Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 16(1), 80-95, 2026
Journal of the Institute of Science and Technology, 16(1), 80-95, 2026

ISSN: 2146-0574, eISSN: 2536-4618

Biology **DOI: 10.21597/jist.1836750**

Research Article

Received: 05.12.2025 **Accepted: 05.02.2026**

To Cite: Huyut, M. T., & Velichko, A. (2026). Stability-Bound Binary Rule Search: A General Workflow for Explainable Prediction on Binarized Clinical Data. *Journal of the Institute of Science and Technology*, 16(1), 80-95.

Stability-Bound Binary Rule Search: A General Workflow for Explainable Prediction on Binarized Clinical Data

Mehmet Tahir HUYUT[1*], Andrei Velichko[2]

Highlights:

- Three-variable rules provided high accuracy and strong clinical interpretability.
- In the cardiac data, CP, CA_0, and thallium findings were consistently prominent in the model.
- In the hepatitis C data, AST-centered rules demonstrated high discriminatory power and stability.

Keywords:

- Explainable Clinical Prediction
- Binary Rule Search
- Stability-Bound Rule Score Binarized Clinical Data
- Interpretable Decision Rules

ABSTRACT:

Explainable prediction is increasingly required in clinical decision support, especially when models must generalize across institutions. We present a stability-bound binary rule search workflow that operates on fully binarized clinical data and expresses decisions as sparse, human-readable rules. Clinical variables are converted into 0/1 indicators using clinically meaningful thresholds, so that each rule corresponds to a binary mask over a small set of interpretable features. A Binary Rule Search (BRS) engine explores conjunctions of up to four predictors (k=1–4), and candidate rules are evaluated by the Matthews-correlation-coefficient (MCC) on development and validation splits. Robustness is summarized by the Stability-Bound-Rule-Score (SBRS), a geometric-style combination of the lower 95% confidence bounds of MCC in both splits. The workflow was applied to two open-access datasets: a heart attack dataset (303 patients) and a hepatitis C dataset (615 patients). In the heart attack data, a four-feature rule combining age 55–64 years, typical chest pain, absence of angiographically stenosed vessels (CA = 0) and a reversible thallium perfusion defect achieved MCC 0.71 and 0.73 in the development and validation sets, with SBRS = 1.59. In the hepatitis C data, rules built from elevated aspartate aminotransferase together with intermediate or high alkaline phosphatase and increased bilirubin reached MCC 0.75 and 0.84, with SBRS = 1.67. Because all predictors are binarized, the final rules can be displayed as compact binary mask plots or implemented as short checklists and look-up tables. Overall, this stability-bound binary rule search workflow yields sparse, stable and clinically interpretable rule sets for cardiovascular risk stratification and chronic liver disease screening.

[1]Mehmet Tahir HUYUT (**Orcid ID:** 0000-0002-2564-991X), Department of Biostatistics and Medical Informatics, Faculty of Medicine, Erzincan Binali Yıldırım University, 24000 Erzincan, Türkiye
[2]Andrei VELİCHKO (Orcid ID: 0000-0001-8760-316X), Petrozavodsk State University, 33 Lenin Ave., 185910 Petrozavodsk, Russia

***Corresponding Author:** Mehmet Tahir HUYUT, e-mail: tahir.huyut@erzincan.edu.tr

Ethics Committee Approval: All datasets used in the analyses are obtained from anonymized, open-access sources, and do not contain any personally identifiable information. Therefore, the study does not inherently involve intervention with human participants or processing of personal data and does not require ethics committee approval as required by national and international ethics committees.

INTRODUCTION

The need for predictive models that are both reliable and easy to understand is becoming increasingly evident in clinical research. While the high accuracy offered by modern machine learning methods is a clear advantage, clinicians still require tools that are "known to work" and whose behaviour can be inspected in routine decision-making. Traditional scoring systems partially satisfy this need by providing transparent risk estimates, but they typically incorporate only a limited number of variables, do not capture higher-order interactions, and may show substantial performance degradation when applied outside the development cohort (Steyerberg, 2019). In contrast, more complex methods such as gradient-boosted ensembles or deep learning can achieve excellent accuracy, yet often fall short of the level of explainability and portability required to establish trust in high-stakes clinical settings (Caruana et al., 2015; Rudin, 2019). This persistent imbalance between performance and interpretability motivates the development of new methodological frameworks.

The structure of clinical data further complicates this problem. Real-world datasets usually contain heterogeneous measurements—dichotomous, categorical, continuous, and ordinal—and substantial inter-institutional variation in how variables are recorded. Many predictors, however, can be meaningfully converted into binary indicators using clinically motivated thresholds, such as age ≥65 years, crossing a laboratory reference limit, or coding an imaging finding as present versus absent (Harrell, 2015). Although binarization is sometimes viewed as an oversimplification, it can in fact harmonize variables across datasets, sharpen the investigation of interactions, and naturally align model outputs with "if–then" decision rules. Nonetheless, few existing approaches are designed to operate natively on binarized inputs while explicitly searching over multi-variable interactions in a transparent manner.

To address this gap, we propose an explainable rule discovery workflow based on stability-bound binary rule search applied to fully binarized clinical data. At its core lies a Binary Rule Search (BRS) engine that systematically scans combinations of binary predictors and their corresponding bit patterns, yielding sparse conjunctions that can be read as human-interpretable rules. Each candidate rule is evaluated by the Matthews correlation coefficient (MCC) on separate development and validation subsets. Robustness to sampling variability is captured by a Stability-Bound Rule Score (SBRS), which aggregates the lower 95% confidence bounds of MCC across splits into a single conservative criterion. This stability-focused perspective follows the broader principles of resampling-based, veridical data analysis, where patterns are prioritised only when they reappear consistently across perturbed versions of the data (Yu & Kumbier, 2020).

Overall, the proposed stability-bound binary rule search workflow aims to generate sparse, clinically usable rules that maintain much of the predictive performance of more complex models without sacrificing interpretability. By operating on binarized data and expressing decisions as compact sets of threshold-based rules, the approach is designed to yield models that clinicians can readily interpret, that are reproducible across settings, and that are methodologically aligned with current expectations for transparent and accountable AI in medicine.

MATERIALS AND METHODS

Study Design and Datasets

This study was designed to test the performance of a stability-bound binary rule search workflow built on the Binary Rule Search (BRS) engine, applied to binarized clinical data. In this context, the datasets required for heart attack risk and hepatitis C disease prediction were obtained from the open-

access "Kaggle" database. The "Heart attack risk" dataset, consisting of 13 characteristics of 303 patients, is summarized in Table 1, and the "Hepatitis C" dataset, consisting of 12 characteristics of 615 patients, is summarized in Table 2. Data addresses are indicated in the table headings. This study aims to evaluate the performance of a statistical and methodological approach that does not involve any primary data collection. All datasets used in the analyses are obtained from anonymized, open-access sources, and do not contain any personally identifiable information. Therefore, the study does not inherently involve intervention with human participants or processing of personal data and does not require ethics committee approval as required by national and international ethics committees.

Table 1. Heart Attack Dataset (https://www.kaggle.com/datasets/pritsheta/heart-attack).

No	Features	Explanation
1	Age	Participant's age (years).
2	Gender	1 = male; 0 = female
3	Chest Pain	Four categories: Type of Chest Pain -- 1: typical angina (all criteria present) -- 2: atypical angina (two of three criteria met) -- 3: non-anginal pain (less than one criterion met) -- 4: asymptomatic (none of the criteria met)
4	Resting Blood Pressure	Resting Blood Pressure (in mmHg, on admission)
5	Cholesterol	Serum cholesterol in mg/dL
6	Fbs	Fasting blood sugar > 120 mg/dL (likely diabetic) 1 = true; 0 = false
7	RestECG	Resting electrocardiogram results. Three categories are available. Value 0: normal; Value 1: ST-T wave abnormality (T wave inversions and/or ST elevation or depression > 0.05 mV); Value 2: Indicates probable or definite left ventricular hypertrophy according to Estes criteria.
8	MaxHR	The highest number of beats per minute your heart can reach during strenuous exercise at full strength.
9	Exang	Exercise-induced angina (1 = yes; 0 = no)
10	Oldpeak	ST depression compared to rest with exercise (in mm, obtained by subtracting the lowest ST segment points during exercise and rest).
11	Slope	The slope of the peak exercise ST segment. ST-T abnormalities are considered an important indicator of the presence of ischemia. Three categories are available. Value 1: Upward sloping, Value 2: Straight, Value 3: Downward sloping
12	Ca	Number of major vessels colored by fluoroscopy (0-3). The major cardiac vessels are: aorta, superior vena cava, inferior vena cava, pulmonary artery (oxygen-poor blood to the lungs), pulmonary veins (oxygen-rich blood to the heart), and coronary arteries (supply blood to the heart tissue).
13	HEART ATTACK	Three categories: 0 = normal; 1 = fixed defect (heart tissue cannot absorb thallium both under stress and at rest); 2 = reversible defect (heart tissue cannot absorb thallium only during the exercise portion of the test).
14	Diagnostic classification	0 = no disease, 1 = disease present

Table 2. Hepatitis C Dataset (https://www.kaggle.com/datasets/fedesoriano/hepatitis-c-dataset)

No	Features	Explanation
1	Age	Participant's age (years).
2	Gender	1 = male; 0 = female
3	ALB	Albumin (Major plasma protein synthesized in the liver; low levels may indicate hepatic dysfunction).
4	ALP	Alkaline Phosphatase (an enzyme increased in liver and biliary tract diseases; may be an indicator of cholestasis).
5	ALT	Alanine Aminotransferase (Sensitive biochemical indicator of hepatocellular damage).
6	AST	Aspartate Aminotransferase (Enzyme found in the liver and other tissues, indicator of hepatic damage).

Table 2 (continued) Hepatitis C Dataset (https://www.kaggle.com/datasets/fedesoriano/hepatitis-c-dataset)

No	Features	Explanation
7	BIL	Total Bilirubin (Product of heme catabolism; increases in liver dysfunction or cholestasis).
8	CHE	Cholinesterase (an enzyme that reflects the synthetic capacity of the liver; low levels may indicate liver failure).
9	CHOL	Cholesterol (Total serum cholesterol level may decrease due to decreased synthesis in liver diseases).
10	CREA	Creatinine (Kidney function indicator; may also be related to metabolic status).
11	GGT	Gamma Glutamyl Transferase (liver and biliary tract enzyme; associated with alcohol use and cholestasis).
12	PROT	Total Protein (Total serum protein level (alb + glob); reflects liver synthesis function and immunological status).
13	Diagnostic classification	0 = no disease (Blood Donor, Suspected Blood Donor), 1 = disease present (Hepatitis, Fibrosis, Cirrhosis).

Binary Feature Construction

For both datasets, we transformed the original variables into sets of binary (0/1) predictors. Categorical predictors were encoded as indicator variables, and continuous predictors were discretized into clinically or distribution-based ranges and then encoded as 0/1 indicators. For variables where zero values or missing codes represented physiologically impossible measurements, we treated such values as missing and created explicit “Missing” indicators. Binary outcomes were kept in their original form (0 vs. 1). After construction, we removed predictors with no variability (columns with the same value for all individuals).

Where indicated, the resulting binary design matrices were further split into development and validation subsets using stratified sampling, and the splits were post-processed to avoid features that were constant within one subset but not the other.

Heart Attack Dataset

Outcome: The original binary outcome variable target was retained without modification, with 0 denoting absence of clinically diagnosed coronary heart disease, and 1 denoting presence of disease.

Categorical predictors: The following categorical predictors were encoded as sets of 0/1 indicator variables: Gender: Encoded as two indicators: male and female. Chest pain type (cp): Encoded as four indicators corresponding to the original categories: typical angina, atypical angina, non-angina pain, and asymptomatic. Fasting blood sugar (fbs): Encoded as two indicators: fasting blood sugar >120 mg/dL vs. ≤120 mg/dL (normal). Resting ECG (restecg): Encoded as three indicators: normal, ST–T wave abnormality, and left ventricular hypertrophy. Exercise-induced angina (exang): Encoded as two indicators: presence vs. absence of angina during exercise. ST segment slope (slope): Encoded as three indicators: upsloping, flat, and downsloping ST segment. Number of major vessels (ca:) Encoded as three indicators: 0 vessels, 1 vessel, and ≥2 vessels with visible narrowing. Thallium stress test (thal): Encoded as four indicators: normal perfusion, fixed defect, reversible defect, and unknown.

Continuous predictors: Continuous predictors were discretized into clinically meaningful ranges and encoded as 0/1 indicators: Age (years): <45, 45–54, 55–64, and ≥65 years. Resting blood pressure (trestbps, mm Hg): <120, 120–139, 140–159, and ≥160 mm Hg. Total cholesterol (chol, mg/dL): <200, 200–239, and ≥240 mg/dL. Maximum heart rate (thalach, beats per minute): <120, 120–149, and ≥150 bpm. ST depression at exercise vs rest (oldpeak, mm): =0, (0, 2.0), and ≥2.0 mm. Each interval was represented by a separate 0/1 indicator. The original continuous variables were not used directly in the binary models.

Final design matrix: All indicator variables, together with the binary outcome target, were combined into a single binary design matrix. Predictors with no variability (all 0 or all 1 in the entire dataset) were removed. The final matrix was saved as Heart_Attack_Dataset_binarized.csv using the semicolon (';') as the field separator.

"Hepatitis C" Dataset

Outcome and initial preprocessing: The Hepatitis C dataset included a patient identifier and a diagnosis category. The patient ID column (Unnamed: 0) was discarded. A binary outcome variable Target was defined from the categorical diagnosis Category:

Healthy (Target = 0): Records labelled as "Blood Donor" or "Suspected Blood Donor".

Patient (Target = 1): Records labelled as Hepatitis, Fibrosis, or Cirrhosis.

Demographic predictors: Age (years): Discretized into <40, 40–49, 50–59, and ≥60 years, each represented by a separate 0/1 indicator. Gender: Two indicators: male and female (based on the original "m" / "f" coding).

Laboratory measurements and missingness: All laboratory variables were discretized into clinically or distribution-motivated ranges. For each variable, an explicit missingness indicator was created, and NaN values (if present) were not included in the range-based indicators:

Albumin (ALB, g/L): <39, 39–45, and ≥45, plus ALB_Missing. Alkaline phosphatase (ALP, IU/L): <55, 55–80, and ≥80, plus ALP_Missing. Alanine aminotransferase (ALT, IU/L): <20, 20–35, and ≥35, plus ALT_Missing. Aspartate aminotransferase (AST, IU/L): <25, 25–40, and ≥40, plus AST_Missing. Bilirubin (BIL, dataset-specific units:) <7, 7–20, and ≥20, plus BIL_Missing. Cholinesterase (CHE): <7.5, 7.5–9.5, and ≥9.5, plus CHE_Missing. Cholesterol (CHOL, mmol/L): <5.2, 5.2–6.2, and ≥6.2, plus CHOL_Missing. Creatinine (CREA, μmol/L): <70, 70–90, and ≥90, plus CREA_Missing. Gamma-glutamyl transferase (GGT, IU/L): <20, 20–60, and ≥60, plus GGT_Missing. Total protein (PROT, g/L): <70, 70–75, and ≥75, plus PROT_Missing.

Each interval was encoded as a separate 0/1 indicator. Missing values contributed only to the corresponding <Feature>_Missing indicator.

Train–Validation Split and Post-Processing

All binary predictors and the outcome Target were combined into a single design matrix. The data were then split into development (70%) and validation (30%) subsets using stratified sampling on Target. The following post-processing strategy was applied:

- Observations were swapped between development and validation subsets when this helped ensure that each binary predictor with variability in the full dataset had both 0 and 1 represented in each subset;
- Predictors with no variability in the full dataset (all 0 or all 1) were removed from all matrices;
- If a predictor remained constant in both subsets after swapping, it was also removed.

Handling Auxiliary Missing-Value Indicators

For all datasets, binary features explicitly encoding missingness (i.e. variables whose names contained the suffix "_Missing", such as Cholesterol_Missing, Glucose_Missing, BMI_Missing, etc.) were treated as auxiliary technical indicators rather than clinically meaningful predictors. During the analysis of the BRS results, we therefore excluded these auxiliary features from the final rule sets: any candidate rule that contained at least one "_Missing" feature was discarded, and only rules composed of non-missing, clinically interpretable predictors were retained for presentation and interpretation.

Binary Rule Search and Mask Optimization

This study aimed to obtain a comprehensive binary rule set using only clinically clearly defined features, optimized for performance, and supported by stability and significance analyses. To this end, continuous and categorical variables were first converted to binary indicators, and then possible rule combinations were exhaustively or stochastically searched using the custom-developed Rust-based Binary Rule Search (brs.exe) tool.

Rule performance was evaluated throughout the process using the Matthews correlation coefficient (MCC). Additionally, the Stability-Bound Rule Score (SBRS), defined as the geometric mean of the 95% lower confidence limits of the MCC obtained by bootstrapping on the development (DEV) and hold-out validation (VALID) sets, was calculated. This metric not only prioritizes accurate classification but also rules robust to sampling variation. The subsequent Python-based stages aggregate candidate rules from each resampling run, recompute their performance on the full development and validation datasets, and compile structured reports and figures. Optional modules can compute permutation-based importance measures, but these were not used in the present experiments.

Given a binarized dataset and a binary outcome, we search over all combinations of (k) binary predictors (feature subsets) and all possible assignments of bit-patterns (states) to the positive class. Each such assignment defines an interpretable decision rule ("mask"). For a fixed combination of features, the optimal mask is the one maximizing MCC on the DEV data.

Implementation highlights:

- BRS operates directly on the semicolon-delimited CSV file, identifies features with exactly two non-missing levels, encodes them as bits, and preserves row order.
- For small subset sizes (e.g. ($k \leq 5$)), all (2^k) masks can be explored exhaustively; for larger (k), we use a stochastic hill-climbing search with multi-bit flips and random restarts, bounded by a fixed number of mask evaluations.
- For each combination, BRS stores: feature indices, the selected state-to-class mapping (binary mask), and the contingency table required to recompute MCC and other metrics.
- This stage is implemented in Rust (brs.exe) for efficiency and parallelized across combinations.

Ensemble Construction and Stability-Bound Scoring

To mitigate overfitting to a single split, we treat the DEV cohort as a source of multiple resampled training–validation splits and track how often particular feature sets and masks reappear among top performers. Candidate rules are then re-evaluated on the *full* DEV and VALID cohorts with bootstrap uncertainty quantification.

Implementation highlights:

- Multiple stratified 80/20 splits of DEV are generated; BRS is run on each split to obtain top-ranked rules across (k) values.
- Python utilities aggregate these outputs (run_1–run_3 logic): masks are grouped by feature set and configuration, and their performance is recomputed on the full DEV and VALID datasets.
- For every rule we estimate a percentile-based 95% confidence interval for MCC on DEV and VALID using fast multinomial bootstrap of the confusion matrix.
- The *Stability-Bound Rule Score* (SBRS) is defined as

$$\mathrm{SBRS} = \sqrt{(L95_{dev} + 1)(L95_{val} + 1)} \quad (1)$$

where ($L95_{dev}$ and $L95_{val}$) are the lower 95% bounds of MCC on DEV and VALID. Adding 1 shifts MCC to a non-negative range for multiplicative aggregation and emphasizes conservative, reproducible rules.

- Final shortlists (e.g. top (N) rules per (k)) are exported for downstream analysis.

Workflow Overview and Reporting

For practical use, the components described above are assembled into a single stability-bound binary rule search workflow (Figure 1). It consists of the following stages:

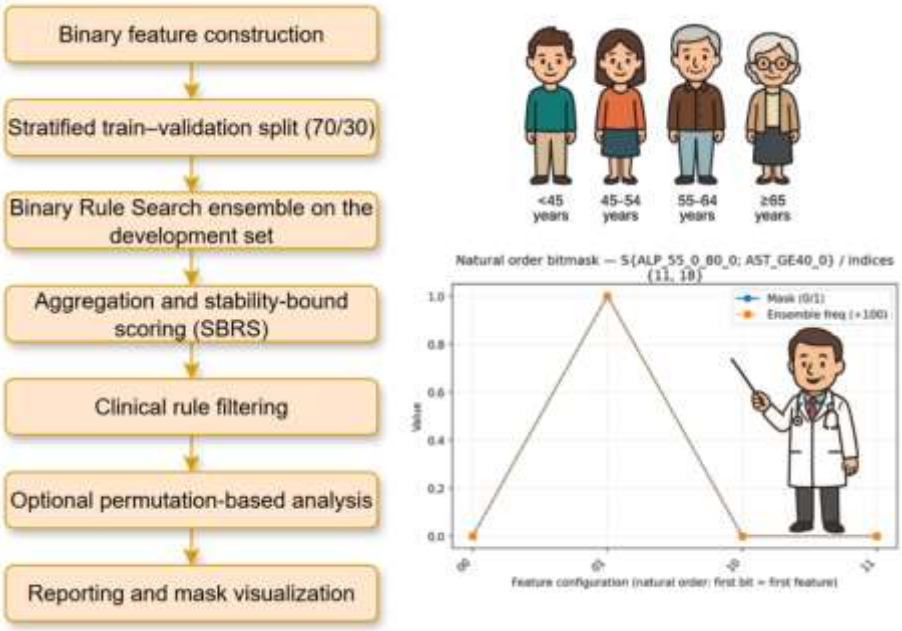


Figure 1. Overview of the stability-bound binary rule search workflow. Left: main pipeline from binary feature construction and stratified 70/30 train–validation split through Binary Rule Search, stability-bound scoring (SBRS), clinical rule filtering, optional permutation analysis and final reporting. Top right: example discretization of age into four clinically meaningful intervals (<45, 45–54, 55–64 and ≥65 years) used to create binary predictors. Bottom right: illustrative "mask vs. outcome pattern" plot for a selected rule, highlighting how the learned binary mask can be interpreted by a clinician as a simple graphical decision aid

- *Binary feature construction:* The full clinical dataset is transformed into a fully binarized design matrix: continuous and categorical variables are converted into 0/1 indicators using clinically meaningful thresholds and category encodings. The original continuous variables are not used directly in subsequent models.
- *Stratified train–validation split (70/30):* The binarized dataset is split into development (70%) and validation (30%) subsets using stratified sampling on the binary outcome. Additional post-processing ensures that every binary predictor with variability in the full dataset has both 0 and 1 represented in each subset; predictors that remain constant are removed.
- *Binary Rule Search ensemble on the development set:* On the development subset, the Binary Rule Search (BRS) engine explores conjunctions of a small number of binary predictors for several rule sizes k. For each k, BRS is run repeatedly on multiple 80/20 resamples of the development data, producing an ensemble of high-performing candidate rules.
- *Aggregation and stability-bound scoring (SBRS):* Candidate rules from all resampled runs are aggregated by feature combination and mask. For each distinct rule, the MCC and bootstrap-based 95% confidence intervals are calculated on the development and validation sets. The lower

confidence bounds are then combined into a Stability-Bound Rule Score (SBRS), which ranks rules that remain both accurate and robust across splits.

- *Clinical rule filtering:* From this ranked shortlist, only clinically interpretable rules are retained. Rules involving explicit missing-value indicators are discarded, and the remaining rules are reviewed to ensure that their structure is compatible with existing clinical knowledge and intended use.
- *Optional permutation-based analysis:* An optional module can quantify rule-level or global feature contributions by permuting individual predictors and measuring the resulting drop in SBRS. This functionality is available in the software but was not used in the present experiments.
- *Reporting and mask visualization:* The final output consists of compact tables and binary mask plots summarizing the top rules and their SBRS-based performance. These artifacts can be directly reused in figures, supplementary material, or downstream decision-support tools and form the basis of the graphical workflow shown in Figure 1.

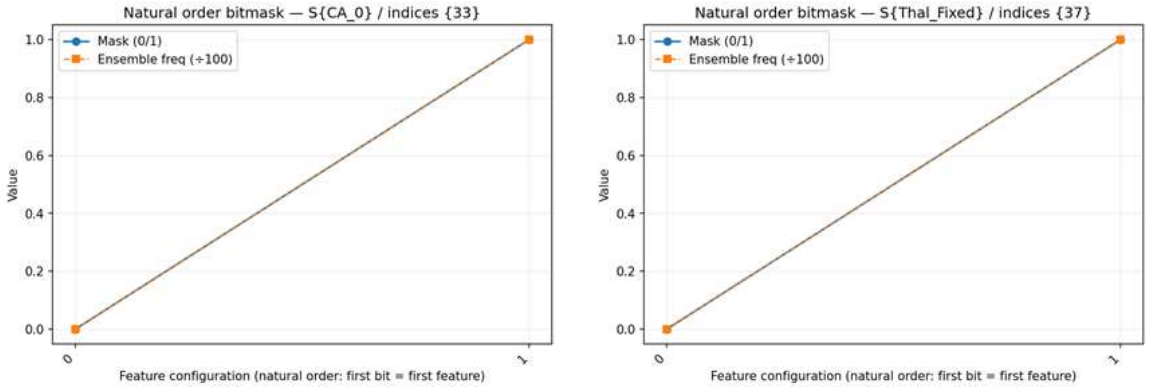
RESULTS AND DISCUSSION

Results stability-bound binary rule search for the "Heart Attack" dataset

The stability-bound binary rule search workflow was first applied to the binarized Heart Attack dataset, using conjunctions of up to four binary predictors (k = 1–4). For each rule, we calculated the MCC on the development and validation sets and derived a Stability-Bound Rule Score (SBRS) from the lower 95% confidence bounds of both. Rules were then ranked by SBRS to identify those that were both accurate and robust to resampling.

Single-feature rules (k=1):

Several single predictors already showed moderate discrimination. The best-performing rule was the indicator of "no stenosed vessels" (CA_0), with MCC =0.40 (95% CI 0.27–0.52) in the development set and 0.62 (95% CI 0.44–0.76) in the validation set, yielding SBRS=1.35. A rule based on a fixed perfusion defect on thallium imaging (Thal_Fixed) achieved MCC =0.57 (95% CI 0.45–0.68) in development and 0.42 (95% CI 0.23–0.61) in validation, with SBRS=1.34. Single-feature rules involving typical chest pain or reversible thallium defects reached similar MCC values but slightly lower SBRS, reflecting less stable performance across splits (Figure 2).



(a) (b)

Figure 2. Single-feature rules ($k = 1$) on the Heart Attack dataset and their optimal BRS masks: (a) rule based on CA_0; (b) rule based on Thal_Fixed

Pairwise rules (k=2):

Introducing a second predictor improved both accuracy and stability. All top-ranked two-feature rules combined the absence of angiographic stenosis (CA_0) with either chest-pain descriptors or exercise-induced angina. The conjunction CP_Typical and CA_0 reached MCC =0.57 (95% CI 0.46–0.66) in the development set and 0.57 (95% CI 0.43–0.70) in validation, with SBRS=1.44. A rule combining exercise-induced angina with the absence of stenosed vessels (Exang_Yes and CA_0) achieved MCC =0.51 (95% CI 0.40–0.62) and 0.62 (95% CI 0.45–0.76) in the development and validation sets, respectively, with SBRS=1.43. Rules involving non-anginal chest pain with CA_0 had slightly lower MCC in development but comparable performance in validation, and SBRS values around 1.42 (Figure 3).

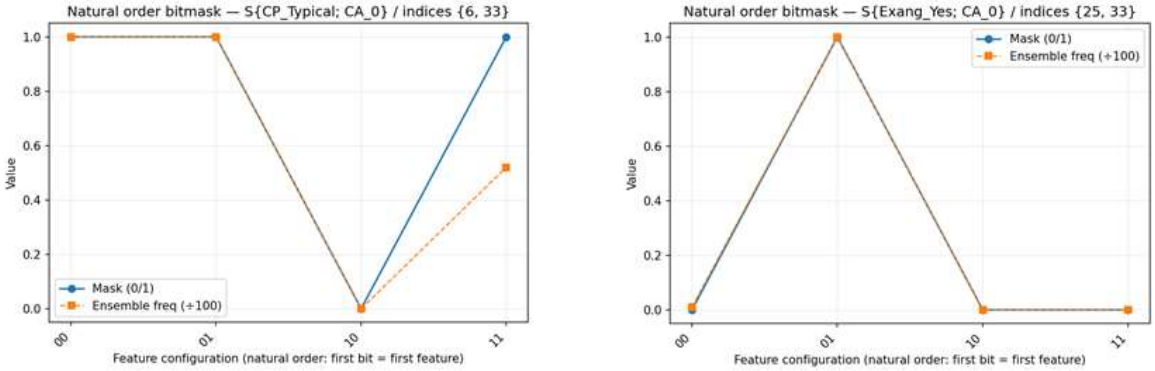


Figure 3. Pairwise rules (k=2) on the Heart Attack dataset and their optimal BRS masks: (a) CP_Typical and CA_0; (b) Exang_Yes and CA_0

Three-feature rules (k=3):

For three-feature combinations, the best solutions consistently involved a chest pain type, angiographic status, and a thallium category. The top-ranked rule CP_Typical and CA_0 and Thal_Fixed achieved MCC =0.72 (95% CI 0.63–0.81) in the development data and 0.65 (95% CI 0.49–0.79) in validation, with SBRS=1.56. A closely competing rule CP_Typical and CA_0 and Thal_Reversible reached MCC =0.70 (95% CI 0.60–0.78) in development and again 0.65 (95% CI 0.49–0.79) in validation, with SBRS=1.54. Triplet rules built around non-anginal chest pain with CA_0 and thallium findings yielded only slightly lower MCC and SBRS values (≈1.52), confirming that the core interaction between chest-pain pattern, coronary anatomy and myocardial perfusion is strongly supported by the data (Figure 4).

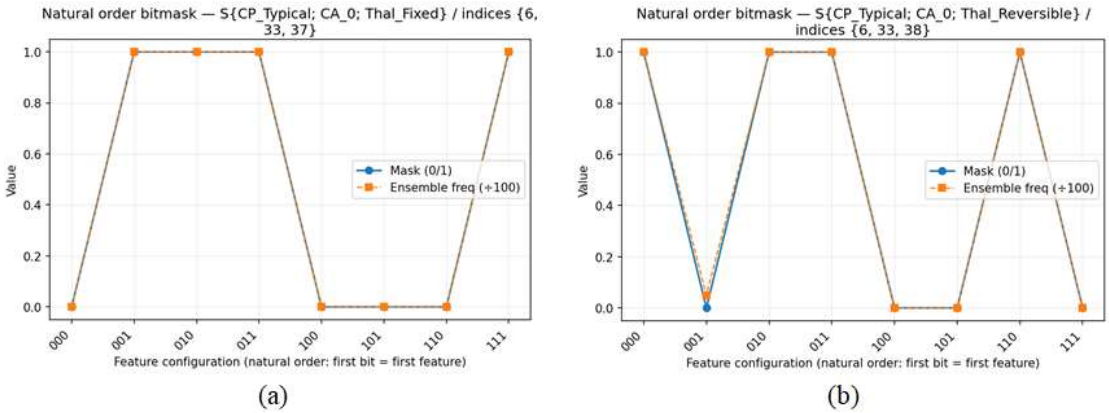


Figure 4. Three-feature rules (k=3) on the Heart Attack dataset and their optimal BRS masks: (a) CP_Typical and CA_0 and Thal_Fixed; (b) CP_Typical and CA_0 and Thal_Reversible

Four-feature rules (k=4):

Adding a fourth feature further refined these patterns but led to only modest gains in stability-bound performance, indicating a near-plateau in predictive accuracy. The highest-ranked rule, combining age 55–64 years, typical chest pain, absence of angiographically stenosed vessels and a reversible thallium defect (Age_55_64 and CP_Typical and CA_0 and Thal_Reversible), achieved MCC =0.71 (95% CI 0.61–0.80) in the development set and 0.73 (95% CI 0.58–0.87) in validation, with SBRS=1.59. Alternative four-feature rules that incorporated a low maximal heart rate (MaxHR_LT120) or intermediate cholesterol (Chol_200_239) alongside CA_0, CP_Typical and thallium categories produced very similar MCC values (development ≈0.73, validation ≈0.66–0.67) and SBRS in the range 1.57–1.58 (Figure 5).

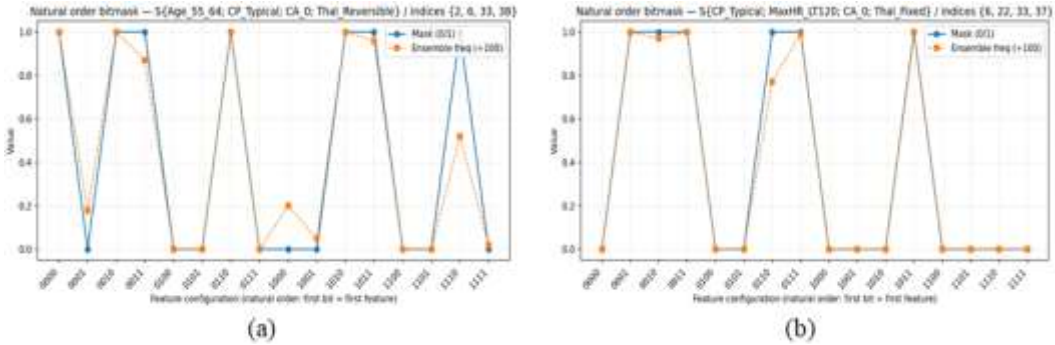


Figure 5. Four-feature rules (k=4) on the Heart Attack dataset and their optimal BRS masks: (a) Age_55_64 and CP_Typical and CA_0 and Thal_Reversible; (b) an alternative four-feature rule including heart rate (MaxHR_LT120) together with CP_Typical, CA_0 and thallium categories

Across all rule sizes, the most stable high-performing rules consistently involved CA_0 together with typical or non-anginal chest pain patterns and thallium test findings. This yields a compact and clinically interpretable core for risk stratification in this dataset, with SBRS increasing from approximately 1.35 for the best single-feature rule to about 1.59 for the best four-feature rule, and only minor gains beyond k=3.

Results stability-bound binary rule search for the “Hepatitis C” dataset

The stability-bound binary rule search workflow was next applied to the binarized Hepatitis C prediction dataset, targeting discrimination between healthy blood donors (including “suspected donors”) and patients with hepatitis, fibrosis or cirrhosis. Conjunctions of up to four binary predictors (k=1,…,4) were explored. For each candidate rule, we computed the Matthews correlation coefficient (MCC) on the development and validation subsets and derived a Stability-Bound Rule Score (SBRS) from the lower 95% confidence bounds of MCC in both splits. Rules were ranked by SBRS, so that high-scoring rules were simultaneously accurate and robust to resampling. Auxiliary missing-value indicators (ALP_Missing and CHOL_Missing) were treated as technical features during preprocessing and did not appear in the final clinical rule set.

Single-feature rules (k=1)

For single-feature rules, elevated aspartate aminotransferase was the dominant predictor. The indicator of high AST (AST_GE40_0) achieved MCC =0.62 (95% CI 0.50–0.72) in the development set and 0.70 (95% CI 0.53–0.84) in the validation set, with SBRS=1.51. Gamma-glutamyl transferase in the highest range (GGT_GE60_0) provided moderate discrimination, with MCC =0.40 (95% CI 0.28–0.52) in development and 0.41 (95% CI 0.22–0.59) in validation, and SBRS=1.25. Other single-feature rules, such as low AST (AST_LT25_0) or intermediate ALP (ALP_55_0_80_0), had lower

MCC and SBRS values (around 1.20–1.23). Overall, these results confirm that markers of hepatocellular injury and cholestasis are informative even when used as single-variable rules (Figure 6).

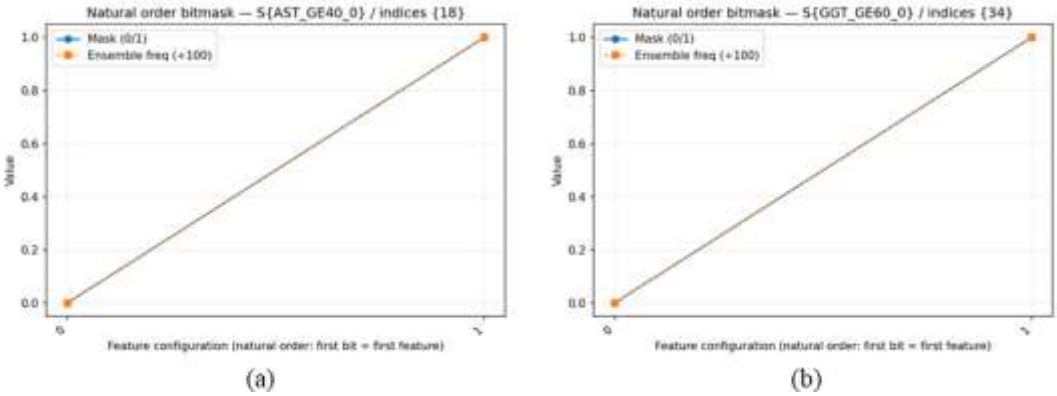


Figure 6. Single-feature rules (k=1) on the Hepatitis C dataset and their optimal BRS masks: (a) rule based on AST_GE40_0; (b) rule based on GGT_GE60_0

Pairwise rules (k=2)

Introducing a second variable markedly improved performance and stability. All top-ranked two-feature rules were centred on AST_GE40_0 combined with additional markers of liver injury or cholestasis. The best-performing rule, ALP_55_0_80_0 and AST_GE40_0, reached MCC =0.68 (95% CI 0.57–0.79) in the development set and 0.79 (95% CI 0.64–0.91) in the validation set, with SBRS=1.61. Rules combining elevated AST with low GGT (GGT_LT20_0), low bilirubin (BIL_LT7_0) or intermediate GGT (GGT_20_0_60_0) yielded MCC values in the range 0.63–0.67 on development and 0.67–0.72 on validation, with SBRS between 1.52 and 1.54 (Figure 7).

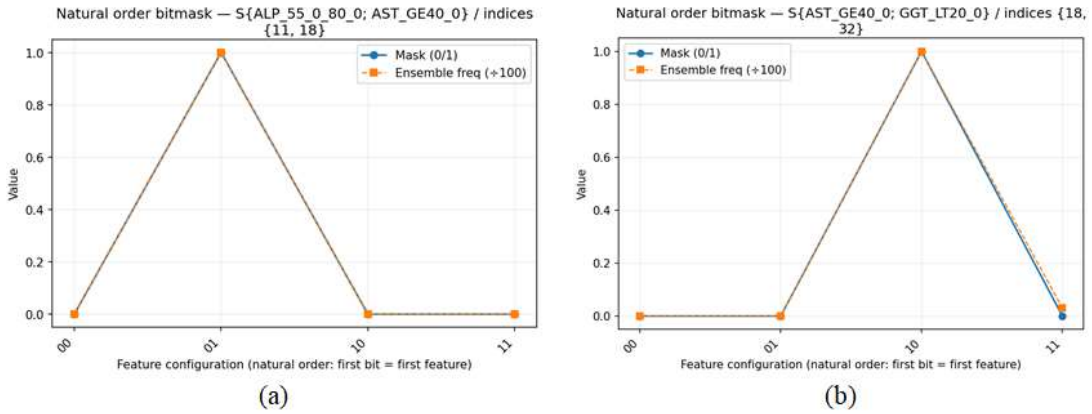
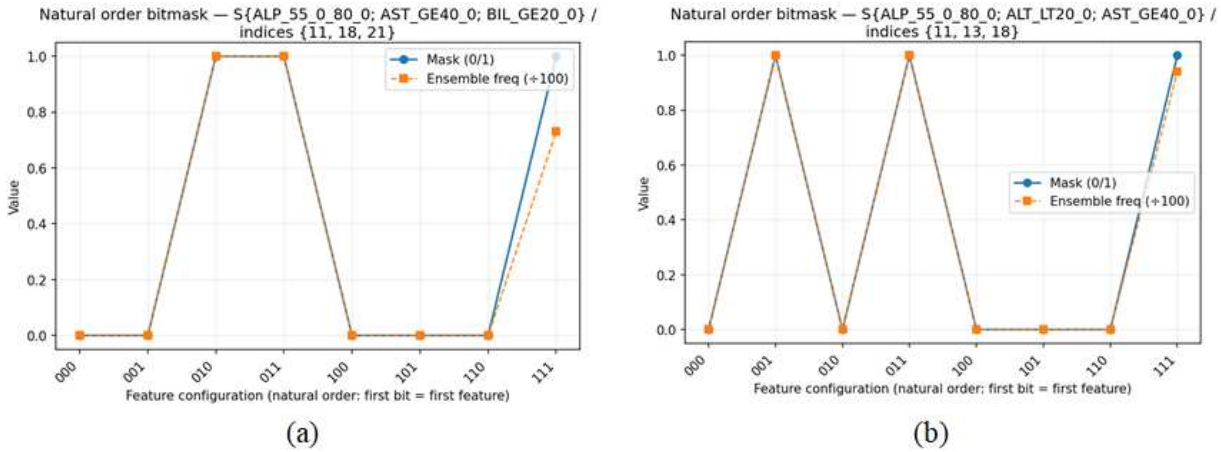


Figure 7. Pairwise rules (k=2) on the Hepatitis C dataset and their optimal BRS masks: (a) ALP_55_0_80_0 and AST_GE40_0; (b) AST_GE40_0 and GGT_LT20_0

Three-feature rules (k=3)

For three-feature rules, the best combinations continued to use AST_GE40_0 as a core component, augmented by alkaline phosphatase and markers of more advanced liver dysfunction. The top-ranked rule ALP_55_0_80_0;AST_GE40_0;BIL_GE20_0 achieved MCC =0.69 (95% CI 0.57–0.79) in the development set and 0.82 (95% CI 0.68–0.94) in the validation set, with SBRS=1.63. A closely competing rule ALP_55_0_80_0;ALT_LT20_0;AST_GE40_0 reached MCC =0.71 (95% CI 0.60–0.81) in development and 0.79 (95% CI 0.64–0.92) in validation, with SBRS=1.62. Another rule involving low cholinesterase (ALP_55_0_80_0;AST_GE40_0;CHE_LT7_5) exhibited similar MCC values and SBRS≈1.61. These results indicate that combining hepatocellular injury (AST), cholestasis

(ALP, bilirubin) and reduced synthetic function (ALT, CHE) effectively reflects the clinical progression from healthy donors to chronic liver disease (Figure 8).

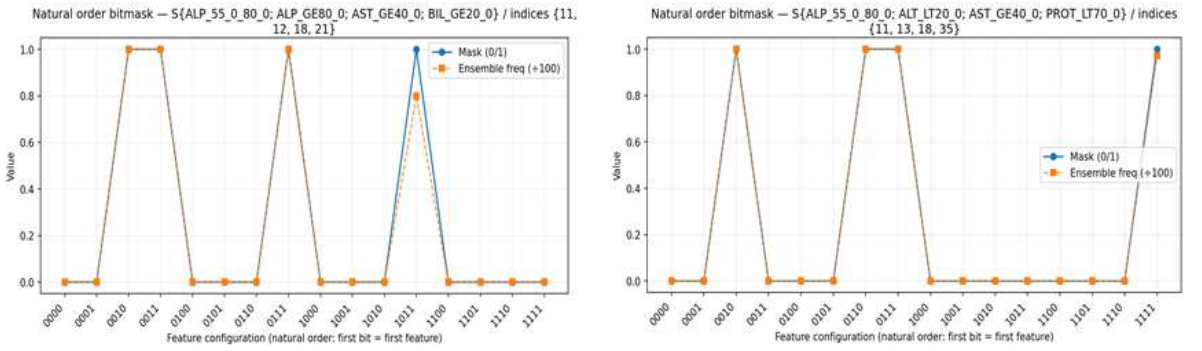


(a) (b)

Figure 8. Three-feature rules (k=3) on the Hepatitis C dataset and their optimal BRS masks: (a) ALP_55_0_80_0;AST_GE40_0;BIL_GE20_0 (b) ALP_55_0_80_0;ALT_LT20_0;AST_GE40_0

Four-feature rules (k=4)

Four-feature rules provided an additional but more modest gain in stability-bound performance, indicating a near-plateau beyond k=3. The best rule by SBRS, ALP_55_0_80_0; ALP_GE80_0; AST_GE40_0; BIL_GE20_0, achieved MCC =0.75 (95% CI 0.64–0.84) in the development set and 0.84 (95% CI 0.71–0.95) in the validation set, with SBRS=1.67. Alternative four-feature rules built around the same AST–ALP core, combined with low ALT or low total protein (e.g. ALP_55_0_80_0;ALT_LT20_0;AST_GE40_0;PROT_LT70_0), produced very similar MCC values (development ≈0.77, validation ≈0.79–0.82) and SBRS in the range 1.65–1.66 (Figure 9). These rules identify patients with marked hepatocellular injury, sustained ALP elevation, and additional signs of impaired bilirubin processing or synthetic function.



(a) (b)

Figure 9. Four-feature rules (k=4) on the Hepatitis C dataset and their optimal BRS masks: (a) ALP_55_0_80_0;ALP_GE80_0;AST_GE40_0;BIL_GE20_0; (b) an alternative four-feature rule including low total protein (ALP_55_0_80_0;ALT_LT20_0;AST_GE40_0;PROT_LT70_0)

Across all rule sizes, the most stable high-performing rules consistently involved elevated AST (AST_GE40_0), intermediate or high alkaline phosphatase (ALP_55_0_80_0 and/or ALP_GE80_0), and markers of cholestasis or reduced liver function such as high bilirubin (BIL_GE20_0), low cholinesterase (CHE_LT7_5) and low total protein (PROT_LT70_0). In this dataset, SBRS increased from approximately 1.51 for the best single-feature rule to about 1.67 for the best four-feature rule, while the corresponding rule sets remained sparse and clinically interpretable.

DISCUSSION

This study evaluated the classification performance of an explainability-focused workflow built on the Binary Rule Search (BRS) engine on both heart attack risk and hepatitis C datasets. We observed that the approach maintained high accuracy while preserving the clinical relevance of simple, rule-based decision structures. The results demonstrate that decision models with understandable mechanisms offer a powerful alternative not only in terms of accuracy but also in terms of clinical interpretability (Caruana et al., 2015; Holzinger et al., 2019). The literature has reported that rule-based models play a confidence-enhancing role, particularly in critical medical decisions (Lakkaraju et al., 2016; Doshi-Velez & Kim, 2017; Shortliffe & Sepúlveda, 2018; Rudin, 2019). In this context, the fact that the BRS-based workflow maintains the "transparent model–high performance" balance makes the current findings compatible with the existing literature.

In this study, the performance pattern of the BRS-based workflow across different rule sizes is in strong agreement with both the clinical literature and explanatory modelling principles. The significant discrimination obtained from CA_0 and thallium findings, even with univariate rules, is consistent with studies emphasizing that coronary anatomy and myocardial perfusion indicators are key determinants of heart attack risk (Dilsizian & Ficaro, 2011; Patel et al., 2017; Gulati et al., 2021). However, the most striking advantage of the workflow is its ability to automatically rediscover these well-known clinical indicators not only in isolation but also in informatively complementary and clinically plausible combinations. In particular, the recurring core structure of CP_Typical, CA_0 and thallium subtypes in two- and three-feature rules directly reflects the diagnostic value of the symptom–anatomy–perfusion triad, frequently emphasized in the literature (Amsterdam et al., 2014; Foy et al., 2015). Moreover, the marked increase in MCC values for these rules compared with univariate solutions demonstrates that BRS can capture interactions in a simple and interpretable manner. Moreover, prior work has shown that sparse and interpretable rule-based models tend to reduce overfitting risk in clinical prediction tasks by limiting model complexity and improving generalization across heterogeneous patient populations (Steyerberg et al., 2019; Yu & Kumbier, 2020). The limited performance gain with four-feature rules supports earlier findings that added complexity in clinical decision models is not always useful and may introduce noise (Angelino et al., 2018; Rudin, 2019; Molnar, 2022). In this context, the BRS-based workflow's ability to select compact yet high-performing decision structures provides a clear advantage for explainable and clinically usable AI. The model's repeated selection of concise rules involving CA_0, typical chest pain and thallium outcomes is not merely data-driven; it reflects a dependable framework consistent with clinical reasoning. Still, some studies note that symptom-based models alone may show variable performance and reduced sensitivity in certain age groups (Foy et al., 2015). The proposed workflow's strength is its capacity to distinguish specific clinical profiles using sparse rules, potentially reducing overfitting. However, larger samples are needed to confirm whether accuracy remains stable in more heterogeneous subgroups.

The results obtained on the hepatitis C dataset are remarkably consistent with existing literature on biochemical markers of liver injury, demonstrating the workflow's ability to capture clinically meaningful patterns with minimal model complexity. The prominence of AST as the most dominant predictor among univariate rules is strongly supported by studies reporting AST as one of the most sensitive indicators of hepatocellular injury (Pratt & Kaplan, 2000; Giannini et al., 2005; European Association for the Study of the Liver [EASL], 2018). The stability-bound binary rule search systematically paired AST with complementary markers such as ALP, bilirubin or GGT in bivariate

and trivariate rules. This is in line with literature emphasising the diagnostic value of concurrent findings of hepatocellular injury and cholestasis in liver disease (Schuppan & Afdhal, 2008; Rockey et al., 2009; Cui et al., 2018). In particular, the separation achieved by three-feature rules such as ALP_55_0_80_0–AST_GE40_0–BIL_GE20_0 with high MCC and SBRS values supports the observation that increased bilirubin tends to indicate more advanced disease or functional impairment. The minimal performance gain from four-feature rules aligns with explainable AI literature suggesting that unnecessary features do not always enhance diagnostic accuracy (Lipton, 2018; Rudin, 2019). Here, the advantage of the stability-bound scoring scheme becomes evident: it preserves high MCC values while favouring rules built only on genuinely informative biochemical markers. The workflow therefore differentiates donors from hepatitis cases using compact, interpretable and clinically meaningful rules. In this sense, unlike regression-based or other "black-box" models, the stability-bound binary rule search framework can be transparently incorporated into clinical decision-making.

In both datasets, the highest stability-bound performance occurred with rules of size k = 3, suggesting that larger rules add little classification benefit. This aligns with prior work stressing the need to balance simplicity and generalisation in rule-based models (Angelino et al., 2018; James et al., 2021). Still, more complex clinical scenarios may require larger rules, indicating that the workflow might need further optimisation across different contexts.

Overall, the results show that the BRS-based, stability-bound workflow provides strong predictive and explainable performance for both cardiovascular risk stratification and hepatitis-related chronic liver disease. However, both datasets were analysed as binary classifications; separate validation is needed for multi-class settings. Future work should therefore examine rule stability in prospective data and evaluate the framework's adaptability to diverse clinical environments and disease profiles.

CONCLUSION

This study demonstrated that a stability-bound binary rule search workflow built on the Binary Rule Search (BRS) engine can provide both high predictive performance and strong explainability when classifying heart attack risk and hepatitis C–related chronic liver disease. In both datasets, the highest stability-bound performance was obtained with sparse, clinically meaningful three-feature rules (k=3), which are easy to interpret and align with established clinical reasoning, increasing their potential to support fast and transparent decision-making at the point of care. These findings support the broader view that transparent rule-based models can be reliable not only in terms of machine learning accuracy but also in terms of clinical applicability.

Future work should test the generalization capacity of the proposed workflow in larger, multicenter cohorts, extend it to multi-class disease settings and prospective data streams, and evaluate its integration into real-world clinical workflows. Overall, the stability-bound binary rule search workflow appears to be a transparent and effective option for medical decision support systems, addressing the need for accountable and interpretable AI solutions in clinical practice.

Conflict of Interest

The authors declare no conflict of interest.

Author's Contributions

Conceptualization, M.T.H. and A.V.; methodology, M.T.H. and A.V.; software, A.V.; validation, M.T.H. and A.V.; formal analysis, M.T.H.; investigation, A.V.; resources, M.T.H.; data curation, M.T.H.; writing—original draft preparation, M.T.H. and A.V.; writing—review and editing,

M.T.H. and A.V.; visualization, A.V.; supervision, M.T.H. All authors have read and agreed to the published version of the manuscript.

REFERENCES

Amsterdam, E. A., Wenger, N. K., Brindis, R. G., et al. (2014). 2014 AHA/ACC Guideline for the Management of Patients With Non–ST-Elevation Acute Coronary Syndromes. Circulation, 130(25), e344–e426. https://doi.org/10.1016/j.jacc.2014.09.017.

Angelino, E., Larus-Stone, N., Alabi, D., et al. (2018). Learning Certifiably Optimal Rule Lists for Categorical Data. Journal of Machine Learning Research, 19, 1-43. https://www.jmlr.org/papers/v19/17-716.html.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare. *Proceedings of the 21st ACM SIGKDD*, 1721–1730. https://doi.org/10.1145/2783258.2788613.

Cui, Y., Wang, Y., Wang, Y., & Liu, J. (2018). Serum liver enzyme levels and hepatitis C virus infection: A systematic review and meta-analysis. Journal of Clinical Laboratory Analysis, 32(2), e22215. https://doi.org/10.1002/jcla.25127.

Dilsizian, V., & Ficaro, E. P. (2011). Cardiac SPECT imaging: State-of-the-art and future directions. Journal of Nuclear Cardiology, 18(6), 1026–1043. https://doi.org/10.1007/s12350-011-9480-1.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. https://doi.org/10.48550/arXiv.1702.08608.

European Association for the Study of the Liver (EASL). (2018). EASL clinical practice guidelines: Management of hepatitis C virus infection. *Journal of Hepatology*, 69(2), 461–511. https://doi.org/10.1016/j.jhep.2018.03.026.

Foy, A. J., Liu, G., Davidson, W. R., et al. (2015). Comparative effectiveness of diagnostic testing strategies in emergency department patients with chest pain. BMJ, 351, h5447. https://doi.org/10.1001/jamainternmed.2014.7657.

Giannini, E. G., Testa, R., & Savarino, V. (2005). Liver enzyme alteration: A guide for clinicians. Canadian Medical Association Journal, 172(3), 367–379. https://doi.org/10.1503/cmaj.1040752.

Gulati, M., Levy, P. D., Mukherjee, D., et al. (2021). 2021 AHA/ACC Guideline for the Evaluation and Diagnosis of Chest Pain. Circulation, 144(22), e368–e454. https://doi.org/10.1161/CIR.0000000000001029.

Harrell FE Jr. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2nd ed. Springer; New York, NY; 2015. https://doi.org/10.1007/978-3-319-19425-7.

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2019). What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923*. https://doi.org/10.48550/arXiv.1712.09923.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1.

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1675–1684). ACM. https://doi.org/10.1145/2939672.2939874.

Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36–43. https://doi.org/10.1145/3233231.

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). Lulu.com.

Patel, M. R., et al. (2017). Low diagnostic yield of elective coronary angiography. *New England Journal of Medicine*, 362(10), 886–895. https://doi.org/10.1056/NEJMoa0907272.

Pratt, D. S., & Kaplan, M. M. (2000). Evaluation of abnormal liver-enzyme results in asymptomatic patients. New England Journal of Medicine, 342(17), 1266–1271. https://doi.org/10.1056/NEJM200004273421707.

Rockey, D. C., Caldwell, S. H., Goodman, Z. D., Nelson, R. C., & Smith, A. D. (2009). Liver biopsy. *Hepatology*, 49(3), 1017–1044. https://doi.org/10.1002/hep.22742.

Rudin, C. (2019). Interpretable machine learning for high-stakes decisions. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x.

Schuppan, D., & Afdhal, N. H. (2008). Liver cirrhosis. The Lancet, 371(9615), 838–851. https://doi.org/10.1016/S0140-6736(08)60383-9.

Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *JAMA*, 320(21), 2199–2200. https://doi.org/10.1001/jama.2018.17163.

Steyerberg, E. W. (2019). Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating (2nd ed.). Springer. https://doi.org/10.1007/978-3-030-16399-0.

Yu, B., & Kumbier, K. (2020). Veridical data science. Proceedings of the National Academy of Sciences, 117(8), 3920–3929. https://doi.org/10.1073/pnas.1901326117.