doi:10.22100/ijhs.v12i3.1351
Original Article

SJMS 2026;12(3):39-49
sjms.shmu.ac.ir

SJMS
Shahroud Journal of Medical Sciences

Providing an Evaluation Model for Medical Machine Learning in the Case of Heart Disease

Fatemeh Shahhosseiny[1], Kimia Zarooj Hosseini[2], Fatemeh Ahouz[3], Amin Gulabpour[4*]

[1] Master of Science, Department of Statistics, Allameh Tabatabaee University, Tehran, Iran.
[2] Student Research Committee, PhD Student in Medical Informatics, Department of Information Technology and Health Management, Faculty of Management and Medical Informatics, Iran University of Medical Sciences, Tehran, Iran.
[3] Department of Computer Engineering, Faculty of Energy and Data Sciences, Behbahan Khatam Alanbia University of Technology, Behbahan, Iran.
[4] Department of Health Informatics Technology, School of Allied Medical Sciences, Shahroud University of Medical Sciences, Shahroud, Iran.

Received: 2 September 2025
Accepted: 11 November 2025

Abstract

Background: *Cardiovascular diseases, the global number one killer, require early diagnosis to reduce premature mortality and enhance quality of life. Decision tree algorithms, whose transparency and credibility are highly valued, were used in order to capture intelligible diagnostic rules for the prediction of heart disease. They were validated and tested by doctors as clinically acceptable.*

Method: *This study experimented on a heart disease data set with statistical tests, splitting it 80:20 into training and test set for distribution studies. A decision tree method generated diagnostic rules from the training set, and PPV or NPV and Support for each rule were calculated. Rules with value less than threshold were removed, and the remaining rules were tested, recalculating PPV/NPV and Support. Non-compliant rules were removed, and clinicians reviewed final rules for clinical usability.*

Result: *This study statistically analyzed a heart disease dataset, splitting it 80:20 into training and test sets, with distributions validated. A decision tree algorithm generated diagnostic rules from the training set, assessed for positive predictive value (PPV) or negative predictive value (NPV) and Support. Rules below thresholds were discarded, and non-compliant adjusted rules were eliminated. Physicians evaluated the final rules for clinical acceptability.*

Conclusion: *This article highlights the crucial role of expert-based qualitative evaluation in validating and optimizing decision tree-induced rules. Optimization rules are accepted and satisfy more than original rules, as shown through comparisons of expert ratings. The findings underscore the necessity of model accuracy, interpretability, and clinical acceptability for the implementation of AI systems in health care.*

Keywords: Heart disease, Decision tree, Artificial intelligence.
***Corresponding to:** A Gulabpour, **Email:** a.golabpour@shmu.ac.ir

Please cite this paper as: Shahhosseiny F, Zarooj Hosseini K, Ahouz F, Gulabpour A. Providing an Evaluation Model for Medical Machine Learning in the Case of Heart Disease. Shahroud Journal of Medical Sciences 2026;12(3):39-49.

Introduction

Heart disease is frequently cited as the leading cause of death in numerous countries[1, 2]. According to the World Health Organization (WHO), cardiovascular disease is the leading cause of death worldwide. In 2019, an estimated 17.9 million people died from cardiovascular disease, accounting for 32% of all deaths globally. Additionally, 38% of premature deaths (under 70 years) caused by non-communicable diseases that year were attributed to cardiovascular disease[3]. According to the WHO, approximately 190 million people worldwide live with cardiovascular disease. Between 2006 and 2008, about 1.2 million deaths were reported in Iran, of which nearly 46% were attributed to cardiovascular disease[4]. Cardiovascular disease is expected to cause more than 23 million deaths worldwide by 2030[5].

Early diagnosis of heart disease is essential, as it helps prevent serious complications such as heart attacks, strokes, heart failure, and death[6]. Early recognition of those most at risk and ensuring they receive appropriate treatment can prevent early mortality. Research indicates that timely diagnosis and appropriate management of heart disease not only reduce death rates but also improve patients' quality of life and decrease healthcare costs[1].

In recent years, the use of artificial intelligence (AI) and machine learning (ML) techniques in disease diagnosis has attracted significant attention from researchers[2,7]. The rapid growth of healthcare data[3], and advances in high-performance computing have amplified the utility of AI in addressing clinical challenges[4]. AI, as an advanced data analysis and prediction tool, can efficiently process extensive volumes of medical information, identifying hidden patterns, and, in some cases, and, occasionally, contributing to diagnostic decision-making[8].

AI and ML methods can be classified into two categories: white-box and black-box models. Black-box algorithms often achieve high performance but lack interpretability. Although the research results are promising, the practical application of these models in real-world settings is limited due to several challenges[5]. These challenges include lack of reproducibility[6], as well as limited transparency and interpretability – issues commonly referred to as the medical black box problem[9-11]. In contrast, white-box algorithms generally offer lower predictive performance but are naturally transparent and interpretable. This allows users to examine how outputs are generated and how conclusions are drawn, facilitating more accurate and transparent decision-making[11]. Additionally, white-box models are auditable and accountable, meaning their accuracy and reliability can be evaluated, and they can be held responsible for their outputs[12]. Examples of white-box algorithms include decision trees, fuzzy logic systems, and rule-based models.

Decision tree algorithms are one of the most commonly used interpretable approaches in clinical practice and therefore were selected in this study as the foundation for rule extraction. Decision trees do have inherent weaknesses however: their greedy splitting approach can create rules with very low

coverage (i.e., low support). Splits at every internal node are made based on threshold tests for single variables, and leaf terminal nodes make final classifications—e.g., disease or no disease. Since each decision takes one and only one root-to-leaf path and the model averages predictions over all such paths, even rules covering just a small subset of the data can contribute to making up the rule set. To mitigate this problem, our method retrieves only rules with above some minimum coverage; high-coverage rules are then validated by doctors.

A review of related studies highlighted an innovative approach in one study[12], which proposed a hybrid classification model to predict heart disease using the Heart dataset. This model creatively combined decision trees with fuzzy systems to deliver accurate and understandable results. First, the researchers used a random forest algorithm to rank the independent variables based on their importance to the prediction task. The most significant variables were then selected to create diagnostic rules through a decision tree. These clear-cut rules were further refined by the fuzzy system, which generated weighted and interpretable rules. Ultimately, this hybrid model achieved an impressive accuracy of 90.5%.

In a study[12], researchers developed a classification model using weighted association rules to predict heart disease with the Heart dataset. They began by applying feature selection techniques to pinpoint the most relevant attributes. These carefully chosen features were then integrated into an association rule mining process, where weights were assigned based on each feature's frequency and importance across various algorithms. The resulting model delivered an impressive 98% confidence level in diagnosing heart disease.

In a study[14], researchers developed a classification model based on rough set theory to predict heart disease. They explored two innovative approaches: one used similarity measures among independent variables within the framework of classical rough sets, while the other applied a dominant attribute-based method within rough domain sets. These approaches were compared to several well-known ML algorithms, including random forest, logistic regression, naïve Bayes, and support vector machines (SVM), to evaluate their predictive performance. Among the rule-based models, the rough set–based algorithm stood out with superior performance. Meanwhile, among the conventional ML methods, SVM delivered the highest predictive accuracy.

When it comes to heart disease, catching it early and starting treatment quickly can truly make a life-saving difference. That's where AI steps in, with tools to help doctors improve patient outcomes. Knowing how important it is for doctors to understand and trust the decision-making process, researchers have embraced clear, transparent models like decision trees. These user-friendly models shine a light on how predictions are made, which is vital in a medical setting. With this goal in mind, researchers have poured their efforts into creating and testing models for heart disease diagnosis that are not only accurate but also easy for doctors to interpret, giving them confidence to act on these insights and ultimately save more lives.

Materials and Methods

In this study, researchers used a comprehensive heart disease dataset to build a predictive model that could help save lives. The dataset included 302 patient records, with 164 individuals diagnosed with heart disease and 138 without. It featured 13 key factors: five continuous numerical variables, like measurements that vary smoothly, and eight categorical variables, such as ordered or labeled traits. The main outcome tracked was whether or not someone had heart disease.

To make the most of decision tree algorithms, which can be sensitive to how variables relate to outcomes, the researchers calculated p-values for each factor to understand their connection to heart disease. They carefully identified variables with nonlinear relationships, ensuring the model captured the complexity of the data. For the statistical analysis, they used independent t-tests for the continuous variables and chi-square tests for the categorical ones, pinpointing which factors were most significant in predicting heart disease.

The heart disease dataset used in this study was obtained from Kaggle, where it had already undergone preliminary preprocessing and validation. Before analysis, the data were carefully inspected for missing values, inconsistencies, or duplicates. No missing or invalid entries were found, and all variables were confirmed to be within acceptable ranges. Therefore, no additional cleaning or imputation steps were required before normalization and modeling.

To get the data ready for the predictive models, the researchers carefully normalized all independent variables (or features) to a [0,1] range using the Min-Max normalization technique. They chose this approach because many ML algorithms can be thrown off by features with wildly different scales—those with larger numeric ranges might unfairly dominate the model's training, which could weaken its accuracy. By using Min-Max normalization, they ensured that each feature's values were smoothly scaled, mapping the smallest value to zero and the largest to one, while keeping the relative relationships between data points intact.

After normalization of the dataset, the researchers proceeded to compartmentalize the data into training versus testing group and assigned, 80% of the samples to be used for training and held the other 20% for testing model accuracy.

To make sure that the test and training sets were similar in their attributes, they compared their distributions on every feature of the dataset through the Kolmogorov–Smirnov test. They remixed the dataset and retried the test if they realized there were any significant differences. This was performed in cycles until the test and train groups did not have any significant differences, creating a valid and reliable setup for the model[11].

Next, we ran the decision tree algorithm on the training data and built a decision tree, deriving independent rules from each branch of the tree. Just a quick note: we used Python programming to create the decision tree. After that, we calculated two parameters—Positive Predictive Value (PPV) or Negative Predictive Value (NPV), and Support—for each rule. Then we selected the rules based on minimum thresholds for these parameters. (By the way, these minimum values differ depending on the type of disease and are set by a clinical expert.)

The selected rules were classified into two categories: positive and negative. After executing the first stage of this algorithm, if the obtained rules do not meet the minimum thresholds and the desired number of rules (as determined by specialists), the process is repeated until an enough number of rules satisfying these conditions is achieved. If, after repeating the algorithm, the conditions are still not met, the decision lies with the clinical specialist, who may choose to adjust the number of rules, modify the minimum threshold for the metrics, or abandon the process altogether. (Figure 1)

Inputs:

- A dataset containing N samples: $\{X_1,X_2,\ldots,X_N\}$ and their corresponding labels $Y=\{y_1,y_2,\ldots,y_N\}$.
- An initial IF–THEN rule of the form:
- IF (x_i op EV_i) AND (x_j op EV_j) THEN ClassLabel.

Outputs:

- A set of optimized IF–THEN rules, denoted as bSet.

Procedure:

1. Identify the minimum and maximum observed values of the variable x_i in the dataset and denote them as a and b, respectively.
2. Similarly, identify the minimum and maximum values of x_j, denoted as *c*and d.
3. Define the thresholds for rule evaluation:
   - Supt → minimum acceptable support,
   - EFFt → minimum acceptable efficiency.
4. Initialize the output rule set as empty:
5. bSet←∅.
6. For each possible value *v*within the range [a,b]:
   - For each possible value *w*within the range [c,d]:
     - Construct a modified rule $mRule_{v,w}$ as:
     - IF (xi op v) AND (xj op w) THEN ClassLabel.
     - Initialize the following counters for this rule:
       - $Support_{v,w}$: number of samples satisfying the rule's conditions.
       - $TP_{v,w}$: number of correctly classified samples.
       - $EFF_{v,w}$: efficiency of the rule.
7. For every sample $X_k$in the dataset:
   - If the conditions (x_i op v) and (x_j op w) are satisfied,
   - increase $Support_{v,w}$ by one and predict the class label for X_k as ClassLabel.
   - If the predicted label matches the true label y_k,
   - increase $TP_{v,w}$ by one.
8. After all samples are evaluated, compute the rule efficiency as:

$$EFF_{v,w}=\frac{TP_{v,w}}{Support_{v,w}}$$

9. If both conditions are satisfied:
   - $EFF_{v,w}>EFF_t$, and
   - $Support_{v,w}>Sup_t$,
   - Then include the corresponding rule $mRule_{v,w}$ in the output set:
   - $bSet=bSet\cup\{mRule_{v,w}\}$.
10. Repeat this process for all combinations of (v,w) to produce the final optimized rule set bSet.

Figure 1. Pseudocode for rule generation

The rules extracted from the first stage are not optimal, as the decision tree algorithm is greedy and does not guarantee the best possible outcome at the end of the process. Consequently, statistical evaluations become necessary. For these statistical evaluations, the most numerically favorable values were selected from the decision tree[12].

The rules derived from the first stage were evaluated across all possible scenarios. Given that each rule is characterized by two parameters—PPV or NPV, and Support—all possible states of each variable were examined. As a result, new rules were generated, each associated with updated PPV or NPV and Support values. Rules that did not meet the predefined minimum thresholds for PPV or NPV and Support were excluded from the rule set. (Figure 2)

Configuration:

Assume that the optimization process has generated K sets of rules,

$\{bSet_1, bSet_2, \ldots, bSet_K\}$.

Each rule in these sets is characterized by two metrics:

- Support (Sup) – representing the coverage of the rule,
- Efficiency (EFF) – representing its classification accuracy (e.g., PPV or NPV).

Initialize a global set S to collect all non-dominated rules:

$S \leftarrow \emptyset$.

Procedure for Finding Non-Dominated Rules:

1. For each optimized rule set $bSet_i$ (where i=1,2,…,K):
   - Let $R=bSet_i$, and let n be the total number of rules in R.
   - Define:
     - Sup_x: support of rule x
     - EFF_x: efficiency of rule x
   - Initialize an empty set of selected rules:
   - $pSet_i \leftarrow \emptyset$
2. For each rule a in R:
   - Initialize a counter C=0.
   - For every other rule $b \neq a$ in R:
     - If both $Sup_a \leq Sup_b$ and $EFF_a \leq EFF_b$, then increase the counter by one: C=C+1.
   - After all comparisons, if C=0, rule *a* is not dominated by any other rule.
   - In this case, add it to the set $pSet_i$: $pSet_i = pSet_i \cup \{rule_a\}$.
3. After evaluating all rules in $bSet_i$, add the resulting $pSet_i$ to the global set S:
4. $S = S \cup pSet_i$.
5. The final set *S* contains all non-dominated rules extracted from the K optimized rule sets.

Figure 2. Pseudocode for generating new rules

Furthermore, the output rules were evaluated, and for each of these new rules, the values of PPV or NPV and Support were calculated. Ultimately, among the new rules derived from each initial rule, the rule exhibiting the best overall performance based on these parameters was selected as the optimal rule.

In this study, AI algorithms were employed to extract rules relevant to the diagnosis of heart disease. The decision tree, utilizing specific parameter values as splitting criteria, generated a set of rules whose combination indicated the presence or absence of the disease. These values, which served as decision thresholds within the tree structure, were

subsequently utilized as thresholds in the new rules. Each sample was categorized into two groups below or above the threshold value for each parameter to enable the evaluation of these values' performance.

To further gauge the experts' thoughts on the input variables and the newly derived thresholds, we put together a questionnaire that included the extracted rules, ROC curves, and the new thresholds. We asked physicians to score the validity of these rules based on their clinical expertise, using a 0-to-5 scale, and to share their qualitative insights. The survey responses were checked for outliers and analyzed from both quantitative and qualitative angles. With 15 physicians participating, we used a two-tailed paired t-test to look for meaningful differences between the optimal and non-optimal rules. In the end, we examined how well the physicians' opinions aligned with the model's proposed rules to see if they backed the optimal ones or leaned more toward accepting the non-optimal ones.

To evaluate the designed model, the researchers calculated the parameters at three levels training, test, and overall using the relevant formulas and analyzed the results. (Table 1)

Table 1. Formulas for calculating evaluation parameters

Parameter	Calculation formula	Parameter	Calculation formula
Sensitivity	TP / (TP + FN)	**Specificity**	TN / (TN + FP)
Accuracy	(TP + TN) / (TP + FP + TN + FN)	**F-measure**	2 × PPV × Sensitivity / (PPV + Sensitivity)
PPV	TP / (TP + FP)	**NPV**	TN / (TN + FN)
LR+	Sensitivity / (1 - Specificity)	**LR-**	Specificity/ (1 - Sensitivity)
Kappa	[(TP+FN)(TP+FP) + (FN+TN)(TN+FP)] / (P + N)^2		
Phi	[(TP×TN) - (FP×FN)] / √[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]		

Results

After analyzing the data and calculating the p-value to assess the linearity and non-linearity of the variables, the results indicated that, due to the non-significant p-value for the FBS variable, it could not be included in the decision tree. Consequently, this variable was not incorporated into the decision tree rules. (Table 2)

The research team, with the input of experts, determined the minimum thresholds for the five parameters in accordance with Table 3.

Table 2. Descriptive statistics of heart disease data

Row	Variables	Data type	Levels	Heart disease N=164 (54%)	No heart disease N=138 (46%)	Statistical test	P-value
1	Sex	Nominal	Woman	71 (43%)	24 (17%)	Chi-squared	<0.01
			Man	93 (57%)	114 (83%)		
2	Type of chest pain	Ordinal	Typical angina	39 (24%	104 (75%)	Chi-squared	<0.01
			Uncommon angina	41 (25%)	9 (7%)		
			Non-angina pain	69 (42%)	18 (13%)		
			Without symptoms	16 (9%)	7 (5%)		
3	Fasting blood sugar more than 120 mg/dL	Nominal	No	141 (86%)	116 (84%)	Chi-squared	0.76
			Yes	23 (14%)	22 (16%)		
4	Results of resting electrocardiography	Ordinal	Normal	67 (41%)	79 (57%)	Chi-squared	<0.01
			Normality of the wave	96 (58.4%)	56 (41%)		
			Left ventricular hypertrophy	1 (0.6%)	3 (2%)		
5	Slope of the peak exercise ST segment	Ordinal	Inclined	9 (6%)	12 (9%)	Chi-squared	<0.001
			Flat	48 (30%)	91 (66%)		
			Downhill	107 (65%)	35 (25%)		
6	Number of main vessels	Ordinal	0	129 (79%)	45 (33%)	Chi-squared	<0.01
			1	21 (13%)	44 (32%)		
			2	7 (5%)	31 (22%)		
			3	3 (1%)	17 (12%)		
			4	4 (2%)	1 (1%)		
7	Thallium stress test	Ordinal	Error	1 (0.6%)	1 (0.7%)	Chi-squared	<0.01
			Elimination of defects	6 (3.4%)	12 (9%)		
			Normal	130 (80%)	36 (26%)		

8	Angina caused by exercise	Ordinal	Reversible defect No Yes	27 (16%) 141 (86%) 23 (14%)	89 (64.3%) 62 (45%) 76 (55%)	Chi-squared	<0.01

Row	Variables	Data type	Statistical index	Patient	Healthy	Statistical test	P-value
1	Age	Numerical	X̄±σ [Max, Min]	52.4±9.51 [29, 76]	56.6±7.96 [35, 75]	T-test	<0.01
2	Blood pressure at rest	Numerical	X̄±σ [Max, Min]	129.3±16.8 [94, 180]	134.3±18.72 [100, 200]	T-test	0.013
3	Cholesterol	Numerical	X̄±σ [Max, Min]	240.2±47.39 [126, 417]	251.08±49.45 [131, 409]	T-test	0.05
4	Maximum heart rate	Numerical	X̄±σ [Max, Min]	158.4±19.23 [96, 202]	139.1±22.59 [71, 195]	T-test	<0.01
5	Exercise-induced ST depression	Numerical	X̄±σ [Max, Min]	1.58±0.77 [0.0, 4.2]	1.58±1.3 [0.0, 6.2]	T-test	<0.01

Table 3. Algorithm parameters for heart disease data

Row	Parameters	Value
1	Minimum Support	20%
2	Minimum PPV	70%
3	Minimum NPV	70%
4	Number of rules for diagnosing heart disease	3
5	Number of rules for diagnosing no heart disease	2

To design the model, the data were initially divided into training and test sets, and the uniformity of data distribution was verified. Subsequently, the decision tree algorithm was applied to the data. All decision tree rules were evaluated based on the first three parameters listed in Table 2. However, to meet the minimum requirements for parameters 4 and 5, the division of training and test data was repeated three times, and a new decision tree was constructed each time. In the first iteration, out of four generated rules, two were deemed acceptable; in the second iteration, out of eight rules, two were acceptable; and in the third iteration, out of fifteen rules, one was acceptable according to the established criteria. Overall, 18.5% of the rules were acceptable, while 81.5% were deemed unacceptable. The extracted rules from the decision tree consist of several variables and their corresponding threshold values. (Table 4)

Now, considering the threshold of the decision tree rules, the rules are expressed as shown in Table 5.

The precise values of Support, PPV, and NPV for the derived rules are presented in Table 6.

Table 4. Threshold values of variables in the decision tree output

Row	Variable name	Classification
1	Type of chest pain	Low=typical angina High=uncommon angina, non-angina pain, without symptoms
2	Exercise-induced ST depression	Low≤2.35 High>2.35
3	Thallium stress test	Irreversible=error, elimination of defects, normal Reversible=reversible defect
4	Cholesterol	Low≤468 High>468
5	Maximum heart rate	Low≤91.043 High>91.043

Table 5. Rules resulting from the three stages of decision tree execution

Stage		Rules
First	A	If (type of chest pain=low) and (angina caused by exercise=yes) then no disease.
	B	If (type of chest pain=high) and (exercise-induced ST depression=low) then disease.
Second	C	If (type of chest pain=high) and (exercise-induced ST depression=low) and (thallium stress test=irreversible) then disease.
	D	If (type of chest pain=low) and (angina caused by exercise=yes) and (cholesterol=low) then no disease.
Third	E	If (type of chest pain=high) and (exercise-induced ST depression=low) and (thallium stress test=irreversible) and (maximum heart rate=high) then disease.

Table 6. Support, PPV, and NPV values for the rules

Row	Rules	All data		
		Support	NPV	PPV
1	A	27.91%	84.52%	-
2	D	27.24%	85.37%	-
3	B	51.16%	-	80.52%
4	C	51.83%	-	81.41%
5	E	49.16%	-	83.78%

Each derived rule was examined from various perspectives, considering all possible scenarios, and was transformed into a set of new rules. These new rules were evaluated based on two defined parameters (PPV or NPV and Support). Ultimately, from the various scenarios of each rule, the one exhibiting the best performance in terms of these two parameters was selected as the optimal rule. The output of this process was a set of optimized rules and new thresholds, presented in Tables 7 and 8.

Table 7. Determination of threshold values for variables of optimized rules

Row	Variable name	Classification
1	Type of chest pain	Typical=typical angina Uncommon=uncommon angina Non angina=non-angina pain, without symptoms
2	Exercise-induced ST depression	Low≤1.86 Normal 1.86<x≤3.72 High>3.72
3	Thallium stress test	Irreversible=error, elimination of defects, normal Reversible=reversible defect
4	Cholesterol	Low≤126 Normal 126<x≤520.2 High>520.2
5	Maximum heart rate	Low≤71 Normal 71<x≤110.3 High>110.3

Table 8. Rules derived from the execution of the proposed algorithm

Rule number	Rules
A1	If (type of chest pain=typical, uncommon) and (angina caused by exercise=yes) then no disease.
B1	If (type of chest pain=typical, uncommon) and (exercise-induced ST depression=low and normal) then disease.
C1	If (type of chest pain=typical) and (exercise-induced ST depression=low) and (thallium stress test=normal) then disease.
D1	If (type of chest pain=typical, uncommon) and (angina caused by exercise=no) and (cholesterol=normal) then no disease.
E1	If (type of chest pain=typical) and (exercise-induced ST depression=low) and (thallium stress test=normal) and (maximum heart rate=normal) then disease.

To qualitatively evaluate the rules extracted from the decision-making model, a questionnaire was designed in which each initial rule and its optimized versions were provided to 15 experts. They were asked to assign each rule a score ranging from 0 (complete disagreement) to 5 (complete agreement). To enhance the accuracy of the analysis, outlier scores were identified and removed using the Interquartile Range (IQR) method. The final results indicated that, in most cases, the optimized rules achieved higher average scores and garnered greater approval from the experts. For example, the average score for Rule A, after removing outliers, was 2.07, whereas its optimized version (A1) was evaluated at an average of 4.67. Similarly, Rule C, with an average score of 2.40, was improved

to its optimized version (C1) with an average of 3.47. Overall, four optimized rules—A1, C1, D1, and E1—successfully gained the approval of the majority of experts. However, Rule B2, even after optimization, remained unacceptable with a low average score of 0.33 and was discarded. These findings demonstrate that the optimization process not only improved the accuracy and logical coherence of the rules but also significantly enhanced their acceptance among experts. (Table 9)

Table 9. Average scores assigned by experts

Initial rule	Average expert scores	Optimal rule	Average expert scores
A	2.07	A1	4.67
B	0.33	B1	0.33
C	2.40	C1	3.47
D	1.07	D1	4.56
E	2.27	E1	4.60

To evaluate the effectiveness of optimized rules compared to initial rules, a statistical hypothesis test was defined as follows:

Null Hypothesis (H_0): There is no significant difference between the mean scores of initial rules and optimized rules.

Alternative Hypothesis (H_1): There is a significant difference between the mean scores of initial rules and optimized rules.

A two-tailed paired t-test was performed to assess the statistical difference between the scores given to initial rules and those assigned following optimization. The two-tailed paired t-test, therefore, tested difference in means for the assignments of each expert to an initial rule and its corresponding optimized version. It was found that in this case mean scores of optimized rules were significantly greater than mean scorers of initial rules: $t(2)=4.35$, P-value<0.05. Seeing as how the p-value fell below our threshold of 0.05, we were able to reject the null hypothesis. Thus, we reached a conclusion that experts' evaluative differences in scores between benchmark and refined rules are statistically significant; furthermore, it is evident that optimization directly enhanced both acceptance and quality of the rules.

The results derived from the analysis of evaluation parameters, as presented in Table 10, indicate the high performance of the model.

The conformity of the data with the derived rules indicates a high coverage of the obtained rules. (Table 11)

Table 10. Evaluation parameters

	Sensitivity	Specificity	Accuracy	PPV	NPV	LR+	LR-	Kappa P-value	Phi P-value	F-measure
Total	0.86	0.88	0.87	0.86	0.88	6.37	0.14	0.72 0.000	0.725 0.000	0.86
Train	0.96	0.81	0.90	0.88	0.94	5.08	0.05	0.79 0.000	0.793 0.000	0.92
Test	0.78	0.67	0.74	0.81	0.63	2.33	0.33	0.44 0.004	0.439 0.004	0.79

Table 11. Percentage of rule coverage

	Number	Percentage
Total	220.301	73.08%
Train	178.240	74.16%
Test	42.61	68.85%

Discussion

In this study, a team of researchers poured their hearts into building a model to help doctors diagnose heart disease, making sure the model's decision rules were carefully reviewed by experts. Decision trees are great tools, but they can sometimes create a lot of branches[15, 16], leading to rules that don't cover enough ground to be truly helpful. The low coverage level of the rules reduces the performance of the clinical rules, thus reducing the comprehensiveness of the rules[17] and the usability of the model. Therefore, setting minimums for the coverage level and threshold values for PPV or NPV can produce better and more comprehensive rules.

In this study, a decision tree algorithm was used to diagnose heart disease. The implemented tree has 81.5% of rules with low coverage and 18.5% with high coverage, which indicates a high number of rules generated with low coverage.

The criteria for determining the minimum coverage level and the threshold of PPV or NPV are the use of expert opinions using the focus group method[18] and Delphi[19] and other sources, which varies according to the type of disease. For example, the accuracy of the model determined for breast cancer is between 90.86% and 97.53%[20-22], diabetes between 75.57% and 88%[23-25], and stomach cancer between 86.8% and 96.20%[26-28].

AI represents an extensive range of machine-learning techniques and uses clinical data to identify patterns, often trivial in nature, related to disease from which it generates accurate diagnostic models[29]. However, one of the main problems in the clinical acceptance of AI algorithms is the low transparency of black box models, which causes doctors to distrust their decisions. In terms of AI models, the generated rules must also be validated for cost, feasibility, operational efficiency, and real-world integration. In contrast, transparent white-box models, such as those derived from decision trees, produce clear diagnostic rules that can be validated through expert review. This process allows expert-derived rules to gain clinical acceptance, fostering trust and ultimately making it much easier for future integration of AI into clinical practice.

AI systems must undergo clinical assessments before they can be integrated into medical practice. An AI model's strong statistical performance—be it accuracy, sensitivity, or AUROC—does not guarantee its usefulness or safety in a practical clinical environment[30]. Validation through statistics alone is problematic because it often evaluates performance on internal or dated datasets from prior to the present which fail to capture real-world patient population diversity[31]. These algorithms should not be deployed in prospective real-world settings as decision-support systems since doing so would mean using unvalidated and risky integrations into patient care[32].

The clinical rules derived in this study indicate the presence or absence of the disease. For example, if variables such as the type of chest pain and low cholesterol levels are present, the likelihood of heart disease is low. This finding is consistent with the studies conducted by Phillips et al.[33] and Deo et al.[34]. In addition, a negative response to the exercise test also suggests a low probability of heart disease, aligning with the research by Akyuz et al.[35]. Conversely, an increased heart rate increases the likelihood of heart disease, which is in agreement with the findings of Nankhen et al.[36]. In addition, changes in the ST segment, such as depression from the isoelectric line in response to the exercise test, suggest a higher risk of heart disease, corroborating the findings of Fitzgerald et al.[37]. In this study, a positive thallium scan test indicates a higher probability of heart disease, which is consistent with the research by Blumenthal et al.[38].

The rules generated by the white-box algorithm used in this study were transformed into optimized rules, with new threshold values established for them. These rules were subsequently evaluated with the input of specialist physicians. The results indicate that the optimized rules were validated by the specialists, demonstrating the high performance and comprehensiveness of the output rules.

These results suggest that the use of white-box algorithms combined with clinical evaluation can lead to the generation of more reliable and acceptable rules. Consequently, the proposed model in this study represents a step toward enhancing the practical application of ML algorithms in the field of healthcare and the diagnosis of heart diseases.

In the realm of heart disease prediction, several innovative approaches have been proposed, all of which emphasize enhancing the accuracy and interpretability of the diagnosis through advanced ML techniques. A notable study by Yazdani et al. (2021) introduced an algorithm based on Weighted Associative Rule Mining (WARM) to improve heart disease prediction by calculating the strength of significant predictors. Using the UCI heart disease dataset, they achieved an impressive 98% confidence score, prioritizing the most influential features with weighted scores. This approach emphasizes the importance of identifying feature strength over mere significance, aligning with our own study's focus on refining prediction models through feature evaluation and weighted scoring[39].

Similarly, Singh et al. (2022) utilized Rough Set Theory for heart disease prediction, applying both Classical Rough Set Approach (CRSA) and Dominance-based Rough Set Analysis (DRSA) on the Cleveland heart disease dataset. Their work highlighted the advantages of rule-based methods, even though SVM performed slightly better in some cases. The rough set-based approach, however, provided clearer, more interpretable decision rules, which were more easily understood by healthcare practitioners. Their study underlines the significance of handling data inconsistencies through rule-based methodologies, a perspective that aligns well with our approach to improving predictive accuracy[14].

In another related study, Mokeddem et al. (2017) proposed a Fuzzy Classification Model for assessing the risk of Myocardial Infarction (MI), combining Random Forest, C5.0 decision tree, and fuzzy modeling. They tackled the common problem of missing data by introducing a method to handle incomplete information, achieving an accuracy of 90.50% on the UCI heart disease datasets. Their fuzzy logic-based model not only improved the handling of uncertainty and imprecision in medical data but also emphasized feature ranking and missing value handling, mirroring the objectives of our study in making decision support systems more transparent and reliable[40].

In this study, the results obtained from the proposed decision tree model in predicting heart disease demonstrate a high level of accuracy compared to traditional methods and other existing models. The decision tree algorithm effectively handled issues related to missing data and inconsistencies in medical data, providing clear and interpretable decision rules for healthcare professionals. One of the key differences in our work is that, unlike many similar studies, the evaluation of the generated rules was conducted with the input of medical specialists. This ensured that the rules were clinically relevant and practical, making the decision support system more reliable and applicable in real-world clinical settings. Ultimately, this research highlights the potential of decision tree-based models in improving the diagnostic process for heart diseases and emphasizes the importance of expert validation in the development of such models.

Limitations: Due to the greedy nature of the decision tree algorithm, the parameters for variable selection and operators

are not optimized, and the proposed model lacks the capability to optimize these parameters.

Future Work: It is recommended that future research explore other rule-extraction algorithms, such as fuzzy methods. Additionally, it is suggested that other decision tree parameters, similar to those in the proposed model, be optimized.

Ethical Considerations

This study was conducted using a publicly available heart disease dataset obtained from Kaggle. The dataset contained anonymized patient records and did not include any personally identifiable information. Therefore, no direct interaction with participants was involved, and informed consent was not required. The study complied with the principles of confidentiality, privacy protection, and responsible use of artificial intelligence in healthcare research.

Acknowledgment

The authors would like to express their sincere gratitude to the physicians and clinical experts who participated in the evaluation of the extracted diagnostic rules and provided valuable feedback throughout this study. The authors would further like to thank t Research and Technology of the Shahroud University of Medical Sciences for supporting this research.

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article. The authors have no financial, professional, or personal relationships that could have influenced the work reported in this study.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Oude Wolcherink MJ, Behr CM, Pouwels X, Doggen CJM, Koffijberg H. Health Economic Research Assessing the Value of Early Detection of Cardiovascular Disease: A Systematic Review. PharmacoEconomics. 2023;41(10):1183-203. doi: 10.1007/s40273-023-01287-2

2. Sahakyan M, Aung Z, Rahwan T. Explainable Artificial Intelligence for Tabular Data: A Survey. Ieee Access. 2021; 9:135392-422. doi: 10.1109/ACCESS.2021.3116481

3. Josephson CB, Wiebe S. Precision Medicine: Academic dreaming or clinical reality? Epilepsia. 2021;62: S78-S89. doi: 10.1111/epi.16739

4. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. The Lancet. 2020;395(10236):1579-86. doi: 10.1016/S0140-6736(20)30226-9

5. Peng JF, Zou KQ, Zhou M, Teng Y, Zhu XY, Zhang FF, et al. An Explainable Artificial Intelligence Framework for the Deterioration Risk Prediction of Hepatitis Patients. Journal of Medical Systems. 2021;45 (5). doi: 10.1007/s10916-021-01736-5

6. Parziale A, Senatore R, Della Cioppa A, Marcelli A. Cartesian genetic programming for diagnosis of Parkinson disease through handwriting analysis: Performance vs. interpretability issues. Artificial Intelligence in Medicine.2021;111. doi: 10.1016/j.artmed.2020.101984

7. Torshizi R, Karimani EG, Etminani K, Akbarin MM, Jamialahmadi K, Shirdel A, Rahimi H, Allahyari A, Golabpour A, Rafatpanah H. Altered expression of cell cycle regulators in adult T-cell leukemia/lymphoma patients. Reports of Biochemistry and Molecular Biology. 2017;6(1):88.

8. Golabpour A, Shirazi HM, Farahi A, Kootiani AZ, Beigi H. A fuzzy solution based on Memetic algorithms for timetabling. In 2008 International Conference on MultiMedia and Information Technology 2008 (pp. 108-110). IEEE. doi: 10.1109/MMIT.2008.193

9.Wilkinson J, Arnold KF, Murray EJ, van Smeden M, Carr K, Sippy R, et al. Time to reality check the promises of machine learning-powered precision medicine. The Lancet Digital Health. 2020;2(12): e677-e80. doi: 10.1016/S2589-7500(20)30200-4

10. Petch J, Di S, Nelson W. Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. Canadian Journal of Cardiology. 2022;38(2):204-13. doi: 10.1016/j.cjca.2021.09.004

11. Lee CK, Samad M, Hofer I, Cannesson M, Baldi P. Development and validation of an interpretable neural network for prediction of postoperative in-hospital mortality. Npj Digital Medicine. 2021;4 (1). doi: 10.1038/s41746-020-00377-1

12. Welchowski T, Maloney KO, Mitchell R, Schmid M. Techniques to Improve Ecological Interpretability of Black-Box Machine Learning Models. Journal of Agricultural Biological and Environmental Statistics. 2022;27(1):175-97. doi: 10.1007/s13253-021-00479-7

13. Zihni E, Madai VI, Livne M, Galinovic I, Khalil AA, Fiebach JB, et al. Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. PLoS ONE. 2020;15 (4). doi: 10.1371/journal.pone.0231166

14. Singh A, Misra SC, Kumar S. Smart healthcare: rough set theory in predicting heart disease. Advances in Computing, Informatics, Networking and Cybersecurity: A Book Honoring Professor Mohammad S Obaidat's Significant Scientific Contributions: Springer; 2022. p. 155-80. doi: 10.1007/978-3-030-87049-2_5

15. Suthaharan S, Decision Tree Learning. In: Machine Learning Models and Algorithms for Big Data Classification. Integrated Series in Information Systems. 2016:237-69. doi: 10.1007/978-1-4899-7641-3_10

16. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry. 2015;27 (2):13-5.

17. Kannan A, Fries JA, Kramer E, Chen JJ, Shah N, Amatriain X. The accuracy vs. coverage trade-off in patient-facing diagnosis models. AMIA Joint Summits on Translational Science proceedings AMIA Joint Summits on Translational Science. 2020; 2020:298 307.

18. Gundumogula M, Gundumogula MJIJoH, Science S. Importance of focus groups in qualitative research. 2020;8(11):299-302. doi: 10.24940/theijhss/2020/v8/i11/HS2011-082

19. Cuhls K. The Delphi method: an introduction. Delphi methods in the social and health sciences: concepts, applications and case studies: Springer; 2023. p. 3-27. doi: 10.1007/978-3-658-38862-1_1

20. Rai R, Sisodia DS, editors. Real-time data augmentation based transfer learning model for breast cancer diagnosis using histopathological images. Advances in Biomedical Engineering and Technology: Select Proceedings of ICBEST 2018; 2021: Springer. doi: 10.1007/978-981-15-6329-4_39

21. Das AK, Biswas SK, Mandal A, Bhattacharya A, Sanyal SJESwA. Machine Learning based Intelligent System for Breast Cancer Prediction (MLISBCP). 2024; 242:122673. doi: 10.1016/j.eswa.2023.122673

22. Alqudah A, Alqudah AMJIJoR. Sliding window based support vector machine system for classification of breast cancer using histopathological microscopic images. 2022;68(1):59-67. doi: 10.1080/03772063.2019.1583610

23. Agliata A, Giordano D, Bardozzo F, Bottiglieri S, Facchiano A, Tagliaferri RJIJoMS. Machine learning as a support for the diagnosis of type 2 diabetes. 2023;24(7):6775. doi: 10.3390/ijms24076775

24. Pal M, Parija S, Panda G, editors. Improved prediction of diabetes mellitus using machine learning based approach. 2021 2nd International Conference on Range Technology (ICORT); 2021: IEEE. doi: 10.1109/ICORT52730.2021.9581774

25. James DE, Vimina E, editors. Machine learning-based early diabetes prediction. Intelligent Sustainable Systems: Proceedings of ICISS 2021; 2022: Springer. doi: 10.1007/978-981-16-2422-3_52

26. Fan Z, Guo Y, Gu X, Huang R, Miao WJSR. Development and validation of an artificial neural network model for non-invasive gastric cancer screening and diagnosis. 2022;12(1):21795. doi: 10.1038/s41598-022-26477-4

27. Zahmatkesh Zakariaee A, Sadr H, Yamaghani MR. A New Hybrid Method to Detect Risk of Gastric Cancer using Machine Learning Techniques. Journal of AI and Data Mining. 2023;11 (4):505-15.

28. Li C, Liu S, Zhang Q, Wan D, Shen R, Wang Z, et al. Combining Raman spectroscopy and machine learning to assist early diagnosis of gastric cancer. 2023; 287:122049. doi: 10.1016/j.saa.2022.122049

29. Almasinejad P, Golabpour A, Mollakhalili Meybodi MR, Mirzaie K, Khosravi A. A dynamic model for imputing missing medical data: a multiobjective particle swarm optimization algorithm. Journal of Healthcare Engineering. 2021;2021(1):1203726. doi: 10.1155/2021/1203726

30. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Publisher Correction: Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature medicine. 2022;28(10):2218. doi: 10.1038/s41591-022-01951-8

31. Hogg HDJ, Martindale APL, Liu X, Denniston AK. Clinical Evaluation of Artificial Intelligence-Enabled Interventions. Investigative Ophthalmology and Visual Science. 2024;65(10):10. doi: 10.1167/iovs.65.10.10

32. Bin-Jumah MN, Al-Abdan M, Al-Basher G, Alarifi S. Molecular Mechanism of Cytotoxicity, Genotoxicity, and Anticancer Potential of Green Gold Nanoparticles on Human Liver Normal and Cancerous Cells. Dose-response: a publication of International Hormesis Society. 2020;18(2):1559325820912154. doi: 10.1177/1559325820912154

33. Laureano-Phillips J, Robinson RD, Aryal S, Blair S, Wilson D, Boyd K, et al. HEART score risk stratification of low-risk chest pain patients in the emergency department: a systematic review and meta-analysis. 2019;74(2):187-203. doi: 10.1016/j.annemergmed.2018.12.010

34. Doi T, Langsted A, Nordestgaard BGJJotACoC. Elevated remnant cholesterol reclassifies risk of ischemic heart disease and myocardial infarction. 2022;79(24):2383-97. doi: 10.1016/j.jacc.2022.03.384

35. Hermiz C, Sedhai YR. Angina. 2020.

36. Nanchen DJH. Resting heart rate: what is normal? BMJ Publishing Group Ltd and British Cardiovascular Society; 2018. p. 1048-9. doi: 10.1136/heartjnl-2017-312731

37. Fitzgerald BT, Smith E, Scalia GM. What are the prognostic implications and factors relating to exercise induced electrocardiographic ST segment changes in the setting of a non-ischemic stress echocardiogram? 2022; 364:157-61. doi: 10.1016/j.ijcard.2022.06.031

38. Blumenthal RS, Becker DM, Moy TF, Coresh J, Wilder LB, Becker LCJC. Exercise thallium tomography predicts future clinically manifest coronary heart disease in a high-risk asymptomatic population. 1996;93(5):915-23. doi: 10.1161/01.CIR.93.5.915

39. Yazdani A, Varathan KD, Chiam YK, Malik AW, Wan Ahmad WAJBmi, making d. A novel approach for heart disease prediction using strength scores with significant predictors. 2021;21(1):194. doi: 10.1186/s12911-021-01527-5

40. Mokeddem SAJAI. A fuzzy classification model for myocardial infarction risk assessment. 2018;48(5):1233-50.