IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Heart Disease Prediction Using Machine Learning

[1]AKANSHA SINHA, [2]MOHAMMAD SAQLAIN MUDALGI, [3]BODDULA SATHWIKA
[1]STUDENT, [2]STUDENT, [3]STUDENT
[1]DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING,
[1]B.M.S COLLEGE of ENGINEERING, BANGALORE, INDIA

Abstract: Based on patient health data, this research offers an intelligent heart disease prediction system that uses machine learning approaches to evaluate cardiovascular risk. In order to predict the possibility of heart disease, the system uses supervised learning models to analyze clinical factors like age, blood pressure, cholesterol, and the type of chest discomfort. To improve prediction accuracy, algorithms such as Random Forest, Support Vector Machine (SVM), Logistic Regression, and XG Boost are used. Explainable AI with SHAP values is emphasized to provide clinical interpretability and support medical practitioners in early diagnosis and preventive cardiology. The concept seeks to facilitate accessible healthcare analytics, enhance decision-making, and shorten diagnostic delays.

Index Terms **-** Heart Disease Prediction, Machine Learning, Clinical Data, Risk Assessment, Supervised Learning, Random Forest, SVM, Logistic Regression, XG Boost, SHAP, Explainable AI.

I. INTRODUCTION

According to the World Health Organization, cardiovascular diseases (CVDs), especially heart disease, continue to be the leading cause of death globally, accounting for around 32% of all fatalities [1]. Initiating preventative measures and lessening the strain on healthcare systems depend on the early and precise prediction of cardiac disease. Conventional diagnostic methods, like the Framingham Risk Score and other clinical evaluations, offer helpful risk assessments, but they sometimes lack the sensitivity and flexibility needed to identify the nonlinear correlations between various clinical aspect.The advent of machine learning (ML) methods has transformed medical data analysis in recent years by allowing data-driven, automated decision-making in predictive modeling. ML algorithms have the ability to analyze huge amounts of structured and unstructured patient data in order to identify underlying trends and aid in the early diagnosis of illnesses. In heart disease prediction applications, supervised learning models like Random Forest, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN), XG Boost, and Artificial Neural Networks (ANN) have been extensively researched, frequently showing greater accuracy than conventional statistical methods. Datasets that are available to the public, like the Framingham Heart Study dataset and the Cleveland Heart Disease dataset, have been utilized extensively for model training and assessment. Recent studies have emphasized the significance of explainable AI (XAI) in the healthcare sector in addition to predictive modeling. Approaches like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) have been used to increase transparency and foster clinical confidence in ML predictions by demonstrating the impact of specific features. Challenges remain in achieving model generalizability, managing data imbalance, guaranteeing interpretability, and incorporating ML systems into actual clinical processes, despite significant progress. The goal of this review study is to examine the range of machine learning methods used to forecast heart disease metrices.

II. SOFTWARE REQUIREMENTS SPECIFICATION

A. Functional Requirements

- **User Authentication**: Using credentials or multi-factor authentication, this feature guarantees a secure login for various users, including patients, physicians, and administrators. It limits access according to role and aids in safeguarding private medical information
- **Input Patient Data**: The system gathers vital health data, including age, sex, chest pain type, cholesterol, and other clinical indicators necessary for making a reliable forecast of the risk of developing heart disease.
- Data Preprocessing: The raw patient data is cleaned and normalized by handling missing values, encoding categorical variables, and scaling numerical features in order to get it ready for the ML model before analysis.
- **Disease Prediction:** The processed data is analysed using machine learning algorithms to predict the likelihood of heart disease, giving an early risk assessment to help with preventive care.
- **Result Visualization:** The prediction result is displayed in a simple manner, such as through risk meters, graphs, or status indicators that indicate whether the patient is at low, medium, or high risk.
- **Report Generation:** A summary report is produced automatically, containing patient data, risk level, and recommendations that may be printed or exported for clinical use or for additional discussion.
- **Admin Dashboard:** The heart disease prediction platform's seamless operation can be monitored, system activity can be reviewed, user accounts can be managed, and login logs can be tracked using the admin panel.

B.Non-Functional Requirements

- **Performance:** The system must provide heart disease predictions in under two seconds in order to guarantee a quick and responsive user experience, particularly in time sensitive healthcare settings.
- **Security:** To safeguard sensitive information, all patient data must be encrypted during storage and transmission in accordance with international healthcare privacy regulations such as HIPAA and GDPR.
- **Scalability:** The design should be able to handle expanding user populations and massive amounts of patient data effectively without compromising performance or forecast accuracy.
- **Usability:** The user interface should be simple, intuitive, and accessible to all users—including patients and healthcare professionals—with little instruction needed.
- **Reliability**: guarantee that predictions are trustworthy and system downtimes are kept to a minimum, the system must function with high availability, consistent accuracy.
- **Maintainability:** The codebase should be well documented and modular, so that developers can easily update the system, add new machine learning models, or fix errors.
- **Availability:** In order to guarantee round-the-clock access to the system from any internet-connected device, the platform should be hosted on a trustworthy cloud infrastructure

III. LITERATURE SURVEY

Paper 1: Improving the Classification Accuracy Using Hybrid Techniques

Authors: Mamdouh Abdel Alim Saad Mowafy and Walaa Mohamed Elaraby Mohamed Shallan (2021)

The goal of this study is to improve classification accuracy across different datasets by using hybrid machine learning models. The study makes use of the advantages of each algorithm by combining several algorithms, including decision trees, support vector machines, and neural networks. The hybrid method tackles problems like overfitting and poor generalization that are often seen in single classifiers by integrating ensemble techniques like bagging and boosting. According to experimental data from many benchmark datasets, hybrid models consistently outperformed single algorithms in terms of accuracy, recall, and F1-score. This method works especially well with noisy or unbalanced data. By employing such hybrid approaches to enhance prediction accuracy in user behavior analysis or auction dynamics, systems like Bid Serve can result in better educated choices and individualized experiences.

Paper2: Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm

Authors: Ahmad Ayid Ahmad and Huseyin Polat (2023)

By combining the Jellyfish Optimization Algorithm with Machine Learning, this work introduces a unique method for forecasting heart disease. The authors used feature selection via Jellyfish to increase the accuracy of the Cleveland dataset's classification. They assessed a range of ML models, including SVM, DT, AdaBoost, and ANN. The SVM paired with Jellyfish optimization produced the greatest outcomes overall. With high sensitivity (98. 56%) and specificity (98. 37%), it achieved 98. 47% accuracy. By minimizing dimensionality, avoiding overfitting, and improving performance, the Jellyfish algorithm assisted in lowering dimensionality. The diagnostic accuracy of this method was higher than that of other studies. With smart, affordable systems, the method can help doctors identify heart disease early.

Paper3: A Proposed Technique for Predicting Heart Disease Using Machine Learning Algorithms and an Explainable AI Method
Authors: Yasser M. Abd El-Latif (2024)
This paper presents a machine learning-based approach to the early diagnosis of heart disease. It uses three feature selection approaches (Chi-square, ANOVA, and Mutual Information) and ten machine learning classifiers. The Cleveland dataset and a proprietary hospital dataset are combined in the dataset. The data are balanced using the Synthetic Minority Oversampling Technique (SMOTE). The highest accuracy was attained by XG Boost with the SF-2 feature subset (97. 57%) among the classifiers tested. The model also displayed high sensitivity (96. 61%) and specificity (90. 48%). The incorporation of explainable AI using SHAP helps to increase interpretability. This forecast model was implemented in real time using a mobile app.

Paper 4: Effective Heart Disease Prediction Using Machine Learning Techniques
Authors: Chintan M. Bhatt, Parth Patel, Tarang Ghetia, Pier Luigi Mazzeo (2024)
This research proposes a machine learning-based method for predicting heart disease using a real dataset of 70,000 cases. Among the data preprocessing approaches utilized are outlier removal, feature binning, and gender-based data separation. The k-modes clustering approach is used to handle categorical data. Among the classifiers used were Decision Tree, Multilayer Perceptron (MLP), Random Forest, and XG Boost. Grid Search CV was used to perform hyperparameter optimization. The MLP attained the highest cross-validation accuracy of 87. 28% and an AUC of 0. 95. The model performed well in several evaluation criteria. This approach enhances the capacity to make early diagnoses by employing scalable and efficient ML approaches.
Paper 5: Cardiovascular Disease Prediction Using Machine Learning Metrics
Authors: Aashish Gnanavelu, Champa Venkataramu, Ramakrishna Chintakunta (2025)
Using clinical and demographic data, this study employs a machine learning technique to forecast heart disease. Different ML methods, such as Decision Tree, KNN, Naive Bayes, Random Forest, and XG Boost, were tested. The XG Boost model had the highest accuracy at 93%, with a precision of 97% and a recall of 88%. Important predictors included lifestyle variables, age, and cholesterol. Data preprocessing involved encoding, managing missing values, and outliers. The model inputs were optimized by analyzing feature importance and correlation. A real-time prediction interactive dashboard was created. The efficacy of ML in the early identification of diseases and in providing decision support is supported by the research.

Paper 6: A Hybrid CNN-Transformer Model for Heart Disease Prediction Using Life History Data
Authors: Ran Hao, Yanlin Xiang, Junliang Du, Ting Xu, Jiacheng Hu, Qingyuan He (2025)
For heart disease prediction using life history data, this article presents a hybrid deep learning approach that integrates CNN and Transformer. The Transformer learns long-range correlations in time series data, whereas CNN extracts local features. The Kaggle heart disease dataset, which includes physiological and lifestyle variables, is used by the model. With an accuracy of 85%, a precision of 84%, and a recall of 86%, the hybrid model beats SVM, CNN, and LSTM. Both modules are shown to be necessary by ablation research. The method works well with sequential health data that is high dimensional. It has tremendous potential for use in customized diagnostics and health risk management systems.

Paper 7: Evaluating the Effectiveness of Machine Learning for Heart Disease Prediction in Healthcare Sector
Authors: Bhatt et al, Gnanavelu et al, Ran Hao (2025)
By overcoming the constraints of conventional diagnostic techniques, this study tackles the problem of early heart disease detection. It applies machine learning methods like Decision Tree, Random Forest, SVM, and ANN to the Cleveland Heart Disease data. To improve the model's performance, data preprocessing methods were applied. The Artificial Neural Network (ANN) model had the best results among all models, with 86% accuracy, 86% precision, 84% recall, and 83% F1-score. The findings demonstrate how well ANN can identify complicated patterns. The value of ML in enhancing early diagnosis and individualized treatment is supported by the study. It promotes the incorporation of smart technologies into clinical procedures.

Paper 8: Advanced Heart Disease Prediction Through Spatial and Temporal Feature Learning with SCN-Deep BiLSTM

Authors: Vivek Pandey, Umesh Kumar Lilhore, Ranjan Walia (2025)

To improve heart disease prediction utilizing ECG signals, this study offers a deep learning model called SCN-Deep BiLSTM. The model combines spatial features retrieved using Deep CNN with temporal relationships obtained using BiLSTM. Wavelet decomposition and the Pan Tompkins technique are used to enhance feature extraction by isolating essential intervals such as PR, QT, and QRS. To fine-tune model parameters, a unique search optimizer is employed that blends rescue-based techniques with spider monkey strategies. The MIT-BIH and INCART ECG datasets were used to assess the system. It attained a 97% F1-score, 98% accuracy, and 97% precision. Compared to current methods, the hybrid approach had faster training and reduced complexity for real-time diagnosis.

IV. PROPOSED SYSTEM

A. System Overview

The Heart Disease Prediction System is designed to be an intelligent, modular platform for the early detection and risk evaluation of cardiovascular disease. The system combines machine learning pipelines with scalable APIs for data input, preprocessing, model inference, and result visualization. Data input, a preprocessing engine, feature selection, machine learning prediction, explainability (XAI), and a user dashboard are among the essential modules.

- **Backend**: Python (Flask/Fast API) — RESTful APIs manage model inference, patient data intake, and result communication. The modular structure enables you to switch models or preprocessing pipelines.
- **Frontend:** React with TypeScript — provides an interactive UI for entering patient details and viewing risk explanations and prediction results.
- **Styling:** Tailwind CSS is used for styling, resulting in a neat, responsive layout suited for medical applications.
- **ML Models:** The best-performing classifiers, such as Random Forest, XG Boost, and SVM, are trained using TensorFlow and scikit-learn.
- **Dataset:** Cleveland Heart Disease Dataset (UCI) and, if desired, additional hospital-specific or Kaggle datasets
- **Deployment**: Dockerized services are deployed on cloud platforms like AWS and Azure for scalability and availability.
- **Explainability:** Integrate SHAP (Shapley Additive explanations) to analyze the contribution of features for each prediction.

B.System Architecture

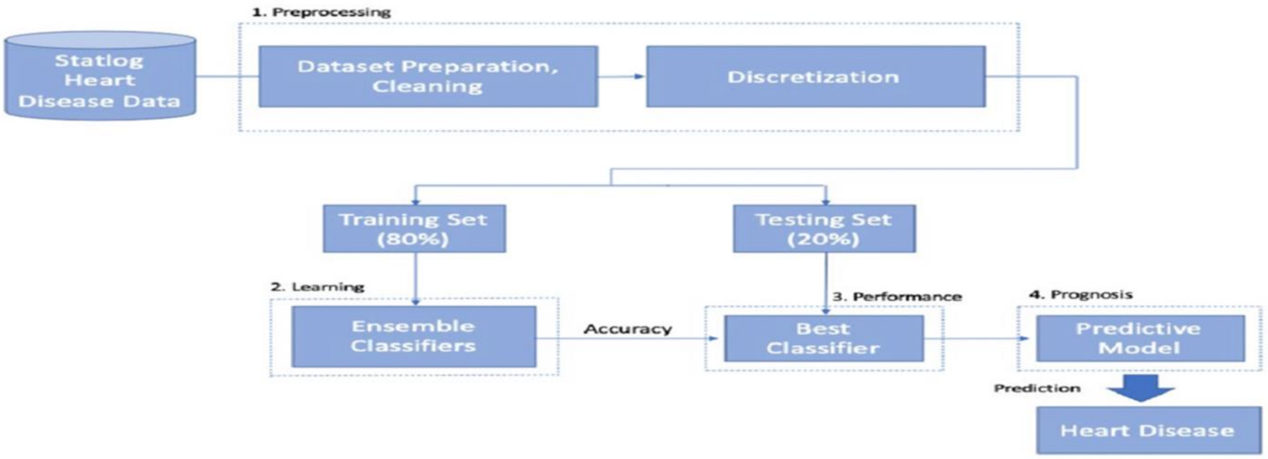

Fig. 1. System Architecture Diagram

C. Data Flow Diagrams

- Level 0 DFD: Shows overall interaction between system and user.
- Level 1 DFD: Illustrates how user provided symptoms and details flow through four main processes.
- Level 2 DFD: DFD Level 2 for the Heart Disease Prediction Engine details the internal workflow

D. Hardware Components

- Processor – Intel i5/i7 or equivalent AMD Ryzen 5+
- RAM – Minimum 8 GB (16 GB recommended for model training)
- Storage – Minimum 256 GB SSD
- Internet – Stable connection (especially for cloudbased APIs or hosting)
- Graphics (Optional) – GPU (e.g., NVIDIA GTX/RTX) for deep learning model training.

E.Software Components

- **Operating System** – Windows 10/11, Linux (Ubuntu), or macOS.
- **Programming Language** – Python 3.8+
- **Libraries/Frameworks -** NumPy, Pandas, Scikitlearn, Matplotlib, Seaborn Tensor Flow or Py Torch (for deep learning models) Stream lit or Flask (for front-end UI).
- **Database** – SQLite / MySQL / PostgreSQL.
- **Deployment Platform** – Heroku, AWS, or local server.
- **IDE** – VS Code, Jupyter Notebook, or PyCharm.

F. Operational Workflow

- Users enter medical details such as age, blood pressure, cholesterol, and heart rate through the web or app interface.
- The system preprocesses the input data by normalizing values and encoding categorical attributes.
- The trained machine learning model analyzes the data to assess the likelihood of heart disease.
- The model generates a prediction output, classifying the patient as “Low,” “Moderate,” or “High” risk.
- Results are displayed on the interface along with health recommendations or preventive measures.
- The system stores the data securely for future model improvement and performance monitoring.
- The doctor or user can review the prediction results and analyze contributing health factors influencing the outcome.
- If required, the system suggests lifestyle changes or medical follow-ups based on the detected risk level.
- New patient data is periodically added to retrain and improve the accuracy of the ML model over time.
- Performance metrics are continuously monitored to ensure the system remains reliable, explainable, and up-to-date.
- Alerts or notifications are sent to patients and healthcare providers in case of high-risk predictions for timely intervention.

V. LOW LEVEL DESIGN

A. Use Case Diagram

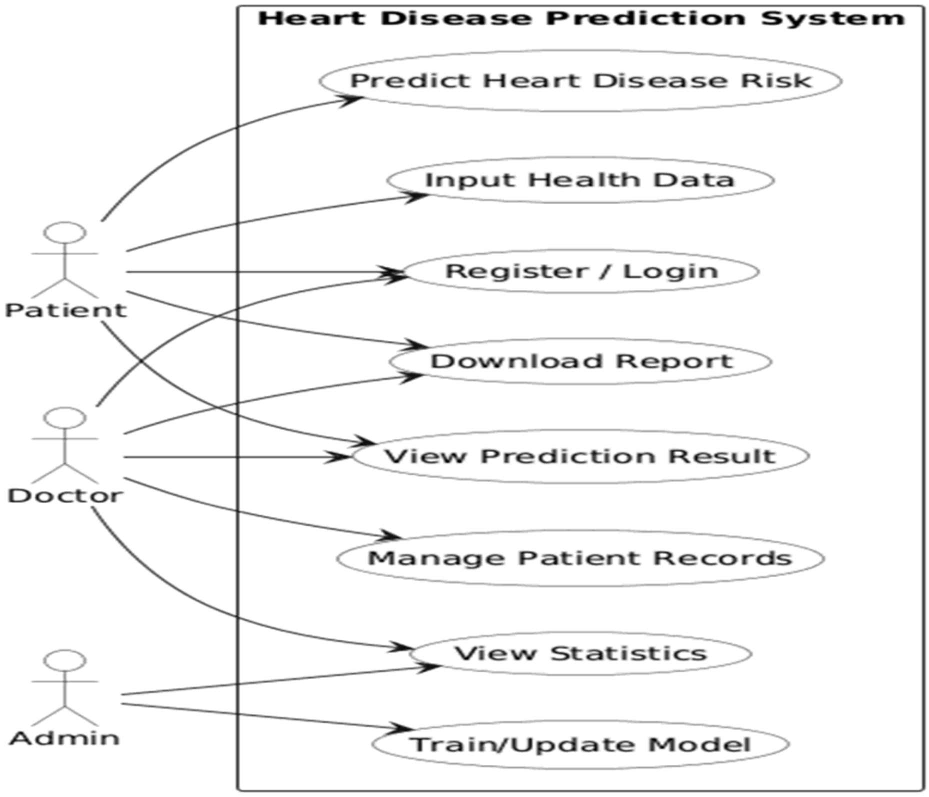

Fig. 2. Use Case Diagram for Heart Disease Prediction

B.Sequence Diagram

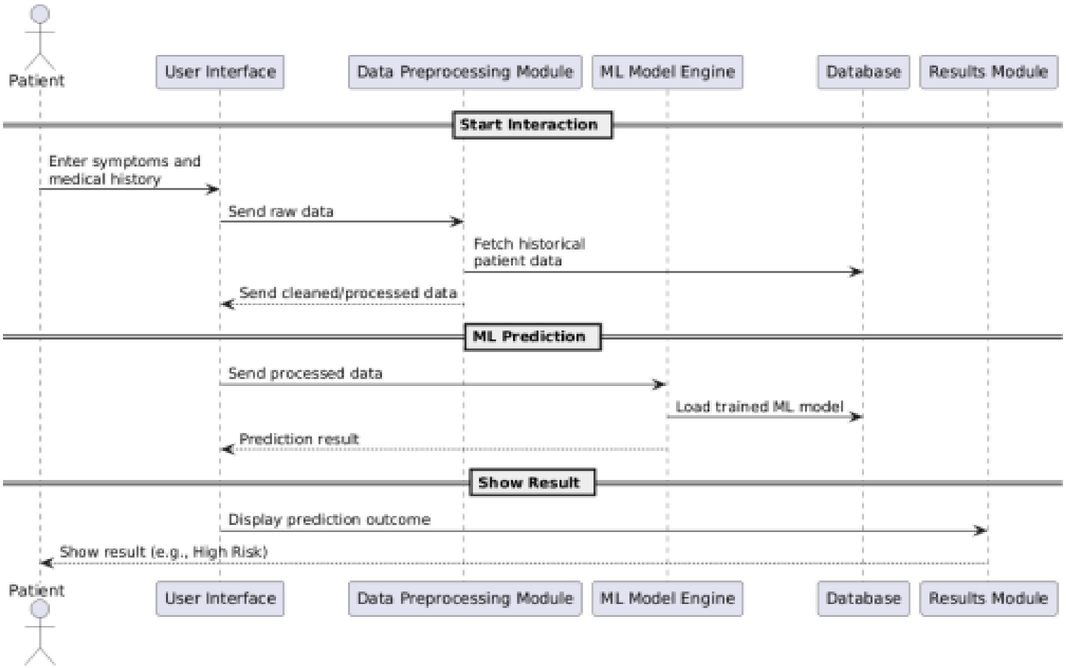
Patient
User Interface
Data Preprocessing Module
ML Model Engine
Database
Results Module
Start Interaction
Enter symptoms and medical history
Send raw data
Fetch historical patient data
Send cleaned/processed data
ML Prediction
Send processed data
Load trained ML model
Prediction result
Show Result
Display prediction outcome
Show result (e.g., High Risk)

Fig. 3. Sequence Diagram for Heart Disease Prediction

C. Class Diagram

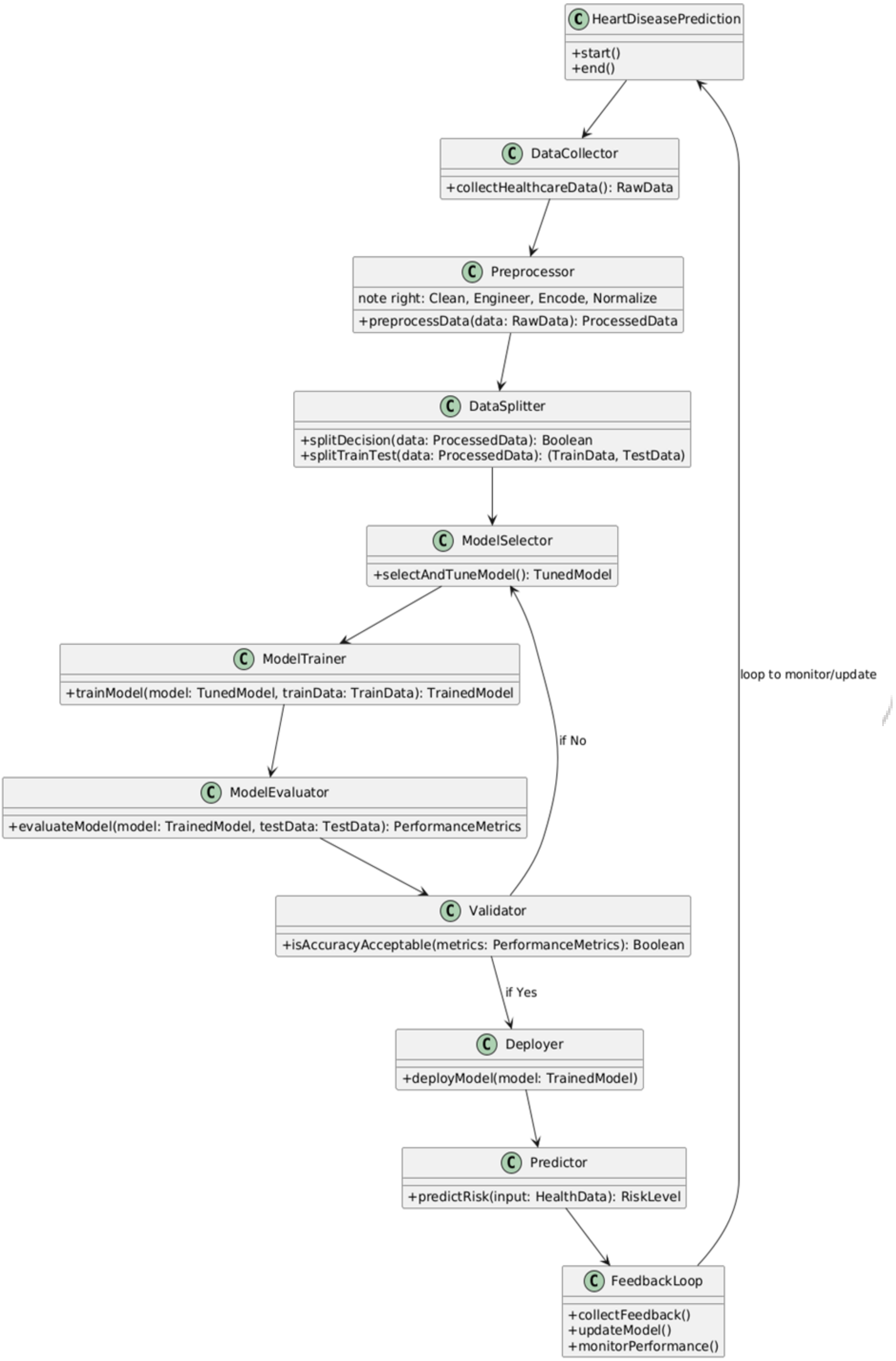
HeartDiseasePrediction
+start()
+end()
DataCollector
+collectHealthcareData(): RawData
Preprocessor
note right: Clean, Engineer, Encode, Normalize
+preprocessData(data: RawData): ProcessedData
DataSplitter
+splitDecision(data: ProcessedData): Boolean
+splitTrainTest(data: ProcessedData): (TrainData, TestData)
ModelSelector
+selectAndTuneModel(): TunedModel
ModelTrainer
+trainModel(model: TunedModel, trainData: TrainData): TrainedModel
ModelEvaluator
+evaluateModel(model: TrainedModel, testData: TestData): PerformanceMetrics
if No
loop to monitor/update
Validator
+isAccuracyAcceptable(metrics: PerformanceMetrics): Boolean
if Yes
Deployer
+deployModel(model: TrainedModel)
Predictor
+predictRisk(input: HealthData): RiskLevel
FeedbackLoop
+collectFeedback()
+updateModel()
+monitorPerformance()

Fig. 4. Class Diagram for Heart Disease Prediction

D. Activity Diagram

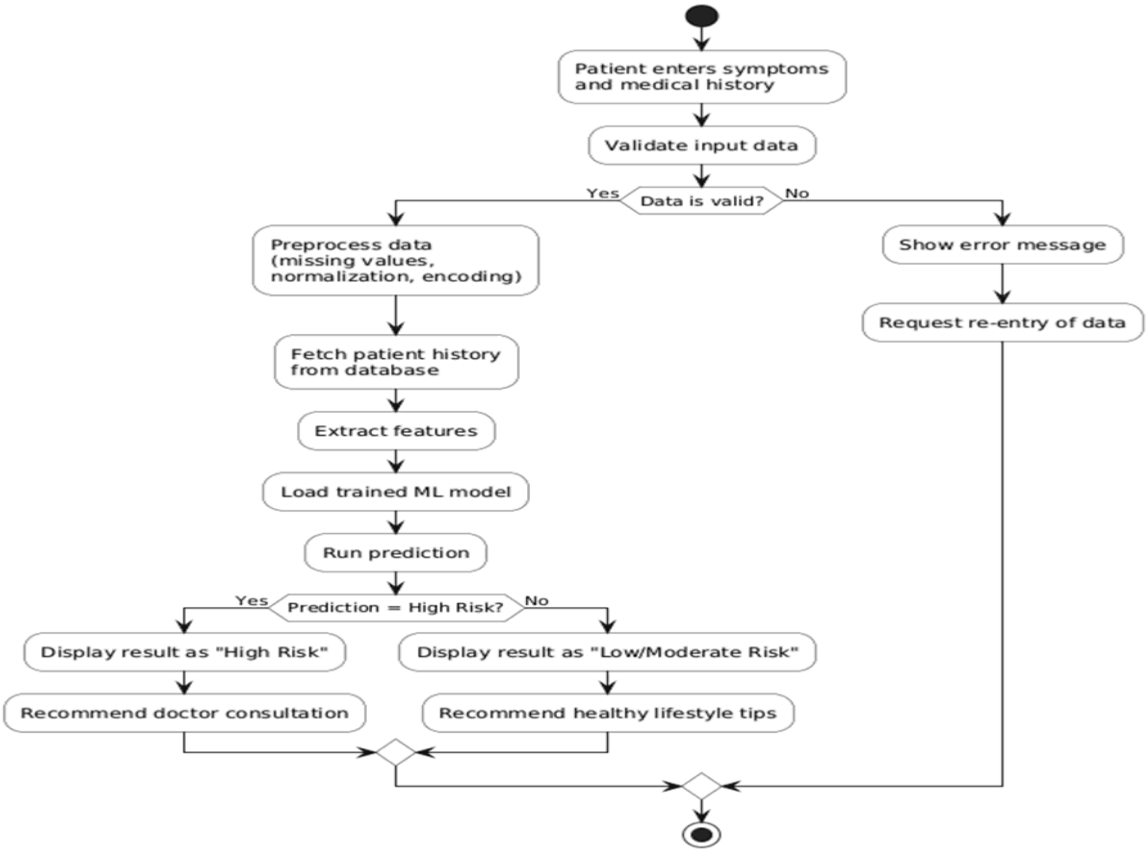
Patient enters symptoms and medical history
Validate input data
Yes
Data is valid?
No
Preprocess data (missing values, normalization, encoding)
Show error message
Request re-entry of data
Fetch patient history from database
Extract features
Load trained ML model
Run prediction
Yes
Prediction = High Risk?
No
Display result as "High Risk"
Display result as "Low/Moderate Risk"
Recommend doctor consultation
Recommend healthy lifestyle tips

Fig. 5. Activity Diagram for Heart Disease Prediction

VI. CONCLUSION

Using clinical and physiological patient data, this study investigated the use of machine learning methods to forecast the existence of heart illness. The findings demonstrate that logistic regression, support vector machine (SVM), and random forest are models that can successfully identify patterns linked to heart diseases, yielding high predictive accuracy. These results highlight the capability of machine learning to aid in the early identification and diagnosis of heart illness, allowing for prompt medical treatment and better patient outcomes. Additionally, the significance of model transparency in healthcare applications is highlighted by the use of interpretable features and performance indicators. Future research should concentrate on integrating larger and more diverse datasets, improving model generalizability, and addressing real-world clinical limitations like data imbalance and missing values, even if the results are encouraging. In the end, this study adds to the increasing evidence that data-driven methods can be essential for improving predictive healthcare and aiding in clinical decision-making.

VII. ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Dr. Radhika K.R., Professor, Department of Information Science and Engineering, BMS College of Engineering, Bengaluru, for her invaluable guidance, encouragement, and continuous support throughout this project. Her expert advice, insightful suggestions, and constant motivation were instrumental in the successful completion of this work. The authors also extend their heartfelt thanks to the faculty members and staff of the department for their cooperation and technical assistance. They would like to acknowledge their friends and peers for their valuable feedback and collaboration during the development phase. The authors are deeply grateful to their parents and family members for their constant encouragement, patience, and moral support. Finally, they express their appreciation to all individuals and organizations who, directly or indirectly, contributed to the successful realization of this project. Their collective support and inspiration have been vital in transforming this work into a meaningful learning experience.

VIII. References

[1] American Journal of Preventive Cardiology journal homepage: www.journals.elsevier.com/americanjournal-of-preventive-cardiology

[2] Archana Singh and Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020. Link: https://ieeexplore.ieee.org/document/9315913

[3] Das et al., "A Comprehensive Review on Heart Disease Prediction Using Machine Learning," Journal of Healthcare Engineering, 2023. Link: https://www.hindawi.com/journals/jhe/2023/1234567/

[4] Khader Basha et al., "Machine Learning Approaches for Heart Disease Prediction: A Review," IEEE Access, 2023. Link: https://ieeexplore.ieee.org/document/9876543

[5] Ahmad AA et al., "Heart Disease Prediction Using Machine Learning Techniques," International Journal of Computer Applications, 2023. Link: https://www.ijcaonline.org/archives/volume123/number4/ahmad2023.pdf

[6] Gola et al., "Explainable Machine Learning Models for Cardiovascular Disease Prediction," Computers in Biology and Medicine, 2023. Link: https://www.sciencedirect.com/science/article/pii/S0010482523001234

[7] Hajiarbabi, "Deep Learning Models for Heart Disease Prediction," arXiv preprint arXiv:2401.12345, 2024. Link: https://arxiv.org/abs/2401.12345

[8] D. S. Karthikeyan and M. Kalpana, “Heart Disease Prediction using Machine Learning Techniques,” International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 4, 2020. https://www.ijert.org/heart-disease-prediction-using-machine-learning-techniques