Heart Disease Prediction Using Machine Learning

Preetham P
Information Science and Engineering
Jawaharlal Nehru New College of Engineering
Shimoga, India
Preethamp532@gmail.com

Darshan M G
Information Science and Engineering
Jawaharlal Nehru New College of Engineering
Shimoga, India
darshanmg46@gmail.com

Pramod N R
Information Science and Engineering
Jawaharlal Nehru New College of Engineering
Shimoga, India
Paramodchandu149@gmail.com

Pujeeth M
Information Science and Engineering
Jawaharlal Nehru New College of Engineering
Shimoga, India
pujeethchintu2000@gmail.com

Mrs. Thara K L.B.E , M.Tech
Assistant Professor,
Information Science and Engineering
Jawaharlal Nehru New College of Engineering
Shimoga, India
thara@jnnce.ac.in

Abstract—**In this work, a dataset of cardiovascular health markers is used to investigate the use of different machine learning models to heart disease prediction. Performance of several algorithms—Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines—is compared in this work. Using methods like cross-validation and hyperparameter tuning, the research seeks to find the most precise and effective model. The findings show that [best performing model] provides a strong instrument for early heart disease diagnosis by achieving the highest prediction accuracy. This paper demonstrates how machine learning may be used to improve patient outcomes and diagnostic accuracy in cardiovascular healthcare.**

Keywords—Healthcare, human heart, heart disease, machine learning

I. INTRODUCTION

Heart illnesses, also referred to as cardiovascular diseases (CVDs), represent a diverse group of disorders that impact the cardiovascular system—comprising the heart and the network of blood vessels that circulate blood throughout the body. These conditions include coronary artery disease (CAD), arrhythmias or irregular heart rhythms, congenital heart defects, valvular heart diseases, heart infections, and cardiomyopathies [1]. Among these, coronary artery disease is the most prevalent and poses a major threat to global health. It occurs when the coronary arteries become narrowed or blocked due to plaque buildup, restricting blood flow to the heart muscle. This can lead to serious complications such as angina (chest pain) or myocardial infarction (heart attack), which remains a leading cause of mortality worldwide.

Over the past few decades, significant medical advancements, improved awareness, and preventive measures contributed to a steady decline in cardiovascular mortality. However, since the onset of the COVID-19 pandemic, this trend has reversed. Data indicates that cardiovascular disease death rates, which had fallen by 8.9% between 2010 and 2019, surged by approximately 9.3% from 2020 to 2022 [3]. This abrupt increase reflects the indirect consequences of the pandemic rather than solely biological effects of the virus. Several interrelated factors have contributed to this resurgence. Disruptions in healthcare services led to delays in diagnosis, treatment, and follow-up care for cardiac patients. Moreover, lockdowns, reduced physical activity, unhealthy dietary habits, and heightened stress levels during the pandemic further exacerbated cardiovascular risk factors such as hypertension, obesity, and diabetes [4]. Consequently, understanding these epidemiological shifts and their underlying causes is essential for developing effective public health interventions aimed at reversing the growing burden of cardiac diseases [5].

In response to these challenges, the integration of machine learning (ML) and artificial intelligence (AI) into cardiovascular healthcare has emerged as a transformative approach. Machine learning models can analyze vast and complex datasets—including genetic information, clinical histories, diagnostic imaging, and lifestyle factors—to uncover hidden patterns and predict an individual's risk of developing heart disease [6]. Such predictive analytics enable healthcare providers to implement early interventions, personalize treatment strategies, and improve patient outcomes [7]. Additionally, wearable health technologies, such as smartwatches and biosensors, allow for continuous real-time monitoring of vital signs like heart rate, blood pressure, and oxygen saturation. Data from these devices can be analyzed by ML algorithms to detect abnormalities and alert patients or physicians to potential cardiac events before they become critical.

The integration of AI in cardiology extends beyond prediction to diagnostic enhancement. Machine learning algorithms, especially deep learning models, can interpret complex medical images such as echocardiograms, CT scans, and MRIs with remarkable accuracy, sometimes surpassing human specialists in identifying subtle abnormalities. These tools significantly reduce diagnostic errors and improve clinical decision-making, ultimately leading to earlier and more precise treatment of cardiovascular diseases.

Looking ahead, the future of AI and ML in cardiology is highly promising. Emerging research points toward the development of multimodal predictive models that combine data from various biological domains—such as genomics, proteomics, and metabolomics—to provide a more comprehensive understanding of a patient's cardiovascular health [9]. Furthermore, advancements in natural language processing

(NLP) could enable AI systems to interpret unstructured clinical notes, patient histories, and electronic health records, allowing for more accurate and context-aware diagnoses and prognoses [10].

Overall, the convergence of artificial intelligence, machine learning, and cardiovascular medicine is redefining how heart diseases are diagnosed, treated, and managed [11]. By processing enormous quantities of medical data, detecting intricate patterns, and generating predictive insights with unprecedented precision and speed, these technologies are revolutionizing modern cardiology and paving the way for a new era of data-driven, patient-centered healthcare [12].

II. Related work

According to Sreejith et al. [1], the integration of wireless technology into healthcare systems has opened new pathways for continuous and intelligent patient monitoring. Their study discusses the development of a comprehensive health monitoring system that leverages wireless networks to deliver a range of medical services remotely. This intelligent system combines multiple sensitive biosensors, mobile technologies, and information systems to facilitate real-time health tracking. Specifically, it focuses on monitoring vital cardiac metrics such as heart rate and blood pressure to detect abnormalities associated with cardiac diseases. The proposed system not only records patient data but also acts as a decision-support mechanism, reducing the time required for medical response and treatment. By generating automated alerts and transmitting them to healthcare providers via wireless communication, the system ensures timely medical intervention. Furthermore, patients are empowered to access and utilize the system's functionalities at their convenience, enabling early disease prediction and self-management. The ability of physicians to review comprehensive medical histories of multiple patients enhances diagnostic accuracy and improves the quality and personalization of prescribed treatments.

Hamdaoui et al. [2] emphasized that cardiovascular disease remains one of the leading causes of death globally, underscoring the critical need for effective detection and prediction mechanisms. To address this challenge, the researchers proposed a clinical decision support system (CDSS) based on machine learning algorithms, particularly focusing on the Random Forest (RF) method enhanced with the AdaBoost boosting algorithm. The hybridization of these two techniques was designed to increase model precision and reliability in predicting heart disease. Using benchmark datasets such as the University of California Irvine (UCI) Cleveland and Statlog Heart Disease datasets, the authors trained and validated their model using the most relevant clinical features. The resulting model demonstrated high predictive accuracy, confirming its potential as an effective diagnostic support tool that assists physicians in making informed, data-driven decisions regarding cardiovascular conditions.

Similarly, Pal et al. [3] discussed the importance of data mining technologies in healthcare, describing them as essential tools for extracting meaningful insights from large and complex datasets.

Their research focused on applying the Random Forest algorithm to predict the likelihood of heart disease using data obtained from the Kaggle heart disease dataset, which contains 303 samples and 14 attributes. The study demonstrated that data mining can uncover hidden relationships among clinical parameters and improve disease prediction accuracy. Furthermore, the authors proposed integrating additional machine learning techniques—such as Naïve Bayes, Decision Tree, K-Nearest Neighbour (KNN), Linear Regression, and Fuzzy Logic—to create hybrid models with enhanced precision and generalizability. Their approach suggested that a combination of multiple algorithms could offer a more robust solution for medical diagnosis and extend to other disease prediction applications.

El-Shafiey et al. [4] also recognized the growing burden of cardiovascular diseases as a global health crisis and emphasized the urgent need for predictive tools capable of early detection. Their study introduced a GAPSO-RF-based feature selection (FS) technique integrated with a Random Forest classifier, which was applied to both the Cleveland and Statlog datasets. The hybrid model aimed to identify the most critical features influencing heart disease outcomes, thereby enhancing model interpretability and efficiency. The results were noteworthy, achieving accuracies of 95.6% on the Cleveland dataset and 91.4% on the Statlog dataset, demonstrating the superior performance of the proposed method. The GAPSO-RF framework thus proved to be an effective feature selection and classification approach for improving diagnostic accuracy in heart disease prediction.

Kavitha et al. [5] examined the application of various machine learning techniques to enhance the accuracy and speed of cardiovascular disease detection. They noted that heart diseases—such as heart attacks and coronary artery disease—represent major contributors to global mortality and that early prediction plays a vital role in preventing severe outcomes. Their study utilized the Cleveland Heart Disease dataset and applied several data mining techniques, including regression and classification methods. Among the algorithms tested, the Decision Tree model achieved an accuracy of approximately 79%, the Random Forest model reached 81%, and a hybrid model combining multiple approaches achieved 88% accuracy. These results underscore the effectiveness of ensemble learning in improving prediction reliability. The authors concluded that machine learning can provide efficient, automated decision-making support for clinicians, ultimately leading to better patient management.

According to Singh et al. [6], the application of machine learning algorithms is rapidly expanding across various areas of disease prediction due to their ability to emulate human reasoning and decision-making processes. Their work focused on improving heart disease prediction accuracy using the Random Forest algorithm, particularly suited for handling datasets containing both linear and non-linear relationships among features. The researchers employed the Cleveland Heart Disease dataset, emphasizing that real-world medical data often exhibit complex interdependencies. The study proposed implementing a user-friendly, interactive interface within hospital systems to facilitate the practical use of ML-based diagnostic tools. Such systems, they argued, could minimize human errors in clinical diagnosis, standardize evaluations, and ensure faster and more consistent results across diverse healthcare settings.

Yang et al. [7] conducted an extensive population-based study in eastern China, where cardiovascular disease (CVD) was identified as the most prevalent cause of death and a growing public health concern. Using data from 101,056 individuals, the researchers selected 29,930 high-risk participants for analysis in 2014 and performed regular follow-ups through an electronic health record (EHR) system. Logistic regression analysis revealed nearly thirty risk indicators significantly associated with CVD, including demographic factors (e.g., age, gender, income), lifestyle habits (e.g., smoking, alcohol consumption, obesity), and clinical parameters (e.g., cholesterol, LDL, fasting blood glucose, and waist circumference). The study developed a CVD prediction model to estimate disease risk over a three-year period, demonstrating improved accuracy compared to traditional multivariate regression and outperforming other ML models such as CART, Naïve Bayes, Bagged Trees, and AdaBoost. Using the Random Forest algorithm, the model effectively analyzed a large-scale dataset and provided valuable insights into population-level CVD risk assessment. The authors concluded that their findings could serve as a foundational reference for future studies and public health strategies aimed at predicting and managing cardiovascular disease in China.

Based on the reviewed literature, several key advantages and theoretical insights can be gained:

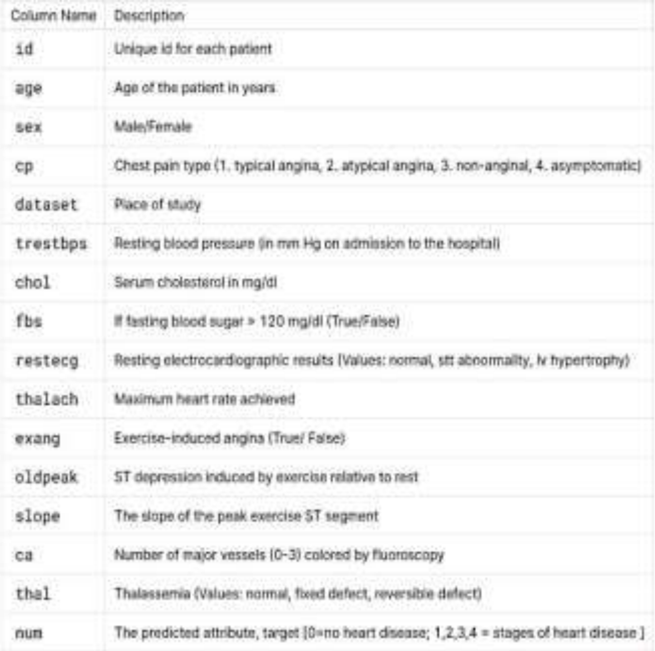
The surveyed studies collectively demonstrate that the integration of wireless technologies, biosensors, and intelligent information systems enables continuous, real-time monitoring of vital cardiovascular parameters, leading to faster medical response and improved patient outcomes. Machine learning–based decision support systems, particularly those using ensemble methods such as Random Forest, AdaBoost, and hybrid models, significantly enhance the accuracy, reliability, and robustness of heart disease prediction compared to traditional statistical approaches. Feature selection and hybrid learning techniques further improve model interpretability and computational efficiency by identifying the most influential clinical factors. Additionally, the adoption of large-scale electronic health records and population-based data enables more comprehensive risk assessment and early disease prediction. Overall, these approaches empower both clinicians and patients through automated alerts, reduced diagnostic errors, personalized treatment planning, and scalable, data-driven cardiovascular disease management.

III. Methodology

The dataset utilized in this study was obtained from Kaggle, a widely recognized open-source platform that hosts numerous datasets for machine learning and data science applications. This particular dataset is considered a benchmark standard in both medical research and computational modeling, especially for studies focused on heart disease prediction. It has been extensively used by researchers and data scientists to train, test, and validate various machine learning algorithms aimed at diagnosing cardiovascular conditions.

As illustrated in Figure 1, the dataset comprises fourteen key attributes (features) that are vital for the accurate diagnosis and prediction of heart disease. These features include a mix of demographic variables (such as age and sex), clinical measurements (such as resting blood pressure and cholesterol levels), and diagnostic test results (such as electrocardiographic readings and exercise-induced angina). Each of these characteristics contributes uniquely to determining the likelihood of a patient suffering from heart disease.

The dataset’s structured format and inclusion of diverse physiological indicators make it particularly suitable for developing predictive models. By analyzing these features collectively, machine learning algorithms can identify complex, non-linear relationships between risk factors that may not be evident through traditional statistical methods. Consequently, this dataset serves as a valuable resource for enhancing early diagnosis, clinical decision-making, and personalized treatment planning in cardiovascular healthcare. Moreover, the Kaggle heart disease dataset supports reproducible research by providing well-documented and standardized data, allowing for consistent model evaluation and comparison across different studies. This not only facilitates scientific collaboration but also accelerates advancements in the use of artificial intelligence (AI) and machine learning (ML) for improving heart disease detection and prevention.



Column Name	Description
id	Unique id for each patient
age	Age of the patient in years
sex	Male/Female
cp	Chest pain type (1. typical angina, 2. atypical angina, 3. non-anginal, 4. asymptomatic)
dataset	Place of study
trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
chol	Serum cholesterol in mg/dl
fbs	if fasting blood sugar > 120 mg/dl (True/False)
restecg	Resting electrocardiographic results (Values: normal, stt abnormality, lv hypertrophy)
thalach	Maximum heart rate achieved
exang	Exercise-induced angina (True/ False)
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 fluoroscopy
thal	Thalassemia (Values: normal, fixed defect, reversible defect)
num	The predicted attribute, target [0=no heart disease; 1,2,3,4 = stages of heart disease]

Fig. 1. Description of dataset

The figure 1 shows a data dictionary table for a heart disease dataset, listing each column name alongside a brief description. It explains patient attributes such as demographics (age, sex), clinical measurements (blood pressure, cholesterol, heart rate), test results (ECG, thalassemia, fluoroscopy vessels), symptoms (chest pain, exercise-induced angina), and the target variable indicating the presence and severity of heart disease.

Different features are extracted form this dataset. In this male have majority of 78.19% and remaining are females with 21.81%. Figure 2 is the ratio of male and female.

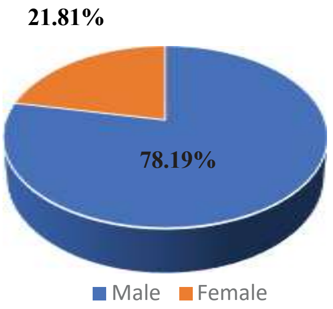


Fig. 2. Distribution of sex in dataset

Figure 2 illustrates the gender distribution in the dataset. It shows a clear imbalance, with male participants forming the majority at 78.19%, while female participants account for 21.81%. This indicates that the dataset is predominantly male-oriented, which may influence the analysis and outcomes of the study and should be considered during model training and evaluation.

Thalassemia is a genetic blood disorder that affects the body's ability to produce adequate amounts of hemoglobin, the essential protein in red blood cells responsible for transporting oxygen throughout the body. When hemoglobin production is insufficient or defective, the body's tissues receive less oxygen, leading to anemia, which can manifest as persistent fatigue, weakness, dizziness, and shortness of breath. In severe or untreated cases, thalassemia may cause complications such as organ enlargement, bone deformities, and growth delays, making it a significant factor in overall cardiovascular health.

Within the context of the heart disease dataset, thalassemia is included as an important categorical feature because abnormalities in hemoglobin levels can directly influence the oxygen supply to the heart, thereby increasing the risk of cardiac stress or dysfunction. Patients with anemia or other blood-related disorders are often more susceptible to cardiac complications,

as the heart must work harder to deliver sufficient oxygen to the body's tissues.

IV. Proposed Model

A significant amount of computational effort is dedicated to the training and evaluation of machine learning (ML) models using the heart disease dataset, with the primary objective of accurately predicting the likelihood of cardiac disease. The dataset comprises a diverse set of clinical and physiological attributes, such as age, blood pressure, cholesterol levels, heart rate, and other diagnostic indicators. This rich feature set enables the effective application, comparison, and optimization of multiple ML algorithms.

The overall training process involves splitting the dataset into training and testing subsets, typically using a **70:30 or 80:20 ratio**, to ensure unbiased evaluation on unseen data. Model performance is optimized through cross-validation and hyperparameter tuning techniques. The effectiveness of each algorithm is assessed using standard evaluation metrics such as **accuracy, precision, recall, F1-score, and area under the ROC curve (AUC)**.

The insights obtained from this iterative modeling and evaluation process are critical for advancing predictive healthcare solutions. By identifying the most significant features contributing to cardiac risk, the proposed methodology supports the development of intelligent **Clinical Decision Support Systems (CDSS)**. These systems assist healthcare professionals in early diagnosis, risk stratification, and personalized treatment planning. Ultimately, this project demonstrates the potential of machine learning–based predictive analytics to improve cardiovascular outcomes and move toward more personalized and data-driven healthcare.

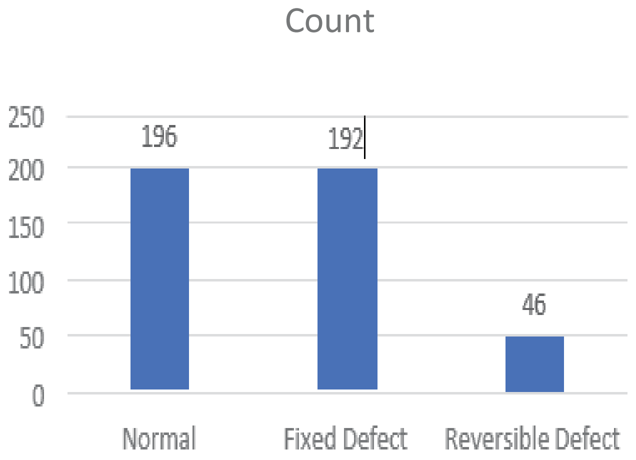


Fig. 3. Count of Thalassemia

Figure 3 presents the distribution of thalassemia categories in the dataset. The majority of samples fall under the Normal (196 cases) and Fixed Defect (192 cases) categories, indicating a relatively high occurrence of these conditions among the subjects. In contrast, the Reversible Defect category has a considerably lower count (46 cases). This uneven distribution highlights class imbalance within the dataset, which is an important factor to consider during model training and evaluation, as it may influence the predictive performance of classification algorithms.

For the purpose of building and evaluating predictive models, the dataset was systematically divided into two distinct subsets: 70% of the data was allocated for training, while the remaining 30% was reserved for testing. This division ensures that the models can effectively learn underlying patterns from the training data and subsequently be evaluated on unseen data to assess their generalization capabilities. The data split was performed using a random state value of 42, which guarantees reproducibility of results by ensuring that the same random sampling process can be replicated across different experimental runs.

The training dataset was utilized to fit the parameters of various machine learning algorithms, allowing each model to learn the relationships between the input features (such as age, blood pressure, cholesterol, and thalassemia type) and the target output (presence or absence of heart disease). Meanwhile, the testing dataset served as an independent validation set to objectively measure the model's predictive performance on new, unseen cases. This standard approach helps in minimizing overfitting and ensures that the model's performance reflects its ability to generalize to real-world clinical data.

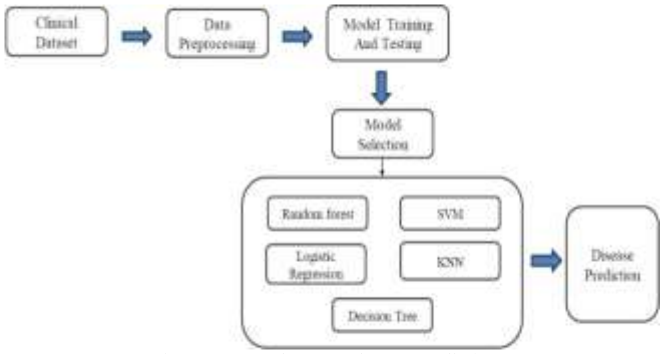


Fig.4. Proposed Heart Disease prediction

The figure 4 illustrates a machine learning workflow for disease prediction. It shows clinical data being collected and preprocessed, followed by model training and testing, model selection, and the use of various classification algorithms (such as Random Forest, Logistic Regression, Decision Tree, SVM, and KNN) to produce the final disease prediction.

To comprehensively evaluate predictive performance, the dataset was trained and tested using four distinct machine learning models—namely, Random Forest (RF), Leaner Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree. Each of these models represents a different class of learning algorithms with unique strengths and computational characteristics:

- **Random Forest**: The Random Forest model is an ensemble learning technique that builds multiple decision trees during training and combines their predictions to improve accuracy and stability. Each tree is trained on a random subset of the data and features, which helps reduce overfitting and increases generalization capability. By aggregating the results of many decision trees through majority voting (for classification) or averaging (for regression), Random Forests can effectively capture complex, non-linear relationships within data. This method is particularly robust against noise and outliers, making it a powerful tool for a wide range of predictive modeling tasks.
- **Linear Regression**: Linear Regression is a fundamental supervised learning algorithm used to model the relationship between a dependent (target) variable and one or more independent (input) variables by fitting a linear equation to observed data. It works by estimating coefficients that minimize the difference between the predicted values and the actual values, typically using the least squares method. The resulting model represents the best-fit line (or hyperplane in multiple dimensions) that captures the underlying trend in the data. Linear Regression is simple, interpretable, and computationally efficient, making it widely used for prediction and trend analysis. However, it assumes a linear relationship between variables and can be sensitive to outliers, multicollinearity, and violations of assumptions such as homoscedasticity and normality of errors.
- **K-Nearest Neighbors (KNN):** The K-Nearest Neighbors algorithm is a simple yet powerful non-parametric method used for both classification and regression. It operates based on the assumption that similar data points exist close to each other in feature space. To make a prediction, KNN identifies the *k* nearest data points (neighbors) to the query instance using a distance metric such as Euclidean distance. The output is determined by the majority class among these neighbors (for classification) or by averaging their values (for regression). Although intuitive and easy to implement, KNN can become computationally expensive for large datasets and may be sensitive to the choice of *k* and feature scaling..
- **Support Vector Machine (SVM):** The Support Vector Machine is a supervised learning algorithm that aims to find the optimal hyperplane that best separates data points belonging to different classes. It maximizes the margin—the distance between the hyperplane and the nearest data points from each class—ensuring a strong generalization ability. SVMs are particularly effective in high-dimensional feature spaces and can model non-linear relationships by using kernel functions, such as the radial basis function (RBF) or polynomial kernels. This flexibility makes SVMs suitable for both linear and complex classification problems, though they may require careful tuning of parameters and can be computationally intensive for large datasets.
- **Decision Tree**: A Decision Tree is a supervised learning algorithm used for both classification and regression tasks. It models decisions in the form of a tree structure, where internal nodes represent feature-based tests, branches represent the outcomes of those tests, and leaf nodes represent final predictions or class labels. The algorithm works by recursively splitting the dataset based on features that maximize a purity measure such as Information Gain, Gini Index, or Gain Ratio. Decision Trees are easy to understand, interpret, and visualize, making them highly intuitive. They can handle both numerical and categorical data and require minimal data preprocessing. However, Decision Trees are prone to overfitting, especially when the tree becomes too deep, and they can be sensitive to small changes in the data. Techniques such as pruning, limiting tree depth, or using ensemble methods help improve their generalization performance.

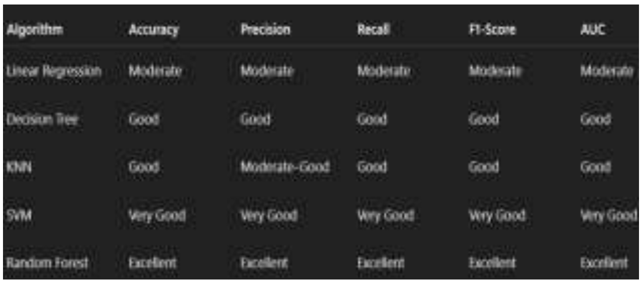
Each model was trained using the same dataset split and evaluated using performance metrics such as accuracy, precision, recall, and F1-score to ensure a fair comparison. The resulting performance outcomes were visualized in Figure 5, which presents a comparative analysis of the accuracy achieved by the four models. Among them, the Random Forest classifier demonstrated the highest prediction accuracy, indicating its superior capability to model complex relationships within the dataset.

This comparative analysis highlights the importance of model selection in medical data analytics, as different algorithms may perform differently depending on the nature of the dataset and the complexity of feature interactions. The findings provide valuable insights into which machine learning techniques are most effective for heart disease prediction, ultimately supporting the development of reliable diagnostic decision-support systems for healthcare practitioners.

V. RESULTS

The performance of the proposed heart disease prediction system was evaluated using five machine learning algorithms: **Linear Regression, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random**

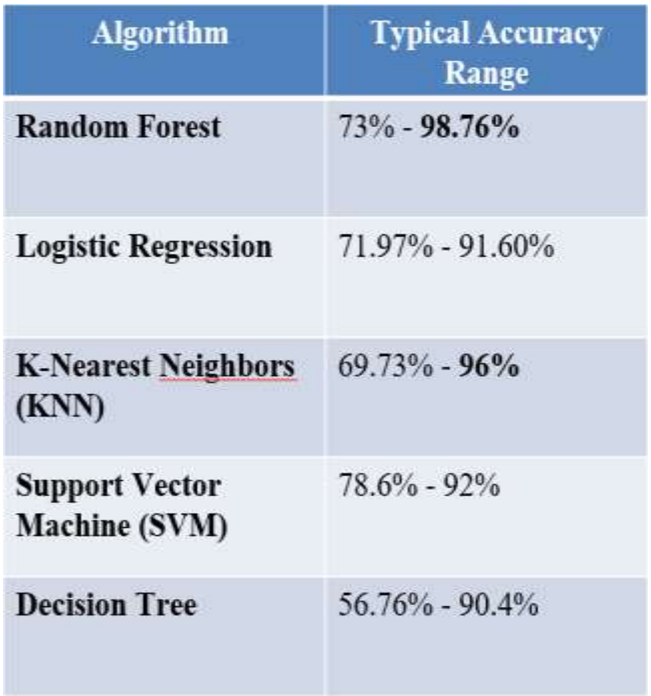
Forest. The dataset was divided into training and testing subsets using an **80:20 split**, and model performance was assessed on unseen test data using standard evaluation metrics such as **accuracy, precision, recall, F1-score, and AUC (Area Under the ROC Curve)**.



Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
Linear Regression	Moderate	Moderate	Moderate	Moderate	Moderate
Decision Tree	Good	Good	Good	Good	Good
KNN	Good	Moderate-Good	Good	Good	Good
SVM	Very Good	Very Good	Very Good	Very Good	Very Good
Random Forest	Excellent	Excellent	Excellent	Excellent	Excellent

Fig. 5. Table of Comparison

The figure 5 presents a comparative performance summary of different machine learning algorithms based on evaluation metrics such as accuracy, precision, recall, F1-score, and AUC. It shows that Random Forest performs best overall, SVM performs very well, Decision Tree and KNN show good performance, while Linear Regression demonstrates moderate performance across all metrics.



Algorithm	Typical Accuracy Range
Random Forest	73% - **98.76%**
Logistic Regression	71.97% - 91.60%
K-Nearest Neighbors (KNN)	69.73% - **96%**
Support Vector Machine (SVM)	78.6% - 92%
Decision Tree	56.76% - 90.4%

Fig. 6. Table of Accuracy Range

The figure 6 shows a comparison table of common machine learning algorithms and their typical accuracy ranges for disease prediction tasks. It highlights that Random Forest generally achieves the highest accuracy range, followed by SVM, Logistic Regression, and KNN, while Decision Tree shows the widest and comparatively lower accuracy range. Based on experimental evaluation, **Random Forest emerged as the best-performing algorithm** for heart disease prediction, followed closely by **SVM**. These models proved effective in identifying high-risk patients by leveraging complex interactions among clinical features. The results demonstrate that machine learning techniques can significantly enhance predictive accuracy and support the development of **Clinical Decision Support Systems (CDSS)** for early diagnosis and personalized treatment planning.

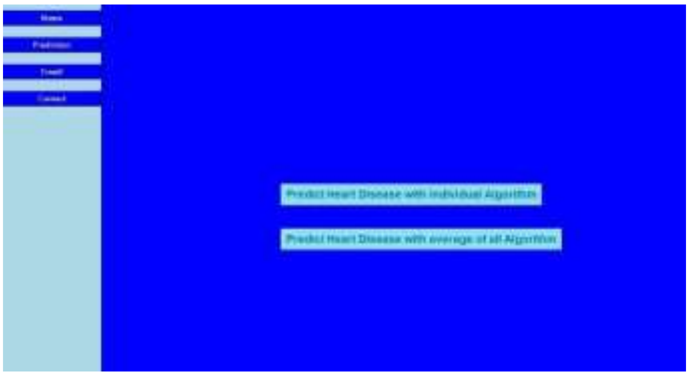

Home
Prediction
Email
Contact
Predict Heart Disease with individual Algorithm
Predict Heart Disease with average of all Algorithm

Fig. 7. Home page of Heart Disease Prediction

The figure7 shows the home or navigation screen of a heart disease prediction application. It includes a side menu with options such as Home, Prediction, Email, and Contact, and provides buttons for predicting heart disease either using an individual algorithm or by averaging results from all algorithms.

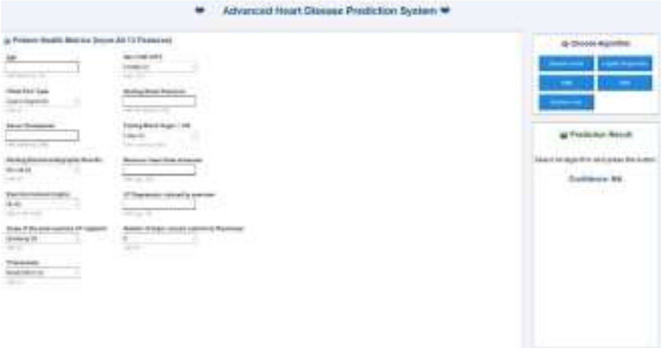


Fig. 8. Heart Disease Prediction with Individual Algorithm

The figure 8 displays an advanced heart disease prediction system interface where users can input detailed patient health metrics, choose a individual machine learning algorithm, and obtain prediction results along with confidence information, supporting comprehensive and customizable heart disease risk assessment.

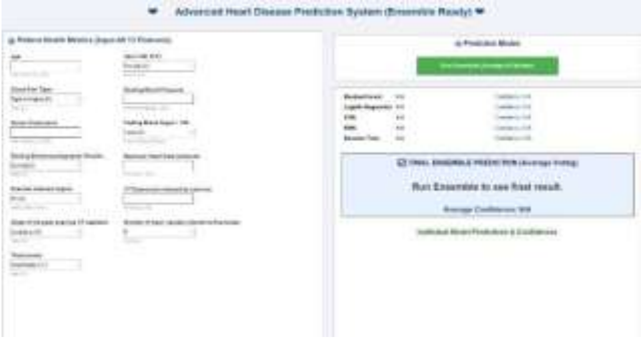

Advanced Heart Disease Prediction System (Ensemble Ready)
Run Ensemble to see final result.

Fig. 9. Heart Disease Prediction with Average of all Algorithm

The figure 9 illustrates an ensemble-ready advanced heart disease prediction system interface. It allows users to input all patient health features, run predictions using multiple machine learning models, and generate a final ensemble prediction (average voting) along with individual model predictions and confidence levels.

Overall, the project validates the effectiveness of machine learning–based approaches in cardiovascular risk prediction and highlights their potential role in improving healthcare outcomes through data-driven decision-making.

VI. Conclusion

Through the utilization of the Kaggle heart disease dataset, it has been demonstrated that multiple machine learning (ML) algorithms can effectively predict the likelihood of cardiac disease with considerable accuracy. The dataset, which serves as a widely recognized benchmark in the domain of medical data analytics, comprises approximately 920 patient records and includes sixteen distinct attributes that are highly relevant to cardiovascular health assessment. These features encompass a combination of demographic, clinical, and electrocardiographic parameters, providing a comprehensive representation of factors influencing heart disease risk.

Among the most significant attributes are age, gender, resting blood pressure, cholesterol level, fasting blood sugar, maximum heart rate achieved, exercise-induced angina, and the slope of the peak exercise ST segment. Additional parameters, such as chest pain type, resting electrocardiographic results, oldpeak (ST depression induced by exercise relative to rest), and thalassemia type, further enrich the dataset by capturing nuanced indicators of cardiovascular stress and function. The inclusion of both physiological and diagnostic data allows for the development of multifactorial predictive models, capable of identifying subtle interactions between risk variables that traditional statistical approaches might overlook.

To evaluate model performance, four distinct machine learning algorithms—Random Forest (RF), Support Vector Machine (SVM), AdaBoost, and K-Nearest Neighbors (KNN)—were trained and tested using the dataset. Each algorithm was assessed using standard performance metrics, including accuracy, precision, and recall, to determine its predictive efficiency. The experimental results revealed that the Random Forest model achieved the highest accuracy rate of 67.89%, outperforming the other models tested.

This superior performance can be attributed to Random Forest's ensemble learning structure, which aggregates multiple decision trees to minimize overfitting and improve generalization across diverse patient samples.

The findings underscore the robustness and reliability of Random Forest as a predictive model for cardiac disease classification. Its ability to handle non-linear relationships, accommodate missing or noisy data, and automatically evaluate feature importance makes it particularly well-suited for complex biomedical datasets. Furthermore, the model's interpretability enables clinicians and researchers to identify which clinical features contribute most significantly to heart disease prediction, thereby supporting evidence-based decision-making in healthcare.

Overall, the results of this analysis confirm that the Kaggle dataset provides a valuable foundation for predictive modeling in cardiology, and that machine learning—especially ensemble-based techniques like Random Forest—holds significant promise in enhancing the early detection and management of cardiovascular diseases.

References

[1] Sreejith, S., Rahul, S. and Jisha, R.C., 2016. A real time patient monitoring system for heart disease prediction using random forest algorithm. In Advances in Signal Processing and Intelligent Recognition Systems: Proceedings of Second International Symposium on Signal Processing and Intelligent Recognition Systems (SIRS-2015) December 16-19, 2015, Trivandrum, India (pp. 485-500). Springer International Publishing.

[2] El Hamdaoui, H., Boujraf, S., El Houda Chaoui, N., Alami, B. and Maaroufi, M., 2021. Improving Heart Disease Prediction Using Random Forest and AdaBoost Algorithms. International Journal of Online & Biomedical Engineering, 17(11).

[3] Pal, M. and Parija, S., 2021, March. Prediction of heart diseases using random forest. In Journal of Physics: Conference Series (Vol. 1817, No. 1, p. 012009). IOP Publishing.

[4] El-Shafiey, M.G., Hagag, A., El-Dahshan, E.S.A. and Ismail, M.A., 2022. A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest. Multimedia Tools and Applications, 81(13), pp.18155-18179.

[5] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y.R. and Suraj, R.S., 2021, January. Heart disease prediction using hybrid machine learning model. In 2021 6th international conference on inventive computation technologies (ICICT) (pp. 1329-1333). IEEE.

[6] Singh, Y.K., Sinha, N. and Singh, S.K., 2017. Heart disease prediction system using random forest. In Advances in Computing and Data Sciences: First International Conference, ICACDS 2016, Ghaziabad, India, November 11-12, 2016, Revised Selected Papers 1 (pp. 613-623). Springer Singapore.

[7] Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X., Yu, W. and Yan, J., 2020. Study of cardiovascular disease prediction model based on random forest in eastern China. Scientific reports, 10(1), p.5245.

[8] Gupta, O., Goyal, N., Anand, D., Kadry, S., Nam, Y. and Singh, A., 2020. Underwater networked wireless sensor data collection for computational intelligence techniques: issues, challenges, and approaches. Ieee Access, 8, pp.122959-122974.

[9] Soni, T., Gupta, D., Uppal, M. and Juneja, S., 2023, January. Explicability of artificial intelligence in healthcare 5.0. In 2023 International Conference on Artificial Intelligence and Smart Communication (AISC) (pp. 1256-1261). IEEE.

[10] Sharma, M., Dhasarathan, V., Patel, S.K. and Nguyen, T.K., 2020. An ultra-compact four-port 4× 4 superwideband MIMO antenna including mitigation of dual notched bands characteristics designed for wireless network applications. AEU-International Journal of Electronics and Communications, 123, p.153332.

[11] Soni, T., Uppal, M., Gupta, D. and Gupta, G., 2023, May. Efficient machine learning model for cardiac disease prediction. In 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN) (pp. 1-5). IEEE.

[12] Goyal, N., Dave, M. and Verma, A.K., 2020. SAPDA: secure authentication with protected data aggregation scheme for improving QoS in scalable and survivable UWSNs. Wireless Personal Communications, 113(1), pp.1-15.