RESEARCH PAPER

Decision Tree-Based Heart Disease Prediction: A Comparative Study Of Id3 And Cart With Java Implementation And Weka Validation

Dr A R JayaSudha[1], S Pradeepkumar[2], T Nagarajan[3], J Jefferson[4], K.S. Suntharasesan[5], K. Sai Prasanna[6], G. Santhosh[7]

[1] Professor / [2] Assistant Professor / [3-7] MCA Final Year Students

Department of Computer Applications (MCA)

Hindusthan College of Engineering and Technology, Coimbatore

Email: sudhahindusthan.backup@gmail.com / pradeepspk1791@gmail.com

Received: 17th Mar, 2026 | Revised: 29th Mar, 2026 | Accepted: 19th Apr, 2026 | Available Online: 5th May, 2026

ABSTRACT

Precise and comprehensible clinical prediction methods are crucial for early identification in cardiovascular medicine. This paper offers a comparative comparison of the ID3 and CART decision tree algorithms for predicting heart disease, utilising the Cleveland Heart Disease Dataset (UCI Repository, n=303). Both methods are analysed via theoretical deduction, Java implementation, and WEKA validation. CART-based REPTree attains an accuracy of 82.4%, surpassing the 72.5% accuracy of ID3-based J48, with enhancements observed in all measures, namely accuracy, recall, precision, and F1-score. The paper illustrates that CART offers enhanced generalisation via binary split and pruning, while preserving the clinical interpretability crucial for decision-support systems.

Keywords: Decision trees, Id3, Cart, Heart disease prediction, Machine learning, Weka, Clinical classification.

How to cite this article: JayaSudha AR, Pradeepkumar S, Nagarajan T, Jefferson J, Suntharasesan KS, Sai Prasanna K, Santhosh G. Decision Tree-Based Heart Disease Prediction: A Comparative Study of ID3 and CART with Java Implementation and WEKA Validation. Int J Drug Deliv Technol. 2026;16(42s): 52-56. DOI: 10.25258/ijddt.16.42s.8

Source of support: Nil.

Conflict of interest: None

1. INTRODUCTION

Cardiovascular illnesses continue to be a primary cause of the worldwide mortality, requiring precise and prompt diagnostic instruments. The Cleveland Coronary Disease Dataset serves as a standard for assessing machine learning classification techniques based on 14 clinical parameters. Machine learning methodologies, especially decision trees, provide an advantageous blend of prediction precision and interpretability—an essential criterion in clinical environments where healthcare professionals must comprehend and have confidence in algorithmic rationale.

Decision tree algorithms are particularly effective for healthcare data because they can manage diverse data types, identify non-linear correlations, and produce comprehensible rule-based models. This paper examines two fundamental algorithms: Iterative Dichotomiser 3 (ID3), which utilises information and entropy gain, and Classification and Regression Tree (CART), which applies the Gini index alongside binary splitting and pruning. A systematic comparison is performed utilising hand derived computation, Java implementation, and WEKA benchmarks.

2. LITERATURE REVIEW

The utilisation of machine learning for healthcare forecasting has advanced considerably. Quinlan (1986) presented the ID3 algorithm, which used entropy-based attribute selection, and subsequently expanded it to C4.5 by incorporating gain ratio normalisation. Breiman et al. (1984) introduced CART utilising Gini impurity and cost-complexity reduction. Kotsiantis (2007) shown that decision trees provide a compromise between efficacy and comprehensibility in medical datasets. Patel et al. (2015) and Sharma et al. (2020) documented robust findings for cardiac disease prediction utilising tree-based classifiers. Nevertheless, the majority of current research depends exclusively on tool-based assessments; few incorporate theoretical analysis, manual execution, and empirical validation – a deficiency our study rectifies.

3. DATASET AND METHODOLOGY

3.1 Dataset Description

The Cleveland Heart Disease Dataset, sourced from the UCI Machine Learning Repository, comprises 303 patient records featuring 13 clinical attributes, including age, sex, chest pain type (cp), resting BP(trestbps), cholesterol (chol), fasting blood sugar (fbs), resting ECG (restecg), maximum heart rate (thalach), exercise-induced angina (exang), ST depression (oldpeak), slope, number of major vessels (ca), and thalassaemia (thal). The binary target variable signifies the existence (1) or nonexistence (0) of heart disease.

Table 1: Dataset Characteristics Summary

Parameter	Value
Dataset Source	UCI Machine Learning Repository
Total Instances	303
Number of Features	13 (+ 1 target)
Target Classes	2 (Disease / No Disease)
Patients with Disease	139 (45.9%)
Patients without Disease	164 (54.1%)
Missing Value Handling	Mode Imputation
Train / Test Split	70% / 30% (212 / 91 instances)

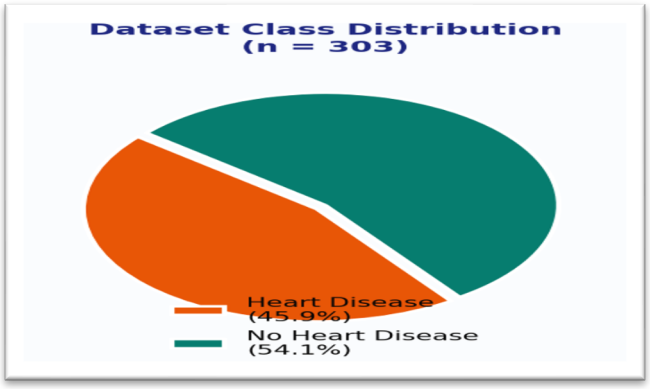

Figure 1: Dataset Class Distribution — 45.9% heart disease vs. 54.1% No Heart Disease (n=303)

3.2 Data Preprocessing

preprocessing entailed designating attribute names pertaining to the raw csv file and transforming the multi-class target variable into a type of binary (0 = no disease, 1 = disease). a limited quantity of absent values (mostly in the ca and thal characteristics) was addressed by mode imputation, maintaining dataset size without considerable information loss. the dataset was divided with a 70/30 train-test split, comprising 212 training scenarios and 91 test instances.

4. ALGORITHMIC FOUNDATIONS

4.1 ID3 Algorithm

ID3 (Iterative Dichotomiser 3) utilises a top-down, greedy approach to pick features that optimise Information Acquisition at each node. Entropy measures the impurity of a dataset, and the property that minimises entropy is selected as the criterion for splitting. ID3 is intrinsically predisposed to favour attributes with numerous distinct values and without a pruning mechanism, hence posing a danger of overfitting on intricate datasets.

Table 2: Core Mathematical Formulas for Decision Tree Algorithms

Metric	Formula	Description
Entropy (ID3)	$H(S) = -\Sigma\, p_i \log_2 p_i$	Measures impurity of dataset
Information Gain	$IG(S,A) = H(S) - \Sigma\, \lvert Sv\rvert/\lvert S\rvert \cdot H(Sv)$	Reduction in entropy after split
Gini Index (CART)	$Gini(S) = 1 - \Sigma\, p_i^2$	Probability of misclassification
Gain Ratio (C4.5)	GR = IG(S,A) / SplitInfo(S,A)	Bias-corrected information gain

4.2 CART Algorithm

CART (Classification and Regression Trees) constructs strictly binary trees using the Gini Index as the splitting criterion. For each candidate split, the weighted Gini impurity of the two resulting subsets is computed; the split minimizing this value is selected. CART applies cost-complexity pruning post-construction to improve generalization. This binary structure, combined with pruning, produces simpler and more balanced trees compared to ID3.

4.3 Comparative Overview

Table 3: Comparative Overview of ID3, C4.5, and CART Algorithms

Feature	ID3	C4.5	CART
Splitting Criterion	Information Gain	Gain Ratio	Gini Index
Tree Structure	Multi-branch	Multi-branch	Binary
Pruning	No	Yes	Yes (cost-complexity)
Continuous Data	Not supported	Supported	Supported
Missing Values	No	Yes	Limited
Overfitting Control	Weak	Moderate	Strong
Interpretability	High	High	Very High

5. MANUAL DERIVATION

To illustrate algorithmic behaviour, a 10-instance sample from the Cleveland dataset is used with binary target (4 disease, 6 no disease).

5.1 ID3 — Entropy and Information Gain

Entropy of the full set: $H(S) = -(0.4 \log_2 0.4 + 0.6 \log_2 0.6) \approx 0.971$. For attribute "Chest Pain (cp)" with subsets cp=1 (4 instances: 3 disease, 1 no disease) and cp=2 (6 instances: 1 disease, 5 no disease):
$H(cp=1) \approx 0.811$, $H(cp=2) \approx 0.650$, Weighted $H = 0.714$, $IG = 0.971 - 0.714 = 0.257$

5.2 CART — Gini Index

$Gini(S) = 1 - (0.4^2 + 0.6^2) = 0.48$. For the same split: $Gini(cp=1) = 0.375$, $Gini(cp=2) = 0.278$, Weighted Gini = 0.314. The attribute minimizing weighted Gini is selected, consistent with ID3 choosing "Chest Pain" in this example.

6. JAVA IMPLEMENTATION

Both algorithms were executed in Java to facilitate a clear, incremental examination of tree creation devoid of tool abstraction. The ID3 algorithm recursively calculates entropy and information gain, determining the optimal attribute for each node. The CART implementation computes the Gini index for all potential binary splits and chooses the partition with the least impurity, employing a depth-based pruning technique after creation.

Essential implementation elements comprise: a loaded data module for CSV parsing, attribute selection algorithms (entropy/Gini), recursive tree construction with termination criteria (pure nodes, empty subsets, maximum depth), and a classification interface. The outputs of the implementation, including the decision tree structure as well as test set predictions, were cross-validated with WEKA findings to verify their accuracy.

- ✓ ID3: entropy + information gain at each node, multi-branch splits, no pruning
- ✓ CART: Gini index, binary splits, simplified depth-constraint post-pruning
- ✓ Both implementations validated against WEKA J48 and REPTree classifiers

7. EXPERIMENTAL RESULTS

7.1 Performance Metrics

Models were evaluated in WEKA using a 70%/30% train-test split. J48 (representing ID3) and REPTree (representing CART) were assessed on accuracy, precision, recall, and F1-score.

Table 4: Classification Performance Comparison (WEKA, 70/30 Split)

Algorithm	Accuracy (%)	Precision	Recall	F1-Score
J48 (ID3)	72.5	0.725	0.725	0.725
REPTree (CART)	82.4	0.825	0.824	0.824

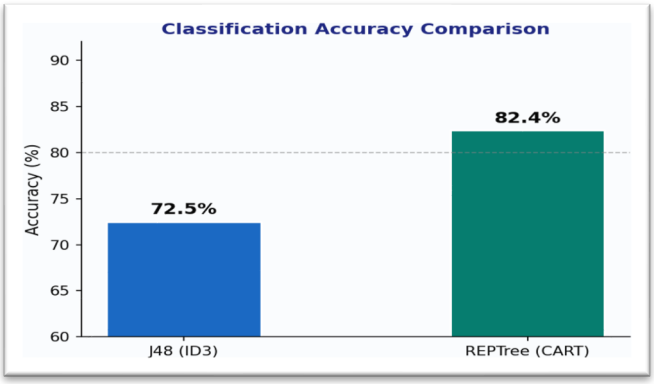

Figure 2: Accuracy Comparison — REPTree (CART) achieves 82.4% vs J48 (ID3) at 72.5%

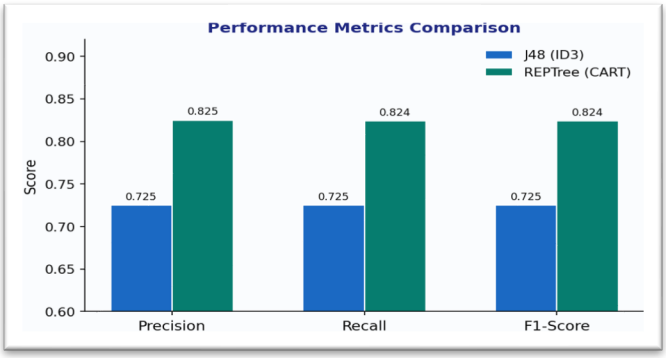

Figure 3: Precision, Recall, and F1-Score Comparison between J48 and REPTree

7.2 Confusion Matrix Analysis

Table 5: Confusion Matrix Summary — TP, TN, FP, FN and Error Rates

Metric	TP	TN	FP	FN	Err%
J48 (ID3)	33	33	13	12	27.5
REPTree (CART)	38	37	9	7	17.6

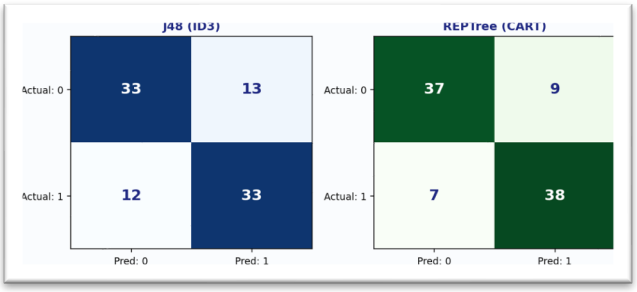

Figure 4: Confusion Matrix Heatmaps for J48 (left) and REPTree (right)

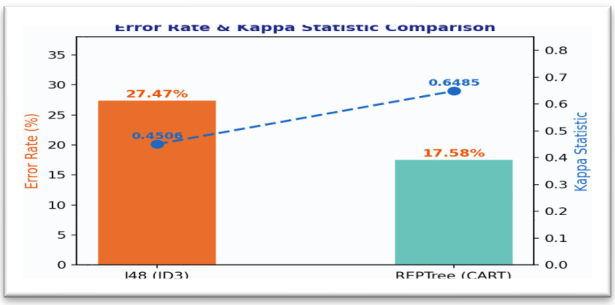

Figure 5: Error Rate and Kappa Statistic Comparison — REPTree reduces error by ~36%

7.3 WEKA Model Structures

The J48 model generated a tree with 15 leaves and 29 nodes, splitting primarily on thal, ca, trestbps, Age, Sex, oldpeak, exang, and fbs. The REPTree produced a considerably more compact tree with only 11 nodes, splitting on thal, ca, exang, Age, and oldpeak — reflecting the effectiveness of CART's pruning in reducing model complexity while improving generalization.

8. DISCUSSION

The findings indicate that the CART-based REPTree model surpasses the ID3-based J48 classifier in all performance metrics. The 9.9 percentage-point enhancement in accuracy (82.4% compared to 72.5%) and 36% decrease in misclassifications (16 versus 25) illustrate the practical superiority of CART for clinical prediction tasks.

This benefit is due to CART's binary splitting method and cost-complexity pruning, which mitigate overfitting by limiting tree expansion to the most predictive divisions. The multi-branch structure of ID3, when not pruned, sometimes overfits the training data, hence diminishing its generalisation ability on novel test examples. Moreover, CART's Gini-based criterion exhibits computational efficiency and reduced vulnerability to attribute-value cardinality bias in comparison to information gain.

From a therapeutic standpoint, both models offer clear, rule-based decision pathways that clinicians can examine and authenticate. The more compact shape of the REPTree (11 nodes compared to 29) is especially beneficial in a healthcare context, as it offers fewer branching conditions for medical professionals to analyse. The principal predictors reported in both models—thalassemia type (thal), number of major vessels (ca), exercise-induced angina (exang), and ST depression (oldpeak)—correspond with recognised clinical risk factors for coronary artery disease.

9. CONCLUSION

This research conducted a rigorous comparison of the ID3 and CART decision tree algorithms for predicting cardiac illness, incorporating theoretical analysis, Java

implementation, and experimental validation using WEKA. The CART-based REPTree attained an accuracy of 82.4% with a Kappa value of 0.65, markedly surpassing the ID3-based J48, which achieved 72.5% accuracy and a Kappa of 0.45. The CART model's concise tree structure and reduced mistake rate render it more appropriate for clinical decision support systems, where accuracy and interpretability are essential.

Subsequent research should investigate ensemble techniques (Random Forest, Gradient Boosting), deep learning frameworks, and expanded multi-centre datasets to enhance predictive efficacy. Integrating explainability frameworks like SHAP values with decision trees could enhance the alignment between machine learning efficacy and clinical reliability.

REFERENCES

[1] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Wadsworth, 1984.

[3] T. M. Mitchell, Machine Learning. McGraw-Hill, 1997.

[4] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2005.

[5] D. Dua and C. Graff, "UCI Machine Learning Repository," 2019. [Online]. Available: https://archive.ics.uci.edu/ml

[6] S. B. Kotsiantis, "Supervised machine learning: A review," Artificial Intelligence Review, 2007.

[7] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," American Journal of Cardiology, 1989.

[8] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Elsevier, 2011.

[9] J. Patel et al., "Heart disease prediction using machine learning," International Journal of Computer Applications, 2015.

[10] P. Sharma et al., "Comparative analysis of classification algorithms," Journal of Healthcare Engineering, 2020.

[11] E. Frank, M. Hall, and I. Witten, The WEKA Workbench. Morgan Kaufmann, 2016.

[12] C. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.