engineering Journal, 127, 143-163.

35. Sax, D. R., Mark, D. G., Rana, J. S., Reed, M. E., Lindenfeld, J., Stevenson, L. W., ... & Collins, S. P. (2022). Current emergency department disposition of patients with acute heart failure: an opportunity for improvement. Journal of Cardiac Failure, 28(10), 1545-1559.

36. Chuzi, S., Saylor, M. A., Allen, L. A., Desai, A. S., Feder, S., Goldstein, N. E., ... & WARRAICH, H. J. (2025). Integration of palliative care into heart failure care: consensus-based recommendations from the heart failure Society of America. Journal of cardiac failure, 31(3), 559-573.

37. Tian, P., Liang, L., Zhao, X., Huang, B., Feng, J., Huang, L., ... & Zhang, Y. (2023). Machine learning for mortality prediction in patients with

38. Ozcan, M., & Peker, S. (2023). A classification and regression tree algorithm for heart disease modeling and prediction. Healthcare Analytics, 3, 100130.

39. Adekkanattu, P., Rasmussen, L. V., Pacheco, J. A., Kabariti, J., Stone, D. J., Yu, Y., ... & Pathak, J. (2023). Prediction of left ventricular ejection fraction changes in heart failure patients using machine learning and electronic health records: a multi-site study. Scientific reports, 13(1), 294.

40. Moreno-Sánchez, P. A. (2023). Improvement of a prediction model for heart failure survival through explainable artificial intelligence. Frontiers in cardiovascular medicine, 10, 1219586.

41. Kumar, A. (2021). Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting. Journal of Engineering and Applied Science, 69, 1-12.

LIBRARY CLEARANCE

PLAGARISM REPORT

212-35-737

ORIGINALITY REPORT

24%	19%	18%	15%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Islamic University of Technology Student Paper	1%
2	jeas.springeropen.com Internet Source	1%
3	Submitted to Midlands State University Student Paper	1%
4	www.mdpi.com Internet Source	1%
5	peerj.com Internet Source	1%
6	arxiv.org Internet Source	<1%
7	pmc.ncbi.nlm.nih.gov Internet Source	<1%
8	studentsrepo.um.edu.my Internet Source	<1%
9	Mohid Qadeer, Rizwan Ayaz, Muhammad Ikhsan Thohir. "Heart Failure Prediction Through a Comparative Study of Machine Learning and Deep Learning Models", The 7th International Global Conference Series on ICT Integration in Technical Education & Smart Society, 2025 Publication	

10	Submitted to Coventry University Student Paper	<1%
11	doria.fi Internet Source	<1%
12	Md Zonayed, Rumana Tasnim, Sayma Sultana Jhara, Mariam Akter Mimona, Molla Rashied Hussein, Md Hosne Mobarak, Umme Salma. "Machine Learning and IoT in Healthcare: Recent Advancements, Challenges & Future Direction", Advances in Biomarker Sciences and Technology, 2025 Publication	<1%
13	umpir.ump.edu.my Internet Source	<1%
14	studenttheses.uu.nl Internet Source	<1%
15	www.coursehero.com Internet Source	<1%
16	Submitted to Sheffield Hallam University Student Paper	<1%
17	Submitted to Bahcesehir University Student Paper	<1%
18	Submitted to Liverpool John Moores University Student Paper	<1%
19	cosmosscholars.com Internet Source	<1%
20	cris.iucc.ac.il Internet Source	<1%

21	Submitted to American Sentinel University Student Paper	<1%
22	Sotiris Bersimis, Polychronis Economou, Athanasios Rakitzis. "Statistical Methods and Applications in Systems Assurance and Quality - Sotiris Bersimis, Polychronis Economou, and Athanasios Rakitzis", CRC Press, 2025 Publication	<1%
	www.frontiersin.org Internet Source	
23	vital.seals.ac.za:8080 Internet Source	<1%
24	Submitted to Manipal International University Student Paper	<1%
25	Submitted to University of Huddersfield Student Paper	<1%
26	www.aasmr.org Internet Source	<1%
27	www.nature.com Internet Source	<1%
28	Submitted to Ashesi University Student Paper	<1%
29	faculty.daffodilvarsity.edu.bd Internet Source	<1%
30	publications.eai.eu Internet Source	<1%
31	www.journalpressindia.com Internet Source	<1%
32		<1%

33	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Computer Science Engineering", CRC Press, 2024 Publication	<1%
34	assets-eu.researchsquare.com Internet Source	<1%
35	Submitted to Adtalem Global Education Student Paper	<1%
36	documents.mx Internet Source	<1%
37	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1%
38	Submitted to University of East London Student Paper	<1%
39	Submitted to Visvesvaraya National Institute of Technology Student Paper	<1%
40	cris.uev.edu.pe Internet Source	<1%
41	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Recent Trends in Intelligent Computing and Communication", CRC Press, 2025 Publication	<1%
42	Submitted to Dublin Business School Student Paper	<1%
43	Submitted to Southern New Hampshire University - Continuing Education Student Paper	<1%
	journals.icapsr.com	

44	Internet Source	<1%
45	link.springer.com Internet Source	<1%
46	"Soft Computing: Theories and Applications", Springer Science and Business Media LLC, 2025 Publication	<1%
47	Submitted to Angeles College Student Paper	<1%
48	Submitted to University of Witwatersrand Student Paper	<1%
49	Submitted to Vrije Universiteit Brussel Student Paper	<1%
50	sarcouncil.com Internet Source	<1%
51	www.geeksforgeeks.org Internet Source	<1%
52	researchoutput.csu.edu.au Internet Source	<1%
53	idss.iocspublisher.org Internet Source	<1%
54	Submitted to Aspen University Student Paper	<1%
55	Submitted to Canterbury Christ Church University Student Paper	<1%
56	Submitted to Swinburne University of Technology Student Paper	<1%

57	Submitted to Alexandru Ioan Cuza University of Iasi Student Paper	<1%
58	F M Javed Mehedi Shamrat, Majdi Khalid, Thamir M. Qadah, Majed Farrash, Hanan Alshanbari. "An explainable multi-objective hybrid machine learning model for reducing heart failure mortality", PeerJ Computer Science, 2025 Publication	<1%
	Submitted to Murdoch University Student Paper	
59	Submitted to University of Wolverhampton Student Paper	<1%
60	Submitted to Mae Fah Luang University Student Paper	<1%
61	ajaronline.com Internet Source	<1%
62	uyo87.co-aol.com Internet Source	<1%
63	www.calaveraslafco.org Internet Source	<1%
64	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1%
65	Submitted to Dundalk Institute of Technology Student Paper	<1%
66	wiredspace.wits.ac.za Internet Source	<1%
67		

68	"Data Mining and Information Security", Springer Science and Business Media LLC, 2025 Publication	<1%
69	www.fastercapital.com Internet Source	<1%
70	"Fifth Congress on Intelligent Systems", Springer Science and Business Media LLC, 2025 Publication	<1%
71	Jin Liu, Xiao-Lan Xu, Bin Wang, Yue Xiao et al. "TaHAK1 promotes salt tolerance via synergistic modulation of K+/Na+ ion homeostasis and auxin signaling in rice", Plant Physiology and Biochemistry, 2025 Publication	<1%
	Submitted to Queen Mary and Westfield College Student Paper	
72	Submitted to University of Strathclyde Student Paper	<1%
73	www.analyticsvidhya.com Internet Source	<1%
74	Submitted to University of Teesside Student Paper	<1%
75	thesai.org Internet Source	<1%
76	"Proceedings of Second International Conference on Advances in Computer	<1%
77	Engineering and Communication Systems",	<1%

	Springer Science and Business Media LLC, 2022 Publication	
78	Delaram Kahrobaei, Enrique Domínguez, Reza Soroushmehr. "Artificial Intelligence in Healthcare and Medicine", CRC Press, 2022 Publication	<1%
79	Sushil Kamboj, Pardeep Singh Tiwana. "Innovations in Computing", CRC Press, 2025 Publication	<1%
80	essay.utwente.nl Internet Source	<1%
81	ijsrst.com Internet Source	<1%
82	www.giiresearch.com Internet Source	<1%
83	www.medrxiv.org Internet Source	<1%
84	"Empowering software engineering automation through explainable Ai", Elsevier BV, 2025 Publication	<1%
85	Daphin Lilda S, Jayaparvathy R. "Effective cardiac disease classification using FS-XGB and GWO approach", Medical Engineering & Physics, 2024 Publication	<1%
86	Mohammad H Alshayeji, Sa'ed Abed. "Heart disease prediction by tabular modeling with deep learning network and interpretability",	<1%

Machine Learning: Science and Technology, 2025
Publication

87 Renu Sharma, Dilip Kumar Mishra, Satyanarayan Bhuyan. "Prospects of Science, Technology and Applications - A Compendium of Symposium", CRC Press, 2024 <1%
Publication

88 Saurav Mishra. "A Comparative Study for Time-to-Event Analysis and Survival Prediction for Heart Failure Condition using Machine Learning Techniques", Journal of Electronics, Electromedical Engineering, and Medical Informatics, 2022 <1%
Publication

Shrikaant Kulkarni, P. William, Vijaya Prakash, Jaiprakash Narain Dwivedi. "Leveraging Artificial Intelligence in Cloud, Edge, Fog and Mobile Computing", CRC Press, 2025
Publication

89 Submitted to UCL <1%
Student Paper

Submitted to University of Stirling
Student Paper

90 Wasswa Shafik, Adel Ben Youssef, Chithirai Pon Selvan, Pushan Kumar Dutta. "Sustainable Healthcare Systems in Africa - Technologies, Practices, and Management", Routledge, 2025 <1%
Publication

91 <1%

92 B. K. Tripathy, Hari Seetha. "Explainable, Interpretable, and Transparent AI Systems", <1%

93 <1%

CRC Press, 2024
Publication

94 Submitted to CSU, San Jose State University <1%
Student Paper

95 Mohamed Abdel-Basset, Hossam Hawash, Laila Abdel-Fatah. "Artificial Intelligence and Internet of Things in Smart Farming", CRC Press, 2024 <1%
Publication

96 Submitted to National College of Ireland <1%
Student Paper

97 Sreedhar, Vikram Rajapura. "Development of Artificial Intelligence Algorithms in Cardiovascular Research: Case Studies in Atrial Fibrillation And Myocardial Infarction", Liverpool John Moores University (United Kingdom) <1%
Publication

98 bmccardiovascdisord.biomedcentral.com <1%
Internet Source

99 doaj.org <1%
Internet Source

100 Submitted to Bentley College <1%
Student Paper

101 Mokheleli, Tsholofelo Diphoko. "A Comparison of Machine Learning Techniques for Predicting Mental Health Disorders.", University of Johannesburg (South Africa), 2024 <1%
Publication

102 betterread.com.au

	Internet Source	<1%
103	jmir.org Internet Source	<1%
104	desklib.com Internet Source	<1%
105	faculty.ksu.edu.sa Internet Source	<1%
106	globalcardiologyscienceandpractice.com Internet Source	<1%
107	journals.sagepub.com Internet Source	<1%
108	Submitted to Daffodil International University Student Paper	<1%
109	S. R. Reeja, Bore Gowda, Y. S. Rammohan, Ganesan Prabu Sankar, G. Jayalatha. "Engineering Science and Technology: Innovations for the Future", CRC Press, 2025 Publication	<1%
110	theses.ncl.ac.uk Internet Source	<1%
111	Abdulrahman Abdullah Bahashwan, Rosdiazli Ibrahim, Madiah Omar, Temitope Ibrahim Amosa. "Supervised learning-based multi-site lean blowout prediction for dry low emission gas turbine", Expert Systems with Applications, 2023 Publication	<1%
112	Ananda Sutradhar, Sharmin Akter, F M Javed Mehedi Shamrat, Pronab Ghosh et al.	<1%

	"Advancing thyroid care: An accurate trustworthy diagnostics system with interpretable AI and hybrid machine learning techniques", Heliyon, 2024 Publication	
113	Augusto César de Camargo Neto. "Evaluating the robustness of AI-based vocal biomarkers against real-world noise - toward regulatory standards and compliance", Universidade de São Paulo. Agência de Bibliotecas e Coleções Digitais, 2025 Publication	<1%
	Submitted to University of Leeds Student Paper	
114	discovery-pp.ucl.ac.uk Internet Source	<1%
115	edepot.wur.nl Internet Source	<1%
116	harvest.usask.ca Internet Source	<1%
117	theses.gla.ac.uk Internet Source	<1%
118	www.bomberbot.com Internet Source	<1%
119	www.tfzr.rs Internet Source	<1%
120	"Modern Practices and Trends in Expert Applications and Security", Springer Science and Business Media LLC, 2025 Publication	<1%
121		

122	Hemant Kumar Saini, Sita Rani, Mariya Ouaissa, Mariyam Ouaissa, Zakaria Abou El Houda, Hajar Moudoud. "Digital Forensics in Next-Generation Internet for Medical Things - Balancing Security and Sustainability", Routledge, 2025 Publication	<1%
123	Komal Kumar Napa, Rajkumar Govindarajan, S. Sathya, J. Senthil Murugan, Bindu Kolappa Pillai Vijayammal. "Comparative analysis of explainable machine learning models for cardiovascular risk stratification using clinical data and shapley additive explanations", Intelligence-Based Medicine, 2025 Publication	<1%
124	Shih-Wei Wu, Cheng-Cheng Li, Te-Nien Chien, Chuan-Mei Chu. "Integrating Structured and Unstructured Data with BERTopic and Machine Learning: A Comprehensive Predictive Model for Mortality in ICU Heart Failure Patients", Applied Sciences, 2024 Publication	<1%
	core.ac.uk Internet Source	
	crd.simad.edu.so Internet Source	
125	dokumen.pub Internet Source	<1%
126	eprints.maynoothuniversity.ie Internet Source	<1%
127	iieta.org Internet Source	<1%
128		<1%
129		

		<1%
130	injoit.org Internet Source	<1%
131	www.preprints.org Internet Source	<1%
132	"ICT: Applications and Social Interfaces", Springer Science and Business Media LLC, 2025 Publication	<1%
133	Shi, Anni. "Kinetics and Applications of On- Surface Topochemical Polymerization of Diacetylene Striped Phases.", Purdue University, 2023 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

ACCOUNT CLEARANCE