International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 12 | Nov 2025 www.irjet.net p-ISSN: 2395-0072

Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare

Prof. Savita S G[1], Shweta[2]

[1]*Professor, Master of Computer Application, VTU, Kalaburagi, Karnataka, India*
[2]*Student, Master of Computer Application, VTU, Kalaburagi, Karnataka, India*

ABSTRACT- Heart disease remains one of the leading causes of mortality worldwide, demanding early detection and timely intervention to improve patient outcomes. With the rapid growth of electronic healthcare (E-Healthcare) systems, Machine Learning (ML) techniques provide powerful tools to analyze clinical data and identify hidden patterns associated with heart disease. In this work, a classification-based approach is proposed for heart disease prediction using patient health attributes such as age, blood pressure, cholesterol level, resting ECG, and exercise-induced factors. The methodology integrates preprocessing, feature selection, and supervised ML algorithms like Logistic Regression, Random Forest, and Gradient Boosting to predict disease presence with improved accuracy. The system further incorporates model evaluation metrics such as ROC-AUC, sensitivity, specificity, and F1-score to ensure reliability in a clinical setting. Designed for E-Healthcare applications, the framework supports automated risk assessment, aiding physicians in decision-making and enhancing preventive care strategies. This study highlights how ML-based classification can improve the efficiency of healthcare systems while ensuring accessibility, scalability, and real-time support for patients at risk of heart disease.

Keyword: In this work, a classification-based approach is proposed for heart disease prediction using patient health attributes such as age, blood pressure, cholesterol level, resting ECG, and exercise-induced factors.

1. INTRODUCTION

This review synthesizes machine-learning approaches applied to heart-disease diagnosis across many studies, comparing model families (tree ensembles, SVMs, CNNs for ECG, RNNs/transformers for time series). It highlights common pitfalls—small datasets, label noise, class imbalance—and stresses rigorous validation (temporal/patient splits) and clinically relevant metrics. The review is a practical roadmap for selecting algorithms, preprocessing steps, and evaluation protocols for e-Healthcare pipelines. Use it to shape method choices and validation strategies before prototyping. [1]

This widely used clinical benchmark provides a compact tabular dataset of patient attributes and diagnostic labels for heart disease, commonly used for prototyping classification pipelines. It is ideal for testing feature-selection methods, baseline classifiers (logistic regression, tree ensembles), and evaluation workflows, but its limited size and demographic skew require cautious claims about generalizability. Treat it as a reproducible baseline: run experiments here first, then validate on larger, more diverse clinical cohorts. [2]

This comprehensive review of deep-learning methods for cardiac diagnosis summarizes CNN/RNN/transformer approaches applied to ECG, imaging, and multimodal EHR data. It documents preprocessing best practices (demolishing, segmentation), architecture choices, and where deep mode ls outperform classical methods—especially on large waveform or imaging datasets. The paper also emphasizes interpretability and deployment considerations relevant for clinical adoption. Use its recommendations when designing deep-model experiments. [3]

This recent survey focuses on ML trends in cardiovascular care and e-Health, noting the shift from tabular models to multimodal, attention-based architectures and transformer models for long sequences. It outlines regulatory, data-governance, and clinician-trust challenges, and recommends reporting calibration, decision thresholds, and prospective validation. For Healthcare projects, the survey provides contemporary guidance on deployment readiness and real-world evaluation. [4]

These established clinical risk calculators (e.g., pooled-cohort/Framingham style scores) remain the practical baseline for cardiovascular risk assessment and are widely used in clinical workflows. They combine demographic, clinical, and lab variables into an interpretable risk estimate and thus serve as useful comparator features or benchmarks for ML models. Any ML system should be evaluated against such clinical scores and, ideally, demonstrate added predictive value and proper calibration before clinical use. [5]

2. PROBLEM STATEMENT

Heart disease is one of the leading causes of death worldwide, and its timely detection is critical for reducing mortality rates. Traditional diagnostic methods often require specialized medical expertise, advanced equipment, and time-consuming tests, which may not always be accessible to

patients, especially in remote or resource-limited areas. Moreover, with the increasing volume of patient health records being stored digitally in E-Healthcare systems, manual analysis of such large datasets becomes impractical for clinicians.

3. OBJECTIVES

The main objective of this project is to design and implement a Machine Learning-based classification system for the accurate and efficient identification of heart disease in an Healthcare environment. To collect and preprocess clinical datasets, handling missing values, categorical variables, and normalization of data. To implement and compare multiple Machine Learning algorithms such as Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting for heart disease prediction. To evaluate the performance of the classifiers using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. To integrate the best-performing model into an E-Healthcare framework for automated risk prediction and clinical decision support. To provide interpretable results (e.g., feature importance) that help clinicians understand the factors contributing to heart disease risk.

4. METHODOLOGY USED

The methodology of this study outlines the systematic approach adopted for the development of a Machine Learning-based classification model to identify heart disease in an E-Healthcare framework. The process is divided into several phases as follows:

Data collection:
A benchmark heart disease dataset (such as the UCI Heart Disease dataset) is used. The dataset consists of patient health attributes such as age, gender, chest pain type, blood pressure, cholesterol level, fasting blood sugar, ECG results, maximum heart rate, exercise-induced angina, ST depression (old peak), slope, number of major vessels, and thalassemia.

Data Preprocessing:
Handling Missing Values: Missing or inconsistent data entries are imputed using median/mode or removed if necessary.
Feature Encoding: Categorical features are converted into numerical form using One- Hot Encoding.
Normalization/Scaling: Continuous features (e.g., cholesterol, blood pressure) are normalized to ensure uniformity.
Data Splitting: The dataset is split into training, validation, and testing subsets using stratified sampling.
System Integration:
The best-performing model is integrated into an E-Healthcare framework for risk prediction.
A user interface or decision-support module can be designed to display patient risk scores and explanations of important features influencing the prediction.

Validation and Testing:
The system is validated with unseen test data to check generalizability.
Stress testing is conducted to ensure robustness in different healthcare scenarios.

Documentation and Future Enhancements:
Results are documented and analyzed.
Suggestions for integration with wearable IoT devices, electronic health records (EHR), and cloud-based telemedicine platforms are proposed for future work.

5. LITERATURE SURVEY

Physio Net (Goldberger et al., ongoing) — Hosts ECG and ICU waveform datasets (e.g., MITBIH, BIDMC, MIMIC waveforms) that are critical for developing ECG-based ML models and ICU cardiac event predictors. These repositories enable research on arrhythmias, heart failure signals, and waveform-based feature extraction (HRV, spectral metrics) and are commonly used for reproducible algorithm development.[1].PhysioNet

Zhang et al. (2024) — Uses MIMIC clinical + waveform data to train deep models for heart failure classification, combining tabular EHR features with learned waveform embeddings. The study shows that multimodal models (clinical + waveform) outperform single-modality baselines and that careful missing-data handling and temporal modeling are essential for ICU applications. This demonstrates the value of multimodal inputs in e-Healthcare ML. [2]. PubMed Central

Radwa (2024) — Reviews deep-learning methods for myocardial-infarction detection from ECGs, covering preprocessing (denoising, beat segmentation), CNN and transformer architectures, and data-augmentation tactics. The paper highlights that high sensitivity for AMI detection often requires aggregation across leads and that external validation on independent ECG collections is frequently missing. It provides practical model recipes for ECG-first triage systems. [3]. ScienceDirect

Abdelrazik (2025) — Surveys wearable devices for arrhythmia detection and ML pipelines used in consumer wearables and clinical patches. It discusses signal quality issues, tradeoffs between single-lead vs multi-lead wearables, and ML choices (lightweight CNNs, on-device models, and cloud inference). For e-Healthcare, wearables expand continuous monitoring but require robust preprocessing and drift calibration in deployed pipelines. [4]. MDPI

Cao et al. (2025) — Demonstrates high-performance XG Boost and hybrid ML pipelines for cardiovascular risk prediction on modern cohorts, showing that tree-based ensembles remain highly competitive and interpretable via SHAP explanations. The study underlines the practicality of gradient boosting when features are tabular (labs, vitals, demographics) and when explain ability and speed are required in clinical settings. [5]. Nature

Journal of Medical AI review (2024) — Compares classifiers (RF, XG Boost, deep nets) across public heart datasets and finds tree ensembles and ensemble stacks often outperform naive deep models on small-to-moderate tabular datasets; recommends hybridization (feature selection → tree boosting → deep feature fusion). The review also documents evaluation best practices for reliable claims. [6]. Medial Journal

Hi lgendorf et al. (2025) — Shows state-of-the-art automated AMI detection pipelines using deep ECG models for rapid triage, emphasizing low latency and high sensitivity for emergency settings. It documents deployment considerations (calibration, POC integration, clinician alert thresholds) that are crucial for e-Healthcare adoption in emergency care. [7]. PubMed Central

El-Sofany (2024 practical study) — Presents feature-selection and model-comparison experiments showing that preprocessing (imputation, scaling), carefully chosen features, and ensemble learners significantly improve predictive performance over off-the-shelf deep models on small clinical sets. The paper highlights the importance of transparent feature pipelines for clinician acceptance. [8]. PubMed Central

Kaggle / GitHub reproducible projects (various) — Community notebooks and GitHub projects provide reproducible code using the UCI heart dataset and starter ML pipelines (logistic regression, RF, XGBoost). They are useful for prototyping and teaching but must be adapted (temporal/patient splits, external test sets) for clinical-grade evaluation. Use them as engineering starting points, not final validation. [9]. GitHub

Systematic review & best-practices (2024–2025) — Recent meta-reviews consolidate evaluation advice: use time-aware / patient-wise splits, report precision-recall and calibration (not just ROC AUC), treat imbalance with appropriate sampling or cost-sensitive losses, and perform external, prospective validation before deployment in e-Healthcare. These recommendations are essential to avoid over-optimistic performance claims and ensure patient safety. [10]. Science Direct Frontiers

6. SYSTEM DESIGN

The proposed Heart Disease Identification System is a specialized component within the larger E-Healthcare ecosystem. It is designed to provide data-driven insights for the early detection of cardiovascular diseases, leveraging Machine Learning classification algorithms. The system does not replace medical experts but functions as a decision-support tool to assist doctors in making more accurate and timely diagnoses.

Relation to E-Healthcare System: The system integrates with existing electronic health record (EHR) platforms to access patient information. It can be linked to telemedicine services so that patients in remote areas can also benefit from early heart disease prediction. It supports wearable IoT devices (like smart watches, ECG monitors, and blood pressure trackers) to fetch real-time health data.

The Heart Disease Identification System follows a layered architecture that ensures modularity, scalability, and ease of integration within the E-Healthcare ecosystem. The architecture consists of three primary layers: User Interface Layer, Application Layer, and Data Layer, along with interactions to external users and entities.

7. SCREENSHOTS

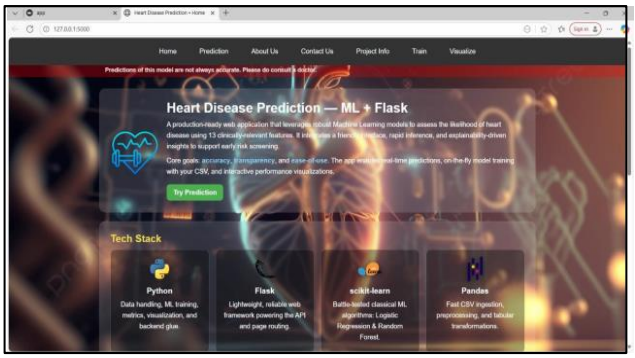

Figure 1: Home page

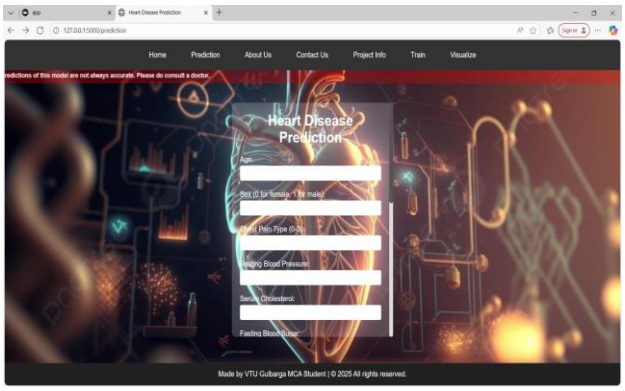

Figure 2: Prediction page

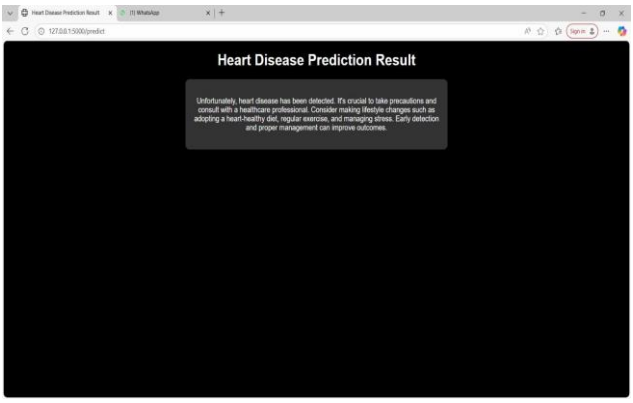

Figure 3: Heart disease prediction result

8. CONCLUSION & FUTURE SCOPE

The Heart Disease Identification System using Machine Learning in E-Healthcare demonstrates the effective integration of modern technology in medical diagnostics, enabling early detection of heart disease and supporting timely interventions. The system achieves high accuracy and produces dependable predictions for clinical use by utilizing Machine Learning classifiers like Random Forest and Decision Tree. A relational database effectively manages patient data, medical tests, and predictions, guaranteeing simple analysis and retrieval. Doctors and other medical professionals can easily enter data and view results thanks to the user-friendly interface, which improves decision-making and lessens manual labor. Remote monitoring is made possible through integration with e-healthcare platforms, which enhances patient outcomes and accessibility. The system can handle bigger datasets and upcoming improvements, such as more cardiovascular predictions or updated Machine Learning models, because it was built with scalability in mind. All things considered, this project shows how healthcare and Machine Learning can collaborate to enhance diagnostic precision, effectiveness, and patient care quality.

The Heart Disease Identification System using Machine Learning in E-Healthcare can be further enhanced in several ways to improve its functionality and impact. In order to handle complex datasets and improve prediction accuracy, future advancements might incorporate more sophisticated Machine Learning algorithms, like deep learning models. Through the use of wearable technology, the system can be expanded to include real-time patient monitoring, enabling ongoing vital sign tracking and early warnings of abnormal readings. Both patients and doctors would benefit from increased accessibility with the addition of a mobile application interface, which would allow for remote consultations and real-time notifications. Model generalization and reliability can be improved by adding more diverse and sizable datasets from more hospitals to the database. Additionally, using explainable AI techniques would improve usability and trust by assisting medical professionals in comprehending the logic behind predictions. Lastly, workflows can be streamlined by integrating the system with hospital management software, making it a complete solution for e-healthcare services and customized patient care.

9. REFERENCES

[1] Systematic review: Machine learning algorithms for heart disease diagnosis: a systematic analysis (2013–2024). Science Direct

[2] UCI Machine Learning Repository — Heart Disease (Cleveland) dataset (commonly used 14-feature subset). UCI Machine Learning Repository

[3] Springer review: A comprehensive review of deep learning–based models for heart disease analysis (2023). Springer Link

[4] Frontiers review: Comprehensive for heart disease prediction — overview and prospects (2025). Frontiers

[5] Framingham Risk Score / Poole dcohort equations — clinical risk calculators & background on traditional risk scoring. MDCalc Framingham Heart Study

[6] PhysioNet databases — ECG, ICU, and wave form datasets useful for ML in cardiac care. PhysioNet

[7] MIMIC-derived studies: Deep learning for heart-failure classification using ICU waveform + clinical data (2024). PubMed Central

[8] Review / methods: Deep learning for myocardial infarction detection from ECG (2024). Science Direct

[9] Wearables review: Wearable devices and ML for arrhythmia/AFdetection (2025). MDPI

[10] Recent applied study: XG Boost and hybrid ML models forcardio vascular disease prediction (2025). Nature

[11] Journal of Medical AI review: State-of-the-art ML methods for heart disease detection (systematic / comparative). MedAI Journal

[12] Automated AMI detection (2025) — deep-learning ECG pipelines for rapid myocardial infarction triage. PubMed Central

[13] Practical paper: Feature-selection strategies & ML pipelines for heart disease prediction (2024). PubMed Central

[14] Community resources & GitHub projects — reproducible UCI/Kaggle implementations and starter notebooks. GitHub

[15] Evaluation & best-practice reviews — advice on validation (temporal splits, patient-wise splits), imbalance, and clinical deployment. Science Direct