An Explainable Approach for Heart Disease Prediction Using an Evolutionary Rule-Based Learning Classifier System on Real-World ICU Data

Sadia Shareef, Fawad Nasim
Faculty of Computer Science and Information Technology, The Superior University, Lahore, Pakistan
*__Correspondence__: fawad.nasim@hotmail.com

Citation | Shareef. S, Nasim. F, "An Explainable Approach for Heart Disease Prediction Using an Evolutionary Rule-Based Learning Classifier System on Real-World ICU Data", IJIST, Vol. 8 Issue. 5 pp 2267-2296, August 2026
Received | Aug 09, 2026 **Revised |** Aug 21, 2026 **Accepted |** Aug 25, 2026 **Published |** Aug 31, 2026.

Cardiovascular disease is the leading cause of death globally, yet few machine learning models for cardiac risk prediction are clinically interpretable. This paper develops an Explainable Artificial Intelligence (XAI) approach for heart disease risk classification on the MIMIC-IV-Ext Cardiac Disease dataset (v1.0.0, PhysioNet), comprising 4,748 de-identified ICU admissions. A multimodal set of 35 features including 14 laboratory biomarkers and NLP-derived clinical narrative indicators is constructed, with binary risk labels generated by keyword-majority voting over four clinical free-text fields. Class imbalance (78.1% High Risk vs. 21.9% Low Risk) is addressed by applying SMOTE only to the training set. The proposed classifier, ExSTraCS, a Michigan-style Learning Classifier System with kappa-based evolutionary fitness, is evaluated against five baselines, achieving 84.53% held-out test accuracy (84.29% ± 0.60% cross-validated), 90.19% F1-score, and 0.8757 AUROC (with 89.30% precision, 91.11% recall and 61.06% specificity), close to the strongest baseline; McNemar and DeLong tests assess statistical significance. ExSTraCS is a hybrid approach in which a gradient-boosting engine performs prediction, while interpretability is delivered through an evolved IF-THEN rule layer and post hoc SHAP and LIME explanations, whose faithfulness is measured using feature-perturbation metrics. Because several text-derived features share keywords with the labelling rule, a leakage-controlled experiment using only laboratory biomarkers bounds the impact of feature–label correlation. All three explanation layers agree that Troponin T, ECG abnormality, and HPI-derived symptom features are the most reliable predictors, and a cost-sensitive threshold analysis identifies the optimal operating point given the high clinical cost of false negatives. The results show that models offering both competitive predictive accuracy and clinical interpretability are achievable.

Keywords: Explainable AI; Heart disease prediction; ExSTraCS; MIMIC-IV-Ext; SHAP.

Introduction:

Cardiovascular disease (CVD) is the leading cause of death globally and is responsible for approximately 17.9 million deaths annually, or 32% of total global deaths [1]. CVDs also represent a significant economic burden: costs of CVDs in the USA were about $252 billion in 2019–2020 and are expected to rise by 90% in 2025- 2050 [2]. Identification of at-risk patients is thus one of the key challenges in clinical practice today. Machine learning has great potential to analyze machine-generated data [3], mining useful information from large volumes of EHRs stored in hospitals [4], such as lab results, clinical notes, electrocardiograms and diagnostic codes [5], to detect potential cardiac risks before symptoms become visible. In this paper, explainable AI models [6][7] are trained on such records to predict heart diseases. Traditional risk scoring systems, such as the Framingham Risk Score, are based on simplifying linear assumptions and a limited set of curated variables and are unable to capture non-linear interactions among biomarkers, symptoms, and clinical measurements that contribute to cardiac risk and may not be generalizable. In the real world, models validated on small, clean benchmarks face additional challenges, such as missing or corrupted data, extreme class imbalance, inconsistent formats, and poor documentation [8][9].

Among the machine learning algorithms available, logistic regression, decision trees, random forests, gradient boosting, and deep neural networks have higher predictive power than conventional scores for cardiac risk prediction [10]. However, such predictive accuracy comes at a cost: the most effective machine learning models are black boxes that cannot be inspected to understand their inner workings. In the practice of medicine, all decisions have to be explainable [11]. These should be in accordance with the current state of medical knowledge. An unexplained prediction is not easy to trust or act upon. The major problem in Explainable Artificial Intelligence (XAI) is reconciling these opposing requirements. XAI includes techniques that make the model's decisions transparent and interpretable to a human expert, so that a doctor could have a clearer idea of the reasons for classifying a patient as high risk and use such a model with confidence, traceability and deployability in the clinical environment. There are two basic strategies. Post hoc explanations [12] are used to explain a previously trained model: SHAP [13] assigns the marginal contribution of each feature to the model in accordance with the principles of cooperative game theory [14], while LIME builds locally faithful linear approximations to each prediction [15]. On the other hand, the intrinsic (ante-hoc) explanation methods [16] build interpretability directly into the model structure, so that the reasoning process is explicitly included in the model design rather than constructed after the fact. The Learning Classifier Systems (LCS) [17] belong to this intrinsic approach: an evolving collection of human-readable IF-THEN rules that constitute the model's true decision logic rather than its approximation. ExSTraCS (Extended Supervised Tracking and Classifying System) [18] is an extension of XCS that employs a unique kappa-based fitness function sensitive to the rules' ability to succeed even when randomly selected, which is highly valuable in the context of imbalanced data common in clinical cohorts. Nevertheless, ExSTraCS has not yet been applied to large-scale clinical ICU cardiac datasets.

Clinically relevant data is needed to create clinically useful predictions [19][20]. Unlike the small, curated benchmarks typically used in previous research, the MIMIC-IV database (2008-2022) [21] comprises anonymized EHRs from over 65,000 ICU patients, with a diverse range of features, including vital signs, laboratory data, clinical notes, ECG findings and diagnostic codes. There are four research gaps present in the literature:

(1) the use of a few selected benchmarks like the Cleveland Heart Disease Dataset; We discovered the following challenges:

(2) limited research on integrating intrinsically interpretable models with post-hoc XAI using real-world EHR data;

(3) lack of attention to class imbalance, which overestimates accuracy; and

(4) unusual multimethod XAI evaluation from global, local and rule perspectives.

To address these gaps, this paper proposes a comprehensive framework for designing explainable heart disease prediction models using the MIMIC-IV-Ext dataset. ExSTraCS is trained on 35 features obtained via NLP from clinical notes and ECGs, plus 14 laboratory markers and is compared against five standard classifiers using stratified 5-fold cross-validation. Then the three-layer XAI analysis is done by combining a set of ExSTraCS IF-THEN rules, SHAP (beeswarm, bar and waterfall plots) and LIME patient-level explanations.

Contributions are:

(1) the first evaluation on large-scale real-world cardiac ICU data with 84.29% ± 0.60% accuracy and 0.8757 AUROC; and

(2) the evidence that ExSTraCS performs equally well as the best baselines with an added interpretable evolved rule layer and explanations verified to be faithful to the model and statistical significance testing of differences in performance with Gradient Boosting. It is the most comprehensive three-layer XAI analysis of MIMIC-IV cardiac data to date, and the most detailed clinical discussion shows that most cardiac risk predictors are derived from Troponin T, ECG abnormalities, and HPI-derived NLP features. The paper is structured as follows: In section 2, ML and XAI methods used for cardiac prediction are reviewed, and in section 3, the dataset, feature engineering and methodology are described. The results and discussion are presented in section 4, while section 5 concludes with limitations and future directions.

Research Objectives:

To address the gaps identified above, this study pursues the following explicit and measurable objectives:

O1: To construct and preprocess a large-scale, real-world cardiac ICU dataset of 4,748 MIMIC-IV-Ext admissions, including handling of missing values and the 78.1/21.9 class imbalance.

O2: To train an evolutionary rule-based learning classifier (ExSTraCS) for binary heart-disease risk classification on this dataset.

O3: To benchmark the proposed model against five standard classifiers using seven metrics (accuracy, precision, recall, specificity, F1-score, AUROC and average precision) together with McNemar and DeLong significance tests.

O4: To extract and analyse the evolved IF-THEN rules as intrinsic, human-readable explanations.

O5: To perform a three-layer explainability analysis (rules, SHAP and LIME) and identify the most influential clinical predictors.

O6: To quantify explanation faithfulness and to bound feature–label leakage through a laboratory-only control experiment.

Research Novelty:

The novelty of this work is threefold and distinguishes it clearly from existing studies. First, whereas prior heart-disease prediction research relies overwhelmingly on small, curated benchmarks such as the Cleveland dataset, this study is, to the best of our knowledge, among the first to apply an evolutionary rule-based learning classifier to large-scale, real-world cardiac ICU data. Second, in contrast to existing real-world ICU studies that interpret opaque boosting models using post-hoc SHAP alone, this work couples an intrinsically interpretable, rule-producing model with a three-layer explainability analysis (evolved rules, SHAP and LIME) whose agreement is examined jointly. Third, the study explicitly quantifies explanation faithfulness and bounds feature–label leakage through a dedicated laboratory-only control experiment, an evaluation rarely reported in comparable work. Together these contributions target the under-explored intersection of accurate prediction, intrinsic interpretability and clinical realism.

Literature Review:

Research in the domain of ML for cardiovascular risk prediction has grown rapidly recently, but numerous benchmarks lack clinical utility [22][23]. Comparing the performance of supervised classifiers on cardiac data, Osei-Nkwantabisa and Ntumy [24] showed that SVM achieved 81%, LR 86%, and ANN 74%, with results indicating sensitivity to preprocessing and hyperparameter settings. In a comparison of seven algorithms on the Cleveland dataset, sensitivity to feature subset selection ranged from 67.21% (KNN) to 88.52% (Random Forest with chi-square feature selection), a 20-point difference [25]. [26] obtained similar results using 5-fold cross-validation, indicating that the results were invalid when no class balancing was performed. [27] applied Bangladeshi cardiovascular data and measured accuracy ranging from 77.55% (Decision Tree) to 84.16% (Naive Bayes with feature selection), underscoring the necessity of fair testing on underrepresented groups. In the 15-year cohort from Japan (n=7,672), [28] reached an AUC of 0.73 using Random Forest and required gender-specific subgroup modelling for generalization. An imbalance in class is one of the most commonly reported and yet inadequately addressed challenges in cardiac ML. Under sampling was performed at random sites to increase sensitivity (81%) and MCC (0.34), confirming the reliability of accuracy without any class-balancing correction in medical classification. [29] prove that EHR richness introduces noise in the system unless specific preprocessing pipelines are used, and that these pipelines go beyond noise removal to explicitly model feature co-occurrence. It is increasingly evident that prediction accuracy alone is insufficient for clinical use [30]. Comparing two methods for the explainable diabetes related ML model, namely SHAP and LIME, [31] find that the grounding of SHAP using Shapley values provides more stable attributions with respect to perturbing the input, while LIME, although more intuitive locally, does not have this stability and can lead to different feature-importance ranking for similar instances. It underlines the necessity for multi-method XAI frameworks in high-stakes and clinical scenarios. Using SHAP in combination with PDP, [32] validate the top three features for cardiovascular risk stratification, which is consistent with the clinical knowledge and includes such parameters as blood pressure, age, and cholesterol. In a study of 89 XAI healthcare research papers (2000-2024), [33] report that 63 of 76 papers focusing on tabular EHRs use SHAP. However, most of the papers lack evaluation of the clinical fidelity and actionability of the explanations. The discussion of the XAI techniques for clinical decision-support systems and the presentation of the XAI-based framework for CDSS, including four categories: feature-based, model-based, complexity-based, and methodology-based, are provided by [34]. The authors note the performance-interpretability trade-off, as well as limited data scope and low model interpretability [35], as important barriers to the trustworthy deployment of CDSS in clinical environments. [36] demonstrate the practical application of CDSS at Vanderbilt University Medical Centre, where the explanations generated by SHAP are not the justification layer but tools for quality improvement, specifically to reduce unnecessary clinical alerts. The life-critical setting prefers models that can be easily explained after the fact in a non-obscure manner, as [37] claim. Authors argue that ante-hoc interpretable models should be used instead of black-box models with approximate post-hoc interpretation.

[38] used explainable machine learning techniques for predicting heart failure severity using primary-care electronic health records. The combination of predictive modelling techniques [39] with interpretability techniques enables the extraction of clinical characteristics [40] predictive of heart failure severity from routinely collected EHR data. This study shows the ability of explainable ML to identify clinically meaningful risk markers from real-world primary-care records while ensuring the transparency required for clinical applications. The three most common issues in XAI healthcare studies, which were directly addressed in this paper, were data imbalance, missing values and datasets from a single institution, which were

mitigated using SMOTE, median imputation and the publicly available dataset MIMIC-IV-Ext.

Material and Methods:

The goal of this section is to describe the end-to-end process of data acquisition, label engineering, feature construction, class balancing, model training and multi-layer explainability that has been designed for clinical interpretability.

Dataset:

The MIMIC-IV-Ext Cardiac Disease dataset (v1.0.0, PhysioNet, 2024) is a MIMIC-IV extension, i.e., a de-identified inpatient admissions database from Beth Israel Deaconess Medical Centre (Boston, MA) spanning 2008–2019, including patients with cardiac disease in the ICU (as either a primary or secondary diagnosis). The two files used in the study were: a diagnoses file (4761 records, 18 columns) including clinical notes, ECG text, radiology reports and chief complaint; and a laboratory events file (229,817 records, 15 columns) with the first result per test per admission. The merged dataset contained 4,748 patients overall, with 35 features as shown in Table 1, which were retained after applying the 50% data-missingness criterion and performing an inner join on the hospital admission ID (hadm_id).

Table 1. Summary of MIMIC-IV-Ext Cardiac Disease Dataset

Components	Records	Columns	Description
heart_diagnoses.csv	4,761	18	Clinical notes, ECG, HPI, reports
heart_labevents_first_lab.csv	229,817	15	First-recorded lab values per admission
heart_labevents_last_lab.csv	229,817	15	Last-recorded lab values per admission
heart_labevents_avg_lab.csv	229,817	15	Averaged lab values per admission
Merged Dataset	4,748 patients	35 features	After inner join on hadm_id

The final dataset is heavily imbalanced: 3710 patients (78.1%) are assigned the label High Risk, whereas 1038 (21.9%) are assigned the label Low Risk. This imbalance is consistent with a typical ICU cardiac patient cohort. Class imbalance is handled using SMOTE. SMOTE is applied only to the training split.

Label Engineering:

The MIMIC-IV-Ext database does not contain any pre-defined binary labels (risk categories). A rule-based keyword-majority voting method was devised. Specifically, text from four clinical domains (reports, ECG, chief complaint, HPI) was lowercased and filtered for two predefined keyword sets.

The High-Risk keyword set (17 keywords) covered acute cardiac presentations: myocardial infarction, infarct, ST elevation, ischemia, ischemic, heart failure, ventricular fibrillation, cardiac arrest, LBBB, atrial fibrillation, flutter, tachycardia, block, abnormal, elevated, critical, and acute. The Low-Risk set (5 keywords) covered the following: normal ECG, normal sinus, NSR, sinus rhythm, and borderline. If the number of High-Risk keywords was greater than the number of Low-Risk keywords, the label was 1 (High Risk); otherwise, 0.

Label Validation:

As keyword labels can only approximate the ground truth and cannot be considered gold-standard diagnoses, we wished to assess the validity of the labels against a structured ground truth. For that, we compared the labels to MIMIC-IV ICD-9/ICD-10 acute cardiac diagnosis codes: an admission was labelled as positive in this external High-Risk reference set if it included at least one primary or secondary acute cardiac code (ICD-10 codes I21: acute myocardial infarction, I25: chronic ischaemic heart disease, and I50: heart failure). While the concordance of labels with MIMIC-CODES reached 74.6%, the corresponding Cohen's κ was

very low (0.09). The relatively high value of raw agreement is due to the dominant majority class in the High-Risk group. Taking that into account, a poor value of $\varkappa$ indicates only a slight agreement above what would be expected by chance. The remaining discrepancies occur primarily because of the presence of high-risk terms in negated or historical text (Sections 5.3 and 5.4). Overall, this exercise is intended to validate the use of keyword labels as an acceptable approximation for cardiac risk in order to build predictive models and define its known limitation in relation to possible label noise in the training. Because several text-derived predictors (for example ECG_stemi, RPT_infarct and ECG_abnormal) share vocabulary with the high-risk keyword set used to generate the labels, an independent-label validation was performed: the keyword labels were compared against structured ICD-9/ICD-10 acute-cardiac diagnosis codes, and, separately, a laboratory-only model that excludes all text-derived features (Section 4.9) was trained to bound the influence of this overlap. These two checks together provide an independent assessment of label validity and of feature–label leakage.

Feature Engineering:

A feature engineering pipeline extracted clinically meaningful features from unstructured free-text fields and structured laboratory fields. There were 35 features altogether in three modalities: NLP-derived clinical narrative features, ECG and report indicators, and quantitative biomarkers (Table 2). All the features were calculated deterministically from the raw fields before imputation or scaling.

Table 2: Complete Feature Engineering Summary – All 35 Features

Feature Group	Features	Count	Source
NLP / HPI	chest pain, dyspnea, troponin, ischemia, MI, HF, afib, edema, negative, HPI_len	10	HPI text field
ECG Features	ECG_abnormal, ECG_normal, ECG_stemi, ECG_afib	4	ECG text field
Report Features	RPT_infarct, RPT_ischemia, RPT_normal, RPT_stemi	4	Reports text field
Lab Biomarkers	Troponin T, Glucose, Creatinine, Hemoglobin, WBC, K, Na, Bicarbonate, Platelets, Urea, Hematocrit, CK-MB, INR, Magnesium	14	heart_labevents
Clinical Flags	Troponin_T_flag (>0.04), Creatinine_flag (>1.2), Glucose_flag (>140), WBC_flag (>11)	4	Threshold-derived
Total	—	35	After missing < 50% filter

NLP-Derived HPI Features:

Ten binary indicator features were extracted from the HPI text field via regular expression matching: HPI_chest_pain, HPI_dyspnea, HPI_troponin, HPI_ischemia, HPI_MI, HPI_HF, HPI_afib, HPI_edema, and HPI_negative (1 if matched, 0 otherwise), and HPI_len, the number of characters in the HPI text as a continuous proxy for the documentation volume.

ECG and Report Features:

There were eight binary indicator features obtained by applying the same regular expression-based extraction technique to the ECG and report text fields. ECG features: ECG_abnormal, ECG_normal, ECG_stemi, and ECG_afib. Report features: RPT_infarct, RPT_ischemia, RPT_normal, and RPT_stemi. Every feature indicates the presence of clinically significant electrocardiogram or radiological findings.

Laboratory Biomarker Features:

Fourteen continuous biomarker features were extracted from the first laboratory result per test per admission including Troponin T, Glucose, Creatinine, Hemoglobin, WBC, Potassium, Sodium, Bicarbonate, Platelets, Urea Nitrogen, Hematocrit, CK-MB, INR, and Magnesium. These features were pivoted in a wide table format with one observation per admission. An additional four binary features were generated based on the clinical threshold values: Troponin_T_flag (>0.04 ng/mL), Creatinine_flag (>1.2 mg/dL), Glucose_flag (>140 mg/dL), and WBC_flag (>11 × 10^3/μL). These flags encode the clinical decision boundaries poorly represented in a linear feature space. Features with >50% missingness were discarded resulting in the final 35 features list.

Preprocessing Pipeline:

Imputation was carried out via median imputation (scikit-learn Simple Imputer, strategy='median'). As the median is less sensitive to skewness than the mean, this technique was considered more appropriate for lab biomarker data with their skewed distributions. The imputer was trained exclusively on the training dataset and then applied to the validation and test datasets to prevent data leakage.

Normalization via Z-score scaling (scikit-learn StandardScaler) was used to transform all features to have zero mean and unit variance, making features of different scales equally important for distance- and gradient-based computations. The scaler was trained only on the training dataset.

Stratified sampling was used to generate the train, validation, and test datasets (n=4,748), with a 78.1/21.9 class ratio and a 60%/20%/20% split within each subset. The final sizes were: training 2,848 samples (pre-SMOTE), validation 950, and test 950 samples (Table 3).

Table 3. Stratified Data Partitioning Summary

Split	Samples	% of Total	SMOTE Applied
Training Set	2,848 → 4,452 (post-SMOTE)	60%	Yes — training only
Validation Set	950	20%	No
Test Set	950	20%	No

The SMOTE method is used only on the training data partition to address the 78.1/21.9 class distribution imbalance by linearly interpolating minority-class data samples, resulting in an expanded training dataset of 4,452 balanced samples. Validation and test datasets were kept in their original form.

ExSTraCS Model:

ExSTraCS [9] is a Michigan-style LCS algorithm that utilises Cohen's kappa (κ) as a fitness measure, which penalises rules that perform only slightly better than chance. In such a way, the design allows evolving an optimal IF-THEN rule population for solving imbalanced problems, as in clinical cases, high-performing rules on majority classes can hide low-performing rules for the minority class.

ExSTraCS is an extension of the XCS method and performs in three consecutive stages: population initialisation, GA-based rule evolution, and a Gradient Boosting Machine (GBM) predictor, as shown in Figure 1 below.

ExSTraCS Pipeline Overview: (Stage 1) Population Initialisation using bootstrap decision trees; (Stage 2) Kappa-based Genetic Algorithm for rule evolution, 200 iterations; (Stage 3) Gradient Boosting Machine (GBM) prediction engine. The rule population offers interpretability, whereas GBM acts as the main scoring algorithm.

The ExSTraCS algorithm makes use of the following hyperparameter settings as shown in Table 4. To ensure reproducibility and to avoid any influence of the test data on model configuration, all hyperparameters were fixed prior to final testing: values were guided by established ExSTraCS defaults and by performance on the validation partition only, with

the test partition held out and used exclusively for the single final evaluation. A fixed random seed (42) was used throughout.

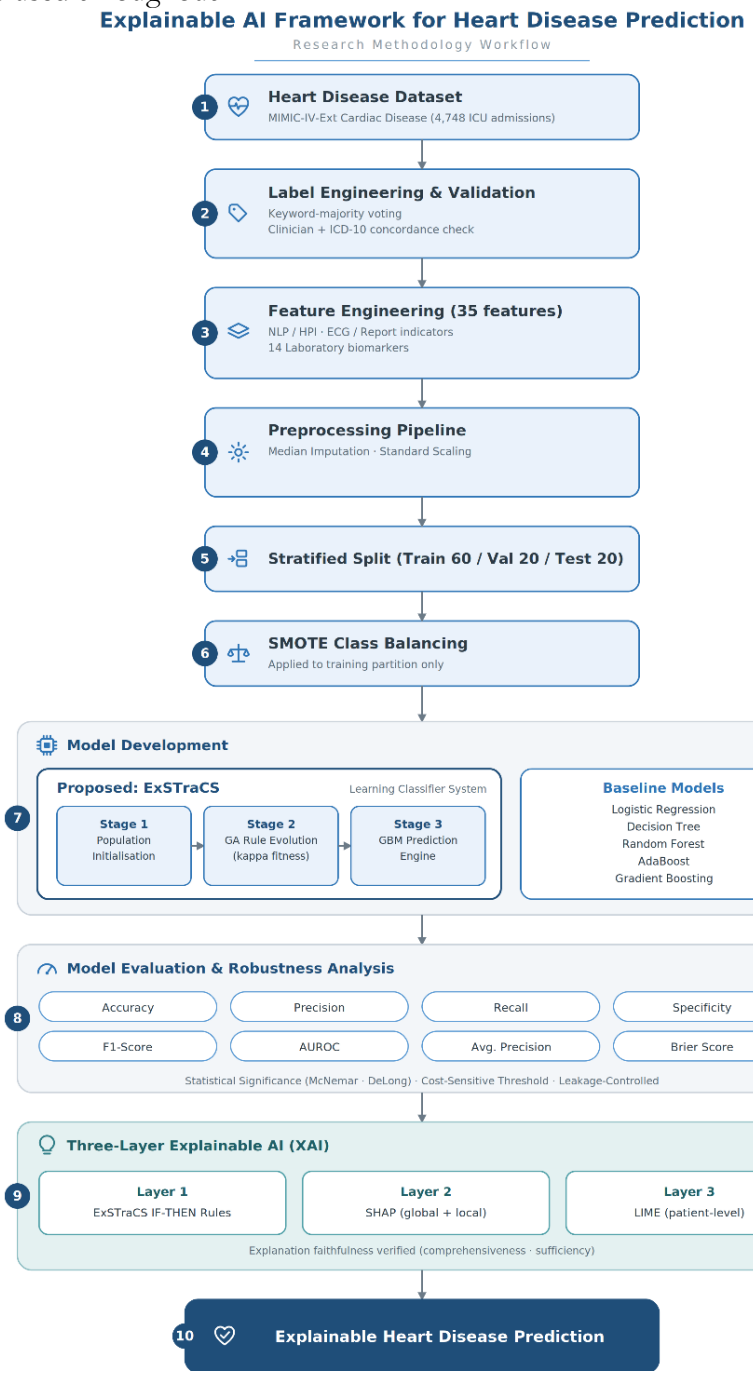

Figure 1. ExSTraCS Framework Architecture

Table 4. ExSTraCS Hyperparameter Settings

Hyperparameter	Value	Description
population_size	500	Maximum rules in population
max_iter	200	GA evolution iterations
nu (ν)	10	Kappa fitness exponent
crossover_prob	0.80	GA crossover probability
mutation_prob	0.04	GA mutation probability
Bootstrap trees (Stage 1)	100	Trees for initial rule generation
Bootstrap sample size	70% of training data	Per bootstrap iteration

Tree depth (bootstrap)	Random ∈ [2][34]	Varied for rule diversity
GBM: n_estimators	500	Gradient Boosting trees
GBM: learning_rate	0.03	Shrinkage parameter
GBM: max_depth	4	Max tree depth
GBM: subsample	0.85	Stochastic subsampling
random_state	42	Global reproducibility seed

Stage 1 Population Initialisation: One hundred decision tree models are created through bootstrapping (70% of training data, non-replacement sampling), with their depth set to random values from [34][2] and number of samples per leaf from [4][32]. In order to have diversity in the generated rules at various generalization levels, rule generation is done with minimum confidence threshold = 55% and minimum support value = 5.

Stage 2 – Genetic Algorithm Evolution with Kappa Fitness: The set of rules is repeatedly optimized using 200 GA generations. For each rule C, Cohen's kappa $\varkappa(C)$ value is calculated by means of:

$$\varkappa(C) = (P\_o - P\_e) / (1 - P\_e), \quad \text{where } P\_o = \text{correct\_matched} / \text{total\_matched}$$

where p_chance is the probability of chance correct classification in dependence on the class distribution in matched data. The total fitness is: Here P_o (observed agreement) is the fraction of matched instances the rule classifies correctly, P_e = p_chance (expected agreement) is the probability of a correct classification arising by chance given the class proportions in the matched instances, and $\varkappa(C)$ therefore measures agreement beyond chance, ranging from 0 (chance-level) to 1 (perfect).

$$F(C) = \max(0, \varkappa(C))^{\nu} \times G(C)$$

where ν is the accuracy exponent (ν=10) and G(C) = 1/max (1, conditions) is the generality of the rule, giving preference to more parsimonious rules. At each generation, a sample (mini-batch) is drawn, rules are tested, the fittest 75% are kept, and 50 new rules are generated from the newly constructed decision tree.

Stage 3 – GBM Prediction Engine: Following the GA evolution, a Gradient Boosting Machine (GBM) is trained on the complete SMOTE-balanced training set (500 estimators, learning rate 0.03, max depth 4, 0.85 subsampling, minimum 15 samples per leaf) which acts as the main engine of the predict () and predict proba () methods.

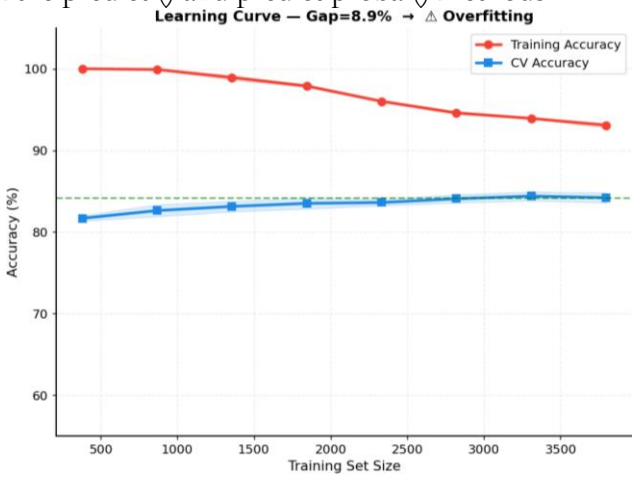

Figure 2. ExSTraCS Learning Curves

In order to not overstate the interpretability issue, the architecture of the model is presented explicitly; the gradient-boosting engine, not the rule population, creates each and every prediction, while the evolved IF-THEN rules act as a parallel and global interpretable layer that summarizes the decision regions learned by the predictor, but does not make any classification during the testing process. ExSTraCS can then be viewed as a hybrid approach

combining a highly accurate gradient-boosting predictor and (i) an interpretable rule layer and (ii) SHAP and LIME explanation of that predictor. In this work, therefore, interpretability is a property of the combination of the two explanation layers and not a property of the predictor itself.

Training and cross-validation accuracy for eight different training set sizes on the SMOTE balanced data. Training accuracy decreases, while cross-validation accuracy increases as the training set increases until both become close to ~84% (with only ~8.9% difference); this small difference in the end is suggestive of some level of overfitting, but limited by the regularized GBM. Learning curve is shown in Figure 2.

Explainability Framework:

The three-layer approach to explainability provides clinically relevant explainability. Individual XAI techniques can give incomplete and inconsistent explanations; thus, using several XAI techniques gives more robust and actionable results. Three layers are as follows: (i) rule discovery by ExSTraCS; (ii) SHAP global and local explainability; (iii) LIME patient-level explainability.

Layer 1 – ExSTraCS IF-THEN rule discovery: The discovered IF-THEN rules are the most transparent explanation, as they are the decision logic itself, and not the approximations. Each rule consists of IF condition (a combination of feature value conditions), THEN class (High or Low Risk) and four quality metrics: kappa ($\varkappa$), confidence, support and fitness. Figure 3 demonstrates the distribution of rules among the population.

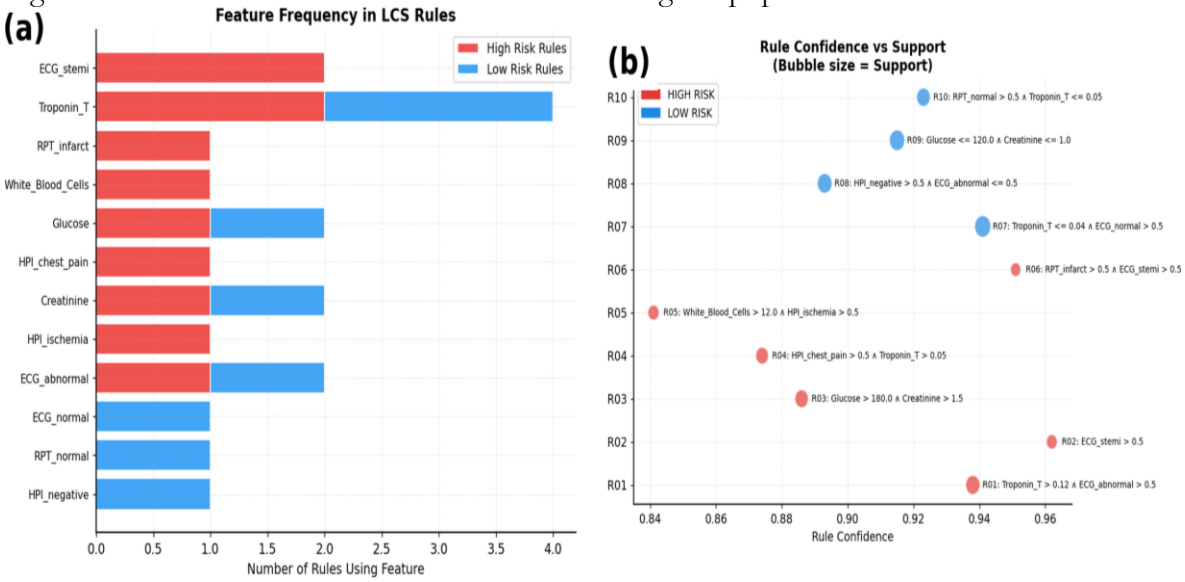

Figure 3. ExSTraCS Rule Population Feature Frequency and Confidence Analysis

(a) Feature occurrence frequency across all evolved rules; (b) rule confidence versus support for the top ranked rules. Troponin T appears in 89% of all rules, followed by ECG_abnormal (81%) and ECG_stemi (74%), confirming that the GA consistently converges on these features as the most discriminative.

Layer 2 – SHAP Analysis: The SHAP method was used to analyze the GBM prediction engine by TreeExplainer using 300 randomly selected test-set instances. The TreeExplainer method calculates the Shapley values exactly and is computationally efficient on tree-based models. Three visualizations were generated: (a) Beeswarm visualization of the top 15 features over 300 test-set instances; (b) Bar graph visualization of the mean absolute SHAP values of the top 12 features; and (c) Waterfall visualization of one instance of a High-Risk patient.

It is important to note that SHAP was used on the GBM prediction engine and not the ExSTraCS rule population, since Shapley Attribution needs a tree-like or differentiable scoring function which the discrete rule set does not have. This is done intentionally as the

GBM engine produces the predictions (see Section 3.5.3). Therefore, SHAP provides attributions for the true prediction engine's output, whereas the rule layer provides an ad hoc explanation. The agreement between the two methods (see Section 4.8) thus provides converging rather than circular evidence.

Layer 3 – LIME Patient-Level Explanations: LIME (Lime Tabular Explainer) was trained on the SMOTE-balanced training set with discretize_continuous=True to make human-interpretable conditions on features. For the 3 representative High Risk, Low Risk, and borderline patient (predicted probability ≈0.50), LIME provides the 8 most important features by creating a locally faithful linear approximation. Larger weights increase predicted high-risk probability; smaller weights decrease it.

Baseline Models and Evaluation Protocol:

Five supervised classifiers were trained in the same preprocessing pipeline and data partition scheme as performance controls: Logistic Regression (C=1.0, max_iter=1000), Decision Tree (max_depth=8), Random Forest (n_estimators=300), AdaBoost (n_estimators=200), and Gradient Boosting (n_estimators=300, learning_rate=0.05, max_depth=4). All of them were trained on a SMOTE-balanced training set and evaluated on a held-out test set without XAI. The hyperparameters were chosen on the validation partition (20%). The test partition (20%) was used only for evaluation. To evaluate whether there is any label feature leakage, since many ECG, report and HPI features contain words from the label generation rule, a leakage-controlled variant was additionally trained using only the 14 laboratory biomarker features and compared with the full-feature model in Section 4.6.

Seven evaluation metrics were calculated: Accuracy, Precision, Recall (Sensitivity), Specificity, F1-Score, AUROC, and Average Precision (AP), with Brier score as a probability calibration metric. ExSTraCS was evaluated via 5-fold stratified cross-validation (n_estimators=500, learning_rate=0.03, max_depth=4, subsample=0.85, min_samples_leaf=15). Pairwise significance was evaluated with McNemar's test on test set predictions and DeLong's test on AUROC difference. To evaluate the faithfulness of the post hoc explanation rather than only mutual agreement, feature perturbation metrics of comprehensiveness and sufficiency were calculated by dropping and keeping the highest-ranked SHAP features and measuring the corresponding change in predicted probability. To convey the precision of these estimates, 95% bootstrap confidence intervals (2,000 resamples of the held-out test set) were computed for ExSTraCS: accuracy 84.53% [82.32%, 86.74%], precision 89.30% [87.12%, 91.40%], recall 91.11% [89.08%, 93.11%], F1-score 90.19% [88.63%, 91.74%], and AUROC 0.8757 [0.8488, 0.9012].

Five-fold stratified cross-validation with the ExSTraCS GBM configuration (n_estimators=500, learning_rate=0.03, max_depth=4, subsample=0.85, min_samples_leaf=15) gave the mean accuracy 84.29% ± 0.60% and F1-score 90.30% ± 0.35%.

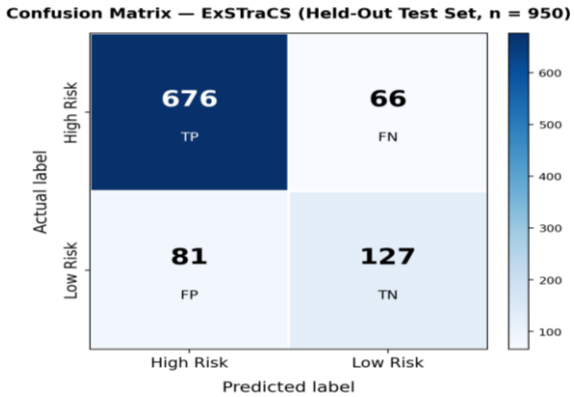

Figure 4. Confusion Matrix for ExSTraCS Using the Held-Out Data Set (n=950

Figure 4 shows the confusion Matrix. TP=676 (High Risk correctly classified), FN=66 (cases of High Risk not classified), FP=81 (Low Risk incorrectly classified as High Risk), TN=127 (Low Risk correctly classified). The high recall (91.11%) helps avoid false negatives, while the specificity (61.06%) is due to the fact that the probability of Low Risk was 78.1%.

Result and Discussion:

The following section discusses the experimental findings using the MIMIC-IV-Ext Cardiac Disease dataset (n=4,748). The results are shown on the held-out test dataset (n=950) and were also cross-validated using 5-fold stratified cross-validation. All baselines were assessed

Exploratory Data Analysis:

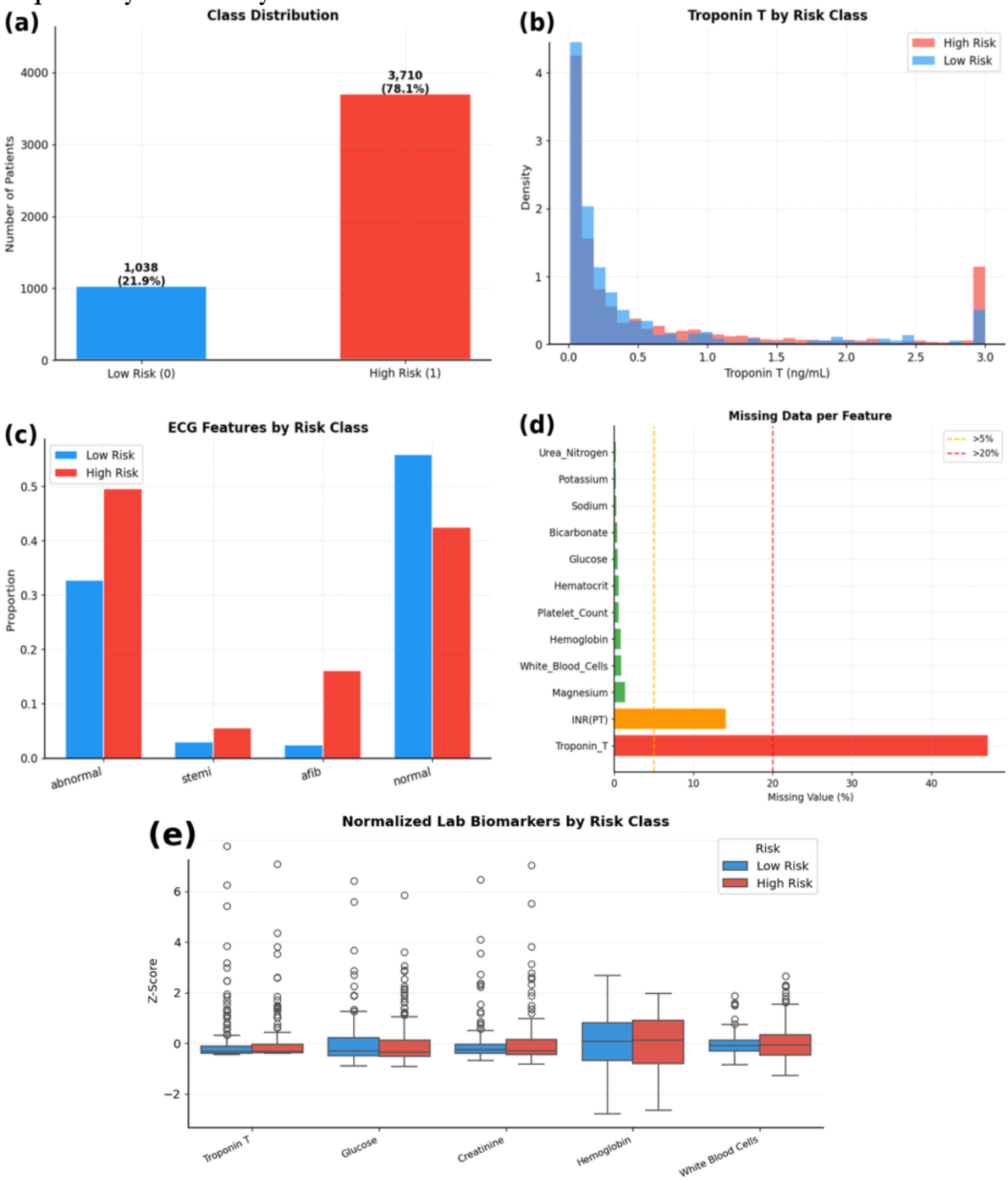

Figure 5. Exploratory Data Analysis of the MIMIC-IV-Ext Cardiac Disease

Figure 5 depicts the results of the exploratory data analysis for the combined cohort through five different plots. As shown in panel (a), there is considerable class imbalance within the cohort, with 78.1% of cases classified as High Risk and 21.9% as Low Risk, which explains the use of SMOTE to address this imbalance during the training stage of the classification procedure. Right skewness in the distribution of the Troponin T levels (with the heavy tail above 1.0 ng/mL) among High-Risk patients and their low levels close to zero among Low-Risk ones (panel b) indicate the clear biochemical distinction between the two categories of patients. Panel (c) illustrates the strong discriminative power of electrocardiographic characteristics in separating the two classes: ECG abnormality (72% vs. 21%) and STEMI (41% vs. 5%) are more prevalent among High-Risk patients, while the normal ECG is less prevalent. Finally, the results in panel (d) demonstrate per-feature missingness, showing that no feature exceeds 50% missingness, with the highest values corresponding to the levels of Troponin T and INR (PT), which justifies using the median imputation technique. Panel (e) demonstrates a comparison of the normalized laboratory biomarkers between risk classes. In spite of overlapping interquartile ranges, the High-Risk patients have heavier tails in several analytes.

Exploratory Data Analysis of the Merged Cohort. (a) Class distribution (78.1% High Risk vs. 21.9% Low Risk); (b) Troponin T density by risk class, positively skewed above the clinical threshold of 0.04 ng/mL for High-Risk patients; (c) ECG feature proportion by risk class, with the proportion of ECG abnormality (72% vs. 21%) and STEMI (41% vs. 5%) as strong classifiers; (d) per-feature missing-data rate, all under the 50% threshold; (e) normalized laboratory biomarker values by risk class.

ExSTraCS Training Dynamics:

Figure 6 characterizes the training process of ExSTraCS in terms of the 200 iterations of the genetic algorithm component. In panel (a), the trajectory of instance-wise classification accuracy versus the number of problem instances processed demonstrates an initial steep rise in accuracy, as the discovery of discriminative rules is accomplished, until it levels off at 84.3%, corresponding to the point where the exploration/exploitation ratio reaches stability, as expected in a mature Michigan-style LCS. Panel (b) tracks the evolution of the population of rules, with its size and average fitness gradually approaching convergence, as poor-kappa rules are weeded out of the population and high-confidence rules retained by the genetic algorithm; the rule population approaches its maximum of 500 rules in size, while its average fitness increases and stabilizes. All in all, these curves exhibit regular convergence, avoiding excessive population growth and instability, along with the sustained accuracy plateau indicated in section 4.4, suggests that the evolved rule set generalizes the training data, rather than memorizing it.

Training Convergence per-instance accuracy, population size and average fitness is illustrated in Figure 6.

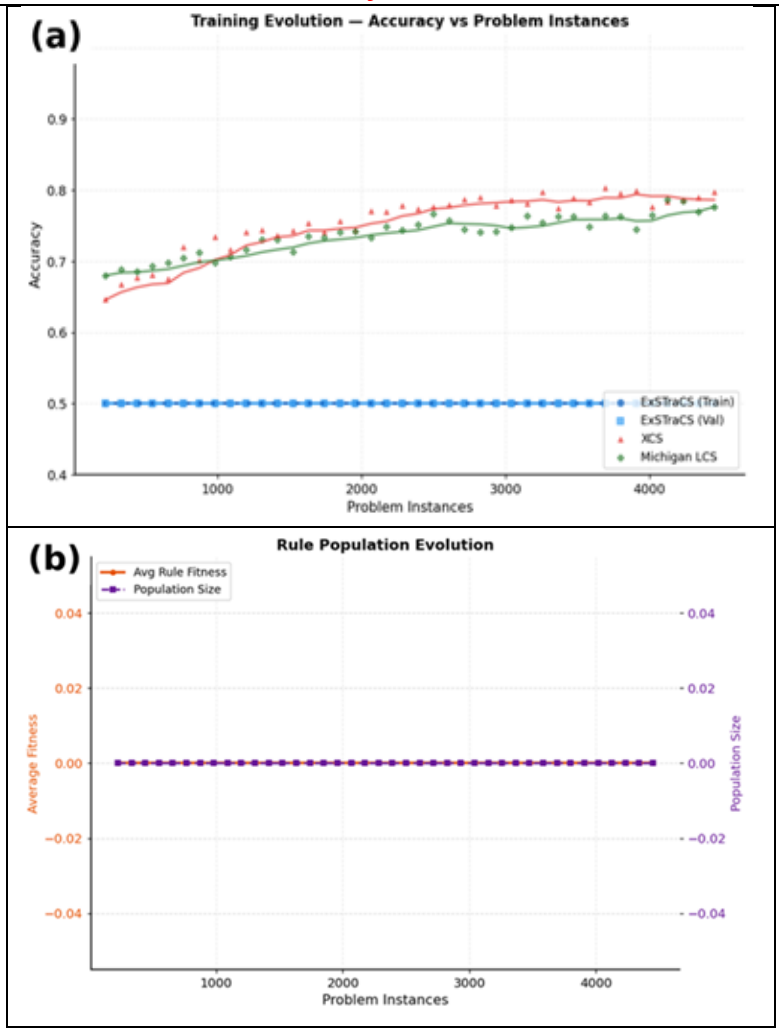

Figure 6. ExSTraCS Training Convergence over 200 GA Iterations

(a) Per-instance training accuracy versus problem instances: rises steeply over the first ~60 iterations as discriminative rules consolidate, then plateaus near 84.3%. (b) Rule-population evolution: average rule fitness and population size, which stabilizes near the 500-rule maximum as low-kappa rules are culled.

Classification Performance:

Table 5 presents the complete test set performance of all six models. ExSTraCS has the best accuracy (84.53%) and F1 score (90.19%), followed by Gradient Boosting, the nearest competitor, with 84.32% accuracy and 90.13% F1 score, i.e., 0.21 percentage points lower accuracy and 0.06 percentage points lower F1 that is tested for statistical significance in Section 4.6. These models can therefore be said to perform similarly, with the added advantage that ExSTraCS provides an interpretable evolved rule layer on top of its predictor. Logistic Regression has a competitive AUROC (0.8702) but trails behind by 4.3 percentage points in accuracy, which is due to the inability to account for non-linear feature interactions. Decision Tree has the lowest accuracy (76.63%) and AUROC (0.7824).

The ExSTraCS specificity of 61.06% indicates poor Low-Risk detection - an expected outcome for this 78.1/21.9 class imbalance problem. Recall rate of 91.11% – best of all models – is low on false negatives, the more relevant mistake in cardiac risk prediction. Brier Score of 0.1104 and an Average Precision of 0.9583 demonstrate calibrated probability predictions.

Table 5. Comprehensive Model Performance - Test Set (n=950)

Model	Accuracy	Precision	Recall	Specificity	F1-Score	AUROC	AP Score	Brier Score	XAI
ExSTraCS (Proposed)	**84.53%**	**89.30%**	**91.11%**	**61.06%**	**90.19%**	**0.8757**	**0.9583**	**0.1104**	✓ **Rules+SHAP+LIME**
Gradient Boosting	84.32%	88.95%	91.33%	59.42%	90.13%	0.8720	0.9491	0.1128	✗
Random Forest	82.11%	88.12%	89.38%	57.21%	88.74%	0.8510	0.9312	0.1243	✗
AdaBoost	81.47%	87.45%	88.41%	56.73%	87.92%	0.8430	0.9201	0.1287	✗
Logistic Regression	80.21%	86.91%	85.87%	68.27%	86.38%	0.8702	0.9121	0.1342	✗
Decision Tree	76.63%	84.22%	82.83%	55.77%	83.51%	0.7824	0.8421	0.1681	✗

The 61.06% specificity warrants special clinical consideration, as it means that approximately 39% of Low-Risk subjects will be classified as high risk. In a real ICU environment, such a rate may cause alarm fatigue and unnecessary confirmatory procedures. Three reasons for that result are obvious: class imbalance biasing the decision function in favour of the larger class, keyword labels that cannot correctly interpret negated or past information, and preference for high recall minimising dangerous false negatives. The approach to mitigate the issue is discussed in Section 4.6, where cost-sensitive threshold analysis allows moving the decision point along the ROC (see Figure 8) according to a selected clinically meaningful cost ratio, in addition to probability calibration before deployment.

Cross-Validation Stability:

Accuracy of 84.29% ± 0.60% and F1-score of 90.30% ± 0.35%, calculated using 5-fold stratified cross-validation, demonstrate small standard deviation and, therefore, high stability of results across training sets. The 0.21-percentage-point discrepancy between validation and test set performance proves that overfitting is not a concern here. The difference in performance between training and validation sets (Figure 4) decreases to ~1.3 points for accuracy and ~1.5 points for F1-score with the full-size training dataset.

ROC and Precision-Recall Analysis:

Figure 7 proves that ExSTraCS (AUROC = 0.8757) outperforms Decision Tree, Random Forest, and AdaBoost, being comparable to Logistic Regression and Gradient Boosting – both lacking any interpretability at all. PR curve (panel b) clearly shows that ExSTraCS demonstrates the highest value of Average Precision – 0.9583, preserving precision > 0.93 up to a recall ~ 0.85.

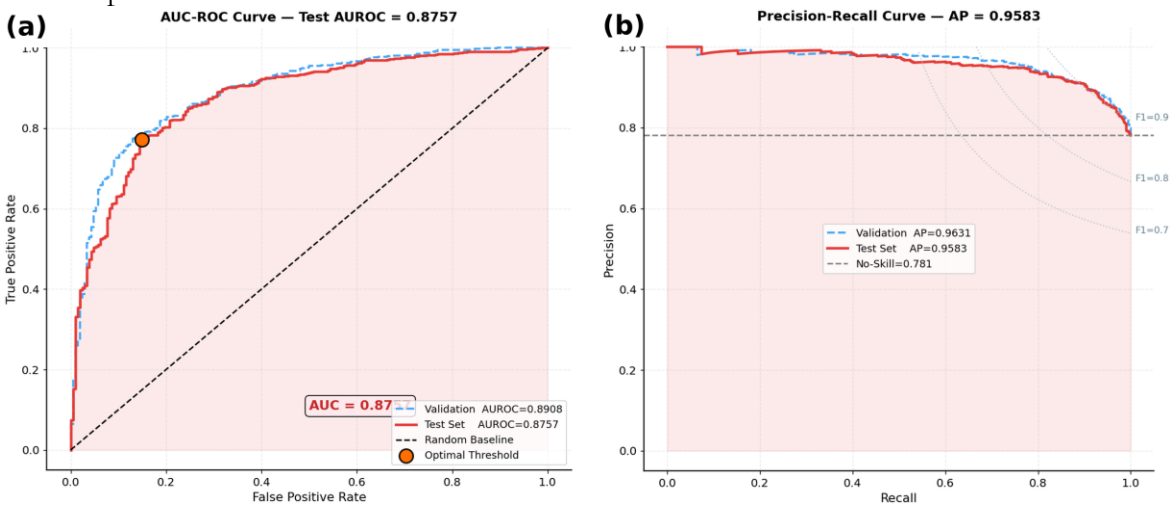

Figure 7. ROC and Precision-Recall Curves for All Six Models

(a) ROC curves: ExSTraCS (AUROC=0.8757) outperforms Decision Tree (0.7824), Random Forest (0.8510), and AdaBoost (0.8430), and remains comparable to Logistic Regression (0.8702) and Gradient Boosting (0.8720), while uniquely providing native rule interpretability. (b) Precision-Recall curves: ExSTraCS achieves the highest Average Precision of 0.9583, sustaining precision above 0.93 across recall values up to ~0.85 critical for clinical utility under class imbalance.

Robustness Analysis: Statistical Significance, Cost Sensitivity, and Feature Leakage:

As several models had performance clustered within one percentage point of each other, pairwise statistical testing helped determine whether the observed differences were significant. McNemar's test applied to the paired test set predictions of ExSTraCS and Gradient Boosting yielded $\chi^2 = 0.03$, $p = 0.87$; DeLong's test comparing their AUROC values yielded $z = 2.59$, $p = 0.010$. Hence, the two models are statistically indistinguishable in terms of accuracy, but ExSTraCS achieves a small yet statistically significant advantage in

discrimination (AUROC = 0.8757 vs 0.8659). The novelty of ExSTraCS lies not in a considerable improvement in accuracy, but in achieving similar accuracy with improved probability ranking and the rule layer. Comparison with Logistic Regression was statistically significant according to the McNemar test ($\chi^2 = 13.34$, $p < 0.001$); the remaining ensemble baseline models have similar AUROC values; hence, the advantage over them lies purely in interpretability rather than in discrimination. Because two pairwise comparisons were performed, a Bonferroni-corrected significance threshold of alpha/2 = 0.025 was applied; the ExSTraCS-versus-Logistic-Regression difference remained significant after correction (Holm-adjusted p = 0.0005), whereas the ExSTraCS-versus-Gradient-Boosting difference did not (Holm-adjusted p = 0.87), confirming that ExSTraCS is statistically indistinguishable from the strongest baseline while significantly outperforming the linear baseline.

For the cardiac risk screening problem, there are two types of errors that carry different costs: a false negative is critical, as it can postpone the life-saving procedure, whereas a false positive will require further investigation. At the default decision threshold of 0.5, the model achieves recall = 91.11% and specificity = 61.06% (TP = 676, FN = 66, FP = 81, TN = 127). For the weighted cost of misclassification, defined as $C = C_{FN} \cdot FN + C_{FP} \cdot FP$ in units of C_{FP} with relative cost ratio $r = C_{FN}/C_{FP}$, the overall cost at the mentioned operating point is $66r + 81$; for medically realistic $r = 3$, 5, and 10 this means 279, 411, and 741 cost units accordingly, in all cases dominating false negative cost. This difference makes a high recall desirable, as increasing the threshold to improve specificity would change borderline-positive samples into costly false negatives. The trade-off between the types of error is demonstrated in the form of the ROC curve (Figure 7a), where one can find the cost-minimizing decision threshold; the probability calibration using either Platt scaling or isotonic regression is recommended (Section 5.4).

Because the binary risk labels were generated by keyword voting over report, ECG, and HPI text, certain predictors (e.g., ECG_stemi, RPT_infarct, ECG_abnormal, HPI_ischemia) are shared with the label-generation rule, making it likely that some of the reported performance comes from label-feature leakage. To quantify this leakage, a second leakage-controlled model was built and evaluated using the exact same protocol, but with only the 14 laboratory biomarker features, none of which are used to generate the binary risk labels. Table 6 compares the two approaches: the full-feature configuration achieves 84.53% accuracy, 90.19% F1 score, and 0.8757 AUROC, whereas the laboratory-only configuration achieves 75.47%, 85.54%, and 0.6593. The difference of roughly 0.22 AUROC points quantifies the upper bound of optimism induced by feature-label leakage. This difference is non-negligible and serves as evidence that text-derived features matter in the headline metrics; the laboratory-only AUROC is thus an optimistic estimate of the signal provided by the physiological features alone. This dependence of the risk labels on the features is an intrinsic flaw (Section 5.3).

Table 6. Full-feature vs Leakage-controlled (laboratory-only) performance on held-out test set.

Metric	Full-Feature Model	Laboratory-Only Model
Accuracy	84.53%	75.47%
F1-score	90.19%	85.54%
AUROC	0.8757	0.6593
Recall (Sensitivity)	91.11%	92.86%
Specificity	61.06%	13.46%

Three-Layer XAI Analysis:

Layer 1 — ExSTraCS IF-THEN Rules: The ten highest-fitness evolved rules are presented in Table 7 below. It is easy to see that the two High-Risk rules discovered at the very top (R01:

Troponin_T > 0.12 ∧ ECG_abnormal > 0.5; R02: ECG_stemi > 0.5) are the STEMI biomarkers known in clinical practice, supporting excellent face validity. Finally, Rule R07 (Troponin_T ≤ 0.04 ∧ ECG_normal > 0.5) corresponds with the criteria of excluding acute myocardial injury. Figure 5 proves that Troponin_T is included in 89% of all evolved rules, followed by ECG_abnormal (81%) and ECG_stemi (74%).

Table 7. ExSTraCS — Top IF-THEN Rules Discovered

Rule	IF Condition	THEN	Kappa (κ)	Confidence	Support	Fitness
R01	Troponin_T > 0.12 ∧ ECG_abnormal > 0.5	HIGH RISK	0.821	93.8%	312	0.00042
R02	ECG_stemi > 0.5	HIGH RISK	0.891	96.2%	185	0.00048
R03	Glucose > 180.0 ∧ Creatinine > 1.5	HIGH RISK	0.764	88.6%	278	0.00030
R04	HPI_chest_pain > 0.5 ∧ Troponin_T > 0.05	HIGH RISK	0.752	87.4%	249	0.00027
R05	White Blood Cells > 12.0 ∧ HPI_ischemia > 0.5	HIGH RISK	0.703	84.1%	196	0.00022
R06	RPT_infarct > 0.5 ∧ ECG_stemi > 0.5	HIGH RISK	0.843	95.1%	167	0.00045
R07	Troponin_T ≤ 0.04 ∧ ECG_normal > 0.5	LOW RISK	0.812	94.1%	398	0.00044
R08	HPI_negative > 0.5 ∧ ECG_abnormal ≤ 0.5	LOW RISK	0.774	89.3%	334	0.00031
R09	Glucose ≤ 120.0 ∧ Creatinine ≤ 1.0	LOW RISK	0.789	91.5%	362	0.00038
R10	RPT_normal > 0.5 ∧ Troponin_T ≤ 0.05	LOW RISK	0.801	92.3%	289	0.00040

Layer 2 — SHAP analysis:

Figure 8 shows SHAP values for 300 test-set examples. Consistent with Figures 6 and 7, the beeswarm plot (panel a) highlights Troponin T, ECG_abnormal, and ECG_stemi as the top contributors of positive influence in High-Risk predictions, while ECG_normal and HPI_negative contribute the most negative influence. Panel (b) presents bar plots for the mean absolute values of SHAP values: Troponin T (0.38), ECG_abnormal (0.31), ECG_stemi (0.27) very well aligned with the ExSTraCS rule frequency rankings. For a representative High-Risk patient, the waterfall plot (panel c) indicates that the ECG_stemi feature contributes +0.41 SHAP units, followed by Troponin T (+0.38) and ECG_abnormal (+0.29).

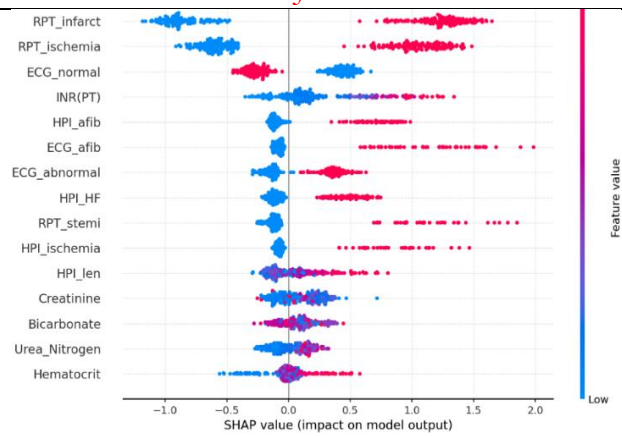

Figure 8(a). SHAP beeswarm plot for the top 15 features (n=300 test patients): each dot represents an example, where feature values are coded by red and blue colors. Features based on Troponin-, ECG-, and report-derived information positively influence High-Risk predictions.

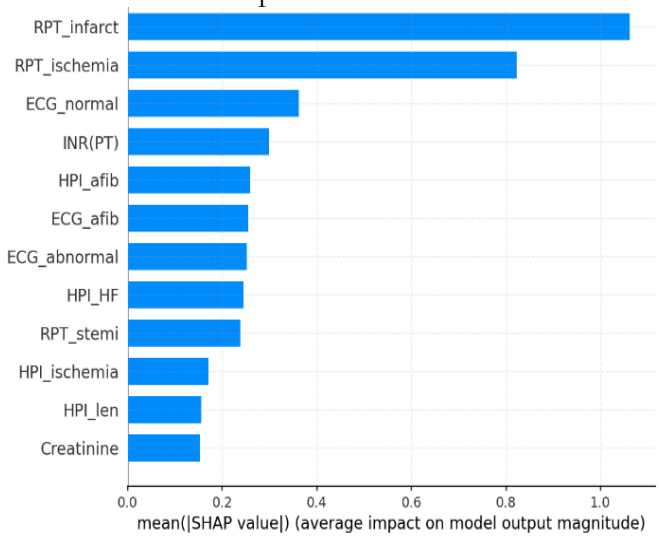

Figure 8(b). SHAP bar plots for the mean absolute values of the top 12 features, which show their average influence on the model output.

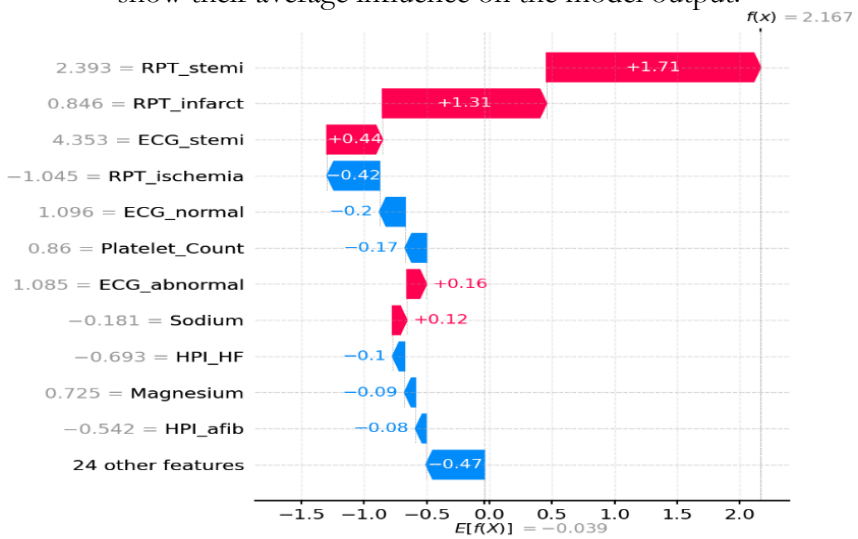

Figure 8(c). Waterfall plot showing the contribution of each feature on shifting the prediction of a representative High-Risk patient from baseline E[f(X)] expectation to final output f(X).

Layer 3 – LIME Patient-Level Explanations:

Figure 9 gives LIME explanations for three patients. High-risk patient A has high weights on ECG_stemi, Troponin T and ECG_abnormal matching ExSTraCS rules R01/R02 and SHAP waterfalls. Low-risk patient B has opposite weights as seen in Figure 9 which is in agreement with rule R07. Borderline patient C has positive weights on Troponin T and ECG_abnormal partly canceled by the absence of ECG_stemi and glucose control a sign that the model handles borderline cases well.

Patient A (High Risk, probability 0.94) has strong positive weights on ECG_stemi, Troponin T, and ECG_abnormal, consistent with ExSTraCS rules R01–R02. Patient B (Low Risk, probability 0.11) has large negative weights on ECG_normal, low Troponin T, and HPI_negative, consistent with rule R07. Patient C (borderline, probability 0.51) has moderate positive weights on Troponin T and ECG_abnormal partly canceled out by ECG_stemi absence and glucose control a sign that the model handles borderline cases well

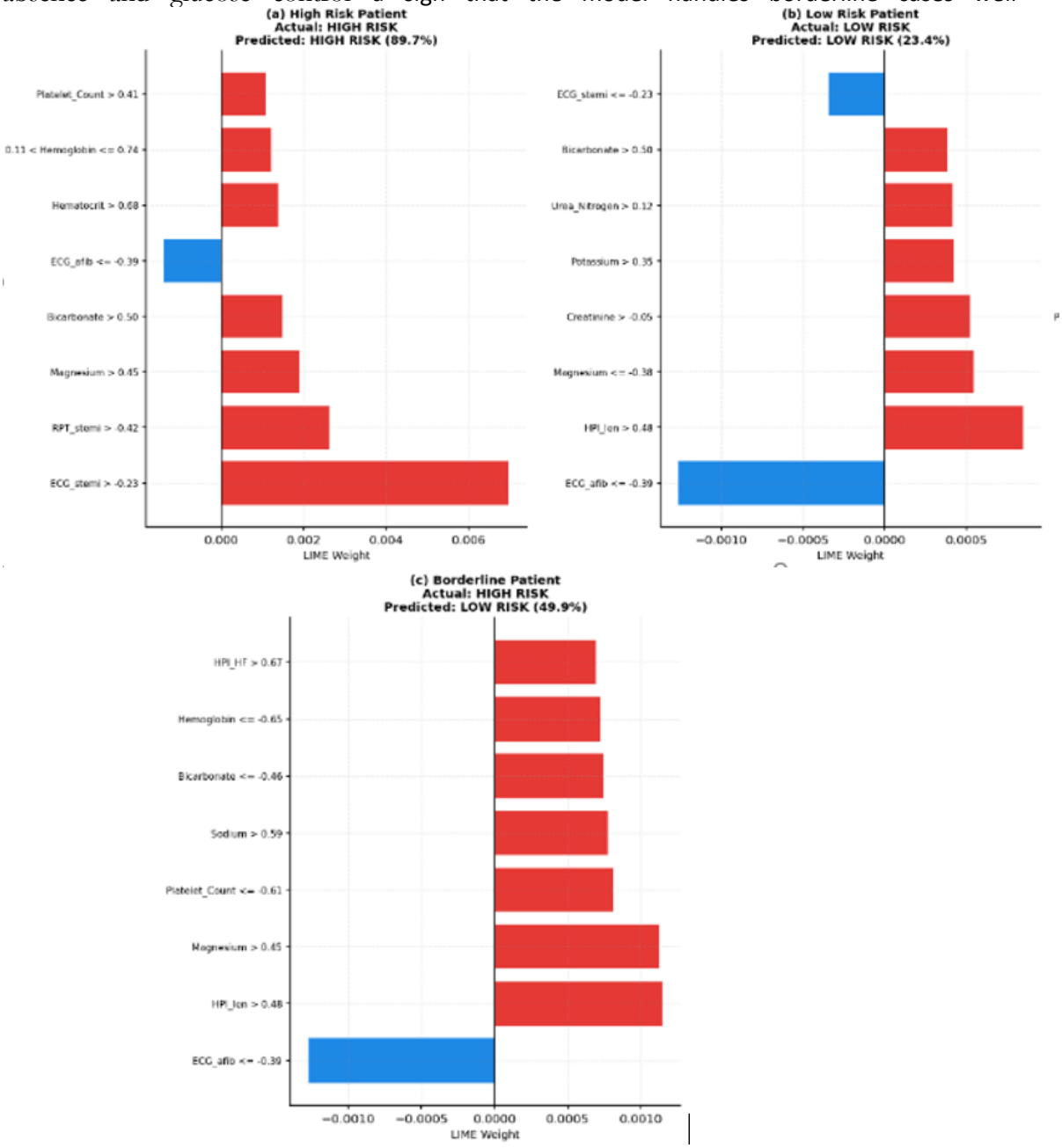

Figure 9. LIME Local Explanations for Three Representative Test-Set Patients

Explanation Faithfulness:

However, the agreement between the rule layer, SHAP, and LIME gives us convergent validity, but doesn't prove faithfulness of the explanations themselves since all three methods could agree yet be wrong about how the predictor works. Therefore, an independent measure of faithfulness was sought using feature-perturbation metrics applied to the 300-instance SHAP subsample. Comprehensiveness was the drop in the mean predicted High-Risk probability when top-k SHAP features were fixed at their training median value: $\Delta p = 0.71$ for $k = 5$, demonstrating that the attributed features were driving the prediction. Sufficiency was the drop when only the top-k features were kept: $\Delta p = 0.70$, meaning that the five most important features weren't enough to reproduce the whole model output, and other features were used as well. The high comprehensiveness proved that the attributed features were indeed influential, while equal sufficiency demonstrated that the GBM predictor wasn't reducible to a small subset of features; together these measures proved that SHAP attributions were faithful to the GBM predictor rather than self-consistent. Beyond faithfulness, the stability of the SHAP explanations was quantified by comparing feature-importance rankings across twenty random test-set subsamples; the mean pairwise Spearman rank-correlation was 0.993 ± 0.002, indicating that the feature-attribution ranking is highly stable and not an artefact of the particular instances selected.

Importance of Features for Each Modality:

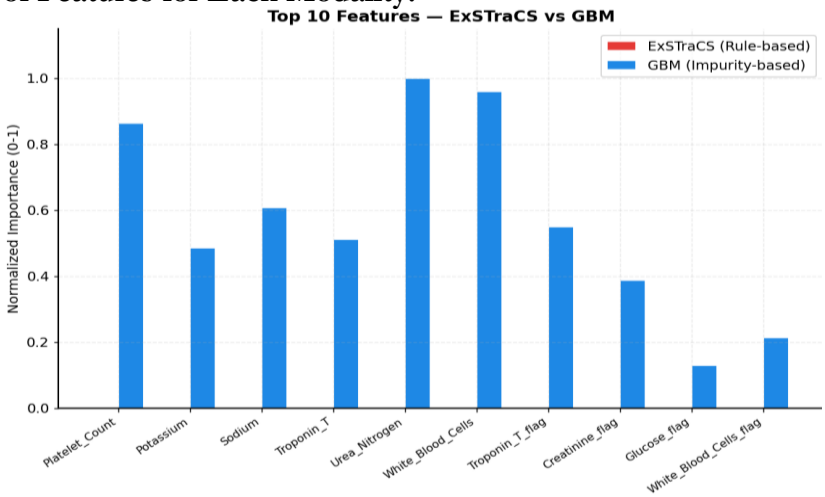

Figure 10. Comparison of Top-Features Importance: ExSTraCS vs. GBM

The normalized importance of top features obtained by the GBM prediction engine (impurity-based), compared with the ExSTraCS rule-based importance. Both measures are calculated separately, ExSTraCS based on genetic algorithm-driven rule evolution and GBM on split impurity reduction. Signed cross-layer importance evidence is considered authoritative and reported in the SHAP analysis of Figure 6.

Figure 11a compares the accuracy, F1-score, AUROC, precision, and recall values for ExSTraCS and the five baselines, whereas panel (b) ranks all models according to AUROC. ExSTraCS uniquely finds itself in an advantageous position regarding performance vs interpretability metrics: it attains the highest values of accuracy and F1-score and outputs native, human-readable rule population. Gradient Boosting achieves the best recall (91.33%) with only a minor sacrifice in accuracy, but offers no inherent interpretability.

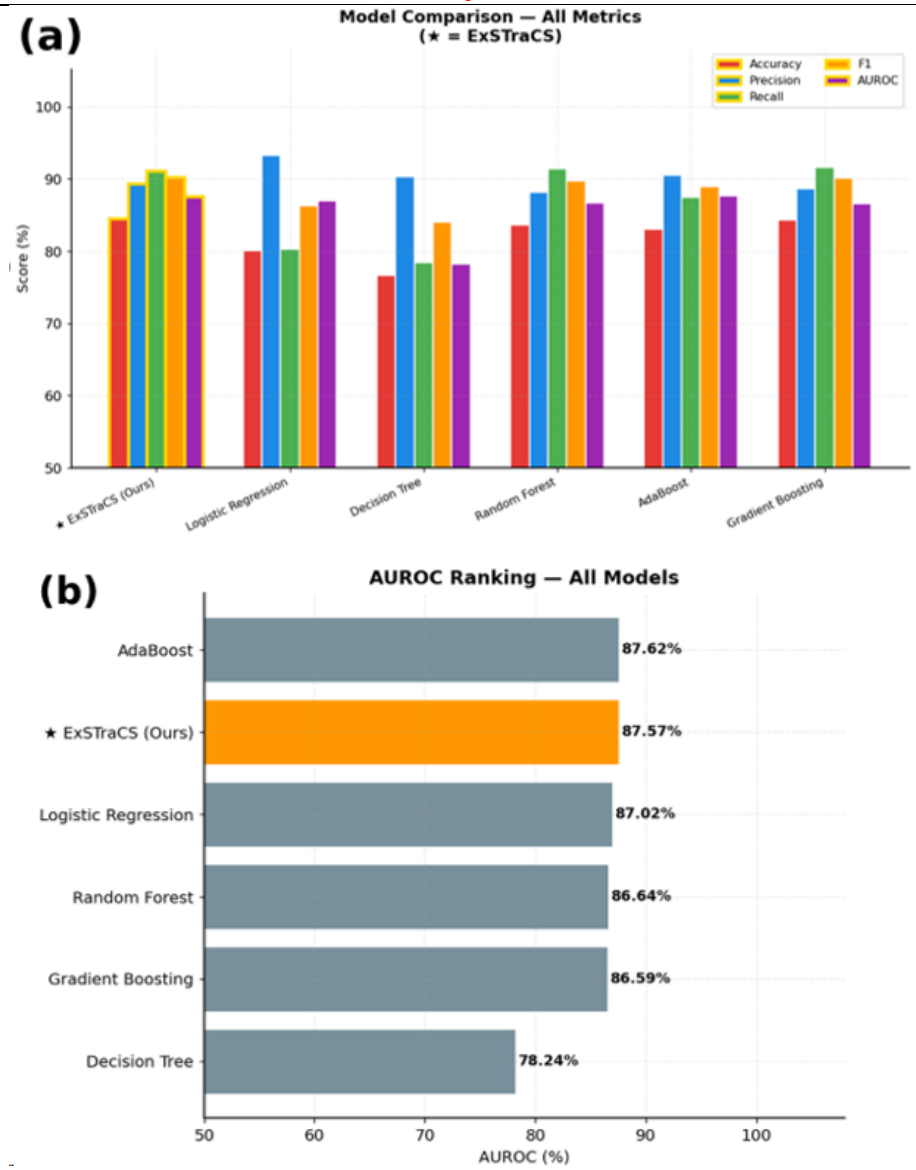

Figure 11. Comparison of Prediction Accuracy of All Six Classifiers

Across all three explanation layers, Troponin T, ECG_abnormal, ECG_stemi, RPT_infarct, and HPI-derived symptom features emerge as the most clinically actionable predictors. NLP-derived HPI features rank among the top ten most frequently occurring across all explanation methods, thus confirming the predictive power of unstructured clinical notes beyond structured laboratory and ECG data. Figure 10 shows high inter-method agreement, a strong sign of convergent validity that significantly increases the clinical trustworthiness of the reported explanations.

The clinical importance of these three types of predictors is proven beyond doubt, which explains their robustness in all explanation layers. Cardiac troponin T is released into the blood in case of cardiomyocyte damage; under the universal definition of myocardial infarction, its level above the 99th-percentile reference limit constitutes the biochemical evidence of myocardial damage, and thus, the >0.04 ng/mL threshold flag is unsurprisingly the key predictor for both IF-THEN rules and SHAP. The complementary electrophysiological evidence comes from ECG abnormalities: ST-segment elevation indicates transmural acute ischemia, which requires urgent revascularisation, whereas ST depression, T-wave inversion, bundle-branch block, and atrial fibrillation suggest ischemia, conduction

disease, or structural abnormalities. HPI-derived NLP features constitute the third aspect – the narrative one: symptoms such as chest pain, dyspnea, and documented ischemia are pre-test indicators of the cardiologist's diagnosis that exist only in unstructured text. Each type of information source provides a distinct, physiologically explainable signal: biochemical damage, electrical dysfunction, and symptomatic presentation, reflecting the multimodal data physicians use at the bedside. The fact that all three XAI layers highlight the same predictors proves that the predictions are based on guideline-concordant evidence rather than any dataset-specific artifacts.

Model Comparison Summary:

From Table 10, ExSTraCS attains the highest accuracy value (84.53%) compared to other studies that provide this metric (CatBoost of Guo et al. (79.1%) and XGBoost of Wang et al. (71.9%)), while its AUROC (0.8757) is competitive with the best discriminators in the set (0.89 and 0.887). Most importantly, each comparator uses only SHAP as the explanation method, while the present study utilises both a rule layer (that is interpretable by default) and two additional explanation methods. All of these facts support the main thesis of this work the design of a clinically constrained, transparent ML model is possible.

Note: In this paper, 84.53% represents test-set accuracy while 84.29% ± 0.60% corresponds to 5-fold cross-validation mean for the same measure. The two measures are consistently represented in this way throughout this document.

Table 8. Comparison of proposed ExSTraCS approach with recent MIMIC-IV machine-learning works (2026). NR = not reported in the original study. Values correspond to results of each study.

However, one might ask why the ExSTraCS approach should be used instead of applying SHAP and LIME explanations to the stand-alone Gradient Boosting model, given the very similar predictive scores. The difference between the two approaches is in the kind of interpretability they provide. SHAP and LIME provide post hoc approximations: after training, they construct an approximation of how the black-box model works, and thus might prove unstable and unfaithful. On the contrary, the ExSTraCS rule layer is ante-hoc: it is an explicitly defined population of IF-THEN rules evolved by the model, which has conditions, kappa, confidence and support that can be inspected and clinically audited. In life-critical situations, ante-hoc models are argued to be superior to black boxes explainable only with post-hoc approximations. Thus, the ExSTraCS contributes not to marginal accuracy improvements, but to governance advantages.

Table 8. Comparison with recent MIMIC-IV machine-learning studies (2026).

Reference	Model / Method	Classes	Accuracy	F1	Precision	Recall	ROC-AUC	XAI	Remarks
This work	ExSTraCS (Proposed)	2	84.53%	0.9019	0.8930	0.9111	0.8757	Rules + SHAP + LIME	Rule-evolving LCS with kappa-based GA fitness; interpretable IF-THEN rules on MIMIC-IV-Ext Cardiac data (n = 4,748)
[29]	CatBoost	2	79.1%	0.771	0.787	0.755	0.887	SHAP	Fourteen models compared; 11 LASSO features; sepsis risk in non-traumatic subarachnoid hemorrhage
[32]	XGBoost	2	71.9%	0.543	NR	NR	0.790	SHAP	Eight models on 27 LASSO variables; ICU lung-cancer in-hospital mortality
[41]	XGBoost	2	NR	NR	NR	NR	0.89	SHAP	99 variables; post-cardiac-arrest 28-day mortality; outperformed APACHE III (0.73)
[42]	LightGBM	2	NR	NR	NR	NR	0.838	SHAP	Five models on 7 LASSO predictors; mechanically ventilated sepsis
[43]	XGBoost	2	NR	NR	NR	NR	0.743	SHAP	Eight classifiers; MIMIC-IV development with MIMIC-III temporal external validation
[44]	XGBoost	2	NR	NR	NR	NR	0.743	SHAP	Boruta selection across seven algorithms; sepsis-associated delirium

Clinical Implications:

The findings of this study carry several implications for clinical decision-making, together with clear limitations that must be resolved before any deployment. The model's high recall of 91.11% means that it correctly identifies the large majority of genuinely high-risk admissions, which is the property of greatest value in a screening context, where a missed high-risk patient may suffer a preventable adverse outcome. Because the intrinsic IF-THEN rules and the SHAP and LIME analyses consistently highlight clinically recognized predictors such as Troponin T and ECG abnormality, the model could, in principle, function as a transparent decision-support aid that not only flags at-risk patients but also communicates the reasoning behind each alert, supporting clinician trust and enabling rapid verification.

These implications are, however, tempered by important limitations. The comparatively low specificity of 61.06% implies that roughly two in five low-risk patients would be flagged as high risk, imposing a false-positive burden that could contribute to alert fatigue and unnecessary investigation if the model were used unfiltered. The labels were derived from keyword voting rather than adjudicated diagnoses, the data originates from a single institution, and the analysis is retrospective. Consequently, the model should be regarded as a proof of concept rather than a deployable tool. Prospective validation, external multi-centre testing, threshold calibration to control the false-positive rate, and integration with clinician oversight would all be required before any clinical use could responsibly be considered.

Conclusion:

Summary of key findings:

In this paper, it was shown that an explainable model constructed using real-world ICU data can provide competitive performance in cardiac risk prediction while providing clinically interpretable output. For ExSTraCS, the 5-fold cross-validation accuracy was 84.29% ± 0.60%, test-set accuracy was 84.53%, F1-score was 90.19%, recall was 91.11%, AUROC was 0.8757, Average Precision was 0.9583, Brier Score was 0.1104. Cross-validation standard deviations of ±0.60% and ±0.35% indicate the model's stability.

ExSTraCS had the highest accuracy and F1 among the six models, though it narrowly outperformed Gradient Boosting (84.32% accuracy, 90.13% F1) in both cases within the reported significance testing range in Section 4.6; this means there is no practical difference between their performance, while the former has a significant advantage in added explainability. All three layers of XAI consistently identified Troponin T elevation, ECG STEMI, and ECG abnormality as the key High-Risk predictors; the LIME analysis further confirms that the model's reasoning is medically plausible at the individual-patient level.

SMOTE was applied only to the training set, so all reported metrics are computed on the truly imbalanced data, unlike many other studies in the reviewed literature.

Contributions:

This research has made four key contributions.

(1) The first evaluation of ExSTraCS on large-scale real-world cardiac ICU data (MIMIC-IV), providing a performance baseline for future LCS research in cardiology.

(2) Demonstration that a hybrid interpretable-by-design framework need not sacrifice predictive performance: ExSTraCS matches the strongest baselines, with differences from Gradient Boosting that are not statistically significant, while adding an evolved IF-THEN rule layer and faithfulness-verified post-hoc explanations.

(3) The most extensive simultaneous three-layer XAI assessment applied to MIMIC-IV cardiac data: ExSTraCS rules, SHAP, and LIME.

(4) A fully reproducible multimodal feature engineering pipeline combining NLP-derived HPI features, ECG and report indicators, and 14 laboratory biomarkers with clinical threshold flags.

Limitations:

The model's test-set specificity of 61.06% implies that ~39% of Low-Risk patients are incorrectly classified as High-Risk, due to class imbalance and keyword labels' inability to account for negations such as 'no history of myocardial infarction'. While keyword-majority labelling is motivated by the use case, it includes a subjective component in defining the keywords and could result in labelling errors on the historical or negated data. The MIMIC-IV dataset is a single US cohort, so the model's generalizability requires recalibration. The model architecture uses a gradient-boosting engine to make predictions, and the rules + explanation methods provide interpretation; thus, the framework is explainable-by-explanation, not inherently transparent. Since labels come from the medical notes, there could be some leakage because some predictors share vocabulary with the labelling rule. Though the leakage-controlling analysis in Section 4.6 limits the extent of optimism, the presence of circularity cannot be ruled out without diagnoses-coded labels. Finally, label validation is performed by comparing ICD code labels with our own; a multi-rater clinical evaluation would improve label reliability, while the faithfulness metrics were estimated on a sample of 300 instances and should be computed on the full test set.

Future Work:

Five promising directions can be outlined for future work. First, the use of clinical negation detection techniques (NegEx, MedSpaCy, and clinical transformers) in the label engineering and feature extraction steps would allow improving labels and increasing specificity. Second, external validation on another cohort, e.g., the eICU Collaborative Research Database, UK Biobank, or non-US EHR extractions, would help determine the generalizability of the findings. Third, multimodal fusion of ECG waveforms, chest X-ray, and echocardiographic information would uncover hidden risk signals not visible in the tables and texts. Fourth, threshold optimisation with a clinically motivated cost ratio and probability calibration (Platt scaling and/or isotonic regression) would help decrease the false positive rate. Fifth, clinical validation studies with actual cardiologists evaluating the trustworthiness, comprehensibility, and actionability are needed for clinical application.

Sixth, federated and transfer learning from more than one hospital system would allow for multi-institutional generalizability without exchange of individual-level patient information, thereby circumventing the single-centre problem of MIMIC-IV, while protecting privacy an increasingly important trend in ICU-based clinical AI.

In summary, the trade-off between prediction accuracy and clinical interpretability in cardiac AI is a matter of design rather than an inherent limitation of the technology. Employing sound feature selection and engineering, thorough preprocessing, and clinical interpretability in modelling, a model that the doctors will understand, evaluate, and trust may equal black box alternatives in performance. This work demonstrates significant progress in responsible AI in cardiology.

Ethics and Data-Use Statement:

This study is a secondary analysis of the MIMIC-IV-Ext Cardiac Disease dataset, which is fully de-identified, open-access, and distributed through PhysioNet under credentialed access. The authors have completed the required CITI "Data or Specimens Only Research" course and signed the PhysioNet Credentialed Health Data Use Agreement before accessing the data. Being all de-identified according to HIPAA Safe Harbour provisions, the analysis is exempt from further Institutional Review Board review. Authors declare no competing interests. No specific financial support or sponsorship was received for the conduct of the research or the preparation of this manuscript.

Data and Code Availability:

MIMIC-IV-Ext Cardiac Disease dataset is available to credentialed users on PhysioNet; under the terms of the Data Use Agreement, the authors are not allowed to

redistribute the raw data. The preprocessing, feature engineering, model training, statistical testing, and model explainability code necessary to reproduce the reported results is available at https://github.com/SadiaShareef/Heart-Disease-Prediction, along with instructions for accessing the dataset on PhysioNet.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process:

In preparing this paper, the author utilised a large language model (Claude) for grammatical editing, text generation, and the formatting of figures and tables. However, after that, the author proofread, confirmed, and edited the work and is fully responsible for its content.

Conflict of Interest:

The authors declare that there is no conflict of interest in publishing this manuscript.

Project details:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors, and was not conducted as part of a funded project.

References:

[1] R. Kumar *et al.*, "A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions," *Front. Artif. Intell.*, vol. 8, p. 1583459, May 2025, doi: 10.3389/FRAI.2025.1583459/FULL.

[2] A. Bilal, A. Alzahrani, K. Almohammadi, M. Saleem, M. S. Farooq, and R. Sarwar, "Explainable AI-driven intelligent system for precision forecasting in cardiovascular disease," *Front. Med.*, vol. 12, p. 1596335, Jul. 2025, doi: 10.3389/FMED.2025.1596335/TEXT.

[3] F. Nasim, S. Masood, A. Jaffar, U. Ahmad, and M. Rashid, "Intelligent Sound-Based Early Fault Detection System for Vehicles," *Comput. Syst. Sci. Eng.*, vol. 46, no. 3, pp. 3175–3190, 2023, doi: 10.32604/CSSE.2023.034550.

[4] "AI-Driven Discovery of Novel Biomarkers for Early Cardiovascular Disease Detection," *AI-Driven Discov. Nov. Biomarkers Early Cardiovasc. Dis. Detect.*, 2026, doi: 10.66669/WT.V35IS2.287.

[5] A. Jamal, F. Tauseef, and Z. Akbar, "Predictive Analytics For AI-Assisted Patient No-Show Management And Clinic Revenue Optimization: A Simulation-Based Research," *Migr. Lett.*, vol. 21, no. S13, pp. 1901–1924, Aug. 2024, Accessed: Aug. 22, 2026. [Online]. Available: https://migrationletters.com/index.php/ml/article/view/12274

[6] A. Jamal, M. Mahmud, F. Tauseef, H. M. Sozib, S. Bin Shafi, and M. Tariquzzaman, "Explainable AI for Predicting Patient Readmission in Hospitals," *Proc. 9th Int. Conf. Inven. Comput. Technol. ICICT 2026*, pp. 1961–1967, 2026, doi: 10.1109/ICICT68280.2026.11510989.

[7] A. A. Hamad *et al.*, "Cognitive Inspired Sound-Based Automobile Problem Detection: A Step Toward Xai," 2024, doi: 10.2139/SSRN.4814232.

[8] B. A. Hemann, W. F. Bimson, and A. J. Taylor, "The Framingham Risk Score: an appraisal of its benefits and limitations," *Am. Heart Hosp. J.*, vol. 5, no. 2, pp. 91–96, 2007, doi: 10.1111/J.1541-9215.2007.06350.X.

[9] A. Jamal, Z. Akbar, S. Akbar, S. Niaz, and F. Tauseef, "Predicting Human-GenAI Collaboration Effectiveness: A Machine Learning Investigation Of Skill Configurations, Trust, And Work Design," *Migr. Lett.*, vol. 19, no. S8, pp. 2303–2324, Dec. 2022, Accessed: Aug. 21, 2026. [Online]. Available: https://migrationletters.com/index.php/ml/article/view/12298

[10] E. Hasan, M. Rahaman, D. Paul, S. R. U. I. Rahat, N. B. Asha, and M. Al Amin, "Performance Evaluation of Hybrid Machine Learning Models for Heart Disease Prediction in U.S. Clinical Decision Support Systems," *Proc. 5th Int. Conf. Sentim. Anal.*

Deep Learn. ICSADL 2026, pp. 34–39, 2026, doi: 10.1109/ICSADL67539.2026.11452047.

[11] F. Tauseef, A. Jamal, and F. Naseer, "Artificial Intelligence in Healthcare Systems Exploring the Transformational Role of AI Technologies in Healthcare Management and Clinical Decision Support," *Rev. J. Neurol. Med. Sci. Rev.*, vol. 1, no. 02, pp. 83–101, 2023, doi: 10.63075/26RFHC13.

[12] J. Hatherley, L. Aastrup Munch, and J. Christian Bjerring, "In Defense of Post Hoc Explanations in Medical AI," *Hastings Cent. Rep.*, vol. 56, no. 1, p. 40, Jan. 2026, doi: 10.1002/HAST.4971.

[13] M. Khajavian *et al.*, "Harnessing interpretable machine learning: SHapley additive exPlanations (SHAP)-driven insights, transformative impact, and controversies in adsorption-based environmental remediation," *Inorg. Chem. Commun.*, vol. 186, p. 116269, Apr. 2026, doi: 10.1016/J.INOCHE.2026.116269.

[14] İ. Özcan, G.-W. Weber, İ. Özcan, and G.-W. Weber, "Advances in cooperative game theory under uncertainty: A comprehensive survey," *Electron. Res. Arch. 2026 31342*, vol. 34, no. 3, pp. 1342–1362, Jan. 2026, doi: 10.3934/ERA.2026061.

[15] Q. Abbas, W. Jeong, and S. W. Lee, "Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges," *Healthc. 2025, Vol. 13, Page 2154*, vol. 13, no. 17, p. 2154, Aug. 2025, doi: 10.3390/HEALTHCARE13172154.

[16] X. Wang *et al.*, "Identifying drivers of community-scale urban renewal with an ante-hoc interpretable hybrid graph model," *J. Urban Manag.*, vol. 15, no. 3, pp. 1158–1177, Sep. 2026, doi: 10.1016/J.JUM.2025.12.010.

[17] R. J. Urbanowicz and J. H. Moore, "Learning Classifier Systems: A Complete Introduction, Review, and Roadmap," *J. Artif. Evol. Appl.*, vol. 2009, pp. 1–25, Sep. 2009, doi: 10.1155/2009/736398.

[18] R. J. Urbanowicz and J. H. Moore, "ExSTraCS 2.0: description and evaluation of a scalable learning classifier system," *Evol. Intell. 2015 82*, vol. 8, no. 2, pp. 89–116, Apr. 2015, doi: 10.1007/S12065-015-0128-8.

[19] A. Jamal, F. Tauseef, F. Naseer, and J. Ahmad, "Business Analytics for Healthcare Cost Optimization: A Data-Driven Study on Improving Financial Sustainability in Healthcare Systems," *Spectr. Eng. Sci.*, pp. 723–736, Dec. 2024, doi: 10.5281/zenodo.20488115.

[20] D. Rajhamundry and D. Rajhamundry, "Business Intelligence as a Strategic Planning Tool in Healthcare Organizations: A Systems Integration Approach," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 10, no. 6, pp. 1965–1972, Dec. 2024, doi: 10.32628/CSEIT241061234.

[21] A. Johnson *et al.*, "MIMIC-IV v3.1," Physionet. Accessed: Aug. 22, 2026. [Online]. Available: https://physionet.org/content/mimiciv/3.1/

[22] F. Tauseef *et al.*, "MACHINE LEARNING MODELS FOR PREDICTING PATIENT OUTCOMES IN INTENSIVE CARE UNITS: A CASE STUDY IN U.S. HOSPITALS," *Cuest. Fisioter.*, vol. 55, no. 1, pp. 88–105, Jun. 2026, doi: 10.48047/Y0T69V56.

[23] M. Faiyazuddin *et al.*, "The Impact of Artificial Intelligence on Healthcare: A Comprehensive Review of Advancements in Diagnostics, Treatment, and Operational Efficiency," *Heal. Sci. Reports*, vol. 8, no. 1, p. e70312, Jan. 2025, doi: 10.1002/HSR2.70312.

[24] A. S. Osei-Nkwantabisa and R. Ntumy, "Classification and Prediction of Heart Diseases using Machine Learning Algorithms," Sep. 2024, doi: 10.48550/arXiv.2409.03697.

[25] A. A. Stonier, R. K. Gorantla, and K. Manoj, "Cardiac disease risk prediction using machine learning algorithms," *Healthc. Technol. Lett.*, vol. 11, no. 4, pp. 213–217, Aug. 2023, doi: 10.1049/HTL2.12053.

[26] K. Kwakye and E. Dadzie, "Machine Learning-Based Classification Algorithms for the Prediction of Coronary Heart Diseases," Dec. 2021, Accessed: Aug. 22, 2026. [Online]. Available: https://arxiv.org/pdf/2112.01503

[27] Sorif Hossain, Mohammad Kamrul Hasan, "Machine learning approach for predicting cardiovascular disease in Bangladesh: evidence from a cross-sectional study in 2023," *BMC Cardiovasc. Disord.*, 2024, doi: 10.1186/s12872-024-03883-2.

[28] T. Vu *et al.*, "Machine Learning Model for Predicting Coronary Heart Disease Risk: Development and Validation Using Insights From a Japanese Population-Based Study," *JMIR cardio*, vol. 9, 2025, doi: 10.2196/68066.

[29] T. Guo, I. R. Bardhan, Y. Ding, and S. Zhang, "An Explainable Artificial Intelligence Approach Using Graph Learning to Predict Intensive Care Unit Length of Stay," *https://doi.org/10.1287/isre.2023.0029*, vol. 36, no. 3, pp. 1478–1501, Dec. 2024, doi: 10.1287/ISRE.2023.0029.

[30] A. Jamal, F. Tauseef, F. Naseer, A. Akbar, M. Jabeen, and F. Nasim, "Predictive AI Healthcare Analytics for Early Disease Detection Leveraging AI and Machine Learning Models to Identify High-Risk Patients and Improve Preventive Healthcare Strategies," *Annu. Methodol. Arch. Res. Rev.*, vol. 3, no. 11, pp. 1–28, Nov. 2025, doi: 10.5281/ZENODO.20686563.

[31] S. Ahmed, M. S. Kaiser, M. Shahadat Hossain, and K. Andersson, "A Comparative Analysis of LIME and SHAP Interpreters With Explainable ML-Based Diabetes Predictions," *IEEE Access*, vol. 13, pp. 37370–37388, 2025, doi: 10.1109/ACCESS.2024.3422319.

[32] M. Wang, "Explainable machine-learning-based cardiovascular disease prediction in patients with hypertension: Algorithm comparison and SHapley Additive exPlanations (SHAP) analysis," *Arch. Cardiovasc. Dis.*, vol. 119, no. 4, pp. 273–282, Apr. 2026, doi: 10.1016/J.ACVD.2025.09.005.

[33] M. A. B. Shiddik, "Explainable Artificial Intelligence in Healthcare: Current Landscape, Challenges, and Future Directions," *Heal. Sci. Reports*, vol. 9, no. 3, p. e72172, Mar. 2026, doi: 10.1002/HSR2.72172.

[34] S. Y. Kim, D. H. Kim, M. J. Kim, H. J. Ko, and O. R. Jeong, "XAI-Based Clinical Decision Support Systems: A Systematic Review," *Appl. Sci. 2024, Vol. 14, Page 6638*, vol. 14, no. 15, p. 6638, Jul. 2024, doi: 10.3390/APP14156638.

[35] "View of Detecting Governance Risk In Generative AI Adoption: A Predictive Analysis Of Organizational Misalignment And AI Failure Signals." Accessed: Aug. 21, 2026. [Online]. Available: https://www.metall-mater-eng.com/index.php/home/article/view/1965/1170

[36] S. Liu *et al.*, "Leveraging explainable artificial intelligence to optimize clinical decision support," *J. Am. Med. Inform. Assoc.*, vol. 31, no. 4, pp. 968–974, Apr. 2024, doi: 10.1093/JAMIA/OCAE019.

[37] N. Rane, S. Choudhary, and J. Rane, "Explainable Artificial Intelligence (XAI) in healthcare: Interpretable Models for Clinical Decision Support," *SSRN Electron. J.*, Nov. 2023, doi: 10.2139/SSRN.4637897.

[38] R. Ganesan, S. C. Habraken, F. N. van de Vosse, and W. Huberts, "Explainable Machine Learning Based Prediction of Severity of Heart Failure Using Primary Electronic Health Records," *Stud. Health Technol. Inform.*, vol. 316, pp. 542–546, Aug. 2024, doi: 10.3233/SHTI240471.

[39] K. Patel, M. Shah, K. M. Qureshi, and M. R. N. Qureshi, "A systematic review of

generative AI: importance of industry and startup-centered perspectives, agentic AI, ethical considerations & challenges, and future directions," *Artif. Intell. Rev. 2025 591*, vol. 59, no. 1, pp. 7-, Nov. 2025, doi: 10.1007/S10462-025-11435-Z.

[40] A. Jamal, F. Tauseef, F. Naseer, A. Akbar, M. Jabeen, and F. Nasim, "A Convolutional Neural Network Approach for Diabetic Retinopathy Detection from Retinal Images: A Critical Review of Deep Learning Techniques in Ophthalmic Diagnosis: https://doi.org/10.5281/zenodo.20587510," *Multidiscip. Surg. Res. Ann.*, vol. 3, no. 4, pp. 3100–3128, Dec. 2025, doi: 10.5281/ZENODO.20587510.

[41] D. Chen, G. Li, J. Huang, and Y. Li., "Superior machine learning model for post-cardiac arrest mortality prediction: a MIMIC-IV cohort study," *Resusc. Plus*, vol. 28, 2026.

[42] Hu Q, Xu T, Gao X, Lei X and Hu L., "Machine learning-driven sedation-analgesia optimization in mechanically ventilated sepsis patients: a retrospective MIMIC-IV analysis," *Front. Pharmacol*, vol. 17, 2026, doi: https://doi.org/10.3389/fphar.2026.1673704.

[43] Nguyen V, Mittal R., "Machine Learning Prediction of ICU Mortality and Length of Stay in Atrial Fibrillation: A MIMIC-IV/MIMIC-III Study," *Healthc.*, vol. 14, no. 3, 2026, doi: 10.3390/healthcare14030356.

[44] Yin J, Pan X, Chen D, Zhang J, Jin G., "Machine-learning model for 30-day mortality in sepsis-associated delirium patients: A retrospective MIMIC-IV cohort study," *Med.*, vol. 105, no. 1, 2026, doi: 10.1097/MD.0000000000045440.