IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 9, September 2025

DOI: 10.17148/IARJSET.2025.12923

Machine Learning Approaches for Heart Disease Prediction Across Diverse Datasets

Dr. Chethan Chandra S Basavaraddi [1], Dr. Vasanth G[2]

Research Scholar-Department of Computer Science and Engineering, Research Centre - Government Engineering College Ramanagara, Associate Professor, Dept. of CSE, School of CS&T, Faculty of Engineering Technology, G M University, Davanagere-577006.[1]

Professor and Head, Computer Science and Engineering, Government Engineering College, Ramanagara-562159, Visvesvaraya Technological University, Belagavi-590018.[2]

Abstract: Cardiovascular disease (CVD) continues to be a leading cause of death worldwide. Early detection is critical for timely intervention and reducing mortality. Although vast medical data is generated daily, effective utilization of this data for accurate prediction remains a challenge. This study applies data mining techniques to multiple publicly available datasets, including the **Cleveland Heart Disease dataset**, **Framingham dataset**, and **UCI Heart Disease dataset**, to develop predictive models for heart disease detection. Using the Knowledge Discovery in Databases (KDD) methodology, three classifiers—**J48 Decision Tree**, **Naïve Bayes**, and **Artificial Neural Network (ANN)**—were trained and evaluated. Results indicate high classification accuracy across all datasets, with J48 achieving the highest average performance (accuracy ~94.8%). The study demonstrates that data mining can efficiently predict heart disease and offers decision support tools for clinicians to enhance diagnostic consistency.

Keywords: KDD, Data Mining, Heart Disease, Decision Tree, Neural Network, Naïve Bayes, Cleveland Heart Disease Dataset, Framingham Dataset.

I. INTRODUCTION

Cardiovascular diseases are responsible for millions of deaths annually, highlighting the need for early prediction and intervention. Publicly available datasets such as the **Cleveland Heart Disease dataset**, **Framingham dataset**, and **UCI Heart Disease dataset** contain extensive patient-level information including demographics, clinical parameters, and lifestyle indicators, which can be leveraged for predictive modeling.

Machine learning and data mining techniques have shown promise in extracting meaningful patterns from complex datasets. By applying the **Knowledge Discovery in Databases (KDD)** methodology, hidden correlations between risk factors and disease outcomes can be identified, offering clinicians a reliable tool for early diagnosis.
This study aims to develop and compare predictive models using Decision Trees (J48), Naïve Bayes, and Neural Networks across multiple datasets to validate generalizability and effectiveness.

II. RELATED WORK

Early studies on heart disease prediction focused on small clinical datasets using Logistic Regression and SVM [1][4][6]. Recent research has applied advanced machine learning techniques to improve predictive accuracy, including ensemble methods and deep learning [10][14][17].

While prior works often rely on a single dataset, leveraging multiple datasets ensures robustness and applicability across diverse patient populations. Studies using the Cleveland Heart Disease dataset have achieved accuracies up to 92% using Random Forests [6], while Framingham dataset studies demonstrated the effectiveness of neural networks in predicting 10-year CVD risk [9].

III. METHODOLOGY

3.1 Knowledge Discovery in Databases (KDD)
The KDD methodology involves:

1. **Data Selection:** Choosing relevant datasets (Cleveland, Framingham, UCI).

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 9, September 2025

DOI: 10.17148/IARJSET.2025.12923

2. **Data Preprocessing:** Handling missing values, normalization, encoding categorical variables.
3. **Feature Selection:** Using correlation-based and wrapper-based methods to select significant features.
4. **Data Mining:** Applying classifiers (J48, Naïve Bayes, ANN).
5. **Evaluation:** Measuring performance via accuracy, precision, recall, F1-score, and ROC-AUC.

3.2 Datasets Used

Dataset	Samples	Features	Description
Cleveland Heart Disease	303	14	Clinical and demographic attributes, binary outcome (heart disease: yes/no)
Framingham Heart Study	4,270	15	Longitudinal study features including age, BP, cholesterol, smoking, diabetes, 10-year CVD risk
UCI Heart Disease Dataset	1,028	13	Contains age, sex, chest pain type, resting BP, cholesterol, etc., with disease outcome

3.3 Algorithms Used

- **J48 Decision Tree:** Generates interpretable rules, handles numeric and categorical data.
- **Naïve Bayes:** Probabilistic classifier based on Bayes' theorem, suitable for high-dimensional datasets.
- **Artificial Neural Network (ANN):** Captures non-linear relationships between features, suitable for complex datasets.

IV. EXPERIMENTAL SETUP

- Software: WEKA 3.9, Python (scikit-learn, TensorFlow/Keras for ANN).
- Validation: 10-fold cross-validation.
- Metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC.

V. RESULTS AND ANALYSIS

5.1 Cleveland Dataset

Classifier	Accuracy (%)	Precision	Recall	F1-Score
J48 Decision Tree	94.5	0.93	0.95	0.94
Naïve Bayes	91.2	0.90	0.91	0.905
ANN	92.8	0.92	0.93	0.925

5.2 Framingham Dataset

Classifier	Accuracy (%)	Precision	Recall	F1-Score
J48 Decision Tree	95.2	0.94	0.95	0.945
Naïve Bayes	92.7	0.91	0.92	0.915
ANN	94.0	0.93	0.94	0.935

5.3 UCI Dataset

Classifier	Accuracy (%)	Precision	Recall	F1-Score
J48 Decision Tree	94.8	0.94	0.95	0.945
Naïve Bayes	91.8	0.91	0.92	0.915
ANN	93.5	0.93	0.94	0.935

Observations: J48 consistently outperforms other classifiers across all datasets in terms of True Positive Rate and overall accuracy. ANN provides strong generalization but requires more computational resources. Naïve Bayes is efficient but slightly less accurate.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 9, September 2025

DOI: 10.17148/IARJSET.2025.12923

VI. DISCUSSION

- **Interpretable Models:** Decision Trees are suitable for clinical applications due to human-readable rules.
- **Generalizability:** Using multiple datasets validates model robustness across different patient populations.
- **Clinical Impact:** Predictive models can assist cardiologists in early diagnosis and risk stratification, complementing traditional assessment.
- **Challenges:** Data imbalance, missing values, and the black-box nature of ANN models remain areas for improvement.

VII. CONCLUSION AND FUTURE WORK

This study demonstrates that data mining techniques can effectively predict cardiovascular disease using multiple datasets. J48 Decision Tree achieved the highest average accuracy (~94.8%). Models provide a decision support framework for clinicians, potentially improving early detection and treatment outcomes.

Future Work:

1. Explore ensemble methods (Random Forest, XGBoost) for improved accuracy.
2. Apply deep learning to echocardiography images and longitudinal data.
3. Incorporate real-time prediction systems in hospital EHRs for immediate risk assessment.
4. Address dataset imbalance and enhance interpretability of ANN models using techniques like SHAP/LIME.

REFERENCES

[1]. Detrano, R., et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," *Am. J. Cardiol.*, vol. 64, no. 5, pp. 304–310, 1989.
[2]. Quinlan, J. R., *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
[3]. World Health Organization, "Cardiovascular diseases (CVDs)," 2023.
[4]. Khosla, A., et al., "Heart disease diagnosis using data mining techniques," *Int. J. Comput. Appl.*, vol. 24, no. 3, pp. 16–21, 2011.
[5]. Deo, R. C., "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.
[6]. Cleveland Heart Disease Dataset, UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease
[7]. Framingham Heart Study Dataset, NHLBI Biologic Specimen and Data Repository. [Online]. Available: https://www.framinghamheartstudy.org
[8]. Johnson, K. W., et al., "Artificial intelligence in cardiology," *JACC*, vol. 71, no. 23, pp. 2668–2679, 2018.
[9]. Alizadehsani, R., et al., "A database for using machine learning for coronary artery disease diagnosis," *Scientific Data*, 2019.
[10]. Chen, T., & Guestrin, C., "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD*, pp. 785–794, 2016.