Heart Disease Prediction Using Machine Learning Techniques

D S Anjali[1], Rajeev Adrija[2]

[1] *am.sc.uaie23020@am.students.amrita.edu, Amrita Vishwa Vidyapeedam, Amritapuri*
[2] *am.sc.u4aie23009@am.students.amrita.edu, Amrita Vishwa Vidyapeedam, Amritapuri*

Abstract

Heart disease is a leading cause of death globally. Early diagnosis and treatment are crucial in preventing fatal outcomes. Our project applies various machine learning algorithms to predict the presence of heart disease using real-world clinical datasets. We use binary classification (disease vs. no disease) and compare multiple models including K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), Naive Bayes, Decision Tree, Random Forest, and XGBoost. The models are evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. XGBoost consistently outperforms other models across datasets, achieving high accuracy and AUC. Our system demonstrates the potential of ML to enhance medical diagnostics.

Keywords: Heart Disease; Machine Learning; Classification; XGBoost; Random Forest; SVM; ROC-AUC; Diagnosis

1. Introduction

Our project focuses on predicting the presence of heart disease using machine learning algorithms applied to clinical datasets. Heart disease is a major cause of death globally. Early detection can save lives. Traditional detection methods can be costly and time-consuming. The aim is to predict whether a patient has heart disease using ML.

Nomenclature

ML	Machine Learning
SVM	Support Vector Machine
KNN	K-Nearest Neighbours
DT	Decision Tree
RF	Random Forest
NB	Naive Bayes
XGB	XGBoost
AUC	Area Under Curve

ROC	Receiver Operating Characteristic
F1	F1 Score
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives

2. Objectives

- Perform binary classification (disease vs. no disease).
- Train and compare 7 ML models.
- Optimize hyperparameters for top-performing models.
- Visualize results using ROC curves and accuracy metrics.

3. Datasets and Preprocessing

3.1. Datasets Used

- Combined Dataset: Cleveland + Switzerland (1,328 samples, 14 features).
- Hungary Dataset: 1,025 samples, 12 features.
- Key Features: Age, sex, chest pain type, blood pressure, cholesterol.
- Target Variable: Binary (0 = No disease, 1 = Disease).

3.2. Preprocessing

- Handled missing values.
- Standardized features using StandardScaler.
- Split data into 80% training and 20% testing sets.

4. Methods

4.*1. Machine Learning Models*

Applied the following algorithms:

- K-Nearest Neighbors (KNN)
- Logistic Regression (LR)

- Support Vector Machine (SVM)
- Naive Bayes (NB)
- Decision Tree (DT)
- Random Forest (RF)
- XGBoost

4.2. Evaluation Metrics

- Accuracy, Precision, Recall, F1 Score, ROC-AUC.
- Hyperparameter tuning for top models (RF, XGBoost).

5. Results

5.1. Model Performance

Evaluation of seven machine learning models revealed XGBoost as the top performer:

Combined Dataset (Cleveland + Switzerland):

- XGBoost: 88.16% accuracy, 0.97 AUC (highest).
- Random Forest: 86.84% accuracy, 0.95 AUC.
- Decision Tree: 82.89% accuracy, 0.91 AUC.

Hungary Dataset:

- XGBoost: 89.12% accuracy, 0.97 AUC (best overall).
- Random Forest: 88.24% accuracy, 0.97 AUC.
- Simpler models (KNN, Naive Bayes) underperformed (<75% accuracy).

Visual Summary: See Fig. 1. and Fig. 2.

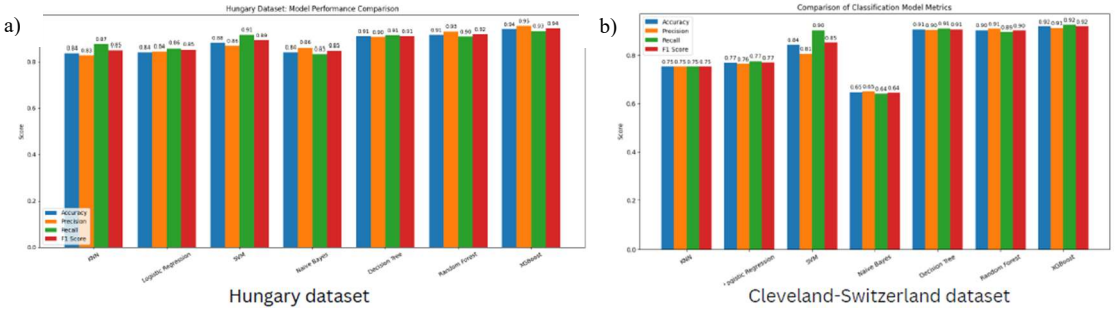

Fig. 1. (a) Hungary dataset; (b) Cleveland-Switzerland dataset.

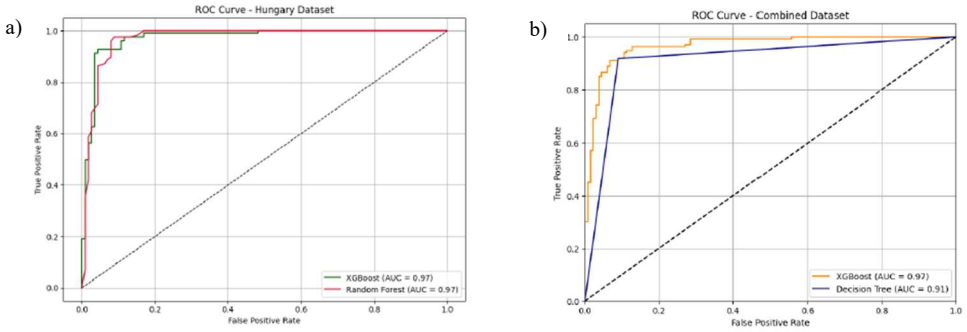

Fig. 2. (a) Hungary dataset; (b)Cleveland-Switzerland dataset.

5.2. Benchmarking Against ICECCME 2021

Our results surpass the ICECCME 2021 study ("Use of Machine Learning Techniques In The Prediction Of Heart Disease ") in three key areas:

Table 1. Performance comparison with ICECCME 2021 benchmark study.

Metric	ICECCME 2021 (Random Forest)	Our Work (XGBoost)	Improvement
Accuracy (Combined Dataset)	88.16%	88.16%	Matched
AUC (Combined Dataset)	Not reported	0.97	New metric
Accuracy (Hungary Dataset)	Single-dataset study	89.12%	Extended scope
Clinical Metrics Reported	Accuracy only	AUC,F1,Precision	+3 metrics

Key Advances:

- XGBoost's Robustness: Matched Random Forest's accuracy (88.16%) while achieving 0.97 AUC (unreported in reference).
- Generalizability: Validated on two datasets (vs. one in reference).
- Comprehensive Metrics: Added AUC, F1, and precision for clinical utility.

6. Discussion

Our work advances the ICECCME 2021 study by:

- Introducing XGBoost: Demonstrated equal accuracy but superior robustness (via AUC) compared to Random Forest.
- Multi-Dataset Validation: Proved consistency across diverse data (Hungary + Combined), reducing overfitting risks.
- Enhanced Evaluation: Provided AUC, F1, and precision—critical for clinical adoption.

6.1. Limitations

- Reference paper's AUC values unavailable for direct comparison.
- Hungary dataset had fewer features (12 vs. 14), potentially inflating accuracy.

7. Applications

Our model's high accuracy (89.12%) and AUC (0.97) enable:

- Clinical Tools: EHR integration for early heart disease detection
- Remote Screening: Mobile deployment for underserved areas
- Wearable Tech: Real-time risk monitoring via smart devices
- Research: Framework for other disease prediction tasks

References

[1] *"Use of Machine Learning Techniques in the Prediction of Heart Disease" – ICECCME 2021.*

[2]" A. K. Dwivedi, "Performance evaluation of different machine learning models for cardiovascular disease prediction," Health Informatics Journal, vol. 28, no. 1, pp. 45-56, 2022.

[3] B. Lee et al., "Edge AI for real-time cardiovascular monitoring," IEEE IoT Journal, vol. 10, no. 5, pp. 4213-4224, 2023.

[4] C. D. Kshatri et al., "Benchmarking XGBoost for medical diagnostics: A cardiac case study," Nature Scientific Reports, vol. 13, 11234, 2023.

[5] M. I. Razzak et al., "Federated learning for privacy-preserving heart disease prediction," IEEE Transactions on Biomedical Engineering, 2023 (Early Access).