09.30.2025 JOURNAL OF

Systematic and Modern Science Research (JSMSR) Vol. 9 No.9

MACHINE LEARNING TECHNIQUES FOR CARDIOVASCULAR DISEASE PREDICTION

ASEGUNLOLUWA E. BABALOLA; & TEKENA SOLOMON

Department of Computing, Anchor University Lagos

DOI Link: https://doi.org/10.70382/bejsmsr.v9i9.013

ABSTRACT

Cardiovascular disease (CVD) continues to be the leading cause of death globally, accounting for millions of lives annually. Early and accurate prediction of CVD is critical for prevention and intervention, yet existing clinical approaches often fail to capture the complex interactions among multiple risk factors. While machine learning (ML) has been increasingly applied to CVD prediction, many studies rely on single datasets, limited sample sizes, or overlook systematic evaluation across different algorithms. This study addresses these limitations by employing a merged dataset of 920 patient records, created by integrating four well-known heart disease datasets from Cleveland, Hungarian, Switzerland, and Long Beach VA based on 11 common features. The integration enhances data diversity and improves model generalization compared to single-dataset approaches. Four supervised ML algorithms, namely, Logistic Regression, Naïve Bayes, Decision Tree, and K-Nearest Neighbour (KNN), were systematically implemented and evaluated. Data preprocessing involved feature scaling for non-tree-based algorithms and tailored encoding strategies for categorical

Introduction

Cardiovascular disease (CVD), also referred to as heart disease, is a major global health challenge and one of the foremost causes of death and disability. According to the World Health Organization (2021), more than 12 million deaths occur worldwide each year due to CVD. CVD encompasses a wide range of disorders involving the heart and blood vessels, such as coronary artery disease, stroke, peripheral artery disease, congenital heart conditions, and endocarditis. Because the heart is central to sustaining life, any disruption in its function can trigger serious systemic consequences.

Over the past few decades, the global burden of cardiovascular diseases has continued to rise. Despite advances in treatment and preventive care, CVD remains among the top five causes of

E-ISSN 3027-2939 P-ISSN 3026-8397

variables. Model training and validation were carried out using Stratified 5-Fold Cross-Validation, ensuring balanced representation of classes across folds. The results shows KNN algorithm consistently outperformed all models, achieving 91.6% accuracy with precision and recall of 92%, highlighting its robustness in capturing non-linear decision boundaries in heterogeneous datasets. Importantly, the study shows that integrating multiple datasets strengthens predictive power and improves the reliability of ML models in cardiovascular diagnosis. Findings demonstrate that simple yet effective models such as KNN, when applied to enriched datasets, can deliver clinically meaningful insights and serve as practical tools for early risk detection.

Keywords: Heart disease prediction; Machine learning; Logistic Regression; Naïve Bayes; Decision Tree; K-Nearest Neighbour;

morbidity and mortality worldwide (Virani et al., 2021). This growing prevalence is linked to aging populations, urbanization, sedentary lifestyles, and unhealthy dietary patterns. Alarmingly, CVD is often described as a silent killer since it can progress without obvious symptoms until it reaches a critical stage (Mensah et al., 2019). As a result, many individuals only seek care after the disease has already caused significant damage, highlighting the urgent need for early detection.

Research has consistently shown that many cardiovascular conditions can be prevented or delayed by addressing modifiable risk factors. High blood pressure, diabetes, obesity, smoking, poor diet, and physical inactivity are the most common contributors to CVD (Benjamin et al., 2019). Identifying these risk factors in advance makes it possible to implement lifestyle interventions and medical treatments that significantly reduce the chance of adverse outcomes. Therefore, predicting the risk of CVD at an early stage is critical for improving patient outcomes and reducing healthcare costs.

To support this goal, data-driven approaches have gained prominence in clinical sciences. Information mining and machine learning (ML) technologies have demonstrated strong potential in analyzing health data and improving disease diagnosis. ML models are particularly well suited to handling the large, complex, and heterogeneous datasets generated by modern healthcare systems (Shinde & Shah, 2018). They can uncover hidden patterns, learn from historical patient data, and generate predictions that aid physicians in clinical decision-making.

Unlike traditional statistical methods that often assume linear relationships among variables, ML models are capable of capturing complex, non-linear interactions among risk factors. This makes them more flexible and accurate in modeling the multifaceted

nature of cardiovascular diseases (Rajkomar et al., 2019). Moreover, the integration of ML into healthcare promises not only better diagnostic accuracy but also faster risk assessment, enabling physicians to allocate resources more efficiently and personalize treatment strategies.

While several studies have applied ML to heart disease prediction, many rely on limited single datasets or lack systematic evaluation across multiple algorithms. This reduces generalizability and limits clinical applicability. This study aims to address these limitations by applying and comparing four machine learning algorithms - Logistic Regression, Naïve Bayes, Decision Tree, and K-Nearest Neighbour (KNN), on an integrated dataset compiled from four well-known heart disease datasets. Although CVD can manifest in many different forms, a standard set of predictors such as age, cholesterol levels, blood pressure, and smoking status often determine risk. Utilizing ML to analyze these variables can help healthcare professionals take preventive measures in time, ultimately reducing the burden of CVD and improving quality of life.

Literature Review

Machine learning (ML), a subset of artificial intelligence, focuses on building systems that can learn patterns from historical data and use those patterns to make predictions on new inputs. Rather than relying solely on explicit programming, ML algorithms infer relationships within training data and generalize these inferences to unseen instances (Topol, 2023). In the medical domain, ML has shown strong promise in augmenting clinical decision-making by identifying subtle associations that might escape human intuition (Jiang et al., 2023).

In practice, a typical ML workflow begins with data preparation: cleaning, handling missing values, and encoding features appropriately. Then, the data is partitioned into training and testing subsets. The training set is used to fit models, which learn the mapping from features to outcomes, while the testing set evaluates the model's generalization. Performance metrics such as Area Under the Curve (AUC)- Receiver Operating Characteristic (ROC), precision, recall, and F1-score are then used to measure effectiveness in classification tasks (Jiang et al., 2023).

Because healthcare data often suffers class imbalance (fewer positive cases than negatives), evaluation approaches like stratified cross-validation are particularly important to preserve the class distribution in each fold and avoid biased estimates (Topol, 2023). ML offers a powerful framework for heart disease prediction that continues to improve with newer methods and larger datasets by combining rigorous preprocessing, robust algorithm selection, and careful evaluation (Zhou et al., 2024).

Machine Learning Techniques

a) **Logistic Regression:** Logistic regression is a supervised classification method commonly used in clinical predictive modeling to estimate the probability of a binary outcome (e.g., disease vs. no disease) from a set of predictor variables. It uses the logistic (sigmoid) function to map linear combinations of independent variables into a probability ranging between 0 and 1 (Zabor et al., 2022).

   In healthcare research, logistic regression remains popular for its interpretability: model coefficients can be converted into odds ratios, which clinicians can use to understand how each predictor is associated with outcome risk (Yu & Wu, 2023). Recent reviews emphasize its continued relevance in disease risk prediction, diagnosis, and treatment outcome modeling, while also noting challenges such as multicollinearity, sample imbalance, and non-linearity in relationships among predictors (Asif et al., 2025).

b) **Naïve Bayes:** Naïve Bayes is a probabilistic machine learning algorithm widely applied to classification problems. It is based on Bayes' theorem, which provides a mathematical framework for estimating the probability of a class given observed features. Despite its simplicity, Naïve Bayes has remained a powerful and efficient algorithm across domains, including healthcare, finance, and text mining (Bose et al., 2023).

   The algorithm makes a key assumption of conditional independence, meaning that the presence of one feature is assumed to be independent of the presence of another given the class label. While this assumption is often unrealistic in real-world datasets, Naïve Bayes still performs remarkably well in many applications because it simplifies computation while retaining predictive accuracy (Kumar & Singh, 2024).

   In medical prediction tasks, Naïve Bayes has been used effectively for disease risk classification, including cardiovascular disease. Its advantages include fast training time, low computational cost, and the ability to handle both categorical and continuous features. Studies have also highlighted its robustness with smaller datasets and noisy data, making it suitable for clinical applications where data may be limited or imperfect (Ali et al., 2025). However, its performance may decrease when strong correlations exist among predictors, as this violates the independence assumption.

c) **Decision Tree:** Decision Trees are supervised learning models that classify data by repeatedly splitting it into subsets based on feature values until a decision is reached. Their hierarchical structure consists of nodes representing conditions, branches representing decision paths, and terminal leaves corresponding to predicted outcomes. This design makes them easy to interpret and particularly

useful in medical decision-making, where transparency is vital (Sivapalan et al., 2023).

In cardiovascular disease prediction, Decision Trees are valued for their ability to handle both numerical and categorical inputs without requiring extensive preprocessing (Dey et al., 2020). However, their major drawback is a tendency to overfit, especially when the tree grows deep and complex. Recent studies highlight that while standalone Decision Trees may underperform compared to other machine learning models, they remain relevant as building blocks for ensemble approaches such as Random Forest and Gradient Boosting, which significantly enhance accuracy and robustness (Mahajan et al., 2023).

d) **K-Nearest Neighbour (KNN):** K-Nearest Neighbour (KNN) is a simple but effective non-parametric algorithm used for classification. It works by comparing new instances with the most similar cases in the training set, assigning the class label based on the majority vote of the k closest neighbors. The algorithm's strength lies in its ability to capture non-linear relationships without assuming an underlying data distribution (Halder et al., 2024).

KNN has been applied successfully in medical prediction tasks, including heart disease detection, due to its flexibility and competitive performance on well-prepared datasets. Its accuracy, however, depends heavily on the choice of k and the distance metric used. Moreover, it can become computationally expensive with large datasets because predictions require distance calculations for every new sample. Recent reviews show that, despite these limitations, KNN continues to perform well compared to more complex models, making it a reliable baseline in healthcare analytics (Ali et al., 2025).

Methodology

The methodology adopted for this study follows a structured data science pipeline, as shown in the system architecture. The process consists of several sequential steps: exploratory data analysis, data preprocessing, model training, and performance evaluation depicted in fig. 1.

a) **Dataset Collection:** The study uses a dataset containing patient health records, including demographic, clinical, and lifestyle-related attributes. These features represent potential risk factors for cardiovascular disease and serve as the foundation for model development.

b) **Exploratory Data Analysis (EDA):** Exploratory data analysis was performed to understand the distribution of features, detect missing values, and identify potential outliers. Visualization techniques such as histograms, box plots, and correlation heatmaps were applied to uncover relationships among variables and assess their relevance to cardiovascular risk prediction.

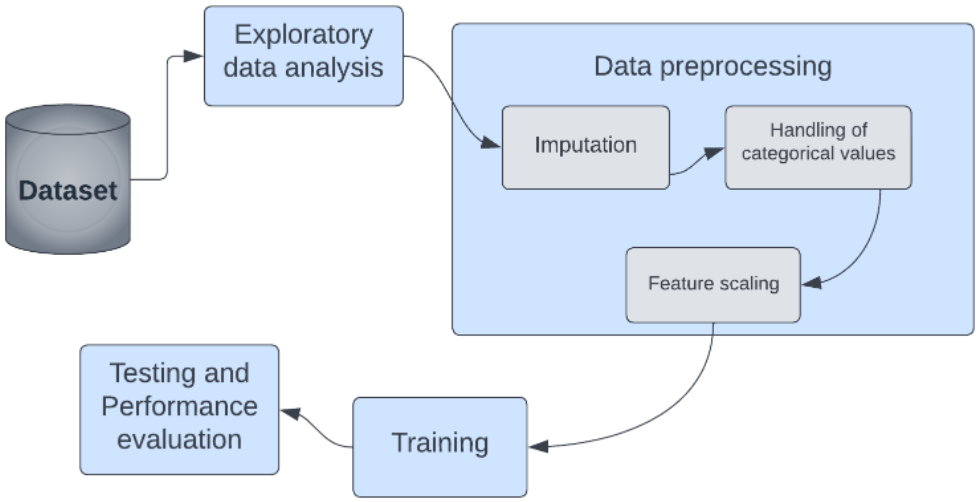

Fig. 1: System Architecture

c) **Data Preprocessing:** Since raw medical data is often incomplete and inconsistent, preprocessing was essential to improve data quality before model training. The preprocessing steps included:

i. **Imputation:** Missing values in the dataset were replaced using appropriate strategies, such as mean/median substitution for numerical attributes and mode substitution for categorical variables.

ii. **Feature Scaling**: Scaling will be conducted using the MinMaxScaler class in the scikit-learn library. This technique will normalize continuous variables to a range of 0–1 and will be applied only to non-tree-based algorithms which are sensitive to feature magnitudes. Tree-based algorithms such as decision trees and random forests will not require scaling.

iii. **Handling of Categorical Values**: Categorical variables will be encoded differently depending on the model type. Label Encoding will be applied for the Decision Tree model, as tree-based algorithms can naturally handle categorical splits. For non-tree-based algorithms, One-Hot Encoding will be applied using the scikit-learn OneHotEncoder to ensure correct binary representation of categorical attributes.

d) **Model Training:** Following preprocessing, the dataset would be divided into training and testing subsets. Various machine learning algorithms were implemented to build predictive models, including Decision Tree, Naïve Bayes, Logistic Regression, and K-Nearest Neighbour (KNN). The training stage involved learning patterns from the input features and mapping them to the target output (presence or absence of cardiovascular disease). To evaluate model reliability and reduce overfitting, Stratified K-fold cross-validation will be

used. In this approach, the dataset will be divided into k folds while preserving the original class distribution across each fold. Each model will be trained on k-1 folds and validated on the remaining fold, with the process repeated until every fold has been used for validation. The average performance across folds will provide a robust estimate of model generalization

e) **Testing and Performance Evaluation:** The trained models will be evaluated using the testing dataset to assess their generalization ability. Performance metrics such as accuracy, precision, and recall will used to compare models.

Results and Discussion

The dataset used in this study was obtained from Kaggle, where multiple existing datasets were merged into a single collection as shown in fig 2. Specifically, four well-known heart disease datasets were combined based on 11 common features:

a) Cleveland dataset: 303 observations
b) Hungarian dataset: 294 observations
c) Switzerland dataset: 123 observations
d) Long Beach VA dataset: 200 observations

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.000	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.000	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.000	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.500	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.000	Up	0

Fig. 2: Heart Disease Dataset

This integration produced a more comprehensive dataset with a total of 920 observations. By combining these sources, the dataset captured a wider variety of patient records across different populations and clinical contexts, making it more robust and representative. Since the features were harmonized across datasets, the model could learn from a diverse sample, thereby improving generalization and predictive performance.

Exploratory Data Analysis

Exploratory data analysis (EDA) was carried out to understand the structure of the dataset and identify important patterns before model training. The merged dataset contained 920 patient records with 11 common features. An initial inspection revealed that heart disease was more prevalent among middle-aged and older individuals. As shown in Fig. 3, the distribution of age indicates that cases began to increase noticeably

from age 45, peaking between 55 and 60 years. This trend is consistent with clinical knowledge that cardiovascular risk escalates with age.

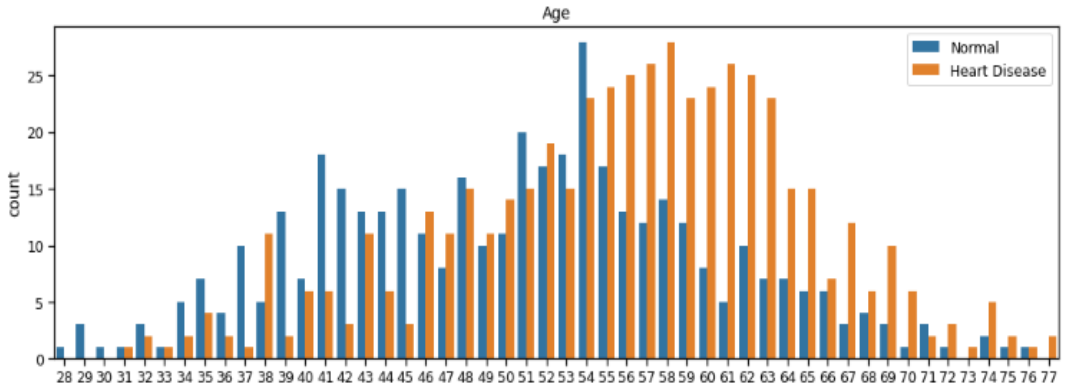

Fig 3: Age Distribution

Beyond age, other clinical and lifestyle-related variables also showed clear differences between patients with and without heart disease. For instance, fasting blood sugar (Fig. 4) demonstrated that while most patients had values ≤ 120 mg/dl, those with elevated levels were more likely to belong to the heart disease category. This supports the well-documented relationship between diabetes and cardiovascular complications.

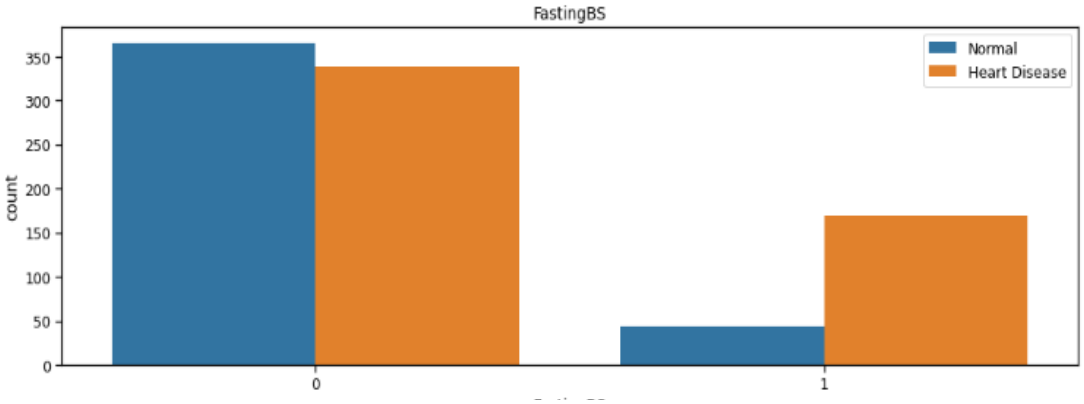

Fig 4: Fasting Blood Sugar

Demographic factors further reinforced these trends. Analysis of the sex variable (Fig. 5) showed that males were both more numerous in the dataset and more frequently diagnosed with heart disease compared to females. Although fewer female patients were represented, the majority were in the normal category, aligning with previous findings

that men are at higher risk at younger ages while women's risk increases post-menopause.

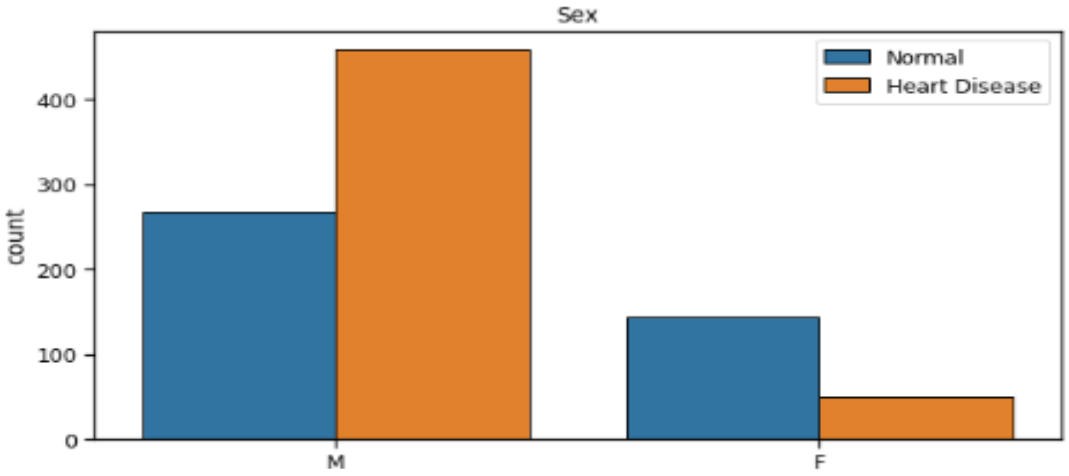

Fig. 5: Sex Distribution

Symptoms also played a crucial role in distinguishing the two groups. Chest pain type (Fig. 6) emerged as one of the strongest indicators. Patients classified with asymptomatic chest pain (ASY) were predominantly diagnosed with heart disease, whereas other categories such as typical angina (TA), atypical angina (ATA), and non-anginal pain (NAP) were more associated with normal cases. This underscores the diagnostic relevance of chest pain as a clinical predictor.

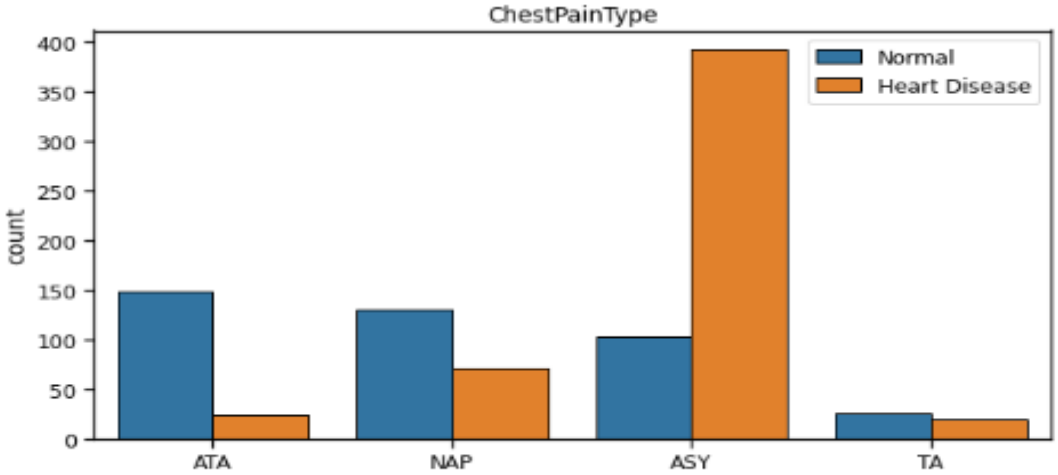

Fig. 5: Chest pain type

Finally, correlation analysis (Fig. 6) provided further evidence of the interplay between variables. Age correlated negatively with maximum heart rate (–0.38), reflecting the

natural decline in cardiovascular capacity with age. Heart disease itself showed positive associations with age (0.28), fasting blood sugar (0.27), and ST depression (oldpeak) (0.40), while being negatively correlated with maximum heart rate (–0.40). These patterns aligned closely with established clinical expectations and demonstrated that the dataset retained strong predictive signals for cardiovascular risk.

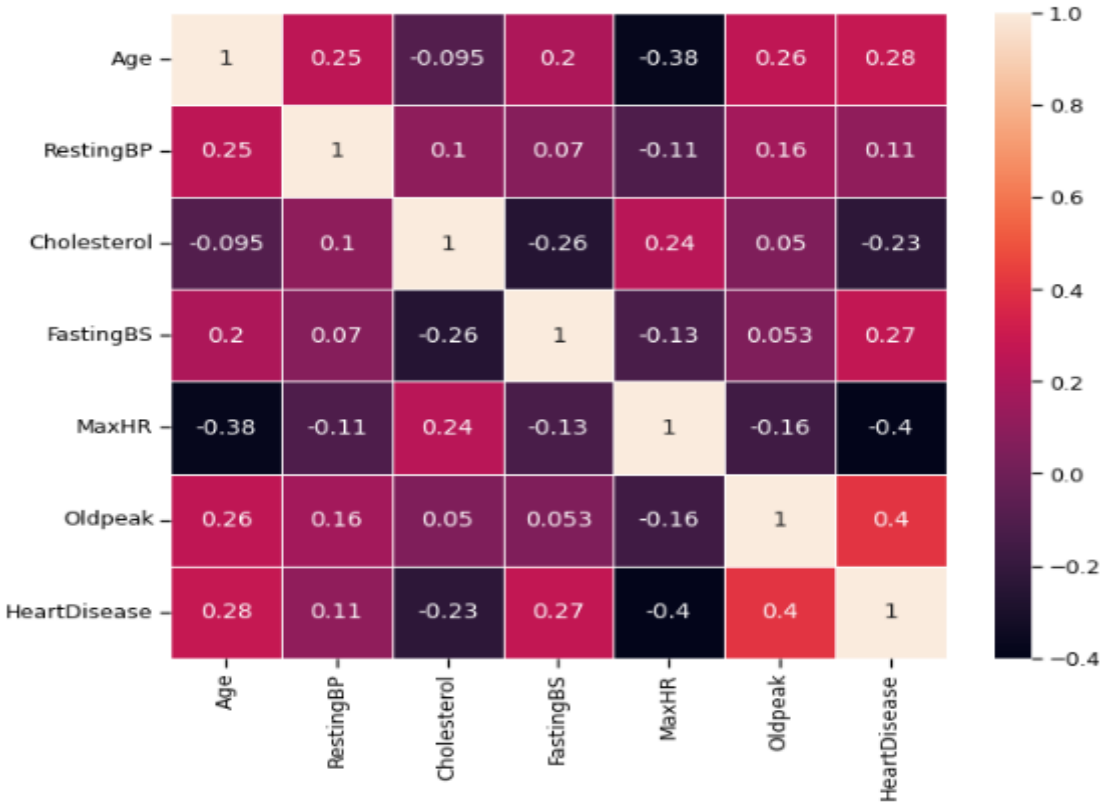

Fig. 7: Heat map correlation

a) **Data Preprocessing**

Before model training, the dataset was examined for missing values using summary functions, and the results confirmed that there were no null values across the 920 records. Consequently, imputation was not required, and the dataset was considered complete for analysis.

Although imputation was unnecessary, other preprocessing steps were applied to prepare the data for modeling. Feature scaling was carried out using the MinMaxScaler class from the scikit-learn library. This transformation normalized the continuous variables to a range between 0 and 1, ensuring that non-tree-based algorithms such as Logistic Regression, Naïve Bayes, and K-

Nearest Neighbour could train effectively without bias toward attributes with larger numerical ranges.

For categorical variables, different encoding strategies were applied depending on the algorithm. Label Encoding was used for the Decision Tree model, as tree-based methods can directly handle ordinal relationships in categorical data. In contrast, One-Hot Encoding was employed for the non-tree-based algorithms, converting categorical variables into binary indicator variables. These transformations ensured that the models could process categorical attributes without introducing spurious ordering.

b) Training, Testing and Evaluation

The dataset was evaluated using Stratified K-Fold cross-validation with 5 folds, which ensured that the distribution of positive (heart disease) and negative (normal) cases was maintained across all splits. In each fold, four partitions of the dataset were used for training and one for testing, and the process was repeated until every fold had served as the test set. This approach reduced bias due to class imbalance and provided a more reliable assessment of model performance.

Four machine learning algorithms were trained and tested on the dataset: Decision Tree, Naïve Bayes, Logistic Regression, and K-Nearest Neighbour (KNN). Their performance was measured using the area under the ROC curve (AUC-ROC), along with precision, recall, and F1-score derived from each fold. The results across the five folds were averaged to obtain robust estimates of generalization ability. These metrics were chosen because they provide a balanced view of how well each model detects heart disease while minimizing false predictions, which is critical in healthcare applications.

The results revealed notable differences across the four algorithms shown in fig. 8 – fig. 10. The Decision Tree model achieved the lowest overall performance, with a precision and recall of 0.75 and an accuracy of 72%. This suggests that while the model was able to capture some patterns in the data, it was prone to both false positives and false negatives, reflecting its tendency to overfit without ensemble enhancement.

In contrast, both Naïve Bayes and Logistic Regression demonstrated stronger performance. Naïve Bayes recorded a precision of 0.86, recall of 0.83, and accuracy of 88%, showing that it was effective at identifying heart disease cases despite its simplicity. Logistic Regression performed slightly better, with a precision and recall of 0.88 and the same accuracy of 88%. Logistic Regression has the additional advantage of interpretability, making it useful for understanding the contribution of risk factors to heart disease prediction.

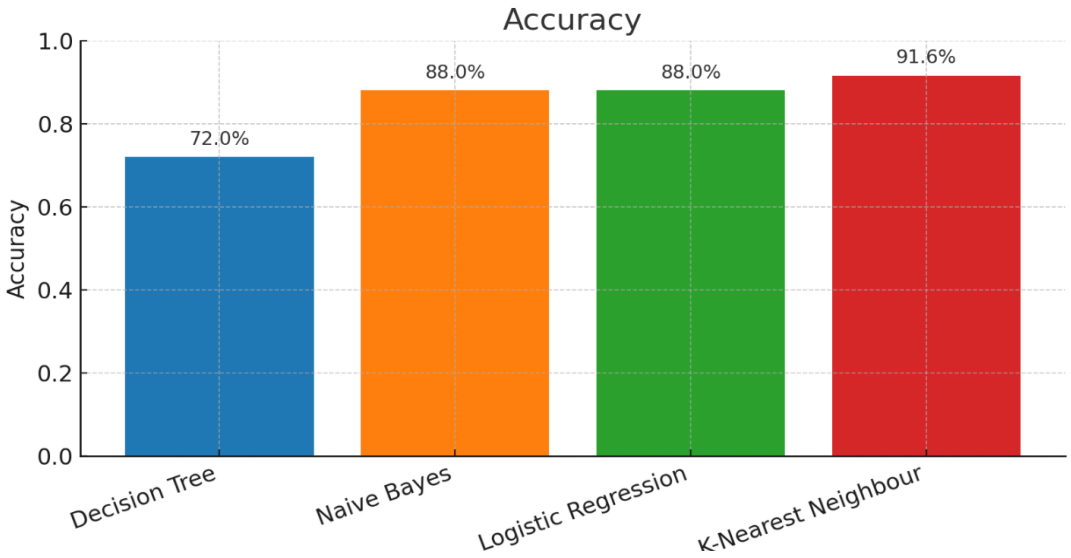

Fig. 8: Accuracy

Among all models, the K-Nearest Neighbour (KNN) algorithm produced the best results. It achieved a precision of 0.92, recall of 0.92, and accuracy of 91.6%, outperforming the others across all metrics. This indicates that KNN was highly effective in distinguishing between normal and heart disease cases, minimizing both false positives and false negatives. Its performance demonstrates its suitability for this dataset, although its computational cost may increase with larger datasets.

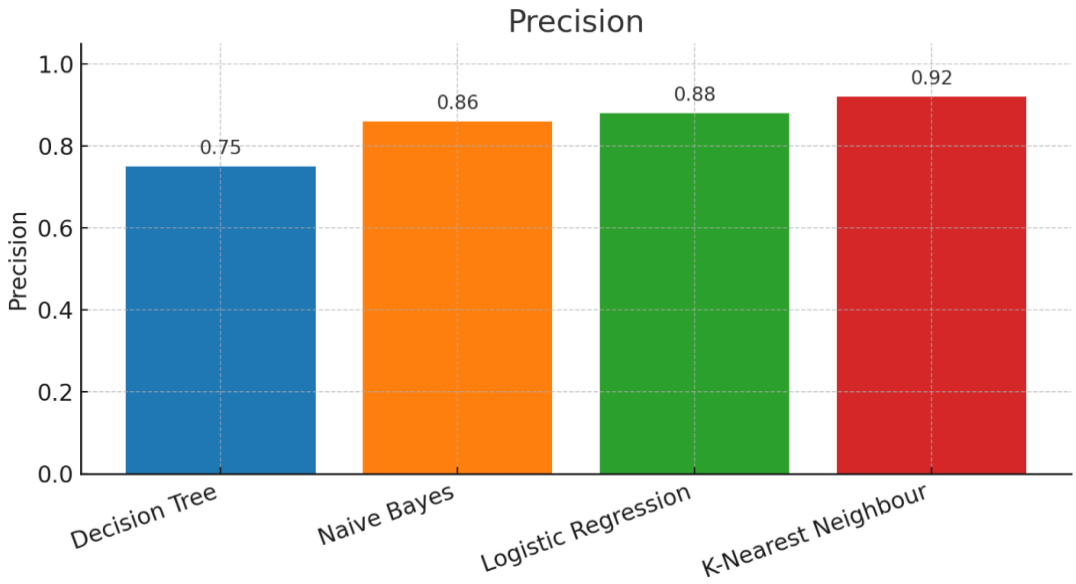

Fig. 9: Precision

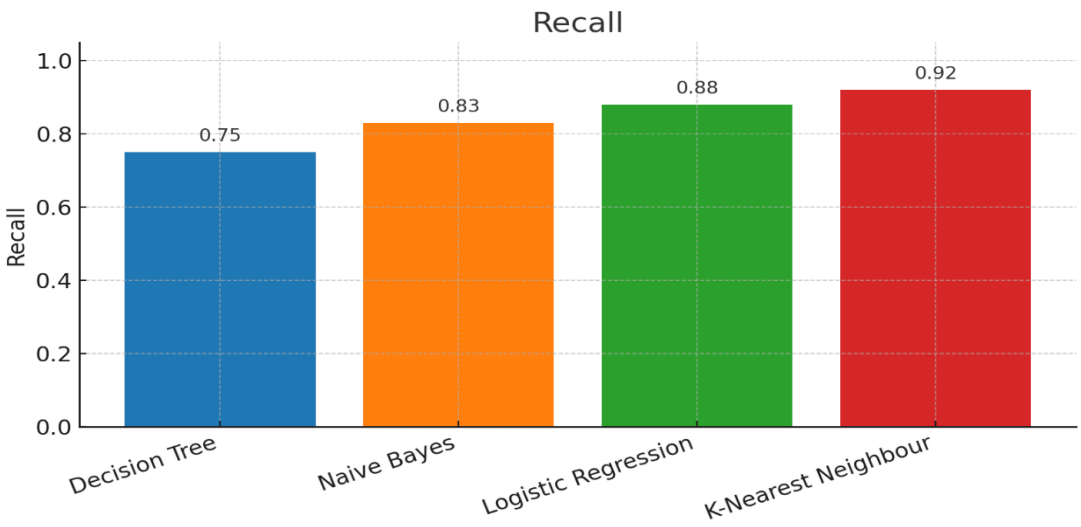

Fig. 10: Recall

Conclusion

This study applied four supervised machine learning algorithms - Logistic Regression, Naïve Bayes, Decision Tree, and K-Nearest Neighbour (KNN) to an integrated dataset of 920 patient records compiled from four well-known heart disease datasets. After preprocessing and stratified 5-fold cross-validation, the models were evaluated using accuracy, precision, and recall. The results showed that KNN achieved the highest performance (91.6% accuracy, precision and recall of 0.92), followed by Logistic Regression and Naïve Bayes (88% accuracy), while Decision Tree performed the weakest (72% accuracy).

The novelty of this work lies in combining multiple datasets into a single harmonized resource and systematically comparing classical algorithms under robust validation. The findings suggest that even relatively simple models, when trained on a diverse dataset, can achieve clinically meaningful performance in cardiovascular disease prediction

References

Ali, A., et al. (2025). A comprehensive review of deep learning-based models for heart disease prediction. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-024-10899-9

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

Bhardwaj, R., & Sharma, A. (2023). Application of K-nearest neighbor algorithm in medical diagnosis: A systematic review. *Health Information Science and Systems, 11*(3), 250–264. https://doi.org/10.1007/s13755-023-00267-w

Bose, S., Ray, S., & Ghosh, A. (2023). Naïve Bayes classifiers in healthcare: A systematic review of applications and challenges. *Health Information Science and Systems, 11*(2), 145–158. https://doi.org/10.1007/s13755-023-00252-3

Dey, D., Haque, M. S., Islam, M., Shammy, S. S., Mayen, S. A., Noor, S. T. A., … Uddin, J. (2025). The proper application of logistic regression model in complex survey data: A systematic review. *BMC Medical Research Methodology, 25*(15). https://doi.org/10.1186/s12874-024-02454-5

Dey, S., Dey, T., & Das, A. (2020). Prediction of heart disease using machine learning algorithms. *International Journal of Engineering Research & Technology, 9*(6), 1064–1068. https://doi.org/10.17577/IJERTV9IS060540

Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, *11*(1), 113.

Jiang, F., et al. (2023). Artificial intelligence and machine learning in clinical medicine. *New England Journal of Medicine, 388*(13), 1201–1208. https://doi.org/10.1056/NEJMra2302038

Kumar, P., & Singh, V. (2024). Machine learning techniques for disease diagnosis: Recent advances and applications. *BMC Medical Informatics and Decision Making, 24*(18), 211–225. https://doi.org/10.1186/s12911-024-02251-6

Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023). Ensemble learning for disease prediction: A review. In *Healthcare* (Vol. 11, No. 12, p. 1808). MDPI.

Mensah, G. A., Roth, G. A., & Fuster, V. (2019). The global burden of cardiovascular diseases and risk factors: 2020 and beyond. *Journal of the American College of Cardiology, 74*(20), 2529–2532. https://doi.org/10.1016/j.jacc.2019.10.009

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

Shinde, P. P., & Shah, S. (2018). A review of machine learning and deep learning applications. *Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 1–6. https://doi.org/10.1109/ICCUBEA.2018.8697857

Sivapalan, S., Wijesinghe, C. J., & Dissanayake, M. (2023). Decision tree algorithms for medical applications: A survey of recent developments. *Informatics in Medicine Unlocked, 41*, 101291. https://doi.org/10.1016/j.imu.2023.101291

Topol, E. (2023). Machine learning in medicine: what clinicians should know. *SMJ*

Virani, S. S., Alonso, A., Aparicio, H. J., Benjamin, E. J., Bittencourt, M. S., Callaway, C. W., … Tsao, C. W. (2021). Heart disease and stroke statistics—2021 update: A report from the American Heart Association. *Circulation, 143*(8), e254–e743. https://doi.org/10.1161/CIR.0000000000000950

World Health Organization. (2021). *Cardiovascular diseases (CVDs)*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Yu, D., & Wu, H. (2023). Variable importance evaluation with personalized odds ratio for machine learning model interpretability with applications to electronic health records-based mortality prediction. *Statistics in Medicine*, *42*(6), 761-780.

Zabor, E. C., Reddy, C. A., Tendulkar, R. D., & Patil, S. (2022). Logistic regression in clinical studies. *International Journal of Radiation Oncology* Biology* Physics*, *112*(2), 271-277.

Zhou, X., et al. (2024). A comprehensive review of deep learning-based models for heart disease prediction. *Springer*.