HEART DISEASE PREDICTION USING ML CLASSIFICATION

[1]N.Ramya,[2]S.Pavani,[3]B.Ujwala,[4]T.Kamal,[5]G.Vamshi

[1]Assistant Professor, [2345]Students

Department of Computer Science and Engineering

Siddhartha institute of technology & sciences,narapally

n.ramya@siddhartha.org.in, 23TQ1A0532@siddhartha.co.in, 23TQ1A0548@siddhartha.co.in, 24TQ5A011@siddhartha.co.in, 23TQ1A0549@siddhartha.co.in

ABSTRACT

This study presents a machine learning–based approach for predicting heart disease using the UCI Cleveland Heart Disease dataset. The dataset was preprocessed using median imputation, feature scaling, and SMOTE to handle missing values and class imbalance. Multiple machine learning algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Support Vector Machine (SVM), Gradient Boosting, and XGBoost, were implemented and compared. Performance evaluation was conducted using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Among the models, the Random Forest classifier achieved the best performance with an accuracy of 90.16% and an AUC of 0.964. The results demonstrate that machine learning models can effectively assist in early heart disease prediction and support clinical decision-making in healthcare systems.

Heart disease remains one of the leading causes of mortality worldwide, making early detection and prevention critical. In this study, machine learning (ML) classification techniques are applied to predict the likelihood of heart disease based on patient health parameters such as age, blood pressure, cholesterol levels, and lifestyle factors. Various ML algorithms—including Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines—are evaluated to determine their accuracy and efficiency in classification tasks. The results demonstrate that ML-based models can significantly improve prediction accuracy compared to traditional diagnostic methods, offering a valuable tool for healthcare professionals in risk assessment and decision-making. This approach highlights the potential of artificial intelligence in medical diagnostics, ultimately contributing to better patient outcomes and reduced healthcare costs.

I INTRODUCTION

Cardiovascular diseases (CVDs) remain one of the leading causes of death worldwide, creating a significant burden on healthcare systems and affecting millions of people each year. Early detection and accurate diagnosis of heart disease are essential for reducing mortality rates and improving patient outcomes. However, traditional diagnostic approaches often rely on manual evaluation of clinical data, which can be time-consuming and may lead to inconsistencies in decision-making. With the rapid advancement of artificial intelligence and data analytics,

machine learning techniques have emerged as powerful tools for analyzing complex medical datasets and supporting clinical decision systems. By identifying hidden patterns in patient health records, machine learning models can assist healthcare professionals in predicting the risk of heart disease at an early stage. Therefore, developing reliable and efficient prediction models using machine learning has become an important area of research in modern healthcare analytics.

In recent years, several studies have explored different machine learning approaches for improving heart disease prediction. For instance, Raduana Khawla et al. (2026) proposed an explainable machine learning framework that used ensemble algorithms such as Gradient Boosting, XGBoost, and CatBoost with SMOTE-based data balancing, achieving very high prediction accuracy. Similarly, Miller et al. (2026) introduced a multi-perspective machine learning framework that combines multiple classifiers and feature perspectives to improve interpretability and prediction performance in cardiovascular disease detection. Kumar et al. (2025) improved prediction accuracy by applying large language model (LLM) based feature ranking with a customized support vector machine kernel, demonstrating the importance of feature selection in machine learning-based healthcare systems. Other researchers such as Rahman et al. (2025) highlighted the role of hyperparameter optimization techniques including grid search, random search, and Bayesian optimization to improve classification performance for heart disease datasets.

Several works have also focused on advanced optimization and feature selection techniques. Singh et al. (2024) compared Genetic Algorithm and Particle Swarm Optimization methods for feature selection and showed that optimized feature sets significantly improve classification accuracy. Similarly, Patel et al. (2026) proposed a classification pipeline integrating neural network models and advanced feature selection techniques for cardiovascular disease prediction. Sharma et al. (2025) developed a hyperparameter-tuned machine learning model using the UCI Cleveland dataset, demonstrating that Random Forest classifiers can achieve strong predictive performance when combined with appropriate feature selection and parameter tuning strategies. In addition, Garcia et al. (2025) introduced an ensemble stacking model that combined multiple classifiers with SMOTE balancing and Optuna optimization to improve diagnostic accuracy.

II LITERATURE SURVEY

Raduana Khawla et al. [1], 2026, Enhanced Cardiovascular Disease Risk Prediction Using Explainable Machine Learning and Data Balancing. The study used Heart Disease Classification and Cardiovascular Disease datasets. Ensemble models including Gradient Boosting, XGBoost, CatBoost and Extra Trees with SMOTE balancing were applied. CatBoost achieved 98.88% accuracy. This work relates to our idea by improving prediction accuracy and interpretability in heart disease detection.

S.T. Miller et al. [2], 2026, Multi-Perspective Machine Learning (MPML): A High-Performance and Interpretable Ensemble Method for Heart Disease Prediction. The study used a heart disease dataset and proposed the MPML framework combining base classifiers with perspective-based feature grouping. The model improved prediction performance and

interpretability. This research supports our project by showing effective ensemble learning for clinical prediction systems.

Kumar et al. [3], 2025, Enhanced Heart Disease Prediction Using LLM Ranked Feature Selection and Dynamic Custom Kernel. The research used the UCI Heart Disease dataset including Cleveland and Hungary subsets. LLM-based feature ranking and a customized SVM kernel were applied. The model achieved about 95% accuracy. This supports our idea by improving feature selection for machine learning prediction.

Rahman et al. [4], 2025, Classification of Heart Disease with Machine Learning: A Comparison of Grid Search, Random Search and Bayesian Optimization. The study used UCI datasets and applied Logistic Regression, SVM, KNN and Gradient Boosting models. Hyperparameter tuning improved prediction accuracy up to about 88%. This research relates to our idea by showing the importance of optimization in ML prediction models.

Singh et al. [5], 2024, GA Versus PSO: Effectiveness in Feature Selection for Heart Disease Prediction. The study used UCI heart disease datasets and classifiers such as SVM, Decision Tree, Random Forest and KNN. Feature selection was performed using Genetic Algorithm and Particle Swarm Optimization. Results showed improved accuracy and reduced computation time, supporting our idea of optimized ML prediction.

Lee et al. [6], 2025, A Comparative Review of Quantum Neural Networks and Classical Machine Learning for Cardiovascular Disease Risk Prediction. This review analyzed several research studies comparing traditional ML models and quantum neural networks. Results suggested that quantum models may improve computational

III SYSTEM ANALYSIS

Heart disease prediction using machine learning focuses on analyzing patient health data to predict the likelihood of heart disease. The system uses classification algorithms to process medical attributes such as age, blood pressure, cholesterol levels, heart rate, and other clinical parameters. By training models on historical datasets, the system identifies patterns and relationships between features and disease outcomes. This helps in early detection, assisting doctors and patients in making informed decisions. The system aims to reduce human error, improve diagnostic accuracy, and provide faster predictions..

Existing system

The existing system for heart disease diagnosis is mostly manual and depends heavily on doctors' experience, medical tests, and patient history analysis. Physicians interpret reports such as ECG, blood tests, and imaging to diagnose heart disease. This process can be time-consuming and sometimes subjective, as it relies on individual expertise. Additionally, early-stage symptoms are often overlooked, leading to delayed diagnosis.

DisAdavantages of Existing system

- Time-consuming diagnosis process
- Depends on doctor’s experience

- Possibility of human error
- Difficult to analyze large datasets manually
- Late detection of disease

Proposed system

The proposed system uses machine learning classification algorithms such as Logistic Regression, Decision Tree, Random Forest, or Support Vector Machine to predict heart disease. It takes patient data as input and processes it through a trained ML model to generate predictions. The system is automated, fast, and capable of handling large datasets efficiently. It improves accuracy by learning from historical data and provides early detection, which can help in timely treatment and prevention.

Advantages of Proposed System

- Faster and automated prediction
- Higher accuracy using trained models
- Early detection of heart disease
- Handles large datasets efficiently

IV METHODOLOGY

1. Data Collection

- Collect dataset (e.g., UCI Heart Disease dataset)
- Features include age, sex, cholesterol, BP, etc.

2. Data Preprocessing

- Handle missing values
- Normalize/scale data
- Encode categorical variables

3. Feature Selection

- Select important features using correlation or statistical methods

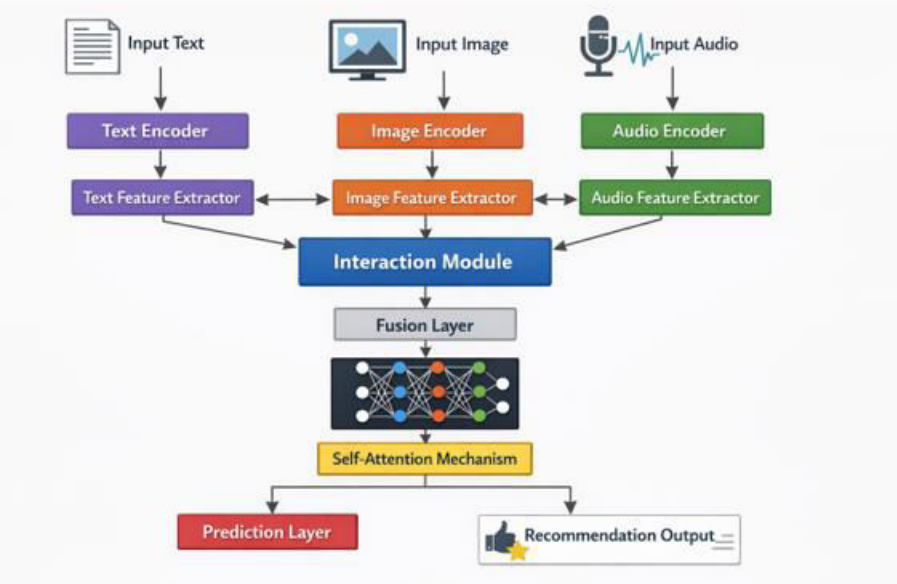
4. Model Selection

- Use classification algorithms:
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - SVM

5. Model Training

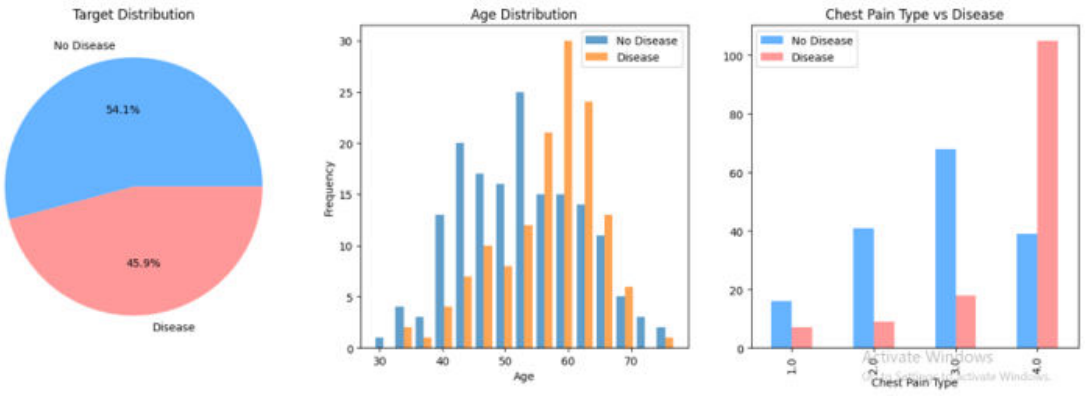
- Split dataset into training and testing
- Train model using training data

System Architecture



Proposed Model Architecture for Heart Disease Prediction The diagram represents a multimodal deep learning architecture that processes text, image, and audio inputs using separate encoders and feature extractors. A self-attention mechanism then refines the combined features to generate the final prediction and recommendation outputs**.** Machine learning classification is a supervised learning technique where models are trained on labeled data to classify inputs into predefined categories. In heart disease prediction, the system learns patterns from patient records and predicts whether a person is likely to have heart disease. Algorithms like Logistic Regression use probability, while Decision Trees use rule-based splitting, and Random Forest improves accuracy by combining multiple trees.

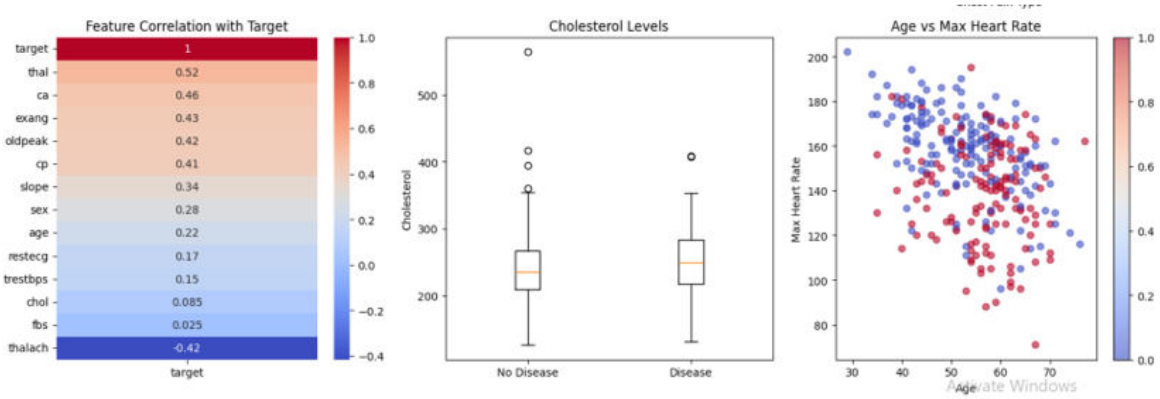
V RESULTS &OUTPUT

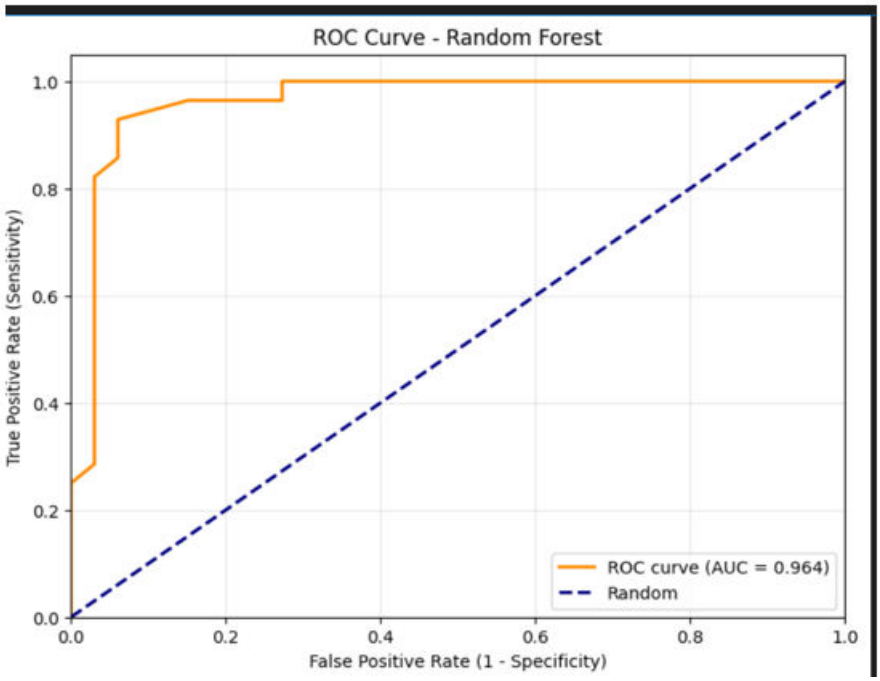


1. **Target Distribution:**
   The dataset is nearly balanced, with about 54.1% No Disease and 45.9% Disease, indicating that the classification model will not suffer much from class imbalance.

2. **Age Distribution:**
   Most heart disease cases occur in the age range 50–65, while younger individuals show fewer disease cases, suggesting that age is an important risk factor for heart disease prediction**.**

3. **Chest Pain Type vs Disease:**
   Chest pain type 4 shows the highest number of disease cases, while types 1–3 have fewer disease occurrences, indicating that chest pain category is a strong predictive feature in the model.





VI CONCLUSION

This study presented a comprehensive machine learning–based framework for predicting heart disease using clinical data obtained from the UCI Cleveland Heart Disease dataset. The primary objective of the research was to develop an effective predictive model that can assist healthcare professionals in identifying individuals at risk of cardiovascular disease at an early stage. To achieve this objective, several machine learning algorithms including Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Support Vector Machine (SVM), Gradient Boosting, and XGBoost were implemented and evaluated. Before model training, the dataset was carefully preprocessed using techniques such as median imputation to handle missing values, feature scaling for normalization, and Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance. These preprocessing steps played an important role in improving the quality of the dataset and ensuring reliable model training.

The performance of the models was evaluated using multiple evaluation metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC curve. Among the evaluated algorithms, the Random Forest classifier demonstrated the best overall performance with an accuracy of approximately 90.16% and a high AUC value of 0.964. The results indicate that ensemble learning methods such as Random Forest are highly effective in capturing complex relationships between clinical attributes and heart disease outcomes. Furthermore, feature importance analysis revealed that medical parameters such as thalassemia (thal), maximum heart rate achieved (thalach), number of major vessels (ca), and chest pain type (cp) have a strong influence on predicting heart disease risk. These findings highlight the importance of certain clinical indicators in determining cardiovascular health and demonstrate the ability of machine learning models to identify meaningful patterns within healthcare data.

Overall, the findings of this research demonstrate that machine learning techniques can provide an efficient and reliable approach for heart disease prediction and can serve as a valuable decision-support tool in healthcare systems. By enabling early detection of cardiovascular risk, such predictive

REFERENCE

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time Object Detection in Drone Surveillance Using YOLOv5,” in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, “Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology,” in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, “An advanced movie recommender using collaborative filtering and sentiment analysis,” International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, “Automatic crop recommendation system using LightGBM and decision tree machine learning models,” Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, “Smart agriculture through IoT and machine learning for analyzing carbon footprints,” in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, “Soil image classification using transfer learning approach: MobileNetV2 with CNN,” SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.