As per UGC guidelines an electronic bar code is provided to secure your paper

International Journal for Modern Trends in Science and Technology
Volume 12, Issue 06, pages 308-324.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue06.html
DOI: https://doi.org/10.5281/zenodo.20739563

Heart Disease Prediction by Using a Stacked Ensemble Learning Approach

U. Jenny Grace, Satti Sri Divya Sai Lakshmi, Nadapana Himasri, Gundabathula Satyanarayana

Department of Information Technology Swarnandhra College of Engineering & Technology, Seetharamapuram, Narsapur, Andhra Pradesh, INDIA.

To Cite this Article
U. Jenny Grace, Satti Sri Divya Sai Lakshmi, Nadapana Himasri & Gundabathula Satyanarayana (2026). Heart Disease Prediction by Using a Stacked Ensemble Learning Approach. International Journal for Modern Trends in Science and Technology, 12(06), 308-324. https://doi.org/10.5281/zenodo.20739563

Article Info
Received: 16 May 2026; Revised: 07 May 2026; Accepted: 12 June 2026.

KEYWORDS

Heart Disease Prediction, Machine Learning, Stacking Ensemble Learning, Logistic Regression, Decision Tree, Random Forest, Clinical Decision Support System, Medical Data Analysis, Predictive Analytics, Healthcare Analytics, Classification Models, Data Preprocessing

ABSTRACT

Heart disease is one of the leading causes of mortality worldwide, making early and accurate prediction essential for effective diagnosis and treatment. This project presents a Heart Disease Prediction System using a Stacking Ensemble Learning Approach, which integrates multiple machine learning models to enhance predictive performance. The proposed system utilizes Logistic Regression, Decision Tree, and Random Forest as base learners to analyse key medical attributes such as age, blood pressure, cholesterol levels, heart rate, and other clinical indicators. Logistic Regression provides a strong baseline through probabilistic classification, while Decision Tree captures nonlinear relationships within the data. Random Forest further improves robustness and reduces overfitting by aggregating multiple decision trees. The outputs of these base models are combined using a meta-classifier in the stacking ensemble framework to generate the final prediction. Data preprocessing techniques including normalization and handling of missing values are applied to improve model reliability. Experimental results demonstrate that the stacking ensemble approach outperforms individual classifiers in terms of accuracy and stability. The proposed system acts as an effective clinical decision support tool for early heart disease detection, aiding healthcare professionals in timely and informed decision-making.

1. INTRODUCTION

Heart disease is one of the leading causes of mortality worldwide and continues to pose a major challenge to global healthcare systems. The increasing number of patients affected by cardiovascular diseases highlights the urgent need for accurate, efficient, and early prediction methods.

Early detection of heart disease plays a crucial role in preventing severe complications and reducing mortality rates by enabling timely medical intervention. However, traditional diagnostic methods largely depend on manual analysis of clinical data and physician expertise, which may not effectively capture complex patterns and hidden relationships present in medical datasets. With the rapid advancement of artificial intelligence (AI), particularly in the domain of machine learning (ML), significant progress has been made in the development of automated disease prediction systems.

Machine learning techniques have demonstrated the ability to analyze large volumes of healthcare data, identify patterns, and generate predictive insights with higher accuracy and efficiency compared to conventional methods.

These techniques enable healthcare professionals to make informed decisions by providing data-driven support systems. Studies have shown that machine learning models can significantly improve diagnostic accuracy and assist in early-stage detection of heart disease, thereby enhancing clinical outcomes.

Despite the effectiveness of individual machine learning algorithms, single-model approaches often suffer from limitations such as overfitting, bias, and reduced generalization capability when applied to real-world datasets. Different algorithms capture different aspects of data patterns, and relying on a single model may not provide optimal performance.

To overcome these limitations, ensemble learning techniques have been introduced, which combine multiple models to improve prediction accuracy and robustness This project focuses on the design and implementation of a heart disease prediction system using a stacked ensemble learning approach. In this approach, multiple base models such as Random Forest, Support Vector Machine (SVM), and XGBoost are trained independently on the dataset to learn diverse patterns and relationships.

These base models generate predictions that are further combined using a meta-learner, specifically Logistic Regression, which produces the final output. The stacked ensemble technique leverages the strengths of each individual algorithm and improves overall model performance by learning how to optimally combine their predictions.

1.1 OVERVIEW OF THE PROJECT

The primary goal of this project is to design and implement a highly accurate and reliable heart disease prediction system by leveraging stacked ensemble learning techniques. The system aims to assist in early diagnosis by analyzing patient health parameters and predicting the likelihood of heart disease using advanced machine learning models. Early detection of cardiovascular disease is crucial, as timely intervention can significantly reduce mortality rates and improve patient outcomes.

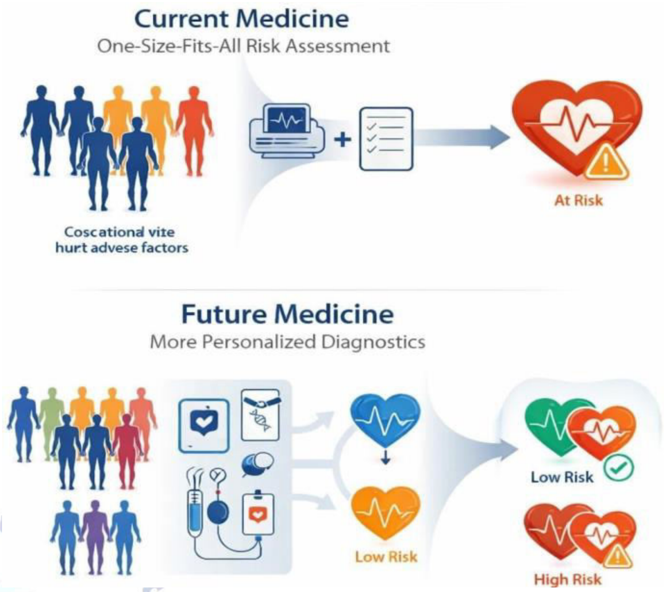

Fig 1.1: The Evolution of Personalized Medicine

In recent years, machine learning has emerged as a powerful tool in healthcare for analyzing complex medical datasets and identifying hidden patterns that are difficult to detect using traditional methods. However, individual machine learning models often face challenges such as overfitting, bias, and limited generalization capability when applied to real-world clinical data. To address these limitations, ensemble learning techniques have been widely adopted, as they combine multiple models to improve prediction accuracy, robustness, and stability. Studies have shown that ensemble approaches, particularly stacking, consistently outperform individual models in heart disease prediction tasks by effectively integrating the strengths of multiple This project utilizes a stacked ensemble learning approach in which multiple base models, including Random Forest, Support Vector Machine (SVM), and XGBoost, are trained independently on the dataset. Each model learns different patterns and relationships within the data, contributing unique predictive insights. The outputs of these base models are then combined using a meta-learner, specifically Logistic Regression, which generates the final prediction. This layered architecture enhances the model's ability to capture complex nonlinear relationships and improves overall predictive performance.

The project follows a systematic workflow that includes data preprocessing, feature engineering, model training, and evaluation. Data preprocessing involves handling missing values, encoding categorical variables, and scaling features to ensure consistency and accuracy. The dataset is then divided into training and testing sets, and multiple models are trained to learn diverse feature interactions. The stacking mechanism integrates these models to produce a more generalized and accurate prediction system.

Furthermore, the developed system is integrated into a user-friendly web application, allowing users to input medical parameters and receive real-time predictions regarding heart disease risk. The performance of the system is evaluated using standard metrics such as accuracy, precision, recall, and F1-score, ensuring the reliability and effectiveness of the model. Overall, this project demonstrates the practical application of stacked ensemble learning in healthcare and highlights its potential in improving the accuracy of heart disease prediction systems. By combining multiple machine learning algorithms into a unified framework, the proposed system provides a robust and efficient solution for early diagnosis and clinical decision support.

1.2 IMPORTANCE OF EARLY PREDICTION

Early prediction of heart disease plays a critical role in modern healthcare, as cardiovascular diseases often develop gradually and remain undetected until severe complications arise. In many cases, individuals do not experience noticeable symptoms during the initial stages, and diagnosis is only made after the occurrence of major events such as heart attacks or strokes. Therefore, identifying the risk of heart disease at an early stage is essential for preventing disease progression and improving patient survival rates. Studies indicate that early detection enables timely intervention, which significantly reduces mortality and enhances overall health outcomes. One of the major advantages of early prediction is the ability to initiate preventive measures before the condition becomes critical. By identifying high-risk individuals in advance, healthcare professionals can recommend lifestyle modifications such as diet control, regular exercise, and medication management. This proactive approach helps in reducing the severity of the disease and prevents irreversible damage to the heart. Research shows that predictive models can support clinicians in recognizing at-risk patients early, allowing them to implement preventive strategies and improve long-term cardiovascular health.

Another important aspect of early prediction is the improvement in diagnostic accuracy through the use of advanced computational techniques. Traditional diagnostic methods often rely on limited clinical parameters and linear models, which may fail to capture complex relationships within medical data. Machine learning-based prediction systems, on the other hand, can analyze large volumes of patient data and identify hidden patterns that are not easily detectable by human analysis. These systems enable more accurate and data-driven decision- making, thereby enhancing the reliability of heart disease diagnosis.

Early prediction also contributes to reducing healthcare costs and improving the efficiency of medical services. Treating heart disease at an advanced stage often requires expensive procedures such as surgeries, long-term hospitalization, and intensive care. In contrast, early detection allows for cost-effective treatment through preventive care and routine monitoring. Furthermore, it reduces the burden on healthcare systems by minimizing emergency cases and improving resource allocation.

In addition, early prediction systems play a vital role in improving the quality of life for patients. Individuals diagnosed at an early stage can manage their condition effectively and maintain a normal lifestyle with minimal restrictions. Advanced machine learning models further enhance this process by enabling remote monitoring and continuous risk assessment, making healthcare more accessible and efficient. These technologies have the potential to transform traditional healthcare into a more proactive and preventive system.

Overall, the importance of early prediction lies in its ability to prevent disease progression, improve diagnostic accuracy, reduce healthcare costs, and enhance patient quality of life. The integration of machine learning techniques into early prediction systems provides a powerful solution for addressing the growing challenges of heart disease and supports the development of intelligent healthcare systems.

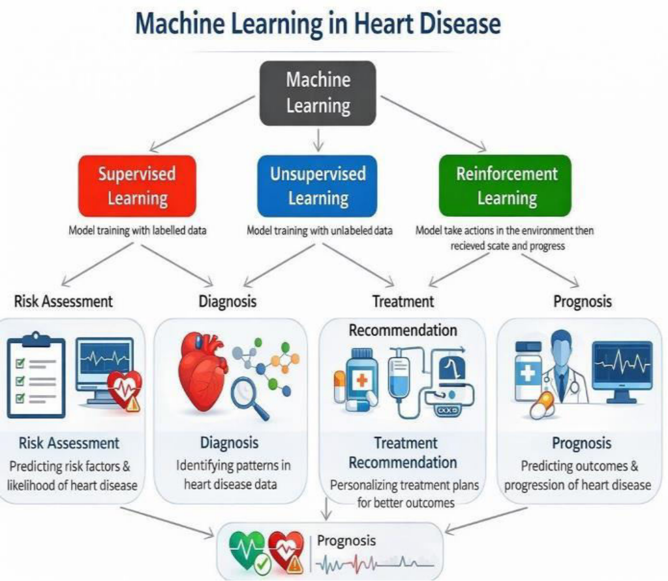

Fig 1.2: Machine Learning in Heart Disease Prediction

1.3 SCOPE OF THE PROJECT

The scope of this project extends to the development of an intelligent and efficient heart disease prediction system that can be applied in real-world healthcare environments for early diagnosis and risk assessment. With cardiovascular diseases being one of the leading causes of death globally, there is a growing need for scalable and accurate prediction systems that can assist medical professionals in identifying high-risk patients at an early stage. The proposed system addresses this need by leveraging machine learning and stacked ensemble learning techniques to provide reliable and data-driven predictions.

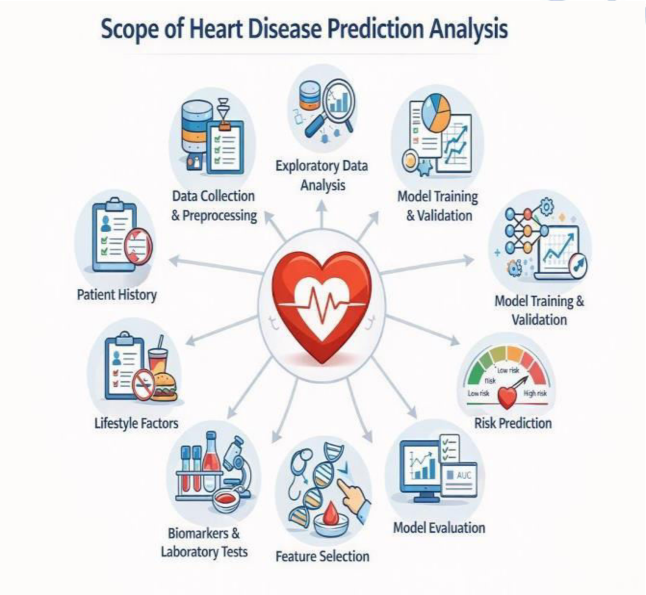

Fig 1.3: Scope of Heart Disease Prediction Analysis

In real-world clinical settings, the system can be used as a decision support tool for doctors and healthcare practitioners. By analyzing patient data such as age, cholesterol levels, blood pressure, and other medical attributes, the model can predict the likelihood of heart disease and assist clinicians in making informed decisions. Early prediction models enable healthcare providers to identify high-risk individuals before the onset of severe symptoms, allowing timely preventive measures and improving patient outcomes. This makes the system highly valuable in hospitals, diagnostic centers, and telemedicine platforms.

The project also has significant applicability in preventive healthcare and remote monitoring systems. With the increasing use of digital health technologies, the model can be integrated into web or mobile applications, enabling users to assess their heart disease risk from anywhere. This is particularly useful in rural and underserved areas where access to specialized medical facilities is limited. Machine learning-based systems can process large volumes of health data efficiently and provide quick predictions, thereby reducing the dependency on manual diagnosis and improving accessibility to healthcare services.

Another important aspect of the project's scope is its ability to handle complex and nonlinear relationships within medical data. Traditional statistical models often fail to capture these relationships, leading to less accurate predictions. In contrast, machine learning techniques, especially ensemble models, can analyze multiple risk factors simultaneously and improve prediction accuracy. Research indicates that AI-based predictive systems enhance diagnostic accuracy and reduce errors in clinical decision-making by identifying hidden patterns in patient data. This highlights the practical importance of the proposed system in real-world medical applications.

Furthermore, the system can be extended for integration with electronic health record (EHR) systems, enabling continuous monitoring and analysis of patient health data. It can also be adapted for predicting other diseases by modifying the dataset and retraining the model, making it a scalable and flexible solution. The use of stacked ensemble learning further enhances the system's robustness by combining the strengths of multiple machine learning algorithms, ensuring better generalization and reliability in diverse real-world scenarios.

Despite its advantages, the scope of the project is currently limited to prediction based on available structured data and does not replace professional medical diagnosis. However, it serves as a powerful supportive tool that can assist healthcare professionals in improving diagnostic efficiency and patient care.

1.4 ENSEMBLE LEARNING CONCEPT

Ensemble learning is a machine learning paradigm in which multiple models, often referred to as base learners, are combined to solve a particular problem and improve overall predictive performance. Instead of relying on a single model, ensemble methods integrate the predictions of several models to produce a more accurate and robust output. The fundamental idea behind ensemble learning is that different models may capture different patterns or relationships within the data, and combining them can reduce errors caused by bias, variance, or noise. Research studies have shown that ensemble techniques significantly outperform individual models in complex prediction tasks by enhancing generalization capability and stability. There are several types of ensemble learning techniques, including bagging, boosting, and stacking. Bagging (Bootstrap Aggregating) focuses on reducing variance by training multiple models on different subsets of the dataset and averaging their predictions. Boosting, on the other hand, aims to reduce bias by sequentially training models where each new model focuses on correcting the errors of the previous one. Stacking, which is the primary technique used in this project, combines multiple base models and uses an additional model, known as a meta-learner, to learn how to best combine their predictions. Among these techniques, stacking is considered highly effective for complex datasets because it leverages the strengths of different algorithms simultaneously.

In this project, the ensemble learning concept is implemented using a stacked ensemble learning approach. Multiple base models, including Random Forest, Support Vector Machine (SVM), and XGBoost, are trained independently on the heart disease dataset. Each of these models has unique characteristics: Random Forest captures nonlinear relationships through decision trees, SVM is effective in high-dimensional spaces and finds optimal decision boundaries, and XGBoost enhances performance through gradient boosting techniques. By training these models separately, the system ensures that diverse patterns and relationships within the data are captured effectively.

The predictions generated by these base models are then combined using a meta-learner, specifically Logistic Regression. This meta-model takes the outputs of the base models as input features and learns how to optimally combine them to produce the final prediction. This layered structure forms the core of the stacked ensemble approach, where the first layer consists of base learners and the second layer consists of the meta-learner. This approach reduces the limitations of individual models and improves overall prediction accuracy. The implementation of ensemble learning in this project involves several steps, including data preprocessing, model training, prediction generation, and stacking. Initially, the dataset is preprocessed to handle missing values, encode categorical variables, and scale numerical features. The processed data is then used to train the base models independently. Once trained, these models generate predictions on the validation dataset, which are then used as input for training the meta-learner. Finally, the meta-learner produces the final prediction, which represents the combined decision of all base models.

Overall, the use of ensemble learning in this project enhances the reliability and robustness of the heart disease prediction system. By combining multiple machine learning algorithms, the system achieves higher accuracy, better generalization, and improved resistance to overfitting. This makes ensemble learning a powerful approach for developing intelligent healthcare systems capable of delivering accurate and consistent predictions in real-world scenarios.

1.5 STACKED ENSEMBLE LEARNING

Stacked ensemble learning, commonly referred to as stacking, is an advanced ensemble technique that combines multiple machine learning models in a layered architecture to improve predictive performance. Unlike traditional ensemble methods such as bagging and boosting, which primarily rely on averaging or sequential learning, stacking introduces a meta-learning strategy where the outputs of several base models are used as inputs for a higher-level model known as the meta-learner. This approach enables the system to learn how to optimally combine the strengths of individual models, thereby enhancing overall accuracy and generalization.

In a typical stacked ensemble framework, the model is organized into two levels. The first level consists of multiple base learners that are trained independently on the same dataset. These models generate predictions based on their individual learning capabilities. The second level consists of a meta-learner that takes the predictions of the base models as input features and produces the final output. This layered structure allows the model to capture complex nonlinear relationships and reduce errors associated with individual algorithms. Research studies have demonstrated that stacked ensemble models achieve superior performance in heart disease prediction by effectively integrating diverse classifiers and reducing overfitting. In this project, the stacked ensemble learning approach is implemented using three base models: Random Forest, Support Vector Machine (SVM), and XGBoost. Each of these algorithms contributes unique strengths to the prediction process. Random Forest is effective in handling nonlinear relationships and reducing variance through multiple decision trees. SVM excels in finding optimal decision boundaries, especially in high-dimensional spaces. XGBoost, a powerful boosting algorithm, enhances model performance by correcting errors iteratively and improving accuracy. By combining these models, the system ensures that different aspects of the dataset are captured effectively.

The predictions generated by these base models are then passed to a meta-learner, specifically Logistic Regression, which serves as the final decision-making model. The meta-learner analyzes the outputs of the base models and learns the optimal way to combine them to produce a more accurate prediction. This process is often supported by techniques such as cross-validation to ensure that the model does not overfit the training data and maintains good generalization on unseen data. Studies show that such two-layer stacking architectures significantly improve classification accuracy and reliability in medical prediction systems. The implementation of the stacked ensemble model in this project follows a structured workflow. Initially, the dataset undergoes preprocessing, including handling missing values, encoding categorical features, and scaling numerical attributes. The processed data is then used to train the base models independently. Once trained, these models generate predictions on validation data, which are used as inputs for training the meta-learner. Finally, during prediction, the input data is passed through the base models, and their outputs are combined by the meta-learner to produce the final result. One of the key advantages of the stacked ensemble approach is its ability to improve prediction accuracy by leveraging the complementary strengths of different algorithms. It reduces both bias and variance, making the model more robust and reliable. Additionally, it enhances the system's ability to handle complex and high-dimensional medical data, which is essential for accurate heart disease prediction. Experimental studies have reported that stacked models can achieve higher accuracy, precision, recall, and AUC scores compared to single-model approaches, making them highly suitable for healthcare applications.

2. LITERATURE SURVEY

The literature review provides an overview of existing health monitoring systems and the technologies used in modern healthcare applications. Traditionally, patient monitoring systems were limited to hospitals and relied on manual observation and basic medical devices. These systems required continuous supervision by healthcare professionals and were not suitable for real-time or remote monitoring. These systems use various sensors to measure physiological parameters such as heart rate, temperature, blood pressure, and ECG signals. Although these systems improved the accuracy of measurements, many of them lacked connectivity and real-time data sharing capabilities.

2.1 MACHINE LEARNING APPROACHES FOR HEART DISEASE PREDICTION

Machine learning (ML) has emerged as a powerful computational approach for the prediction and diagnosis of heart disease by analyzing complex medical datasets. Cardiovascular diseases are among the leading causes of mortality worldwide, and accurate prediction systems are essential for early diagnosis and effective treatment. Machine learning techniques enable the extraction of meaningful patterns from large-scale healthcare data, facilitating data- driven decision-making and improving diagnostic accuracy. Studies indicate that ML-based prediction models significantly enhance early detection, which helps reduce mortality rates and improves patient outcomes

Various supervised machine learning algorithms have been widely used for heart disease prediction. These include Logistic Regression, Decision Trees, Naïve Bayes and Support Vector Machines (SVM). Logistic Regression is commonly used for binary classification problems and provides interpretable results, making it suitable for medical applications. Decision Trees are effective in handling both categorical and numerical data and offer clear decision rules. Naïve Bayes is a probabilistic classifier that performs well with smaller datasets, while KNN is a distance-based algorithm that classifies data based on similarity measures. SVM is particularly effective in high-dimensional spaces and is capable of finding optimal decision boundaries between classes. These algorithms have been extensively applied in heart disease prediction systems and have demonstrated satisfactory performance across various datasets.

In addition to traditional machine learning algorithms, advanced models such as Random Forest and Gradient Boosting techniques have gained significant attention. Random Forest is an ensemble method that constructs multiple decision trees and combines their outputs to improve accuracy and reduce overfitting. Similarly, boosting algorithms like XGBoost iteratively improve model performance by correcting errors from previous iterations. These advanced models are particularly effective in capturing nonlinear relationships and complex interactions among features, which are common in medical datasets. Research shows that ensemble-based models often achieve higher accuracy compared to individual classifiers in heart disease prediction tasks.

Machine learning approaches for heart disease prediction typically follow a structured workflow that includes data preprocessing, feature selection, model training, and evaluation. Data preprocessing involves handling missing values, removing noise, encoding categorical variables, and scaling numerical features to ensure data consistency. Feature selection techniques are used to identify the most relevant attributes that contribute to prediction accuracy. The processed dataset is then divided into training and testing sets, and machine learning models are trained to learn patterns within the data. The performance of these models is evaluated using metrics such as accuracy, precision, recall, and F1-score.

Despite their effectiveness, individual machine learning models have certain limitations. They may suffer from overfitting, lack of generalization, and inability to capture all aspects of complex medical data. Additionally, real-world healthcare datasets often contain noise, missing values, and imbalanced classes, which can affect model performance. To address these challenges, researchers have increasingly focused on ensemble and hybrid approaches that combine multiple models to improve prediction accuracy and robustness.

Overall, machine learning approaches have significantly transformed the field of heart disease prediction by providing efficient, scalable, and accurate solutions. These techniques enable early diagnosis, support clinical decision-making, and contribute to the development of intelligent healthcare systems. The limitations of individual models have led to the adoption of advanced techniques such as ensemble learning, which forms the foundation of the proposed stacked ensemble approach used in this project.

2.2 EXISTING PREDICTION MODELS

Existing prediction models for heart disease primarily rely on traditional statistical methods and individual machine learning algorithms to analyze patient data and determine the likelihood of cardiovascular disease. These models use clinical attributes such as age, blood pressure, cholesterol levels, chest pain type, and other health indicators to perform classification tasks. Over the years, several approaches have been developed, each with its own strengths and limitations.

One of the earliest approaches used in heart disease prediction is statistical modeling, particularly Logistic Regression. This method is widely used due to its simplicity, interpretability, and effectiveness in binary classification problems. Logistic Regression models establish a relationship between input features and the probability of disease occurrence. However, these models are limited in handling complex nonlinear relationships within medical datasets.

In addition to statistical models, various machine learning algorithms have been applied in existing systems. Commonly used models include Decision

Trees, Naïve Bayes, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM). Decision Trees provide a hierarchical structure for decision-making and are easy to interpret, while Naïve Bayes uses probabilistic principles for classification. KNN classifies data based on similarity measures, and SVM identifies optimal decision boundaries between classes. These models have demonstrated moderate accuracy in heart disease prediction tasks and are widely used in early research studies.

With the advancement of machine learning, more sophisticated models such as Random Forest and Gradient Boosting techniques have been introduced. Random Forest, an ensemble- based method, constructs multiple decision trees and combines their outputs to improve prediction accuracy and reduce overfitting. Similarly, boosting algorithms like XGBoost iteratively enhance performance by correcting errors from previous models. Studies have shown that Random Forest can achieve higher accuracy compared to other individual models, reaching up to 88.5% accuracy in heart disease prediction tasks.

Another limitation of existing models is the lack of integration with real-time applications and decision-support systems. Many research models are developed and tested in controlled environments but are not deployed in practical healthcare systems. This reduces their usability in real-world scenarios where quick and reliable predictions are essential.

2.3 USE OF ENSEMBLE LEARNING IN HEALTHCARE

Ensemble learning has become an essential approach in modern healthcare systems for improving the accuracy and reliability of disease prediction models. In medical applications, data is often complex, high-dimensional, and contains nonlinear relationships, making it difficult for a single machine learning model to achieve optimal performance. Ensemble learning addresses this challenge by combining multiple models to produce a more accurate and stable prediction. The fundamental objective of ensemble methods is to reduce prediction errors caused by bias, variance, and noise, thereby enhancing overall model performance. Studies indicate that ensemble techniques consistently provide better predictive accuracy than individual classifiers in disease prediction tasks.

n healthcare, ensemble learning is widely used for the prediction and diagnosis of various diseases, including heart disease, diabetes, cancer, and neurological disorders. These models analyze large volumes of patient data, such as clinical records, laboratory results, and physiological measurements, to identify patterns and assess disease risk. Ensemble methods integrate multiple machine learning algorithms, such as decision trees, support vector machines, and neural networks, to capture different aspects of the data. This integration enables the system to provide more reliable and consistent predictions, which is crucial in clinical decision-making. Research shows that ensemble approaches are particularly effective in handling complex and imbalanced healthcare datasets, leading to improved diagnostic accuracy and better patient outcomes.

One of the key applications of ensemble learning in healthcare is early disease prediction and risk assessment. By analyzing patient data at an early stage, ensemble models can identify individuals at high risk of developing diseases and support preventive healthcare measures. This is especially important in the case of heart disease, where early detection can significantly reduce mortality rates. Ensemble models are capable of capturing both linear and nonlinear relationships within the data, making them highly suitable for predicting complex medical conditions.

2.4 LIMITATIONS OF EXISTING APPROACHES

Despite significant advancements in machine learning and healthcare analytics, existing approaches for heart disease prediction still face several limitations that affect their performance and real-world applicability. These limitations arise due to the complexity of medical data, the constraints of individual models, and challenges in deploying predictive systems in clinical environments.

One of the major limitations of existing approaches is their heavy dependence on the quality and availability of datasets. Most machine learning models are trained on publicly available datasets, which may not accurately represent diverse patient populations. Variations in demographics, lifestyle factors, and medical conditions across different regions can significantly impact model performance. As a result, models trained on limited or biased datasets may fail to generalize effectively when applied to real-world scenarios.

Another critical issue is the problem of class imbalance in medical datasets. In many cases, the number of patients without heart disease is significantly higher than those with the disease. This imbalance can lead to biased models that perform well for the majority class but fail to accurately predict the minority class, which is often the most important. Consequently, the reliability of predictions decreases, especially in critical healthcare applications where accurate detection of positive cases is essential.

Existing approaches also suffer from overfitting and lack of generalization. Many models perform well on training data but fail to maintain the same level of accuracy on unseen data. This occurs when the model learns noise or specific patterns from the training dataset rather than capturing generalizable features. Research indicates that overfitting is a common issue in both individual and ensemble models, particularly when trained on small or highly specific datasets.

2.5 SUMMARY OF LITERATURE REVIEW

The literature review highlights the significant role of machine learning techniques in improving the prediction and diagnosis of heart disease. Various traditional machine learning models, including Logistic Regression, Decision Trees, Naïve Bayes, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM), have been extensively studied and applied in heart disease prediction systems. While these models provide reasonable accuracy and interpretability, they often face limitations in handling complex and nonlinear relationships present in medical datasets.

To overcome the limitations of individual models, ensemble learning techniques such as Random Forest and XGBoost have been introduced, demonstrating improved performance by combining multiple models. These approaches enhance prediction accuracy, reduce overfitting, and provide better generalization compared to single-model methods. However, many existing systems still rely on limited ensemble strategies and lack the ability to fully integrate the strengths of diverse algorithms.

The review also emphasizes the growing importance of stacked ensemble learning as an advanced approach for improving predictive performance. Stacking combines multiple base models and utilizes a meta-learner to generate final predictions, resulting in higher accuracy and robustness. Research studies have consistently shown that stacked ensemble models outperform traditional machine learning and basic ensemble methods in heart disease prediction tasks.

Despite these advancements, existing approaches face several challenges, including dependency on high-quality datasets, class imbalance issues, overfitting, lack of interpretability, and limited real-world deployment. These limitations indicate the need for more robust and scalable solutions that can effectively handle complex medical data and provide.

3. PROPOSED METHODOLOGY

The proposed system is a Heart Disease Prediction System developed using a Stacked Ensemble Learning approach. The primary goal of the system is to improve the accuracy and reliability of heart disease prediction by combining multiple machine learning models instead of relying on a single classifier.

In traditional prediction systems, individual algorithms such as Logistic Regression or Decision Tree are used independently. However, single models may suffer from limitations such as overfitting, bias, and poor generalization. To overcome these issues, the proposed system integrates multiple base models and combines their predictions using a meta-learner.

Working of the Proposed System

The system consists of two levels:

- Level 1 – Base Models

The base models used in this project include:

- Random Forest
- Support Vector Machine (SVM)
- XGBoost

Each model is trained independently on the same preprocessed dataset. These models analyze medical attributes such as:

- Age
- Gender
- Cholesterol Level
- Blood Pressure
- Chest Pain Type
- Maximum Heart Rate
- Fasting Blood Sugar

- Exercise-Induced Angina

Each base model produces a prediction (0 – No Disease, 1 – Disease).

- Level 2 – Meta Model

The predictions from the base models are given as input to a Logistic Regression model, which acts as the meta-learner. The meta-model learns how to optimally combine base model outputs and generates the final prediction.

Features of Proposed System

- Uses stacked ensemble learning for higher accuracy.
- Reduces overfitting.
- Handles nonlinear medical data.
- Provides real-time prediction through a web interface.
- Acts as a Clinical Decision Support System (CDSS).

Advantages

- Improved prediction accuracy.
- Better generalization.
- Robust performance.
- User-friendly web application.
- Scalable for future enhancements.

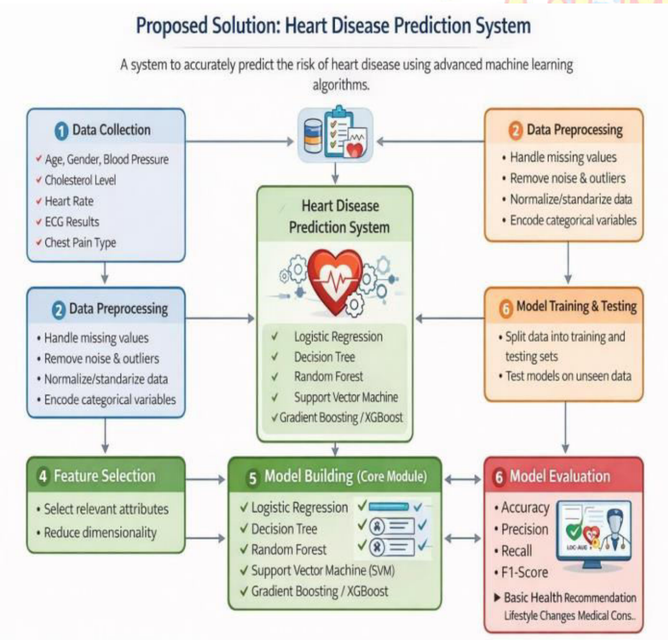

Fig 3. Block Diagram of Purposed System

In the proposed approach, multiple base models are employed to analyze the dataset and capture diverse patterns within the medical data. The base models used in this system include Random Forest, Support Vector Machine (SVM), and XGBoost. Each of these algorithms has unique characteristics that contribute to the overall performance of the system. Random Forest is effective in handling nonlinear relationships and reducing variance through multiple decision trees. SVM is capable of identifying optimal decision boundaries in high-dimensional spaces, while XGBoost improves performance through gradient boosting by correcting errors iteratively. By training these models independently, the system ensures that different aspects of the dataset are learned effectively.

The key feature of the proposed system is the implementation of stacked ensemble learning. In this approach, the predictions generated by the base models are not used directly as final outputs. Instead, they are passed as input to a higher-level model known as the meta-learner. In this project, Logistic Regression is used as the meta-learner, which learns how to optimally combine the predictions of the base models. This layered architecture improves the overall prediction performance by reducing the limitations of individual models and capturing complex relationships within the data. Research studies show that stacked ensemble models significantly improve accuracy and generalization compared to single-model approaches.

The proposed system follows a structured workflow that begins with data preprocessing. The dataset is first cleaned by handling missing values, removing inconsistencies, and encoding categorical variables. Numerical features are scaled to ensure uniformity and improve model performance. After preprocessing, the dataset is divided into training and testing sets. The base models are trained independently using the training data, and their predictions are generated. These predictions are then used as input features for training the meta-learner. During the prediction phase, user input is collected through a web-based interface. The input data is preprocessed and passed to the trained base models, which generate individual predictions.

Another important aspect of the proposed system is its integration into a user-friendly web application. The application allows users to input their medical parameters and obtain real- time predictions. This makes the system accessible and practical for both healthcare professionals and general users. Additionally, the system can be extended to support real-time monitoring, cloud-based storage, and integration with healthcare platforms.

The proposed system offers several advantages over existing approaches. It improves prediction accuracy by combining multiple models, reduces overfitting through better generalization, and provides stable and consistent results. Ensemble learning techniques are known to enhance model performance by reducing variance and bias, making them highly effective for complex datasets such as medical data.

Overall, the proposed system provides a robust and efficient solution for heart disease prediction. By utilizing stacked ensemble learning, it addresses the limitations of traditional methods and delivers improved performance, making it a valuable tool for early diagnosis and clinical decision support in healthcare environments.

3.1 SYSTEM ARCHITECTURE

The system architecture diagram represents the structural organization of the Heart Disease Prediction System. It shows the major components (layers) of the system and their responsibilities.

1. Data Input Layer

This component represents the entry point of the system. It is responsible for collecting patient health parameters required for prediction. These parameters include clinical attributes such as age, cholesterol level, blood pressure, chest pain type, and other medical indicators.

This layer ensures that the data entering the system is structured and ready for further processing.

2. Data Preprocessing Layer

This component represents the data preparation module of the system.

Its role is to transform raw medical data into a suitable format for machine learning algorithms. It performs operations such as:

- Cleaning the dataset
- Handling missing values
- Encoding categorical attributes
- Scaling numerical features

This layer ensures data quality and consistency.

3. Base Model Layer

This layer represents the first level of machine learning models in the stacking architecture.

It contains multiple independent classifiers such as:

- Random Forest
- Support Vector Machine (SVM)
- XGBoost

Each model independently analyzes the input data and produces a prediction. These models operate in parallel and contribute diverse learning patterns.

- 4. Meta Model Layer

This component represents the stacking mechanism of the system.

Logistic Regression is used as the meta-learner. It receives the predictions generated by the base models and combines them intelligently to produce the final classification output.

This layer improves overall accuracy and reduces bias and variance.

- 5. Application Layer

This layer represents the user interface of the system.

It is implemented using a web-based framework (Streamlit). It allows users to:

- Enter medical data
- Submit data for prediction
- View prediction results
- See confidence probability

It acts as the bridge between the user and the backend models.

- 6. User Layer

This component represents the end users of the system.

The users include:

- Doctors
- Healthcare professionals
- Patients

They interact with the system to obtain heart disease risk predictions.

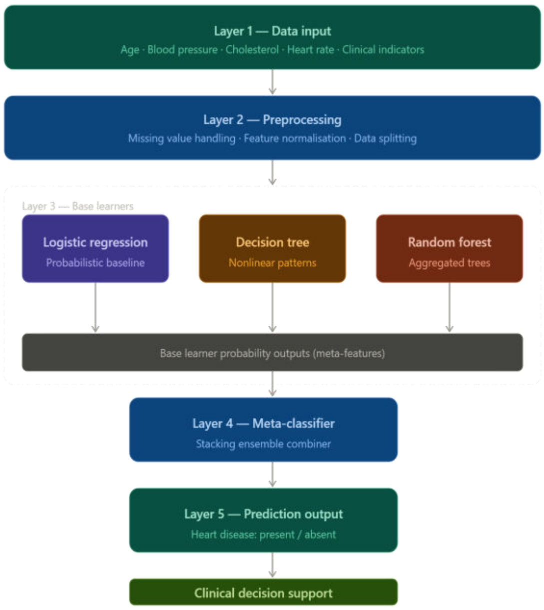

Fig 3.1 Proposed System Architecture

3.1.1 Overview

The System Architecture of the Heart Disease Prediction System using Stacked Ensemble Learning follows a layered modular approach. The architecture ensures proper data flow, separation of concerns, scalability, and improved maintainability.

The system consists of six major layers:

1. Data Input Layer
2. Data Preprocessing Layer
3. Base Model Layer
4. Meta Model Layer
5. Application Layer
6. User Layer

3.1.2 Architectural Components

1. Data Input Layer

The Data Input Layer is responsible for collecting patient health parameters. Data is collected through a web-based interface.

Input Parameters Include:

- Age
- Gender
- Chest Pain Type
- Resting Blood Pressure
- Cholesterol Level
- Fasting Blood Sugar
- Resting ECG
- Maximum Heart Rate
- Exercise-Induced Angina
- ST Depression
- Thalassemia

This layer ensures structured data collection for further processing.

2. Data Preprocessing Layer

The Data Preprocessing Layer prepares raw medical data for machine learning models.

Functions:

- Handling missing values
- Encoding categorical variables
- Feature scaling (Normalization / Standardization)
- Data splitting (Training and Testing)

Proper preprocessing improves model accuracy and reliability.

3. Base Model Layer (Level–0 Models)

This layer consists of multiple machine learning algorithms trained independently.

Base Models Used:

- Random Forest
- Support Vector Machine (SVM)
- XGBoost

Each model analyzes input features and generates independent predictions.

4. Meta Model Layer (Level–1 Model)

The Meta Model Layer implements the stacking mechanism.

- Logistic Regression is used as the meta-learner.
- It takes predictions from base models as input.
- Produces final classification output (High Risk / Low Risk).

This layer enhances overall prediction accuracy and reduces overfitting.

5. Application Layer

The Application Layer provides user interaction through a web interface developed using Streamlit.

Functions:

- Accept user input
- Send input for prediction
- Display final result
- Show probability/confidence score

6. User Layer

The User Layer consists of:

- Doctors
- Healthcare Professionals
- Patients

Users interact with the system to receive heart disease risk assessment.

3.1.3 System Architecture Workflow

1. User enters medical data.
2. Data is validated and preprocessed.
3. Preprocessed data is sent to base models.
4. Base models generate predictions.
5. Predictions are passed to the meta-model.
6. Final result is displayed to the user.

3.2 USE CASE DIAGRAM

The Use Case Diagram represents the functional behavior of the Heart Disease Prediction System and illustrates how external users interact with the system. It provides a high-level view of the system's capabilities by identifying the actors involved and the services or functions the system offers. The primary purpose of the use case diagram is to define the system's functional requirements in a clear and structured manner.

In this system, the main actor is the User, who may be a doctor, healthcare professional, or patient. The user interacts with the system to assess the risk of heart disease based on medical parameters. The diagram focuses only on external interactions and does not describe internal processing or technical implementation details.

One of the primary use cases is entering medical details. In this functionality, the user provides required health-related information such as age, blood pressure, cholesterol level, chest pain type, and other clinical attributes. This use case ensures that the system receives necessary input data for prediction.

Another important use case is validating and preprocessing input data. The system verifies whether the entered information is complete and within acceptable ranges. Although preprocessing is handled internally, from a functional perspective it ensures that the data is properly prepared before prediction is performed.

The core use case of the system is predicting heart disease. In this functionality, the system analyzes the processed input data using the stacked ensemble learning model and determines whether the individual is at high or low risk of heart disease. This represents the central objective of the entire system.

The final use case involves displaying the prediction result and providing recommendations. The system presents the outcome in a clear and understandable format and may offer precautionary suggestions if a high risk is detected. These use cases collectively define how the user interacts with the system and outline the functional scope of the application

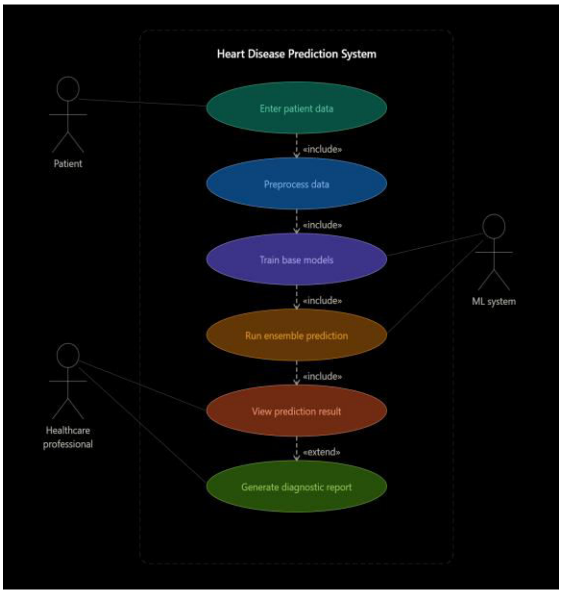

Fig 3.2 Use Case Diagram

3.2.1 Introduction

The Use Case Diagram represents the functional behavior of the system and illustrates how users interact with the system.

3.2.2 Actors

Primary Actor:

- User (Doctor / Patient)

3.2.3 Use Cases

1. Enter Medical Details
2. Validate Input Data
3. Preprocess Data
4. Predict Heart Disease
5. View Prediction Result
6. Receive Recommendation

3.2.4 Use Case Description

- Use Case 1: Enter Medical Details

The user inputs health parameters through the web interface.

- Use Case 2: Validate Input

The system verifies whether the data entered is complete and within valid range.

- Use Case 3: Preprocess Data

The system performs encoding, scaling, and cleaning of input data.

- Use Case 4: Predict Heart Disease

The stacked ensemble model analyzes the processed data and predicts the risk.

- Use Case 5: View Prediction Result

The system displays:

- High Risk of Heart Disease

or

- Low Risk of Heart Disease
- Use Case 6: Receive Recommendation

If high risk is detected, the system provides precautionary suggestions.

3.2.5 Use Case Flow (Sequential)

1. User → Enters Data
2. System → Validates Data
3. System → Preprocesses Data
4. System → Runs Stacked Model
5. System → Displays Result
6. System → Provides Recommendations

3.3 CLASS DIAGRAM

The Class Diagram represents the static structural view of the Heart Disease Prediction System. It illustrates the system's classes, their attributes, methods, and the relationships among them. Unlike behavioral diagrams, the class diagram focuses on how the system is organized internally and how different components are logically structured.

In this system, the User class represents the external entity that interacts with the application. It contains attributes related to patient medical details such as age, cholesterol level, blood pressure, and other health parameters. The methods associated with this class allow users to enter and submit their data to the system for analysis.

The Data Preprocessing class is responsible for preparing the input data before it is used by machine learning models. It includes methods for cleaning data, encoding categorical variables, normalizing numerical values, and organizing the dataset. This class ensures that the data is structured and suitable for accurate model training and prediction.

The Base Model class represents the first level of machine learning models used in the stacking architecture. It includes classifiers such as Logistic Regression, Decision Tree, and Random Forest. The methods within this class handle model training and generation of intermediate predictions, which are later passed to the meta-classifier.

The Meta Model class acts as the higher-level decision-making component in the stacked ensemble structure. It receives prediction outputs from the base models and combines them to generate a final classification result. This class enhances the overall prediction accuracy by learning how to optimally integrate base model outputs.

Finally, the Result or Output class manages the presentation of prediction outcomes to the user. It contains methods for displaying the final risk assessment and probability score. The relationships between these classes reflect a structured and modular design, ensuring clear separation of responsibilities and efficient system organization.

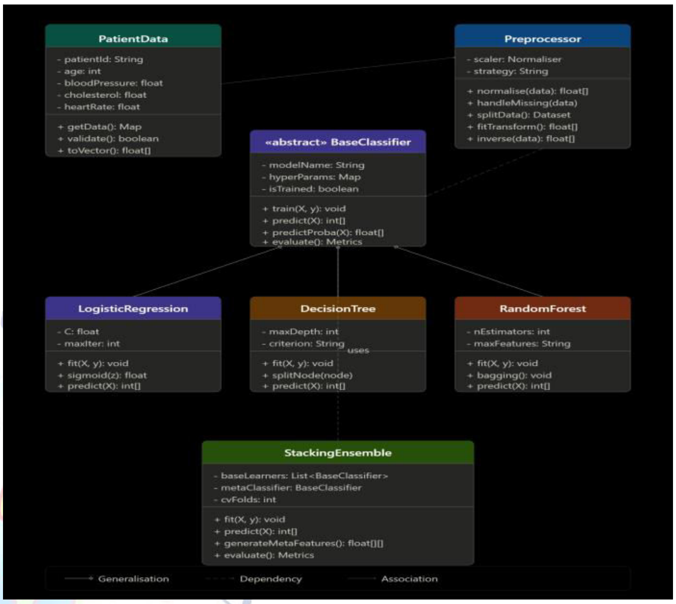

Fig 3.3 Class Diagram

3.3.1 Introduction

The Class Diagram represents the static structure of the system. It defines system classes, attributes, methods, and relationships.

3.3.2 Classes Description

1. User Class

Represents the system user.

Attributes:

- age
- gender
- cholesterol
- bloodPressure
- heartRate
- chestPainType

Methods:

- enterData()
- submitData()

2. DataPreprocessing Class

Handles data cleaning and transformation.

Methods:

- cleanData()

- encodeData()
- normalizeData()
- splitData()

3. BaseModel Class

Implements Level–0 models.

Attributes:

- randomForestModel
- svmModel
- xgboostModel

Methods:

- trainModels()
- predictBaseModels()

4. MetaModel Class

Implements Level–1 model.

Attribute:

- logisticRegressionModel

Methods:

- trainMetaModel()
- predictFinalOutput()

5. Recommendation Processor Class

Provides health suggestions based on risk level.

Methods:

- analyzeRisk()
- suggestPrecautions()
- suggestMedicalConsultation()

6. Result Class

Displays prediction result.

Methods:

- displayResult()
- showProbability()

3.3.3 Relationship Between Classes

- User → provides input to → DataPreprocessing
- DataPreprocessing → sends data to → BaseModel
- BaseModel → sends predictions to → MetaModel
- MetaModel → sends output to → RecommendationProcessor
- RecommendationProcessor → sends result to → Result

3.3.4 Design Benefits

- Clear separation of responsibilities
- Modular architecture
- Easy maintenance and extension
- Improved scalability
- Supports integration of additional models

RESULTS

Fig 4.1: Low Risk

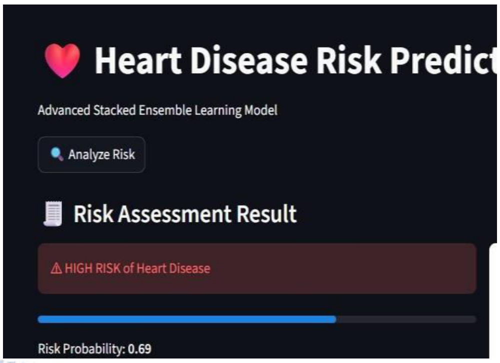

Fig 4.2: High Risk

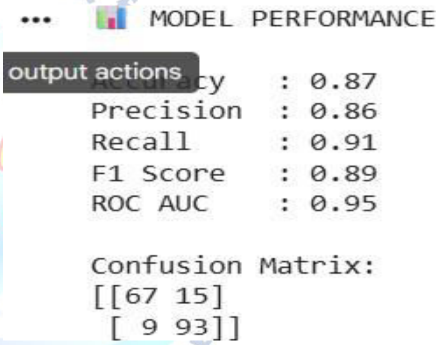

Fig 4.3: Metrics Performance

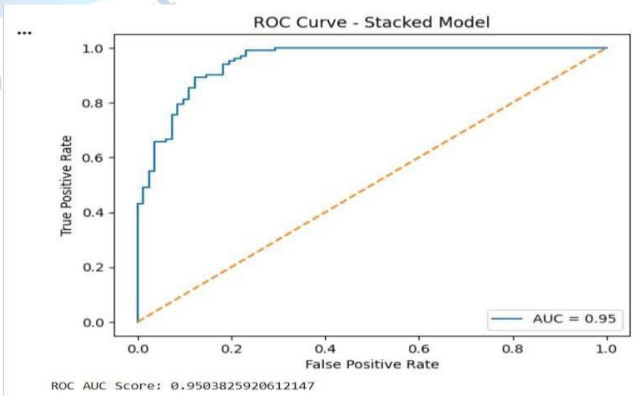

Fig 4.4: ROC Curve Analysis

5. CONCLUSION

The Heart Disease Prediction Analysis system represents a significant advancement in the application of machine learning in the healthcare domain. By leveraging data-driven techniques, the system effectively predicts the likelihood of heart disease based on various medical parameters, enabling early detection and preventive care. The model demonstrates strong performance, achieving an accuracy of approximately –90%, as validated through evaluation metrics such as the classification report and confusion matrix.

The system processes user-specific health data, including age, blood pressure, cholesterol levels, chest pain type, and other clinical attributes, to generate reliable predictions. By analyzing these parameters, the model provides a clear assessment of whether an individual is at risk of heart disease, assisting in timely medical decision-making.

Furthermore, the implementation of machine learning algorithms enhances the system's ability to identify complex patterns within the data. The confusion matrix analysis indicates that the model can effectively distinguish between patients with and without heart disease, thereby minimizing prediction errors and improving reliability.

The results highlight the potential of AI-based systems in supporting healthcare professionals by providing quick and accurate risk assessments. This reduces manual effort, improves efficiency, and helps in prioritizing patients who may require immediate medical attention Although the system serves as a supportive tool rather than a replacement for professional diagnosis, it establishes a strong foundation for future advancements in intelligent healthcare systems. With further improvements, integration of real-time data, and clinical validation, such systems can play a crucial role in preventive healthcare and early diagnosis.

FUTURE ENHANCEMENT

To further improve the Heart Disease Prediction Analysis system, several important enhancements can be implemented. First, integrating real-time health monitoring data from wearable devices such as smartwatches and fitness trackers will enable continuous tracking of vital parameters like heart rate and blood pressure, making the predictions more dynamic and accurate. Expanding the system to include a wider range of cardiovascular conditions and related diseases will improve its capability to provide comprehensive health risk analysis. This will ensure that the system is not limited to basic heart disease prediction but can support broader diagnostic insights. Incorporating Explainable AI (XAI) techniques is another key improvement. This will allow users and healthcare professionals to understand how the model arrives at its predictions, increasing transparency, trust, and usability of the system. Additionally, integrating personalized health recommendations—such as diet plans, exercise routines, and lifestyle modifications—based on prediction results can make the system more practical and beneficial for users in managing their health. Another enhancement involves connecting the system with electronic health records (EHRs) to utilize historical patient data for more accurate and personalized predictions. This will significantly improve the reliability of the system.

Deploying the application in real-world environments and collecting feedback from users and medical professionals will help in continuous improvement and performance optimization. Finally, adding multi-language support will make the system accessible to a wider population, ensuring usability across different regions and language groups.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Breiman, L. (1996). Stacked regressions. Machine Learning, 24(1), 49–64. https://doi.org/10.1007/BF00117832
[2] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1
[3] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., et al. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64(5), 304–310.
[4] UCI Machine Learning Repository. (2019). Heart Disease Dataset. https://archive.ics.uci.edu/ml/datasets/heart+disease
[5] Kaur, H., Kumar, V., & Kumar, V. (2021). Predicting heart disease using machine learning algorithms. Procedia Computer Science, 167, 1654–1663. https://doi.org/10.1016/j.procs.2020.03.380
[6] 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
[7] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939785
[8] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.
[9] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43–48.

[10] Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. IEEE/ACS International Conference on Computer Systems and Applications. https://doi.org/10.1109/AICCSA.2008.4493524

[11] Shah, D., Patel, S., Bharti, S. K., & Shah, A. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1(6), 345. https://doi.org/10.1007/s42979-020-00365-y

[12] Polat, K., & Güneş, S. (2007). An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of heart disease. Digital Signal Processing, 17(4), 601–611.

[13] Khan, M. A., et al. (2020). An ensemble learning approach for heart disease prediction using machine learning. IEEE Access, 8, 144950–144963. https://doi.org/10.1109/ACCESS.2020.3013490