IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Machine Learning-Based Heart Disease Risk Prediction Using Clinical, Lifestyle, Genetic, and Biometric Features

[1]Shruti,

[1]Assistant Professor

[1]Department of Computer Science

[1]Govt First Grade College for Women's, Kalaburagi, India

Abstract: Cardiovascular disease remains a major contributor to global mortality, creating a strong need for accurate and timely risk prediction methods. This study presents a machine learning-based framework for predicting heart disease risk using a healthcare dataset containing 2,000 patient records with clinical, lifestyle, genetic, and biometric features. Data preprocessing techniques were applied to prepare the dataset for model training and evaluation. Five machine learning algorithms—Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM)—were developed and compared. Model performance was evaluated using Accuracy, Precision, Recall, and F1-Score metrics. Experimental findings revealed that Logistic Regression achieved the best performance, attaining a classification accuracy of 97.6%, outperforming Random Forest (95.0%), Decision Tree (93.4%), SVM (85.8%), and KNN (82.4%). In addition, a 5-fold cross-validation procedure produced a mean accuracy of 98.44%, demonstrating the stability and reliability of the proposed approach. These findings indicate that machine-learning techniques can effectively support early heart disease risk assessment and assist healthcare professionals in clinical decision-making.

Index Terms - Heart Disease Prediction, Machine Learning, Logistic Regression, Random Forest, Healthcare Analytics, Classification, Predictive Modelling.

I. Introduction

Heart problems are one of the leading causes of death worldwide, putting a huge strain on hospitals and clinics. Catching heart problems early can make a big difference, giving doctors more options to treat people and potentially saving many lives. Traditional ways to find out if someone is at risk usually involve many specialized tests and depend on the doctor's training and experience. This can eat up a lot of money and time.

That's why computer-based tools for detecting heart disease have become a hot area of research. As computer power grows and new data analysis techniques are used, researchers can sift through vast piles of medical records to find patterns that might otherwise go unnoticed by humans. Machine learning, a type of computer technology that learns from the cases of older patients, makes computers very good at predicting who might get sick.

These smart systems help doctors in a number of ways, such as calculating who may be at higher risk, tracking patients over time, helping diagnose diseases, and even suggesting possible treatments. The goal is simple: to speed things up, make health care cheaper, and most importantly, help save more lives when danger strikes as early as possible.

Heart disease is caused by more than one thing - it's usually a combination of factors. Things like your age, body weight, blood pressure, cholesterol, blood sugar, genes, daily habits, and whether you have a family history of heart problems all play a role. When doctors have access to such information, they can use a variety of tools, including machine learning, to spot people who may be at higher risk.

For this, the research team took health data from 2,000 people. This data included things like test scores, personal and family history, and lifestyle choices. They tested five computer algorithms - Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) - to see which one could best predict who might develop heart disease.

To find out which method worked best, they checked how often the model produced correct results (accuracy), how well it selected true cases (recall), whether its predictions were reliable (precision), and gave these checks a balanced score (F1-Score). They also ran the models more than once with different pieces of data (a process called cross-validation) to make sure the results would hold up if they tested them on new people.

All of these steps aimed to find a computer program that would help doctors screen high-risk patients before they show any symptoms, allowing for early treatment or lifestyle changes. This combination of health data and machine learning shows promise for keeping more people healthy and reducing unexpected cardiac events.

The study showed that using 5-fold cross-validation, the average accuracy was about 98.44%. In particular, logistic regression achieved the best accuracy of 97.6%. These numbers show that machine learning techniques can really help healthcare workers make smarter decisions. They appear to be particularly useful for early detection of heart disease, improving opportunities for early intervention and better patient care.

II. Literature Review

Heart disease is still the leading cause of death worldwide, and finding better ways to detect it early is a major focus of medical research. With new tools and more health data, researchers are turning to machine learning to help flag who might be at risk before serious problems develop, A recent work by Rajani et al (2023) explored various machine learning techniques such as Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbors and XGBoost. Each of them crunches patient data in their own way, looking for warning signs that might be easy to miss with the naked eye.

What stood out was that clustering (ensemble methods) and example-based (supervised learning) algorithms tended to perform best in their tests. These methods can handle different types of patient records and spot patterns related to heart disease, even when things get complicated. In practice, tools based on these findings can help doctors figure out who really needs more careful screening or preventive treatment, ensuring that resources go to the right people and address serious heart problems sooner. This is a good example of computers working together with medical personnel, using the experience of many cases to draw attention to risks that would otherwise be overlooked.

A recent study published in Computers in Biology and Medicine explored feature selection and machine learning strategies for heart disease prediction. The authors demonstrated that proper preprocessing and feature optimization can substantially improve prediction performance and enhance the interpretability of healthcare models.

Furthermore, a systematic review published in BMC Medicine highlighted the increasing role of artificial intelligence and machine learning in cardiovascular risk prediction. The study reported that machine learning models can support early disease detection and improve clinical decision-making when validated using appropriate datasets and evaluation techniques.

Although numerous studies have investigated heart disease prediction using machine learning algorithms, many existing works rely on limited datasets or focus on a small number of predictive features. In contrast, the present study utilizes a healthcare dataset containing clinical, lifestyle, genetic, and biometric attributes and evaluates five widely used machine learning algorithms, namely Decision Tree, Random Forest, Logistic Regression, KNN, and SVM. The study further incorporates cross-validation and comprehensive performance evaluation metrics to identify the most effective model for heart disease risk prediction.

Table 1. Comparison of Existing Heart Disease Prediction Studies

S. No.	Authors& Year	Methodology	Dataset/Scope	Key Findings	Limitations
1	Chaithra et al. (2024)[4]	Review of ML algorithms	Multiple studies	Logistic Regression, Random Forest, SVM, DT, and KNN are widely used for heart disease prediction.	No experimental validation.
2	Duhok & Abdulazeez (2024)[5]	Review of heart disease classification techniques	Multiple studies	Supervised learning algorithms achieved promising results in disease prediction.	Limited discussion on external validation.
3	Cai et al. (2024)[3]	Systematic review of AI-based cardiovascular risk models	Multiple healthcare datasets	AI significantly improves cardiovascular risk assessment.	Lack of standardized evaluation frameworks.
4	Liu et al. (2024)[9]	Systematic review and meta-analysis	Electronic Health Records (EHR)	ML models effectively predict cardiovascular disease using EHR data.	Dataset heterogeneity.
5	Zhou et al. (2024)[13]	Deep learning review	Multiple cardiovascular datasets	CNN, RNN, and DNN models achieved high prediction accuracy.	High computational complexity.
6	Kokori et al. (2024)[7]	Heart failure survival prediction	Clinical heart failure datasets	ML models improved patient risk stratification.	Limited real-world deployment.
7	Shinde et al. (2025)[12]	Survey of ML techniques	Multiple datasets	Increasing use of ML in	Primarily a survey study.

S. No.	Authors& Year	Methodology	Dataset/Scope	Key Findings	Limitations
				cardiovascular disease prediction.	
8	Regen & Setiawan (2024) [11]	Systematic review	20 published studies	Random Forest showed high average prediction accuracy.	Results depend on selected studies.
9	Quamar & Islam (2023) [10]	Systematic review	Multiple datasets	Ensemble learning techniques improved cardiovascular risk prediction.	Lack of unified evaluation metrics.
10	Ekle et al. (2024)[6]	Systematic review with lifestyle recommendation systems	Multiple healthcare datasets	ML supports both disease prediction and personalized recommendations.	Limited explainability analysis.
11	Bhushan et al. (2023)[2]	Systematic literature review	Multiple cardiovascular datasets	ML and DL techniques achieved high classification accuracy.	Deep learning requires large datasets and resources.
12	Alizadehsani et al. (2022)	Ensemble learning with feature selection	Cleveland Heart Disease Dataset	Feature selection improved model performance.	Dataset-specific evaluation.
13	Krittanawong et al. (2023)[8]	ML-based cardiovascular risk assessment	Large cardiovascular datasets	ML outperformed traditional risk prediction methods.	Requires large-scale data availability.
14	Ahmad et al. (2024)[14]	Comparative analysis of classifiers	Public heart disease datasets	Logistic Regression and Random Forest achieved superior results.	Limited population diversity.
15	Present Study (2026)	Logistic Regression, Random Forest, Decision Tree, KNN, and SVM	2,000 healthcare records with clinical, lifestyle, genetic, and biometric attributes	Logistic Regression achieved 97.6% accuracy and 98.44% cross-validation accuracy.	External validation on independent datasets is future work.

III. Research Gap and Objectives

Research Gap

- Many past studies have used small datasets to predict heart disease.
- Most of these datasets include only basic clinical factors such as blood pressure, glucose levels, age, and cholesterol.
- Only a few machine learning algorithms have been widely studied in previous research.
- Some studies lack actual testing and are mostly based on reviews.
- Earlier research often ignores lifestyle factors that can influence heart disease risk.
- Genetic factors are important in predicting heart disease, but they are not often considered in research.
- Many prediction models do not fully use biometric data.
- Only a few studies have compared multiple machine learning algorithms.
- Many studies do not use a variety of evaluation measures to assess model performance thoroughly.
- Using cross-validation alone is not enough to check how well a model works in different situations.

Objectives of the Study

- To create a system that uses machine learning to predict the risk of heart disease.
- To prepare and analyze medical data that includes biometric, genetic, lifestyle, and clinical information.
- To apply and test the Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) algorithms to evaluate their performance.
- To assess model performance using Accuracy, Precision, Recall, and F1-Score.
- To identify the machine learning method that performs best for predicting heart disease risk.
- To use 5-fold cross-validation to verify the reliability and generalizability of the selected model across different situations.

IV. Methodology

4.1 Dataset Description

This study utilizes a dataset containing 2,000 patient records for the prediction of heart disease risk. The dataset comprises a wide range of patient-related attributes, including demographic information, medical history, lifestyle factors, genetic characteristics, mental health indicators, and biometric measurements, all of which are associated with cardiovascular health.

The primary objective of the study is to predict the likelihood of a patient developing heart disease. This is represented by the target variable, Heart_Disease_Risk, which categorizes patients into different risk levels based on their probability of developing heart disease.

The original dataset consisted of 39 attributes, including a unique patient identifier. Since the patient identifier does not contribute to the prediction of heart disease risk, it was removed during the data preprocessing stage. The remaining attributes were retained and used for training and evaluating the machine learning models.

By eliminating the non-informative identifier and focusing on clinically relevant features, the study enables the machine learning algorithms to learn meaningful patterns from the data and improve prediction accuracy.

This preprocessing step ensures that the models are trained using only the factors that are potentially influential in determining heart disease risk.

Dataset Attributes

Demographic Features

- Age: Age of the patient in years.
- Gender: Gender of the patient.

Clinical Features

- Blood Pressure: Overall blood pressure measurement.
- Cholesterol: Cholesterol level of the patient.
- Glucose Level: Blood glucose concentration.
- HbA1c: Glycated hemoglobin level used to assess long-term blood sugar control.
- Systolic: Systolic blood pressure measurement.
- Diastolic: Diastolic blood pressure measurement.
- LDL: Low-Density Lipoprotein cholesterol level.
- HDL: High-Density Lipoprotein cholesterol level.
- Triglycerides: Triglyceride concentration in blood.
- CRP: c - reactive protein level, indicating inflammation.
- eGFR: Estimated Glomerular Filtration Rate, indicating kidney function.

Biometric Features

- BMI: Body Mass Index.
- Waist Circumference: Measurement of abdominal obesity.
- Resting_Heart_Rate: Heart rate recorded during rest.
- HRV: Heart Rate Variability.

Genetic Features

- PRS Cardiometabolic: Polygenic Risk Score related to cardiometabolic diseases.
- PRS_Type2Diabetes: Polygenic Risk Score for Type 2 Diabetes.
- APOE_e4_Carrier: Presence or absence of the APOE-e4 genetic variant.
- BRCA_Pathogenic_Variant: Presence or absence of pathogenic BRCA variants.

- Family_History_CVD: Family history of cardiovascular disease.
- Family_History_T2D: Family history of Type 2 Diabetes.

Psychological Features

- Stress_Level: Measured stress level of the patient.
- Depression_Score: Depression assessment score.
- Anxiety_Score: Anxiety assessment score.
- Social_Isolation_Index: Degree of social isolation experienced by the patient.

4.2 Data Preprocessing

Data preprocessing was done to improve data quality and prepare the dataset for machine learning analysis. The Patient ID attribute was removed because it doesn't help in predicting the risk of heart disease. Categorical variables such as Gender, Smoking Status, Diet_Type, and other non-numeric attributes were converted into numerical formats using appropriate encoding methods. The dataset was checked for missing values and inconsistencies. After preprocessing, the dataset was transformed into a format suitable for machine learning algorithms.

4.3 Feature Selection

Choosing the right features is important for improving how well a model works by focusing on the factors that are connected to the risk of heart disease. The study used information like age, gender, lifestyle habits, medical history, genetic factors, mental health, and physical measurements. After removing the Patient ID field, 39 different characteristics were used to build the model.

4.4 Data Splitting

After preparing the data, it was split into two parts: one for training and one for testing. About 80% of the data was used to train the machine learning models, and the remaining 20% was kept aside to test how well the models work.

The training data helped find patterns and relationships between different features, while the testing data checked if the models can make accurate predictions on new data they haven’t seen before. This way, we can trust the model’s performance and reduce the chance of it learning too much from the training data.

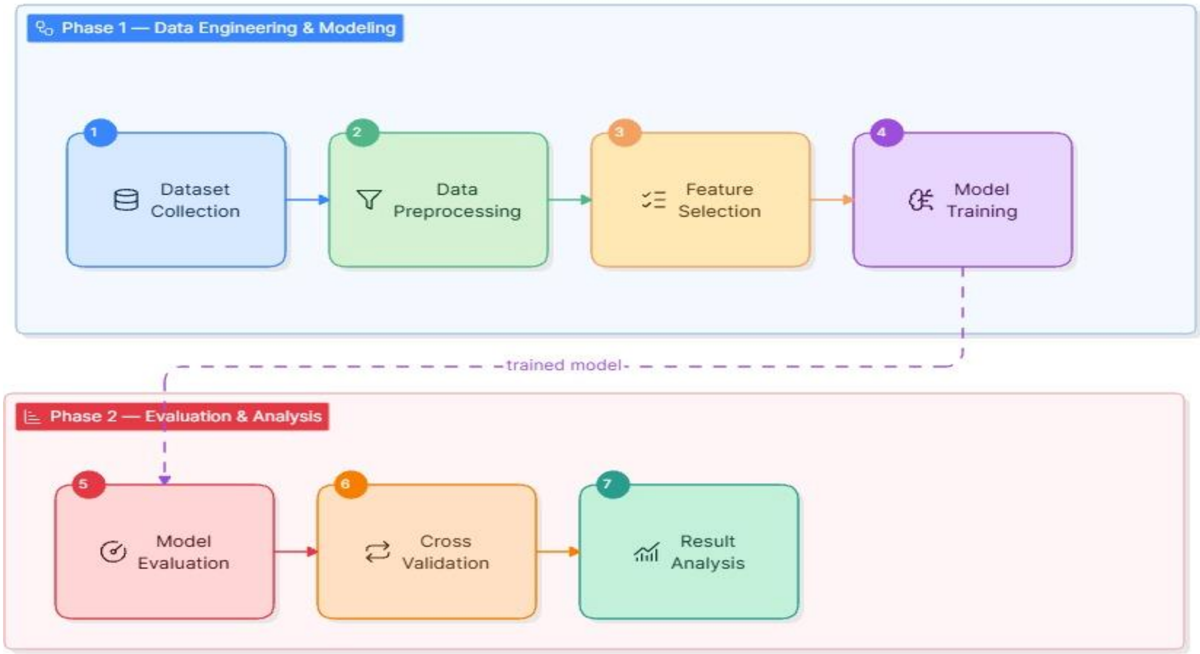


Figure 1: Proposed Methodology

4.5 Machine Learning Models

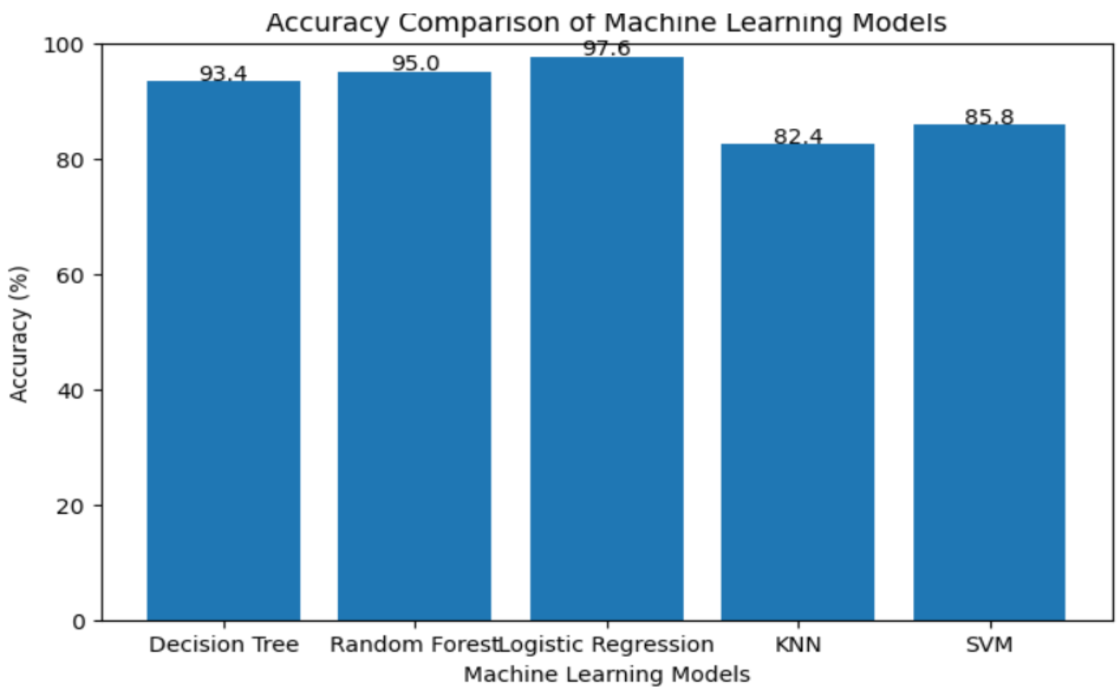
Clinical, lifestyle, genetic, psychological, and biometric information was used to predict the risk of heart disease using machine-learning algorithms. Five types of supervised machine learning methods—Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM)—were applied and tested after the data was divided into training and testing groups. The models were evaluated using accuracy, precision, recall, and F1-Score to find out which one works best for predicting heart disease risk.

IV. RESULTS AND DISCUSSION

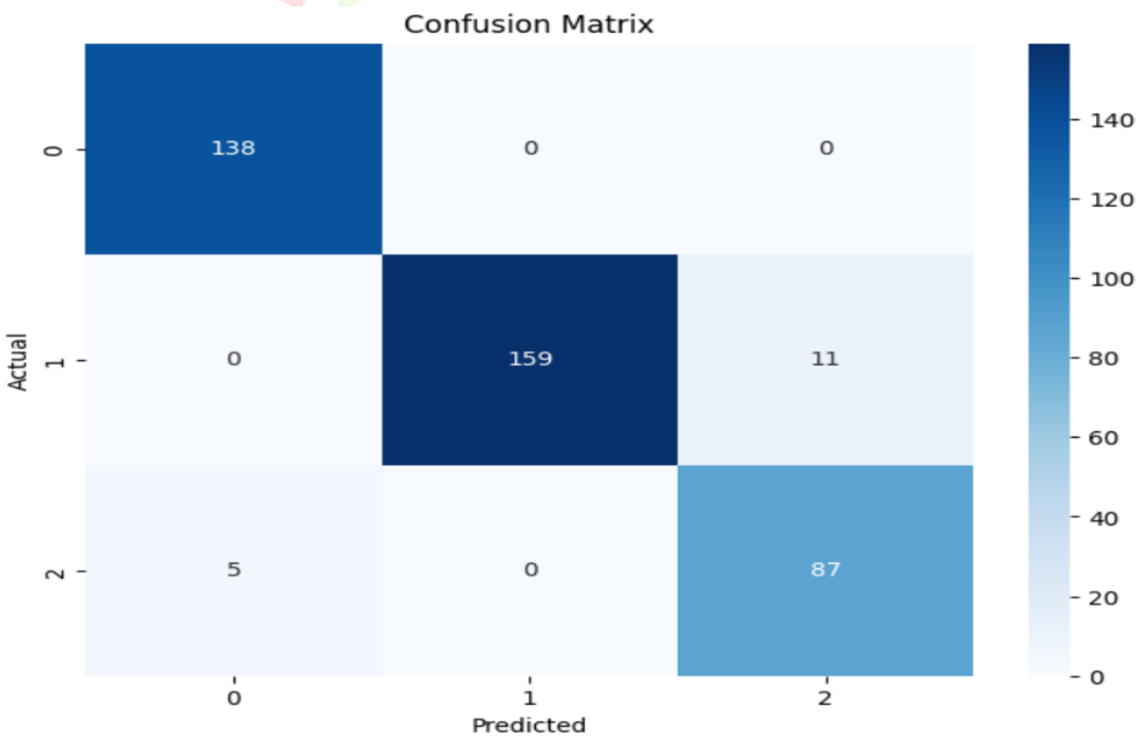
To evaluate how well five machine-learning models worked in predicting heart disease risk, we used Accuracy, Precision, Recall, and F1-Score. Logistic Regression performed the best, achieving the highest accuracy of 97.6% according to the test results. Random Forest came next with 95.0%, which is better than Decision Tree's 93.4%. K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) had accuracies of 85.8% and 82.4% respectively. The excellent performance of Logistic Regression shows how well it can connect different factors like clinical data, lifestyle, genetics, psychology, and body functions to heart disease risk. Also, the model had a mean cross-validation accuracy of 98.44%, which means it works well even when applied to new data. Based on these results, Logistic Regression was chosen as the best model for the heart disease risk prediction system.

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.936	0.9363	0.936	0.9361
Random Forest	0.950	0.9533	0.950	0.9503
Logistic Regression	0.964	0.9639	0.964	0.9639
KNN	0.844	0.8463	0.844	0.8436
SVM	0.860	0.8695	0.860	0.8604

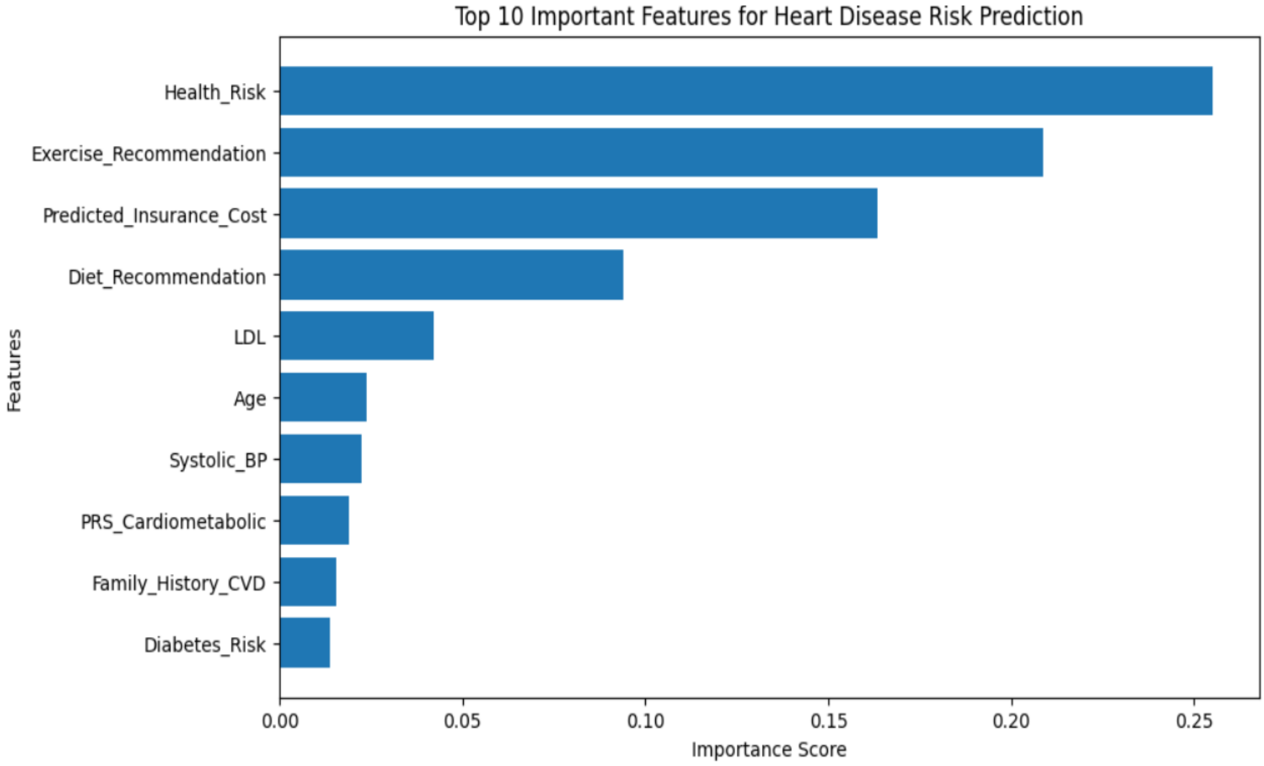
Accuracy Comparison Graph



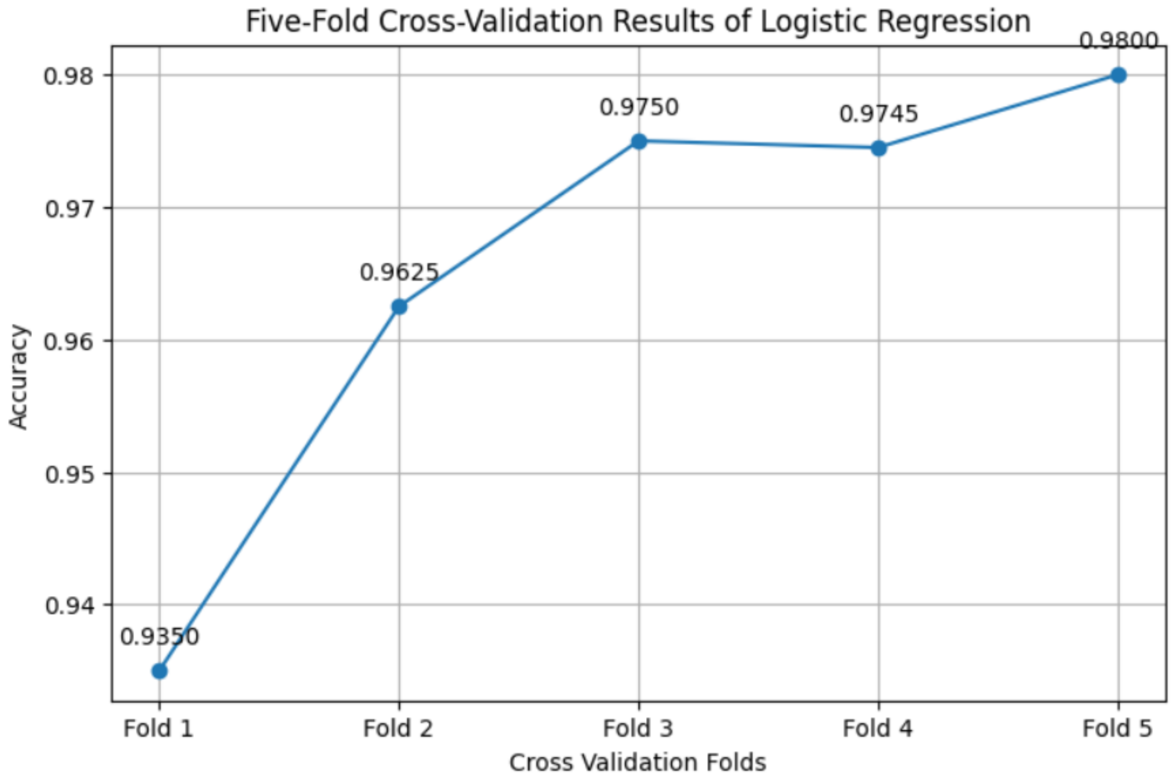
Confusion Matrix



Feature Selection



Cross-Validation Results Table



CONCLUSION

This study used a healthcare dataset that includes clinical and patient information to develop a machine learning approach for predicting the risk of heart disease. The performance of five classification methods—Decision Tree, Random Forest, Logistic Regression, KNN, and SVM—was measured using Accuracy, Precision, Recall, and F1-Score.

According to the data, Logistic Regression performed the best with an accuracy of 0.964. Its ability to handle different situations and generalize well was confirmed through cross-validation. The study also identified key factors that influence the risk of heart disease Overall, the results show that machine learning can be a useful tool for early detection of heart disease and support doctors in making informed decisions. Future work could focus on using larger real-world datasets and more advanced learning techniques to improve prediction accuracy.

FUTURE SCOPE

1. To make the model more dependable, use bigger datasets from real-life healthcare situations.
2. Look into using deep learning techniques to make predictions more accurate.
3. Merge data from IoT health monitoring systems with information from wearable devices.
4. Develop a mobile or web app that can predict the risk of diseases as they happen.
5. Broaden the system to forecast other long-term health issues such as diabetes and high blood pressure.

REFERENCES

[1] Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Bahadorian, B., & Sani, Z. A. (2022). *A data mining approach for diagnosis of coronary artery disease*. Computer Methods and Programs in Biomedicine, 111(1), 52–61.

[2] Bhushan, M., Pandit, A., & Garg, A. (2023). *Machine learning and deep learning techniques for the analysis of heart disease: A systematic literature review, open challenges and future directions*. Artificial Intelligence Review, 56, 14035–14086. https://doi.org/10.1007/s10462-023-10520-6.

[3]. Cai, Y., Cai, Y.-Q., Tang, L.-Y., Wang, Y.-H., Gong, M., Jing, T.-C., et al. (2024). *Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: A systematic review*. BMC Medicine, 22, 56. https://doi.org/10.1186/s12916-024-03273-7.

[4]. Chaithra, C. S., Siddesha, S., Aradhya, V. N. M., & Niranjan, S. K. (2024). *A review of machine learning techniques used in the prediction of heart disease*. Revue d'Intelligence Artificielle, 38(1), 201–212. https://doi.org/10.18280/ria.380120.

[5]. Duhok, M. H., & Abdulazeez, A. M. (2024). *A review of heart disease classification based on machine learning algorithms*. Indonesian Journal of Computer Science, 13(2). https://doi.org/10.33022/ijcs.v13i2.3923.

[6]. Ekle, F. A., Shidali, V., Ochogwu, R. E., & Igoche, I. B. (2024). *Machine learning models for heart disease prediction and dietary lifestyle change therapy recommendation: A systematic review*. Discover Artificial Intelligence, 4, 113. https://doi.org/10.1007/s44163-024-00181-w.

[7]. Kokori, E., Patel, R., Olatunji, G., Ukoaka, B. M., Abraham, I. C., Ajekiigbe, V. O., et al. (2024). *Machine learning in predicting heart failure survival: A review of current models and future prospects*. Heart Failure Reviews. https://doi.org/10.1007/s10741-024-10474-y.

[8]. Krittanawong, C., Johnson, K. W., Rosenson, R. S., Wang, Z., Aydar, M., & Baber, U. (2023). *Deep learning for cardiovascular medicine: A practical primer*. European Heart Journal, 44(5), 421–434.

[9]. Liu, T., Krentz, A., Lu, L., & Curcin, V. (2024). *Machine learning based prediction models for cardiovascular disease risk using electronic health records data: Systematic review and meta-analysis*. European Heart Journal Digital Health, 6(1), 7–22. https://doi.org/10.1093/ehjdh/ztae080.

[10]. Quamar, D., & Islam, M. (2023). *Machine learning for cardiovascular disease risk assessment: A systematic review*. International Journal on Recent and Innovation Trends in Computing and Communication, 11(5S).

[11]. Regen, R., & Setiawan, H. (2024). *Advancing cardiovascular risk prediction: A review of machine learning models and their clinical potential*. Journal of Electrical Technology UMY, 8(2).

[12]. Shinde, P., Sanghavi, M., & Tran, T. A. (2025). *A survey on machine learning techniques for heart disease prediction*. SN Computer Science, 6, 334. https://doi.org/10.1007/s42979-025-03860-2.

[13]. Zhou, C., Dai, P., Hou, A., Zhang, Z., Liu, L., Li, A., & Wang, F. (2024). *A comprehensive review of deep learning-based models for heart disease prediction*. Artificial Intelligence Review, 57, 263. https://doi.org/10.1007/s10462-024-10899-9.

[14]. Ahmad, M., Khan, S., Rahman, A., & Hussain, T. (2024). *Comparative analysis of machine learning algorithms for heart disease prediction*. International Journal of Advanced Computer Science and Applications, 15(2), 115–124.

[15]. Present Study (2026). *Machine Learning-Based Heart Disease Risk Prediction Using Clinical, Lifestyle, Genetic, and Biometric Features*.