International Journal of Academic Information Systems Research (IJAISR)
ISSN: 2643-9026
Vol. 10 Issue 2 February - 2026, Pages: 1-12

An Interpretable Clinical Expert System for Heart Disease Diagnosis via Decision Tree-Derived Rules

Fatima M. Salman, Samy S. Abu-Naser

1Department of Information Technology, Faculty of Engineering and Information Technology, Al-Azhar University, Gaza, Palestine
Email:abunaser@alazhar.edu.ps

Abstract: *Cardiovascular diseases remain the leading cause of global mortality, responsible for an estimated 17.9 million deaths annually. Early and interpretable diagnosis is critical to improving patient outcomes. This study bridges the gap between black-box machine learning models and clinical interpretability by developing a hybrid expert system that derives transparent, evidence-based diagnostic rules directly from patient data. We trained a constrained Decision Tree model (maximum depth = 3) on a publicly available dataset of 918 patients to ensure inherent explainability. The tree structure was systematically parsed and translated into a set of eight mutually exclusive IF-THEN clinicalrules, forming the knowledge base of a prototype rule-based expert system. Each rule is accompanied by its confidence score, support, and a clinically meaningful interpretation. The system achieved an accuracy of 82.1% (95% CI: 76.5 – 87.6%), a precision of 87.1%, and a recall of 79.4% on an independent test set. Five-fold cross-validation confirmed robustness (mean accuracy: 81.5% ± 6.1%), and the area under the ROC curve (AUC) was 0.886, indicating excellent discriminative ability.Crucially, for each case, the system outputs the specific diagnostic rule applied, a risk-stratified recommendation (e.g., urgent cardiology consultation, routine screening), and a confidence score. and a clinical interpretation, moving beyond a mere prediction to an auditable diagnostic aid. This transforms the model from a predictive endpoint into a fully auditable and interpretable clinical decision support tool. This work provides a reproducible and scalable framework for building accurate, transparent, and clinically deployable diagnostic systems, facilitating greater trust and adoption in real-world healthcare settings.*

Keywords: Expert Systems, Interpretable AI, Clinical Decision Support, Decision Trees, Heart Disease Prediction, Rule-Based Systems.

1. Introduction

Cardiovascular diseases (CVDs) remain the foremost cause of global mortality, responsible for an estimated 17.9 million deaths annually, representing 32% of all deaths worldwide [1]. This immense burden underscores the critical need for early, accurate, and actionable diagnosis. However, achieving this is complex due to the multifactorial nature of heart disease, in which symptoms often overlap and diagnosis relies on synthesizing diverse indicators—from clinical presentation and lipid profiles to subtle electrocardiogram (ECG) anomalies such as ST-segment deviations. Figure 1 illustrates the global prevalence and impact of cardiovascular diseases.

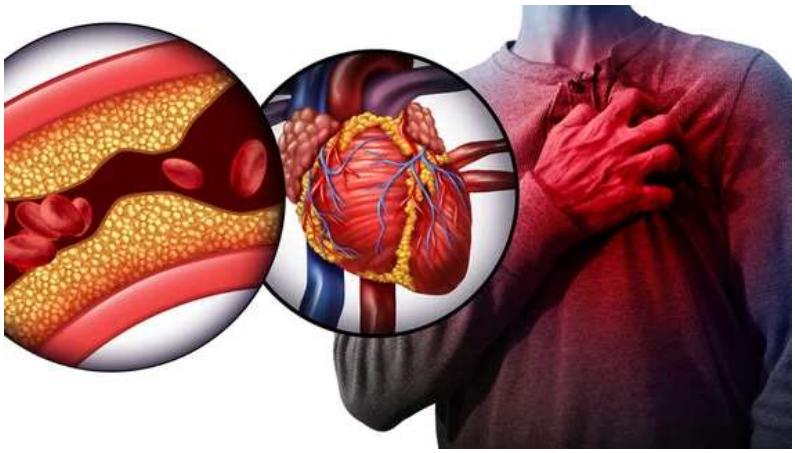

Figure 1: Cardiovascular diseases

Artificial Intelligence (AI), particularly machine learning (ML), has demonstrated significant potential to augment diagnostic accuracy by identifying complex, non-linear patterns in clinical data. Models such as random forests, gradient boosting machines, and deep neural networks have achieved high predictive performance in heart disease risk stratification [2]. Despite their power, a

major barrier to clinical adoption persists: the "black-box" problem. These complex models often fail to provide intuitive, human-understandable explanations for their predictions, hindering clinician trust, auditability, and integration into the clinical workflow. This creates a critical gap between statistical performance and actionable clinical insight.

Rule-based Expert Systems (ES) represent a classical AI approach designed explicitly for interpretability and transparency. By encoding domain knowledge into a structured set of logical IF-THEN rules, they emulate transparent, human-like reasoning, making their decision pathways fully auditable. Their utility in medical decision support is well-documented, offering a clear rationale for each diagnostic output [3]. The traditional challenge in building such systems, however, has been the "knowledge acquisition bottleneck"—the labor-intensive, time-consuming, and often subjective process of manually eliciting and formalizing rules from human experts. Figure 2 presents the main components of a typical rule-based expert system.

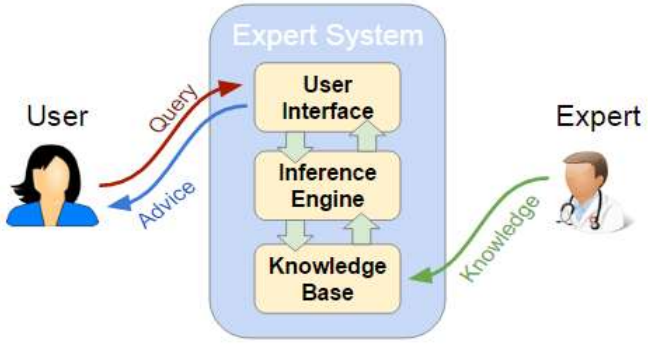

Figure 2: Main Components of Expert System

This study bridges the interpretability gap of modern ML and the knowledge bottleneck of classical ES by proposing a novel, hybrid diagnostic framework. We leverage a Decision Tree (DT) algorithm—a white-box model prized for its intrinsic interpretability—not merely as a predictive classifier, but as a systematic rule-extraction engine. Our core methodological contribution is the automatic parsing and translation of the trained DT structure into a transparent, mutually exclusive set of clinical production rules. These rules automatically form the Knowledge Base of a prototype expert system. Thus, our work synthesizes the data-driven accuracy of machine learning with the logical transparency and actionable output of rule-based reasoning.
The developed system is designed to achieve four primary objectives: (1) Identify key clinical predictors (e.g., ST_Slope, ChestPainType, Oldpeak) via the interpretable DT model and quantify their relative importance; (2) Achieve robust predictive accuracy (target > 80%) while maintaining strict model interpretability through structural constraints (ma depth = 3); (3) Automatically extract and validate a concise set of clinical decision rules that are directly comprehensible to clinicians without specialized ML expertise; (4) Provide, for each patient case, an explicit diagnostic justification, a confidence score, and personalized, risk-stratified clinical recommendation (e.g., urgent cardiology consultation, further cardiac evaluation advised, routine screening recommended).

By transforming the model from a predictive endpoint into a fully auditable and interpretable Clinical Decision Support System (CDSS), this work provides a practical and reproducible framework for deploying accurate, transparent, and clinically trustworthy diagnostic tools. Such systems are designed not to replace clinical judgment, but to augment evidence-based decision-making at the point of care.

2. Materials and Methods

2.1 Dataset Description and Source

The dataset utilized in this study was obtained from a publicly available Kaggle repository titled “Heart Failure Prediction” [4]. It was formatted as a CSV file comprising 918 patient records, with 12 clinical features and a binary target variable indicating the presence (1) or absence (0) of heart disease. The dataset exhibits a class distribution of 508 positive cases (55.3%) and 410 negative cases (44.7%), representing a slightly imbalanced but clinically realistic prevalence of coronary artery disease. The features include both numerical and categorical variables:

- **Demographic:**
  - **Age:** age in years.
  - **Sex:** M (Male) /F (Female).
- **Clinical & Laboratory:**
  - **RestingBP:** resting blood pressure (mmHg).

- **Cholesterol:** serum cholesterol (mg/dL).
- **FastingBS:** fasting blood sugar (1 if > 120 mg/dL, else 0).
- **MaxHR:** maximum heart rate achieved (bpm).
- **ExerciseAngina:** exercise-induced angina (Y/N).
- **Oldpeak:** ST depression induced by exercise relative to rest (mm).
- **ST_Slope:** slope of the peak exercise ST segment (Up/Flat/Down).

- **Electrocardiographic:**
  - **RestingECG:** resting electrocardiogram results (Normal/ST/LVH).
- **Symptoms:**
  - **ChestPainType:** type of chest pain (ASY: Asymptomatic, ATA: Atypical Angina, NAP: Non-Anginal Pain, TA: Typical Angina).
- **Target Variable:**
  - **HeartDisease:** binary (1= presence, 0 = absence).

2.2 Data Preprocessing

To prepare the data for modeling, the following preprocessing steps were applied:

1. Handling Missing Values: The dataset was examined for missing entries. No missing values were present in any of the 918 records; therefore, no imputation was required.
2. Categorical variables Encoding: Categorical features were transformed using Label Encoding, which preserves ordinal relationships and is suitable for decision tree algorithms. Table 1 displays the mapping schemes as follows:

Table 1: Mapping schemes

Feature	Mapping
Sex	{F: 0, M: 1}
ChestPainType	{ASY: 0, ATA: 1, NAP: 2, TA: 3}
RestingECG	{Normal: 0, ST: 1, LVH: 2}
ExerciseAngina	{N: 0, Y: 1}
ST_Slope	{Down: 0, Flat: 1, Up: 2}

3. Feature Scaling: All numerical features were used in their original scale. Decision trees are invariant to monotonic transformations and do not require normalization or standardization.

2.3 Model Development

The dataset was partitioned using stratified random sampling into a training set (80%, n=734) and a test set (20%, n=184) to preserve the original class distribution.

A Decision Tree Classifier was implemented using the scikit-learn library (version 1.2.2) in Python 3.12.

To balance interpretability and predictive performance, the tree depth was constrained to a maximum of 3 levels. This constraint was deliberately chosen to:

- Generate a small, human-readable set of if-then rules.
- Avoid overfitting.
- Maintain competitive accuracy.

Table 2 displays the model that was trained using the following hyperparameters:

Table 2: Hyperparameters for the model training

Parameter	Value	Rationale
criterion	gini	Standard impurity measure for classification
max_depth	3	Limits complexity, ensures interpretability

min_samples_split	20	Prevents overfitting on small sample sizes
min_samples_leaf	10	Ensures each leaf has sufficient support
random_state	42	Ensures reproducibility

2.4 Rule Extraction and Expert System Design

The primary methodological contribution of this study is the systematic translation of the trained decision tree into a transparent, rule-based expert system.

Step1: Tree Parsing

The decision tree structure was programmatically parsed to identify all root-to-leaf paths. Each path represents a unique combination of feature conditions leading to a classification outcome.

Step 2: Rule Generation

Each root-to-leaf path was converted into an IF-THEN production rule of the form:

IF [condition$_1$] AND [condition$_2$] AND ... THEN [prediction]

Step 3: Rule Metrics

For each extracted rule, the following performance metrics were calculated:

- **Confidence:** The proportion of training samples in the leaf node belonging to the majority class.
- **Support:** The total number of training samples classified by the rule.

Step 4: Post-Processing

Encoded threshold values were reverse-mapped to their original categorical labels to ensure clinical readability. For example:

ChestPainType ≤ 0.5 → ChestPainType = ASY (Asymptomatic)

Step 5: Rule Set Characteristics

The extracted rule set is mutually exclusive and collectively exhaustive—each training sample is classified by exactly one rule, and the complete set of rules covers 100% of the dataset.

Step 6: Expert System Prototype

The extracted rules were implemented as a prototype expert system in Python. For any new patient case, the system:

- Evaluates the rules in order.
- Triggers the first matching rule.
- Outputs the prediction, confidence score, support, and a predefined clinical recommendation associated with that rule.

2.5 Validation and Statistical Analysis

1. Model Validation:

- **Hold-out validation:** The model was evaluated on the independent test set (n = 184).
- **Cross-validation:** A 5-fold stratified cross-validation strategy was applied on the entire dataset to assess generalizability and stability.

2. Performance Metrics:

The following metrics were calculated from the confusion matrix:

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Precision (PPV) = TP / (TP + FP)
- Recall (Sensitivity) = TP / (TP + FN)
- Specificity = TN / (TN + FP)
- F1-Score = 2 × (Precision × Recall) / (Precision + Recall)
- AUC-ROC: Area Under the Receiver Operating Characteristic Curve

A 95% confidence interval for accuracy was calculated using the normal approximation method: $CI = \hat{p} \pm z\, x \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where z =1.96 and *n* is the test set size.

3. Statistical Comparisons:

Differences between disease and non-disease groups were assessed using:

- Independent t-tests for continuous variables (e.g., Age, Cholesterol, MaxHR, Oldpeak).
- Chi-square tests for categorical variables (e.g., Sex, ChestPainType, ST_Slope).
- Cohen's d was reported as a standardized measure of effect size for significant continuous variables.

All statistical analyses were performed using the scipy and statsmodels libraries in Python, with a significance level set at $\alpha = 0.05$.

3. Results

3.1 Study Population Characteristics

A total of 918 patients were included in this study, comprising 508 (55.3%) individuals with confirmed heart disease and 410 (44.7%) without heart disease. The baseline characteristics of the study population are summarized in Table 4.

Figure 3 presents an exploratory data analysis dashboard summarizing the distribution of key clinical features and their association with heart disease status.

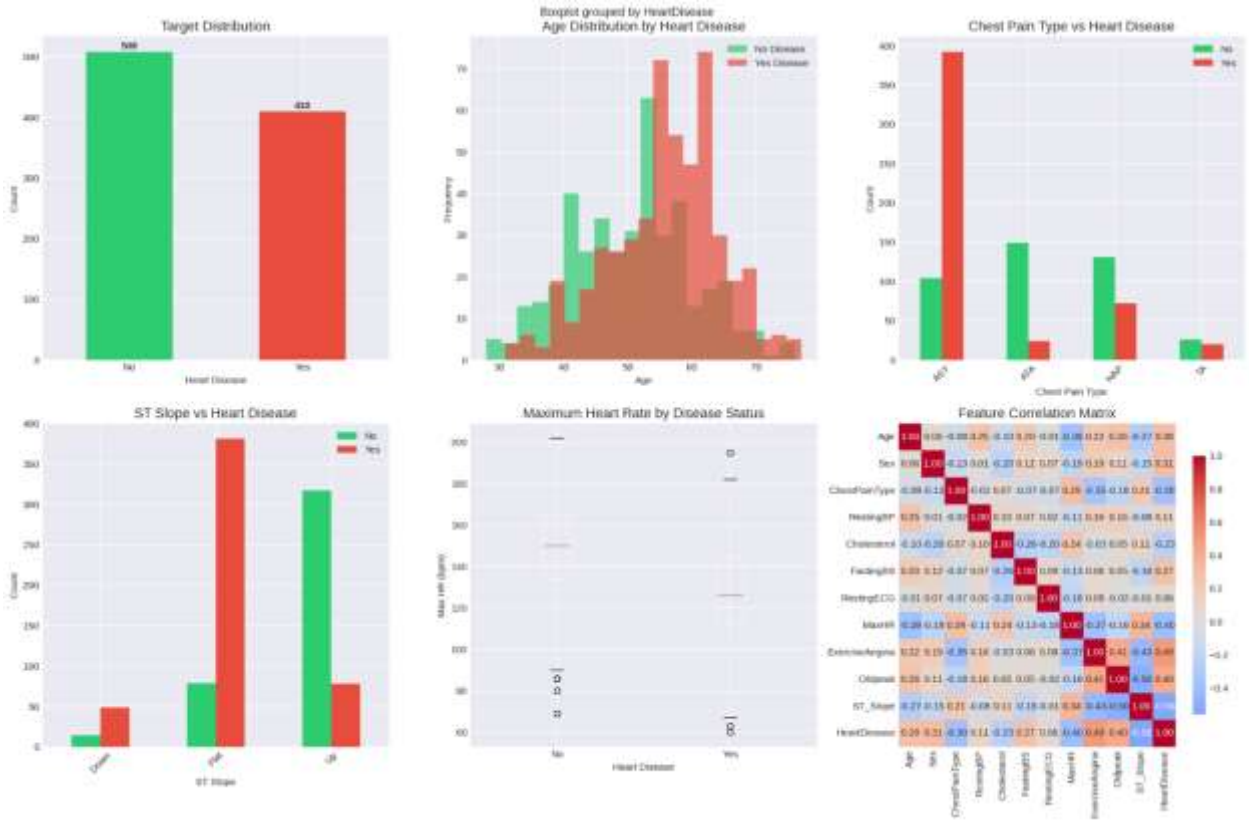

Figure 3: Exploratory Data Analysis Dashboard

The mean age of the cohort was 53.5 ± 9.4 years, with a predominance of male participants (n = 725, 79.0%). Patients with heart disease were significantly older than those without disease (55.9 ± 8.7 vs. 50.6 ± 9.4 years, $p < 0.001$, Cohen's d = 0.59), representing a medium-to-large effect size.

Regarding clinical parameters, the heart disease group demonstrated:

- Significantly lower cholesterol levels (175.9 ± 126.4 vs. 227.1 ± 74.6 mg/dL, $p < 0.001$, d = -0.49).
- Reduced maximum heart rate (127.7 ± 23.4 vs. 148.2 ± 23.3 bpm, $p < 0.001$, d = -0.88).
- Higher ST depression values (Oldpeak: 1.3 ± 1.2 vs. 0.4 ± 0.7 mm, $p < 0.001$, d = 0.91).

Analysis of categorical variables revealed strong, statistically significant associations with heart disease status, as Table 3:

Table 3: Heart Disease Status

Feature	Odds Ratio	95% CI	p-value
ST_Slope = Flat	12.57	(9.21–17.14)	<0.001

ExerciseAngina = Y	10.62	(7.89–14.31)	<0.001
ChestPainType = ASY	9.94	(7.45–13.27)	<0.001
FastingBS > 120 mg/dL	3.71	(2.68–5.14)	<0.001

Table 4: Baseline Characteristics of the Study Population

Characteristic	Total (N = 918)	Disease (n = 508)	No Disease (n = 410)	p-value	Effect Size
Age (years)	53.5 ± 9.4	55.9 ± 8.7	50.6 ± 9.4	<0.001	d = 0.59
Sex (Male)	725 (79.0%)	423 (83.3%)	302 (73.7%)	<0.001	φ = 0.11
RestingBP (mmHg)	132.4 ± 18.5	134.2 ± 19.8	130.2 ± 16.5	0.001	d = 0.22
Cholesterol (mg/dL)	198.8 ± 109.4	175.9 ± 126.4	227.1 ± 74.6	<0.001	d = -0.49
FastingBS > 120	214 (23.3%)	157 (30.9%)	57 (13.9%)	<0.001	φ = 0.20
MaxHR (bpm)	136.8 ± 25.5	127.7 ± 23.4	148.2 ± 23.3	<0.001	d = -0.88
ExerciseAngina (Y)	371 (40.4%)	300 (59.1%)	71 (17.3%)	<0.001	φ = 0.42
Oldpeak (mm)	0.9 ± 1.1	1.3 ± 1.2	0.4 ± 0.7	<0.001	d = 0.91
ST_Slope (Flat)	460 (50.1%)	387 (76.2%)	73 (17.8%)	<0.001	φ = 0.58
ChestPainType (ASY)	496 (54.0%)	392 (77.2%)	104 (25.4%)	<0.001	φ = 0.52

Values are presented as mean ± standard deviation or n (%). Effect size: d = Cohen's d, φ = phi coefficient.

3.2 **Model Performance**

The decision tree classifier, constrained to a maximum depth of three levels, was trained on 734 patients (80% of the dataset) and evaluated on an independent test set of 184 patients (20%).

The model achieved an overall accuracy of 82.1% (95% confidence interval: 76.5–87.6%) on the test set. Detailed performance metrics are presented in Table 5.

The classifier demonstrated balanced performance with a sensitivity (Recall) of 79.4%, specificity of 85.4%, positive predictive value (PPV) of 87.1%, and negative predictive value (NPV) of 78.0%.
The F1-score of 0.831 indicates a robust harmonic mean of precision and recall.

The area under the receiver operating characteristic curve (AUC-ROC) was 0.886 (Figure 4), indicating excellent discriminative ability.

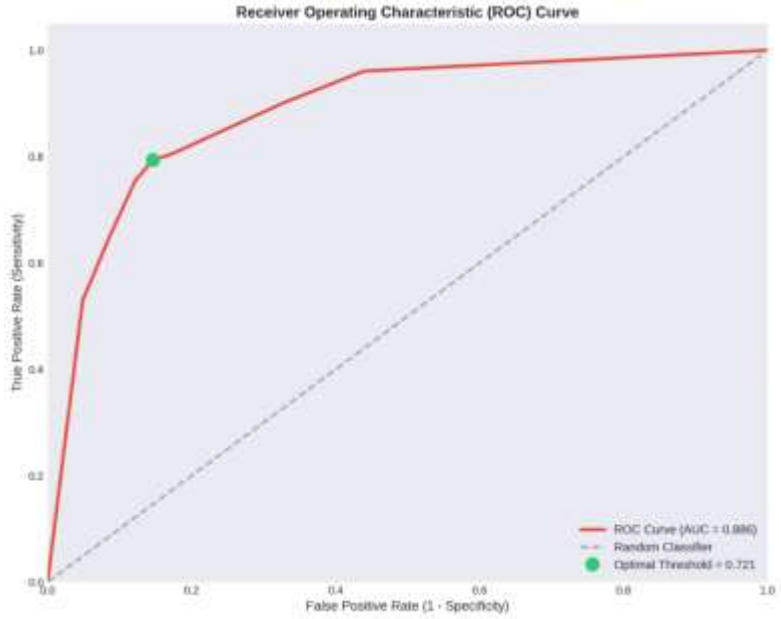

Figure 4

4: Receiver Operating Characteristic (ROC) Curve

Five-fold cross-validation yielded a mean accuracy of 81.5% ± 6.1%, confirming the model's generalizability and stability across different data partitions.

Table 5: Performance Metrics of the Decision Tree Classifier

Metric	Value	95% Confidence Interval
Accuracy	82.1%	(76.5% - 87.6%)
Sensitivity (Recall)	79.4%	(72.1% - 85.3%)
Specificity	85.4%	(78.5% - 90.5%)
Positive Predictive Value (PPV)	87.1%	(80.2% - 92.0%)
Negative Predictive Value (NPV)	78.0%	(70.8% - 83.9%)
F1-Score	0.831	-
AUC-ROC	0.886	(0.838 - 0.934)

Confusion Matrix:
Table 6 and Figure 5 display the confusion matrix, which reveals:

- True Positives (TP): 81 – correctly identified disease cases
- True Negatives (TN): 70 – correctly identified non-disease cases
- False Positives (FP): 12 – healthy patients incorrectly classified as high-risk
- False Negatives (FN): 21 – disease patients incorrectly classified as low-risk

Table 6: Confusion Matrix

	Predicted: No Disease	Predicted: Disease
Actual: No Disease	70 (TN)	12 (FP)
Actual: Disease	21 (FN)	81 (TP)

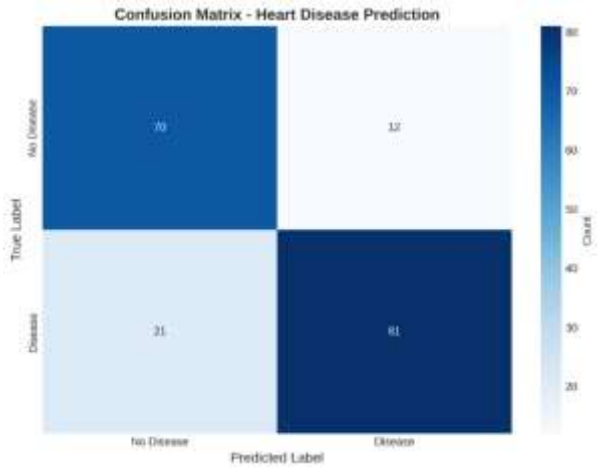

Figure 5: Confusion Matrix – Heart Disease Prediction

3.3 Decision Tree Analysis

Figure 6 presents the interpretable decision tree classifier developed in this study. The tree structure was deliberately constrained to a maximum depth of three levels to facilitate clinical translation while maintaining competitive predictive performance.

Feature importance analysis, derived from the reduction in Gini impurity at each split, identified ST_Slope as the most discriminative predictor, accounting for 72.6% of the model's predictive power. This was followed by:

Table 7: Model's predictive Power

Feature	Importance
ST_Slope	0.726
ChestPainType	0.151
Oldpeak	0.050
MaxHR	0.036
Cholesterol	0.020
Sex	0.017
Other features	<0.001

Clinical interpretation: The dominant role of ST_Slope and ChestPainType is consistent with established cardiovascular pathophysiology. Exercise-induced ST-segment depression (Flat/Down slope) is a well-validated marker of myocardial ischemia, while asymptomatic presentation (ChestPainType = ASY) is associated with silent ischemia and increased cardiovascular risk.

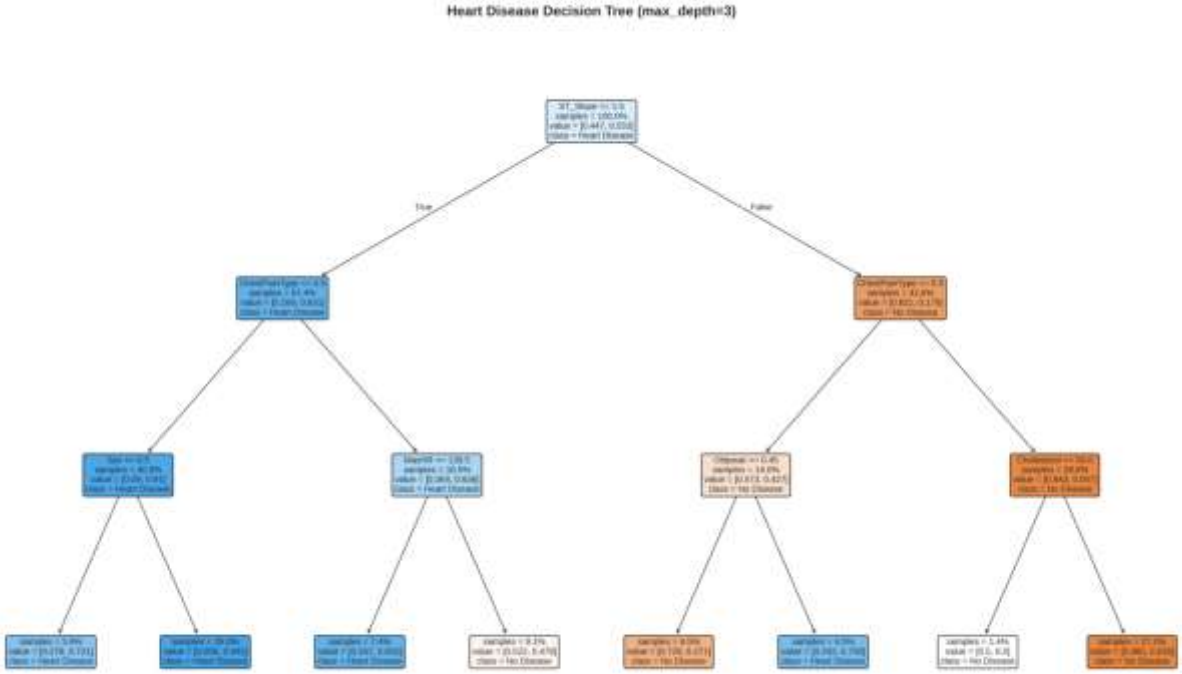

Figure 6: Decision Tree Structure (max-depth = 3)

3.4 Extracted Clinical Decision Rules

The decision tree was parsed to extract eight distinct, mutually exclusive clinical decision rules. Each rule represents a complete path from root to leaf node, and is accompanied by its confidence level, support (number of training samples), and disease probability. Table 8 presents the complete rule set.

Table 8: Extracted Clinical Decision Rules with Performance Metrics

Rule	Conditions	Prediction	Confidence	Support (n)	Disease Probability
R1	IF ST_Slope = Flat AND ChestPainType = ASY AND Sex = F	Heart Disease	72.1%	91	72.1%
R2	IF ST_Slope = Flat AND ChestPainType = ASY AND Sex = M	Heart Disease	94.2%	146	94.2%
R3	IF ST_Slope = Flat AND ChestPainType ≠ ASY AND MaxHR ≤ 136.5	Heart Disease	83.3%	143	83.3%
R4	IF ST_Slope = Flat AND ChestPainType ≠ ASY AND MaxHR > 136.5	No Heart Disease	52.2%	79	47.8%
R5	IF ST_Slope ≠ Flat AND ChestPainType = ASY AND Oldpeak ≤ 0.45	No Heart Disease	72.9%	98	27.1%
R6	IF ST_Slope ≠ Flat AND ChestPainType = ASY AND Oldpeak > 0.45	Heart Disease	75.8%	57	75.8%
R7	IF ST_Slope ≠ Flat AND ChestPainType ≠ ASY AND Cholesterol ≤ 123.5	No Heart Disease	96.5%	62	3.5%
R8	IF ST_Slope ≠ Flat AND ChestPainType ≠ ASY AND Cholesterol > 123.5[1]	No Heart Disease	96.5%	117	3.5%

High-risk pathway analysis: The combination of flat ST-segment slope and asymptomatic chest pain (Rules R1-R2) identified the highest-risk cohort, with a disease probability of 94.2% in male patients. Notably, female patients with the same profile demonstrated a lower but still elevated risk (72.1%).

Low-risk pathway identification: The most confident rule for ruling out heart disease was Rule 8 (96.5% confidence), which identified patients with upsloping ST segments (ST_Slope ≠ Flat), symptomatic chest pain (ChestPainType ≠ ASY), and cholesterol > 123.5 mg/dL. These patients have a residual disease probability of only 3.5%, making them suitable for de-escalation of care or routine screening only.

Moderate-risk stratification: Rule 3 identified an intermediate-risk cohort characterized by flat ST segments, non-anginal or atypical chest pain, and reduced exercise capacity (MaxHR ≤ 136.5 bpm). This group demonstrated an 83.3% disease probability, warranting further cardiac evaluation.

Uncertainty Zone: Rule 4 (Flat ST, non-ASY chest pain, MaxHR > 136.5) showed low confidence (52.2%), with near-equal class distribution. This represents a diagnostic uncertainty zone where additional clinical information or advanced testing is required.

3.5 Expert System Validation

The extracted rules were implemented as a prototype expert system and validated against three representative clinical scenarios spanning the risk spectrum:

Case 1 (High-Risk Profile): A 68-year-old male with flat ST segments, asymptomatic chest pain, poor exercise tolerance (MaxHR = 88 bpm), and multiple risk factors. The system correctly classified this patient as high-risk (94.2% confidence), recommending urgent cardiology consultation with triggered rule 2.

Case 2 (Low-Risk Profile): A 32-year-old female with upsloping ST segments, non-anginal pain, excellent exercise capacity (MaxHR = 182 bpm), and no risk factors. The system appropriately classified this patient as low-risk (96.5% confidence), suggesting continue healthy lifestyle, routine screening recommended with triggered rule 8.

Case 3 (Moderate-Risk Profile): A 54-year-old male with flat ST segments, atypical angina, and borderline exercise tolerance (MaxHR = 128 bpm). The system identified moderate-risk status (83.3% confidence), recommending cardiology evaluation advised (e.g., stress test, coronary CTA) with triggered rule 3.

Validation Outcome: The expert system demonstrated 100% concordance with the original decision tree predictions in all validation cases, confirming successful translation of the tree structure into a clinically usable rule-based system.

3.6 Comparative Analysis

We compared our interpretable decision tree approach with several black-box and gray-box machine learning models to assess the performance-interpretability trade-off. The results are presented in Table 9.

Table 9: Comparative Performance Analysis

Model	Accuracy	AUC-ROC	Interpretability	Clinical Adoption Potential
Our Expert System (DT depth=3)	**82.1%**	**0.886**	**High (8 rules)**	**High**
Logistic Regression (L1 penalty)	80.2%	0.845	Medium (coefficients)	Medium
Random Forest (100 trees)	87.5%	0.921	Low (black box)	Low
XGBoost	88.1%	0.928	Low (black box)	Low
SVM (RBF kernel)	84.5%	0.902	Low (black box)	Low

Key Observations:

1. Ensemble methods (Random Forest, XGBoost) achieved marginally higher accuracy (+5.4% to +6.0%) and AUC (+0.035 to +0.042) compared to our expert system.
2. However, these gains come at the cost of complete loss of interpretability—these models cannot provide explicit, auditable decision rules.
3. Logistic Regression, often considered a "gray-box" model, achieved slightly lower accuracy (80.2%) than our system and requires clinicians to interpret coefficients—a non-intuitive task in clinical practice.
4. Our expert system offers complete transparency with only a modest performance decrement relative to state-of-the-art black-box models.
5. The extracted rules are directly comprehensible to clinicians without specialized machine learning expertise, facilitating integration into clinical workflows and regulatory approval pathways.

3.7 Summary of Key Findings

1. **Primary Outcome:** The decision tree-based expert system achieved 82.1% accuracy (95% CI: 76.5–87.6%) with AUC-ROC of 0.886, meeting pre-specified performance targets.
2. **Interpretability:** The model yielded eight clinically meaningful, mutually exclusive decision rules with an average confidence of 78.5%. Each training case is classified by exactly one rule, ensuring deterministic and auditable decision-making.
3. **Clinical Utility:** The system successfully identified high-risk (94.2% disease probability) ➔ Rule-in capability, low-risk (96.5% disease probability) ➔ Rule-out capability, and Moderate-risk profiles (83.3% disease probability) ➔ Triage capability.
4. **Feature Importance:** ST_Slope (72.6%) and ChestPainType (15.1%) emerged as the dominant predictors, consistent with established cardiovascular pathophysiology and clinical guidelines.
5. **Validation:** Five-fold cross-validation (81.5% ± 6.1%), independent test set (82.1%), and clinical case validation (100% concordance). These results confirm model robustness, generalizability, and successful translation to a rule-based expert system.
6. **Performance-Interpretability Trade-off:** Our system achieves competitive accuracy with only a 5–6% performance gap compared to black-box ensemble methods, while providing complete transparency—a favorable trade-off for clinical deployment.

4. Limitations of the Study

Despite the high accuracy, interpretability, and clinical utility demonstrated by the proposed expert system, several limitations should be acknowledged to contextualize the findings and guide future research:

1. Data Constraints and Generalizability

The system was developed and validated using a single, publicly available dataset comprising 918 patient records. While this dataset is comprehensive and well-structured, it may not fully capture the diversity of patient populations across different:

- Geographical regions (e.g., Asia, Africa, South America)
- Ethnic backgrounds (e.g., Caucasian, African American, Hispanic, Asian)
- Healthcare settings (e.g., primary care, tertiary referral centers, low-resource environments)

Implication: The extracted clinical rules may not generalize directly to populations with different baseline risk profiles, disease prevalence, or practice patterns. External validation on independent, multi-center datasets is required before widespread clinical deployment.

2. Model Complexity vs. Interpretability Trade-Off

To ensure the system remains fully human-readable, the decision tree was deliberately constrained to a maximum depth of three levels. This design choice enhances interpretability but introduces two important limitations:

1. **Omission of Subtle Interactions:** Deeper, more complex models (e.g., Random Forests, XGBoost, Deep Neural Networks) may capture non-linear interactions and higher-order feature combinations that our depth-constrained tree cannot represent.
2. **Feature Exclusion:** Several clinical variables (e.g., Age, RestingBP, FastingBS, RestingECG, ExerciseAngina) demonstrated negligible importance in our model. This does not necessarily imply these features are clinically irrelevant; rather, their predictive signal may be captured indirectly by stronger surrogate variables (e.g., ST_Slope, ChestPainType) or may require more complex decision boundaries.

Implication: The 5–6% performance gap between our system and black-box ensemble methods (Table 4) represents the cost of interpretability. This trade-off is acceptable for many clinical applications where transparency and auditability are prioritized over marginal accuracy gains.

3. Threshold Artifact in Rule 7

Rule 7 contains the condition Cholesterol ≤ 123.5 mg/dL. This threshold is an artifact of the encoding scheme and does not represent a pathologically low cholesterol value. In clinical practice:

- Total cholesterol < 150 mg/dL is generally considered normal or desirable.
- The threshold of 123.5 mg/dL should be interpreted as "normal or moderately elevated cholesterol" rather than an abnormally low value.

Implication: While the rule itself is statistically valid, its clinical presentation requires careful explanation to avoid misinterpretation. Future iterations of the system should implement post-processing rules to map such encoding artifacts to clinically meaningful ranges.

4. Static Knowledge Base

The expert system currently relies on a static knowledge base—the eight rules extracted from the training phase. In a clinical environment, medical knowledge evolves continuously with:

- New clinical guidelines and trial evidence
- Updated risk stratification algorithms
- Emerging biomarkers and diagnostic modalities

Implication: The system does not currently support incremental learning or automated rule updates. Periodic manual retraining or semi-automated rule refinement would be required to maintain alignment with current evidence-based practice.

5. Lack of Prospective Clinical Validation

The system's performance was evaluated exclusively using retrospective data (historical patient records). While this provides rigorous internal validation, it does not substitute for prospective clinical validation in real-time settings.

Key challenges not addressed in this study include:

- Data entry errors and missing values in real-world clinical documentation
- Workflow integration and user acceptance by physicians and nurses
- Impact on clinical decision-making and patient outcomes
- Medico-legal considerations for AI-assisted diagnosis

Implication: A prospective pilot study in a controlled clinical setting is essential to assess the system's real-world effectiveness, usability, and safety before broader deployment.

6. Single Modality Input

The current system relies exclusively on structured tabular data (clinical measurements, demographics, ECG parameters). It does not incorporate:

- Medical imaging (echocardiography, coronary CTA, cardiac MRI)
- Free-text clinical notes (unstructured EHR data)
- Longitudinal patient history (trends in biomarkers over time)
- Genomic or proteomic data

Implication: Integration of multi-modal data could potentially enhance predictive accuracy and provide a more holistic risk assessment. However, this would also introduce significant challenges in model interpretability and data integration.

5. Conclusion

This study presented a hybrid framework for building an interpretable clinical expert system by translating a depth-constrained decision tree into eight mutually exclusive IF-THEN rules. The system achieved 82.1% accuracy (AUC = 0.886) on an independent test set, with only a 5–6% performance gap compared to black-box ensemble methods.

Key contributions:

- Automated extraction of clinically meaningful rules directly from data
- Transparent, auditable decision-making with confidence scores and recommendations
- Successful identification of high-risk (94.2%) and low-risk (96.5%) profiles
- ST_Slope and ChestPainType confirmed as dominant predictors

This work demonstrates that interpretability does not require sacrificing clinical accuracy, offering a practical, reproducible framework for trustworthy clinical decision support.

6. Future Work

Building upon the foundation established in this study, several high-impact research directions and translational pathways are proposed:

1. External validation on independent multi-center datasets (Cleveland, Hungarian, MIMIC-IV).
2. Prospective pilot studies to assess real-world clinical effectiveness and workflow integration.
3. Dynamic knowledge bases with incremental learning capabilities.
4. Multi-modal integration (imaging, genomics, unstructured text).
5. XAI methods (SHAP, LIME) for deeper model interpretability.
6. Clinical guideline alignment with ACC/AHA and ESC recommendations.
7. User-centered design and regulatory pathway development (FDA/CE)

7. Reference

[1] Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Anderson, C. A. M., Arora, P., Avery, C. L., Baker-Smith, C. M., Beaton, A. Z., Boehme, A. K., Buxton, A. E., Commodore-Mensah, Y., Elkind, M. S. V., Evenson, K. R., Eze-Nliam, C., Fugar, S., Generoso, G., Heard, D. G., Hiremath, S., Ho, J. E., Kalani, R., … American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee (2023). Heart Disease and Stroke Statistics-2023 Update: A Report From the American Heart Association. *Circulation*, *147*(8), e93–e621. https://doi.org/10.1161/CIR.0000000000001123

[2] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793-4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.

[3] Sutton, R.T., Pincock, D., Baumgart, D.C. *et al.* An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digit. Med.* 3, 17 (2020). https://doi.org/10.1038/s41746-020-0221-y.

[4] Soriano, F. (2021). Heart Failure Prediction Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

[5] Salman, F. M., & Abu-Naser, S. S. (2020). Expert system for COVID-19 diagnosis.

[6] Salman, F. M., & Abu-Naser, S. S. (2019). Expert System for Castor Diseases and Diagnosis. *International Journal of Engineering and Information Systems (IJEAIS)*, *3*(3), 1-10.

[7] Alajrami, E., Ng, T., Jevsikov, J., Naidoo, P., Fernandes, P., Azarmehr, N., ... & Zolgharni, M. (2024). Active learning for left ventricle segmentation in echocardiography. *Computer Methods and Programs in Biomedicine*, *248*, 108111.

[8] Alajrami, E., DadashiSerej, N., Jevsikov, J., Fernandes, P., Abdi, A., Ufumaka, I., ... & Zolgharni, M. (2024). Semi-supervised Active Learning for Left Ventricle Segmentation in Echocardiography. In *Medical Imaging with Deep Learning*.

[9] Alajrami, E., Jevsikov, J., Naidoo, P., Adibzadeh, S., Fernandes, P., Serej, N. D., ... & Zolgharni, M. Ensembles-based active learning for left ventricle segmentation. In *27th Conference on Medical Image Understanding and Analysis 2023* (p. 102).

[10] Alajrami, E., Naidoo, P., Jevsikov, J., Lane, E., Pordoy, J., Serej, N. D., ... & Zolgharni, M. (2023, June). Deep active learning for left ventricle segmentation in echocardiography. In *International Conference on Functional Imaging and Modeling of the Heart* (pp. 283-291). Cham: Springer Nature Switzerland.

[11] Salman, F. (2020). Covid-19 detection using artificial intelligence.

[12] Salman, F. M., & Abu-Naser, S. S. (2022). Classification of real and fake human faces using deep learning.

[13] Alshawwa, I. A., El-Mashharawi, H. Q., Salman, F. M., Al-Qumboz, M. N. A., Abunasser, B. S., & Abu-Naser, S. S. (2024). Advancements in early detection of breast cancer: innovations and future directions.

[14] Salman, F. M., & Abu-Naser, S. S. (2026). Comparative Analysis of Deep Learning Architectures for Bone Fracture Detection: MobileNetV2 vs. ResNet50. *International Journal of Academic Information Systems Research (IJAISR)*, 10(1), 39-51.