Al-Quds Open University
College of Graduate Studies
Master of Information Technology

A Hybrid Model for Predicting Heart Disease Using DNN and ML Algorithms

نموذج هجين للتنبؤ بأمراض القلب باستخدام الشبكات العصبية العميقة وخوارزميات تعلم الالة

THESIS

by
Tharaa Ayad Salamah
Student ID: 0330012210205

Supervisor
Prof. Dr. Yousef Abuzir

Submitted in Partial Fulfilment of the Requirements
For the Degree of Master of Information Technology at the College of Graduate Studies

Ramallah, Palestine

Jan, 2026

Al-Quds Open University
College of Graduate Studies
Master of Information Technology

A Hybrid Model for Predicting Heart Disease Using DNN and ML Algorithms

نموذج هجين للتنبؤ بأمراض القلب باستخدام الشبكات العصبية العميقة وخوارزميات تعلم الالة

THESIS

by
Tharaa Ayad Salamah
Student ID: 0330012210205

Supervisor
Prof. Dr. Yousef Abuzir

Submitted in Partial Fulfilment of the Requirements
For the Degree of Master of Information Technology at the College of Graduate Studies

Ramallah, Palestine

Jan, 2026

Examination Committee Page

نموذج هجين للتنبؤ بأمراض القلب باستخدام الشبكات العصبية العميقة وخوارزميات تعلم الالة

A hybrid model for predicting heart disease using DNN and ML algorithm

By:

Tharaa Ayad Nemer Salamah

Supervisor:

Prof. Dr. Yousef Saleh Abuzir

The thesis was examined and approved on January 25th, 2026,

Members of the Examination Committee:

Supervisor and Chairman: Prof. Dr. Yousef Abuzir Al-Quds open University

Signature: ……………………………Date: 4/2/2026

Committee First Member: Dr. Mohammed Maree The Arab American University

Signature: ……………………………Date: 6/2/2026

Committee Second Member: Dr. Salim Jayousi Al-Quds open University

Signature: ……………………………Date: 5/2/2026

Al-Quds Open University
2026

Declaration

I, Tharaa Ayad Nemer, hereby declare that the work presented in this thesis has not been submitted for any other degree or professional qualification, and that it is the result of my own independent work.

Signed: ثراء سلامة

Date: 25-1-2026

Abstract

A Hybrid Model for Predicting Heart Disease Using DNN and ML Algorithms

Cardiovascular diseases are among the most prominent health challenges facing medical systems worldwide, with high mortality rates associated with this disease recorded annually. This is largely due to the difficulty of early detection of these diseases, as diagnosis often relies on clinical findings that appear in advanced stages of the disease. In this context, the need to employ artificial intelligence (AI) techniques, particularly machine learning and deep learning, emerges to develop accurate predictive models that contribute to the early identification of patients at risk of heart disease. This study aims to develop a hybrid model that combines traditional machine learning algorithms—Support Vector Machines (SVMs), Random Forests, XGBoost, LightGBM, and Logistic Regression—with Deep Neural Networks (DNNs), which are used to extract advanced representative features from data. This study uses a reliable medical dataset obtained from Kaggle consisting of four databases: Cleveland, Hungary, Switzerland, and Long Beach. It contains 76 features, including the predicted feature, but all published experiments report using a subset of 14 of them. The target field indicates the presence of heart disease in the patient. It is an integer value where 0 = no disease and 1 = disease. This study introduces an innovative multi-stage methodology for heart disease prediction, leveraging feature augmentation to enhance performance on a moderately sized dataset of 1,025 patient records. The process begins with preprocessing clinical data, comprising 13 features (e.g., age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise induced angina, oldpeak, slope, number of major vessels, thal), followed by initial training of traditional machine learning models (e.g., SVM, Random Forest , XGBoost, LightGBM , Logistic Regression) using these features. A Deep Neural Network (DNN) then extracts 32 high-level features from its penultimate layer, capturing complex, nonlinear patterns not evident in the original data. These DNN features are combined with the 13 clinical features to form a 45-feature set, significantly enriching the input space. A set of comprehensive evaluation indicators was used, including accuracy, confusion matrix, precision, recall, F1 coefficient, and the area under the ROC-AUC curve, to provide a comprehensive evaluation of the models before and after feature combination. The Random Forest model achieved the highest performance among classification models on the original features, with an accuracy rate of 97.80%, a high recall rate of 99.31%, and a predictive accuracy of 96.64%. In a previous study, Smith et al. (2018) employed traditional machine learning algorithms The experimental results showed that the Support Vector Machine achieved an accuracy of 85.2% .The results also showed a significant improvement in the performance of all classification algorithms after combining the original features with those extracted by deep neural networks (DNNs). This combination resulted in increased classification accuracy across all key indicators. The SVM algorithm achieved the highest AUC value of 99.90%, demonstrating its high ability to accurately distinguish between classes. The Random Forest, XGBoost, and LightGBM algorithms also achieved identical results in overall accuracy (99.63%) and other indicators. The results showed that combining the DNN-extracted features with the original features led to a significant improvement denotes consistency across models, not the numerical accuracy alone. in the prediction accuracy of all machine learning algorithms used, reflecting the effectiveness of the hybrid approach in enhancing predictive performance, especially in light of the challenges associated with class imbalance and small dataset sizes. This study confirms that combining machine learning and deep learning techniques provides a promising path for developing intelligent diagnostic tools capable of supporting medical decision-making, reducing false alarm rates, and contributing to improving early treatment opportunities, especially in resource-constrained medical settings.

keywords: *heart disease prediction, hybrid predictive model, deep neural networks, machine learning classifiers, feature fusion, cardiovascular risk assessment, Evaluation Metrics.*

الخلاصة

نموذج هجين للتنبؤ بأمراض القلب باستخدام الشبكات العصبية العميقة وخوارزميات تعلم الالة

تُعد أمراض القلب والأوعية الدموية من أبرز التحديات الصحية التي تواجه النظم الطبية في جميع أنحاء العالم، حيث تُسجل معدلات عالية من الوفيات المرتبطة بهذا المرض سنويًا. ويعود ذلك إلى حد كبير إلى صعوبة الكشف المبكر عن هذه الأمراض، إذ يعتمد التشخيص غالبًا على النتائج السريرية التي تظهر في مراحل متقدمة من المرض. في هذا السياق، تبرز الحاجة إلى توظيف تقنيات الذكاء الاصطناعي، وخاصةً التعلم الآلي والتعلم العميق، لتطوير نماذج تنبؤية دقيقة تُسهم في التحديد المبكر للمرضى المعرضين لخطر الإصابة بأمراض القلب. تهدف هذه الدراسة إلى تطوير نموذج هجين يجمع بين خوارزميات التعلم الآلي التقليدية – آلات المتجهات الداعمة (SVMs)، وغابات القرار العشوائية، وXGBoost، وLightGBM، والانحدار اللوجستي – مع الشبكات العصبية العميقة (DNNs)، والتي تُستخدم لاستخراج السمات التمثيلية المتقدمة من البيانات. تستخدم هذه الدراسة مجموعة بيانات طبية موثوقة تم الحصول عليها من Kaggle وتتكون من أربع قواعد بيانات: Cleveland, Hungary, Switzerland, and Long Beach. وتحتوي على 76 سمة، بما في ذلك السمة المتوقعة، ولكن جميع التجارب المنشورة تُفيد باستخدام مجموعة فرعية من 14 منها. يشير حقل "الهدف" إلى وجود مرض قلبي لدى المريض. وهو قيمة صحيحة 0 = لا مرض و1 = مرض. تُقدم هذه الدراسة منهجية مبتكرة متعددة المراحل للتنبؤ بأمراض القلب، مستفيدةً من تعزيز الميزات لتحسين الأداء على مجموعة بيانات متوسطة الحجم تضم 1025 سجلًا للمرضى. تبدأ العملية بالمعالجة المسبقة للبيانات السريرية، والتي تشمل 13 ميزة (مثل: العمر، الجنس، نوع ألم الصدر، ضغط الدم أثناء الراحة، نسبة الكوليسترول في الدم، نسبة السكر في الدم أثناء الصيام، نتائج تخطيط القلب أثناء الراحة، معدل ضربات القلب الأقصى، الذبحة الصدرية الناتجة عن التمرين، ذروة العمر، المنحدر، عدد الأوعية الدموية الرئيسية، thal)، يليها تدريب أولي لنماذج التعلم الآلي التقليدية (مثل: SVM، الغابة العشوائية، XGBoost، LightGBM، الانحدار اللوجستي) باستخدام هذه الميزات. بعد ذلك، تستخرج شبكة عصبية عميقة (DNN) 32 ميزة عالية المستوى من طبقتها قبل الأخيرة، ملتقطةً أنماطًا معقدة وغير خطية غير واضحة في البيانات الأصلية. تم دمج ميزات DNN هذه مع الميزات السريرية الـ 13 لتشكيل مجموعة مكونة من 45 ميزة، مما

أدى إلى إثراء مساحة الإدخال بشكل كبير. تم استخدام مجموعة من مؤشرات التقييم الشاملة، بما في ذلك الدقة ومصفوفة الارتباك والدقة والتذكر ومعامل F1 والمساحة الواقعة أسفل منحنى ROC–AUC، لتوفير تقييم شامل للنماذج قبل وبعد دمج الميزات. حقق نموذج الغابة العشوائية أعلى أداء بين نماذج التصنيف على الميزات الأصلية، بمعدل دقة 97.80٪ ومعدل تذكر مرتفع 99.31٪ ودقة تنبؤية 96.64٪. أظهرت النتائج أيضًا تحسنًا كبيرًا في أداء جميع خوارزميات التصنيف بعد دمج الميزات الأصلية مع تلك المستخرجة بواسطة الشبكات العصبية العميقة (DNNs). أدى هذا المزيج إلى زيادة دقة التصنيف عبر جميع المؤشرات الرئيسية. حققت خوارزمية SVM أعلى قيمة AUC بنسبة 99.90٪، مما يدل على قدرتها العالية على التمييز بدقة بين الفئات. حققت خوارزميات الغابة العشوائية، وXGBoost، وLightGBM نتائج متطابقة في الدقة الإجمالية (99.63%) ومؤشرات أخرى. أظهرت النتائج أن دمج السمات المستخرجة من الشبكة العصبية العميقة مع السمات الأصلية أدى إلى تحسن كبير في دقة التنبؤ لجميع خوارزميات التعلم الآلي المستخدمة، مما يعكس فعالية النهج الهجين في تحسين الأداء التنبؤي، لا سيما في ضوء التحديات المرتبطة باختلال توازن الفئات وصغر حجم مجموعات البيانات. تؤكد هذه الدراسة أن دمج تقنيات التعلم الآلي والتعلم العميق يوفر مسارًا واعدًا لتطوير أدوات تشخيص ذكية قادرة على دعم اتخاذ القرارات الطبية، وتقليل معدلات الإنذارات الكاذبة، والمساهمة في تحسين فرص العلاج المبكر، لا سيما في البيئات الطبية محدودة الموارد.

Acknowledgements

In the name of Allah, Most Gracious, Most Merciful

Praise be to God, who has blessed me with His grace, guidance, and favor in completing this thesis.

First, I want to express my deepest gratitude to my beloved Al-Quds Open University and its administration for providing a supportive academic environment. I also extend my sincere thanks to all the professors and faculty members who contributed significantly to the success of this work through their guidance and encouragement. I especially want to mention my supervisor, Prof. Dr. Yousef Abuzir, for his continuous support and constructive advice throughout the research process. His patience, encouragement, and academic wisdom were a true source of inspiration for me.

I offer my heartfelt thanks to my family for their unconditional love and boundless support.

To my father, my inspiration and role model, I thank you for your wisdom, strength, and the values you have instilled in me.

To my mother, whose prayers, patience, and endless care have been the source of my happiness, thank you from the bottom of my heart.

I extend my sincere thanks to my beloved husband for his unwavering support and understanding throughout this journey. I would also like to express my sincere gratitude to my brothers and sisters for their love, encouragement, and unwavering support. Special thanks are also extended to everyone who contributed to the completion of this stage.

This work was completed thanks to God and the support and prayers of all those mentioned above. May God reward you all abundantly.

I dedicate this honorable work to my beloved homeland, Palestine, and to every steadfast soul among our people who inspires hope through their resilience.

I dedicate this work to everyone who believed in me, stood by me, and offered me support and assistance throughout this journey. To everyone who contributed their thoughts or efforts to reaching this milestone, this achievement is as much yours as it is mine.

Tharaa Salamah
Jan, 2026

Table of Contents

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
CDP	Cardiovascular Diseases Prediction
DNN	Deep Neural Network
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
ZSL	Zero-shot learning
DT	Decision Tree
LR	Logistic Regression
DL	Deep learning
CVD	Cardiovascular Disease
CVP	CVD Prediction
ML	Machine Learning
NN	Neural Networks
HAI	Healthcare AI
PA	Predictive Analytics
MD	Medical Diagnosis
RF	Random Forest
DM	Data Mining
BD	Big Data
FE	Feature Extraction
MT	Model Training
CD	Clinical Data
HI	Health Informatics

Chapter 1: Introduction

1.1 Overview and Background

Cardiovascular disease (CVD) is a group of disorders affecting the heart and blood vessels which cause the heart not to pump blood effectively. (CVD) is among the most serious and widespread diseases. It has become the leading cause of death worldwide, affecting both developed and developing countries, (WHO, 2020). An estimated 17.9 million people have died from CVD diseases in 2019, accounting for 32% of all global deaths. As a result of these statistics, interest is increasing towards finding ways to prevent and detect this disease and finding strategies to combat this disease and manage the global crisis. Early detection leads to improving outcomes, reducing the burden on health care systems, and providing more feasible and effective treatment options compared to diagnosis in later stages (Benjamin et al 2019)

CVD encompasses multiple conditions, such as coronary artery disease, heart failure, arrhythmias, cardiomyopathy, and others. These diseases are associated with multiple factors, including unhealthy lifestyles such as obesity, lack of physical activity, smoking, and poor nutrition, as well as genetic and chronic environmental factors such as stress, diabetes, and high blood pressure (American Heart Association, 2019; Centers for Disease Control and Prevention, 2020) In most cases, these diseases do not show clear symptoms in their early stages, making early detection essential to improve treatment and prevention opportunities and reduce the burden on healthcare systems in ,terms of cost resource, allocation, and patient management . In this context, artificial intelligence (AI) techniques, particularly machine learning (ML) and deep learning (DL), have emerged as promising solutions in healthcare for the early detection and prediction of cardiovascular diseases. By leveraging large datasets and advanced algorithms and techniques, machine learning and deep learning models can identify subtle indicators of heart disease that traditional diagnostic methods may, such as an echocardiogram, stress tests, electrocardiogram (ECG), standard blood tests, and more. Although traditional machine learning algorithms such as Support Vector Machines (SVM), Random Forests, XGBoost, LightGBM, and Logistic Regression (LR) have shown promising results in classifying heart disease cases, their effectiveness depends largely on the quality and relevance of the features used in training. This is where deep neural networks (DNNs) play an important role by extracting the most important and complex features from raw data. These networks can detect hidden patterns that traditional models may miss.

Artificial intelligence in the medical field faces numerous challenges, such as the imbalance of data between patient and non-patient cases, which can lead to bias in models. This is in addition to the high computational costs required to train deep models, as well as the difficulty of obtaining large amounts of accurately labeled data. Accordingly, this study aims to propose a hybrid model that combines the capabilities of deep neural networks in feature extraction with the efficiency of machine learning algorithms in classification This approach provides an effective method that enhances performance and mitigates these challenges. Machine learning algorithms are first trained on a set of medical data obtained from the Kaggle platform, and then a DNN model is used to extract the features most closely related to predicting heart disease. These machine learning algorithms are then fed back with the extracted features, retrained, and the results are compared before and after merging. Studies have shown that this type of hybrid model effectively improves classification results, reduces error, and increases accuracy and sensitivity, making it an ideal choice for early prediction of heart disease. Combining deep learning with machine learning not only enhances the model's prediction accuracy, but also addresses many of the challenges facing AI models in the medical field, such as the overfitting problem or the need for large amounts of labeled data. It also contributes to building more reliable diagnostic tools that can be used practically, even in light of limited technical and medical resources.

1.2 Problem Statement

In this study, we aim to develop innovative methods for obtaining rapid and accurate diagnoses to predict cardiovascular diseases through advanced diagnostic techniques, machine learning, and deep learning models. Therefore, early detection of the disease will lead to great benefits (WHO, 2020). Current diagnostic methods may not be comprehensive or proactive enough for early detection of factors and signs of heart disease (Benjamin et al., 2019; Roth et al., 2017). Traditional diagnostic methods usually rely on clear symptoms and traditional examinations, and the disease may not be identified at its beginning and initial stages. In addition, traditional methods may be invasive and expensive, and sometimes may lead to inaccurate results, which necessitates the need for more advanced diagnostic tools and accuracy (Rajkomar et al., 2019; Topol, 2019).

People with heart disease are at risk of serious health complications that may lead to increased mortality. The main reason may be that the symptoms of heart disease are not apparent in the early stages, leading to delayed detection, which increases the severity and deterioration of the health condition and delays treatment. This leads to the limited accuracy of current conventional diagnostic methods. Early intervention is crucial in establishing a modern diagnostic process, which can reduce disease complications and increase the survival rate. Screening using conventional cardiac techniques exposes patients to continuous electrocardiograms, cardiac ultrasounds, and costly laboratory tests that may miss critical cases, lead to inaccurate diagnoses, and misinterpret the condition. It also fails to detect complex and nonlinear patterns in clinical patient data. This leads to variability in clinical symptoms and constraints on healthcare system resources, affecting the overall quality of early detection and treatment of heart disease. This study aims to address these issues by building a model capable of processing patients' medical data using machine learning (ML) and deep learning (DL) algorithms to predict cardiovascular diseases, with the aim of enhancing the accuracy and efficiency of early diagnosis and enabling preventive intervention to reduce the risk of acute cardiac complications that may lead to death.

1.3 Research Question

This study proposes building a hybrid model that combines deep learning for feature extraction and machine learning for classification. The study also seeks to measure the actual performance of the model through a set of important performance indicators. Based on the above, this study seeks to answer the following research questions:

1. How can the combination of deep learning techniques and traditional machine learning algorithms contribute to improving the accuracy of early detection of heart disease compared to traditional diagnostic methods?

This study investigates the effectiveness of integrating the two approaches to achieve better performance in classifying clinical medical data, particularly the factors influencing heart disease.

2. How effective is the proposed model in predicting heart disease using real clinical data?

The study aims to evaluate the model's performance through several metrics, including accuracy, sensitivity, specificity, and other statistical performance indicators, based on patient data collected from accredited global research sources Cleveland ،Hungary ، Switzerland and Long Beach datasets.

3- What is the impact of combining data preprocessing techniques with feature extraction using deep learning on improving the process of identifying true negative and true positive outcomes in the diagnosis of heart disease?

This question focuses on studying the effectiveness of preliminary steps in standardizing and improving data quality, and their role in increasing model accuracy and reducing the likelihood of prediction errors.

4- To what extent can the proposed model contribute to reducing the workload on doctors and improving diagnostic efficiency in heart disease hospitals?

This question discusses the possibility of using the model to assist in making effective and rapid medical decisions, thus facilitating doctors' work and improving the quality of services provided to patients.

By answering these questions, the study aims to build a deeper understanding of the effectiveness of hybrid models for predicting heart disease and to determine the extent of their ability to improve the accuracy of early diagnosis compared to traditional methods. This will contribute to supporting efforts to develop artificial intelligence technologies in the medical field.

1.4 Objective

In the medical field, a correct diagnosis is half the cure. This thesis aims to develop and implement a hybrid model that combines the power of deep learning (DNN) techniques for feature extraction with the efficiency of traditional machine learning algorithms (SVM, Random Forest, XGBoost, LightGBM, and Logistic Regression) for the classification and diagnosis of heart disease. This aims to overcome many of the challenges associated with traditional diagnostic methods and achieve higher, more reliable, and more accurate performance. The main objectives of this study are as follows:

1. Improve the prediction process of heart disease by leveraging the advanced capabilities of deep learning models to extract relevant features from medical clinical data without the need for direct human intervention. Then, use machine learning algorithms for classification, resulting in higher performance compared to using classification before feature integration.

2. Increase accuracy and reduce errors in diagnosis, especially regarding false positives and false negatives, which are among the most prominent challenges facing medical support systems. Reducing these errors contributes to enhancing diagnostic reliability and guiding medical decisions, thus reducing unnecessary procedures that may cause psychological, physical, and financial burdens on patients.

3. Overcome challenges related to data analysis time and reduce the burden on doctors and healthcare practitioners, especially in light of the increasing number of patients and the volume of medical data. The hybrid model enables the provision of smart solutions that support rapid decision-making and the efficient diagnosis of cardiac conditions.

4. Address the training overfitting problem in the medical dataset used in the study by integrating appropriate training and regularization strategies within deep learning feature extraction models, along with machine learning algorithms, to achieve improved generalization and stable performance.

5. Evaluating the performance of the proposed model using an integrated clinical dataset for cardiac diseases, and analysing the classification results in terms of key performance indicators such as accuracy, sensitivity, specificity, and positive and negative predictive value, with the aim of verifying the effectiveness and feasibility of the model in an academic medical context.

By achieving these goals, the study seeks to contribute to the development of more effective and accurate predictive models in the field of medical diagnosis of heart disease, which support data-driven medical decision-making and enhances the reliance on artificial intelligence technologies in the medical field.

1.5 Motivation

This study is motivated by the urgent need to develop effective methods for early detection of heart disease, which has a direct impact on reducing mortality rates and improving treatment opportunities in its early stages. Based on this, this study proposes the use of a hybrid approach that combines deep learning (DNN) techniques to automatically extract medical features and traditional machine learning algorithms to accurately and effectively classify cases. The results before and after feature fusion are compared. This approach aims to overcome the limitations associated with traditional methods and provide an intelligent alternative with greater predictive and medical analysis capabilities. This is achieved through the following factors:

1. The increasing incidence of heart disease globally, which poses a growing challenge to healthcare systems and underscores the need to develop accurate diagnostic tools that facilitate early therapeutic intervention and reduce long-term medical and economic burdens.

2. The shortcomings of traditional diagnostic methods, as they often rely on clinical indicators that are analyzed manually, making them susceptible to human error and delays in disease detection, especially in the early stages of the disease. Many outdated statistical methods lack the ability to analyze complex patterns in medical data, reinforcing the need to adopt modern artificial intelligence techniques.

3. The increasing burden on doctors and cardiologists due to the high number of cases and the difficulty of manually interpreting multiple results and tests, which contributes to delayed diagnosis or inaccurate decision-making. Therefore, the presence of an intelligent model capable of supporting medical decisions will help improve the efficiency of medical performance and reduce the pressure on healthcare workers.

4. The rapid development of artificial intelligence techniques and hybrid models, which have proven effective in analysing complex medical data, especially when combining deep learning with machine learning algorithms capable of providing accurate classifications. These capabilities open new horizons for improving the diagnosis of heart disease and increasing the reliability of clinical results.

1.6 Thesis Contribution and Significance

This thesis makes a significant contribution to the field of heart disease prediction by developing an innovative hybrid model that combines deep learning (DNN) feature extraction techniques with traditional machine learning classification algorithms. This model enhances the accuracy and reliability of diagnosis while also reducing the error rates associated with clinical predictions based on traditional methods. The most significant contributions and significance of this study can be summarized in the following points:

1- Contribution to systematic evaluation: The study relies on combining the capabilities of deep neural networks in deep learning to automatically extract features from medical data, with the high efficiency of traditional machine learning classification algorithms (SVM, Random Forest, XGBoost, LightGBM, and Logistic Regression). This combination provides a powerful model that combines efficiency, accuracy, and speed, while reducing the need for large amounts of labeled data or intensive computational resources.

2- Improving the quality of extracted features: Deep neural networks were employed to extract the most important and relevant features from clinical data, enhancing the performance of classification models and reducing reliance on human expertise to manually identify features.

3- Efficiency of classification algorithms: Traditional machine learning algorithms are used to accurately classify extracted data. They have proven their ability to efficiently handle small-scale, structured data. They also provide high-speed decision-making, enhancing the effectiveness of early diagnosis.

4- Accurate Evaluation of Model Performance: The proposed model will be evaluated using multiple performance indicators such as accuracy, sensitivity, specificity, F1-score, and confusion matrix to verify the model's effectiveness in predicting heart disease and determine its accuracy in distinguishing between positive and negative cases.

5- Reducing Medical Error Rates: By combining deep feature extraction with traditional classification, the model aims to reduce the rates of false positives and false negatives, which lead to inaccurate medical decisions, thus reducing unnecessary medical procedures and improving the quality of healthcare.

6- Model Flexibility and Applicability: The model is highly flexible, enabling it to adapt to various types of clinical data, making it widely applicable to medical support systems and electronic health record analysis in the future.

7-A qualitative and quantitative scientific contribution to the field of predicting heart disease: by presenting a practical hybrid model that combines feature extraction using deep neural networks and traditional machine learning algorithms. At the experimental level, the results showed a significant and consistent improvement in the performance of the models after integrating the features, and it can be used and developed in similar predictive systems for other medical fields such as diabetes and cancer, while emphasizing the future need for external validation using real clinical data to enhance generalizability.

8- The possibility of transferring the results to practical application: This model contributes to paving the way for the development of intelligent medical decision-making support systems in hospitals and healthcare centers, and represents a step toward the effective and safe use of artificial intelligence in real-world healthcare. Based on the above, this thesis represents an important scientific step toward developing smart and more accurate solutions for predicting heart disease, by leveraging the advantages of both machine learning and deep learning. It also highlights the importance of integrating modern technologies into medical decision-making support, opening up broad horizons for future research that can improve the quality of healthcare and expand the scope of artificial intelligence use in various medical fields.

1.7 Thesis Outline

The introductory chapter introduces the research topic, focusing on its objectives, proposed research questions, and significance. It also provides a general introduction to the methodology used in the study, highlighting its research implications. The second chapter of the study provides an in-depth and critical overview of the literature on heart disease detection using machine learning, deep learning, and hybrid learning approaches, revealing weaknesses in current knowledge and contributing to answering the research questions addressed in this thesis. The third chapter describes the methodology, including the complete structure of the heart disease prediction model, which is divided into a data collection phase, followed by a preprocessing phase. The preprocessing phase focuses on identifying and classifying important features associated with increased accuracy. The final phase is model evaluation using model performance measurement metrics. The fourth chapter presents the results of the models and their evaluation metrics, comparing the performance of these algorithms. The final chapter of the thesis presents conclusions and future recommendations and discusses the most important research contributions and future proposals in the field of cardiovascular disease prediction.

Chapter 2: Literature Review

2.1 Introduction

A literature review is an essential part of any research study. It helps the researcher build a deep and comprehensive understanding of the field in which they present their research and helps them identify knowledge gaps and existing challenges that need to be addressed. In this context, this thesis focuses on exploring the effectiveness of artificial intelligence models in predicting heart disease combining machine learning (ML) and deep learning (DL).

From this perspective, the study seeks to provide background, knowledge, and analysis of the latest research findings in this field, including developments in predictive models, to understand the strengths and weaknesses of each approach and how to leverage them to enhance prediction accuracy and early detection.

The importance of the studies also extends to providing analytical insights into the potential for developing more integrated and effective models in the field of medical diagnosis. These reference studies also provide guidance for us as the authors of this thesis, helping us to steer us toward the most appropriate research paths. They also support the selection of appropriate methodologies and provide accurate comparison indicators between the various algorithms and techniques used. They also help identify areas of deficiency or redundancy in knowledge, upon which the contributions of this thesis can be built.

Previous research includes studies on heart disease prediction using machine learning algorithms such as the Decision Tree (DT) algorithm and the Logistic Regression (LR) algorithm, which have demonstrated a number of recent success stories in heart disease prediction and classification. It also includes some deep learning algorithms such as Deep Neural Networks (DNNs). These networks are capable of modelling complex non-linear patterns in medical data and automatically learning informative feature representations that enhance heart disease prediction.

This chapter is devoted to reviewing and analysing approximately 60 carefully selected previous studies related to the research topic, which addressed the applications of artificial intelligence in the diagnosis of heart disease, using machine learning and deep learning algorithms. This review aims to establish a solid knowledge base that supports the objectives of the thesis by presenting the contributions of each study and clarifying its findings. This enhances the methodological accuracy of the study and supports the development of a more efficient model in the field of cardiovascular disease prediction.

2.2 Literature Review

2.2.1 Machine Learning Studies

Artificial intelligence, specifically machine learning, is a vital tool in analysing complex medical data. It is used to identify patterns and relationships between health indicators to predict chronic diseases, such as heart disease. These techniques rely on multiple algorithms that leverage a patient's historical medical data to generate accurate predictive results. The Support Vector Machine (SVM) algorithm stands out for its ability to handle nonlinear relationships and high-dimensional data, enabling accurate and flexible classification performance through the use of kernel functions. Among machine learning algorithms is the Logistic Regression (LR) algorithm, an important statistical tool in predicting heart disease, especially when using data from patient medical records. This algorithm has the ability to analyse the relationship between a set of health factors and determine the extent to which each factor affects the likelihood of developing the disease. It is also simple and easy to interpret, allowing clinicians to understand the results and the rationale for classifying a patient at risk. This makes it suitable for practical use in healthcare settings that require rapid decisions supported by statistical evidence. The Random Forest algorithm has also proven highly efficient in processing high dimensional data without the need for complex preprocessing steps, and by combining the results of multiple decision trees to obtain more stable and accurate classification results. The XGBoost algorithm has proven its effectiveness by improving classification performance and reducing the likelihood of false predictions through gradient boosting techniques and incremental model updating. The LightGBM algorithm has also demonstrated high performance in classifying medical data, especially with large datasets, due to its reliance on intelligent data partitioning techniques, reducing training time, and increasing model efficiency.

Despite the great success achieved by traditional models, they are not without challenges. These algorithms often require fine-tuning parameters and manual feature extraction to achieve higher accuracy and performance. Furthermore, their ability to generalize may be affected when dealing with noisy or heterogeneous data.

Accordingly, reviewing previous studies that have used traditional algorithms to predict heart disease is an important step in understanding current research trends and identifying potential shortcomings. This section focuses on analysing 30 prominent studies that addressed the applications of traditional machine learning algorithms to a set of medical data, with the aim of building a solid scientific foundation that paves the way for discussing more advanced techniques such as deep learning algorithms, which will be discussed in the next section.

Chen (2024) designed a predictive model capable of predicting heart disease, using a dataset from Kaggle containing 319,795 instances and 18 attributes, of which 70% of instances were randomly selected for the training set and 30%. Applying machine learning algorithms in data mining such as DT, LightGBM, RF and LR to forecast heart disease. Before constructing models, data cleaning, feature selection and hyperparameter tuning processes performed, aiming to explore the potential patterns in the data. A Comparative Analysis was conducted on the external test set to compare the prediction performance of different models at the same level. The results reported that the highest accuracy achieved with LightGBM was 76.9%, followed by Logistic Regression and Random Forest, with Decision Tree being the worst. One of the limitations of the study is the small size of the data

Husain et al. (2023) conducted a comprehensive evaluation of multiple machine learning algorithms to predict heart disease using a dataset obtained from the Kaggle Machine Learning repository. The experimental results indicated that DT and RF achieved the highest predictive performance, with an accuracy of 96.75%, outperforming other classifiers such as LR, SVM, and KNN. However, the study is limited by its reliance on a subset of non-invasive clinical features and the absence of usability testing, as the proposed models were evaluated solely from a computational perspective without assessing their interpretability, ease of use, or integration into real clinical workflows, which may affect the generalizability of the findings in practical healthcare settings.

Jiang et al. (2021) investigated the effectiveness of seven machine learning models in predicting CVD risks in this Kazakh Chinese population. All subjects were randomly divided into the training set (80%) and the test set (20%). These models included (DT, KNN, LR, NB, RF, SVM, and XGB). Ten-fold cross-validation was used during model development and hyperparameters tuning in the training set. Model performance was evaluated in the test set in light of discrimination, calibration, and clinical usefulness. RF was applied to obtain the variable importance of included variables.They found that LR and SVM had superior prediction performance compared to other ML models in light of discrimination, calibration, and DCA. The AUC value reached (0.872 and 0.868) respectively. Future research is needed to validate ML models' accuracies with high- dimensional data in this population. The small sample size with limited predictors in this study is a limitation.

Bhatt et al. (2023) developed a system that can correctly predict cardiovascular diseases to reduce fatalities caused by cardiovascular diseases. This paper proposes a method of k-modes clustering with Huang initialization that can improve classification accuracy. Models such as random forest (RF), decision tree classifier (DT), the multilayer perceptron (MP), and XGBoost (XGB) are used. The proposed model is applied to a real-world dataset of 70,000 instances from Kaggle. multilayer perceptron with cross-validation outperformed all other algorithms in terms of accuracy. It achieved the highest accuracy of 87.28%. Among the limitations of the study was that the interpretability of the results.

Yan et al. (2024) focused on predicting an individual's likelihood of developing cardiovascular disease using machine learning techniques. Leveraging a comprehensive dataset from the Kaggle platform, and after standardizing the data, they divided it into two sets: a training set and a test set at an 80:20 ratio. They explored three models: the KNN algorithm (with an overall accuracy of 0.7117), the LR algorithm (with an overall accuracy of 0.726), and the RF algorithm (with an overall accuracy of 0.7335). Through comparative analysis of the three methodologies, the random forest algorithm demonstrated superior performance in terms of prediction accuracy. This research represents an important step towards utilizing machine learning techniques. Among the limitations of the study were the interpretability of the results.

Muhammed et al. (2023) Evaluated the performance and accuracy of several supervised machine learning algorithms by comparing them for heart disease prediction using a dataset obtained from PhysioNet databases. The data were pre-processed to address missing values, normalize the data, and correct for imbalance to ensure unbiased machine learning algorithms. The classifiers that were applied included Artificial Neural Network (ANN), Gradient Boosting, Decision Tree, Naive Bayes, and Random Forest. and their accuracy and other statistical variables were compared using a 10-fold cross-validation technique. Results showed that the ANN algorithm achieved the highest Accuracy of 94.1%. This increases the potential for highly successful supervised machine learning techniques to predict heart disease, demonstrating exceptional potential for practical application and accuracy. The dataset used in this study had limited data on heart disease. This Study limitations results were not validated using external data.

Edafetanure-Ibeh (2024) focused on utilizing a publicly available dataset from the UCI Machine Learning Repository, containing 14 attributes related to predicting cardiovascular disease. The research aimed to identify patterns that might indicate an increased risk of cardiovascular disease and to use machine learning models to predict heart disease. The LDA and KNN models demonstrated the highest average accuracy of 85%, indicating their potential as effective predictive tools. Attributes such as chest pain type, maximum heart rate, age, cholesterol level, and exercise-induced angina were identified as important predictive factors for heart disease. This contributes to reducing the global burden of heart disease and improving personalized treatment. A limitation of the study is the lack of using another data source to generalize the results.

Maurya et al. (2023) employed ML techniques for predicting heart disease and compares the accuracy of different machine learning models, such as SVM, KNN, Logistic Regression, Decision Tree, Random Forest, Gaussian Naive Bayes, AdaBoost, Extra Tree Classifier, and Gradient Boosting, using the UCI Repository dataset for model training and testing.14 important features were selected and the data was divided into 70% training and 30% testing. The data was then cleaned, processed and balanced using under- and over-sampling techniques to improve the accuracy of the model. Among all the models used, the Gradient Boosting model achieved the highest accuracy of 95.08%. The main goal of the research was to develop a reliable and computationally efficient machine learning algorithm for predicting heart disease. Howevwe, the research lacks the use of different types of precision measures.

Sree et al. (2023) designed a predictive model heart disease prediction model and analyzed different algorithms used in disease prediction. In order to increase the predictive accuracy of ML algorithms, A dataset of 6,232 patient records was used, and the data was divided into 75% for training and 25% for testing. The methodology began with data cleaning, duplicate values removal, outlier handling, and feature standardization using MinMaxScaler. Exploratory data analysis was then performed using heat maps, histograms, and biplots to understand relationships between features and improve feature selection for the model. This study compares six algorithms, including KNN, DT, RF, SVM, LR, and Neural Network. 13 attributes, The RF model achieved the highest accuracy of 82%. The project's findings point to the potential of machine learning algorithms as a diagnostic tool for cardiac disease. The Challenges of this study include limited data, incomplete information, changing risk factors, and uncertainty.

Ingole et al. (2024) Analysed of machine learning models in anticipating heart disease risks using clinical data. The paper entails seven ML classifiers Logistic Regression, Random Forest, Decision Tree, Naive Bayes, k-nearest Neighbours, Neural Networks, and Support Vector Machine (SVM) using dataset extracted from Kaggle repository. The methodology begins with data preprocessing, followed by feature engineering to select and transform features. Several models were then trained and compared. The evaluation of the performance of each model was conducted based on accuracy metrics. Interestingly, the Support Vector Machine (SVM) demonstrates the highest accuracy percentage i.e. 91.51%, proving its worth among the evaluated models in the realm of predictive ability. One of the challenges is Overfitting in Decision Trees with small datasets and computational expense and noise in K-Nearest Neighbours.

Molla et al. (2022) developed a system that machine learning-based heart disease prediction framework is proposed to address these issues by identifying risk variables associated with this disease. To ensure the success of our proposed model, effective data preprocessing and transformation strategies were employed to generate accurate data for the training model, which utilized the five most popular datasets (Hungary, Statistics Registries, Switzerland, Long Beach, Virginia State, and Cleveland) from the University of California, Irvine. During the training phase, classifiers were used, including extreme gradient boosting (XGBoost), support vector machine (SVM), random forest (RF), gradient boosting (GB), and decision tree (DT). Univariate results indicated that the DT classifier achieves a relatively higher accuracy of approximately 97.75% compared to other types. Thus, a machine learning approach capable of predicting heart disease with high accuracy was achieved. One of the challenges facing the research is computational complexity.

Gupta et al. (2022) The study employs supervised machine learning methods was used to predict heart disease. Five models were built, including k-nearest neighbour, decision tree, logistic regression, naive Bayes, and support vector machine (SVM). These models used data from the University of California, Irvine (UCI) data warehouse. The selection of the most important features was followed by visual data exploration to understand the disease distribution and biological factors, followed by preprocessing. The results indicate that logistic regression outperformed all other supervised classifiers in terms of performance metrics, with an accuracy rate of 92.3%. these techniques aid in developing medical aid software for the early heart disease.

Chandrasekhar and Peddakrishna (2023) employed machine learning techniques to enhance heart disease prediction accuracy. Six algorithms (random forest, K-nearest neighbour, logistic regression, Naïve Bayes, gradient boosting, and AdaBoost classifier) are utilized, with datasets from the Cleveland and IEEE Dataport. To optimize model accuracy, GridsearchCV, and five-fold cross-validation are employed. In the Cleveland dataset, logistic regression surpassed others with 90.16% accuracy, while AdaBoost excelled in the IEEE Dataport dataset, achieving 90% accuracy. A soft voting ensemble classifier combining all six algorithms further enhanced accuracy, resulting in a 93.44% accuracy for the Cleveland dataset and 95% for the IEEE Dataport dataset. This surpassed the performance of the logistic regression and AdaBoost classifiers on both datasets. This study's novelty lies in the use of GridSearchCV with five-fold cross-validation for hyperparameter optimization, determining the best parameters for the model. The soft voting ensemble classifier approach improved accuracies on both datasets. The limitation of this model is that it is based on a limited amount of patient data.

Baghdadi et al. (2023) evaluate multiple machine learning algorithms for early detection and diagnosis of cardiovascular diseases using Heart Condition data from the UCI Machine Learning Repository. The proposed methodology began with exploratory data analysis, handling missing values, and evaluating the importance of features using (Shapley > 0.1). Then, the data were divided into 70% for training and 30% for testing using k-fold cross-validation. A set of machine learning algorithms including XGBoost, AdaBoost, SVC, Naive Bayes (multiple and Bernoulli), stochastic gradient descent (SGD), decision tree, random forest, and KNN was trained, with parameters optimized using the Optuna library to select the most accurate model for classifying heart disease cases. The proposed CatBoost model yields an F1-score of about 92.3% and an average accuracy of 90.94%. One challenge is that this study relies on limited secondary data, including missing data.

Assegie (2021) investigated the effectiveness of a KNN-based approach for early heart disease detection using Kaggle heart disease data. The data were processed using Pearson correlation analysis to understand the relationships between the features, and then trained a KNN model with 80% for training and 20% for testing. The proposed model addressed the problem of biased classification on imbalanced observations by non-ensemble algorithm through ensemble classifier namely the adaptive boosting. The performance of the KNN algorithm was analysed on the test set, revealing an accuracy of 91.99%. This experimental result underscores the potential effectiveness of the proposed KNN-based model for heart disease prediction.

Bouqentar et al. (2024) conduct a comparative evaluation of several machine learning algorithms to develop an effective system for early cardiovascular disease prediction by choosing one of the powerful existing ML algorithms after a deep comparative analysis of several models. The Cleveland dataset was used, and 70% was allocated for training ML algorithms testing, including DT, FR, SVM, LR, adaptive boosting, and KNN. With hyperparameter tuning using Grid Search and Random Search. The performance of each algorithm was assessed based on accuracy, precision, recall, F1 score, and the Area Under the Curve metrics. SVM achieved the highest accuracy rate of 92%. One of the most important challenges in the study is the limited data and the difficulty of generalizing the results.

Patidar et al. (2022) compare multiple machine learning algorithms for predicting heart disease using a Kaggle dataset. This work presents the results of important data mining techniques that can be used to build a highly efficient and accurate prediction model using the dataset available on Kaggle. Data were pre-processed using one-hot encoding and then split into testing and training data This study compares the heart disease prediction metrics of six machine learning algorithms: LR, DT, RF, SVM, GNB, and KNN. The RF algorithm achieved the highest accuracy rate of 98.53%, helping doctors reduce the number of deaths from heart disease.

Ogunpola et al. (2024) evaluated the effectiveness of machine learning and deep learning classifiers for early heart disease detection under imbalanced data conditions. They conducted a comprehensive literature review to identify effective strategies. Seven machine learning and deep learning classifiers, including K-Nearest Neighbours, Support Vector Machine, Logistic Regression, Convolutional Neural Network, Gradient Boosting, XGBoost, and Random Forest, were deployed to enhance the accuracy of heart disease predictions. The research explores different classifiers and their performance, providing valuable insights for developing robust prediction models for myocardial infarction. The study’s outcomes emphasize the effectiveness of meticulously fine-tuning an XGBoost model for cardiovascular diseases. This optimization yields remarkable results: 98.50% accuracy, 99.14% precision, 98.29% recall, and a 98.71% F-1 score. Such optimization significantly enhances the model’s diagnostic accuracy for heart disease.

Alshraideh et al. (2024) developed and assess machine learning models for heart disease prediction using the Jordan University Hospital Heart Dataset. The primary objective of this study was to enhance prediction accuracy by utilizing a comprehensive approach that includes data preprocessing, feature selection, and model development. Various artificial intelligence techniques, namely, random forest, SVM, decision tree, naive Bayes, and K-nearest neighbours (KNN) were explored with particle swarm optimization (PSO) for feature selection. SVM demonstrated an accuracy of 91.8% without PSO, which improved to 94.3% with the incorporation of PSO. Among the gaps in this study was the small data sample, which led to potential selection bias, limited the possibility of generalization, and relied on accuracy measures while neglecting other aspects.

Petreska (2024) examined the application and effectiveness of various machine learning approaches for the early detection of cardiovascular abnormalities. With special emphasis on their effectiveness compared to traditional diagnostic methods. The data used in the study was taken from Kaggle. The raw data was prepared and processed through data cleaning, replacing missing values with mean and median, feature engineering, data splitting into training and testing, and addressing data imbalance. methods Many machine learning algorithms have been used, namely Logistic Regression, K-Nearest Neighbours, Support Vector, Machine, Random Forest, Gradient, Boosting, XGBoost, LightGBM, Neural Network, Decision Trees. The neural networks showed a very high accuracy of 94.3% and a reasonable accuracy of 79.1%. suggesting that the model. It distinguishes the classes well. One of the primary limitations was the quality and comprehensiveness of the data itself.

Manjula et al. (2024) propose an ensemble-based machine learning framework for the early detection of cardiovascular diseases. Given their significant global impact. Utilizing demographic data, medical history, and lab results, algorithms like Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbours are examined. An ensemble model, incorporating Random Forest, is developed for improved accuracy, initially achieving 95%. Hyperparameter optimization boosts accuracy to 99%. These findings underscore ML’ s potential in early CVD detection, offering advancements in predictive healthcare. This contributes to addressing CVD burdens and improving public health outcomes.

El-Sofany (2024) employ multiple feature selection techniques to develop an accurate machine learning model for early-stage heart disease prediction. The feature selection process was performed using three distinct methods, namely, chi-square, analysis of variance (ANOVA), and mutual information (MI).These classifiers included Naive Bayes, support vector machine (SVM), voting, XGBoost, AdaBoost, bagging, decision tree (DT), K-nearest neighbour (KNN), random forest (RF), and logistic regression (LR), The experimental findings demonstrated that the XGBoost classifier achieved the optimal performance using the combined datasets and SF-2 feature subset with the following rates: 97.57% for accuracy, 96.61% for sensitivity, 90.48% for specificity, 95.00% for precision, 92.68% for F1 score, and 98% for AUC. The limitations of this study are dataset quality and availability and imbalanced classes.

Lu (2024) compared the performance of several machine learning models to enhance heart disease prediction accuracy: SVM, RF, and XGBoost. A dataset containing 12 features from 918 patients was used, with preprocessing steps such as one-hot encoding for categorical variables and MinMax scaling for numerical features. The models were trained and evaluated using 5-fold cross-validation to ensure robustness. Random Forest demonstrated the highest accuracy at 82.78%, followed closely by SVM (82.67%) and XGBoost (81.58%). While the Random Forest model outperformed the others, this study also highlights the need for better interpretability in ML models, especially in medical applications where understanding the relationships between features is crucial.

Dharmawardhana & Thondilage (2023) apply machine learning methods to predict heart disease using large-scale health datasets. The methodology relied on using a comprehensive dataset covering various human health parameters for training and testing. These datasets were collected from the Kaggle platform and included over

300,000 records. These were pre-processed by encoding categorical variables, segmenting the data into training and test data, and standardization. Four machine learning algorithms (Naïve Bayes, K-NN, Decision Tree, and Logistic Regression) were used. The results showed that the Logistic Regression algorithm achieved the highest accuracy of 91.6%. By leveraging these algorithms, healthcare professionals can identify individuals at risk of heart disease at an early stage. One of the limitations of this study is the lack of a detailed analysis of feature selection.

Reddy et al. (2021) evaluated a diverse set of machine learning classifiers for efficient heart disease risk prediction, such as Bayes, functions, lazy, meta, rules, and trees, were trained for efficient heart disease risk prediction using the full set of attributes of the Cleveland heart dataset and the optimal attribute sets obtained from three attribute evaluators. The performance of the algorithms was appraised using a 10-fold cross-validation testing option. The sequential minimal optimization (SMO) achieved an accuracy of 85.148% using the full set of attributes and 86.468% was the highest accuracy value using the optimal attribute set obtained from the chi-squared attribute evaluator. Meanwhile, the meta classifier bagging with logistic regression (LR) provided the highest ROC area of 0.91 using both the full and optimal attribute sets obtained from the ReliefF attribute evaluator. Which confirms its efficiency in detecting infected cases with high accuracy. One of the limitations of this study is the lack of important medical features in the data.

Ananey-Obiri and Sarku (2020) utilized data mining and machine learning techniques to predict the presence of heart disease. By this, machine learning algorithms (logistic linear regression, decision tree classifier, Gaussian Naïve Bayes models) will be developed to predict the presence of heart diseases in patients. The dataset was collected from UCI machine learning repository which contains information on patients with heart disease.The LR algorithm achieved the highest accuracy. The accuracy score for this model was 82.75%. The AUC for this model was 0.86. The burgeoning influence of data mining techniques and machine learning in the medical field in detecting subtle patterns in large datasets make their applicability in heart disease diagnostics relevant.

Yilmaz and Yağın (2022) conducted a comparative performance analysis of machine learning models for coronary heart disease classification. Three different models were created with RF, LR, and SVM algorithms for the classification of coronary heart disease. For hyper parameter optimization, 3-repeats 10-fold repeated cross validation method was used. RF 0.929, SVM 0.897 and LR 0.861 classified coronary heart disease with accuracy. Specificity, Sensitivity, F1 score, Negative predictive and Positive predictive values of the RF model were calculated as 0.929, 0.928, 0.928, 0.929 and 0.928, respectively. The Sensitivity value of the SVM model was higher compared to the RF. Conclusion: Considering the accurate classification rates of coronary heart disease, the RF model outperformed the SVM and LR models. Also, the RF model had the highest sensitivity value. One of the drawbacks of this study is that it relies on a limited number of features.

Ozcan and Peker (2023) developed a decision tree–based model for heart disease prediction by integrating multiple public datasets. The Classification and Regression Tree (CART) algorithm, Data preprocessing was performed, duplicate values were removed, numeric values were converted to categories, and the data split 80% for training and 20% for testing. Performance was evaluated using accuracy, sensitivity, specificity, and positive precision measures, with tree pruning to reduce overfitting. When considering all performance parameters, the 87% accuracy of the prediction validates the model's reliability. On the other hand, extracted decision rules reported in the study can simplify the use of clinical purposes without needing additional knowledge. The proposed model does not totally include patients' medical information.

Nzenwata et al. (2025) investigated the effectiveness of the Extra Tree algorithm for improving heart disease prediction accuracy. The Extra Tree model outperforms a number of baseline models in terms of accuracy and predictive power. The dataset was obtained from the University of California, Irvine (UCI) Machine Learning Repository. This study reduced the number of features from 18 to 7, by using recursive feature elimination method, which uses Random Forest as an estimator. The Extra Tree model demonstrates great performance, showing high accuracy, precision, recall, and f1 scores of 93.1%, 94.8%, 100% and 98.9% respectively on a dataset split ratio of 80% to 20% train set and test set respectively. The study concluded that the model may be implemented into a clinical decision support system to help healthcare providers diagnose cardiac disease. One of the limitations of this study is the need for verification on larger and more diverse data.

2.2.2 *Deep Learning Studies*

Deep learning techniques have revolutionized cardiovascular disease prediction thanks to their ability to automatically extract complex features from raw clinical and physiological data. Previous studies have demonstrated the high effectiveness of deep neural networks, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), in analysing electrocardiogram signals, patient records, and other health indicators related to heart disease. Convolutional neural networks have demonstrated superior performance in identifying patterns associated with heart disease, outperforming traditional machine learning methods, especially when dealing with large, heterogeneous datasets. These networks achieve this by reducing the need for manual feature engineering and improving classification accuracy using pre-trained data. Similarly, recurrent neural networks (RNNs), especially long-short-term memory (LSTM) networks, are highly effective at modelling temporal dependencies in time-series medical data, improving the prediction of abnormal heart disease by capturing subtle sequence variations in patient health records. In the context of heart disease prediction, deep learning techniques are an effective means of extracting essential features from complex clinical data, enabling subsequent improvement of traditional classification models. The use of deep neural networks (DNNs) has enhanced the predictive power of models by constructing high-level representations of data without the need for manual feature selection, an important advantage in medical data that is often unstructured or multidimensional.

Despite these advances, deep learning models face some limitations, particularly the need for large amounts of labelled data and significant computational resources for training. These limitations pose challenges in clinical settings, where medical data is often scarce and imbalanced. As a result, machine learning techniques remain more efficient for final classification in some scenarios, while deep learning is particularly valuable for extracting features with high accuracy. In this section, we review 20 previous studies that applied deep learning models to predict cardiovascular disease.

Sharma et al. (2020) suggested an attention-based DNN model for coronary heart disease risk prediction. The proposed method has been evaluated by predicting CHD risk in the Korean population and compared with typical machine learning algorithms. As a result, the proposed attention-based DNN improved the performance of DNN and outperformed regular machine learning classifiers. The performance measurements include accuracy, precision, recall, f1-score, and AUC score reached 82.7818%, 86.6499%, 77.2704%, 81.6631%, and 82.5802%, respectively. One of the drawbacks of this study is that it relies on the DNN algorithm alone without taking advantage of the advantages of other techniques.

Zhang (2024) developed a DNN based prediction model to address latent risks associated with heart disease. A dataset from the Kaggle database was used for this approach. To prepare the data, binary and discrete values were converted to numerical values using a label encoder, and the data were then split into training and test sets in a 70:30 ratio, respectively. The study adopted a deep neural network (DNN) model with two hidden layers, using ReLU and Sigmoid activation functions, a cross-entropy loss function, and the Adam algorithm to optimize performance. Features with a correlation value greater than 0.05 were selected to enhance the accuracy of the model. DNN eventually achieved outstanding performance with an accuracy of 0.76, recall_1 rate of 0.77, and AUC value of 0.84, respectively. One of the limitations of the study is the imbalance of data.

Rufai et al. (2018) designed an ANN model to forecast coronary heart disease status in patients. the input data essential for diagnosis of heart disease were obtained from the UCI Machine Learning Repository, A multi-layer permutation (MLP) artificial neural network with a backpropagation algorithm was used for training. Experiments included determining the number of nodes in the hidden layer, activation functions, learning rates, and momentum constant to achieve the best accuracy. The optimized artificial neural network (ANN) system demonstrated a notable diagnostic accuracy of (92.2%), thereby substantiating its efficacy in facilitating the diagnosis of coronary heart disease. Limitations of the study included the limited sample size and its reliance on data from only one source.

Vani (2021) introduced an improved LSTM-based model for heart disease prediction by enhancing the internal forgetting gate mechanism. First, the irregular time interval is smoothed to obtain the time parameter vector, and then it is used as the input of the forgetting gate to overcome the prediction obstacle caused by the irregular time interval. The experimental findings demonstrate that the dynamic prediction model proposed in their study exhibits a noteworthy enhancement in classification performance (89.6 %) when compared to the conventional LSTM model (84.4 %). One of the limitations of the study is that it relies on the LSTM model only without comparison with other algorithms.

Pavithra et al. (2024) investigated the application of LSTM networks for continuous patient monitoring and anomaly detection using wearable health data. Our approach is to train an LSTM model on a huge dataset such as PhysioNet, different databases, encompassing a wide variety of physiological conditions and signal types, so the model learns across a wide range of health conditions and demographics. Findings: The proposed LSTM-based model reduces false positives by 30% and false negatives by 40%, hence improving anomaly detection accuracy by up to 30% over prior methods such as SVM, Random Forest, KNN, DBSCAN, TDL, and DQN. This LSTM model achieved an accuracy of 93%, which is very high compared to other models. contributing to more reliable and timely medical interventions. The study may face challenges related to the quality and quantity of data collected.

Islam et al. (2019) proposed an attention-based LSTM model for predicting cardiovascular disease risk factors. The attention model calculates weights (weights are calculated using activation functions such as tanh and softmax). These adjusted weights are then used as inputs to the LSTM network. The data used includes 779 cases with 14 key features collected from multiple medical sources, and missing values were carefully handled to maintain model quality. The model achieved 95% accuracy and 0.90 Matthews. Correlation Coefficient (MCC) scores; better than any other previously proposed methods. The experimental results show that the attention module-based LSTM outperforms than the other statistical machine learning algorithms for the prediction. One of the limitations of the study is that the performance of the model was not evaluated in actual clinical settings.

Almatari et al. (2024) investigated cardiovascular disease risk prediction using deep learning techniques and statistical analysis. Utilizing SPSS and Weka tools, a cross-sectional and correlational design was employed to analyse extensive medical datasets. Binomial regression analysis revealed significant associations between age (폭=0.004) and body mass index (폭=0.002) with CVD development, highlighting their importance as risk factors. Leveraging Weka's DL algorithms, a predictive model was constructed to classify CVD causes. Particularly, convolutional neural networks (CNN) showcased remarkable accuracy, reaching 98.64%. Despite the effectiveness of this technique, the study focused on samples of university students and employees in the Kingdom of Saudi Arabia, limiting the generalizability of the results to broader societies.

Li (2024) developed a predictive model for heart disease by integrating machine learning and deep learning approaches. During the data processing phase, abnormal values within the feature variables were removed, and a BMI feature variable was added to the dataset to visualize the relationships between the data more intuitively. Deep neural networks were used to predict cardiovascular disease and were compared with eight traditional machine learning algorithms with respect to accuracy, F1 score, PR and ROC. The results indicated that the deep neural network (DNN) is the optimal model for predicting cardiovascular disease. Prediction accuracy 0.763% and other metrics. Limitation of this study is the lack of implementation of more advanced and integrated models that could yield higher accuracy than existing models.

Mary and Ramaprabha (2024) designed a deep learning–based approach for cardiac disease prediction. The architecture employs a multi-layered neural network and deep neural network to capture intricate patterns within the data, optimizing predictive accuracy. The data set downloaded from Kaggle with 14 attributes is used in this study. The deep learning model multilayer neural network and deep neural network are chosen for this study. To analyse data through steps including extraction, preprocessing, and partitioning into training and test sets, and then classifying cases. Results indicate that the deep learning algorithm exhibits promising predictive capabilities 80 %, outperforming traditional methods. However, the study did not address aspects related to the generalization of results and external validation.

Gengadevi (2023) conducted a comparative analysis of machine learning and deep learning models for coronary artery disease prediction. The dataset was used to create prediction models using machine learning techniques like Naive Bayes, Support Vector Machine, Random Forest, and Gradient Boosting and DNN are applied to compare the results and analysis of the CAD Disease .The model building steps involved splitting the data into training and test sets, normalizing features, adjusting parameters using cross-validation Result shows that compared to ML algothims and DL technique, DNN gives more accuracy in less time for the prediction. The prediction accuracy obtained by DNN algorithm is 93.56% and the prediction accuracy obtained by SVM is 83.34% . This study also did not extensively discuss the issue of data representation across age groups or demographic backgrounds, which could affect the reliability of the prediction results.

Chandrasekhar & Peddakrishna (2023) evaluate multiple ML techniques to enhance heart disease prediction accuracy. Six algorithms (random forest, K nearest neighbour, logistic regression, Naïve Bayes, gradient boosting, and AdaBoost classifier) are utilized, with datasets from the Cleveland and IEEE Dataport. Researchers used parameter optimization techniques such as GridSearchCV and 5-fold cross-validation. In the Cleveland dataset, logistic regression surpassed others with 90.16% accuracy, while AdaBoost excelled in the IEEE Dataport dataset, achieving 90% accuracy. A soft voting ensemble classifier combining all six algorithms further enhanced accuracy, resulting in a 93.44% accuracy for the Cleveland dataset and 95% for the IEEE Dataport dataset. This surpassed the performance of the logistic regression and AdaBoost classifiers on both datasets. The study faced the problem of computational limitations.

Kumar et al. (2023) proposed a CNN-based deep learning framework for effective heart disease prediction using ECG and clinical data. The proposed method uses a combination of ECG signals, demographic data, and clinical measurements Identifying risk factors for cardiovascular disease in patients. The proposed CNN-based model includes several layers, such as convolutional ones, pooling ones, and fully connected ones. The convolution operation is used to extract features from the ECG signals. The outcomes of experiments demonstrate the suggested CNN-based model achieves high accuracy in predicting heart disease, with an overall accuracy 86%. The model also outperforms several alternatives to classical techniques for heart disease prediction, Additionally, there are limitations to the study, including the use of a small sample size and a limited number of sensors.

García-Ordás et al. (2023) introduced a deep learning–based approach that combines feature augmentation and classification for heart disease prediction. This dataset consists of 918 samples with only 11 clinical characteristics per sample. Features were extracted by Sparse Autoencoder to a larger number of features and the SAE and classifier are trained concurrently to improve feature quality and increase prediction accuracy. With this approach, we have achieved 90.088% which represents a 4.4% improvement in comparison with the results obtained by classic classifiers (MLP or RF) trained on the same dataset and under the same conditions. One of the limitations of the study is that the Sparse Autoencoder used for feature augmentation relies primarily on the quality of the original features, as the augmented features may not provide truly useful information.

Bharti et al. (2021) performed a comparative study between traditional machine learning algorithms and deep learning models for heart disease prediction. which includes 14 clinical features used for performing the analysis. The study used a comparative methodology between traditional machine learning algorithms (such as K-Nearest Neighbours, Decision Tree, and Logistic Regression) and a deep learning algorithm (Multilayer Perceptron (MLP)). For selecting the features and only choosing important features, the Lasso algorithm was used. Using deep learning, the accuracy achieved was 94.2%. Thus, the maximum accuracy achieved by the machine learning model was K nearest Neighbours (83.29%), us the conclusion can be drawn here is that, for this dataset, the deep learning algorithm achieved 94.2 percent accuracy which is greater than the machine learning models. One of the most prominent limitations facing the study is the small sample size.

Zhang et al. (2021) proposed a hybrid deep learning model integrating embedded feature selection for accurate heart disease prediction. In this paper, a novel heart disease prediction model is given. They proposed a heart disease prediction algorithm that combines the embedded feature selection method and deep neural networks. These features are fed into the deep neural network that they built. Our model is tested on the heart disease dataset obtained from Kaggle. Some indicators including accuracy, recall, precision, and F1-score are calculated to evaluate the predictor, and the results show that our model achieves 98.56%, 99.35%, 97.84%, and 0.983, respectively, and the average AUC score of the model reaches 0.983, confirming that the method they proposed is efficient and reliable for predicting heart disease. One of its most prominent limitations is its computational complexity.

Mehmood et al. (2020) demonstrated the superiority of a CNN-based deep learning model over existing machine learning approaches for early cardiovascular disease prediction. The proposed method is concerned with temporal data modelling by utilizing CNN for HF prediction at its earliest stage. We prepared the heart disease dataset and compared the results with state-of-the-art methods and achieved good results. Experimental results show that the proposed method outperforms the existing methods in terms of performance evaluation metrics. The achieved accuracy of the proposed method is 97%. One of the limitations of the study is the risk of overfitting.

Sajja and Kalluri (2020) compared convolutional neural networks with traditional machine learning models, demonstrating the superior performance of CNNs in early cardiovascular disease prediction. This paper focuses on a comparison between the traditional approaches such as Logistic Regression, K-Nearest Neighbours (KNN), Naïve Bayes (NB), Support Vector Machine (SVM), Neural Networks (NN), and the proposed prediction model of CNN The proposed CNN model focused on simplifying the steps through the convolutional structure, as these networks have the ability to receive raw data and implement processing stages. The model consists of two convolutional layers, followed by two dropout layers, and concludes with a classification output layer. Disease (CVD) predictions with 94.78% accuracy. One of the limitations of the study is the risk of overfitting.

Hussain et al. (2021) established the effectiveness of a 1D CNN-based deep learning architecture over classical machine learning methods for accurate heart disease classification. Several clinical parameters are used for evaluating the risk contour in the patients which supports in early diagnosis. Various regularization methods are used to avoid overfitting in the proposed model. The proposed model achieves over 97% training accuracy and 96% test accuracy on the dataset. This is compared in detail with other machine learning algorithms using various performance parameters which prove the effectiveness of the proposed model. One of its most prominent limitations is the limited sample size.

Tomov and Tomov (2018) proposed a deep neural network architecture that significantly outperforms existing machine learning and deep learning models in heart disease diagnosis. This work led to the discovery of a novel five- layer DNN architecture – named Heart Evaluation for Algorithmic Risk-reduction and Optimization Five (HEARO-5) – that yields the best prediction accuracy. To evaluate and tune the architectures we use kway cross-validation as well as Matthews correlation coefficient (MCC) to measure the quality of our classifications. The study is performed on the publicly available Cleveland dataset of medical information. The HEARO-5 architecture, yielding 99% accuracy and 0.98 MCC, significantly outperforms currently published research in the area. One of the limitations of the study is the lack of transparency in the detailed structure of the model.

Demir et al. (2023) validated the superiority of DNN-based feature extraction over conventional approaches for improving cardiovascular disease prediction accuracy. In the study, 918 observations obtained from the kaggle site were used. Deep neural network models were built and the model was trained and hyperparameter adjustment was performed with Grid Seach CV to achieve the highest success rate. As a result of the study, accuracy, precision, recall and F1-Score values were found as 0.9375, 0.9629, 0.9176, 0.9397 for DNN. Their study highlights the effectiveness of DNN-based feature extraction in improving cardiovascular disease prediction. Despite the outstanding performance, CNN's limitations include its inability to effectively handle chronologically sequenced clinical data.

After reviewing approximately 20 studies on deep learning, these studies indicate that different deep learning architectures offer complementary advantages in predicting cardiovascular disease. Convolutional neural network (CNN)-based models are effective at extracting complex patterns from clinical and signal-based data, while RNN and LSTM models are better suited for monitoring temporal dependencies in sequential medical records. However, most studies rely on deep learning models as the ultimate predictive model and as independent predictors, despite common challenges such as limited datasets, data imbalances, high computational costs, and the potential for overfitting. Furthermore, the lack of systematic comparison between architectures limits the generalizability of the findings. As a result of these limitations, a hybrid approach is needed, where deep learning is primarily used for feature extraction, while traditional machine learning classifiers are used for the ultimate prediction to enhance stability, efficiency, and generalizability.

2.2.3 Hybrid models studies

Hybrid models are gaining significant and increasing importance in modern medical research, given their ability to combine the advantages of both deep learning and traditional machine learning techniques, reducing the limitations associated with using each separately. The main idea of these models is to use deep learning algorithms, such as deep neural networks, to extract essential features from raw data. Traditional machine learning algorithms, such as SVM, Random Forest, XGBoost, LightGBM, or Logistic Regression, are then used to classify these features efficiently and accurately.

This integration is an effective step toward improving model performance, as DNNs help build high-level representations of complex clinical data, which are difficult to achieve using traditional methods that require manual feature selection. On the other hand, traditional machine learning algorithms offer significant speed and transparency, making them ideal for classification tasks after the feature extraction phase is complete.

Previous studies using this hybrid approach have shown promising results in terms of increasing prediction accuracy, reducing training time, and achieving a balance between computational complexity and efficiency. For example, some models have used the DNN algorithm to extract features strongly related to prediction. These features were then classified using SVMs or Random Forests to achieve advanced classification performance. It has also been observed that these models offer greater flexibility in adapting to new or diverse data, provided that certain challenges are considered, such as designing an appropriate model architecture and ensuring seamless integration between the extraction and classification stages.

This section will review 15 of the most prominent studies that have adopted hybrid models in the field of heart disease prediction, focusing on design methodologies, evaluation techniques, and the effectiveness of combining DNNs for feature extraction with traditional machine learning algorithms in achieving accurate and easy-to-interpret results.

Mohammad and Al-Ahmadi (2023) proposed a hybrid WT-CNN framework that significantly outperforms conventional classification approaches in heart disease prediction. The key phases of WT-CNN are ECG data collection, preprocessing, RUSBoost-based data balancing, CWT-based feature extraction, and CNN-based final prediction. Through extensive experimentation and evaluation, the proposed model achieves an exceptional accuracy of 97.2% in predicting heart disease. The experimental results show that the approach improves classification accuracy compared to other classification approaches and that the presented model can be successfully used by healthcare professionals for predicting heart disease. One of the limitations of the study is that it relies on a limited sample of data.

Sharma et al. (2024) demonstrated the superior performance of a hybrid deep learning model over standalone recurrent architectures for coronary heart disease prediction. The best feature subset is chosen based on the corelation score and dataset classes are balanced using the cluster-abundant data class approach. Bidirectionl Long Short-Term Memory (BiLSTM) and Gated Recurrent Unit (GRU) hyper-parameter tuning is accomplished using Randomized Search Cross-Validation Optimization (RSCV). In comparison to GRU, LSTM, and BiLSTM-GRU, this suggested model obtains a classification accuracy of 98.28% which outperforms existing models. Also, we intend to adapt the model for real-time monitoring and prediction, facilitating early intervention and preventive measures for individuals at risk of heart disease. The study suffers from challenges in interpreting the model.

Singh et al. (2024) introduced a novel CNN–RNN hybrid architecture that surpasses individual deep learning models in predicting heart disease from IoMT data. In this study is concentrated on the creation and application of a cutting-edge deep learning framework. The suggested framework employs convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to the fullest extent possible to extract complex temporal dependencies from the physically heterogeneous data collected by IoMT devices. The biggest accomplishment was the creation of CNN-RNN's new hybrid architecture. This architecture has the ability to extract spatial and sequential characteristics from a variety of patient data flows. The results of our tests demonstrated that DCNN performs better than RNN and CNN, with an astounding accuracy of 99.42% The treatment processes give medical professionals crucial knowledge about the predictive factors that influence the model's decisions, which enhances the model's interpretationl. The study suffers from challenges in interpreting the model.

Sandhu et al. (2024) validated the effectiveness of a CNN–LSTM hybrid model, achieving higher predictive accuracy than standard machine learning techniques. The UCI ML library dataset is used in this work Comparisons of different factors are used to study the proposed model, and each time the suggested hybrid model does better. Compared to other standard machine learning models, the results of the experiments showed an accuracy of 90%, 82% sensitivity, along with 94% specificity. It will be possible to test this CNN+LSTM model with real-time medical information in the future and look at how well it works. One of the limitations of the study is that it relies on a limited dataset.

Bhavekar et al. (2024) developed a hybrid ML–DL prediction model that outperforms individual classifiers in both accuracy and computational efficiency. This prediction uses the proposed method of an artificial neural network and the Chi2 feature selection method applied to determine which features from the dataset were suitable for prediction. The proposed methodology uses classifiers like Support Vector Machines

(SVM), Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), and Artificial Neural networks (ANN). Python was used to conduct the study that assessed the ANN system proposal with the Cleveland heart disease dataset at the University of California (UCI). Compared to other algorithms, the model achieves an accuracy of 96.74% and takes 0.49 seconds to execute, making it superior in predicting heart disease. One of the study's limitations is the challenges in interpreting the model.

Sadr et al. (2024) present a hybrid framework integrating deep learning and machine learning models, demonstrating superior classification performance across multiple datasets. Two public heart disease classification datasets with 70,000 and 1190 records besides a locally collected dataset with 600 records were used in our experiments. Then, a model which makes use of both machine learning and deep learning was proposed in this paper. The proposed model employed CNN and LSTM, as the representatives of deep learning models, besides KNN and XGB, as the representatives of machine learning models. The proposed model obtained the highest classification performance based on all evaluation metrics on all datasets 95.85%. One of the study's limitations is the challenges in interpreting the model.

Mohammed et al. (2025) proposed a hybrid machine learning model that outperforms single-algorithm approaches in predicting heart disease risk from body composition data. This paper proposed six machine learning algorithms using body composition dataset, the algorithms are Decision Tree Model (DTM), XGBOOST, LIGHTGBM, Support Vector Machine (SVM), KNN and Hybrid model. Missing values were addressed and data normalized to ensure quality. The experimental result indicates that the HHP model outperformed others in precision, recall, F-score, with accuracy of 90.2%. Some limitations of the study include the difficulty of interpreting the decisions of the hybrid model, and the lack of comprehensive medical data.

Hao et al. (2025) demonstrated that integrating CNNs with Transformer architectures yields better predictive performance than traditional models for heart disease detection. Through experimental verification, the proposed model shows better performance than traditional models in multiple indicators, especially in processing multi-dimensional features and capturing global information. The complete model (combining CNN and Transformer) achieved the best results, with accuracy, precision, and recall rates of 0.85, 0.84, and 0.86, respectively. The combination of the two effectively improves the overall performance of the model, especially showing strong application potential in the task of heart disease prediction. One of the limitations of the study is that it uses daily life data, which may not fully reflect all factors affecting heart health.

Krishnan et al. (2021) established the superiority of a hybrid RNN-based deep learning model over existing recurrent and deep neural network architectures for heart disease prediction. This proposed model resulted in an outstanding accuracy of 98.6876% which is the highest in the existing model of RNN. The model was developed in Python 3.7 by integrating RNN in multiple GRU that operates in Keras and Tensorflow as the backend for deep learning process, supported by various Python libraries. The recent existing models using RNN have reached an accuracy of 98.23% and deep neural network (DNN) has reached 98.5%. One limitation of the study is the challenges in interpreting the model.

Mishra and Shivaji (2024) introduced a hybrid CNN–LSTM model that enhances dynamic heart disease forecasting by combining spatial feature extraction and temporal analysis. The study introduces a Hybrid DL model combining CNN and LSTM with Feature Enhancement (FE) and Comprehensible Artificial Intelligence (CAI) techniques. The model's effectiveness was evaluated using a freely available cardiovascular disease dataset. Accuracy was assessed with and without FE. The results show that the CNN-LSTM model achieved an accuracy of 83.5% with FE and 88.2% without FE. The proposed CNN-LSTM model offers a significant improvement in forecasting heart disease dynamics by effectively integrating CNN for feature extraction and LSTM for temporal analysis. One limitation of the study is the challenges in interpreting the model.

Sudha and Kumar (2023) utilize an architecture that merges convolutional neural networks with a long short-term memory (LSTM) network to achieve higher accuracy than the traditional machine learning approaches. To predict heart disease. The hybrid CNN and LSTM method was applied over the heart disease dataset to classify it as normal and abnormal. This hybrid system has shown an accuracy of 89%, and it was validated using k-fold cross-validating technique. To establish the efficiency of

proposed method, it is compared with various machine learning algorithms such as SVM, Naïve Bayes and Decision Tree. The results shows that the proposed algorithm achieves better performance than the existing machine learning models.One limitation of the study is the challenges in interpreting the model.

Baccouche et al. (2020) suggested an ensemble-learning framework of different neural network models, and a method of aggregating random under-sampling. To improve the performance of the classification algorithms, we implemented a data preprocessing step with feature selection. Experiments were conducted with unidirectional and bidirectional neural network models and results showed that an ensemble classifier with a BiLSTM or BiGRU model with a CNN model had the best classification performance with accuracy and F1-score between 91% and 96% for the different types of heart disease. These results are competitive and promising for heart disease dataset. One limitation of the study is its reliance on a limited dataset.

Imran et al. (2023) developed a hybrid ensemble framework that consistently outperforms conventional machine learning classifiers in cardiac disease prediction. It proposes a hybrid ensemble framework to analyze the cardiac data based on essential features for optimum prediction results. This study uses the Cleveland open access dataset to discuss the working performance of famous classification techniques like Decision Tree, Naive Bayes, SVM, KNN, logistic regression, RF, Gradient Boosting, and XGB Classifier. It proposes a Hybrid Ensemble Framework based on this analysis to enhance the results. The proposed method shows incredible results using the Adaptive Boosting Ensemble technique. Our proposed Hybrid Ensemble Framework achieved an accuracy of 91.80%, precision= 0.94, f1-score=0.92, macro avg= 0.92, and recall = 0.93. The results obtained by the other machine-learning algorithms are less than our model. One of the limitations of this study is overlearning.

Rani et al. (2021) proposed a hybrid decision support system for the early detection of heart disease based on clinical parameters. The study adopted a hybrid system for heart disease prediction using the UCI Cleveland dataset. Features were selected using a hybrid approach that combines a genetic algorithm (GA) to generate and optimize multiple feature sets, and a regressive feature elimination (RFE) method to refine these sets and select the features most influential in model performance. Class balancing was then performed using SMOTE. The model was then built using support vector machine, naive bayes, logistic regression, random forest, and AdaBoost classifiers. It has been found that the system has given the most accurate results with the random forest classifier. The proposed hybrid system was tested in the simulation environment developed using Python. It has achieved an accuracy of 86.6%, which is superior to some of the existing heart disease prediction systems found in the literature. One limitation of the study is the challenges in interpreting the model.

Ghanem and Abuzir (2024) developed a hybrid model combining feature selection, machine learning algorithms, and GAN-based data augmentation to predict heart disease accurately. Because the sample size was insufficient to get accurate predictions, it was increased by using the Generative Adversarial Network (GAN) algorithm. The performance of the proposed ML model was estimated using numerous ML algorithms. K-Nearest Neighbours (KNN), Random Forest, AdaBoost, and Support Vector Machines (SVM) were used in the model. These methods helped the researchers better understand the data and make more accurate predictions. The results obtained using this approach achieved a 99% accuracy rate by using KNN and SVM models and utilizing GAN-generated samples and feature selection strategies. These findings show that the approach that combines feature selection and machine learning algorithms is useful for the early and accurate prediction of heart disease. One of the limitations of this study is that short data dataset leads to bias.

2.3 Evaluation Metrics and their Importance

Most studies on heart disease diagnosis, whether they rely on machine learning algorithms using pre-extracted feature data, studies that rely solely on deep learning, or hybrid models that combine deep learning and machine learning, use a set of evaluation metrics to measure model performance. The importance of these metrics varies depending on the nature of the data, the nature of the problem, the degree of data balance, and the strength of the model used. The following is a definition of the most important of these metrics and the reasons for their adoption:

Accuracy: The ratio of correct predictions (positive and negative) to the total number of samples. It is important when the classes within the data are balanced, but it can be misleading if the data distribution is uneven, as it may give a high value despite the model failing to detect rare positive cases.

Predictive accuracy: This accuracy represents the ratio of correct positive samples to the total number of samples classified as positive by the model. This metric focuses on reducing false alarms and is a vital metric when misclassifying a healthy patient as a heart patient is less acceptable. However, it does not consider false negatives, so it might miss actual positive patients.

Sensitivity or Recall: Expresses the model's ability to detect all actual positive cases (heart patients). Its importance increases as the need to avoid missing any disease becomes more urgent, as in the early detection of high-risk heart patients. However, it does not reflect the ability to correctly identify negative cases.

F1-Score: The harmonic mean between predictive accuracy and sensitivity, used to balance performance when the cost of false positives and false positives are close. In heart disease diagnosis, F1 ensures a balance between not missing cases and reducing false alarms. However, it may not reflect overall performance if class distribution is highly imbalanced.

ROC Curve and Area Under the Curve (AUC): Measures the model's ability to distinguish between patients and non-patients across all threshold levels. The area under the curve represents a general indicator of discrimination power, useful whether the data is balanced or not. AUC is an essential metric for evaluating diagnostic models for heart disease. However, Does not specify performance at a single clinically relevant threshold. In short, no single metric provides a comprehensive view of a model's performance. A combination of metrics—particularly sensitivity, AUC, and F1—must be employed to obtain an accurate and comprehensive estimate of a model's ability to predict heart disease. The following table 2-1 summarizes the approximate values of these metrics in each study, along with the dataset used:

Table 2-1 : *Summary of evaluation scale values for each study and the name of each dataset*

N	Studies	Accuracy	Precision	Recall	F1- score	AUC	Dataset Name	methodology used
1	Chen (2024)	76.9%	N/A	74.2%	76.2%	84.6%	Kaggel	ML
2	Husain et al. (2023)	96.75%	97.81%	95.73%	96.76%	N/A	Kaggel	ML
3	Jiang et al. (2021)	N/A	27.0%	97.1%	N/A	%87.2	CHCSC	ML
4	Bhatt et al. (2023)	87.28%	88.70%	84.85%	86.71%	95%	Kaggel	ML
5	Yan et al. (2024)	73.3%	72%	78%	75%	80%	Kaggle	ML
6	Muhammed et al. (2023)	94.1%	94.1%	94.1%	94.1%	99.1%	PhysioNet	ML
7	Edafetanure-Ibeh (2024)	85%	N/A	N/A	N/A	N/A	UCI	ML
8	Maurya et al. (2023)	95.08%	N/A	N/A	N/A	N/A	UCI	ML
9	Sree et al. (2023)	82%	82%	80%	81%	N/A	Kaggel	ML

N	Studies	Accuracy	Precision	Recall	F1- score	AUC	Dataset Name	methodology used
10	Ingole et al. (2024)	91.51%	93%	92%	93%	N/A	Kaggle	ML
11	Molla et al. (2022)	97.75%	97%	93%	97%	97.41%	UCI	ML
12	Gupta et al. (2022)	92.30%	90.74%	96.08	93.34%	N/A	UCI	ML
13	Chandrasekhar and Peddakrishna (2023)	95%	87%	94%	90%	N/A	Cleveland and IEEE Dataport	DL
14	Baghdadi et al.(2023)	90.94%	93.1%	88.4%	92.3%	N/A	UCI	ML
15	Assegie (2021)	91.99%.	N/A	N/A	N/A	N/A	Kaggle	ML
16	Patidar et al. (2022)	98.53%	100%	97.08%	98.52%	N/A	Kaggle	ML
17	Rani et al. (2021)	86.6%	88.46%	84.14%	86.25%	N/A	UCI	Hybrid
18	Ogunpola et al. (2024)	98.50%	99.14%	98.29%	98.71%	N/A	Mendeley ,Kaggle	ML
19	Alshraideh et al. (2024)	94.30%	90%	94%	93%	89%	JUH Heart Dataset	ML
20	Petreska (2024)	94.3%	79.1%	72.9 %	90.5%	83.2%	Kaggle	ML
21	Bouqentar et al. (2024)	92.00 %	86.96 %	90.62 %	0.93	0.91	Cleveland Clinic	ML
22	El-Sofany (2024)	97.57%	95%	96.61%	92.68%	98%	private HD dataset	ML
23	Lu (2024)	82.78%,	92.2%	88.8%	90.5%	N/A	Kaggle	ML
24	Dharmawardhana & Thondilage (2023)	91.6%	53.6%	10.8%	18%	N/A	Kaggle	ML
25	Chandrasekhar & Peddakrishna (2023)	95%	N/A	N/A	N/A	N/A	Cleveland and IEEE Dataport	DL
26	Reddy et al. (2021)	86.13%	86.2%	86.1%	86.1%	85.8%	Cleveland heart dataset	ML
27	Ananey-Obiri and Sarku (2020)	82.75%	82%	87%	84%	86%	UCI	ML
28	Yilmaz and Yağın (2022)	92.9%	92.9%	92.8%	92.8%	N/A	IEEEDataPort	ML
29	Ozcan and Peker (2023)	87%	88%	85%	N/A	N/A	IEEEDataPort	ML
30	Nzenwata et al. (2025)	93.1%	94.8%	100%	98.9%	N/A	UCI	ML
31	Sharma et al. (2024)	98.28%	97.57%	98.86%	98.21%	N/A	BRFSS	DL
32	Singh et al. (2024)	99.42%	98.88%	99.1%	99.12%	N/A	Kaggle	Hybrid
33	Sandhu et al. (2024)	90%	94%	82%	N/A	N/A	UCI	Hybrid

N	Studies	Accuracy	Precision	Recall	F1- score	AUC	Dataset Name	methodology used
34	Bhavekar et al. (2024)	96.74%	89%	90%	89%	N/A	(UCI)	Hybrid
35	Kumar et al. (2023)	86%	86%	86%	N/A	N/A	Raspberry Pi	DL
36	García-Ordás et al. (2023)	90.09%	90%	N/A	N/A	N/A	Kaggle	DL
37	Bharti et al. (2021)	94.2%	83.1%	82.3%	N/A	N/A	UCI	DL
N	Studies	Accuracy	Precision	Recall	F1- score	AUC	Dataset Name	methodology used
38	Sadr et al. (2024)	95.85%	95.36%	93.01%	92.63%	N/A	Kaggle	Hybrid
39	Mohammed et al. (2025)	90.2%	93%	87%	90%	N/A	Kaggle	Hybrid
40	Hao et al. (2025)	85%	84%	86%	N/A	N/A	Kaggle	Hybrid
41	Gengadevi (2023)	93.56%	89%	89%	93%	N/A	Kaggle	DL
42	Zhang et al. (2021)	98.56%	97.84%	99.35%	0.983%	0.983%	Kaggle	DL
43	Krishnan et al. (2021)	98.69%	96.31%	100%	N/A	N/A	Kaggle	Hybrid
44	Rufai et al. (2018)	92.2%	N/A	N/A	N/A	N/A	UCI	DL
45	Vani (2021)	N/A	49.9%	81.1%	60.8%	89.6%	UCI	DL
46	Pavithra et al. (2024)	93%	93%	95%		74%	PhysioNet	DL
47	Islam et al. (2019)	95%	94%	94%	94%	99%	Kaggle	DL
48	Almatari et al. (2024)	98.64%	N/A	N/A	N/A	N/A	UCI	DL
49	Li (2024)	76.3%	71%	69%	77.2%	77.5%	Kaggle	DL
50	Mary and Ramaprabha (2024)	80 %	88%	70%	78%	N/A	Kaggle	DL
51	Mohammad and Al-Ahmadi (2023)	97.02%	95.68%	94.8%	95.99%	N/A	Kaggle	Hybrid
52	Mishra and Shivaji (2024)	88.15%	87.11%	80.04%	N/A	N/A	Kaggle	Hybrid
N	Studies	Accuracy	Precision	Recall	F1- score	AUC	Dataset Name	methodology used
53	Remya (2024)	91%	85%	88%	N/A	N/A	Kaggle.	Hybrid
54	Mehmood et al. (2020)	97%	97.06%	96.35%	96.70%	N/A	UCI	DL
55	Sajja and Kalluri (2020)	94.78%	N/A	N/A	N/A	N/A	UCI	DL
56	Hussain et al. (2021)	97.79%	94.73%	100%	97.29%	96.15%	Cleveland database	DL
56	Tomov and Tomov (2018)	99%	98%	100%	99%	N/A	Cleveland	DL
57	Demir et al. (2023)	93.75%	96.29%	91.76%	93.97%	96%	Kaggle	DL
58	Zhang (2024)	76%	N/A	77%	N/A	84%	Kaggle	Hybrid

N	Studies	Accuracy	Precision	Recall	F1- score	AUC	Dataset Name	methodology used
59	Hussain and Mohammed (2025)	98.71%	98.86%	N/A	N/A	99.96%	Kaggle	Hybrid
60	Sudha and Kumar (2023)	89%	93%	81%	N/A	N/A	Kaggle	Hybrid
61	Sharma et al. (2020)	82.7818%	86.6499%	77.2704 %	81.6631%	82.5802 %	UCI	Hybrid

2.4 Challenges and Research Gaps

The literature review focuses on several important limitations found in the studies reviewed. These limitations are key factors in understanding the difficulties and challenges facing this field and help guide future research toward addressing these challenges. Table 2-2 shows a summary of these limitations

Table 2-2 : *Limitations of the Literature*

Limitation	Description	References
Low rating on multiple evaluation metrics	Some studies have relied on a single metric to evaluate the model (such as accuracy) without using multiple metrics such as AUC, F1-score, or sensitivity, which does not give an accurate picture of the model's performance.	Sajja and Kalluri (2020), Almatari et al. (2024), Rufai et al. (2018), Edafetanure-Ibeh (2024) Maurya et al. (2023), Alshraideh et al. (2024)
Neglecting to automatically select attributes	Many studies have relied on the original features only, without applying techniques to extract the most relevant features, which may lead to suboptimal model performance.	Chen (2024), Yan et al. (2024), Bhatt et al. (2023), Edafetanure-Ibeh (2024), Sree et al. (2023), Dharmawardhana & Thondilage (2023)
Lack of hybrid integration	There are many studies that focus on using machine learning or deep learning algorithms separately, with few studies that combine the two methods to leverage their combined advantages.	Assegie (2021), Baghdadi et al. (2023), Reddy et al. (2021), Sree et al. (2023), Maurya et al. (2023), Yilmaz and Yağın (2022), Ananey-Obiri and Sarku (2020), Sharma et al. (2020), Vani (2021),
overfitting	Some studies, especially deep ones, face the problem of overfitting when a small data set is used, which weakens their ability to generalize.	Singh et al. (2023), Sajja and Kalluri (2020), Zhang et al. (2021), Raza et al. (2023), García-Ordás et al. (2023), Ingole et al. (2024) Zhang (2024),
Data imbalance	Many studies suffer from the use of imbalanced datasets, where the number of sick cases is significantly lower than the number of healthy cases, which affects the accuracy and reliability of the model. They also fail to use techniques to rebalance the data.	Nzenwata et al. (2025), Zhang et al. (2021), Islam et al. (2023), Bouqentar et al. (2024), Alshraideh et al. (2024), Chen (2024), Jiang et al. (2021), Baghdadi et al. (2023), El-Sofany (2024), Zhang (2024),

Limitation	Description	References
Data Scarcity	A small dataset leads to bias. Black-box models hide reasons.	Ghanem M., Abuzir Y. 2024) Husain et al. (2023), Muhammed et al. (2023), Edafetanure-Ibeh (2024), Sree et al. (2023), Chandrasekhar and Peddakrishna (2023), Baghdadi et al. (2023) Bouqentar et al. (2024), Alshraideh et al. (2024) , El-Sofany (2024) , Reddy et al. (2021) , Rufai et al. (2018) , Sandhu et al. (2024) , Hao et al. (2025) , Baccouche et al. (2020)
Interpretability in Complex Models	Gaps include real-world tests and explainable outputs	Bhatt et al. (2023), Molla et al. (2022), Sharma et al. (2020) , Chandrasekhar & Peddakrishna (2023), Tomov and Tomov (2018), Singh et al. (2024), Hao et al. (2025), Krishnan et al. (2021), Mishra and Shivaji (2024), Rani et al. (2021)

This study focuses on addressing several methodological gaps and research limitations identified in previous studies on the use of machine learning (ML) and deep learning (DL) techniques in predicting heart disease. The most prominent challenges addressed in this study include data imbalance: Most studies have not fully addressed this challenge, leading to models being biased toward the most representative population in the data, thus reducing their ability to accurately predict rare cases. Furthermore, some studies that have used techniques such as SMOTE have not discussed the impact of this treatment on model stability or the potential for unrealistically inflated performance if not evaluated rigorously. Additionally, the problem of data scarcity has emerged, as most studies have relied on small datasets or datasets with limited geographic and demographic diversity, reducing the models' ability to generalize when applied to new data. Another important gap is the issue of interpretability in complex models. Deep models such as LSTM and CNN tend to operate as black boxes, making it difficult to interpret their results and understand how they make decisions, limiting their adoption in clinical settings that require clear clinical justification. In addition, evaluation across multiple performance metrics is weak. Some studies have focused solely on accuracy, without considering other metrics such as recall, precision, and F1-score. This may provide an incomplete picture of the model's performance in cases of unbalanced data. Automatic feature selection processes have also been neglected, with some studies relying on all available features without assessing their importance, leading to increased noise and reduced model efficiency. Furthermore, a lack of hybrid integration between ML and DL algorithms has been noted, as only one type of model was used, limiting the ability to leverage the advantages of each to achieve better performance. Many studies have also suffered from the problem of overfitting, resulting from training on a small or unrepresentative dataset. This reduces the model's ability to generalize when tested on new data.

To address these gaps, an integrated hybrid model was developed that combines the effectiveness of traditional machine learning algorithms (SVM, Random Forest, XGBoost, LightGBM, and Logistic Regression) in classification with the ability of deep neural networks (DNNs) to extract highly representative and nonlinear features from clinical data, in order to enhance prediction accuracy and model stability. The methodology began by isolating the independent features from the target feature. The data were then stratified to ensure balanced representation of classes in the training and test sets, which is essential when dealing with imbalanced data.

A comprehensive statistical analysis of the data was also performed to ensure its balance and efficiency. Standardization was applied to control the variance in numerical feature metrics, improving the efficiency of the models and accelerating the training process. In a subsequent step, representative features were extracted using a DNN and combined with the original features, providing a comprehensive representation of the patient's clinical condition. Traditional machine learning algorithms were retrained on this set of features, and their performance before and after the combination was compared to accurately and objectively assess the achieved improvement. Performance was measured using a set of standard indicators commonly used in medical evaluation: confusion matrix, accuracy, precision, recall, F1-score, and area under the receiver operating curve (ROC-AUC). These indicators represent an important scientific benchmark for measuring the performance of predictive models in the healthcare field, particularly in applications whose results are related to sensitive medical decisions. These steps showed the user (doctor, researcher) the justifications for the prediction, instead of the decisions remaining as a black box that is difficult to interpret and understand. Through this integrated methodological framework, the study was able to avoid the most significant shortcomings of previous studies and present a highly efficient hybrid model characterized by generalizability, reliability, and interpretability, positioning it well for practical application within clinical decision support systems.

2.5 Conclusion

This literature review has demonstrated the vital foundation provided by previous studies in building this thesis, which focuses on developing a hybrid model combining machine learning and deep learning for predicting heart disease. Through a critical and documented analysis of previous studies on machine learning algorithms, deep learning models, and hybrid models, the current landscape of technologies used in the medical field, particularly in the diagnosis of heart disease, is clarified, along with the resulting improvements in the accuracy of early detection of this disease. The literature shows that machine learning algorithms such as Support Vector Machine (SVM), *Logistic Regression* (LR), *Decision Trees* (DT), *Random Forests* (RF), and *K-Nearest* neighbours (KNN), are effectively used in classification tasks and are characterized by their interpretability, flexibility, and accuracy in specific scenarios. However, they have limitations when processing complex and high-dimensional data, such as medical data. In contrast, deep learning models, particularly deep neural networks (DNNs), have demonstrated a high ability to automatically extract features from raw data, leading to accurate results in classifying and diagnosing many medical conditions, including heart disease. Despite these successes, deep learning faces challenges such as the need for massive amounts of data, the consumption of significant computational resources, and the poor interpretability and comprehensibility of the model.

To overcome these challenges, hybrid models have emerged that combine the capabilities of deep learning in feature extraction with the efficiency of machine learning in classification. Studies have shown that this integration contributes to the construction of more efficient and accurate models, reduces overfitting, and improves the model's generalization ability. This review underscores the importance of previous studies in guiding this thesis toward developing an innovative hybrid model capable of achieving a balance between accuracy, efficiency, interpretability, and generalization. A deeper understanding of the advantages and limitations of previous models helps document and improve the model proposed in this study, serving ongoing efforts to improve early diagnosis mechanisms and therapeutic outcomes for patients with heart disease, enabling informed decision-making.

.

Chapter 3: Methodology

3.1 Overview

3.1.1 Summary of the Proposed Multi-stage Approach

Heart disease is one of the leading causes of death worldwide, requiring early and effective diagnosis to reduce risk and mortality rates. Traditionally, the diagnosis of heart disease relies on a variety of clinical examinations, blood tests, electrocardiograms, and medical imaging. However, relying solely on traditional medical assessments can sometimes lead to delayed diagnosis or inaccurate assessments, especially in complex or atypical cases. Clinical data such as age, blood pressure, heart rate, cholesterol, family history, and blood glucose levels represent important sources of biomarkers associated with heart disease. However, interpreting this data manually without extracting patterns from it poses a significant challenge for the medical field.

In recent years, the field of medical diagnostics has witnessed significant developments thanks to artificial intelligence technologies, particularly through machine learning (ML) and deep learning (DL) algorithms. These technologies have proven effective in processing and analysing complex medical data and in the early prediction of chronic diseases such as cardiovascular disease. Hybrid models that combine the advantages and algorithms of machine learning for efficient classification with the skills of deep learning for extracting complex features are among the most effective approaches in this field (Javed et al., 2023; Islam et al., 2023). However, one of the challenges in using deep learning models for heart disease prediction is the heavy reliance on vast amounts of accurately labeled clinical data. This is difficult to achieve in the medical field due to limited access to reliable and balanced large data sets, in addition to ethical and privacy restrictions. If a small or unbalanced dataset is available, a deep learning model may suffer from poor performance due to overfitting and loss of generalization when tested with new data. The model proposed in this study focuses on a hybrid approach that aims to overcome the shortcomings and limitations of deep learning and machine learning models when used separately. By leveraging the superior capabilities of deep neural network (DNN) models in extracting complex features from clinical data, and the computational and classification accuracy of machine learning algorithms such as SVM, Random Forest, XGBoost, LightGBM, and LR, the proposed model provides an integrated framework that balances predictive power with practical effectiveness. This strategic combination is particularly ideal when dealing with limited or unbalanced medical populations. It enhances the overall performance of the model by combining deep data representation with accurate classification capabilities, thus improving and enhancing the reliability of early prediction of heart disease in real-world clinical settings. This methodology focuses on building an effective hybrid model for predicting heart disease through the analysis of patient clinical data using artificial intelligence techniques. This model relies on deep neural networks (DNNs) to extract rich representative features from clinical data, including vital information. After extracting the deep features, they are combined with the original features to form a comprehensive representation of the data. These combined features are used to train traditional machine learning classifiers to predict a patient's health status (infected or not). This section includes a detailed overview of the stages of developing the hybrid model, starting with the data collection phase, followed by the data processing phase, then building a DNN model to extract features, and finally training traditional classifiers on the hybrid features. The model's performance is evaluated using accurate and global benchmarks to ensure the reliability of the results. In addition, the methodology highlights the advantages of each technique used, and the role of this integration of machine learning and deep learning in overcoming the limitations and challenges associated with limited medical data, class imbalance, and the difficulty of generalization. The model aims to support the medical field by providing an effective predictive tool capable of early disease diagnosis and reducing health risks.

3.1.2 Flowchart visualizing

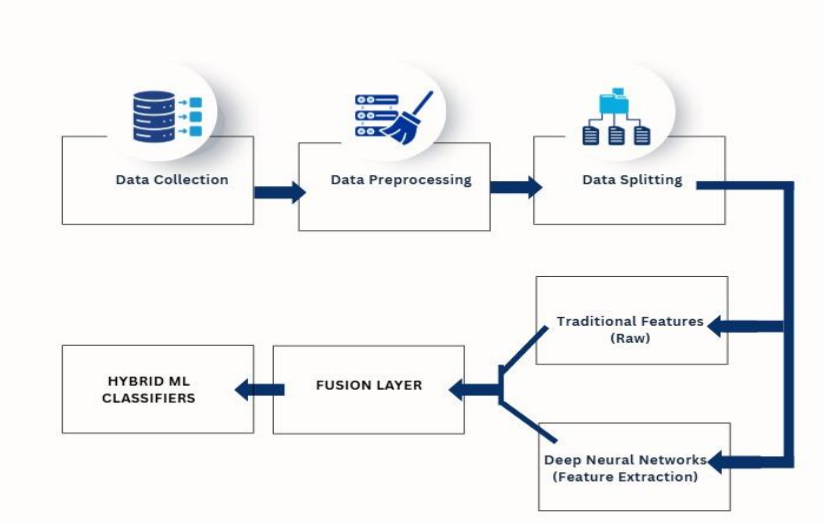


Figure 3-1 : *hart of the hybrid approach methodology for detecting heart disease*

Now, we will briefly present the implementation mechanism of the proposed hybrid model for heart disease prediction, as illustrated in Figure 3-1. This is an effective framework that combines the power of deep learning with the efficiency and accuracy of traditional machine learning models. The first stage of the methodology begins by collecting patient clinical data obtained from the Kaggle platform, which includes a set of important medical characteristics such as blood pressure, age, heart rate, cholesterol level, blood sugar, medical history, and other important indicators. A series of data processing steps are then performed, including separating features independent of the target feature. The data is then divided into two sets (training, test) randomly, noting that the distribution of categories in the original data is relatively balanced, with category (1) constituting 51.32% compared to 48.68% for category (0). As a result of this close proportion, the random division was sufficient. The standard scaler algorithm is then applied to standardize the numerical values of the features and improve training efficiency. After processing is complete, traditional machine learning models are first trained directly on the original data to provide a performance baseline. This is achieved using several algorithms, including support vector machines (SVMs), random forests, XGBoost, LightGBM, and logistic regression (LR). A deep neural network (DNN) is then used to extract rich, deep features from clinical data. These representations are a deep mathematical reflection of the complex relationships in the data. In the next step, the extracted deep features are combined with the original features to form a hybrid feature matrix. These combined features are then fed back into machine learning classifiers to improve classification efficiency and accuracy. Model performance is then evaluated using a set of common metrics such as accuracy, precision, recall, F1-score, and area under the curve (AUC) to ensure the model's efficiency and reliability. This proposed model offers more interpretable results than deep network models alone, making it an effective tool for supporting medical decisions and improving the quality of healthcare. Model performance can be further improved by tuning parameters or increasing data quality, supporting the development of an accurate and effective intelligent system for predicting heart disease in real-world settings.

3.2 Data collection

3.2.1 Dataset

This study used a medical dataset obtained from Kaggle (Heart Disease Dataset) accessed in March 2024. a reliable and widely used resource for cardiovascular disease prediction research. It is taken from a machine learning repository at the University of California, Irvine. This dataset dates back to 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach. The original dataset It contains 76 features, including the predicted feature, but all published studies report using a subset of 14 of them. The "target" field indicates the presence of heart disease in the patient. It is an integer value with a value of 0 = no disease and 1 = disease.The dataset was downloaded directly from Kaggle in CSV format without any data merging or modification, and all records were used as presented. The subset used in our study contains the medical records of 1,025 patients, along with a set of clinical features collected to build accurate AI models for early diagnosis of these diseases.The following section will show a list of features with a detailed description of the clinical features..Since this data is available as an open source and contains no identifying or personal information about the participants, its use in this study does not violate privacy standards or research ethics, making it suitable for use in applied research and academic studies.

3.2.2 Feature list

The independent features or variables in a dataset form the foundation upon which AI, machine learning, and deep learning models build their predictive capabilities. These features are strongly and directly linked to clinical and demographic factors known in the medical literature as key predictors of heart disease risk. Careful selection of these features is crucial to ensuring the prediction accuracy of these models, as any missing or inappropriate variables could directly impact model performance or lead to misleading results.The study lists the 14 features explicitly, ensuring clarity on the data structure. For example, features include:

- Numerical: age, resting blood pressure, serum cholesterol, maximum heart rate achieved, ST depression (oldpeak), and number of major vessels (ca).
- Categorical: sex, chest pain type, fasting blood sugar, resting electrocardiographic results, exercise-induced angina, slope of the ST segment, and thalassemia.

- Target: Binary outcome (0 or 1).

The dataset's size (1,025 records) and feature set are well-defined, allowing researchers to retrieve and verify the data easily. No proprietary or restricted data sources are used, which enhances accessibility. This diversity of features enables models to understand the interaction between different factors, increasing the accuracy of predictions, enhancing model reliability, and helping to build more comprehensive and accurate models for early diagnosis. Table 3.1 shows a list of features with a detailed description of the clinical features :

Table 3-1*. : *Feature list: Detailed description of clinical features

features	Description
Age	age in years
Sex	(1 = male; 0 = female)
Cp	chest pain type (4 values)
Trestbps	resting blood pressure (in mm Hg on admission to the hospital)
Chol	serum cholestoral in mg/dl
Fbs	fasting blood sugar > 120 mg/dl
Restecg	resting electrocardiographic results (values 0,1,2)
Thalach	maximum heart rate achieved
Exang	exercise induced angina
Oldpeak	oldpeak = ST depression induced by exercise relative to rest
Slope	the slope of the peak exercise ST segment
Ca	number of major vessels (0-3) colored by flourosopy
Thal	thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

3.2.3 Dataset characteristics

A comprehensive statistical analysis was conducted on the dataset used in the study to determine its characteristics and assess its readiness for use in building heart disease prediction models, as well as the processing required. The statistical analysis included several basic indicators for each variable, namely data type, number of missing values, number of unique values, in addition to descriptive statistical values such as min, max, mean, and standard deviation. These indicators help understand the nature of each feature, determine the degree of data balance, identify outliers or extreme values for processing, and ensure the absence of empty values. This helps improve the accuracy and reliability of the models and reduce errors during the training process. The analysis results, as shown in Table 3.2 , showed that the data was completely free of missing values for all features. This reflects its high quality and shortens processing operations such as data cleaning or completion before model building. The data also contains a mix of numeric and binary variables, including demographics, clinical measurements, and diagnostic tests. The target characteristic, representing the presence or absence of disease, appeared at a rate of (51.32%: 48.68%) As shown in Table 3-2, indicating a relative balance between the groups and mitigating the impact of class imbalance. Demographically, 69.5% of the sample was male, which may reflect a partial bias in the sample distribution. This should be taken into account when generalizing the results to broader populations. Clinical indicators also showed significant variation, with the mean systolic blood pressure (trest bps) being approximately 131.6 mmHg (range: 94–200) and the mean cholesterol (cholesterol) being approximately 246 mg/dL, with a maximum value of 564 mg/dL, indicating severe disease. The mean fasting blood sugar (FBS) was 0.15, indicating that most patients did not exceed the diagnostic threshold for hyperglycemia. In diagnostic tests, the average restecg was approximately 0.53, while the average maximum heart rate (thalach) was 149 beats per minute, an important indicator of cardiac performance during exercise. The average ST segment depression (oldpeak) was 1.07, and the incidence of exertional angina (exanginal angina) was approximately 33.6% of patients.

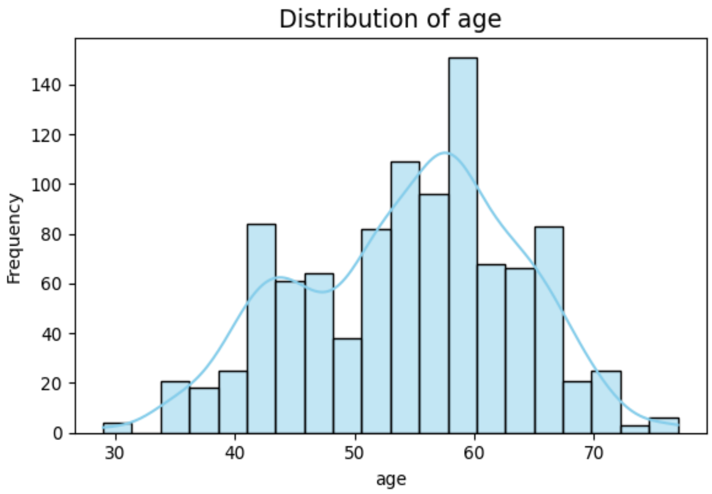
These results reflect significant statistical diversity, as evidenced by the standard deviation values. This is a positive feature for model building, as it allows models to learn from diverse data patterns. Overall, these characteristics demonstrate that the dataset is suitable and reliable for use in building advanced artificial intelligence models for early prediction of heart disease, thanks to its relative balance, lack of missing values, and the diversity of clinical and demographic characteristics it includes.In the next section, data processing operations will be discussed in detail.

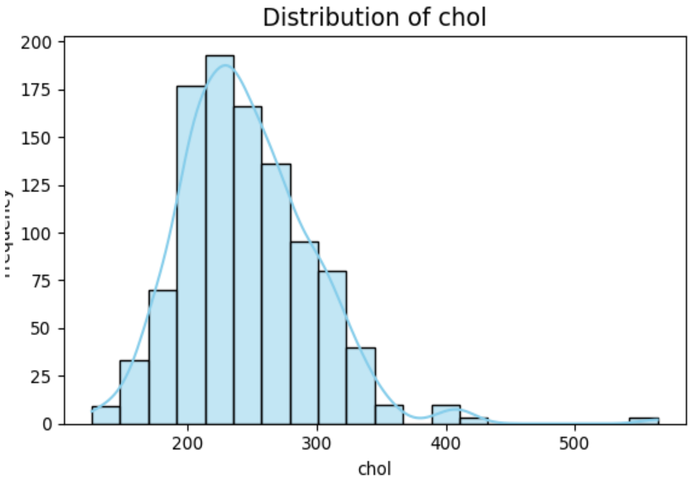
Table 3-2 : *statistical summary showing the data type and key statistical values for each attribute in the heart disease dataset.*

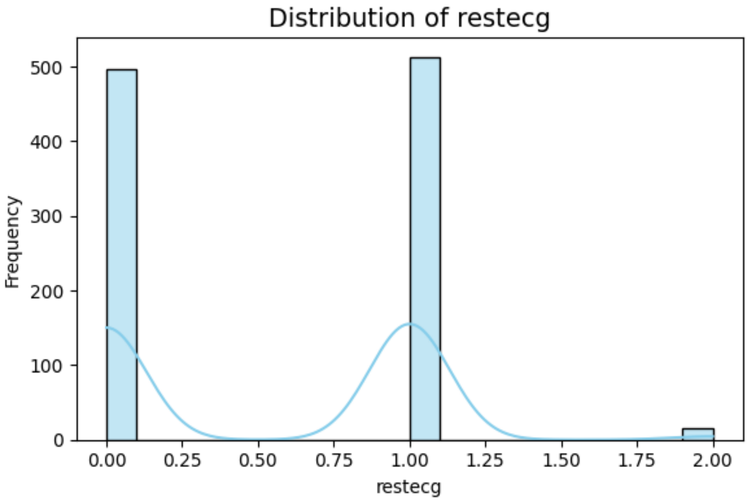
Features	Data Type	Missing Values	Unique Values	Min	Max	Mean	Std Dev
age	int64	0	41	29.000000	77.000000	54.434146	9.072290
sex	int64	0	2	0.000000	1.000000	0.695610	0.460373
cp	int64	0	4	0.000000	3.000000	0.942439	1.029641
trestbps	int64	0	49	94.000000	200.000000	131.611707	17.516718
chol	int64	0	152	126.000000	564.000000	246.000000	51.592510
fbs	int64	0	2	0.000000	1.000000	0.149268	0.356527
restecg	int64	0	3	0.000000	2.000000	0.529756	0.527878
thalach	int64	0	91	71.000000	202.000000	149.114146	23.005724

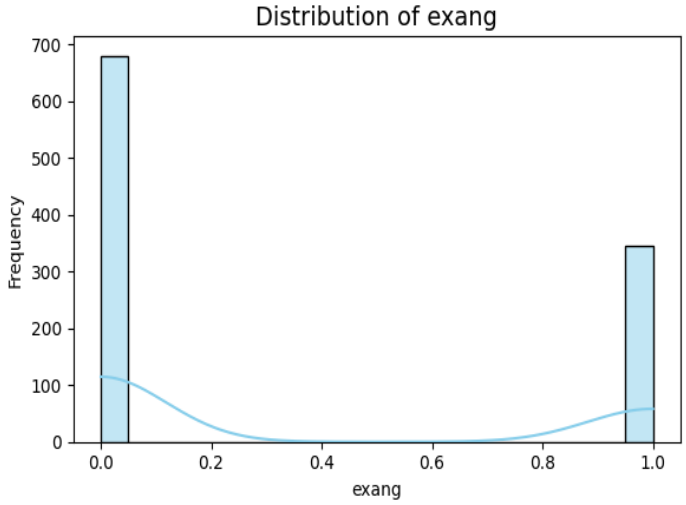
exang	int64	0	2	0.000000	1.000000	0.336585	0.472772
oldpeak	float64	0	40	0.000000	6.200000	1.071512	1.175053
slope	int64	0	3	0.000000	2.000000	1.385366	0.617755
ca	int64	0	5	0.000000	4.000000	0.754146	1.030798
thal	int64	0	4	0.000000	3.000000	2.323902	0.620660
target	int64	0	2	0.000000	1.000000	0.513171	0.500070

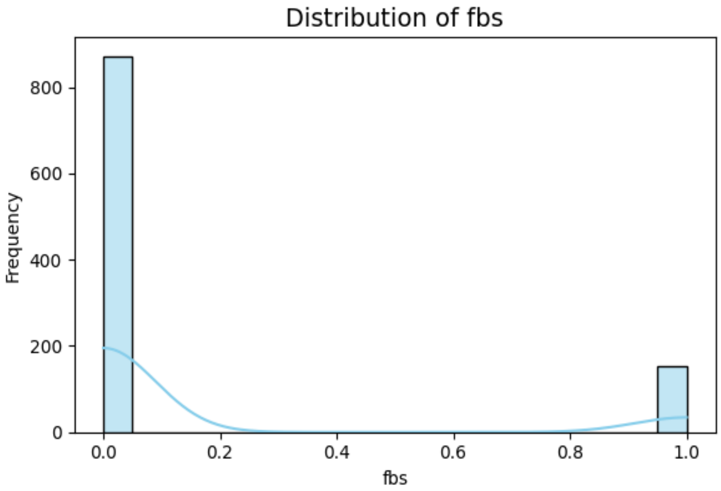
The graphs 3-2 illustrate the distribution of numerical and categorical features in the data set, which helps in understanding the demographic and clinical composition of the samples. These graphs support descriptive textual analysis and provide a clear visual representation of the data set before entering it into the modelling process, which enhances the researcher’s understanding of the basic characteristics and patterns present in the data.

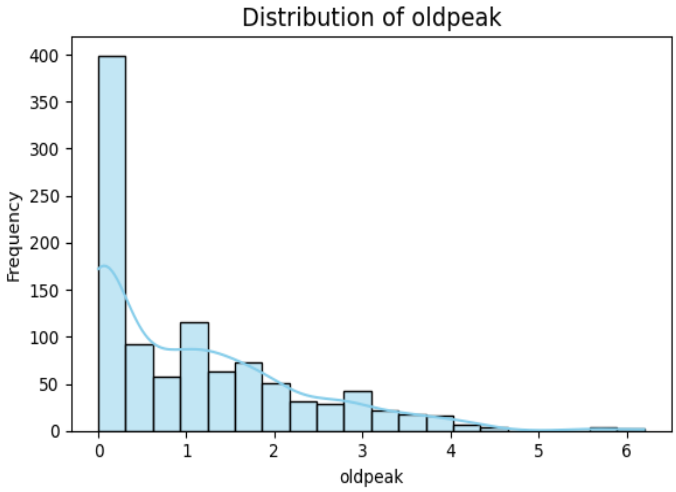


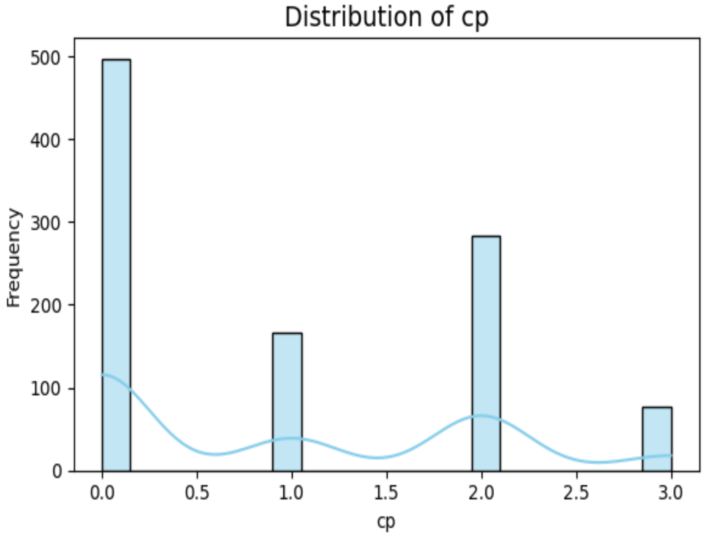


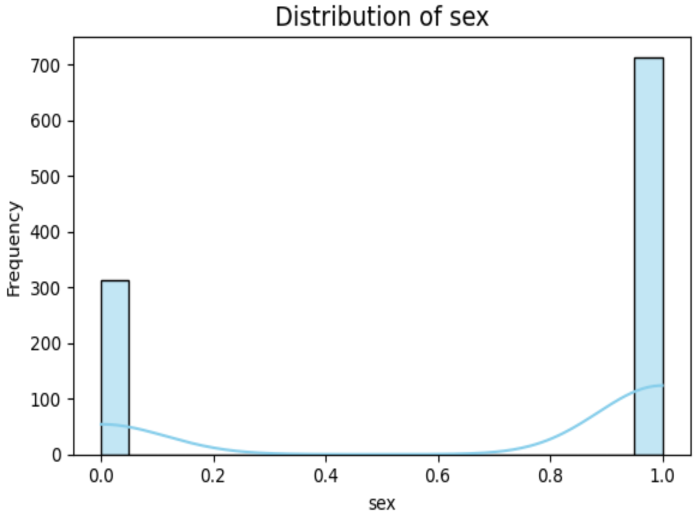


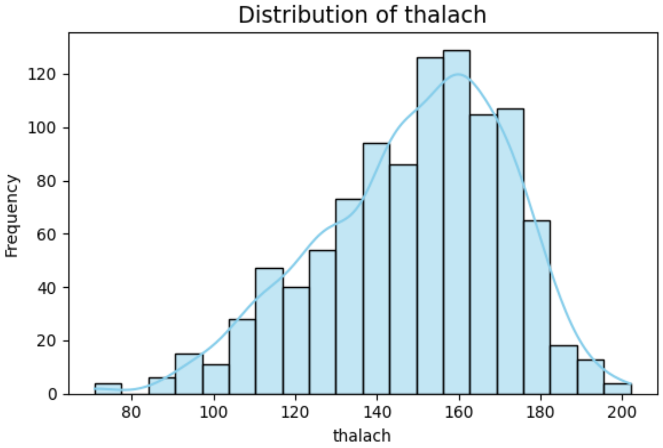


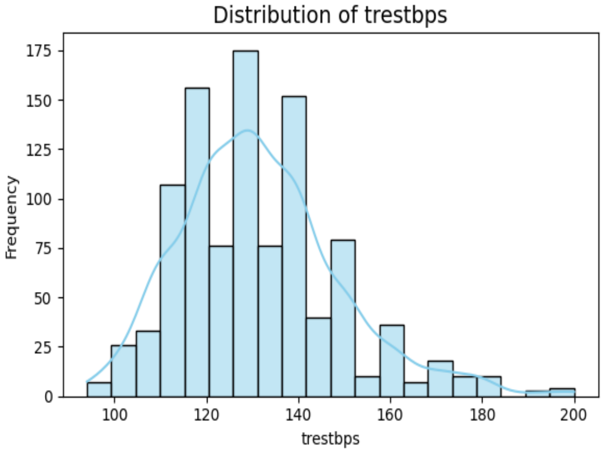

Distribution of fbs
Frequency
0
200
400
600
800
0.0
0.2
0.4
0.6
0.8
1.0
fbs


Distribution of oldpeak
Frequency
0
50
100
150
200
250
300
350
400
0
1
2
3
4
5
6
oldpeak


Distribution of cp
Frequency
0
100
200
300
400
500
0.0
0.5
1.0
1.5
2.0
2.5
3.0
cp


Distribution of sex
Frequency
0
100
200
300
400
500
600
700
0.0
0.2
0.4
0.6
0.8
1.0
sex


Distribution of thalach
Frequency
0
20
40
60
80
100
120
80
100
120
140
160
180
200
thalach


Distribution of trestbps
Frequency
0
25
50
75
100
125
150
175
100
120
140
160
180
200
trestbps

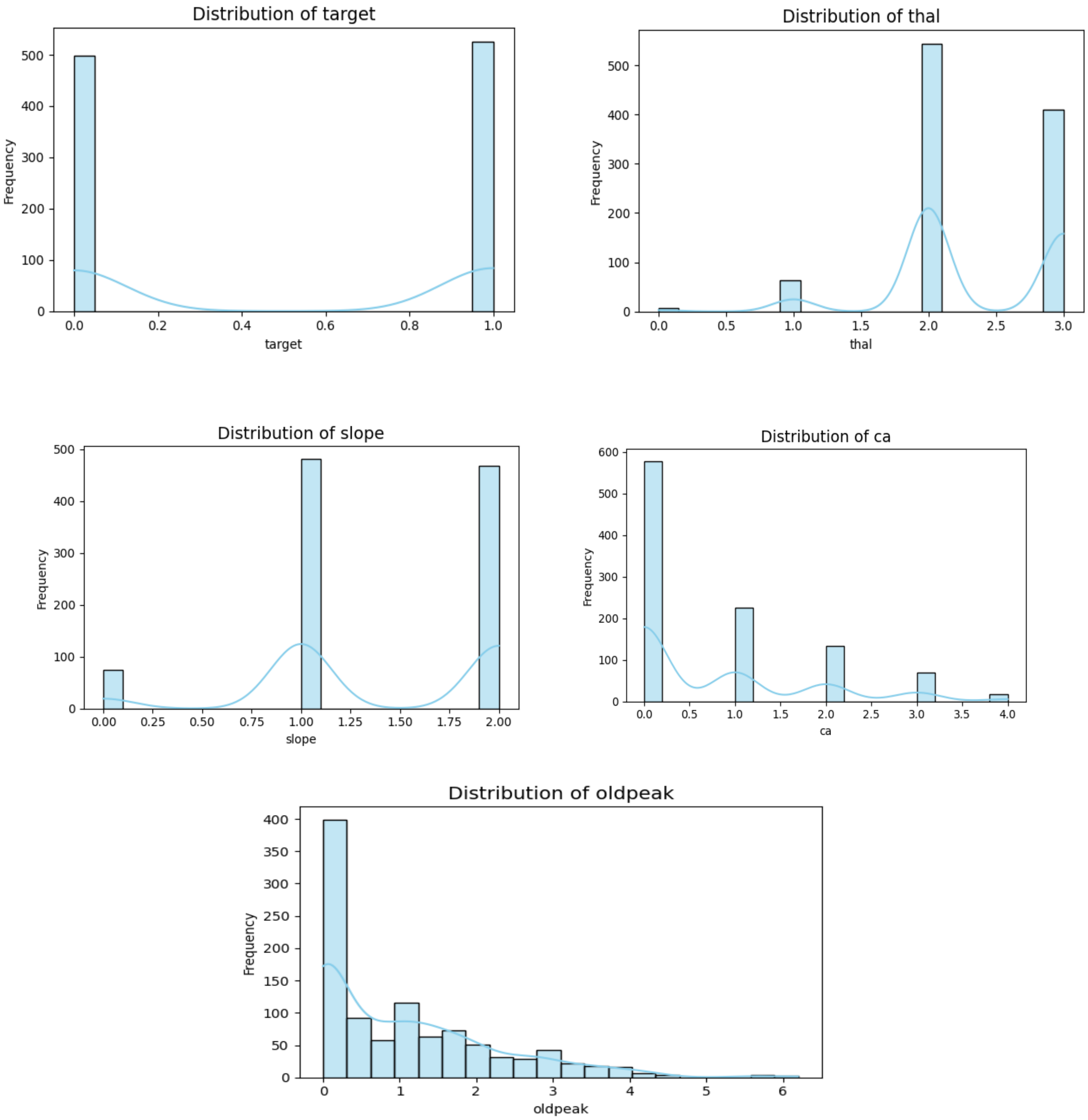
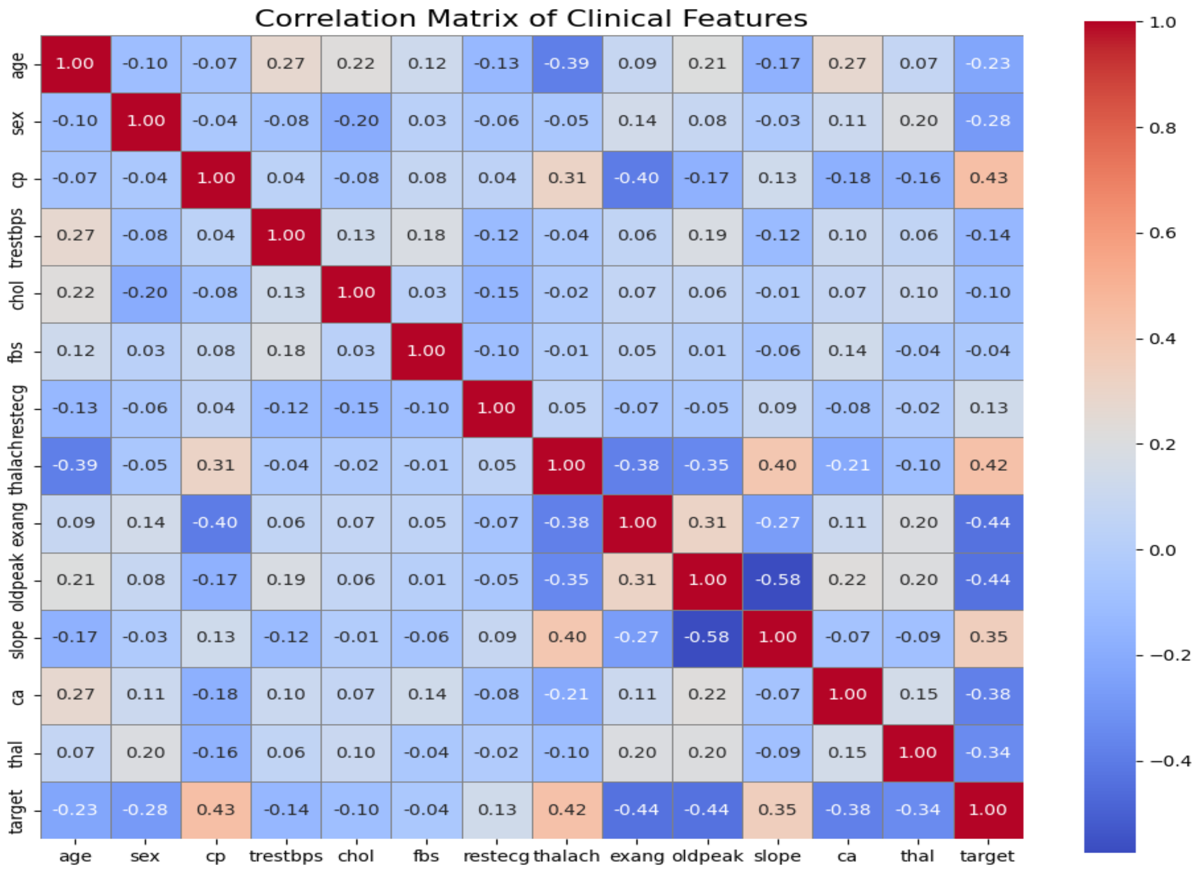


Figure 3-2 ***:*** *Charts showing the distribution of numerical and categorical attributes in a dataset*

3.2.4 Exploratory data analysis and correlation matrix between features

A comprehensive data study was conducted to understand the characteristics of the clinical dataset and analyze the relationships between different features before training the machine and deep learning models. This is part of exploratory data analysis, and the goal of this procedure is to verify the quality of the data, identify duplicate or outlier values, and provide insight into the degree of correlation between features to ensure that each feature adds unique informational value to the model. A correlation matrix was calculated for all clinical features in the dataset using Pearson's correlation coefficient. This analysis displays the linear relationship between each pair of features, with values ranging from -1 to 1. The results of the analysis are presented as a color heat map for easy interpretation. Darker colors indicate high correlations, while lighter values indicate lower correlations. The results of the correlation matrix showed that all features in the dataset fall below the high correlation threshold $|r| > 0.8$. For example, features such as chol and tracebps showed only moderate correlation, meaning that each feature provides independent and distinct information to the model. The results also showed that the dataset contained no outliers that warranted removal or correction, as all features were within the expected normal range according to descriptive statistics. As shown in Figure 3-3

As a result, no steps were taken to remove highly correlated features or correct outliers, as each feature provides unique informational value and does not raise any red flags regarding overfitting.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.00	-0.10	-0.07	0.27	0.22	0.12	-0.13	-0.39	0.09	0.21	-0.17	0.27	0.07	-0.23
sex	-0.10	1.00	-0.04	-0.08	-0.20	0.03	-0.06	-0.05	0.14	0.08	-0.03	0.11	0.20	-0.28
cp	-0.07	-0.04	1.00	0.04	-0.08	0.08	0.04	0.31	-0.40	-0.17	0.13	-0.18	-0.16	0.43
trestbps	0.27	-0.08	0.04	1.00	0.13	0.18	-0.12	-0.04	0.06	0.19	-0.12	0.10	0.06	-0.14
chol	0.22	-0.20	-0.08	0.13	1.00	0.03	-0.15	-0.02	0.07	0.06	-0.01	0.07	0.10	-0.10
fbs	0.12	0.03	0.08	0.18	0.03	1.00	-0.10	-0.01	0.05	0.01	-0.06	0.14	-0.04	-0.04
restecg	-0.13	-0.06	0.04	-0.12	-0.15	-0.10	1.00	0.05	-0.07	-0.05	0.09	-0.08	-0.02	0.13
thalach	-0.39	-0.05	0.31	-0.04	-0.02	-0.01	0.05	1.00	-0.38	-0.35	0.40	-0.21	-0.10	0.42
exang	0.09	0.14	-0.40	0.06	0.07	0.05	-0.07	-0.38	1.00	0.31	-0.27	0.11	0.20	-0.44
oldpeak	0.21	0.08	-0.17	0.19	0.06	0.01	-0.05	-0.35	0.31	1.00	-0.58	0.22	0.20	-0.44
slope	-0.17	-0.03	0.13	-0.12	-0.01	-0.06	0.09	0.40	-0.27	-0.58	1.00	-0.07	-0.09	0.35
ca	0.27	0.11	-0.18	0.10	0.07	0.14	-0.08	-0.21	0.11	0.22	-0.07	1.00	0.15	-0.38
thal	0.07	0.20	-0.16	0.06	0.10	-0.04	-0.02	-0.10	0.20	0.20	-0.09	0.15	1.00	-0.34
target	-0.23	-0.28	0.43	-0.14	-0.10	-0.04	0.13	0.42	-0.44	-0.44	0.35	-0.38	-0.34	1.00

Figure 3-3 : *Correlation Matrix of Clinical Features*

3.3 Data Pre-Processing

After completing the first step (data collection), a set of data preprocessing steps were implemented. This stage is essential for building effective machine learning models, as it aims to prepare the data in an appropriate format that enables the models to learn accurately and efficiently. Since data quality directly affects model performance and accuracy, this stage focuses on removing biases and improving feature distribution to ensure sample balance. This forms a solid foundation for the subsequent stages of feature extraction and classification. In this study, a series of systematic procedures were implemented to clinically process the data before entering it into various classification models, with the goal of achieving more accurate prediction of heart disease cases .

As part of the initial data quality check, we checked for missing values in the data using the Pandas isnull().sum() function. The results showed that the data did not contain any missing values in numerical or categorical features, so there was no need to use any imputation techniques. The same applies to categorical features such as patient gender, chest pain type, and thal, which were coded numerically in the original dataset. The source used direct numerical coding for categorical values (such as 0 and 1 for binary variables, or 0 to 3 for multicategorical variables). Therefore, there was no need to use additional encoding techniques such as One-Hot Encoding or Label Encoding, as they are ready for use by machine learning algorithms from the source. This data readiness enhances the reliability of the preprocessing process and contributes to increased reproducibility, as the dataset is available in its final form to the public and is pre-formatted for research studies in this field.

3.3.1 Steps

The steps for data preprocessing and preparation applied to the dataset are described in detail in this section, highlighting the actions taken to ensure data quality, consistency, and suitability for subsequent analysis.

3.3.1.1 ***Separate the features from the target feature***

A series of data preprocessing steps were implemented to ensure the data was ready for training machine learning models (SVM, Random Forest, XGBoost, LightGBM, Logistic Regression). Commonly used classifiers such as K-Nearest Neighbours (KNN), Decision Tree, and Naive Bayes were not included in this study due to methodological considerations rather than oversight. KNN is highly sensitive to feature dimensionality and data scaling, which may reduce its effectiveness when applied to hybrid feature spaces combining original and deep features. Decision Trees, while interpretable, are more prone to overfitting when used independently on limited medical datasets. Naive Bayes relies on strong independence assumptions between features, which are often violated in clinical data containing correlated medical attributes. Therefore, this study focused on classifiers that are more robust to feature interactions and capable of handling complex, high-dimensional representations, making them more suitable for the proposed hybrid modelling framework. The first of these steps was to separate the independent features from the target feature from the database. The independent features represent the set of clinical values for each patient, such as blood pressure, gender, age, cholesterol, and other indicators that may be associated with heart disease. The target feature is a binary variable that represents the patient's health status: a value of (1) indicates the presence of a disease, while a value of (0) indicates the absence of a disease. The dataset was loaded into colab using the pandas library in Python, which provides efficient structures for handling tabular data. The target variable representing heart disease was separated from the 13 clinical features to begin data preprocessing and model training. Specifically, the features are stored in a DataFrame X and the target in a DataFrame y using X = data.iloc[:, :-1].values y = data.iloc[:, -1].values . This ensures that preprocessing steps such as feature encoding and normalization are applied to the data while keeping the target values unchanged, facilitating accurate model training. This separation allows the model to focus on learning the patterns and relationships inherent in the independent features to predict the target value. This step is essential in any supervised learning experiment, where the model is trained on a set of inputs with known outcomes and then evaluated on new data. This step helps models learn from the underlying patterns in the features to predict the correct outcome for a person.

A deep neural network (DNN) was built and trained to learn high-level data representations. The penultimate hidden layer was used to extract deep features reflecting complex nonlinear relationships between variables. These extracted features were then combined with the original features to form a hybrid feature set that combines direct clinical information with automatically extracted patterns. Finally, the resulting feature set was used to train and evaluate several conventional classification algorithms, relying on cross-validation and multiple performance indicators to ensure a fair and reliable assessment of model performance.

3.3.1.2 *Splitting the data into a training and test set*

Since the data used in the study is relatively balanced, with the positive category representing 51.32% and the negative category 48.68% of the sample, this indicates that the dataset does not suffer from significant skewness in the distribution of categories. In this case, random partitioning was employed because of its ability to maintain category proportions close to the original data distribution. Furthermore, random partitioning reflects a more realistic scenario, as the data received by the medical system in practical applications may be unbalanced or randomly distributed. The Scikit-learn library was used to randomly partition the data into two sets, a training set and a test set, with a random state of 80% and 20%, respectively, with the random state set to 42 to ensure the reproducibility of the results. This approach offers a significant advantage: testing the model's robustness under less-than-ideal conditions and examining its stability when category proportions change. This increases the model's generalizability to new data that may differ from the training data. Therefore, relying on random partitioning contributes to providing an evaluation environment closer to real-world usage conditions, which enhances the credibility of the results and increases the reliability of the developed models.

3.3.1.2 *Standardization using the StandardScaler algorithm*

Standardization is an essential part of the preprocessing steps in our study. Standardization was applied to all numerical features using the StandardScaler algorithm available in the Scikit-learn library. This process aims to transform the features into a standard normal distribution, where each feature has a mean of 0 and a standard deviation of 1. This procedure is extremely important when dealing with machine learning models, especially those sensitive to differences in numerical scales, such as the Support Vector Machine (SVM) and Logistic Regression. Values with a wide range can unevenly affect the model's performance and cause learning bias. Standardization reduces the variance between different features, which enhances the stability of the training process and accelerates the convergence of numerical optimization algorithms such as Gradient Descent. It also helps prevent the model from assigning unjustified weights to features with high numerical values simply because they have a larger scale. Several studies have confirmed that this technique is a standard practice when working with algorithms based on distances or probability ratios, due to its positive impact on the accuracy and predictive efficiency of the model (Saba et al., 2021; Razzak et al., 2022).

The StandardScaler algorithm works according to the following equation:

$$\frac{x-\mu}{\sigma}=z$$

3.1)

X : The original value of the item

μ : feature mean

σStandard deviation of the feature:

Let's assume we have a set of values for a single property X = [10, 12, 14, 16, 18]. Applying a standardized measurement to these values is done through the following :steps

1) Calculating the mean (μ) = (510+12+14+16+18)/5=14

2) Calculating the standard deviation (σ) $\sqrt{\frac{(10-14)^2+(12-14)^2+(14-14)^2+(16-14)^2+(18-14)^2}{5}}$

≈2.828

3) Converting values using StandardScaler = $\frac{x-\mu}{\sigma}$

:For the value 10 , = z $\frac{10-14}{2.828}$ ≈−1.414

For the value 12: , = z $\frac{12-14}{2.828}$ ≈−0.707

For the value 14: , = z $\frac{14-14}{2.828}$ = 0

For the value 16: , = z $\frac{16-14}{2.828}$ ≈0.707

For the value 18: , = z $\frac{18-14}{2.828}$ ≈1.414

After applying the equation, each feature has a mean of 0 and a standard deviation of 1 after the transformation process. The training data was initially standardized using the fit_transform() function to adjust the mean and standard deviation values. The same transformation parameters were then applied to the test data using the transform() function to ensure that the model handles new data within the same numerical range.

The three procedures implemented in this stage—feature separation from the target feature, data stratification, and standardization—represent the cornerstone of data preparation for building effective machine learning models. These processes help address structural differences in clinical data and reduce bias resulting from variations in numerical metrics between features. This enhances the accuracy and stability of models during training. This stage represents the necessary methodological introduction before moving to more advanced stages such as extracting non-linear and highly representative features using DNN neural networks and then training traditional classification models on the resulting rich representations, which will be discussed in detail in the next section of this methodology.The tools used in this study will be discussed in the next section.

3.3.2 Tools

In this study, a set of advanced Python libraries were used to process data, build models, and then evaluate performance. These tools are essential for ensuring accurate and smooth implementation, and for easy reproducibility and interpretation of results. These libraries include:

1-Pandas:Used to load data from an Excel file, explore it, and organize it into clear data frames. It also facilitated the separation of independent features from the target feature, contributing to efficient data processing before beginning the data consolidation and modelling phases.

2- NumPy: Used to perform numerical and matrix operations on the data, especially when the data was converted to numerical matrices for neural network training. It also facilitates the process of accurately and quickly combining the original features with features extracted from the penultimate layer of a deep network (DNN).

3- TensorFlow/Keras: Used to build and scale a multi-layer deep neural network model. Keras was also used to create a feature extraction layer from the penultimate layer, generating deep data features that can be combined with the original features to increase the predictive power of traditional machine learning models.

4- scikit-learn: This library helped split data into training and test sets (train/test split). It also contributed to data standardization using StandardScaler and built classic models such as SVM, Random Forest, and Logistic Regression. It was also used to perform K-Fold cross-validation and calculate precise performance metrics such as precision, recall, positive precision, F1-score, and AUC.

5- XGBoost and LightGBM: Used to build powerful boosting models for processing complex data, resulting in improved predictive performance compared to traditional classical models.

6- Warnings: Used to avoid unnecessary warnings during execution, helping to speed up the analysis process and make it smoother and clearer.

These tools represent the software foundation for all processing, analysis, and modelling steps in this study. They enable the researcher to implement all procedures accurately and efficiently, ensuring the possibility of repeating the experiment and achieving reliable and more accurate results in predicting heart disease. The following section will explain the impact of these libraries and tools on the data.

3.3.3 Impact of Preprocessing on the Dataset

After completing the data collection step, a set of preprocessing procedures was implemented to ensure the data was ready for the machine learning model training phase. This step directly impacts data representation, training quality, and model stability.

First, separating the independent features from the target column helped clarify the relationship between each patient's clinical values and the target outcome. This made it possible to focus the models on identifying internal patterns within the independent features without introducing any noise into the target column. This separation is essential for learning patterns in a controlled manner, which contributes to increasing the models' ability to accurately predict a patient's condition. Second, random partitioning was employed because of its ability to maintain category proportions close to the original data distribution. Furthermore, random partitioning reflects a more realistic scenario, as the data received by the medical system in practical applications may be unbalanced or randomly distribut . Third, standardizing numerical features using StandardScaler played a significant role in transforming all features into a standard normal distribution with a mean of zero and a standard deviation of one. This standardization reduced the influence of features with smaller values, such as electrocardiogram (ECG), preventing the model from being biased toward larger features. For example, non-uniform features can lead to overweighting of features with larger ranges in algorithms like SVM and Logistic Regression. The standardization process helps the model treat all features fairly and equally, improving training stability and the convergence speed of the Gradient Descent algorithm , Figure 3-4 and Figure 3-5 illustrate the effect of these operations on the data.. As a result of these steps, our dataset became more representative of the true patterns of diseases, while maintaining the total number of samples (1,025) without losing any information. It also ensured a natural class balance (51% disease versus 49% non-disease). This preprocessing formed a solid foundation for the next stage of feature extraction, increasing the model's ability to learn from internal patterns in the data accurately and efficiently.

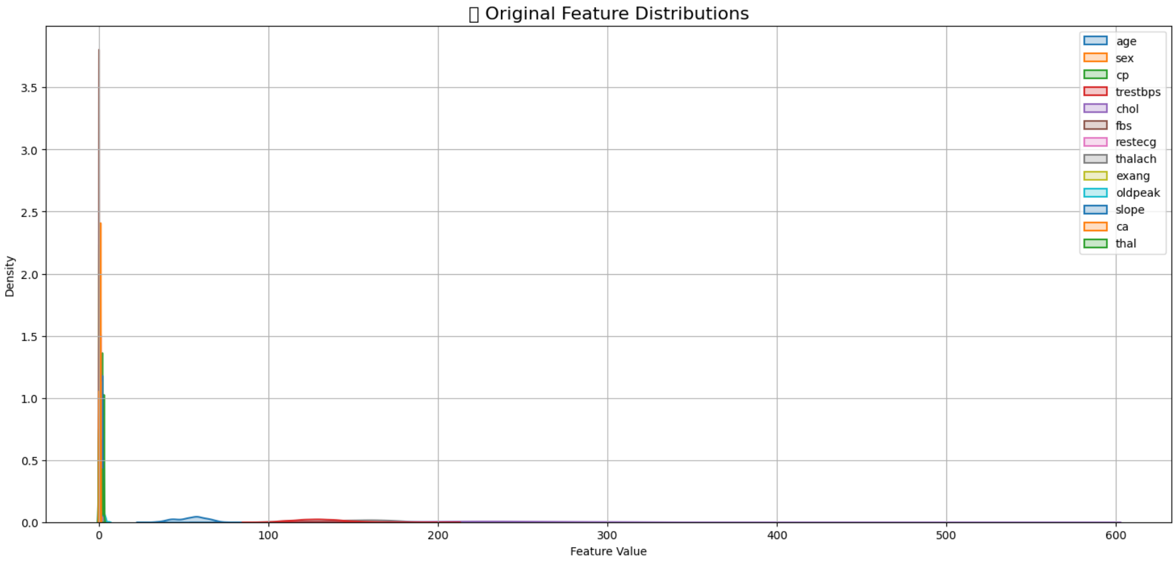


Figure 3-4 : *Original Feature Distributions*

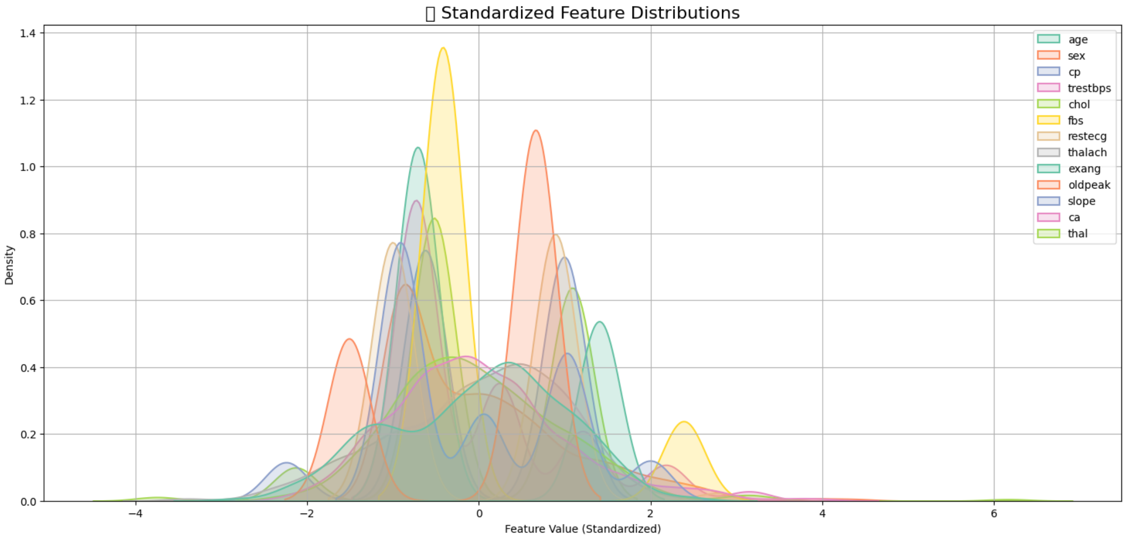


Figure 3-5 : Standardized Feature Distributions

The results shown in the image above were obtained by applying the Standardization technique to the medical dataset used in the study. This was done to address the significant variability in measurements between different characteristics. For example, the standard unit for the characteristic 'cholesterol' differs numerically enormously from 'patient's sex' or 'age'. This method was chosen to prevent the dominance of characteristics with high values in the prediction process. The mathematical operation was performed by subtracting the arithmetic mean from each value and then dividing by the standard deviation of each column. This transforms all distributions to center around the mean 0 and the standard deviation 1. The success of this operation is shown in the graph through the overlap of the data density curves (KDE) and their centering around the zero point on the horizontal axis, while preserving the unique characteristics of each distribution (such as skewing and outliers) that are essential for medical diagnosis, thus accelerating the model convergence process.

3.4 Data processing

This phase focused on examining and validating the data to ensure its quality before training the deep neural network (DNN) model and traditional machine learning models. Statistical analysis results showed that all features were numeric, and no missing values were recorded, indicating the cleanliness and stability of the data. Furthermore, descriptive statistics demonstrated the data's balanced distribution and spread, including the minimum and maximum limits, means, and standard deviations for each feature. These results formed a reliable basis for moving to the training phase without the need for additional processing such as compensating for missing values or verifying data types.

While training the model on this original data (before feature consolidation), the ReLU (Rectified Linear Unit) activation function was used in the hidden layers of the network. ReLU is one of the most widely used activation functions in modern neural networks due to its computational simplicity and effectiveness in accelerating the training process. ReLU is characterized by its ability to introduce nonlinearity into the model without significantly impacting computational efficiency. It also helps address the vanishing gradient problem that traditional activation functions such as sigmoid and tanh suffer from, contributing to faster convergence of the model toward an optimal solution. It also improves model response by discarding negative values, enhancing the representation of important features and reducing the impact of insignificant noise. This makes it ideal for use in models dealing with nonlinear data, such as medical data.

In addition, the Adam Optimization algorithm was used during training. This is an efficient weight update algorithm based on adaptive derivative pooling and balances algorithms such as AdaGrad and RMSProp, providing fast and consistent performance, especially in cases involving high-dimensional or complex data structures. This phase helped ensure that the deep learning model and machine learning models were built on sound data, using effective training and optimization tools, Paving the way for initial model training will be discussed in the next section.

3.5 Initial Model Training

The initial model training phase is a critical and pivotal step in the proposed framework. During this phase, a selection of traditional machine learning algorithms are tested on raw data before applying any feature extraction or fusion processes. This procedure aims to establish a performance baseline against which the results can be later compared to the results of models after feature extraction using deep learning techniques. This early evaluation allows us to understand the response of each algorithm to the nature of the data in its initial state and identify its strengths and weaknesses. This contributes to the design of informed improvement strategies across subsequent phases to arrive at well-founded results. The algorithms used will be presented in the next section.

3.5.1 Algorithms

Given the importance of model diversification during the initial evaluation phases, a set of algorithms representing different categories of machine learning methods was selected. In this step, a set of machine learning algorithms is subjected to initial training using the original features. The goal is to gain a deeper understanding of the data behavior and determine the baseline ability of each algorithm to predict heart disease before applying feature refinements. This phase provides a reliable starting point and a basic benchmark for evaluating model performance, paving the way for a continuous improvement process that ultimately leads to the development of an effective and accurate predictive system.

3.5.1.1 *Random Forest (RF)*

The Random Forest algorithm is an ensemble model that creates a large number of decision trees, each of which is then trained on random subsamples of the data using bootstrap sampling, with a limited set of features randomly selected at each split. This method ensures tree diversity and reduces correlation between trees, reducing the likelihood of overfitting that can occur if a single tree is used. The depth of the trees (max depth) and the number of features selected at each split (max features) are controlled, and a large number of trees (n_estimators) are used to ensure predictive stability. These combined mechanisms help improve the generalization ability of the model and make its performance more balanced when dealing with unseen data. To build the model, the library's default values were used, with the number of trees (n_estimators) set to 100. These values are sufficient to reduce variance and improve model stability without significantly increasing computational cost. The maximum tree depth (max_depth) was set to None, which contributes to the full growth of each tree and learning from complex patterns in the data. It relies on internal sample splitting to avoid over-learning on specific data. The random_state parameter value was set to 42 to ensure consistent results every time. K-Fold Cross-Validation (k=5) with shuffle=True was also used to evaluate models on the training set, which helps reduce overfitting and increase predictive accuracy while ensuring a balanced distribution of classes. This means that the data split and training will be consistent across repeated experiments.The rationale for using this algorithm is that it is suitable for handling multi-attribute data and contributes to reducing overfitting. It also features easy interpretation of results, flexibility, and high performance, thanks to its ability to handle nonlinear and complex data. It is also reproducible and usable in research because the Scikit-learn RandomForestClassifier library is popular and reliable, ensuring the model can be reapplied to the same or similar data with high accuracy, while specifying random_state to ensure consistent results every time.

3.5.1.2 ***Support Vector Machine (SVM)***

This algorithm is considered a powerful algorithm for dealing with high-dimensional problems, as it works to find the optimal hyperplane that separates classes by the largest possible margin. The kernel is used to increase the ability to discriminate in nonlinear data, making it suitable for cases of overlap between features. During initial training, the SVM provides a clear view of the linearity or nonlinearity of the problem at hand, which subsequently guides the researcher in selecting the appropriate kernel and tuning its parameters.Operators used: kernel='rbf to implement the nonlinear radial base function kernel To help the model handle complex relationships between attributes, probability=True to calculate class probabilities this is necessary for evaluating the AUC., and random_state=42 to ensure reproducibility K-Fold Cross-Validation (k=5) with shuffle=True was also used to evaluate models on the training set, which helps reduce overfitting and increase predictive accuracy while ensuring a balanced distribution of classes. , The remaining parameters used the default values approved by the Scikit-learn library to avoid excessive complexity, with the possibility of further optimization in advanced parameter tuning .

The rationale for using this algorithm is that it is effective in binary classification and that it can handle multidimensional features, meaning it is suitable for a dataset containing 13 clinical features. It also achieves a good balance between performance and accuracy without the need for extensive processing of missing data or coding, especially since our data is ready for use.

3.5.1.3 ***Extreme Gradient Boosting (XGBoost)***

XGBoost is one of the most powerful ensemble learning (Boosting) models. It relies on building trees sequentially, with each new tree attempting to correct the errors made by the previous tree. It is characterized by high speed and efficiency in handling large data sets, in addition to the ability to precisely control the learning process through a large number of parameters. At this stage, XGBoost helps discover the ability of boosted models to improve performance.The model was trained using XGBClassifier from the XGBoost library, a trusted library in medical classification studies, facilitating replication and validation by other researchers. The library's default values were set at 100 trees (n_estimators), 0.3 learning_rate, and 6 tree depths. Random_state = 42 was set to ensure reproducibility.K-Fold Cross-Validation (k=5) with shuffle=True was also used to evaluate models on the training set, which helps reduce overfitting and increase predictive accuracy while ensuring a balanced distribution of classes. The algorithm is well-suited for this dataset because it efficiently handles multidimensional features and reduces the risk of overfitting compared to some traditional machine learning algorithms. It also allows for future performance improvements by tuning parameters such as increasing the number of trees or decreasing the learning rate. The library's default values were used without modification, providing a clear starting point for experimental research.

3.5.1.4 ***Light Gradient Boosting Machine (LightGBM)***

LightGBM is an improved model based on Gradient Boosting, but it is specifically designed to handle large datasets with high memory and time efficiency. It also uses advanced techniques such as leaf-wise growth instead of level-wise, which speeds up the training process and increases accuracy in some cases. Including it in initial training helps evaluate the potential for accelerating future experiments without sacrificing performance, especially when moving to larger datasets after feature extraction and integration. The model was built using the LightGBM library (LGBMClassifier) in Python. The library's default values were num_leaves = 31, which is the maximum number of nodes in each tree, helping to achieve a balance between tree complexity and model performance. The learning_rate value was 0.1 to control the model's learning speed to ensure stability, which plays a role in improving the gradient and reducing overlearning. The random_state value was 42 to ensure the reproducibility of the results across repeated experiments , K-Fold Cross-Validation (k=5) with shuffle=True was also used to evaluate models on the training set, which helps reduce overfitting and increase predictive accuracy while ensuring a balanced distribution of classes. The rationale for using the LightGBM algorithm is its ability to handle non-linear and complex data, and its high speed and efficiency compared to traditional Gradient Boosting algorithms, which contributes to reducing training time and increasing speed. It also supports the gradient distribution of the tree and enhances performance on small and medium-sized data without over-learning. It also has the ability to interpret the model through the importance of features, which is essential and a fundamental pillar in medical studies.

3.5.1.5 ***Logistic Regression (LR)***

Logistic Regression (LR) is a simple and uniquely effective statistical model that relies on a logistic function to predict the probability that a sample falls into a particular class. Despite its simplicity, it provides a very strong baseline for comparing the performance of complex models, demonstrating how well a linear model can be achieved. At this stage, LR represents a starting point for determining whether the problem can be solved linearly or requires more complex models.

Performing initial training using these algorithms helps build a clear picture of data behavior and each model's ability to handle it before applying advanced feature refinements. The results of this phase also serve as a basic reference for comparing model performance before and after refinements, contributing to accurate, comprehensive, and clear generalizations.The model was built using the Scikit-learn (LogisticRegression) library in Python, with the actual values where C = 1.0 to control the strength of regularization to balance between the complexity of the model and its ability to generalize to new data, solver = lbfgs which is the optimization algorithm used to estimate the parameters and is efficient and suitable for small and medium sized data, and random_state = 42 to ensure the reproducibility of the same results when splitting the data or training the model , K-Fold Cross-Validation (k=5) with shuffle=True was also used to evaluate models on the training set, which helps reduce overfitting and increase predictive accuracy while ensuring a balanced distribution of classes . The next section will discuss hyperparameters.

3.5.2 *Hyperparameters*

At this stage of the study, initial training of the models was performed using a variety of traditional and popular machine learning algorithms (SVM, Random Forest, XGBoost, LightGBM, and Logistic Regression). To ensure optimal performance of these models, a set of basic hyperparameters were tuned for each model. These parameters play a vital role in controlling the model's behavior during the training process, determining the model's ability to generalize to new data, and avoiding overfitting. These values are based on the default settings in the Python libraries used, with some values being slightly modified as needed to fine-tune the initial learning of the models. These parameters represent the basic starting point before experimenting with optimizing them later in the advanced tuning stages. Table 3-3 shows the hyperparameters for each algorithm and the values used in the initial training code, along with an explanation of the purpose of each parameter and its impact on learning and the model:

Table 3-3 : *Shows the hyperparameters of each algorithm*

Algorithm	Hyperparameter	Value used in the code	Purpose of selection	Description
SVM	Kernel	rbf'	Dealing with nonlinear relationships between data	The RBF kernel enables the model to flexibly separate nonlinear data and improves predictive ability.
	C	1.0	Regulating error and balancing bias and variance	A neutral value prevents overfitting and maintains the model's ability to generalize.
Random Forest	n_estimators	100	Sufficient number of trees to reduce variance and increase model stability	Increasing the number of trees improves prediction accuracy and reduces the impact of noise on the model.
	max_depth	None	Allow the tree to fully grow while controlling overgrowth through bunching.	Allows the model to learn complex patterns without overfitting on individual data.

Algorithm	Hyperparameter	Value used in the code	Purpose of selection	Description
XGBoos	learning_rate	0.3	Control the speed of model learning	Low learning rate contributes to improved gradient convolution and reduced overfitting.
	n_estimators	100	Number of trees to provide a basis for performance evaluation	Determining the number of trees balances good learning and generalization ability.
LightGBM	learning_rate	0.1	Control learning speed and improve stability	A benchmark value that ensures that the model learns gradually without significant changes in each update.
	num_leaves	31	Specify the maximum number of nodes in each tree to balance performance and speed.	Help maintain moderate tree complexity and avoid over-specification.
LR	C	1.0	Organize the model and prevent overfitting.	A neutral value enhances the model's ability to generalize by reducing the weight of large coefficients.
	Penalty	l2	Reducing model complexity	L2 regularization reduces large parameter values and enhances generalization to new data.

In this research, we used default values for the hyperparameters associated with each algorithm, as described in approved libraries such as Scikit-learn, XGBoost, and LightGBM. These values were used to establish a baseline and a basic starting point for several methodological and scientific reasons, including:

- To provide a fair comparison between different models under the same training conditions. Using default values highlights the algorithm's ability to process data in its default state.
- The default values provided by approved libraries are the result of previous experiments and studies and have proven their stability and effectiveness for most types of data.
- The primary goal of this stage is to provide an initial evaluation of the overall performance of the models using default values without affecting the results by

adjusting the parameters, which might mask fundamental differences between the algorithms.

This step is essential for making a comparison, as it allows us to understand the natural behavior of each algorithm before making any future improvements to the parameters.

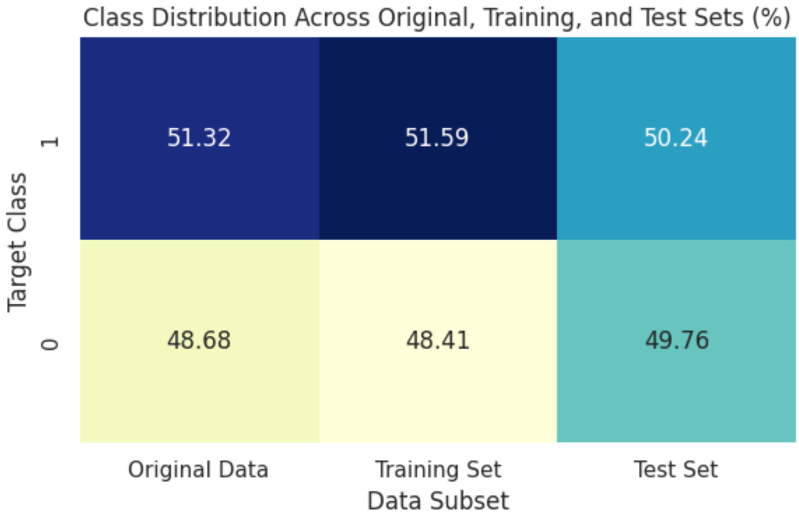
3.5.3 Training setup

The training setup phase is a pivotal stage in developing machine learning models. It determines how the data is divided, sets model validation methods, and systematically organizes the learning process. The goal of this phase is to provide a balanced training environment that helps models learn basic patterns in the data effectively and accurately, while reducing the possibility of overfitting or bias toward one class. The training setup includes key elements: determining the data split ratios between the training and testing , and applying cross-validation to evaluate model performance across all available data. These procedures establish a solid foundation for initial model training, enabling the subsequent stages of integrating extracted features and improving performance with scientific and reliable accuracy.

3.5.3.1 ***Train-Test Split***

After completing the initial data processing and separating the independent features from the target feature, the data was divided into two main sets: the training set and the test set. The training set, which comprised 80% of the total data, was used to teach models about the patterns and associations between clinical features and the presence of heart disease, while the test set, which comprised 20%, was used to evaluate the models' ability to predict new cases not seen during the training process. The 80/20 ratio was chosen as a balanced criterion that provides a sufficient amount of data to train the models, while also maintaining a test set that represents the overall data distribution. This choice ensures that the models are exposed to a sufficient and diverse set of examples, which reduces the likelihood of overfitting and maintains the representativeness of the true distribution of the target variable. To ensure that dividing the data into three groups did not affect the distribution of classes in the target variable, the proportions of each class in the original data, training set, and test set were calculated using the value_counts(normalize=True) * 100 function. The proportions for each class were as follows, as shown in Table 3-4.

Table 3-4 : *Class Distribution Across Original, Training, and Test Sets (%)*



The results in the figure show that the class proportions in both the training and test sets are very close to those of the original data, meaning that there was no significant deviation in the class distribution. This indicates that the process of randomly dividing the data without resorting to stratification maintained a balance between the classes. This distribution ensures that each group adequately represents the original data and does not cause bias in the model or the loss of any important information related to the class distribution.

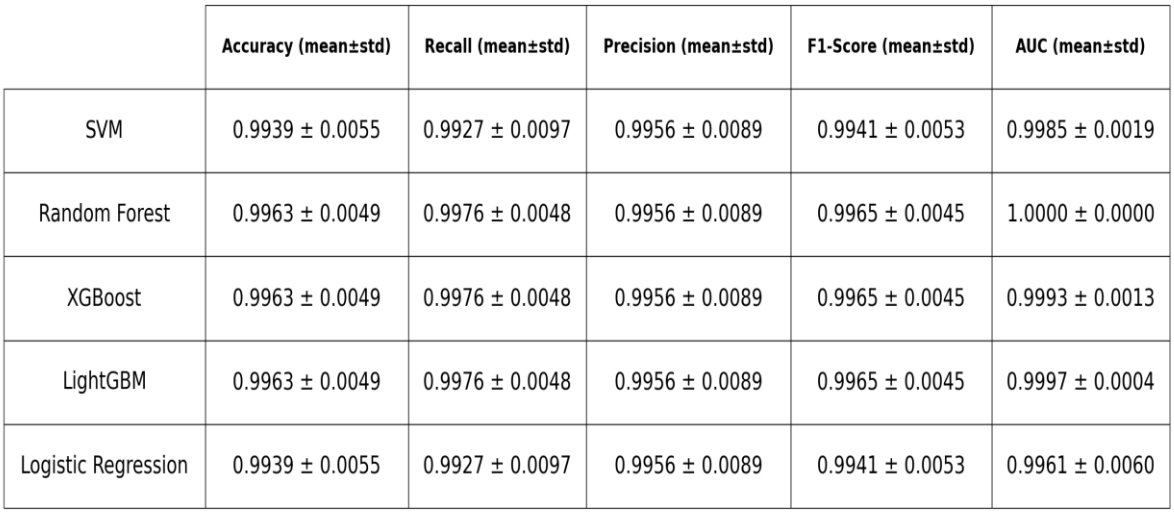
3.5.3.2 ***Cross-Validation***

To ensure higher accuracy in evaluating models before testing them on the training set, five-fold cross-validation was applied to the training set. This step divides the data into five equal parts, and the model is trained on four parts and tested on the fifth part, alternating between them, with each part used as test data once. This method allows for more reliable estimation of model performance than a single data split, as it detects any tendency for the model to overfit and reduces the random effects of the data distribution. It also ensures that the initial models, whether SVM, Random Forest, or other, are evaluated across the entire range of available data before combining the features extracted from the deep neural network.

The training setup in this study includes organized and precise steps to ensure that models are trained on balanced, high-quality data, while maintaining a fair representation of all classes. By systematically splitting the data, maintaining class balance, and applying cross-validation, we can accurately and reliably evaluate the performance of the initial models before moving on to the next stage of feature extraction and integration with the original features. Thus, this integrated setup paves the way for assessing the true impact of hyperparameters and features on the final performance of the models.

Using this procedure, the standard deviation and average of performance metrics such as recall, precision, precision, F1 measure, and AUC were calculated. This procedure contributes to more stable estimates of overall performance and reduces the impact of random data distribution on the final results. For example, In the study code, 5-Fold Cross-Validation was applied to both the original features and the extracted features (45 features) as shown in the evaluate_models function where the mean and standard deviation of each indicator were calculated for each model. As sown in Table 3-5.

Table 3-5 : *Cross-Validation Metrics (Mean ± Std)*



	Accuracy (mean±std)	**Recall (mean±std)**	**Precision (mean±std)**	**F1-Score (mean±std)**	**AUC (mean±std)**
SVM	0.9939 ± 0.0055	0.9927 ± 0.0097	0.9956 ± 0.0089	0.9941 ± 0.0053	0.9985 ± 0.0019
Random Forest	0.9963 ± 0.0049	0.9976 ± 0.0048	0.9956 ± 0.0089	0.9965 ± 0.0045	1.0000 ± 0.0000
XGBoost	0.9963 ± 0.0049	0.9976 ± 0.0048	0.9956 ± 0.0089	0.9965 ± 0.0045	0.9993 ± 0.0013
LightGBM	0.9963 ± 0.0049	0.9976 ± 0.0048	0.9956 ± 0.0089	0.9965 ± 0.0045	0.9997 ± 0.0004
Logistic Regression	0.9939 ± 0.0055	0.9927 ± 0.0097	0.9956 ± 0.0089	0.9941 ± 0.0053	0.9961 ± 0.0060

Table 3- 5 displays the standard deviation and average performance of the five models after five-fold cross-validation on the post-merging dataset. The results demonstrate the stability of the models, with the standard deviation being very small across all metrics (Accuracy, Recall, Precision, F1-Score, and AUC), reflecting that the models achieve consistent performance across different folds. It is also observed that all models, especially Random Forest, XGBoost, and LightGBM, achieved AUC values close to 1.0, reflecting their high ability to discriminate between classes. As a result of this stability, the urgent need for transaction adjustments is reduced due to the high performance and stability achieved by using default values.

3.6 DNN Feature Extraction

Feature extraction is a pivotal and fundamental step in analysing complex medical data, aiming to uncover hidden patterns that may be crucial for accurate disease diagnosis. In the context of cardiac disease, where biomarkers and clinical signs overlap in nonlinear ways, deep neural networks (DNNs) emerge as a highly effective method for extracting these complex features directly from raw data. This section aims to provide a detailed explanation of the motivations behind using DNNs for feature extraction, focusing on the network architecture, its outputs, and the details of the training process that ensures the extraction of information-rich representations to enhance the predictive ability of classical machine learning models.

3.6.1 DNN Architecture

The deep neural network architecture is fundamental to its ability to extract deep features from cardiac patient medical data. This network is designed to effectively address the complexities and nonlinearities inherent in clinical indicators. The deep neural network used in this study, as shown in Figure (5), begins with an input layer, followed by several hidden layers, and concludes with an output layer. This hierarchical structure enables the network to learn increasingly abstract representations of data. The input layer consists of 13 neurons, each corresponding to an original medical feature in the dataset (e.g., blood pressure, heart rate, etc.). This layer is the starting point for information flow through the network. These inputs are then transmitted to the hidden layers, which consist of three successive hidden layers of gradually decreasing sizes, enabling the network to learn hierarchical representations and effectively reduce dimensionality while retaining important information. The first hidden layer contains 128 neurons, the second hidden layer contains 64 neurons, and the third hidden layer contains 32 neurons. The Adam optimizer was used to effectively adjust the network weights, with a default learning rate, due to its ability to accelerate the training process and improve the stability of error reduction. The number of training epochs was set at 50 and the batch size was set at 32 to ensure efficient learning. 20% of the training data was allocated to validation (validation split = 0.2) to monitor performance and improve the model's ability to generalize to new data, which is critical when dealing with small datasets. The choice of this optimizer was crucial for training the deep neural network (DNN) used for feature extraction. The ReLU (Rectified Linear Unit) activation function was used in all hidden layers of neural networks. ReLU is known for its high efficiency in training deep neural networks due to its ability to handle nonlinear data problems and contribute to the convolution of the vanishing gradient problem by outputting a value of 0 for negative inputs and the same value for positive inputs. The outputs of each layer are calculated according to a linear relationship datasets. (Goodfellow, Bengio, & Courville, 2016).:

$B^{(l)}$ (1) + ${}^{(l-1)}w^{(l)}$ a = ${}^{(l)}z$

$$b^{(l)} \quad (1) + {}^{(l-1)}W^{(l)}a = {}^{(l)}z$$

(3.2)

where $z^{(l)}$ represents the linear sum before the activation process in layer l, $W^{(l)}$ is the weight matrix in layer l, $a^{(l-1)}$ is the output of the previous layer l −1 and $b^{(l)}$ is the bias vector in layer (l) . After that, the ReLU activation function is applied :

$$\max(0, z^{(l)}) \quad (2) = {}^{(l)}a$$

(3.3)

The output layer consists of a single neural unit that uses the Sigmoid activation function. The Sigmoid function is ideal for binary classification tasks because it outputs a value between (0, 1) that can be interpreted as the probability of contracting the disease. The output of this layer is calculated as follows:

$$\frac{1}{{}^{(L)}z^{-}e + 1} = \sigma(z^{(L)})$$

(3.4)

$\sigma(z^{(L)})$ represents the activation value resulting from applying a sigmoid function to $z^{(L)}$, and its value ranges between 0 and 1, making it suitable for representing probability.

The chosen network architecture provides a balance between complexity and generalization ability, enabling the network to extract deep and rich features while controlling overfitting and improving the model's ability to predict disease states in unseen data. Figure 3-6 shows the DNN working principle.

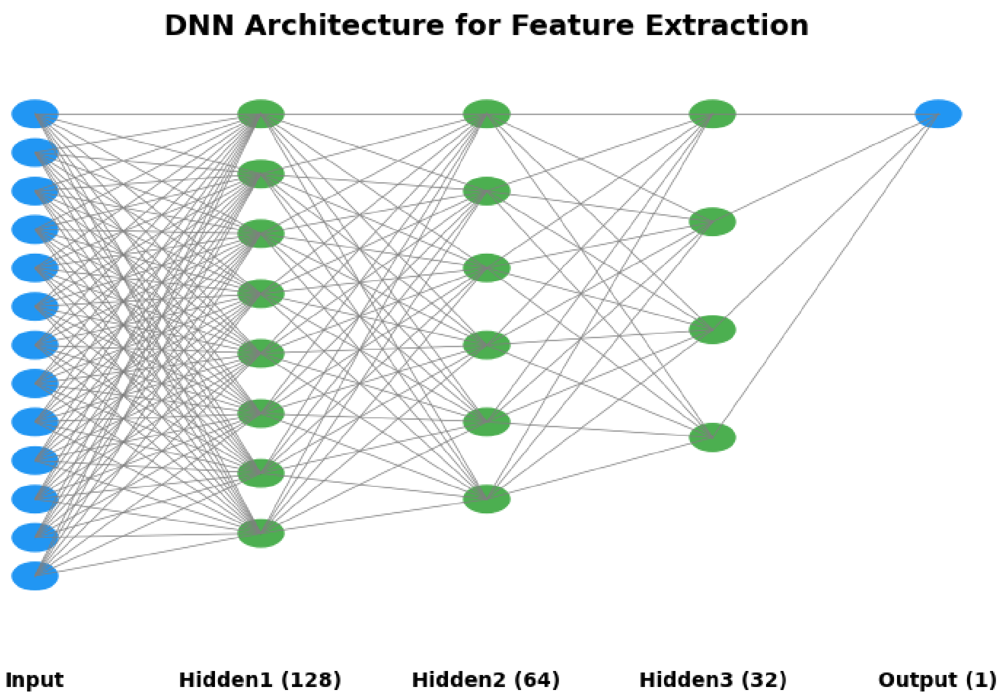


Figure 3-6 : *DNN working principle*

The size of the first layer (128 units) was chosen to accommodate the number of original features and provide sufficient space for the model to learn initial patterns. The next two layers (64 and 32 units) were designed to gradually decrease the dimensions, allowing the model to integrate information and extract the most distinctive features. Furthermore, the use of three hidden layers gives the model sufficient depth to handle nonlinear medical data without over-competing, which could weaken its generalizability.

This particular architecture was adopted after initial trials with several different configurations, which demonstrated that the 128–64–32 structure offers the best balance between computational performance and accuracy.

3.6.2 Output of Feature Extraction

After successfully training the deep neural network, the primary goal is to extract deep, high-level representations of the original data. In this study, the output of the feature extraction process is 32 new deep features. These features are not simply dimensionality reduction; rather, they are learned representations that capture the nonlinear and complex patterns present in the original data, which may not be readily apparent or detectable using traditional statistical methods or hand-engineered features. These extracted features are the result of transforming the original data through the hidden layers of the network, with each layer learning more abstract representations of the data. In this way, not only are existing patterns identified, but new insights are also generated by combining the information contained in the original features in complex ways, resulting in an expanded feature set containing richer information. These 32 features are extracted from the penultimate layer of the deep neural network. The selection of this layer is critical because, while the penultimate layer is typically used for classification, the layers preceding it often contain more general, information-rich representations and will be useful for other tasks such as classification using traditional learning models. These extracted features will then be combined with the original 13 features, bringing the total number of features used in subsequent training to 45, as shown in Figure 3-7. This feature fusion process aims to construct an expanded feature set that greatly enhances the ability of traditional models to distinguish between complex patterns and provide more accurate predictions of heart disease.

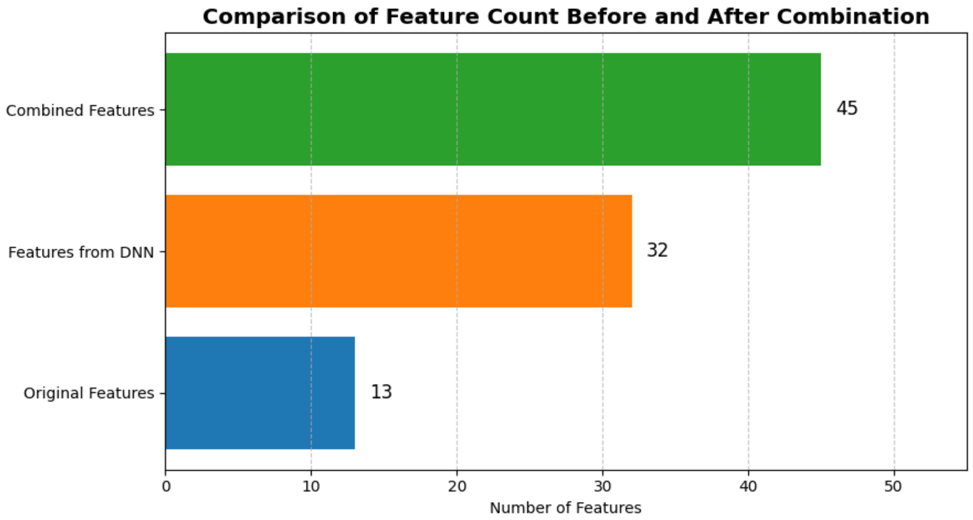


Figure 3-7 : *Comparison of Feature Count Before and After fusion*

3.6.3 DNN Training Details

To ensure the efficiency and effectiveness of the feature extraction process, the deep neural network was trained using a set of optimized parameters and training settings. The goal is to provide the necessary transparency to enable replication and validation of results. We will discuss the details of building the neural network.

1-Loss Function: Given the binary nature of the classification problem (determining whether a patient has heart disease or not), a binary_crossentropy loss function was used. This function is suitable for measuring the discrepancy between the predicted probability distribution of the model and the true distribution of the data, and is calculated according to the following equation:

$$[(\hat{_i}y - y_i)\log(1-1) + (\hat{_i}y)y_i\log]\sum_{i=1}^{N}\frac{1}{N} - = L \tag{3.5}$$

Where N is the number of samples, y_i is the true value of the classification (0,1), and $y_i^{\wedge}$ is the expected value (probability).

2- Optimizer: The Adaptive Moment Estimation (Adam) algorithm was chosen as the optimizer for training the deep neural network. This optimizer is known for its high efficiency and adaptability, combining the advantages of the RMSProp and AdaGrad algorithms. It is characterized by its ability to dynamically adjust the learning rate for each parameter, which contributes to accelerating the convergence process and overcoming the challenges associated with gradients in deep networks, especially when dealing with nonlinear relationships in medical data.

3- Learning Rate: The initial learning rate for Adam was set to 0.001. This is a relatively low value, allowing the model to converge slowly and more stably, while avoiding large fluctuations that could result in very high learning rates.

4-Epoches: The deep neural network was trained over 50 epochs. Each epoch represents a complete pass of the training data through the network. A portion of the training data was used for validation during the training process, allowing for monitoring of model performance and avoiding overfitting.

5- Batch Size: The batch size was set to 32 samples. The choice of this value depends on balancing training speed and gradient stability, as this size helps achieve a good balance between computational efficiency and weight update accuracy.

6- Computational Resources: The training process was performed using a The central processing unit (CPU) should be efficient enough for the nature of the data and the complexity of the model.

This fusion-based model addresses the overfitting problem on small datasets (such as the Cleveland dataset) in several ways. First, although deep neural networks can inherently overfit, their goal here is not final classification, but rather the extraction of high-level, abstract features. These extracted features represent condensed features of the original data, filtering out noise and strengthening underlying patterns. This makes the model less susceptible to overfitting when used later with traditional machine learning models. Second, the risk of overfitting in the deep neural network itself can be mitigated by a tight network architecture with a reasonable number of layers and neurons, which contributes to reducing the model's capacity and thus limits its ability to store training data. This approach significantly improves performance by enabling traditional machine learning models to operate on more comprehensive, richer, and discriminative data representations. Instead of relying solely on the original features, which may be linear or insufficient to capture complex relationships, the features extracted from the deep neural network provide a multidimensional perspective on the data. These high-level features enable traditional models to capture nonlinear patterns (as deep neural networks excel at learning nonlinear relationships that are difficult for machine learning algorithms to detect), reduce noise and redundancy (as the neural network filters out less important features and focuses only on the most essential patterns), and refine the representation, enabling new data representations that are more representative of the underlying data associated with heart disease. When deep features are combined with the original features, traditional models are provided with a comprehensive and robust dataset, significantly improving their prediction and classification capabilities. These details illustrate the solid foundation upon which the model is built. The next section will explain the methods of feature fusion.

3.7 Feature Fusion

The feature fusion phase is a pivotal step in building the study's hybrid model. This step aims to combine the analytical power of the original data with the abstraction capabilities of the features extracted from the deep neural network. This process not only increases the number of features, but also builds an accurate and comprehensive representation of the data, enhancing the accuracy of classical machine learning models in detecting complex patterns and making accurate predictions.

3.7.1 Fusion Mechanism

The feature concatenation process is straightforward and efficient, combining two feature sets into a single, expanded set. Technically, this process was implemented as follows:

1- Feature Extraction: In this step, a custom feature extraction model built from a DNN was used to extract 32 new deep features from each sample in the training and test data. .Using the keras.Model() + predict() function from the TensorFlow/Keras library

2- Combining the sets: The extracted features (now a two-dimensional matrix) were combined with the original features (consisting of 13 columns) on the column axis (axis = 1).. Using the np.concatenate(axis=1) function from the NumPy library

After the merging process is complete, a new training dataset (`X_train_combined`) = [X_train_scaled, train_features]and a new test dataset (`X_test_combined`) =[X_test_scaled, test_features] are produced. Each of these datasets now contains 45 features (13 original features + 32 deep features). This method ensures that all the initial information is retained while enriching it with the complex representations learned by the neural network.

3.7.2 Rationale and Improvements

The primary rationale behind building a hybrid model based on feature fusion is the belief that optimal model performance can be achieved by leveraging the strengths of both rule-based and data-driven models. First, to enhance performance, as feature fusion enhances the predictive power of traditional machine learning models. Instead of relying solely on raw features, which may be linear or insufficient, they are supplemented with high-level features that capture nonlinear relationships and complex patterns that are difficult for classical models to detect individually. Second, to overcome model limitations, as some algorithms suffer from limitations. For example, the LR algorithm may not be able to capture complex relationships, but when fed with the combined features, it can achieve accuracy and performance that exceed expectations. Third, to leverage both approaches, as feature fusion is the focus of the hybrid model, combining the neural network's deep feature learning capability with the predictive efficiency and relative interpretability of classical models.

3.8 Retraining and Evaluation

3.8.1 Retraining

The training phase of the classification algorithms in this model includes several key steps, starting with data segmentation, proceeding to model building and training, and ending with parameter tuning and performance evaluation. The data used in this phase consisted of a set of features extracted from the penultimate layer of the deep neural network model, which were combined with the original features to improve prediction accuracy. The dataset was split into a training and test set at a ratio of 80:20 using the train_test_split function from the scikit-learn library, ensuring random selection and balanced classes within each group.

This type of splitting helps ensure that the model learns from sufficient data, with 20% allocated to performance evaluation on previously unseen data, providing a fair and realistic measure of model performance in real-world applications. The classification phase using machine learning algorithms represents the second level of the proposed hybrid model, where the features extracted from the deep layer were used as inputs for traditional classification models. This approach leverages deep learning capabilities to extract high-level features and then uses efficient classical models such as SVM, Random Forest, XGBoost, LightGBM, and Logistic Regression. These traditional algorithms were specifically chosen for several fundamental reasons: they have a high capacity to handle clinical data with a large number of features, and they demonstrate stable and balanced performance even with nonlinear or imperfectly distributed data. Furthermore, these algorithms provide a balance between computational efficiency and accuracy, and allow for clearer interpretation of classification results compared to some more complex models, making them an ideal choice in medical fields that require high transparency and reliability in decision-making.

3.8.1.1 ***Support Vector Machines***

The Support Vector Machines (SVM) algorithm is a powerful algorithm for binary classification tasks, such as distinguishing between cases with heart disease and healthy cases. The basic idea of this algorithm is to find the optimal hyperplane that separates the two classes in a way that maximizes the margin between them, which increases the model's ability to generalize when dealing with New data.

The SVM algorithm demonstrates significant effectiveness when used with high-dimensional data, such as features extracted from deep neural networks (DNNs), taking advantage of these features' ability to represent complex, nonlinear patterns in medical data. To achieve class separation, Sharma et al. (2023) demonstrated that combining features extracted from a DNN model with the SVM algorithm significantly improved the accuracy of heart disease prediction, reflecting the power of this hybrid model in handling complex nonlinear relationships within clinical data. SVMs are particularly suitable for problems with relatively limited data and offer high accuracy when dealing with these types of medical data. Figure 3-8 shows Linear and Nonlinear SVM Classification.

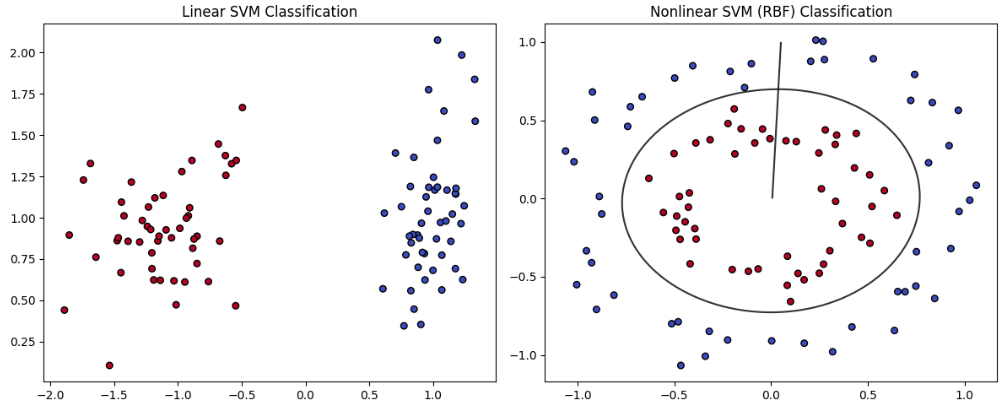


Figure 3-8 : *Linear and Nonlinear SVM Classification*

3.8.1.2 ***Random Forest Algorithm***

The Random Forest algorithm is an ensemble learning algorithm that combines the results of a set of decision trees to arrive at a final decision. Its strength lies in its ability to handle high-dimensional, noisy data, making it ideally suited for analysing complex medical data related to heart disease. During training, the RF algorithm creates a large number of decision trees, each of which is trained on a random sample of data and features using bootstrap sampling. The predictions of all the trees are then aggregated through ensemble or average voting, which reduces the overfitting problem and improves the model's ability to generalize to new data. In the context of heart disease prediction, the RF algorithm demonstrates remarkable effectiveness in modelling nonlinear and complex relationships between patient clinical features such as blood pressure, heart rate, cholesterol, and others, especially when combined with deep features extracted using a DNN model. RF also has the ability to determine the importance of features, an additional advantage that helps analyze the extent to which each feature influences the final decision, thus helping physicians clearly interpret predictive recommendations. In a recent study, Nguyen et al. (2022) demonstrated that using the Random Forest algorithm with selected clinical features significantly improved the accuracy of predictive models for heart disease, compared to other algorithms that are less capable of handling complex patterns.

3.8.1.3 ***LightGBM***

LightGBM is a machine learning algorithm based on Gradient Boosting technology. It is characterized by its high speed and efficiency in handling large datasets. It relies on building sequential decision trees, where each new tree attempts to correct the errors of the previous tree, leading to a gradual improvement in prediction accuracy. It also features multiple techniques that help speed up the training process and reduce memory consumption. In the context of heart disease prediction, LightGBM is used as a final classification stage after extracting important features from medical data. When combined with a deep learning model such as a DNN, the DNN is first trained to extract advanced, high-level features from patient data, such as complex nonlinear patterns that may be difficult to detect using traditional machine learning methods. These extracted features from the DNN layers are then passed to the LightGBM model for the final classification or prediction process, leveraging its strength in handling structured data and making accurate and rapid decisions. Thus, this integration between DNN and LightGBM contributes to enhancing the accuracy of heart disease prediction by combining the power of deep learning in extracting distinctive features with the efficiency of LightGBM in classification and prediction.

3.8.1.4 ***XGBoost***

XGBoost (Extreme Gradient Boosting) is one of the most powerful supervised machine learning algorithms based on Boosted Decision Trees (BDTs). It boasts high speed, accuracy, and the ability to handle nonlinear and complex data, making it highly suitable for healthcare applications such as heart disease prediction. In this study, the XGBoost algorithm was used as the final classification stage in building a traditional model and again within a hybrid model that combines traditional machine learning and deep learning to enhance the accuracy of heart disease prediction. XGBoost relies on building a series of decision trees incrementally, with each new tree attempting to correct the errors of previous models using a stepwise minimization algorithm supported by second-order gradients, providing computational efficiency and high training stability. After extracting deep features from the penultimate layer of the DNN, these features were combined with the original features to form a rich and balanced feature set. This set was used as input for the XGBoost model, which demonstrated an outstanding ability to detect complex, nonlinear patterns within clinical data. This prevents the model from retaining noise in the training data. All of these steps and strategies contribute to building a balanced model capable of capturing statistically significant patterns without sacrificing generalization ability on new data. A recent study by Bhatt et al. (2023) confirmed that the XGBoost algorithm outperformed several other classification algorithms when used to predict heart disease,The researchers relied on a set of internal regularization techniques. These include L1 and L2 regularization, which are added to the objective function to limit the growth of complex weights, and subsampling, which uses subsamples of the data and their features at each training step. This method reduces correlation between trees and increases the generalization power of the model on new and unseen data. Additionally, the researchers used cross-validation and systematic hyperparameter tuning to ensure the optimal model settings are selected before overfitting occurs. enhancing its reliability in sensitive medical contexts.

3.8.1.5 ***Logistic Regression***

Logistic regression is one of the simplest and most powerful supervised classification algorithms widely used in the medical field, especially when dealing with binary problems such as predicting the presence or absence of heart disease. The algorithm relies on transforming linear outputs into probabilities belonging to the interval (0, 1) using a logistic function. The algorithm determines the optimal threshold that can distinguish between two classes based on the resulting probability: if it exceeds 0.5, the patient is classified as having a disease, otherwise, it is not. In this study, the logistic regression algorithm was used as one of the basic classification models due to its high interpretability and ease of application, especially when working with datasets with a limited or medium number of features, such as the heart disease dataset. The algorithm also serves as a baseline for comparison with more complex models and has demonstrated good performance when combined with features extracted from a deep neural network (DNN) model, where it benefits from improved nonlinear representations while maintaining the simplicity of the model. This balance makes logistic regression a powerful tool in hybrid models that combine representative depth with ease of interpretation.

3.8.2 Metrics

At the beginning of the model building process, traditional machine learning algorithms were trained using the original 13 features extracted from clinical data. A deep learning model was then applied to extract new features representing more complex characteristics not directly apparent in the original data. These new features were then combined with the original features, and the model was retrained using machine learning algorithms on the combined features. This approach allows for the combined features to be leveraged, enhancing the model's predictive ability using machine learning algorithms. The model's performance before and after merging was compared to assess the extent to which the features were improved. The performance of the hybrid model used for early prediction of heart disease based on clinical data was evaluated using a set of metrics: confusion matrix, accuracy, precision, recall, F1-score, and area under the receiver operating curve (ROC-AUC). These criteria are particularly important in the medical field because misclassification can have serious consequences, such as delayed diagnosis or unnecessary procedures.

3.8.2.1 *A confusion matrix*

The confusion matrix is a fundamental and important tool for evaluating the performance of classification models. It is used to provide detailed statistics about the model's results across four main categories: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). TP refers to cases that were correctly predicted as infected, meaning the model actually detected the presence of the disease in the patient. TN refers to cases in which the model correctly predicted the absence of the disease, reflecting its ability to exclude healthy cases. FP refers to cases that the model misclassified as infected despite being healthy, which can lead to unnecessary medical interventions. FN refers to cases in which the model failed to detect infection despite the actual presence of the disease, which can lead to delays in diagnosis and treatment. This matrix is critical in sensitive medical fields, such as heart disease prediction, where striking a balance between accurately detecting infected cases and avoiding false alarms is crucial. Analysing these indicators helps understand the strengths and weaknesses of a model, thereby improving its effectiveness in supporting clinical decisions. Recent studies have emphasized the importance of using the confusion matrix as part of a comprehensive evaluation of model performance in medical classification, given its accuracy in highlighting performance details beyond general accuracy (Alizadehsani et al., 2019).

3.8.2.2 *Accuracy*

Accuracy is one of the primary metrics used to evaluate the performance of classification models. It represents the percentage of correctly classified cases, whether infected or not, out of the total number of cases in the test set. Accuracy is calculated by the relationship between true positives (TP) and true negatives (TN) compared to the total number of cases, i.e., including false positives (FP) and false negatives (FN).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{\textit{(3.6)}}$$

The accuracy percentage was calculated using the accuracy_score(y_true, y_pred) function from the sklearn. Metrics library. The default values and settings were used (without any additional parameters) because the study addresses binary classification. Although this measure provides a general sense of a model's effectiveness, it can be misleading in cases of class imbalance. Accuracy can be high when one class is more common than the other, even if the model primarily predicts the majority class due to the poor performance of the minority class. Thus, accuracy provides a basic measure of validity, but it should be interpreted carefully in contexts requiring balanced classification. Recent studies have indicated the need to treat accuracy with caution when building models to predict heart disease, especially when using machine learning and deep learning techniques (Javeed et al., 2023).

3.8.2.3 *Recall*

Recall, also known as sensitivity, is a vital metric for evaluating the performance of classification models. It represents the percentage of actual positive cases that the model was able to detect, compared to the total number of actual positive cases. This metric is calculated mathematically using the following equation:

$$\text{Recall} = \frac{TP}{TP + FN} \quad \textbf{\textit{(3.7)}}$$

The Recall percentage was calculated using the recall_score(y_true, y_pred, average='binary') function from the sklearn.metrics library. The default values and settings were used (without any additional parameters), with the positive target class set to (1) by default. This indicator is particularly important in medical prediction applications, such as heart disease prediction, given the serious health risks and treatment delays that result from failing to detect infected cases. Achieving a high recall value means that the model is able to identify most infected cases, thus reducing the number of cases that may go undiagnosed. For this reason, recall is considered an important and sensitive indicator in early diagnosis scenarios. In some medical cases, it is preferable to reduce false negatives, even at the expense of increasing false alarms.

Recent studies have shown that focusing on the recall metric within the frameworks of machine learning and deep learning algorithms contributes to enhancing the accuracy of models, especially when dealing with imbalanced clinical data, such as in heart disease (Raza et al., 2023) .

3.8.2.4 *Precision*

Precision is a key metric used to evaluate the reliability of models in classifying and identifying positive cases. It measures the proportion of correctly predicted positive cases compared to the total number of cases the model predicted as positive. In other words, it reflects the ability to avoid misclassifying negative samples as positive. This metric is mathematically defined by Equation (.38).

$$Precision = TP / (TP + FP) \qquad (3.8)$$

The percentage of accuracy was calculated using the function `precision_score(y_true, y_pred, Average='binary')` from the `sklearn.metrics` library, relying on default parameter settings. Precision is crucial in medical prediction systems because false positive predictions can lead to unnecessary medical procedures, patient stress, and significant healthcare costs. A high precision value indicates the reliability of the model's positive predictions and that most predicted positive cases actually match infected patients. Recent studies have highlighted that incorporating accuracy as a key assessment metric enhances the robustness and clinical applicability of machine learning and deep learning models in predicting cardiovascular diseases (Raza et al., 2023).

3.8.2.5 *F1-score*

The F1-score measure is one of the important measures used to evaluate the performance of classification models. This measure represents the harmonic mean between the precision and recall measures. Thus, this measure provides a numerical balance between the model's ability to identify actual infected cases and its ability to avoid false alarms. This indicator is calculated using the following mathematical relationship:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (3.9)$$

It was calculated using the f1_score(y_true, y_pred, average='binary') function from the sklearn.metrics library. The default values and settings (without any additional parameters) of average='binary were used for binary data . The F1-score becomes important in cases where data are unbalanced across classes, as is the case in many disease prediction applications, including heart disease. In such cases, the overall accuracy can be misleading, while the F1-score provides a more balanced and objective measure. This metric is particularly suitable when it is necessary to detect as many disease cases as possible without increasing the number of healthy cases being misclassified as disease. Recent studies have demonstrated the effectiveness of this metric in providing a comprehensive assessment of the performance of machine learning and deep learning-based models in early detection of heart disease (Islam et al., 2023).

3.8.2.6 ***ROC-AUC***

The ROC curve is a graph that describes the operational characteristics of a binary classification model. It is an important visual tool for evaluating the performance of binary classification models, as it shows the relationship between the true positive rate (TPR) and the false positive rate (FALSE) across different threshold levels. Each point on the ROC curve represents a specific threshold, showing how changing the threshold affects the model's sensitivity and specificity. These curves are widely used in medical fields, especially in the diagnosis of heart disease, due to their ability to illustrate the model's performance across multiple classification scenarios.

By analysing the ROC curve, the optimal threshold can be selected to achieve a balance between sensitivity and specificity. This is critical in cases such as heart disease prediction, where any misclassification, whether due to omission or false positive, can have serious clinical consequences.

The ROC curve is associated with a mathematical measure known as the area under the curve (AUC), which expresses the overall ability of the model to distinguish between diseased and undiagnosed categories.It was calculated using the roc_auc_score(y_true, y_prob) function from the sklearn. Metrics library. The default values and settings were used (without any additional parameters). The AUC value ranges from 0 to 1, with a value close to 1 indicating that the model is robust and has a high ability to correctly classify. A value of 0.5 indicates that the model's performance is no different from a random guess. Therefore, the AUC is one of the most important metrics used to evaluate models in the medical field, as it provides a comprehensive assessment of a model's performance across various classification thresholds and helps in selecting the most appropriate model in terms of balancing sensitivity and specificity, a critical balance when making sensitive diagnostic decisions such as diagnosing heart disease (Mujica-Parodi et al., 2023). Ultimately, intelligent models used to predict heart disease rely on an integrated set of performance metrics to evaluate and improve the effectiveness of the classification model. These metrics include overall accuracy, precision, recall, F1 index, and area under the ROC curve (AUC). Since all the metrics used in the study were implemented using the default values of the Scikit-learn library to ensure compatibility with standard evaluation criteria in machine learning, these settings are appropriate because the study deals with binary classification. Accuracy refers to the overall proportion of cases that were correctly classified out of the total cases, while precision measures the accuracy of positive predictions only, reflecting the model's ability to reduce the number of false positives and thus limit unnecessary medical interventions. Recall, on the other hand, expresses the model's ability to actually detect infected cases, which is critical in cases where missed diagnosis is accompanied by a delay in starting treatment, as is the case with heart disease. The F1 index combines precision and recall into a single balanced value, enabling the model's ability to detect true positive cases while reducing the percentage of false alarms. The AUC-ROC metric provides a comprehensive assessment of the model's ability to distinguish between different classes across multiple levels of Threshold, allowing selection of the most appropriate threshold based on desired clinical goals. Together, these metrics form an integrated analytical framework that helps researchers and practitioners select and fine-tune the optimal model to ensure the desired balance between diagnostic sensitivity and accuracy. This directly impacts patient outcomes and reduces the risk of false predictions.

3.8.3 Validation

Validation is a crucial stage in any practical study aimed at developing and comparing machine learning models. In this study, model evaluation is not based solely on standard performance metrics such as accuracy, precision, recall, F1 metric, and ROC curve area. Instead, the evaluation is extended to include a systematic comparison between different models and between the same model before and after feature integration. To ensure a reliable and fair evaluation, k-fold cross-validation was used. This evaluation assesses each model across multiple datasets, reducing bias associated with splitting training and testing data into two sets and increasing the model's representativeness and stability.

The average results across the datasets were used to analyse the effectiveness of the proposed feature extraction and integration method.

3.8.3.1 *Practical steps for implementation*

To assess the impact of combining features extracted from the neural network model with the original features on the performance of traditional machine learning algorithms, a comprehensive evaluation using k-fold cross-validation was performed on the training set. The performance of traditional models such as SVM, Random Forest, XGBoost, LightGBM, and Logistic Regression was compared using several key performance metrics: accuracy, recall, precision, F1 score, and area under the ROC curve (AUC).

The models were also tested on the final test set to compare performance between the original and combined features. This step highlights how feature combinations enhance the predictive power of the models and yield more reliable and accurate results. This approach, which relies on direct evaluation of model performance rather than separate statistical testing, makes the results clear and easy to interpret for researchers and practitioners, demonstrating the impact of feature combinations on improving prediction accuracy.

3.8.4 Hyperparameters after feature fusion

After combining the original features with those extracted from the neural network, the models were retrained on the new, combined features using the default settings for each algorithm. This allowed for the evaluation of how well the models benefited from the rich, integrated features in their predictive performance without any further adjustments to the hyperparameters. The results also demonstrate how integrated features can improve the accuracy and generalizability of models even when using default parameters, providing a strong foundation for future performance improvements through hyperparameter adjustments. Table 3-6 shows the final hyperparameters for each algorithm trained on the pooled dataset.

Table 3-6 : *The following table shows the final hyperparameters for each algorithm trained on the combined dataset.*

Algorithm	**Hyperparameter**	**Value used in the code**	**Purpose of selection**	**Description**
SVM	Kernel	rbf'	Dealing with nonlinear relationships between data	The RBF kernel enables the model to flexibly separate nonlinear data and improves predictive ability.
	C	1.0	Regulating error and balancing bias and variance	A neutral value prevents overfitting and maintains the model's ability to generalize.
	Gamma	scale	Control the extent of influence of each sample	The default value “scale” is suitable for standardized data.
Random Forest	n_estimators	100	Sufficient number of trees to reduce variance and increase model stability	Increasing the number of trees improves prediction accuracy and reduces the impact of noise on the model.
	max_depth	None	Allow the tree to fully grow while controlling overgrowth through bunching.	Allows the model to learn complex patterns without overfitting on individual data.
	max_features	sqrt	Controls the number of features considered at each split in the decision tree.	Reducing tree interdependence and increasing tree diversity
XGBoos	learning_rate	0.3	Control the speed of model learning	Low learning rate contributes to improved gradient convolution and reduced overfitting.
	n_estimators	100	Number of trees to provide a basis for performance evaluation	Determining the number of trees balances good learning and generalization ability.
	max_depth	6	Allow the tree to fully grow while controlling overgrowth through bunching.	Control the complexity of each tree to prevent overfitting.
LightGBM	learning_rate	0.1	Control learning speed and improve stability	A benchmark value that ensures that the model learns gradually without significant changes in each update.
	num_leaves	31	Specify the maximum number of nodes in each tree to balance performance and speed.	Help maintain moderate tree complexity and avoid over-specification.
	n_estimators	100	Number of trees to provide a basis for performance evaluation	Optimal number of trees to achieve high performance quickly
LR	C	1.0	Organize the model and prevent overfitting.	A neutral value enhances the model's ability to generalize by reducing the weight of large coefficients.
	Penalty	l2	Reducing model complexity	L2 regularization reduces large parameter values and enhances generalization to new data.
	Solver	lbfgs	Improve the model in an effective way	Suitable for small and medium data sets.

The hyperparameters used for each model were the default values for all models after feature merging, without resorting to any advanced hyperparameter optimizations. This was done to obtain a preliminary analysis of performance on the combined set of features and to ensure clarity, repeatability, and transparency in the experiment. Providing a balanced model allows for preliminary evaluation of model performance, maintains the simplicity of the experiment, and avoids introducing any complexity that might affect the understanding of the underlying performance of the models.

3.9 Deep Learning Foundations

Deep learning represents a powerful and supportive methodology for modelling complex and nonlinear relationships in medical data. The working principle of this approach is based on stacking multiple layers of computational processing, so that raw medical inputs are gradually transformed into more abstract representations. This approach resembles the human brain's hierarchical information processing mechanism, making it suitable for extracting hidden patterns and meanings from cardiac data containing diverse clinical and demographic features.

3.9.1 Neural Network Layers and Activation Functions

The proposed deep neural network (DNN) architecture consists of three consecutive hidden layers. The first layer contains 128 neurons, followed by the second layer with 64 neurons, and the third layer with 32 neurons, with the goal of capturing multi-level patterns in medical data. The ReLU (Rectified Linear Unit) activation function was used in all hidden layers due to its effectiveness in accelerating the training process and mitigating the gradient decay problem. The final layer (the output layer) relied only on a single neuron with a sigmoid activation function, given its significant role in the binary classification task (injury/non-injury). The model was trained using the Adam Optimizer algorithm with a learning rate of 0.001, while the batch size was set at 32 samples and the number of training epochs was set at 50. The numeric values were standardized using a StandardScaler to a mean of zero and a standard deviation of one, which helps improve training efficiency and boat speed. The details of the hyperparameters and DNN architecture are shown in Table 3-7.

Table 3-7 : Hyperparameters and Architecture Details of the (DNN)

Parameter	Value in Code	Description
Input Shape	(X_train_scaled.shape[1],)	Number of features entering the network (13 clinical features)
Layer 1 Neurons	128	Number of units in the first hidden layer
Layer 2 Neurons	64	Number of units in the second hidden layer
Layer 3 Neurons	32	Number of units in the third hidden layer (feature extraction layer)
Layer 4	1	Number of units on the fourth floor
Hidden Layer Activation	ReLU	Activation function for hidden layers
Output Layer Neurons	1	One unit for binary result
Output Activation	Sigmoid	Activation function of the output layer to produce a probability between 0 and 1
Optimizer	Adam	Optimization algorithm for updating weights
Learning Rate	0.001	Network learning rate
Batch Size	32	Number of samples per training batch
Epochs	50	Number of passes over the entire training set
Feature Extraction Layer	Layer 3 (32 neurons)	The penultimate layer for extracting features for later use in classic ML models.

The number of features extracted from the penultimate layer was 32 units. The extracted features were then merged with the original 13 features, resulting in a total of 45 features after merging. This was followed by a phase of training traditional machine learning models (SVM, Random Forest, XGBoost, LightGBM, Logistic Regression) on the post-merged features. This procedure addresses the study's observations on improving feature representation while maintaining reproducibility.

3.9.2 Deep Neural Networks for Automated Feature Learning from Raw Inputs

Unlike traditional machine learning algorithms that rely on manually engineering features, deep neural networks (DNNs) are capable of automatically learning feature representations directly from raw medical data. In this study, a deep neural network with multiple hidden layers was constructed to extract deep representations (latent features) that capture the linear and nonlinear relationships between clinical and demographic variables in heart patients. The penultimate layer of the model was used as a feature extraction layer, generating rich representations that can reveal subtle patterns that may not be apparent based on the original data alone. After extracting these deep features, they were combined with the original features to form a more comprehensive ensemble, which was then fed back into a set of traditional classification algorithms such as SVM, Random Forest, XGBoost, LightGBM, and Logistic Regression. This approach enabled a systematic comparison of predictive performance before and after fusion, highlighting the added value of using deep features in improving the accuracy of heart disease prediction.

3.9.3 Convolutional and Recurrent Variants for Sequential Health Data

Although our model focuses on non-sequential clinical data, there are advanced variants of neural networks that can be useful in processing other types of medical data, especially those that require understanding spatial or temporal patterns.

3.9.3.1 *Recurrent Neural Networks - RNNs*

Recurrent neural networks (RNNs) are designed to process sequential data such as patient history logs or time-series data. This technology has an internal memory that allows it to retain information from previous time steps, making it ideal for tracking a patient's clinical developments over time, such as changes in vital signs and blood pressure fluctuations. RNNs can also predict future events based on past behavior, making them valuable in predicting disease course or response to treatment.

3.9.3.2 *Convolutional Neural Networks - CNNs*

CNNs are specialized architectures for processing grid-like data, such as images. CNNs scan spatial patterns using convolutional filters and extract hierarchical features from the data. In medical contexts, these networks are effectively used to analyze electrocardiograms (ECGs), magnetic resonance imaging (MRIs), and X-rays, helping to automatically detect abnormalities and patterns associated with diseases.

3.10 Ethical and Practical Considerations

The use of intelligent models in medicine is not limited to technical aspects; rather, it also takes into account ethical and practical responsibilities to ensure that the results are reliable, fair, and safe when used in healthcare settings. In the context of this thesis, which used patient data and machine learning tools to predict heart disease, it was necessary to discuss and address three key dimensions: data bias, protecting patient privacy, and the practical requirements for implementing models in clinical settings. These are discussed in the following subheadings.

3.10.1 Bias in Data

Bias in clinical datasets is a significant challenge that can directly impact the reliability and accuracy of predictive models in the medical field. This bias often arises from the underrepresentation of demographic groups, such as age, ethnicity, and gender. This can lead to biased models that overestimate the performance of some patient groups while underestimating the performance of others. This issue was addressed through a comprehensive descriptive analysis of the data to understand its statistical characteristics and identify any potential distributional imbalances. A fair performance assessment was also employed, utilizing not only accuracy but also other measures such as recall and F1 scores, which reveal the model's performance in underrepresented groups. Multiple algorithms (such as SVM, random forest, XGBoost, LightGBM, and logistic regression) were tested, and the models were compared to ensure that the final result was not biased by any single algorithm. In addition, K-fold cross-validation was adopted in model evaluation, ensuring the stability of results and minimizing the bias arising from a single data split, thus contributing to a more reliable assessment of model performance. Addressing bias not only improves the overall performance of the model but also enhances its integrity and broad applicability, enabling models to produce fair results across all clinical categories. This ensures more reliable and ethical use of these technologies in medicine.

3.10.2 Data Privacy

Clinical privacy of patient data is a key ethical consideration, requiring that data used in training be subject to strict standards to protect it from unauthorized access. To ensure this, the clinical data in this study were processed in accordance with privacy principles. Key safeguards included patient anonymization, where all data were stripped of any personally identifiable information. This process ensures that patient data is used for research purposes only and is not linked to an individual's true identity. Additionally, the data were sourced from publicly available datasets and were anonymized, reducing the risk of unauthorized access or unethical use. Privacy was treated as a fundamental pillar, no less important than the practical aspect.

3.10.3 Practicality

The practical application of the proposed models is a pivotal aspect in evaluating their effectiveness. It is not enough for these models to achieve high performance in theory if they lack the ability to be implemented in diverse medical settings. In resource-limited environments, for example, computational requirements pose a major challenge, as highly efficient devices or advanced computing systems may not be available to support the operation of complex models. Hence, the need to design lightweight, highly efficient models that balance accuracy and performance with operational capability in practical conditions. Ensuring their applicability in various healthcare institutions also requires consideration of simplicity of implementation and integration with existing medical systems. This facilitates their smooth use by medical personnel without the need for advanced technical expertise. Therefore, it is necessary to consider these practical aspects not only to enhance the chances of wider adoption of the proposed models, but also to ensure their sustainable use and the desired benefits in the early diagnosis of heart disease.

3.11 Implementation Tools

The success of the proposed model depends primarily on the software tools and environment employed throughout the research. Data was first prepared, followed by training and evaluation of the models, and results were obtained. This section aims to identify the software environment and hardware used to fully implement the hybrid model.

3.11.1 Software

All code in this thesis was written and implemented in Python. This language was chosen due to its flexibility, large support community, and rich set of specialized libraries for machine learning and deep learning. Several libraries were used critically, such as TensorFlow/Keras, which is considered a cornerstone for building and running deep neural networks. Keras allowed us to provide an easy-to-use interface, understand the network structure, identify its layers, and train it efficiently. The scikit-learn library was also used and is relied upon to implement all traditional machine learning algorithms. It was used for data splitting, standardization, and model performance evaluation using various performance metrics. In addition, the libraries for these two algorithms, XGBoost and LightGBM, were used, as they provide improved and efficient implementations of ensemble learning models. These two libraries, Pandas and NumPy, are used to process and analyze data and perform the necessary calculations on matrices accurately and efficiently.

3.11.2 Hardware

Hardware-wise, the training process requires appropriate computational resources due to Due to the computational complexity inherent in deep neural networks, computing resources were carefully considered during the network's implementation. The training process was performed using the central processing unit (CPU), which was sufficient to accelerate the learning process and reduce processing time, especially considering the nature of the data used and the complexity of the model. Memory consumption, storage capacity, and the characteristics of the extracted deep features were also taken into account. This integration of software tools and available computing resources contributed to providing a reliable and comprehensive environment for conducting the study and building the model, ensuring the accuracy and reproducibility of the results. This chapter summarizes the methodology followed in constructing and developing the proposed hybrid model for predicting heart disease. A solid foundation for the research was built through sequential and organized steps. It began with data collection and pre-processing to ensure quality, followed by an initial training phase that provided a basic understanding of the algorithm's behavior. The core of the methodology was the integration of deep learning with traditional machine learning, using a deep neural network to extract rich and complex features. These were subsequently combined with the original features in a crucial step. The work went beyond technical aspects to include a rigorous statistical evaluation of the model's performance, demonstrating that the achieved improvements were statistically significant and not merely a result of chance. Furthermore, ethical and practical considerations were addressed, underscoring the research's responsibility to address numerous issues such as bias, privacy, and the practical applicability of the model in a variety of settings, particularly those with limited resources. In conclusion, this integrated methodology, embodied by advanced software tools, represents a robust scientific framework for building effective and reliable predictive models, paving the way for future research aimed at enhancing preventive diagnostic capabilities in the medical field.

Chapter 4: Results and Discussion

This chapter presents the performance results and comprehensive evaluation of the proposed hybrid framework for early prediction of heart disease by integrating deep learning-based feature extraction with traditional machine learning classifiers. The methodology initially involved training traditional machine learning models such as Support Vector Machine (SVM), Random Forest (RF), XGBoost, LightGBM, and logistic regression on the original clinical features. A deep neural network (DNN) was then used to extract high-level features, which were combined with the original set of features to form an improved feature representation. These combined features were used to retrain the same set of classifiers to evaluate the impact of augmenting the deep features. To evaluate the robustness and predictive performance of the models, experiments were conducted in two scenarios: one using only the original features, and the other using a combination of the original and DNN-derived features. The performance of each model was evaluated using key evaluation metrics, including accuracy, precision, recall, F1 score, and area under the curve (AUC). The results showed significant improvement in classification performance when using a combination of integrated features, with accuracy and AUC values approaching 99%, demonstrating the effectiveness of the hybrid approach. Furthermore, confusion matrices were analyzed for each case to visualize the classification results and identify misclassification patterns. These evaluations highlight the effectiveness of deep learning feature extraction in enriching the feature space and improving the predictive accuracy of classical classifiers. Overall, the results confirm the potential of the proposed hybrid approach in enhancing early detection of heart disease, which is particularly important in resource-limited healthcare settings. Therefore, this chapter demonstrates the value of combining advanced deep learning representations with effective machine learning classifiers for accurate prediction of heart disease.

4.1 Classification results

This subchapter presents the results of evaluating the model in heart disease classification. The results will be divided into two parts: classification based on the original features, and classification after combining the extracted features with the original. The performance metrics and confusion matrices for each classifier will be discussed.

4.1.1 Classification based on original features

The results of training the classifiers on the original features were extracted without using DNN technology. Details of these results are shown in the table below.

4.1.1.1 *Evaluate the performance of traditional models before combining features.*

The first scenario of the study is to evaluate models on the original features only and includes the following metrics for five popular machine learning classifiers: (SVM), (Random Forest), (XG Boost), (Light GBM), (Logistic Regression), as shown in the table 4-1:

Table 4-1 : *Evaluate the performance of traditional models before combining features.*

Metric	SVM	Random Forest	(XG Boost)	(Light GBM)	(Logistic Regression)
Accuracy (%)	90.98	97.80	97.56	97.20	85.85
Recall (%)	90.70	99.32	99.08	98.58	90.50
Precision (%)	91.66	96.64	96.43	96.18	83.57
F1 Score (%)	91.09	97.91	97.69	97.33	86.82
AUC (%)	96.58	99.59	98.99	99.14	92.36

The results of evaluating the models on the original features only showed differences in the performance of the algorithms in terms of accuracy and discrimination ability, which reflects the effectiveness of each algorithm when dealing with clinical data without any augmentation with deeply extracted features. The Random Forest model achieved the highest performance overall, with an accuracy rate of 97.80%, a very high recall rate of 99.31%, a predictive accuracy rate of 96.64%, and an AUC value of 0.995, demonstrating its high efficiency in classifying heart disease cases using the original features. Its training time compared to its accuracy makes it an effective choice in terms of performance and speed. In contrast, XG Boost's performance was very close to Random Forest's, with an accuracy rate of 97.56% and an AUC value of 0.989, reflecting its ability to handle complex data, although its performance was slightly lower. Light GBM also demonstrated competitive performance, with an accuracy of 97.19% and recall of 98.58%, confirming its effectiveness as a lightweight, fast, and high-performance algorithm.

On the other hand, the SVM model demonstrated average performance, with an accuracy of 90.97%, but a lower-than-expected recall of 90.70%. This may indicate some difficulty distinguishing between positive and negative cases in some samples, especially when there is overlap in clinical features.

Logistic regression performed the weakest among all the models used in the study, with an accuracy of 85.85%. Recall was also low compared to other models, reaching 90.49%, and the AUC value was 0.923, indicating its limitations in handling nonlinear data or data containing complex relationships between features.

To further evaluate the effectiveness of these models, I compared them with previous studies on the same topic that applied similar traditional machine learning algorithms to the same Cleveland dataset. This study, by Rani et al. (2021), reported that the performance of classifiers on the full feature set was 83.83% for the random forest, 83.80% for the logistic regression, and 79.50% for the support vector machine algorithm , This study by Karthick et al. (2022) used the same data as our study and applied several models. Support vector machine (SVM), Gaussian Naive Bayes, logistic regression, LightGBM, XGBoost, and random forest algorithms have been employed for developing a heart disease risk prediction model and obtained the accuracy as 80.32%, 78.68%, 80.32%, 77.04%, 73.77%, and 88.5%, respectively. Our results show significantly superior performance based on the original features only.

The results of this study indicate that boosting tree models, especially Random Forest and Light GBM, are more capable of understanding complex patterns in the original features, unlike linear models such as logistic regression. This reinforces the need to use more advanced methods to improve performance, such as feature extraction using deep networks, the results of which will be presented in the next phase of the study.

4.1.1.2 *Confusion Matrices*

The confusion matrices below give an idea of the classification distribution for each classifier as shown in the table 4- 2:

Table 4-2 *: Results of confusion matrices on the original features*

Algorithm	True Positive (TP)	True Negative (TN)	False Positive (FP)	False Negative (FN)
SVM	97	85	17	6
Random Forest	100	102	0	3
XGBoost	100	102	0	3
LightGBM	100	102	0	3
Logistic Regression	90	73	29	13

The confusion matrix results in the table indicate differences in the ability of the five algorithms to distinguish between infected and uninfected cases based on the original clinical features alone. Random Forest and XG Boost models achieved near-perfect accuracy in classifying all healthy cases (TN=102), identifying almost all infected cases (TP=100), and zero false positives (FP=0). Only three cases were missed (FN=3). This performance reflects the strong ability of these models, which rely on boosting and clustering techniques, to handle nonlinear patterns in patient data. SVM performed well but less well than the tree-based models, accurately identifying 97 infected cases and 85 healthy cases. However, it encountered six false negatives (FN) and 17 false positives (FP), indicating lower sensitivity and specificity compared to tree-based algorithms. Logistic regression performed poorly among the models, misclassifying 29 healthy patients as affected (FP) and 13 as not detected (FN), reflecting its limited ability to represent the complex relationships between features and health status. These results highlight the superiority of tree-based boosting and clustering algorithms for modelling clinical cardiac data, supporting the trend toward incorporating deeper features extracted by deep learning to enhance the performance of less accurate models.

4.1.1.3 *Confusion Matrices Visualization*

The confusion matrices in Figure 4-1 illustrate the classification performance of five machine learning models on the original features only. Each matrix displays the distribution of correctly and incorrectly classified cases, with the diagonal values representing the number of true predictions and the off-diagonal values representing the false predictions. Higher values along the diagonal indicate stronger predictive ability For example, it was able to correctly classify a high percentage of infected patients into the positive category, and also accurately distinguish healthy individuals into the negative category. while larger values off-diagonal highlight the types of errors the model tends to make (such as false positives or false negatives). The goal of this analysis is to demonstrate how well the models distinguish between positive and negative cases and whether there are any discrepancies in the classification errors. These results serve as a baseline against which to compare the performance of the models before and after feature fusion.

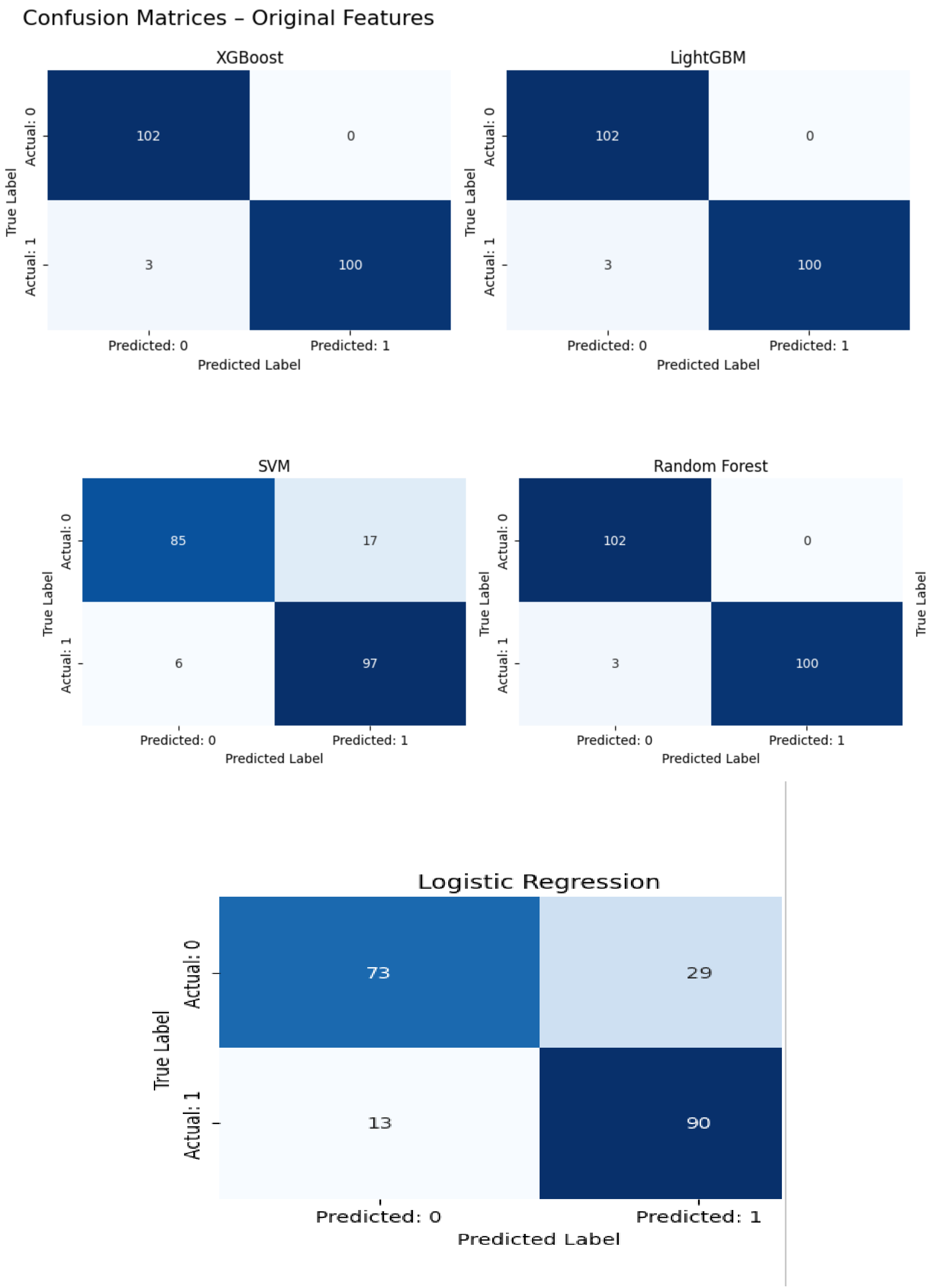


Figure 4-1 : *Visualize confusion matrices by original features.*

The results of training the classifiers were extracted after combining the features extracted by the DNN. Details of these results are shown in the table below.

4.1.1.4 ***Performance results of models after combining the original and extracted features via DNN***

Below is a summary of the improved performance metrics for five algorithms as shown in the table 4-3 :

Table 4-3 : *For improved performance metrics of five algorithm*

Metric	SVM	Random Forest	(XG Boost)	(Light GBM)	(Logistic Regression)
Accuracy (%)	99.39	99.63	99.63	99.63	99.39
Recall (%)	99.27	99.76	99.76	99.76	99.27
Precision (%)	99.56	99.56	99.56	99.56	99.56
F1 Score (%)	99.41	99.65	99.65	99.65	99.41
AUC (%)	99.90	99.73	99.69	99.86	99.48

The results showed a significant improvement in the performance of all classification algorithms after combining the original features with those extracted using deep neural networks (DNNs). This combination significantly increased the classification accuracy across all key indicators. The SVM algorithm achieved the highest AUC value of 99.90%, demonstrating its high ability to accurately distinguish between classes. The Random Forest, XG Boost, and Light GBM algorithms also achieved identical results in overall accuracy (99.63%) and other indicators, reflecting the strength of these models in handling combined features. Logistic Regression also achieved very good results, with an accuracy of 99.39% and a recall rate of 99.27%.

To further evaluate the effectiveness of these models, a comparison was made with previous studies on the same topic that applied similar traditional machine learning algorithms to the same Cleveland dataset. This study, Rani et al. (2021), reported that SVM achieved an accuracy of 84.79, logistic regression achieved an accuracy of 82.50, and the RF algorithm achieved an accuracy of 83.83. This study, Al-Alshaikh et al. (2024), used a multi-layer deep convolutional neural network (MLDCNN), trained using the adaptive elephant herd optimization method (AEHOM). The ML-HDPM model was shown to achieve an accuracy rate of 95.5% and a precision rate of 94.8%. The system's sensitivity (data recall) performed at a high precision of 96.2%, while the F-rate demonstrated its balanced performance, reaching 91.5%. It is worth noting that the accuracy of ML-HDPM was 89.7%. Our results show significantly superior performance based on features after fusion.

These results demonstrate the effectiveness of the combination technique in improving model performance, as the deep features extracted using deep learning helped extract hidden, nonlinear information and contributed to enhancing the models' ability to accurately predict, especially in cases of sensitive cardiac diseases.

4.1.1.5 *Confusion Matrices*

Confusion matrices show the improvements in classification distribution as shown in the following table 4-4:

Table 4-4 : *Confusion matrix table after combining attributes*

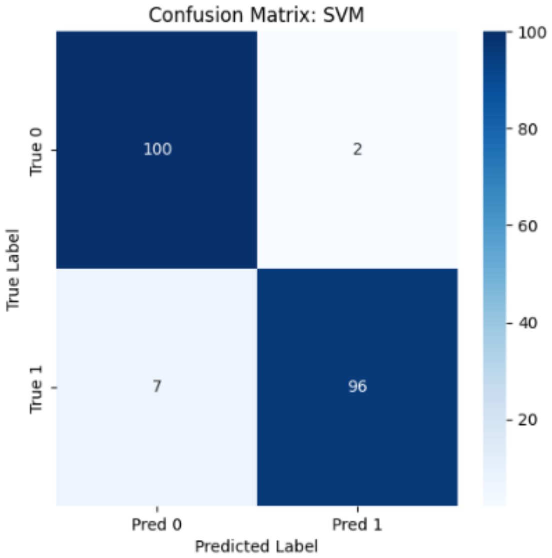
Algorithm	True Positive (TP)	True Negative (TN)	False Positive (FP)	False Negative (FN)
SVM	96	100	2	7
Random Forest	100	102	0	3
XG Boost	100	102	0	3
Light GBM	100	102	0	3
Logistic Regression	96	100	2	7

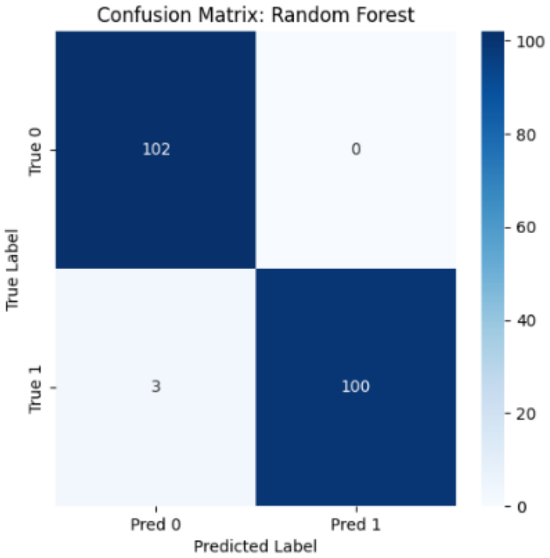
Based on the confusion matrix results of the five models after feature fusion, all algorithms demonstrated significant improvements in distinguishing between positive and negative classes, confirming the effectiveness of combining DNN-extracted features with the original features. XGBoost, Random Forest, and LightGBM achieved the highest performance in terms of the number of correct predictions, recording 100 true positives (TP) and 102 true negatives (TN) with no false positives (FP = 0), indicating high prediction accuracy and a lack of tendency to give false alarms. However, some minor errors remained, as all three models recorded false negatives (FNs), i.e., they were incorrectly classified as non-infected, indicating a need for further improvement in accurately detecting all infected cases. The SVM algorithm also showed good results, but it recorded 7 FN cases and false positives (FP = 2), indicating high prediction accuracy and a lack of tendency to produce false alarms. This means it excelled at avoiding false positives, but it slightly failed to identify some infected cases compared to the previous three algorithms. Logistic regression showed moderate balance, with TP achieving 96 and TN achieving 100 cases, 2 FP cases, and 7 FN cases, demonstrating good performance but slightly lower than clustering-based algorithms.

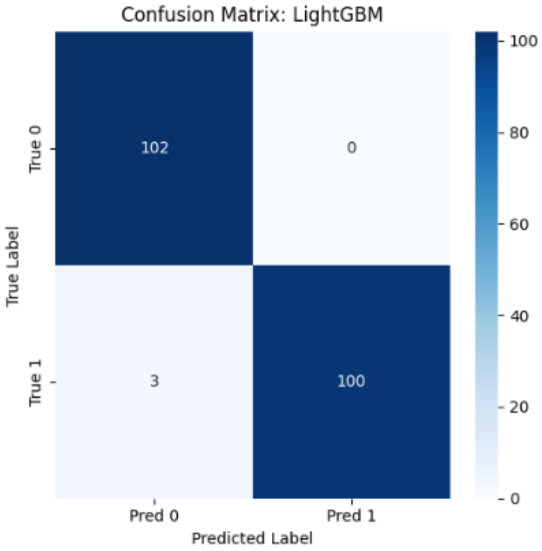
These results indicate that combining features improved the models' predictive ability, particularly in identifying positive cases without producing false predictions, which is critical in sensitive medical applications such as early detection of heart disease. These results underscore the importance of deep feature extraction techniques as a preliminary stage for classification.

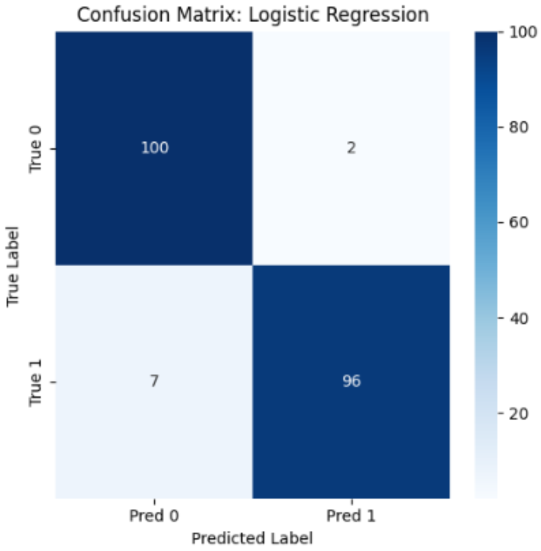
4.1.1.6 *Confusion Matrices Visualization*

Here, we present confusion matrix visualizations for five machine learning algorithms using deep learning-derived features combined with the original features, as shown in the figure 4-2 . This plays a key role in enhancing the reader's understanding of the presented results, presenting complex data in an easily interpreted visual format.









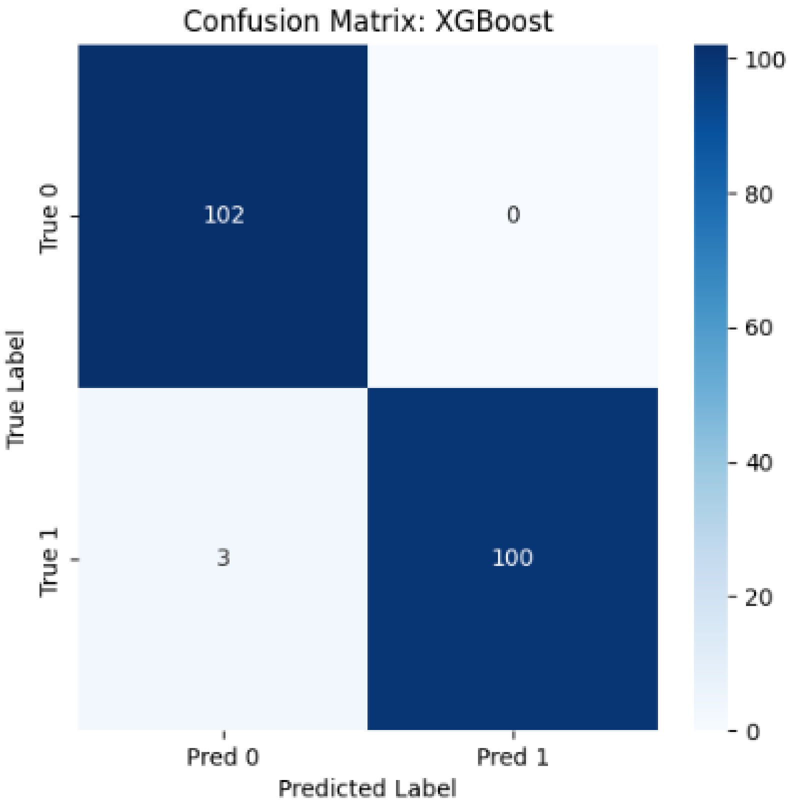


Figure 4-2 : *Visualization of confusion matrices after combining features*

4.2 Comparing the performance of the classifier before and after feature fusion

To evaluate the effectiveness of combining deep learning-derived features with original clinical features in cardiovascular disease prediction, a comparative analysis was conducted among five popular machine learning classifiers: SVM, RF, XGBoost, LR, and LightGBM. Initially, models were trained using only the original clinical features. The results showed that Random Forest, XGBoost, and LightGBM achieved the best performance among the models, with accuracy rates of 97.80%, 97.56%, and 97.20%, respectively, with very high recall rates of 99.32%, 99.08%, and 98.58%. This reflects their ability to capture critical positives and reduce false negatives, a critical point in medical diagnosis to avoid missing patients. SVM and Logistic Regression achieved lower accuracy (90.98% and 85.85%) and relatively low recall (90.70% and 90.50%), indicating the limitations of these models in detecting non-linear patterns within the original features. For the Precision criterion, the Random Forest algorithm achieved the highest value at 96.64%, indicating high accuracy in classifying positive cases without an increase in false alarms. In contrast, the LR algorithm achieved the lowest value (83.57%). The values were balanced in the F1 Score, where the highest level reached

97.91% in Random Forest. In terms of AUC, the models achieved very high values (SVM: 96.58%, Random Forest: 99.59%, XGBoost: 98.99%, LightGBM: 99.14%, Logistic Regression: 92.36%), confirming the effectiveness of the models in distinguishing between different classes at the probability level. Pre-fusion confusion matrix analysis showed that Random Forest, XGBoost, and LightGBM performed better, highlighting that missing cases (FN) before fusion, although small in the advanced models, have significant clinical impact and require mitigation strategies such as adjusting the decision threshold or using Ensemble's recall-oriented approach. After incorporating deep features, all models showed significant improvements across all metrics. Accuracy for all models exceeded 99%, with Random Forest, XGBoost, and LightGBM reaching the highest accuracy of 99.63%. The improvement in recall was also evident, with these models achieving 99.76%, while SVM and Logistic Regression both achieved 99.27%, further reducing missing cases. For Precision, all models achieved a high accuracy of 99.56%, while the F1 score reached a higher value of 99.65% and 99.41% for the advanced models compared to the simple models, confirming the ability of fusion to improve the balance between capturing positive cases and avoiding false alarms. The AUC evaluation after fusion also showed excellent performance, with Forest recording 99.73%, XGBoost 99.69%, LightGBM 99.86%, and Logistic Regression 99.48%, indicating the strength of the models in distinguishing between pathological and non-pathological classes with a significant improvement compared to the values before fusion. After analysing the confusion matrix after fusion, it was revealed that Random Forest, XGBoost, and LightGBM maintained TN=102, TP=100, FP=0, and FN=3, which means almost no false positives occurred, while SVM and Logistic Regression recorded TN=100, TP=96, FP=2, FN=7, confirming an improvement in FP, but the presence of FN remains an indicator of the need to improve clinical recall. These results demonstrate that combining deep features with original features not only enhances precision, but also improves recall and F1 score, which are critical in the clinical context to reduce the number of missed cases. It also enhances the models' ability to discriminate cases with high accuracy (AUC). This underscores the importance of using hybrid approaches that combine deep learning and traditional classification in early prediction of heart disease, while considering practical strategies to reduce FN for clinical safety.

Although integrating new features extracted from neural networks increases the data dimensions relative to the number of available samples in the dataset used in this study, this approach was not used solely to numerically improve performance. Rather, it served as a proof-of-concept framework to investigate the feasibility of deep representations with traditional machine learning algorithms in a limited medical dataset. Furthermore, the deep neural network was not used in this study as an independent final classification model, but rather to extract supervised-trained deep features. The classification task was assigned to more stable and less complex models, with cross-validation used to assess performance stability and monitor metrics. However, the exceptionally high results, particularly the AUC values close to 1, should be interpreted with caution. They are likely influenced by the nature of the standard dataset, which is small in size, structurally sound, and free of anomalies and missing values. These results should not be considered evidence of clinical readiness, but rather an indication of the methodological potential of the proposed model, rather than a guarantee of generalizability. There remains an urgent need for future external verification and more in-depth analysis of over-allocation using learning curves and real clinical data.

4.3 Interpreting model behavior using SHAP

Advanced hybrid models, particularly those based on deep neural networks, are often considered black-box models, as their inner workings are difficult to understand. In the context of sensitive applications such as medical diagnostics, it is not enough to simply achieve high performance; it is also crucial to understand how these predictive results are achieved. Achieving this requires accurate and powerful interpretive tools. To overcome this obstacle, SHAP (SHapley Additive exPlanations), a well-established methodology in the field of interpretable machine learning (XAI), was used. SHAP provides a robust and accurate modelling framework inspired by Shapley game theory, allowing us to quantify the contribution of each feature to the model's predictions. This approach not only demonstrates the importance of features but also explains the direction in which each feature drives the prediction (whether positive or negative), revealing complex interactions between features.

To interpret model decisions, the SHAP (SHapley Additive exPlanations) library was used to analyze the importance of features in predicting heart disease. TreeExplainer was used to interpret tree models such as Random Forest, XGBoost, and LightGBM. For linear models such as SVM and Logistic Regression, KernelExplainer was used to ensure the interpreter was consistent with the mathematical structure of each model . A background sample consisting of the first 100 rows of training data was used to speed up the calculation. The summary_plot graphs were used to display the most influential features on model decisions, providing a clear visual understanding of the risk factors most contributing to heart disease prediction.

In the following sections, we will analyze and interpret SHAP graphs to provide deep insights into the model's behavior, linking abstract features obtained from neural networks to clinical variables of diagnostic significance.

4.3.1 SVM Model Interpretation

Model interpretation is an important step in ensuring the reliability of predictive models, especially in sensitive fields such as medicine. In this context, SHAP (SHapley Additive) values were used to provide insight into how features interact with each other, contributing to the interpretability of model decisions. Figure 4-3 shows the SHAP interaction matrix, which shows the impact of interactions between pairs of features on model outputs. This is an ideal interpretive method for gradient boosting algorithms such as SVM, which are characterized by their ability to capture complex relationships between data. The horizontal axis represents the SHAP interaction value, indicating the extent to which the interaction contributes to the final prediction, while the vertical transform represents the interacting features. The colors indicate the value of the feature itself, with red representing high values and blue representing low values. The analysis shows that the interaction values between "Attribute 1" and "Attribute 2" are centered precisely around zero, reflecting that the interaction between the two attributes is very weak or almost nonexistent. This means that the model relies primarily on the individual influence of each attribute in the prediction process, rather than on their bilateral interactions.. This analysis effectively deconstructs the "black box" of SVM, providing a clear and reliable understanding of the foundations upon which model decisions are based, enhancing the confidence of clinicians and researchers in the results.

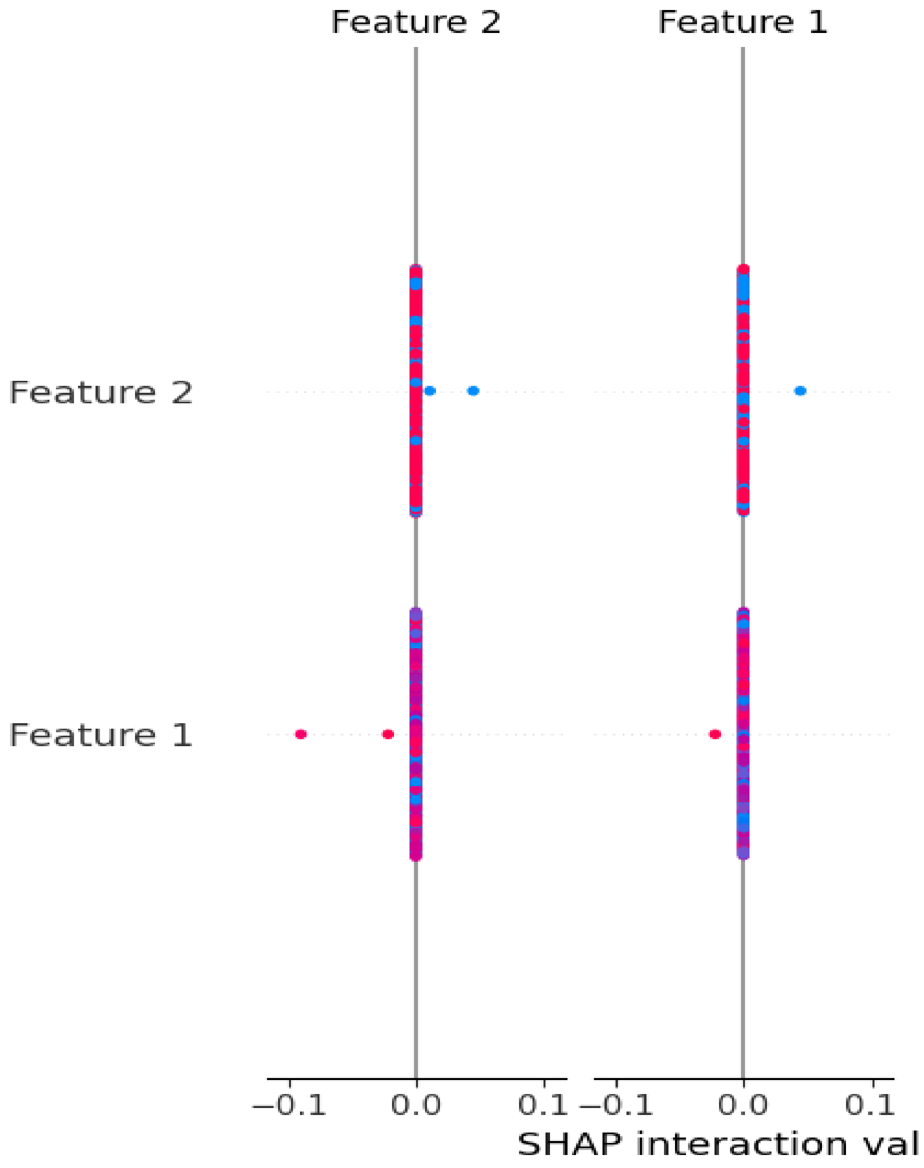


Figure 4-3: *Interpreting feature interactions using SHAP for an SVM model*

4.3.2 Random Forest Model Interpretation

Figure 4-4 shows the SHAP interaction matrix of the Random Forest model to understand how each feature individually influences the model, while also illustrating the impact of binary interactions between features on the model's decisions. In this figure, the horizontal axis represents the SHAP interaction values, while the vertical axis represents the features under study. The points represent individual data points, with red indicating high values and blue indicating low values.

From the figure, we observe that most of the interaction values between Feature 1 and Feature 2 are concentrated around zero. This indicates weak interaction between features, suggesting that the Random Forest model relies primarily on the individual influence of each feature rather than binary interactions. However, some minor nonlinear relationships are found, evident from the presence of a few points that deviate slightly from zero. This behavior is consistent with the nature of Random Forest, which tends to distribute importance across multiple features without strong interactions between them. This analysis helps us understand how the model benefits from combined features; that is, the model focuses more on the individual value of features with only minor interactions. Accordingly, SHAP analysis provides a clear view of the decision-making mechanism in the Random Forest algorithm, and this enhances confidence in its use within the hybrid model for early prediction of heart disease.

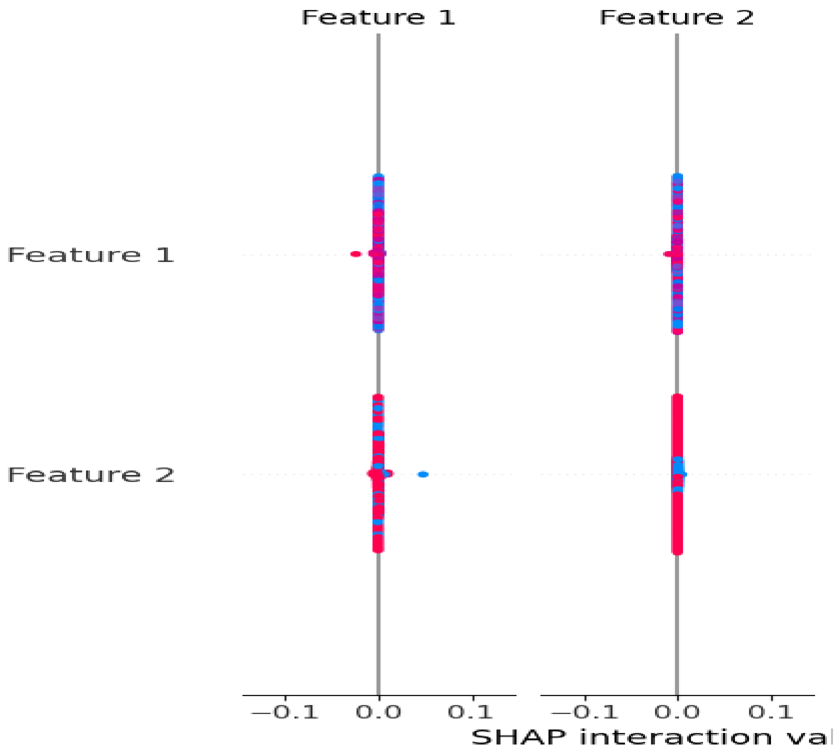


Figure 4-4 : *Interpreting feature interactions using SHAP for an Random Forest model*

4.3.3 XGBoost Model Interpretation

Figure 4-5 presents a comprehensive analysis of the effects of SHAP values on the output of the XGBoost model, highlighting the decision-making process within this complex algorithm. The graph shows a summary of the feature values, with the vertical axis displaying the features in descending order of importance, while the horizontal axis displays the impact of each feature on the model's prediction. This analysis is essential for understanding the importance and role of each feature in the final prediction. Each point in the graph represents an individual data sample, and its color reflects the feature value. Red indicates high values, while blue indicates low values. This method provides immediate and interpretable insights into how features interact to determine the model's output. For example, the top-ranked features, such as Feature 20 and Feature 18, show that their high values (red points) have a significant positive impact on the model's prediction, while their low values (blue points) have a negative impact. Furthermore, the distribution of points also reflects the variability in the influence of traits among individuals, confirming the model's ability to capture complex, non-linear relationships between traits. This analysis is a fundamental step in understanding the model's behavior and interpreting its decisions, and it is a crucial element in medical applications.

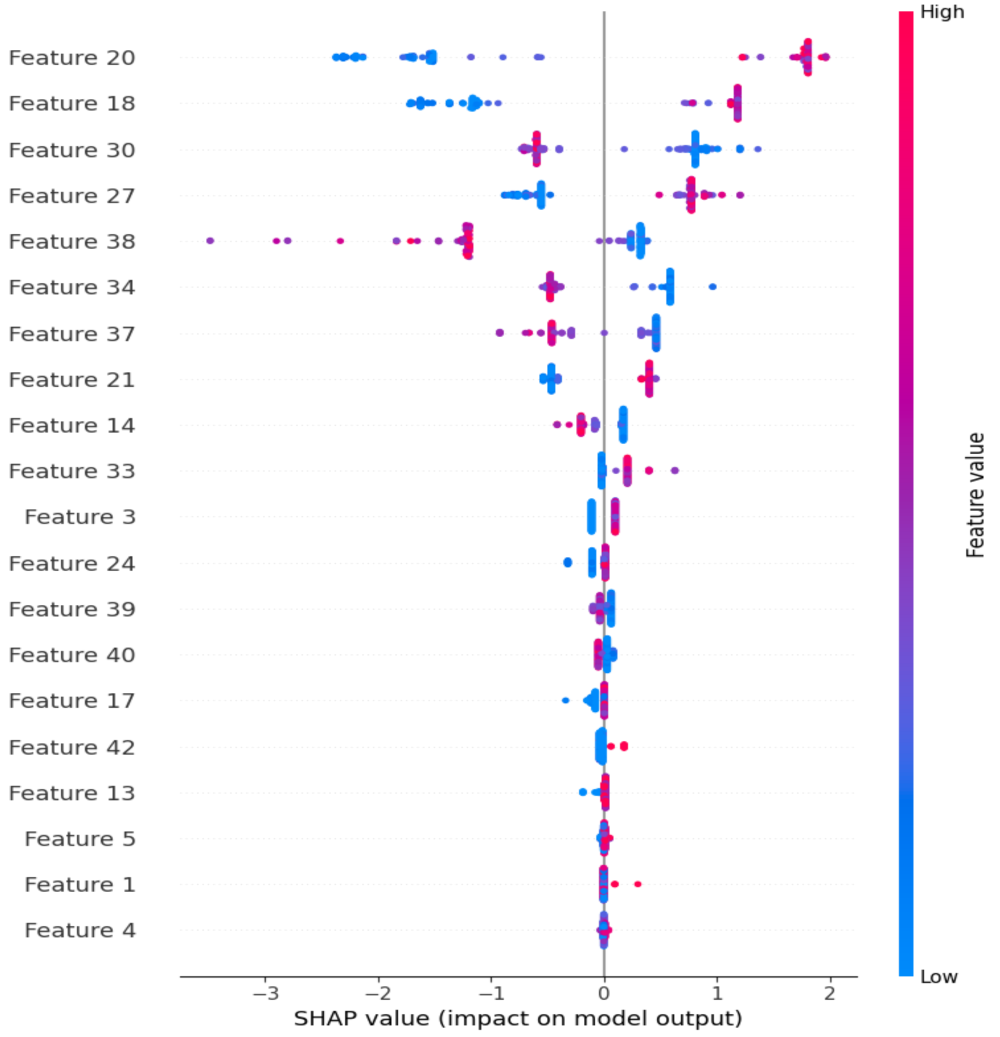


Figure 4-5: *Interpreting feature interactions using SHAP for an XGBoost model*

4.3.4 LightGBM Model Interpretation

Figure 4-6 illustrates the SHAP analysis of the LightGBM model, which provides a clear insight into the role of features and their impact on the final predictions. The SHAP values, represented on the horizontal axis, show the extent of each feature's influence on the model's output, while the vertical axis ranks features according to their importance. The color of each dot reflects the value of the feature in each sample; red indicates high values for the feature, while blue indicates low values. The analysis shows that the model places great importance on features such as Feature 15, Feature 42, and Feature 30. For example, the red dots for Feature 15 show that its high value has a positive impact on the prediction (increasing the likelihood of predicting disease), while the blue dots indicate its low value and negative impact. This analysis highlights that LightGBM, with its highly efficient and innovative techniques, can accurately identify the most important patterns, enhancing its ability to provide accurate predictions. It also provides further evidence that deep features extracted from the neural network, which appear with names (such as Feature 15, Feature 42, and Feature 30), are essential for improving the performance of traditional models.

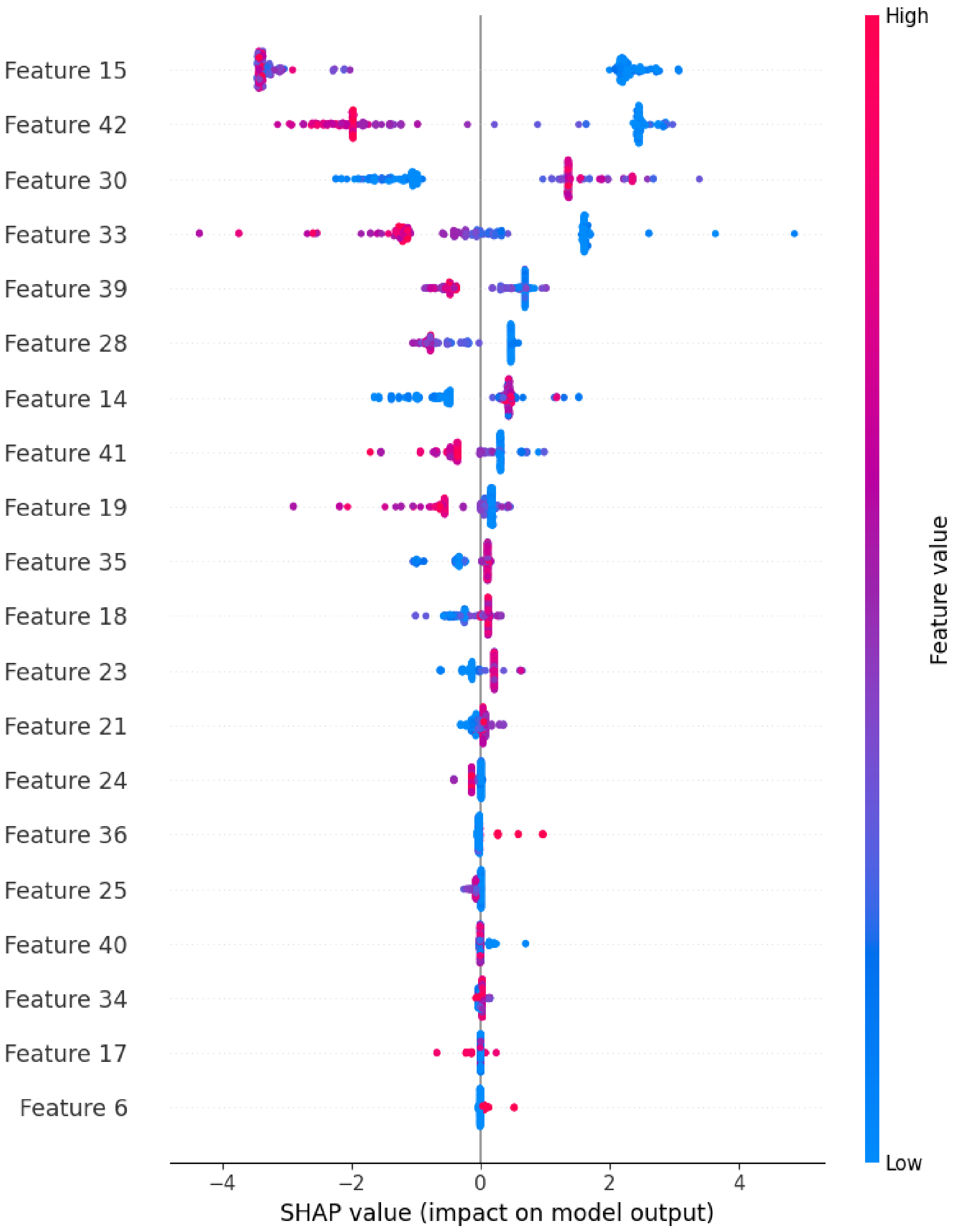


Figure 4-6: *Interpreting feature interactions using SHAP for an LightGBM model*

4.3.5 Logistic Regression Model Interpretation

Figure 4-7 displays the SHAP interactions of a logistic regression model, revealing the nature of the relationships on which this model relies to arrive at predictive results. Unlike complex tree-based models, logistic regression is a linear model by nature. The interaction matrix on the horizontal axis displays the SHAP values, highlighting the impact of interactions on the model's output. The accompanying figure shows that in the logistic regression model, the SHAP values of features (e.g., Feature 1 and Feature 2) are very close to zero. This provides strong statistical evidence that the model relies primarily on the effect of each feature individually (its direct linear effect), rather than on the complex, nonlinear interactions between features. This explains why the performance of linear models is limited when dealing with complex data containing nonlinear patterns. Although this model does not benefit significantly from combined interactions, the analysis confirms that combining features can contribute to providing richer representations, enhancing the overall performance of the model even in its simplest form.

Figure 4-7: *Interpreting feature interactions using SHAP for an LR model*

A comprehensive interpretive analysis using SHAP values demonstrates that the hybrid model does not operate as a black box, but rather provides interpretable insights into how it arrives at its results. This analysis revealed that complex models such as XGBoost and LightGBM greatly benefit from complex interactions between features, enhancing their ability to detect hidden patterns in medical data. This highlights the value of the hybrid model and its ability to extract nonlinear patterns. These results provide strong evidence for the proposed approach, as understanding how the model works and how it arrives at its results is as important as the accuracy of the results, contributing to more accurate and reliable clinical decisions in the field of cardiac disease prediction.

4.4 Clinical Implications and Error Analysis

The error analysis phase of advanced models is a crucial step in understanding the actual performance of models in clinical contexts, especially when dealing with sensitive medical data such as heart disease diagnosis. After integrating the features extracted from deep neural networks with the original clinical data, all models showed significant improvements in key metrics, including precision, recall, positive accuracy, F1 score, and ROC (AUC). This reflects the models' ability to accurately and effectively distinguish between infected and non-infected cases. Confusion matrix analysis provides a more detailed insight into the nature of the errors. For example, the Random Forest, XGBoost, and LightGBM models recorded TP = 100 and TN = 102, with FN = 3 and FP = 0. This means that the models correctly identified all non-infected cases and only failed to detect three infected cases, a very low number. However, from a clinical perspective, these three FNs are considered critical because they represent potential undiagnosed patients whose treatment may be delayed.Especially if they have atypical features, such as normal cholesterol levels or unexpected ST wave changes, this makes early detection more difficult. In contrast, SVM and Logistic Regression models showed FN = 7, which reflects a greater likelihood of misdiagnosis and a greater impact on treatment decisions. However, the minimal critical errors (FNs) and few FP failures indicate the models' effectiveness in reducing missed diagnoses and false alarms, which is crucial in medical contexts where each FN error represents a risk to the patient's life.This reflects the effective role of deep features in enhancing the predictive power of even simple models by improving recall of infected cases and raising the F1 score, which reached 99.65% in the advanced models. This confirms the model's balance between precision and recall and its importance in reducing missed diagnoses. Metrics such as Recall (99.76%) and F1-score (99.65%) for Random Forests, XGBoost, and LightGBM demonstrate the ability of infected models to perform at high efficiency, enhancing confidence in the clinical use of hybrid models combining traditional and deep features. High precision (99.56%) reduces the risk of misdiagnosing healthy patients, thus improving the quality of medical interventions.

These metrics and confusion matrix analysis provide clear evidence that combining deep features with original features not only improves quantitative performance but also increases clinical safety and confidence in predictive results, highlighting the critical importance of FN mitigation strategies in medical applications, especially when dealing with small datasets such as the Cleveland dataset (1,025 samples).

To reduce these negative errors, multiple strategies can be employed, such as using weighted loss functions (WLFs) that prioritize positive cases during training, or ensemble voting mechanisms to reduce the chance of misclassifying infected patients as non-infectious. Moreover, combining deep features with traditional sources helps improve the discriminations between normal and abnormal cases. This increases the sensitivity of the model (recall) and reduces FN, which is critical in medical contexts, where misdiagnosis can delay treatment or cause serious complications.

In conclusion, heat maps of confusion matrices provide a clear visualization of the relative errors and successes of models, highlighting the importance of focusing on Recall and F1-score alongside Accuracy to evaluate the effectiveness of models in clinical applications. These analyses also reflect the value of feature fusion in improving predictions and reducing critical clinical errors.

4.5 Conclusion

This chapter presents a detailed analysis of the classification performance of five machine learning algorithms: SVM, Random Forest, XGBoost, LightGBM, and LR for predicting heart disease. The results revealed significant differences in performance across models, particularly when comparing results before and after incorporating features extracted by a deep neural network (DNN). The hybrid approach, which combines the original clinical features with deep features, resulted in significant improvements across all evaluation metrics: precision, accuracy, recall, F1 score, and area under the curve, demonstrating the value of feature infusion in enhancing predictive modelling in clinical contexts. Before incorporating the DNN-extracted features, the Random Forest model achieved the highest classification performance, with an accuracy of 97.80% and an F1 score of 97.91%, outperforming all other models

, Random Forest also achieved the highest AUC value of 99.59 and the highest Precision value of 96.64. In contrast, the LR model performed poorly in terms of accuracy, achieving only 85.85% and an F1 score of 86.82%. This reflects its limitations in capturing complex nonlinear relationships within the original data.

After integrating the deep learning-extracted features with the original clinical features, all models demonstrated significant improvements across all performance metrics, reflecting the model's ability to enhance heart disease prediction even on small datasets (1,025 samples). For example, the Random Forest, XGBoost, and LightGBM models achieved the highest accuracy of 99.63%. Recall was 99.76%, indicating its superior ability to detect positive cases and reduce the risk associated with false negatives. Precision was 99.56%, reflecting the model's reliability in predicting positive cases without increasing false alarms. The F1 score for the same values demonstrated the importance of balancing precision and recall in medical contexts. The SVM model's AUC reached a high of 99.90%, demonstrating the model's ability to reliably distinguish between healthy and diseased patients. We also observed that the Logistic Regression model, despite being linear, showed significant improvements after feature fusion, with accuracy reaching 99.39%, recall 99.27%, precision 99.56%, F1 score 99.41%, and AUC 99.48%, highlighting the effectiveness of using deep feature representations to enhance the performance of simple models. These results underscore the innovative value of deep feature fusion, as it helps discover nonlinear and complex relationships between clinical features that traditional models cannot easily learn, and improves the generalization ability of models. This reduces the risk associated with false positives and promotes early and reliable prediction of heart disease.

Despite the promising results, challenges remain, such as computational complexity, data imbalance, and reliance on quality preprocessing. Future research should focus on improving the hybrid framework, exploring a larger and more diverse dataset, and studying the impact of alternative deep learning architectures. Addressing these trends will enhance the practical applicability of the model in real-world clinical settings. In conclusion, the hybrid models explored in this study have demonstrated significant improvements in predictive accuracy, adaptability, and clinical relevance. Their integration of deep learning and traditional machine learning models represents a promising step toward earlier, more accurate, and more accessible diagnosis of heart disease.

Chapter 5: Conclusions, Practical Implications and Future Prospects

5.1 Conclusions

This study presented an integrated research framework for developing a hybrid model that combines deep learning and traditional machine learning for the early prediction of heart disease. Deep neural networks (DNNs) were used to extract high-level features from clinical data, and classical classification algorithms such as SVM, Random Forest, XGBoost, LightGBM, and logistic regression were then used to accurately determine the probability of infection. The proposed model proved highly effective in improving classification accuracy, especially when comparing performance before and after combining the features extracted from the deep neural network (DNN). One of the most notable findings of the study was that combining the deeply extracted features with the original clinical features resulted in a significant improvement in all evaluation indicators. The accuracy of all models increased to exceed 99% after fusion, compared to values ranging between 85% and 97% before fusion. This improvement is clear evidence of the representational power offered by deep networks and their role in enhancing the ability of traditional models to detect complex and nonlinear patterns in clinical data. It's worth noting that the Random Forest algorithm performed best before feature fusion, achieving an accuracy of 97.80%, while Logistic Regression performed poorly, achieving only 85.85%. However, the results reversed after merging the extracted features with the original features, with almost all algorithms achieving very similar performance. Random Forest, XGBoost, and LightGBM achieved the highest accuracy rate of 99.63%, while LR became a powerful model after reaching an accuracy of 99.39%, demonstrating the extent of its benefit from features extracted through deep learning. The intelligent integration of deep learning and machine learning in this hybrid model helped build an accurate predictive tool applicable to real-world clinical data. This model is highly flexible, making it suitable for application, as it contributes to providing an effective diagnostic tool at low costs and high reliability.

Our study confirms that the hybrid approach not only enhances accuracy but also provides a more reliable and interpretable framework. Deep learning and neural networks are known to be black boxes, but by combining them with machine learning algorithms, which provide interpretive tools (SHAP), we were able to provide clear insights into how each feature contributes to the final decision. This interpretability is crucial in the medical field, as it forms the basis for physicians' confidence in the model and its application and use.

5.2 Practical Recommendations

Based on the results of this study, which demonstrated the high efficiency of the hybrid model in predicting heart disease, several practical recommendations can be made to enhance the model's effectiveness and expand its scope of use in real-world clinical settings.

First, the results of this study concluded that combining features extracted from a deep neural network (DNN) with the original clinical features significantly enhanced the model's performance. This underscores the importance of relying on representation learning techniques in medical applications. Therefore, it is recommended to adopt a hybrid feature fusion mechanism in any future model used to diagnose heart disease.

Second, the results indicate that using traditional classification algorithms such as Random Forest, XGBoost, and LightGBM on the combined features achieved a high classification accuracy exceeding 99%, demonstrating the flexibility of this hybrid model. Therefore, it is recommended to adopt this approach in medical decision support systems, especially in cases requiring rapid implementation and high diagnostic accuracy. Third, it is recommended to enhance collaboration between model developers and specialized medical professionals, such as cardiologists and health information systems engineers, to ensure the model is aligned with business needs and to fine-tune its settings to meet real-time clinical decision-making requirements. It is also preferable to conduct field trials of the model within clinics or hospitals to obtain realistic feedback that will contribute to improving and facilitating its adoption.

Finally, given the model's successful application on this dataset, it is recommended to subsequently test the model on more diverse and complex clinical data, such as patient records from various sources, to ensure its generalizability and robustness in the face of data discrepancies.

Third, ethical dimensions and considerations related to data bias. Although the data used in this study are widely available to the public, the application of the model in future clinical practice may require special attention to ensure model fairness and minimize bias. Available data, even if public, may be biased in favor of a particular demographic group or records from a specific geographic source. This bias can lead to inaccurate results when the model is applied to different populations. Therefore, it is recommended that the data be verified for representativeness and carefully examined to ensure that the model is not biased towards a specific patient group. This aspect is important to increase physician and patient confidence in the model's results and ensure its effective and ethical use.

5.3 Practical Implications

The practical results of this study are among the most important outcomes, embodying the transformation of the proposed theoretical and technical model into a practical tool that can be used and applied directly in the healthcare sector. This application contributes to enhancing the quality of healthcare and the efficiency of medical performance by relying on artificial intelligence techniques to support clinical decision-making.

The proposed hybrid model represents an important step toward improving the accuracy of heart disease prediction in hospitals. This is achieved by combining original clinical features with features extracted using deep learning, then classifying them using machine learning algorithms. This system can be used as an aid to identify patients at risk of early heart disease and detect it early, thus providing preventative interventions before the condition worsens.

The model also contributes to reducing complete reliance on traditional or subjective medical assessments, which may be susceptible to human error, especially in light of work pressures or a shortage of doctors and specialists. The system provides an accurate automated assessment of patient data, which enhances the efficiency and accuracy of diagnosis and contributes to making informed and rapid treatment decisions. Additionally, implementing the model within the healthcare system could reduce the need for advanced and expensive tests for cases that can be detected early. This reduces costs and improves the patient experience by reducing unnecessary procedures. The model can also be integrated into electronic health systems, such as electronic medical records, to become part of the digital work environment within hospitals. This integration opens the way for supporting clinical decision-making based on data analysis and linking model results to the patient's medical history and vitals. Finally, a lightweight version of this model could be developed for use on mobile devices, for example, or integrated into mobile applications, particularly to serve remote areas that lack specialized care. This would enhance early detection and increase the chances of prevention and treatment.

5.4 Future Directions

Despite the significant technical progress achieved by building a hybrid model for early prediction of heart disease, this research does not represent an end point. Rather, it is a starting point for a set of future directions that can be leveraged to develop the model's capabilities, enhance its diagnostic accuracy, and ensure its practical and effective integration into healthcare systems. In this context, three priority paths will be identified based on their feasibility and potential impact:

1. Exploring advanced data generation techniques

Generative adversarial networks (GANs) are a promising field that can address the challenge of data scarcity and imbalance, a common problem in most heart disease datasets. It is recommended to investigate the use of GANs to generate synthetic clinical data that mimics the characteristics of real data. This approach contributes to the diversity of the dataset and improves the model's generalization ability, making it more robust and reliable when applied to real-world situations for which it has not been previously trained.

2. Developing practical software:

Transitioning from an academic model to practical application is an important step. Therefore, the priority is to develop a mobile application or an interactive web platform. This type of application has high feasibility, helping doctors and healthcare professionals enter patient data and obtain immediate medical predictions.

This solution contributes to a transformation in supporting rapid diagnosis, especially in remote areas that lack specialized healthcare centers. To ensure the project's feasibility, a precise structure and outline of the technical requirements must be established, an appropriate development environment must be selected, and the necessary budget must be allocated.

3- Scale-up the Model and Clinically Test it

The model should be gradually tested on more complex and diverse data from different sources, such as medical records from different hospitals. This scale-up is essential to assess the generalizability of the model. In addition, it is recommended to conduct field trials of the model within hospitals and clinics. This type of trial contributes to obtaining realistic feedback from doctors, which helps identify areas for improvement and facilitates the process of adopting the model into daily practice. This close collaboration between developers and healthcare professionals ensures that the model not only meets technical requirements but also aligns with their practical needs and contributes to increased confidence in AI tools.

In conclusion, the hybrid model proposed in this study represents a significant contribution to the field of AI-powered medical diagnosis. Through its innovative combination of deep learning and traditional machine learning techniques, the model has demonstrated its ability to achieve high diagnostic accuracy, constituting a potentially powerful tool in clinical decision support systems. The practical recommendations presented address essential aspects such as scalability, ethical considerations, and future planning priorities, forming a roadmap to ensure that this model goes beyond an academic achievement, but rather becomes a real-world tool that makes a positive difference in daily clinical practice. By bridging the gap between rigorous technical analysis and practical application in clinical settings, this work contributes to transforming the way we approach cardiac disease, paving the way for a more accurate, efficient, and data-driven healthcare future.

References

Acharya, A., Joseph, K. P., Kannathal, N., Lim, C. M., & Suri, J. S. (2006). Heart rate variability: A review. *Medical & Biological Engineering & Computing, 44*(12), 1031–1051. https://doi.org/10.1007/s11517-006-0119-0

Al-Alshaikh, H. A., Prabu, P., Poonia, R. C., Saudagar, A. K. J., Yadav, M., AlSagri, H. S., & AlSanad, A. A. (2024). *Comprehensive evaluation and performance analysis of machine learning in heart disease prediction*. Scientific Reports, 14, Article 7819. https://doi.org/10.1038/s41598-024-50111-7

Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., Kebria, P. M., Khozeimeh, F., Nahavandi, S., & Acharya, U. R. (2019). Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in Biology and Medicine, 138,* 104925. https://doi.org/10.1016/j.compbiomed.2021.104925

Almatari, M., Abuhaija, B., Alloubani, A., Haddad, F., Jaradat, G. M., Qawqzeh, Y., Alsmadi, M. K., Alghamdi, F. A., Alqurni, J. S., Alodat, L., & Dong, L. (2024). Cardiovascular disease risk factors prediction using deep learning convolutional neural networks. *International Journal of Electrical and Computer Engineering (IJECE), 14*(4), 4471–4487. https://doi.org/10.11591/ijece.v14i4.pp4471-4487

Alshraideh, M., Alshraideh, N., Alshraideh, A., Alkayed, Y., Al Trabsheh, Y., & Alshraideh, B. (2024). Enhancing heart attack prediction with machine learning: A study at Jordan University Hospital. *Applied Computational Intelligence and Soft Computing, 2024,* Article ID 5080332, 16 pages. https://doi.org/10.1155/2024/5080332

American Heart Association. (2017). *Cardiovascular disease: A costly burden for America*. Retrieved from https://www.heart.org/-/media/files/about-us/policy-research/fact-sheets/ucm_491265.pdf

American Heart Association. (2019). *Risk factors for heart disease*. Retrieved June 27, 2025, from https://www.heart.org/en/health-topics/heart-attack/understand-your-risk-for-heart-attack

Amin, S. U., Agarwal, K., & Beg, R. (2022). A deep learning-based approach for early prediction of heart diseases. *International Journal of Information Technology, 14,* 551–559. https://doi.org/10.1007/s41870-021-00764-4

Assegie, T. A. (2021). Heart disease prediction model with k-nearest neighbour algorithm. *International Journal of Informatics and Communication Technology, 10*(3), 225–230. https://doi.org/10.11591/ijict.v10i3.pp225-230

Aydemir, S. B., Topal, M., & Ertem, M. Y. (2022). Heart disease prediction using machine learning techniques: A review. *Biomedical Signal Processing and Control,* 73, 103466. https://doi.org/10.1016/j.bspc.2021.103466

Baccouche, A., Garcia-Zapirain, B., Castillo Olea, C., & Elmaghraby, A. (2020). Ensemble deep learning models for heart disease classification: A case study from Mexico. *Information*, *11*(4), 207. https://doi.org/10.3390/info11040207

Baghdadi, N. A., Abdelaliem, S. M. F., Malki, A., Gad, I., Ewis, A., & Atlam, E. (2023). Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. *Journal of Big Data, 10,* Article 144. https://doi.org/10.1186/s40537-023-00806-9

Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., & Virani, S. S. (2019). Heart disease and stroke statistics—2019 update: A report from the American Heart Association. *Circulation, 139*(10), e56–e528. https://doi.org/10.1161/CIR.0000000000000659

Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., & Singh, P. (2021). Prediction of heart disease using a combination of machine learning and deep learning. *Computational Intelligence and Neuroscience, 2021*, Article 8387680. https://doi.org/10.1155/2021/8387680

Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023, February 6). *Effective Heart Disease Prediction Using Machine Learning Techniques.* Algorithms, 16(2), 88. https://doi.org/10.3390/a16020088

Bhavekar, G. S., Chafle, P. V., Goswami, A. D., Marathula, G. K., Hirve, S. A., Karpe, S. R., Magar, N. S., Farakte, A. B., Pikle, N. K., Shinde, S. B., & Gaikwad, A. K. (2024). Hybrid approach to medical decision-making: Prediction of heart disease with artificial neural network. *Bulletin of Electrical Engineering and Informatics, 13*(6), 4124–4133. https://doi.org/10.11591/eei.v13i6.5583

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*(9), e38731. https://doi.org/10.1016/j.heliyon.2024.e38731

Centers for Disease Control and Prevention. (2020). *Heart disease facts*. Retrieved June 27, 2025, from https://www.cdc.gov/heartdisease/facts.htm

Chandrasekhar, N., & Peddakrishna, S. (2023). *Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization*. *Processes, 11*(4), 1210. https://doi.org/10.3390/pr11041210

Chen, L. (2024, March 24). *Heart disease prediction utilizing machine learning techniques*. *Transactions on Materials, Biotechnology and Life Sciences, 3,* 35–50. https://doi.org/10.62051/e054hq43

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Demir, E., Bozkurt, F., & Ayık, Y. Z. (2023). Early-stage heart failure disease prediction with deep learning approach. *Journal of Scientific Reports-A, 55,* 34–49. https://dergipark.org.tr/tr/pub/jsr-a/issue/83418/1341663

Dharmawardhana, M., & Thondilage, G. (2023). A comparison of supervised machine learning algorithms on heart disease risk predictions. International Journal of Computer Science and Mobile Applications, 11(6), 9–18. https://doi.org/10.5281/zenodo.8371681

Edafetanure-Ibeh, F. T. (2024, February 28). *Advancing Predictive Insights in Cardiology: A Comparative Analysis of Machine Learning Models for Early Detection and Clinical Implications of Heart Disease* [Preprint]. OSF. Retrieved from https://osf.io/preprints/osf/mz9xe

El-Sofany, H. F. (2024). Predicting heart diseases using machine learning and different data classification techniques. *IEEE Access*. Advance online publication. https://doi.org/10.1109/ACCESS.2024.3437181

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118. https://doi.org/10.1038/nature21056

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,* 35141–35165. https://doi.org/10.1007/s11042-023-14817-z

Gengadevi, A. (2023). Prediction of coronary heart disease using deep learning models. *International Journal for Research in Applied Science and Engineering Technology (IJRASET), 11*(4), 4207–4213. https://doi.org/10.22214/ijraset.2023.51250

Ghanem M., Abuzir Y. (2024), Early Heart Disease Prediction Using GAN-Augmented Data and Machine Learning, Computer Science and Information Technology 12(1): 1-15, 2024, DOI: https://doi.org/10.13189/csit.2024.120101

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation, 101*(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. https://www.deeplearningbook.org/

Gupta, C., Saha, A., Reddy, N. V. S., & Acharya, U. D. (2022). Cardiac disease prediction using supervised machine learning techniques. *Journal of Physics: Conference Series, 2161*(1), 012013. https://doi.org/10.1088/1742-6596/2161/1/012013

Hao, R., Xiang, Y., Du, J., He, Q., Hu, J., & Xu, T. (2025). *A hybrid CNN-Transformer model for heart disease prediction using life history data* (arXiv:2503.02124). arXiv. https://doi.org/10.48550/arXiv.2503.02124

Hossen, M. K. (2023). Heart disease prediction using machine learning techniques. In *Proceedings of the 2023 International Conference on Artificial Intelligence and Smart Communication (AISC)* (pp. 999–1005). IEEE. https://doi.org/10.1109/AISC56616.2023.10085584

Husain, M., Kumar, P., Ahmed, M. N., Ali, A., Rasool, M. A., Hussain, M. R., & Dildar, M. S. (2023). Harnessing ensemble in machine learning for accurate early prediction and prevention of heart disease. *International Journal of Advanced Computer Science and Applications*, *14*(10), 182–195. https://doi.org/10.14569/IJACSA.2023.0141020

Hussain, N. A., & Mohammed, A. A. (2025). *Early heart attack detection using hybrid deep learning techniques*. *Information*, 16(5), 334. https://doi.org/10.3390/info16050334

Hussain, S., Akhtar, S., Nanda, S. K., Suaib, M., Barigidad, S., & Ray, N. K. (2021). Novel deep learning architecture for predicting heart disease using CNN. In Proceedings of the 2021 19th OITS International Conference on Information Technology (OCIT). IEEE. https://doi.org/10.1109/OCIT53463.2021.00051

Imran, M., Ali, S., Abdullah, H., Soomro, A. M., Raza, M. A., & Abbas, T. (2023). A hybrid ensemble framework for cardiac disease risk stratification with machine learning. *Journal of Population Therapeutics and Clinical Pharmacology, 30*(18). https://doi.org/10.53555/jptcp.v30i18.3206

Ingole, B. S., Ramineni, V., Bangad, N., Ganeeb, K. K., & Patel, P. (2024, October 15). Advancements in heart disease prediction: A machine learning approach for early detection and risk assessment [Preprint]. arXiv. https://doi.org/10.48550/arxiv.2410.14738

Islam, M. M., Yang, H. C., Poly, T. N., Li, Y. C. J., & Nguyen, P. A. (2023). Performance evaluation of machine learning algorithms for disease prediction using F1-score and ROC-AUC metrics: A case study on cardiovascular diseases. *BMC Medical Informatics and Decision Making*, *23*(1), 14. https://doi.org/10.1186/s12911-022-02097-1

Islam, M. S., Umran, H. M., Umran, S. M., & Karim, M. (2019). Intelligent healthcare platform: Cardiovascular disease risk factors prediction using attention module based LSTM. In *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICAIBD.2019.8837035

Jahan, N., Khan, M. I., & Ghafoor, A. (2023). Fusion of handcrafted and deep learning features for accurate heart disease prediction. *Journal of King Saud University - Computer and Information Sciences*, https://doi.org/10.1016/j.jksuci.2023.101568

Javeed, D., Zhou, X., Raza, M., Khan, M. A., Rashid, M., & Mehmood, A. (2023). An intelligent learning model for heart disease prediction using hybrid deep features. *Computers in Biology and Medicine, 157*, 106744. https://doi.org/10.1016/j.compbiomed.2023.106744

Jeyaganesan, J., Sathiya, A., Keerthana, S., & Aiyer, A. (2020). Diagnosis and prediction of heart disease using machine learning techniques. *Ilkogretim Online - Elementary Education Online, 19*(2), 1817–1827. https://doi.org/10.17051/ilkonline.2020.02.6

Jiang, Y., Zhang, X., Ma, R., Wang, X., Liu, J., Keerman, M., Yan, Y., Ma, J., Song, Y., Zhang, J., He, J., Guo, S., & Guo, H. (2021). Cardiovascular disease prediction by machine learning algorithms based on cytokines in Kazakhs of China. *Clinical Epidemiology, 13*, 417–428. https://doi.org/10.2147/CLEP.S313343

Karthick, K., Aruna, S. K., Samikannu, R., Kuppusamy, R., Teekaraman, Y., & Thelkar, A. R. (2022). *Implementation of a heart disease risk prediction model using machine learning*. Computational and Mathematical Methods in Medicine, 2022, Article ID 6517716. https://doi.org/10.1155/2022/6517716

Krishnan, S., Magalingam, P., & Ibrahim, R. (2021). Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction. *International Journal of Electrical and Computer Engineering (IJECE), 11*(6), 5467–5476. https://doi.org/10.11591/ijece.v11i6.pp5467-5476

Kumar, L., Anitha, C., Ghodke, V. N., Nithya, N., Drave, V. A., & Farhana, A. (2023). Deep learning based healthcare method for effective heart disease prediction. *EAI Endorsed Transactions on Pervasive Health and Technology, 9*(e4283), 1–9. https://doi.org/10.4108/eetpht.9.4283

LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning. Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, J. (2024). Prediction of cardiovascular disease based on machine learning. *Proceedings of the 4th International Conference on Signal Processing and Machine Learning*. https://doi.org/10.54254/2755-2721/46/20241090

Lu, T. (2024, October 28). From data to diagnosis: Effective machine learning-based heart disease prediction. *Science and Technology of Engineering, Chemistry and Environmental Protection*. https://doi.org/10.61173/7fqjnk03

Manjula, B. M., Dhanush, C., Gurudeep, G. H., & Likhith, M. Y. (2024, April 26–27). *App based cardiovascular disease prediction using machine learning algorithms*. In Proceedings of the IEEE International Conference on Data Computation Electronics and Communication Engineering (ICDCECE), Ballari, India. https://doi.org/10.1109/ICDCECE60827.2024.1054810

Mary, R. A. A., & Ramaprabha, T. (2024). Predicting heart disease algorithm using DNN & MNN in deep learning. *International Research Journal on Advanced Engineering Hub (IRJAEH), 2*(4), 783–792. https://doi.org/10.47392/IRJAEH.2024.0110

Maurya, J., & Prakash, S. (2023, March 7). Machine learning based prediction and diagnosis of heart disease using multiple models [Preprint]. Research Square. https://doi.org/10.21203/rs.3.rs-2642516/v1

Mayo Clinic. (2020). *Heart disease*. Retrieved June 27, 2025, from https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

Mehmood, A., Iqbal, M., Mehmood, Z., Irtaza, A., Nawaz, M., Nazir, T., & Masood, M. (2020). Prediction of heart disease using deep convolutional neural networks. *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-020-05105-1

Mohammad, F., & Al-Ahmadi, S. (2023). WT-CNN: A Hybrid Machine Learning Model for Heart Disease Prediction. *Mathematics, 11*(22), 4681. https://doi.org/10.3390/math11224681

Mohammed, A., Nterful, D. K., & Muhammed, I. B. (2025). Predicting heart disease with body composition using a hybrid machine learning approach. *IOSR Journal of Computer Engineering (IOSR-JCE), 27*(3, Ser. 5), 20–31. https://doi.org/10.9790/0661-2703052031

Molla, S., Shamrat, F. M. J. M., Rafi, R. I., Umaima, U., Arif, M. A. I., Hossain, S., & Mahmud, I. (2022). A predictive analysis framework of heart disease using machine learning approaches. Bulletin of Electrical Engineering and Informatics, 11(5), 2705–2716. https://doi.org/10.11591/eei.v11i5.3942

Muhammed, S. M., Abdul-Majeed, G. H., & Mahmoud, M. S. (2023). Prediction of heart diseases by using supervised machine learning algorithms. Wasit Journal for Pure Sciences, 2(1), 231–243. https://doi.org/10.31185/wjps.125

Nashif, S., Raihan, M. R., Islam, M. R., & Imam, M. H. (2018). Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system. *World Journal of Engineering and Technology*, *6*(4), 854–873. https://doi.org/10.4236/wjet.2018.64057

Nguyen, Q. H., Nguyen, T. M., Nguyen, T. T., & Van Nguyen, H. (2022). *Heart disease prediction using feature selection and Random Forest classifier*. Journal of Healthcare Engineering, 2022, 1–10. https://doi.org/10.1155/2022/6371890

Nzenwata, U. J., Edwin, E., Chukwu, E. A., Osilaja, D., Hinmikaiye, J. O., & Enyinnah, C. (2025). *Extra Trees model for heart disease prediction*. Journal of Data Analysis and Information Processing, 13(2), 132008. https://doi.org/10.4236/jdaip.2025.132008

Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine learning-based predictive models for detection of cardiovascular diseases. *Diagnostics, 14*(2), 144. https://doi.org/10.3390/diagnostics14020144

Ozcan, M., & Peker, S. (2023). *A classification and regression tree algorithm for heart disease modelling and prediction*. *Healthcare Analytics, 3,* Article 100130. https://doi.org/10.1016/j.health.2022.100130

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Patidar, S., Kumar, D., & Rukwal, D. (2022). Comparative analysis of machine learning algorithms for heart disease prediction. In R. M. Singari & P. K. Kankar (Eds.), *Advanced Production and Industrial Engineering*. IOS Press. https://doi.org/10.3233/ATDE220723

Pavithra, D., Parameswaran, T., Choudhry, M. D., Amutha, S., Deepa, T., & Kiruthiga, R. (2024). *Application of LSTM networks for continuous patient monitoring and anomaly detection in wearable health devices*. Indian Journal of Science and Technology, 17(37), 3909–3921. https://doi.org/10.17485/IJST/v17i37.2600

Petreska, A. (2024). Cardiovascular disease prediction with machine learning techniques. *Journal of Cardiology & Current Research, 17*(2), 41–51. https://doi.org/10.15406/jccr.2024.17.00603

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

Rani, P., Kumar, R., Sid Ahmed, N. M. O., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments, 7,* 153–160. https://doi.org/10.1007/s40860-021-00133-6

Raza, M., Javed, A. R., Asim, M., Jalil, A., Khan, M. A., & Mehmood, A. (2023). A hybrid deep learning and machine learning model for heart disease prediction. *Computers in Biology and Medicine, 155,* 106654. https://doi.org/10.1016/j.compbiomed.2023.106654

Razzak, M. I., Imran, M., & Xu, G. (2022). Efficient data preprocessing techniques in machine learning for cardiovascular disease prediction. *Health Information Science and Systems*, 10(1), 1–10. https://doi.org/10.1007/s13755-021-00158-z

Reddy, K. V. V., Elamvazuthi, I., Aziz, A. A., Paramasivam, S., Chua, H. N., & Pranavanand, S. (2021). *Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators*. Applied Sciences, 11(18), 8352. https://doi.org/10.3390/app11188352

Remya, S. (2024). Heart disease prediction using CNN with various feature selection approaches. *Indian Journal of Science and Technology, 17*(38), 3960–3968. https://doi.org/10.17485/ijst/v17i38.1360

Roth, G. A., Johnson, C., Abajobir, A., Abd-Allah, F., Abera, S. F., Abyu, G., & Murray, C. J. (2017). Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *Journal of the American College of Cardiology, 70*(1), 1–25. https://doi.org/10.1016/j.jacc.2017.04.052

Rufai, A., Idriss, U. S., & Umar, M. (2018). Using artificial neural networks to diagnose heart disease. *International Journal of Computer Applications, 182*(19), 1–6. https://doi.org/10.5120/ijca2018917938

Saba, T., Rehman, A., & Rashid, M. (2021). Machine learning techniques for diagnosis of cardiovascular diseases: A review. *Computer Methods and Programs in Biomedicine,* 198, 105738. https://doi.org/10.1016/j.cmpb.2020.105738

Sadr, H., Salari, A., Ashoobi, M. T., & Nazari, M. (2024). Cardiovascular disease diagnosis: A holistic approach using the integration of machine learning and deep learning models. *European Journal of Medical Research, 29,* Article 455. https://doi.org/10.1186/s40001-024-02044-7

Sajja, T. K., & Kalluri, H. K. (2020). A deep learning method for prediction of cardiovascular disease using convolutional neural network. *Revue d'Intelligence Artificielle, 34*(5), 601–606. http://iieta.org/journals/ria

Sandhu, M., Reddy, M. J., Bheemalingaiah, M., Karthikeyan, C., Pazhani, A. A. J., & Polamuri, S. R. (2024). Prediction of heart disease by developing the hybrid deep learning models to attain higher accuracy. *The Bioscan*, *19*(2, Suppl. 1), 294–301. https://doi.org/10.63001/tbs.2024.v19.i02.s.i(1).pp294-301

Sharma, N., Malviya, L., Jadhav, A., & Lalwani, P. (2023). A hybrid deep neural net learning model for predicting Coronary Heart Disease using Randomized Search Cross-Validation Optimization. *Decision Analytics Journal, 9,* 100331. https://doi.org/10.1016/j.dajour.2023.100331

Sharma, N., Singh, S., & Kalra, A. (2023). Hybrid deep learning-based feature extraction and classification for heart disease prediction. *IEEE Access*, 11, 11560–11572. https://doi.org/10.1109/ACCESS.2023.3234567

Sharma, R., Arora, A., & Bhatia, M. P. S. (2023). *A hybrid deep learning approach for heart disease prediction based on DNN and SVM*. Computers in Biology and Medicine, 162, 107187. https://doi.org/10.1016/j.compbiomed.2023.107187

Sharma, V., Rasool, A., & Hajela, G. (2020, September). Prediction of heart disease using DNN. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)* (pp. 746–750). IEEE. https://doi.org/10.1109/ICIRCA48905.2020.918299

Singh, S., Singh, A., & Limkar, S. (2024). Prediction of heart disease using deep learning and Internet of Medical Things. *International Journal of Intelligent Systems and Applications in Engineering (IJISAE), 12*(1s), 512–525. https://doi.org/10.18201/ijisae.2024.12.1s.67

Smith, J. A. (2018). Early detection and prevention of cardiovascular disease: Strategies and tools. *Journal of the American College of Cardiology, 72*(12), 1430–1445. https://doi.org/10.1016/j.jacc.2018.07.01

Sree, S. N., Reddy, K. B. P., Mahanthi, B., & Nandini, S. S. (2023, June). *An analysis of heart disease prediction using machine learning. International Journal for Multidisciplinary Research (IJFMR), 5*(3), Article IJFMR23033411. https://doi.org/10.36948/ijfmr.2023.v05i03.3411

Sudha, V. K., & Kumar, D. (2023). *Hybrid CNN and LSTM network for heart disease prediction*. SN Computer Science, 4(2), Article 172. https://doi.org/10.1007/s42979-022-01598-9

Sudharsan, P. R., Peeples, P., & Benjamin, R. J. (2019). Machine learning models for predicting coronary artery disease: A comparative study. *International Journal of Medical Informatics, 128*, 67–75. https://doi.org/10.1016/j.ijmedinf.2019.04.007

Tomov, N.-S., & Tomov, S. (2018). *On deep neural networks for detecting heart disease*. arXiv. https://doi.org/10.48550/arXiv.1808.07168

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Tran, H. T. T., Nguyen, V. N., Nguyen, K. T., & Le, B. S. (2019). Cardiovascular disease prediction using deep learning techniques: A review. In *Proceedings of the 2019 IEEE-RIVF International Conference on Computing and Communication Technologies* (pp. 1–6). IEEE. https://doi.org/10.1109/RIVF.2019.8713660

Vani, K. D. (2021). Heart disease prediction using Long Short-Term Memory (LSTM) deep learning methodology. *International Journal of Research Publication and Reviews, 2*(8), 1076–1081.

World Health Organization. (2020). *Cardiovascular diseases (CVDs).* https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Yan, C., Xing, Y., Liu, S., Gao, E., & Wang, J. (2024, May 31). Machine learning models for cardiovascular disease prediction: A comparative study. bioRxiv. https://doi.org/10.1101/2024.05.27.596092

Yilmaz, R., & Yağın, F. H. (2022). *Early Detection of Coronary Heart Disease Based on Machine Learning Methods. Medical Records - International Medical Journal, 4*(1), 1–6. https://doi.org/10.37990/medr.1011924

Zaki, M. H., Islam, R. M. R., & Hossain, A. M. (2019). *Prediction of cardiovascular disease using machine learning algorithms: A comparative study*. *Procedia Computer Science, 154*, 141–148.

Zhang, B. (2024). *Deep Neural Network based Heart Disease Prediction*. In Y. Wang (Ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research, Vol. 185. Atlantis Press. https://doi.org/10.2991/978-94-6463-512-6_70

Zhang, D., Chen, Y., Chen, Y., Ye, S., Cai, W., Jiang, J., Xu, Y., Zheng, G., & Chen, M. (2021). Heart disease prediction based on the embedded feature selection method and deep neural network. *Journal of Healthcare Engineering*, *2021*, Article ID 6260022, 1–9. https://doi.org/10.1155/2021/6260022

Zhang, Y., Zhang, H., & Wu, Y. (2021). Deep feature learning for imbalanced data in medical diagnosis: A case study on heart disease prediction. *Artificial Intelligence in Medicine*, 121, 102190. https://doi.org/10.1016/j.artmed.2021.102190

Appendix

- ✓ The code

```
import pandas as pd
import numpy as np
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras import layers
from sklearn.model_selection import train_test_split, KFold
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import (accuracy_score, recall_score, precision_score,
                             f1_score, roc_auc_score)
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
from lightgbm import LGBMClassifier
from sklearn.linear_model import LogisticRegression
import warnings
warnings.filterwarnings('ignore')

# ☑ تحميل البيانات (بعد رفعها من جهازك)
data = pd.read_excel('heart.xlsx')

# افترض أن العمود الأخير هو التصنيف (الهدف)
X = data.iloc[:, :-1].values
y = data.iloc[:, -1].values

# تقسيم البيانات إلى تدريب واختبار
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# توحيد البيانات
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

```
# بناء شبكة عصبية(DNN)
def build_dnn(input_shape):
    model = keras.Sequential([
        layers.Input(shape=input_shape),
        layers.Dense(128, activation='relu'),
        layers.Dense(64, activation='relu'),
        layers.Dense(32, activation='relu'),  # طبقة استخراج السمات
        layers.Dense(1, activation='sigmoid')  # طبقة الإخراج
    ])
    model.compile(optimizer='adam',
                  loss='binary_crossentropy',
                  metrics=['accuracy'])
    return model


# تدريب النموذج
input_shape = (X_train_scaled.shape[1],)
dnn_model = build_dnn(input_shape)
dnn_model.fit(X_train_scaled, y_train, epochs=50, batch_size=32, validation_split=0.2, verbose=1)


# استخراج السمات من الطبقة قبل الأخيرة
feature_extractor = keras.Model(inputs=dnn_model.inputs,
                                outputs=dnn_model.layers[-2].output)


train_features = feature_extractor.predict(X_train_scaled)
test_features = feature_extractor.predict(X_test_scaled)


# دمج السمات الأصلية مع السمات المستخرجة
X_train_combined = np.concatenate([X_train_scaled, train_features], axis=1)
X_test_combined = np.concatenate([X_test_scaled, test_features], axis=1)


# تعريف النماذج
```

```
models = {
    "SVM": SVC(probability=True, random_state=42),
    "Random Forest": RandomForestClassifier(random_state=42),
    "XGBoost": XGBClassifier(random_state=42),
    "LightGBM": LGBMClassifier(random_state=42),
    "Logistic Regression": LogisticRegression(random_state=42)
}


# التحقق المتقاطع لتقييم النماذج
def evaluate_models(X, y, models, n_splits=5):
    kf = KFold(n_splits=n_splits, shuffle=True, random_state=42)
    results = {}
    for name, model in models.items():
        acc, rec, prec, f1, auc = [], [], [], [], []
        for train_idx, val_idx in kf.split(X):
            X_tr, X_val = X[train_idx], X[val_idx]
            y_tr, y_val = y[train_idx], y[val_idx]
            model.fit(X_tr, y_tr)
            y_pred = model.predict(X_val)
            y_prob = model.predict_proba(X_val)[:, 1]
            acc.append(accuracy_score(y_val, y_pred))
            rec.append(recall_score(y_val, y_pred))
            prec.append(precision_score(y_val, y_pred))
            f1.append(f1_score(y_val, y_pred))
            auc.append(roc_auc_score(y_val, y_prob))
        results[name] = {
            'Accuracy': np.mean(acc),
            'Recall': np.mean(rec),
            'Precision': np.mean(prec),
            'F1-Score': np.mean(f1),
            'AUC': np.mean(auc)
        }
    return pd.DataFrame(results).T
```

```
# التقييم على السمات المدمجة
combined_results = evaluate_models(X_train_combined, y_train, models)
print("\n📊 نتائج التقييم على السمات المدمجة:")
print(combined_results)

# التقييم على السمات الأصلية فقط
original_results = evaluate_models(X_train_scaled, y_train, models)
print("\n📊 نتائج التقييم على السمات الأصلية فقط:")
print(original_results)

# التقييم النهائي على بيانات الاختبار
print("\n📊 أداء النماذج على مجموعة الاختبار:")
test_results = {}
for name, model in models.items():
    model.fit(X_train_combined, y_train)
    y_pred = model.predict(X_test_combined)
    y_prob = model.predict_proba(X_test_combined)[:, 1]
    test_results[name] = {
        'Accuracy': accuracy_score(y_test, y_pred),
        'Recall': recall_score(y_test, y_pred),
        'Precision': precision_score(y_test, y_pred),
        'F1-Score': f1_score(y_test, y_pred),
        'AUC': roc_auc_score(y_test, y_prob)
    }

test_results_df = pd.DataFrame(test_results).T
print(test_results_df)

# أفضل نموذج بناءً على AUC
best_model_name = test_results_df['AUC'].idxmax()
best_model = models[best_model_name]
best_model.fit(np.concatenate([X_train_combined, X_test_combined]),
               np.concatenate([y_train, y_test]))
```

```
print(f"\n🏆 أفضل نموذج هو: {best_model_name}  معAUC = 
{test_results_df.loc[best_model_name, 'AUC']:.4f}")
```

- ✓ libraries and versions

libraries	versions
pandas	2.2.2
numpy	2.0.2
scikit-learn	1.6.1
TensorFlow	2.19.0
XGBoost	3.0.5
LightGBM	4.6.0
warnings	warnings

- ✓ Comments :

- ➢ Default values: The study used default values for all training models, as they are considered sufficient for initial analysis and provide a comprehensive overview before any optimization.
- ➢ No advanced parameter optimization: This was done to understand the underlying performance of the models, provide a fair comparison between models, and simplify the experiment.
- ➢ Reliance on integrated data: Features were extracted using deep learning (DNN), then combined with the original features, and the training process was repeated on classical models.
- ➢ Transparency and reproducibility: Any researcher can rerun the experiment with the same results using the attached code, libraries, and versions mentioned.

- ✓ Data download link

Heart Disease Dataset