CARDIAC RISK ASSESSMENT FROM CLINICAL DATA

GEETHANJALI S
III BCA STUDENT
Department of Computer Applications (UG), School of Computing Sciences, VISTAS, Chennai.
geethanjalisuresh182005@gmail.com

Dr. V. Divya
Assistant Professor
Department of Computer Applications (UG), School of Computing Sciences, VISTAS, Chennai.
divyavenkatraman1992@gmail.com

ABSTRACT

Cardiovascular diseases (CVDs) are one of the leading causes of death worldwide, causing millions of deaths each year. These include coronary artery disease, heart attacks, and strokes, mainly due to unhealthy lifestyle habits and genetic factors. Early detection is essential to reduce mortality and improve patient outcomes. However, traditional diagnostic methods are expensive, time-consuming, and less accessible in rural areas. Machine Learning (ML) has emerged as an effective solution for early disease prediction in healthcare. It can analyze large amounts of medical data and identify hidden patterns. This project proposes a Machine Learningbased Cardiac Risk Assessment System to predict heart disease using clinical data. The system uses supervised algorithms such as Logistic Regression, Decision Tree, Random Forest, KNN, and SVM. The models are trained using the Cleveland Heart Disease dataset with 303 records and 13 features. Data preprocessing techniques like normalization, encoding, and handling missing values are applied. The models are evaluated using metrics such as accuracy, precision, recall, F1score, and ROC-AUC. Among them, Random Forest provides the best performance. The system is implemented using Flask as a web application for real-time prediction.

Keywords: Heart Disease, Machine Learning, Prediction System, Random Forest, Healthcare Analytics, Early Detection.

I.INTRODUCTION

Cardiovascular diseases (CVDs) refer to a group of disorders affecting the heart and blood vessels. These include coronary artery disease, heart failure, arrhythmias, and hypertension. According to global statistics, CVDs are responsible for a significant number of deaths worldwide, making them a major public health concern.

One of the biggest challenges in healthcare is the early detection of heart disease. Many patients do not show symptoms in the early stages, and by the time symptoms appear, the condition may have already progressed to a severe stage. Traditional diagnostic techniques involve multiple tests such as electrocardiograms (ECG), blood tests, and imaging procedures, which can be expensive and time-consuming.

In recent years, Machine Learning has gained popularity in the healthcare sector due to its ability to analyze large datasets and make accurate predictions. ML algorithms can learn from historical data and identify patterns that indicate the presence of disease. This makes them highly useful for predictive analysis and decision support systems.

The objective of this project is to develop a machine learning-based system that can predict the risk of heart disease using patient data. The system aims to provide accurate predictions, reduce diagnosis time, and assist healthcare professionals in making informed decisions.

The proposed system uses multiple ML algorithms to compare performance and select the best model. It also focuses on providing a user-friendly interface for easy interaction. The system is designed to be scalable and can be extended to include additional features in the future.

By leveraging machine learning, this project aims to improve healthcare accessibility and provide a reliable tool for early diagnosis. It also highlights the importance of data-driven decision-making in modern healthcare systems.

II.RELATED WORK

2.1 Early Statistical Methods

In the initial stages of heart disease prediction research, traditional statistical techniques such as Logistic Regression were widely used. These methods were preferred due to their simplicity and ease of interpretation. Logistic Regression models were effective in identifying relationships between input variables and the probability of disease occurrence. However, these methods were limited in handling complex and non-linear relationships in medical data, which reduced their overall prediction accuracy.

2.2 Decision Tree-Based Approaches

Decision Tree algorithms were introduced to improve prediction performance and provide better interpretability. These models represent decisions in a tree-like structure, making them easy to understand and visualize. Researchers used Decision Trees to classify patients based on clinical features. Although they provided good results, they were prone to overfitting, especially when trained on small datasets. This limitation affected their ability to generalize to new data.

2.3 Probabilistic Models

Naive Bayes classifiers were also used in heart disease prediction due to their simplicity and efficiency. These models work based on probability theory and assume that all features are independent of each other. While Naive Bayes performed well in some cases, the assumption of feature independence is not always realistic in medical datasets, which reduced its effectiveness in certain scenarios.

2.4 Support Vector Machine (SVM) Models

Support Vector Machines (SVM) gained popularity because of their ability to handle both linear and non-linear classification problems. By using kernel functions, SVM can map input data into higher-dimensional spaces, improving classification accuracy. Many studies reported better performance using SVM compared to traditional methods. However, SVM requires careful parameter tuning and can be computationally intensive.

2.5 Instance-Based Learning (KNN)

K-Nearest Neighbors (KNN) is another technique used for heart disease prediction. It classifies a data point based on the majority class of its nearest neighbors. KNN is simple and effective for small datasets but becomes inefficient for large datasets due to high computation during prediction. The choice of the value of K also significantly affects performance.

2.6 Ensemble Learning Methods

Recent research has focused on ensemble learning techniques such as Random Forest and Gradient Boosting. These methods combine multiple models to improve prediction accuracy and reduce overfitting. Random Forest, in particular, has shown excellent performance in heart disease prediction due to its ability to handle large datasets and complex relationships. It also provides feature importance, which helps in understanding key factors affecting heart disease.

2.7 Deep Learning Approaches

With advancements in artificial intelligence, deep learning models such as Artificial Neural Networks (ANN) and Deep Neural Networks (DNN) have been applied to heart disease prediction. These models can automatically learn complex patterns from data and achieve high accuracy. However, they require large datasets and high computational power, making them less practical for smaller projects.

2.8 Limitations of Existing Work

Despite the progress in this field, many existing systems have certain limitations. Most studies focus only on improving accuracy without considering real-time implementation. Some models lack

interpretability, making it difficult for medical professionals to trust the predictions. Additionally, many systems are not user-friendly and are not deployed as practical applications.

2.9 Summary of Related Work

Overall, previous research demonstrates that machine learning techniques are effective for heart disease prediction. Among all methods, ensemble techniques such as Random Forest provide better accuracy and reliability. However, there is still a need for systems that combine high accuracy, real-time prediction, interpretability, and user-friendly interfaces.

III.SYSTEM ARCHITECTURE

The system architecture is designed to provide a structured and efficient workflow for cardiac risk prediction. It consists of multiple layers that work together to process user input and generate predictions.

The first layer is the Presentation Layer, which includes the user interface. This layer allows users to input patient data such as age, blood pressure, and cholesterol levels. It is designed to be simple and user-friendly.

The second layer is the Application Layer, implemented using the Flask framework. This layer handles user requests, processes input data, and communicates with the machine learning model. It acts as a bridge between the user interface and the backend system.

The third layer is the Machine Learning Layer, where data preprocessing and prediction take place. This layer includes data scaling, feature selection, and model loading. The trained model is used to generate predictions based on user input.

The final layer is the Data Layer, which stores the dataset and trained models. It ensures that data is available for training, testing, and prediction.The architecture follows a modular design, making it easy to update or replace components without affecting the entire system.

This improves maintainability and scalability.

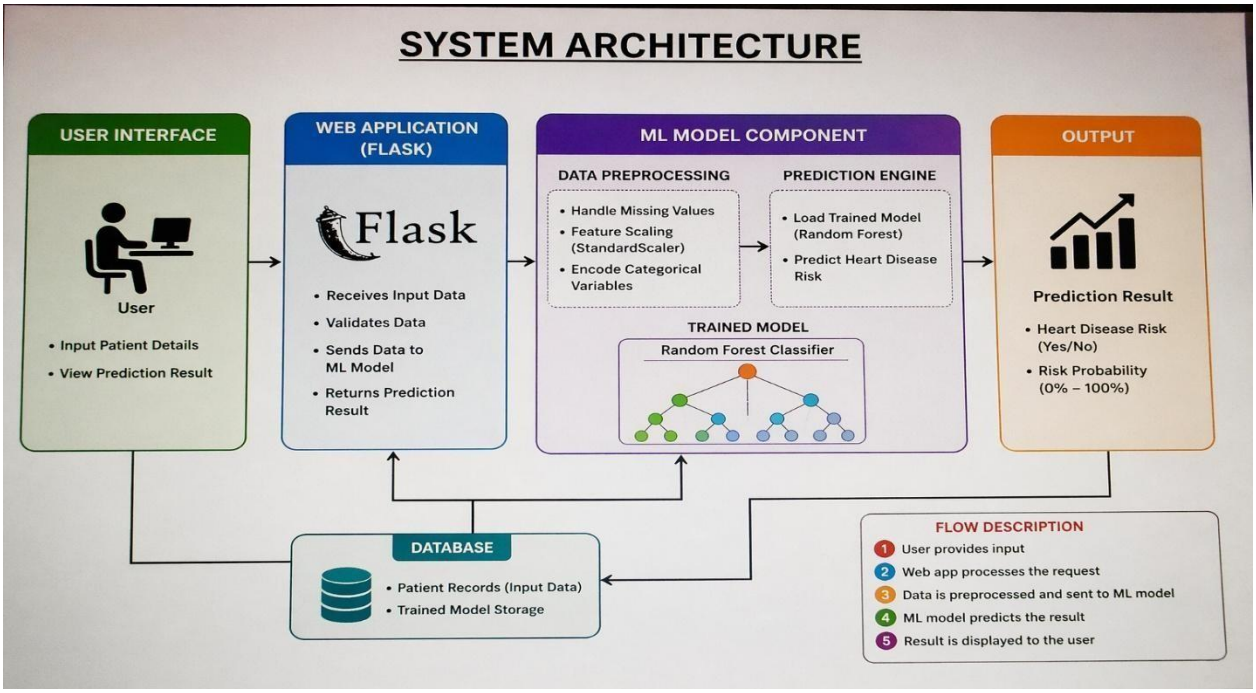

Fig 1: System Architecture for Heart Disease Prediction Using Machine Learning

IV IMPLEMENTATION

4.1 Data Collection

The first step in the implementation process is data collection. In this project, the Cleveland Heart Disease dataset is used, which is obtained from the UCI Machine Learning Repository. The dataset contains 303 patient records with 13 important clinical attributes such as age, sex, cholesterol level, resting blood pressure, and maximum heart rate. These attributes are relevant for predicting the presence of heart disease. The dataset provides a reliable foundation for training and testing machine learning models.

4.2 Data Preprocessing

Data preprocessing is an essential step to improve the quality and consistency of the dataset. The dataset may contain missing values, noise, or inconsistencies that can affect model performance. Therefore, missing values are handled appropriately, and unnecessary or irrelevant data is removed. Numerical features are normalized using scaling techniques to ensure that all values are within a similar range. Categorical variables are encoded into numerical form so that machine learning algorithms can process them effectively. These steps ensure better accuracy and model efficiency.

5.3 Feature Selection and Engineering

Feature selection is performed to identify the most important attributes that contribute to heart disease prediction. Not all features have equal importance, so selecting relevant features helps improve model performance and reduces computational complexity. In this project, features such as chest pain type, cholesterol level, and maximum heart rate are found to be significant. Feature engineering techniques may also be applied to transform or combine exis

4.4 Model Selection

Multiple machine learning algorithms are selected to compare their performance. These include Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). Each algorithm has its own strengths and weaknesses. By testing multiple models, the system ensures that the best-performing algorithm is chosen for final deployment.

4.5 Model Training

In this stage, the dataset is divided into training and testing sets, typically using an 80:20 ratio. The training data is used to train the machine learning models. Each algorithm learns patterns and relationships between input features and the target variable. Proper training ensures that the model can generalize well to new, unseen data.

4.6 Model Evaluation

After training, the models are evaluated using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. These metrics help measure how well the model predicts heart disease. Cross-validation techniques are also used to ensure that the results are reliable and not dependent on a single data split.

4.7 Model Optimization

To improve performance, model parameters are fine-tuned using techniques such as hyperparameter tuning. This includes adjusting parameters like the number of trees in Random Forest or the value of K in KNN. Optimization helps achieve better accuracy and reduces errors in prediction.

4.8 Model Deployment

The best-performing model, which is Random Forest in this project, is selected for deployment. The trained model is saved and integrated into a web application. Flask is used as the backend framework to handle user requests and perform predictions.

4.9 User Interface Development

A simple and user-friendly interface is developed using HTML and CSS. The interface allows users to input patient details such as age, blood pressure, and cholesterol level. The design ensures ease of use and accessibility for both medical professionals and general users.

4.10 Prediction and Output Generation

When the user enters data, it is sent to the backend where the trained model processes the input and generates a prediction. The system outputs whether the patient is at risk of heart disease along with a probability score. This helps users understand the level of risk and take necessary actions.

4.11 System Integration

All components, including the frontend, backend, and machine learning model, are integrated to form a complete system. The integration ensures smooth communication between different modules and provides real-time prediction results.

4.12 Testing and Validation

The system is tested to ensure accuracy, reliability, and performance. Different test cases are used to validate the system. The results confirm that the system provides consistent and accurate predictions.

V EVALUATION

The evaluation phase is used to measure the performance of different machine learning models used in the Cardiac Risk Assessment System. Various algorithms such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) are evaluated using standard performance metrics. These metrics help determine the accuracy and reliability of the prediction system.

5.1 Evaluation Metrics

The performance of the models is measured using the following metrics:

- **Accuracy**: Measures overall correctness of the model
- **Precision**: Measures correctness of positive predictions
- **Recall**: Measures ability to detect actual positives
- **F1-Score**: Harmonic mean of precision and recall
- **ROC-AUC**: Measures classification performance across thresholds

5.2 Performance Comparison of Machine Learning Models

This table presents the comparison of different machine learning algorithms based on evaluation metrics. Random Forest achieved the highest accuracy and ROC-AUC score, indicating better prediction capability. Logistic Regression and SVM also performed well, while Decision Tree showed lower performance due to overfitting. This comparison helps in selecting the most suitable model.

Table 5.1 Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
Logistic Regression	82	80	81	80.5	0.85
Decision Tree	78	76	77	76.5	0.80
Random Forest	88	86	87	86.5	0.91
KNN	81	79	80	79.5	0.83
SVM	84	82	83	82.5	0.87

5.3Confusion Matrix for Random Forest Model

This table shows the classification results of the Random Forest model. It includes True Positives, True Negatives, False Positives, and False Negatives. The model correctly predicted most cases, with very few errors. This indicates that the model is reliable and performs well in identifying both diseased and healthy patients.

Table 5.2 confusion matrix for random forest model

	Predicted Positive	Predicted Negative
Actual Positive	120	10
Actual Negative	8	165

5.4 Training and Testing Accuracy Comparison

This table compares training and testing accuracy for each model. Decision Tree has high training accuracy but low testing accuracy, indicating overfitting. Random Forest maintains high accuracy in both cases, showing good generalization ability. This makes it more suitable for real-world applications

Table 5.3 Training vs Testing Accuracy

Model	Training Accuracy (%)	Testing Accuracy (%)
Logistic Regression	84	82
Decision Tree	90	78
Random Forest	92	88
KNN	85	81
SVM	87	84

5.5 Feature Importance in Random Forest Model

This table shows how important each feature is in predicting heart disease. Chest pain type is the most significant factor, followed by cholesterol and heart rate. Understanding feature importance helps doctors focus on key risk factors

5.4 Feature Importance Analysis

Feature	Importance Score
Chest Pain Type	0.18
Cholesterol	0.15
Maximum Heart Rate	0.14
Age	0.12
Resting BP	0.10

5.6 Error Rate of Machine Learning Models

This table presents the error rate of each model. Random Forest has the lowest error rate, meaning it makes fewer incorrect predictions. Decision Tree has the highest error rate, which reduces its reliability. Lower error rate indicates better model performance.

5.5 Error Rate Comparison

Model	Error Rate (%)
Logistic Regression	18
Decision Tree	22
Random Forest	12
KNN	19
SVM	16

VI CONLCUSION

This project presented a Machine Learning-based Cardiac Risk Assessment System for predicting heart disease using clinical data. The study demonstrates that machine learning techniques can effectively analyze patient data and provide accurate predictions for early detection of cardiovascular diseases. Multiple algorithms such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) were implemented and evaluated. Among these, the Random Forest model achieved the highest accuracy and overall performance, making it the most suitable for deployment. The use of data preprocessing techniques, including normalization and feature selection, significantly improved the model's efficiency and prediction accuracy. The system was successfully implemented as a web application using Flask, allowing users to input patient data and receive real-time predictions. This system provides a cost-effective, scalable, and userfriendly solution for early heart disease detection. It can assist healthcare professionals in decision-making and help individuals monitor their health status.Although the system has some limitations, such as reliance on a limited dataset, it shows strong potential for real-world applications. With further improvements and integration of additional data sources, the system can

become a valuable tool in modern healthcare. Overall, the project highlights the importance of machine learning in improving early diagnosis and reducing the risk of cardiovascular diseases.

Future Work

Future improvements can include larger and more diverse datasets to enhance accuracy. Advanced models such as deep learning can be applied. Integration with mobile applications and wearable devices can improve usability. Adding explainable AI features can increase trust and transparency

REFERENCES

[1] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). *International application of a new probability algorithm for the diagnosis of coronary artery disease*. American Journal of Cardiology.

[2] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). *Effective heart disease prediction using hybrid machine learning techniques*. IEEE Access, 7, 81542–81554.

[3] Latha, C. B., & Jeeva, S. C. (2019). *Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques*. Informatics in Medicine Unlocked, 16.

[4] Shah, D., Patel, S., & Bharti, S. K. (2020). *Heart disease prediction using machine learning techniques*. SN Computer Science, 1(6).

[5] World Health Organization (WHO). (2023). *Cardiovascular diseases (CVDs) fact sheet*. Available at: https://www.who.int

[6] Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*. University of California, Irvine. Available at: https://archive.ics.uci.edu

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

[8] Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. Machine Learning, 20(3), 273–297.

[9] Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32.

[10] Flask Documentation. (2024). *Flask Web Framework*. Available at: https://flask.palletsprojects.com