

Original Article

A Decision Tree – Based Heart Disease Prediction Model with SMOTE and Fractional-Order Feature Transformation

Abirami. M¹, Chandrasekaran. S²

^{1,2}Department of Mathematics, Pachaiyappa's College, University of Madras, Chennai, Tamil Nadu, India.

Corresponding Author : iyaradharma522@gmail.com

Received: 13 July 2026

Revised: 19 August 2026

Accepted: 12 September 2026

Published: 29 September 2026

Abstract - This study compares two machine-learning pipelines for heart disease prediction using lipid-profile variables: Decision Tree with Synthetic Minority Oversampling Technique (SMOTE) and fractional-order feature transformation followed by Decision Tree with SMOTE. A truncated Grünwald–Letnikov-style transformation with fractional order $\alpha = 0.5$ was used to construct an alternative feature representation before classification. The study evaluates accuracy, balanced accuracy, precision, sensitivity, F1-score, ROC-AUC, training time, and Decision Tree complexity. The conventional Decision Tree + SMOTE model achieved higher predictive performance on the analyzed dataset, with an accuracy of 87.96% and ROC-AUC of 0.922, whereas the fractional-order model achieved lower predictive values but a shallower tree and slightly shorter training time. The findings indicate that the fractional-order transformation changed the feature representation and tree structure but did not provide a predictive advantage under the experimental conditions considered. The contribution is therefore an empirical comparison of fractional-order feature representation with a conventional SMOTE-based Decision Tree pipeline for lipid-profile-based heart disease prediction.

Keywords - Decision tree, Fractional – Order Feature Transformation, SMOTE, Heart Disease, ROC-AUC.

1. Introduction

Heart disease is a major health problem, and early risk prediction can support timely clinical assessment and preventive intervention. Lipid -profile parameters, including HDL, LDL, T.CHOL, TRI are commonly used in cardiovascular risk assessment [2]. In clinical prediction datasets, class imbalance can reduce the ability of a classifier to identify minority-class observations. SMOTE is therefore used to generate synthetic minority-class observations within the training pipeline [1,6]. The effectiveness of a fractional-order feature transformation combined with Decision Tree classification and SMOTE has not been sufficiently evaluated for lipid-profile-based heart disease prediction.

Previous studies have investigated SMOTE, Decision Tree classifiers, feature optimization, ensemble learning, and related machine-learning approaches for heart disease prediction [4–10]. Limited work has systematically compared a conventional Decision Tree + SMOTE pipeline with a fractional-order feature transformation + Decision Tree + SMOTE pipeline using the same lipid-profile variables and evaluating both predictive performance and model complexity. This study addresses that gap through a controlled empirical comparison.

Before studies have applied SMOTE, DT, feature transformation, and, ensemble[4–10]. But limited studies have joined for evaluation using SMOTE+FOFT+DT in terms of both prediction performance and model complexity.

This study compares two approaches

- ✓ DT + SMOTE
- ✓ FOFT + DT + SMOTE

A Grünwald – Letnikov based transformation is applied to lipid profile feature before classification[3,4].

The contribution of this study is an empirical comparison of two prediction pipelines using the same lipid-profile variables and evaluation framework. The first pipeline combines a Decision Tree with SMOTE, while the second applies a truncated



Grünwald–Letnikov-style fractional-order feature transformation before SMOTE and Decision Tree classification. Predictive performance, feature importance, training time, and tree structure are jointly examined

In heart disease prediction how FOFT affects the prediction performance and DT structure while comparing with DT + SMOTE the evaluation metrics are considered.

The objective is to determine whether fractional-order feature transformation improves heart disease prediction when combined with SMOTE and a Decision Tree, and to examine its effect on feature importance, training time, and Decision Tree structure relative to the conventional Decision Tree + SMOTE approach.

To determine whether FOFT provides predictive improvement and to improve its performance with DT + SMOTE approach.

2. Literature Review

Bhatt et al. (2023) [4] evaluated machine-learning methods for heart disease prediction and emphasized the role of model optimization. Jaddoa (2023) [5] investigated SMOTE with a Decision Tree for classification using a balanced heart-disease dataset, while Aryuni et al. (2023) [6] examined SMOTE for improving minority-class prediction. Asif et al. (2023) [7] explored ensemble learning with hyperparameter optimization, and Yang et al. (2022) [8] combined feature optimization and SMOTE with XGBoost. Srinivasan et al. (2023) [9] investigated active learning for cardiovascular disease prediction. Amilo et al. (2025) [10] reported a hybrid fractional-order and machine-learning approach for heart disease prediction. These studies show the value of imbalance handling, feature representation, and model optimization, but they do not establish that fractional-order transformation improves a Decision Tree + SMOTE pipeline for lipid-profile-based prediction.

2.1. Machine Learning for Heart Disease Prediction

Machine-learning methods have been increasingly applied to heart disease prediction because they can model nonlinear relationships among clinical variables. The studies cited above demonstrate that classifier selection, feature optimization, and hyperparameter tuning can materially affect predictive performance. For the present study, the Decision Tree is retained because its rule-based structure allows the relationship between lipid-profile variables and predicted classes to be inspected directly.

2.2. SMOTE for Imbalanced Medical Data

SMOTE addresses class imbalance by generating synthetic observations for the minority class rather than simply duplicating existing observations [1,6]. Its use is particularly relevant when minority-class sensitivity is important. To prevent information leakage, SMOTE should be applied only to the training portion of each model-development split, while the test data remain untouched.

2.3. Feature Engineering and Fractional-Order Transformation

Feature engineering changes the representation supplied to a classifier and can improve or reduce predictive performance depending on the information contained in the original variables. Fractional-order transformations provide a flexible representation in which observations or feature components can receive weighted contributions. The Grünwald–Letnikov formulation provides one mathematical basis for fractional differences [3]. In this study, the transformation is treated as a feature-engineering operation rather than as a physiological fractional differential equation because the available measurements are cross-sectional.

2.4. Research Gap and Position of the Present Study

The reviewed literature supports the use of SMOTE for imbalance handling and Decision Tree or advanced ensemble methods for heart disease prediction. However, a direct comparison of the conventional Decision Tree + SMOTE representation with a fractional-order transformed representation, using the same lipid-profile inputs and assessing predictive performance together with tree complexity and feature importance, remains limited. The present study is positioned as an empirical evaluation of this combination rather than as a claim of a new classifier or a new resampling algorithm.

3. Methodology

3.1. Data Preprocessing

The data used in this research were collected from the Thanjavur, Dharmapuri, and Tirupathur districts. The dataset was divided into stratified training and testing subsets so that the class proportions were preserved as far as possible. SMOTE was applied only within the training pipeline to avoid information leakage from synthetic training observations into the test set.

- ✓ Decision Tree + SMOTE
- ✓ Fractional – Order Feature Transformation + Decision Tree + SMOTE

Integrated Prediction Pipeline

The two pipelines were evaluated under the same data-partitioning and model-development framework. For the conventional pipeline, the training data were balanced using SMOTE and then supplied to the Decision Tree. For the fractional-order pipeline, the lipid-profile variables were first transformed using the truncated Grünwald–Letnikov-style representation, after which SMOTE was applied to the transformed training data and the Decision Tree was fitted. The held-out test data were not resampled. This sequence was used to reduce the possibility that synthetic observations or preprocessing information would influence the test evaluation.

Pipeline: original data → stratified train/test split → feature transformation where applicable → SMOTE on training data only → Decision Tree fitting → held-out test prediction → evaluation.

The fractional transformation was implemented as a truncated Grünwald–Letnikov-style feature transformation. The dataset contains cross-sectional measurements rather than longitudinal observations, the method is interpreted as fractional – order feature transformation rather than a physiological fractional differential equation model.

Grünwald–Letnikov Fractional Feature Transformation

A fractional-order feature transformation was applied to the lipid-profile variables before SMOTE and Decision Tree classification. The transformation was implemented as a truncated Grünwald–Letnikov-style fractional difference with fractional order $\alpha = 0.5$ and a finite window. The transformed features were then used as inputs to the classifier to investigate whether the modified feature representation could improve prediction performance. The exact truncation-window length must be reported from the experimental implementation before submission; it is not specified in the current manuscript and should not be inferred.

The truncated Grünwald–Letnikov fractional difference is represented as:

$$D^\alpha x(k) = \sum_{j=0}^m (-1)^j \binom{\alpha}{j} x(k-j)$$

where p_i is the probability of class i and C is the total number of classes.

- α = fractional order.
- $C(\alpha, k)$ = generalized fractional binomial coefficient.
- m = finite truncation-window length.

The generalized coefficient is defined by:

$$C(\alpha, k) = \Gamma(\alpha + 1) / [\Gamma(k + 1) \Gamma(\alpha - k + 1)]$$

Stratified 10-fold cross-validation and grid-search optimization were used for model development. Performance was assessed using accuracy, balanced accuracy, precision, sensitivity, F1-score, ROC-AUC, confusion matrices, training time, tree depth, leaf count, and total node count. The manuscript should report the actual grid-search parameter ranges and selected values in the final version to make the experiment reproducible.

3.2. Decision Tree (DT)

DT is a hierarchical supervised learning algorithm that recursively partitions the dataset into subsets based on feature conditions.

The Gini impurity used for node splitting is defined as:

$$\left[Gini(\mathbb{D}) = 1 - \sum_{i=1}^c p_i^2 \right]$$

where:

Entropy can also be used as a splitting criterion:

$$\left[Entropy(\mathbb{D}) = - \sum_{i=1}^c P_i \log_2(P_i) \right]$$

The information gain is computed as:

$$\left[IG(\mathbb{D}, \mathbb{A}) = Entropy(\mathbb{D}) - \sum_{v \in Values(\mathbb{A})} \frac{|\mathbb{D}_v|}{|\mathbb{D}|} Entropy(\mathbb{D}_v) \right]$$

Decision Trees are easy to interpret and visualize; they are prone to overfitting on complex datasets.

3.3. SMOTE Technique

SMOTE generates synthetic minority-class samples by interpolation between a minority observation and one of its selected nearest neighbors:

$$x_{new} = x_i + \lambda(x_{nn} - x_i)$$

where:

- x_i = minority-class sample.
- x_{nn} = selected nearest-neighbor sample.
- λ = random interpolation coefficient in the interval $[0,1]$.

3.4. Evaluation Metrics

Accuracy: The proportion of all observations correctly classified by the model.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Balanced Accuracy: The average of sensitivity and specificity, useful when outcome classes are imbalanced.

$$Balanced\ Accuracy = (Sensitivity + Specificity) / 2$$

Precision: The proportion of predicted positive cases that are actually positive.

$$Precision = TP / (TP + FP)$$

Sensitivity: The proportion of actual heart disease cases correctly identified by the model

$$Sensitivity = TP / (TP + FN)$$

F1-score: The harmonic mean of precision and sensitivity.

$$F1 = 2 \times (Precision \times Sensitivity) / (Precision + Sensitivity)$$

Confusion matrix: A table showing true positives (TP), true negatives (TN), false positives (FP), and, false negatives (FN).

ROC-AUC curve: A graphical representation of sensitivity against the false-positive rate across different classification thresholds.

Time Complexity: The computational time required to fit the model to the training data.

Model Complexity: The structural size of the Decision Tree, evaluated using tree depth, number of leaves, and total number of nodes.

Feature importance: A model- based score representing how much each input feature contributes to the Decision Tree's impurity reduction. It describes the fitted model's use of variables and should not be interpreted as evidence of causation.

4. Results and Discussion

4.1. Predictive Performance Comparison

The predictive results show a consistent advantage for the conventional Decision Tree + SMOTE pipeline on the analyzed dataset. The reported values are reproduced below so that the comparison is explicit rather than dependent only on the graphical presentation.

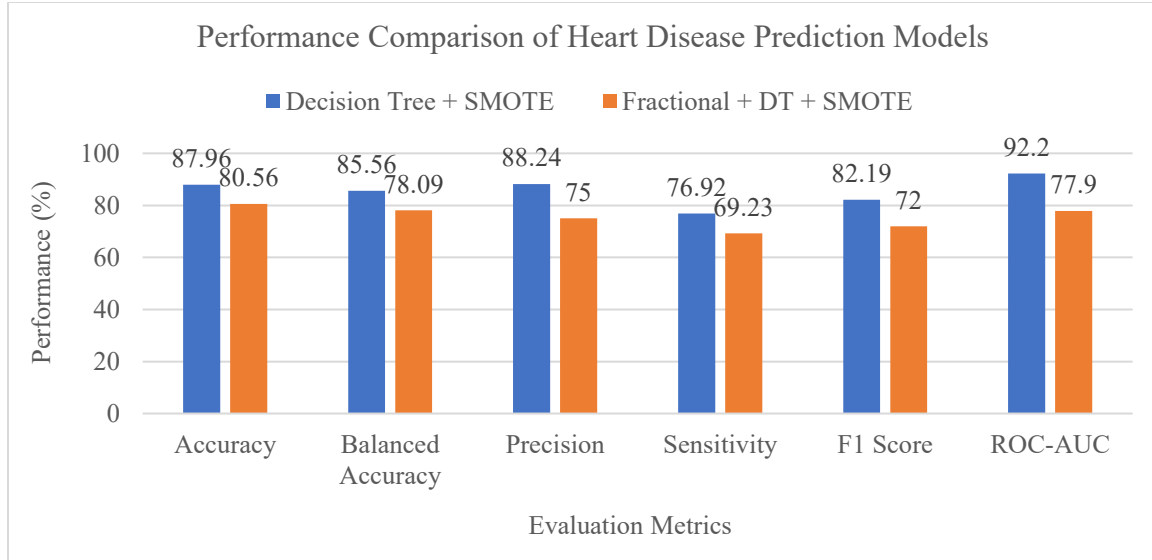


Fig. 1 Performance comparison of the two heart disease prediction models

Decision Tree + SMOTE model outperformed the FOFT + Decision Tree + SMOTE model on every reported predictive metric. The largest difference is observed in ROC-AUC, where the conventional model reached 0.922 compared with 0.779 for the fractional-order model. The results suggest that, for the available lipid-profile representation, the original variables provided more useful discriminatory information than the transformed representation. The lower precision and sensitivity of the fractional-order model also indicate that the transformation did not improve minority-class identification under the reported experimental conditions.

Metric	DT + SMOTE	FOFT + DT + SMOTE
Accuracy	87.96%	80.56%
Balanced Accuracy	85.56%	78.09%
Precision	88.24%	75.00%
Sensitivity	76.92%	69.23%
F1-score	82.19%	72.00%
ROC-AUC	0.922	0.779

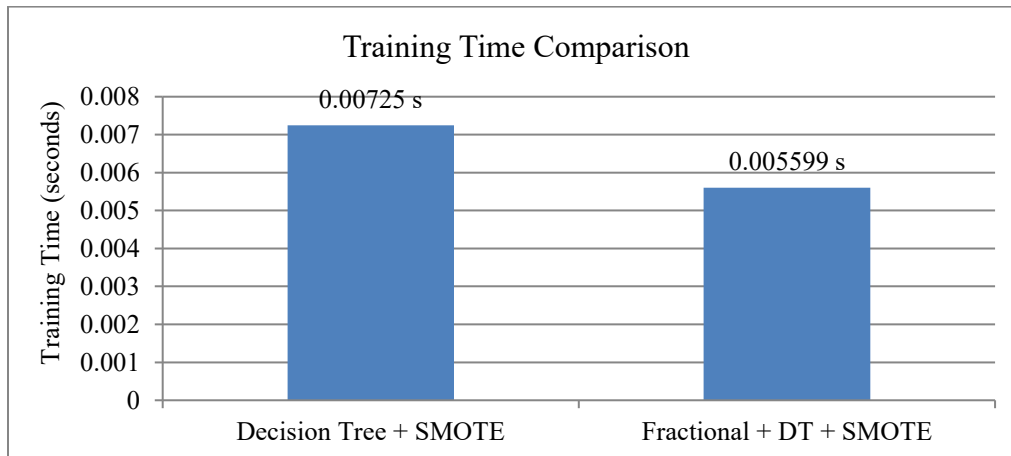


Fig. 2 Training-time comparison of the two models

The training-time results show that both models were computationally fast in the reported run. The fractional-order pipeline required 0.005599 s, compared with 0.007250 s for the conventional pipeline. Although the fractional-order model was faster in this run, the absolute difference was small. Therefore, a meaningful computational advantage should not be claimed without repeated timing trials under the same hardware and software conditions.

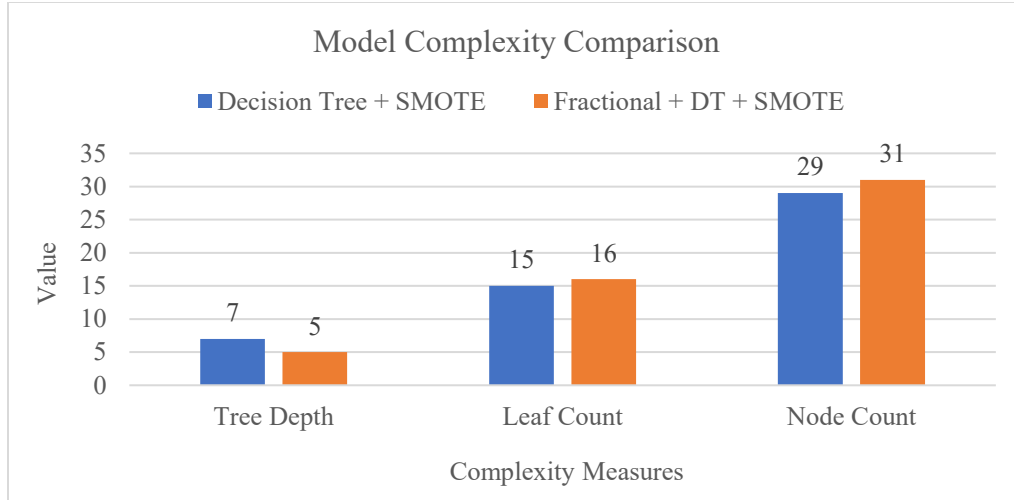


Fig. 3 Decision Tree complexity comparison

The complexity results show that the conventional Decision Tree + SMOTE model had depth 7, 15 leaves, and 29 nodes, whereas the fractional-order model had depth 5, 16 leaves, and 31 nodes. The fractional-order representation therefore produced a shallower tree but slightly more leaves and total nodes. A lower depth alone does not imply lower overall structural complexity; the leaf and node counts should also be considered.

Table 3. Training time and Decision Tree complexity

Measure	DT + SMOTE	FOFT + DT + SMOTE
Training time (s)	0.007250	0.005599
Tree depth	7	5
Leaf count	15	16
Node count	29	31

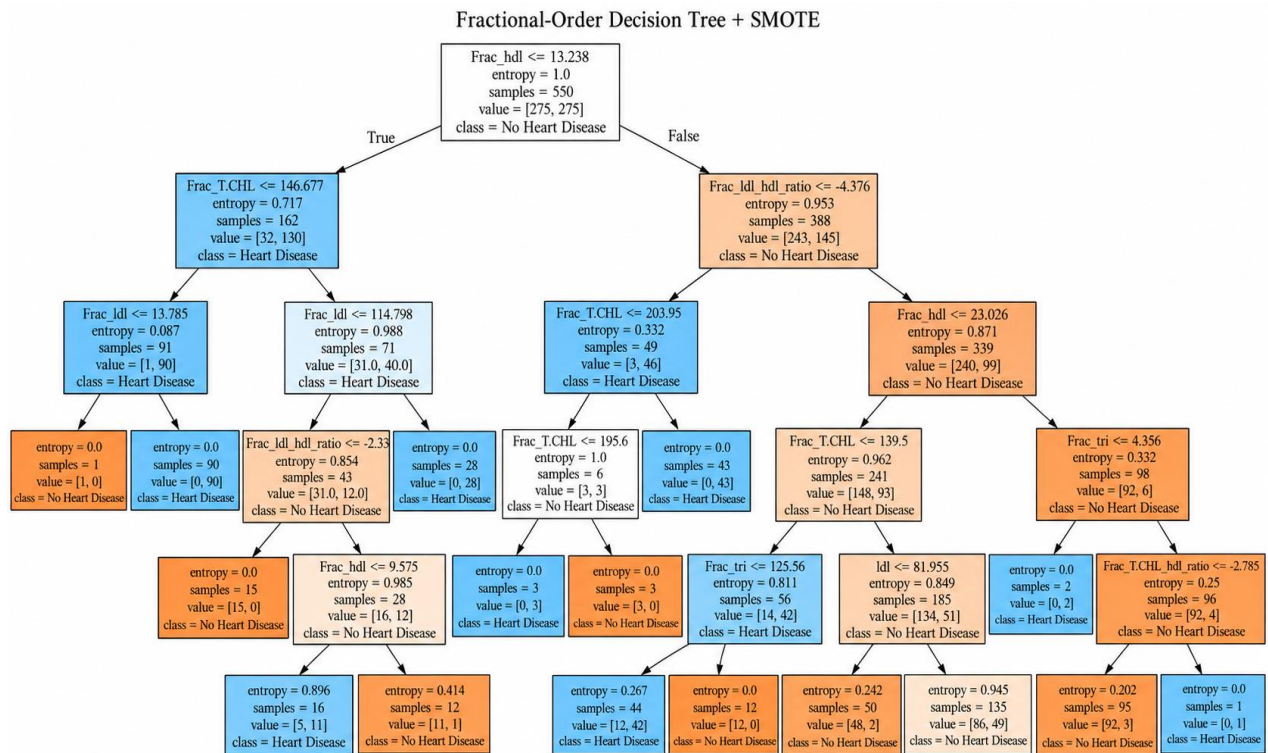


Fig. 4 Decision Tree with SMOTE

The Decision Tree diagram explain the hierarchical classification rules learned from the original lipid-profile variables. HDL appears at the root node, followed by variables such as LDL, TRI, T.CH.L, and lipid ratios. Each node reports a splitting condition and associated impurity, sample count, class distribution, and predicted class, making the fitted rules comparatively interpretable. The reported root condition is $HDL \leq 35.778$.

Component	Description
Model	Decision Tree with SMOTE- balanced data.
Root Node	$HDL \leq 35.778$ (first splitting feature.)
Main Feature	HDL, LDL, TRI, T.CH.L/HDL ratio.
Decision Rule	If condition is True($\leq$) go left; ($>$) go right.
Entropy	Measures node impurity (0=pure node).
Samples	Number of records at each node.
Value	Number of samples in each class. Eg., [class 0, class1]
Leaf Node	Final prediction: heart disease or no heart disease
Blue Node	Predicts heart disease.
Orange Node	Predicts No heart disease.
Purpose	Predict heart disease using original clinical features

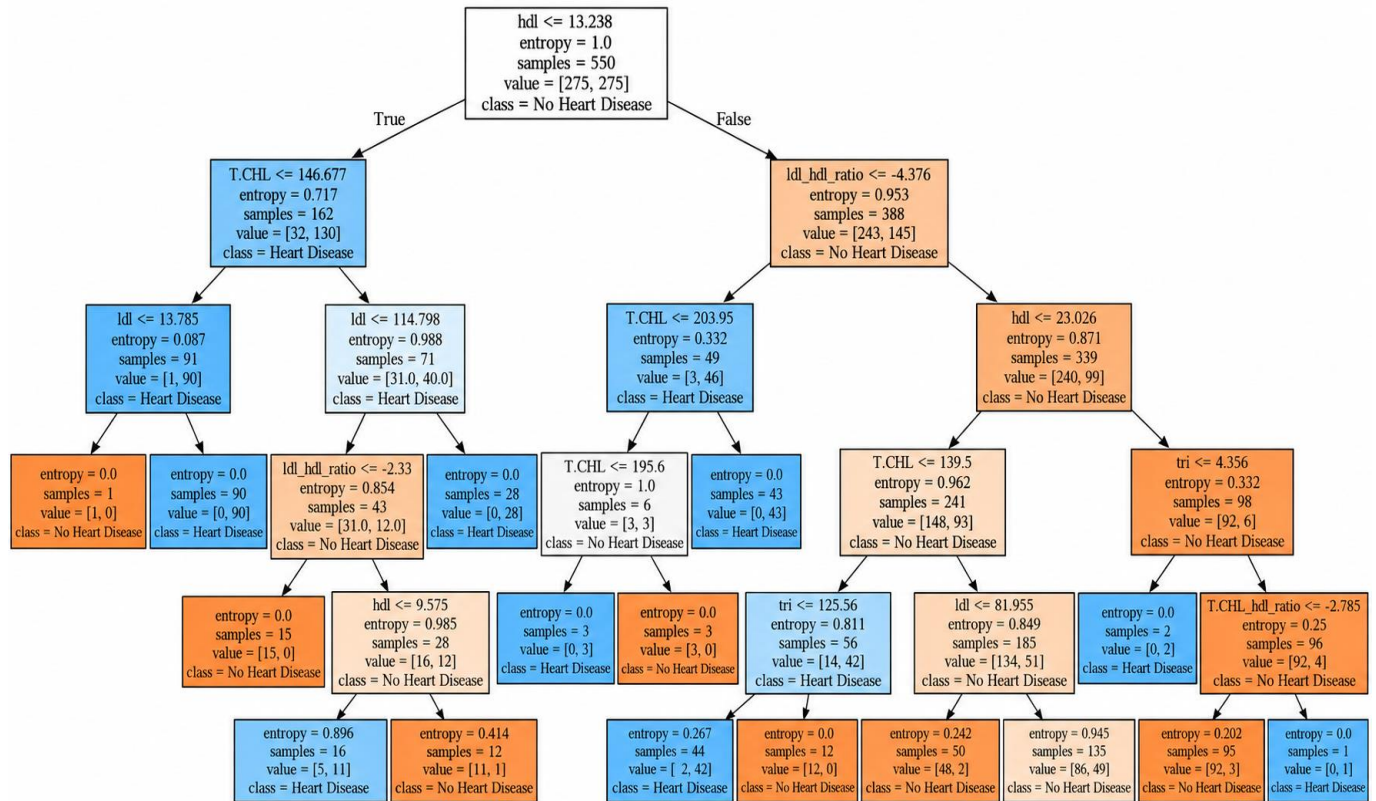


Fig. 5 Fractional-order Decision Tree with SMOTE

The fractional-order tree diagram presents the classification rules learned after feature transformation. The tree uses transformed variables and threshold values that differ from the conventional model. Its maximum depth is lower, but it contains slightly more total nodes. This indicates that the fractional transformation altered the decision structure without improving the overall predictive performance. The reported root condition is $Frac_hdl \leq 13.325$.

Component	Description
Model	Fractional- Order Decision Tree with SMOTE – balanced data.
Root Node	Frac_hdl<= 13.325 (first splitting feature)
Main Features	Frac_hdl, frac_ldl , frac_tri, frac_T.CHL,frac_T.CHL_HDL_ratio
Decision Rule	If condition is true (<=) go left; (>) go right
Entropy	Measures node impurity (0= pure node)
Samples	Number of records at each node
Value	Number of Samples in each class, [class 0, class 1]
Leaf Node	Final Prediction Heart Disease or No heart Disease
Blue Node	Predicts Heart Disease
Orange Node	Predicts No Disease
Purpose	Predict heart disease using fractional – order transformed features.

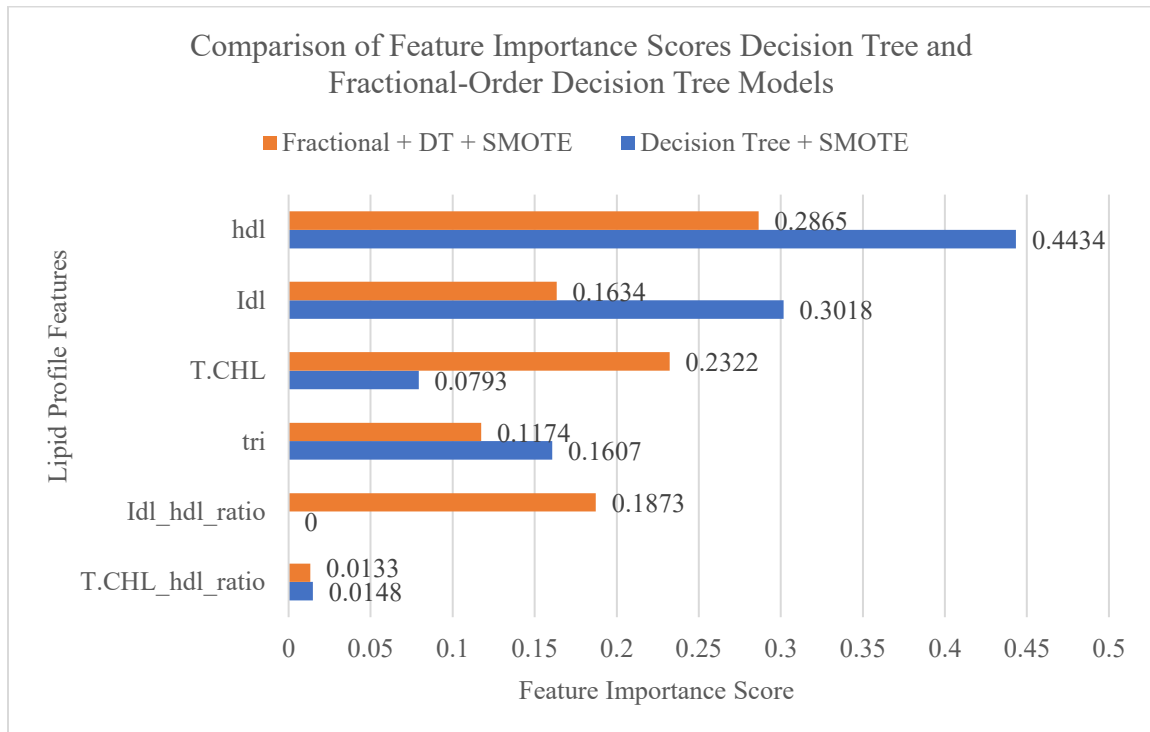


Fig. 6 Feature-importance comparison of the two models

The feature-importance results show that HDL was the strongest predictor in both fitted models. For the conventional model, HDL and LDL had the largest reported importance scores, 0.4434 and 0.3018, respectively. In the fractional-order model, importance was distributed more broadly among transformed HDL, T.CHL, LDL/HDL ratio, and LDL. These values describe how the fitted Decision Tree used the available variables for impurity reduction; they should not be interpreted as evidence that a lipid variable causes heart disease.

Table 4. Reported feature-importance values

Feature	DT + SMOTE	FOFT + DT + SMOTE
HDL	0.4434	0.2865
LDL	0.3018	0.1634
T.CHL	0.2322	0.2232
TRI	0.1607	0.1178
LDL/HDL ratio	0.0000	0.1873
T.CHL/HDL ratio	0.0148	0.0210

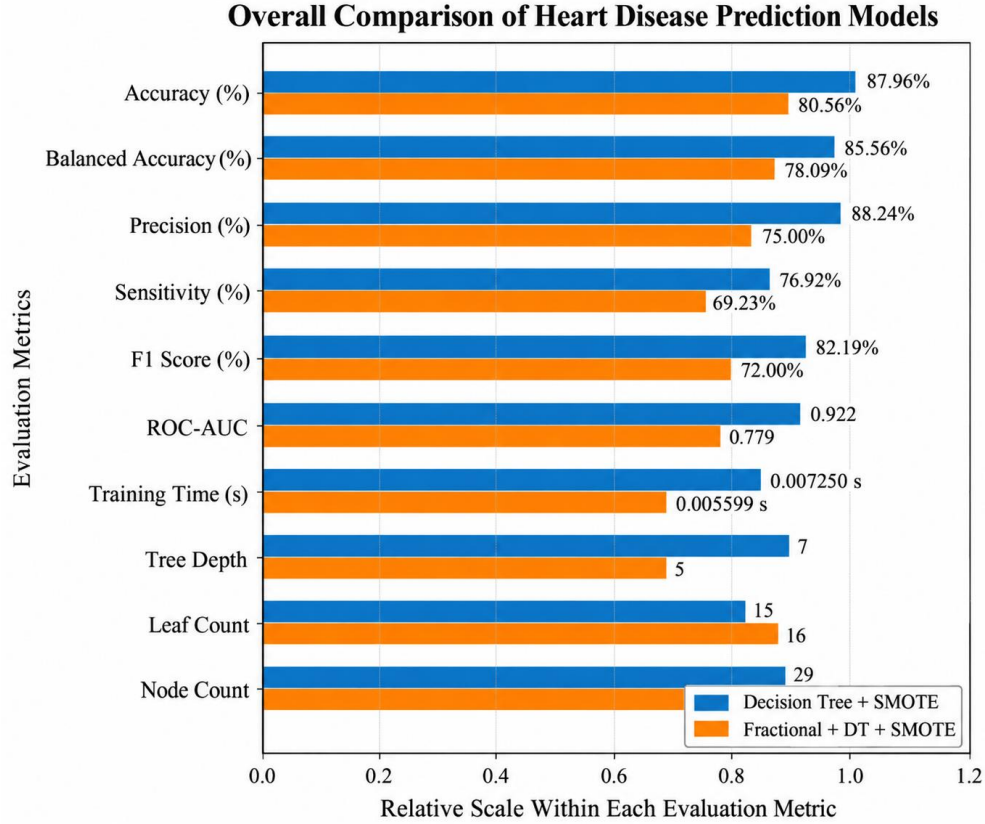


Fig. 7 Overall comparison of the two models

4.2. Comparison with Existing Studies

The present results should be interpreted alongside previous studies rather than as a direct claim of superiority over all published methods. Jaddoa [5] and Aryuni et al. [6] support the use of SMOTE with imbalanced heart-disease data, while Yang et al. [8] and Asif et al. [7] illustrate the use of feature optimization and ensemble approaches. The current study differs in its focus on the direct comparison of a conventional Decision Tree + SMOTE pipeline with a fractional-order transformed pipeline. Because the datasets, preprocessing procedures, and experimental settings reported in the literature are not identical, numerical performance should not be compared as if all studies were conducted under the same conditions.

4.3. Parameter Selection and Reproducibility

The fractional-order parameter was set to $\alpha = 0.5$ in the reported experiment. The final manuscript should additionally state the truncation-window length and the grid-search ranges used for the Decision Tree hyperparameters. If multiple fractional orders or window lengths were tested, those candidate values and the selection criterion should be reported. If $\alpha = 0.5$ and the window length were fixed a priori, that decision should be stated explicitly. These details are necessary for reproducibility.

4.4. Mathematical Interpretation and Limitations

The fractional-order representation depends on the selected fractional order and truncation window. Truncation approximates the corresponding non-truncated fractional difference and therefore introduces a representation that is parameter dependent. In addition, the available measurements are cross-sectional rather than longitudinal; consequently, the transformation should not be interpreted as a physiological fractional differential equation. The present findings also show that a richer transformed representation does not necessarily produce better classification. Future evaluations should examine sensitivity to α and the truncation window using larger and more diverse datasets.

The overall comparison indicates a trade-off between predictive performance and structural characteristics. The Decision Tree + SMOTE model provides stronger accuracy, balanced accuracy, precision, sensitivity, F1-score, and ROC-AUC. The fractional-order model has a slightly shorter training time and lower tree depth, but it also has more leaves and nodes. Consequently, the conventional model provides the stronger overall balance of predictive performance and interpretability for the analyzed dataset.

Study	Approach	Relevance to present study
Bhatt et al. [4]	Machine-learning methods for heart disease	Supports the importance of model selection and optimization.
Jaddoa [5]	SMOTE + Decision Tree	Directly supports imbalance handling with DT.
Aryuni et al. [6]	SMOTE for imbalanced heart-disease data	Supports minority-class improvement.
Asif et al. [7]	Ensemble learning + optimization	Provides an advanced-method comparison context.
Yang et al. [8]	Feature optimization + SMOTE-XGBoost	Provides a feature-optimization and imbalance-handling comparison context.
Srinivasan et al. [9]	Active learning for cardiovascular prediction	Provides a broader ML comparison context.
Amilo et al. [10]	Fractional-order + machine learning	Provides recent fractional-order context.

5. Conclusion

This study compared a conventional Decision Tree + SMOTE model with a truncated Grünwald–Letnikov-style fractional-order feature transformation followed by Decision Tree + SMOTE for heart-disease prediction using lipid-profile variables. The fractional-order transformation was introduced as a feature-engineering operation intended to modify the representation of the lipid variables before classification. On the analyzed dataset, the conventional Decision Tree + SMOTE model achieved higher predictive performance, including an accuracy of 87.96% and ROC-AUC of 0.922. The fractional-order model produced a shallower tree and a slightly shorter training time, but it did not improve predictive accuracy or discrimination. These results indicate that the original lipid-profile representation was more effective for the present Decision Tree classifier than the selected fractional-order representation. The contribution of the study is therefore an empirical assessment of the combined FOFT–SMOTE–Decision Tree pipeline and its effect on prediction and model structure.. Future work should investigate different fractional orders and truncation windows, larger datasets, additional clinical variables, repeated statistical evaluation, and longitudinal observations to determine whether fractional-order representations can provide benefits under other prediction settings.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Funding Statement

This research received no external funding.

References

- [1] Nitesh V. Chawla et al., “SMOTE: Synthetic Minority Over-Sampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Scott M. Grundy et al., “2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol,” *Journal of the American College of Cardiology*, vol. 73, no. 24, pp. e285-e350, 2018. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [3] Rudolf Scherer et al., “The Grünwald–Letnikov Method for Fractional Differential Equations,” *Computers & Mathematics with Applications*, vol. 62, no. 3, pp. 902-917, 2011. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [4] Chintan M. Bhatt et al., “Effective Heart Disease Prediction using Machine Learning Techniques,” *Algorithms*, vol. 16, no. 2, pp. 1-14, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Ahmed Sami Jaddoa, “Heart Disease Prediction System using SMOTE Technique Balanced Dataset and Decision Tree Classifier,” *AIP Conference Proceedings*, vol. 2834, no. 1, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Mediana Aryuni et al., “Imbalanced Learning in Heart Disease Categorization: Improving Minority Class Prediction Accuracy using the SMOTE Algorithm,” *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 23, no. 2, pp. 140-151, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [7] Daniyal Asif et al., “Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization,” *Algorithms*, vol. 16, no. 6, pp. 1-17, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] Jian Yang, and Jinhan Guan, “A Heart Disease Prediction Model based on Feature Optimization and Smote-Xgboost Algorithm,” *Information*, vol. 13, no. 10, pp. 1-15, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] Saravanan Srinivasan et al., “An Active Learning Machine Technique based Prediction of Cardiovascular Heart Disease from UCI-Repository Database,” *Scientific Reports*, vol. 13, no. 1, pp. 1-19, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [10] David Amilo, Khadijeh Sadri, and Evren Hincal, "A Hybrid Approach to Heart Disease Prediction using a Fractional-Order Mathematical Model and Machine Learning Algorithm," *Computer Methods in Biomechanics and Biomedical Engineering*, pp. 1-30, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]