IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 6, February 2025

Impact Factor: 7.67

# Optimizing Machine Learning Models for Heart Disease Prediction Using UCI Cleveland Dataset

**Arjun Patil[1] and Vaishnavi S Gharat[2]**

Assistant Professor and Head Department of IT[1]
Student P.G. Department of IT[2]
Veer Wajekar ASC College, Phunde, Uran

**Abstract:** *Heart disease remains one of the most critical health issues globally, with acute myocardial infarction (AMI) posing serious threats to human life. Early diagnosis can greatly enhance treatment outcomes and prevent fatalities. In this study, we explore the application of machine learning (ML) models to predict heart disease using the UCI Cleveland dataset. Four classifiers—Logistic Regression, Naïve Bayes, Support Vector Machines (SVM), and XGBoost—were evaluated. The work emphasizes preprocessing techniques to mitigate overfitting and ensure data quality. Among the evaluated models, XGBoost demonstrated superior performance with a 92% accuracy and an AUC score of 0.94. This research highlights the significance of ML in healthcare diagnostics and suggests directions for future improvements through feature enhancement and hybrid approaches.*

**Keywords**: *Heart disease*

## I. INTRODUCTION

Cardiovascular diseases (CVDs) remain the leading cause of death globally, with acute myocardial infarction (AMI), or heart attack, being one of the most prevalent and life-threatening forms. AMI results from an interruption in the blood supply to a portion of the heart muscle, causing irreversible tissue damage if not treated promptly. The causes of such blockages often include the accumulation of fatty deposits in coronary arteries or the sudden formation of clots, leading to ischemia.[1-2]

The early diagnosis of AMI is critical for effective treatment and improved patient outcomes. However, traditional diagnostic methods—based on patient history, clinical symptoms, and a few biomarkers—can be subjective and may not account for overlapping risk factors such as diabetes, hypertension, and obesity. These complexities necessitate the integration of advanced computational tools for more accurate and timely prediction.[3-5]

Machine learning (ML), a branch of artificial intelligence, offers powerful tools to discover patterns and relationships within complex medical datasets. By training models on labeled datasets, ML algorithms can learn to identify the risk of heart disease in new, unseen patient data. This predictive capability is especially valuable for decision support systems in clinical environments.

This study employs ML techniques to develop predictive models using the UCI Cleveland Heart Disease dataset—a widely used dataset in cardiac diagnostics research. It focuses on 13 critical numeric features that have been empirically shown to be significant in predicting heart disease. The models investigated include Logistic Regression, Naïve Bayes, Support Vector Machines (SVM), and XGBoost. In addition to model performance, this work emphasizes proper data preprocessing and highlights the importance of mitigating overfitting to ensure model reliability. The ultimate goal is to aid healthcare professionals in early diagnosis, leading to timely intervention and reduced mortality.

## II. MACHINE LEARNING RESEARCH METHODS

### Research Framework and Approach

This research adopted an **active learning** approach, where models iteratively improved through data-driven insights and targeted feedback. Active learning is particularly useful in healthcare settings where labeled data can be expensive or limited, allowing the model to prioritize learning from the most informative data points.

### Role of Electronic Health Records (EHR)

The digitization of medical records in the form of Electronic Health Records (EHR) has greatly increased the volume and accessibility of healthcare data. While EHRs offer rich datasets, they also come with challenges such as data redundancy, heterogeneity, and unstructured formats. This study addressed these issues by focusing on a structured dataset (the UCI Cleveland dataset) and applying rigorous preprocessing steps including normalization, handling of missing values, and outlier removal.

### Modeling Strategy

The research follows a modular pipeline for ML model development, which includes:

- **Data Acquisition**: The dataset was sourced from the UCI Machine Learning Repository.
- **Preprocessing**: Techniques such as imputation for missing values, normalization, and outlier detection were applied.
- **Exploratory Data Analysis (EDA)**: Visualization tools (e.g., heatmaps and boxplots) were used to identify feature correlations and distribution anomalies.
- **Model Training and Evaluation**: Multiple ML models were implemented and evaluated using cross-validation and performance metrics like accuracy, AUC, precision, and recall.

### Highlights of the ML Approach

- **Improved Predictive Accuracy**: Machine learning models were shown to outperform traditional rule-based approaches.
- **Interpretability and Clinical Relevance**: Diagnostic features such as chest pain type, cholesterol levels, and ECG results were emphasized to maintain clinical interpretability.
- **Augmented Clinical Decision-Making**: Integrating ML with cardiovascular treatment plans enhances the decision-making capabilities of healthcare professionals **Literature Survey**

## III. REVIEW SCOPE AND METHODOLOGY

To ground this research in existing literature, a comprehensive literature review spanning the last two decades (2000–2021) was conducted. Databases including **Scopus**, **Web of Science**, and **ScienceDirect** were used for keyword-based searches. Key phrases included: "machine learning for heart disease prediction" "cardiac risk assessment with AI" "optimization of heart disease models"deep learning in cardiovascular diagnostics" A total of 50 papers were retrieved initially. After applying exclusion criteria (removal of duplicate entries, outdated models, or papers lacking robust evaluation), 27 papers were selected for in-depth review.

### Inclusion and Exclusion Criteria

**Inclusion Criteria:**

Research conducted in the past 10 years Use of modern and widely accepted machine learning or deep learning algorithms. Presence of evaluation metrics such as accuracy, precision, recall, and AUC. Clinical relevance of datasets or use of real-world health records.

**Exclusion Criteria:** Studies using outdated algorithms without comparison to modern techniques. Lack of reproducible methodology or code. Articles with vague or missing performance evaluations.

### Key Findings from Literature

The reviewed studies presented diverse approaches and insights:

- **Artificial Neural Networks (ANN)** consistently outperformed traditional models like Decision Trees or KNN in several studies.

- **Principal Component Analysis (PCA)** and **Fast Independent Component Analysis (FastICA)** were effectively used for dimensionality reduction. FastICA achieved an impressive **F1 score of 99.83%** in arrhythmia detection.
- **Hybrid Models**, combining feature selection techniques and ensemble learning (e.g., Random Forest + SVM), offered improved generalizability.

Use of **EHR data** and real-time monitoring tools further enhanced prediction accuracy and applicability in clinical settings One influential study by Fatima and Pasha (2017) provided a thorough survey of ML algorithms applied in disease diagnosis, emphasizing the importance of feature selection and balanced datasets. Another by Singh et al. (2018) highlighted the benefits of reduced feature sets in improving classifier performance, particularly with Extreme Learning Machines.

**Search Strategy:** Research was conducted for the year 2021 using keywords like “machine learning based heart disease prediction,” and “optimization of health disease prediction.” Inclusion criteria involved relevance, modern algorithm use, and domain challenges. Exclusion criteria included outdated or low-evaluation parameter studies. Previous studies have used neural networks, PCA, GDA, and sequential modeling with EHR data. ANN outperformed traditional classifiers, and FastICA showed a 99.83% F1 score for arrhythmia classification.

**Methods:**

- **Study Design:** The methodology consists of six main steps:
- **Data Acquisition:** Data collected from UCI ML repository (Cleveland subset) with 14 attributes.
- **Preprocessing:** Missing values replaced with mean, outliers detected via boxplots, and duplicates removed using dictionary functions. **Integration:** Data modules and libraries were integrated using Python 3.8.3.
- **Exploratory Data Analysis (EDA):** Relationships analyzed via heatmaps and boxplots.**Literature Review:** Previous effective methods reviewed.
- **Model Application:** ML models (SVM, Naïve Bayes, Logistic Regression, XGBoost) applied using Scikit-learn and XGBoost libraries.
- **Data Collection:** The dataset includes data from four locations (Hungary, Switzerland, Cleveland, Long Beach) with 76 attributes. A subset of 14 attributes was used based on previous research relevance.

**ML Models:**
**Logistic Regression:** Supervised model for binary classification using probabilistic outputs. **Naïve Bayes:** Based on Bayes theorem, performs well with high-dimensional data.n**Support Vector Machines:** Maximizes hyperplane margins for classification. **XGBoost:** Ensemble boosting technique combining weak classifiers sequentiall
**Discussion:** Findings validate that smaller datasets are sufficient for ML applications. Normalization and overfitting were addressed. XGBoost achieved the highest accuracy (92%) and AUC (0.94). The results support its robustness for small medical datasets. Future improvements could integrate patient-friendly multimedia platforms for result dissemination. In addition, ensemble and hybrid models can further enhance prediction reliability.

## IV. CONCLUSION

This study compared four ML classifiers on the Cleveland dataset for heart disease prediction. XGBoost outperformed others, achieving highest accuracy and AUC. Future research aims to explore deep learning methods, utilize the full 76 attributes, and integrate datasets for broader applicability. Feature selection will be emphasized for better prediction accuracy.

## REFERENCES

[1]. Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. J. Intell. Learn. Syst. Appl. 2017; 09:1–16.

[2]. Singh RS, Saini BS, Sunkaria RK. Detection of coronary artery disease by reduced features and extreme learning machine. Med. Pharm. Rep. 2018; 91(2):166–175.

[3]. Yaghouby F, Ayatollahi A, Soleimani R. Classification of cardiac abnormalities using reduced features of HRV signal. World Appl. Sci. J. 2009; 6(11):1547–1554.

[4]. Asl BM, Setarehdan SK, Mohebbi M. SVM-based arrhythmia classification using reduced HRV features. Artif. Intell. Med. 2008; 44(1):51–64.

[5]. Zhang D, et al. Integrating feature selection and extraction with deep learning for breast cancer outcomes. IEEE Access. 2018; 6:28936–28944