Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Auditing Machine Learning for Acute Myocardial Infarction: Leakage, Calibration, Label Circularity, and Explainable ECG Evidence

Agughasi Victor Ikechukwu
Dayananda Sagar University
Kirankumar R. Dyawarkonda
Dayananda Sagar University
Chiemeka C. Chinaka
The University of Queensland
Bikila Abebe Ganati
bikila.abebe@ju.edu.et

Jimma University
Diriba Gichile Rundasa
Adama Science and Technology University
Tariku Birhanu Yadesa
Adama Science and Technology University
Yogesh Yogesh
Dayananda Sagar University
Rajesh T.M
Dayananda Sagar University

Article

Keywords: Acute myocardial infarction, Cardiac troponin, Clinical decision support, Dataset auditing, Decision-curve analysis, Explainable AI

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10743818/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Auditing Machine Learning for Acute Myocardial Infarction: Leakage, Calibration, Label Circularity, and Explainable ECG Evidence

Agughasi Victor Ikechukwu[1*], Kirankumar R. Dyawarkonda[2], Chiemeka C. Chinaka[3], Bikila Abebe Ganati[4*], Diriba Gichile Rundasa[5], Tariku Birhanu Yadesa[5], Yogesh[6], Rajesh T.M.[7]

[1.]Department of Computer Science & Medical Engineering, School of Engineering, Dayananda Sagar University, Bengaluru, 562112, Karnataka, India
[2.]Department of Cardiology, Dr. Chandramma Dayananda Sagar Institute of Medical Education and Research (CDSIMER), Dayananda Sagar University, Bengaluru, Karnataka, India
[3.]Queensland Brain Institute, The University of Queensland, St. Lucia, Brisbane, 4072, QLD, Australia
[4.]Department of Computer Science, Faculty of Computing and Informatics, Jimma Institute of Technology, Jimma University, P.O. Box 378, Jimma, Ethiopia
[5.]Department of Computer Science and Engineering, School of Electrical Engineering and Computing, Adama Science and Technology University, Adama, Ethiopia
[6.]Department of Computer Science and Engineering (Artificial Intelligence & Machine Learning), School of Engineering, Dayananda Sagar University, Bengaluru, 562112, Karnataka, India
[7.]Department of Computer Science and Medical Engineering, School of Engineering, Dayananda Sagar University, Bengaluru, 562112, Karnataka, India

Corresponding authors email(s): bikila.abebe@ju.edu.et; victor-csme@dsu.edu.in, https://orcid.org/0000-0002-1175-3089

ABSTRACT

Machine-learning risk scores for cardiovascular disease are proliferating, yet independent appraisals repeatedly find the same three defects: silent data leakage, an absent calibration analysis, and no clinician involvement in defining the outcome being predicted. A model can therefore report a near-perfect area under the curve and still be unusable at the bedside. This paper presents TRUST-AMI, a leakage-audited, calibration-aware, and explainable approach for one narrowly scoped and time-critical decision: the rapid rule-in and rule-out of acute myocardial infarction (AMI) from routinely collected cardiac biomarkers, supported by a separately audited electrocardiogram (ECG) imaging branch. A cohort of 1,319 emergency presentations (positive = 810, negative = 509) described by eight bedside variables was analysed under a protocol in which imputation and scaling are fitted inside cross-validation folds, model selection uses 5-fold stratified cross-validation on the training partition alone, and the held-out test set (n = 264) is scored exactly once. Four classifiers were compared across calibration, clinical operating characteristics, 1,000-sample bootstrap intervals, ablation, decision-curve analysis, and sex and age subgroup analysis. The selected random forest attained a test ROC-AUC of 0.996 (95% CI 0.988–1.000), sensitivity 1.000, specificity 0.961, and agreement with the cardiologist reference of $\kappa = 0.968$. The ECG branch, trained on PTB-XL under patient-disjoint folds, reached a test ROC-AUC of 0.908 (n = 1,517) with Grad-CAM supplying waveform-level evidence. A second, clinically validated four-class ECG image corpus was audited before use, and the audit is itself a result: its 928 released files contain 491 distinct images and 311 distinct patients, and 87.4% of the acute-MI class consists of byte-identical copies. Holding estimator, features, and corpus fixed and varying only what the partition may share, macro-F1 falls from 0.969 to 0.656 and acute-MI F1 from 0.998 to 0.121. The resulting tool is an AMI triage aid within coronary heart disease and is not presented as a general cardiovascular-disease diagnoser. All code, seeds, and results are released publicly, and the study is offered as a reproducible template for reporting clinical prediction models transparently.

Keywords: Acute myocardial infarction, Cardiac troponin, Clinical decision support, Dataset auditing, Decision-curve analysis, Explainable AI

1 INTRODUCTION

Cardiovascular disease accounts for roughly one in three deaths worldwide, and acute myocardial infarction is among its most time-critical presentations. In the emergency department, the decision to admit, discharge, or escalate a patient presenting with chest pain must frequently be reached within a single clinical hour. That decision rests principally on serial cardiac biomarkers: high-sensitivity cardiac troponin and creatine-kinase MB (CK–MB), read alongside vital signs, demographics, and the electrocardiogram [1]–[4]. The cost asymmetry is severe. A missed infarction may be fatal, whereas an unnecessary admission consumes a bed and exposes the patient to iatrogenic risk. Any decision aid entering this setting is therefore judged less by its average accuracy than by the reliability of the probability it emits at the operating threshold actually used.

Machine learning is an attractive instrument for this task because it can combine several weakly informative signals into a single probability that supports, without replacing, clinician judgement [5]–[7]. The published record, however, is uneven in a way that has become well documented. Systematic appraisals of clinical prediction studies report that a substantial fraction are compromised by data leakage[8]: information from the evaluation partition contaminating the fitting procedure through premature scaling, target-aware feature selection, or hyper-parameter search conducted on the test split [9]–[12]. Reporting standards have responded: TRIPOD+AI [13] and the risk-of-bias instrument PROBAST+AI [14] both now require a discrimination metric and a calibration metric, each with confidence intervals. Calibration is nonetheless omitted from a large share of published cardiovascular models, and interpretability and subgroup fairness are addressed inconsistently [15].

The consequence is a literature in which headline discrimination has drifted upward while the evidence that a model would behave safely at the bedside has not kept pace [16]. Three defects recur, and they are distinguishable. Leakage inflates the estimate of a model's performance. Poor calibration leaves discrimination intact but corrupts the probability that a threshold is applied to, so that a model may rank patients correctly and still assign numerically wrong risks [17]. Circularity in the outcome definition inflates the apparent difficulty of the task itself, because the model is partly rediscovering the rule that produced its labels [18]. A study that audits only the first of the three can still mislead.

Biomarker-only reasoning has a further limitation that is clinical rather than statistical [19]. Troponin reports *whether* myocardial injury has occurred, but it is silent on *where* and *how*. The morphological evidence that cardiologists read directly: *ST*-segment deviation, pathological *Q*-waves, *T*-wave inversion resides in the electrocardiogram and is discarded by a tabular model. A system that also reasons over the ECG image, and that can indicate which leads and time-segments carried its evidence, is both more clinically faithful and more amenable to expert audit.

1.1 Scope and clinical positioning

Cardiovascular disease is not a single entity but a family of conditions: coronary heart disease (CHD), cerebrovascular disease, peripheral arterial disease, aortic disease, structural, congenital and valvular disease, functional and conduction disorders such as heart failure and the arrhythmias, and venous thromboembolism, each diagnosed through distinct data modalities including echocardiography, cross-sectional imaging, D-dimer, and the ankle–brachial index [20],[21]. The framework presented here is scoped to a single target within CHD: *the rapid rule-in and rule-out of acute myocardial infarction from routine biomarkers and the ECG*. That boundary is stated explicitly throughout, because a claim of broader coverage would overstate what the available features can support. The scoping is a design decision rather than an omission, and it is defended in **Section 7** on the evidence assembled here.

1.2 Motivation for the present work

This work originates in an interdisciplinary medical-elective collaboration at *Dr. Chandramma Dayananda Sagar Institute of Medical Education and Research* (CDSIMER), Dayananda Sagar University, Bengaluru, Karnataka-India in which modelling choices and the interpretation of biomarker values were reviewed by practising cardiologists and the reference labels were adjudicated under a written protocol. Earlier work by the authors on explainable diagnosis from chest radiographs motivates the interpretability-first stance adopted throughout [22]-[26]. The guiding question is not whether a high area under the curve can be obtained on this cohort: it can, and **Section 6** reports 0.996, but what remains of that number once leakage is excluded by construction, calibration is measured rather than assumed,

the incremental value over existing clinical practice is quantified, and the circularity of the outcome definition is confronted.

1.3 Contributions

The contributions of this study are the following:

i **A leakage-audited framework stated as an explicit protocol.** Algorithm 1 (**Section 4.2**) confines every preprocessing step to training folds, restricts model selection to cross-validation on the training partition, and scores the test set exactly once.

ii **A calibration audit reported alongside discrimination.** Brier score, expected calibration error, and calibration slope are reported with bootstrap intervals in accordance with TRIPOD+AI, and are shown to dissociate sharply from discrimination on this cohort.

iii **An incremental-value analysis against clinical practice.** The model is benchmarked not only against competing learners but against the single-threshold troponin rule it would have to displace, yielding a $+0.071$ AUC gain that is statistically significant and clinically modest.

iv **A cardiologist-authored adjudication protocol with a reliability study.** Reference labels follow the Fourth Universal Definition of Myocardial Infarction, and model–reference agreement is quantified with Cohen's κ.

v **An explicit treatment of biomarker–label circularity.** A construct-validity threat specific to biomarker-defined outcomes is separated from data leakage, quantified through ablation, and carried into the limitations rather than left unstated.

vi **An independently audited ECG imaging branch with Grad-CAM evidence,** trained under patient-disjoint folds, together with a reasoned refusal to report a fused metric that the available unpaired cohorts cannot support.

vii **A pre-fit audit of a clinically validated ECG image corpus, and a measurement of what a leaked partition manufactures.** The four-class corpus of Section 3.5 is shown to contain 491 distinct images in 928 released files and 311 distinct patients behind those images. Holding estimator, features, and corpus fixed and varying only the partition, macro-F_1 falls from 0.969 to 0.656, and F_1 on the acute-MI class from 0.998 to 0.121. A separate ablation shows that the heart rate printed on the report page is on its own worth macro-F_1 0.461, so a model trained on the uncropped page has a non-waveform channel available to it.

viii **A full public release** of code, configuration, fixed seeds, and results, so that every table and figure in this article can be regenerated end to end.

1.4 Organisation

The remainder of the article is organised as follows. **Section 2** reviews related work across five themes and summarises it in Table 1. **Section 3** describes the cohort, the clinical adjudication protocol, and the two ECG corpora, one of which is audited before use. **Section 4** presents the proposed TRUST-AMI framework, its components, and its formal statement. **Section 5** details the training and evaluation methodology. **Section 6** reports five named experiments. **Section 7** provides interpretability and model analysis, and **Section 8** concludes, while stating the future directions.

2 RELATED WORK

The literature bearing on this study divides into **five themes**: *classical machine learning* for tabular cardiovascular cohorts; *biomarker-based rule-out* pathways as the clinical reference; *the methodological literature* on leakage, calibration, and reporting standards; *deep learning* on the electrocardiogram; and *explainability as a precondition* for clinical adoption. Each is reviewed in turn, and **Table 1** consolidates the studies referenced.

2.1 Classical machine learning for tabular cardiovascular cohorts

Gradient-boosted trees dominate recent tabular cardiovascular studies. Dhanka and Maini tune XGBoost with Optuna [27], Cao et al. combine multiple feature-selection stages with particle-swarm optimisation [28], and Naik et al. apply social-group optimisation to random forests and extreme gradient boosting [29]; hybrid oversampling-and-boosting schemes address class imbalance [30].

Reported accuracies cluster at 84–88% with AUCs of 0.92–0.98 on the Cleveland, Statlog, and Cardio collections [5], [31], [32]. Two features of this body of work are relevant here. First, these cohorts encode chronic cardiovascular risk rather than acute infarction, so their operating characteristics do not transfer to emergency triage. Second, the reported gains are typically measured against other learners rather than against the clinical rule in current use, which leaves the incremental value of the modelling unquantified. Tabular foundation models such as TabPFN [33] are beginning to compete on small clinical cohorts, although established methods remain strong at this scale.

2.2 Biomarker pathways as the clinical reference

Guideline-endorsed high-sensitivity troponin pathways achieve admission AUCs of 0.87–0.91 for AMI. Völschow et al. validate a novel high-sensitivity cardiac troponin T assay [1], Koechlin et al. characterise point-of-care performance [2], Giannitsis et al. appraise the case for further accelerating rule-out protocols [3], and Sataranatarajan et al. revisit the baseline cut-offs on which such protocols depend [4]. This literature establishes the benchmark that any biomarker-based model must be measured against, and it is the benchmark adopted in Experiment II (Section 6.2). It also establishes the sensitivity ceiling that motivates the present work: accelerated pathways trade sensitivity for speed, and it is precisely that trade that a learned model might improve.

2.3 Leakage, calibration, and reporting standards

The methodological literature has converged on data leakage as a principal source of optimistic bias. Rosenblatt et al. document its effect on connectome models [8], van de Mortel and van Wingen trace how leaked estimates propagate into meta-analytic summaries [9], Bernett et al. provide guiding questions for biological applications [10], Apicella et al. examine leakage risk in transfer settings [11], and Starcke et al. isolate the interaction between leakage and feature selection in a Parkinson's-disease detection task [12]. In parallel, TRIPOD+AI [13] and PROBAST+AI [14] have made calibration reporting and risk-of-bias appraisal explicit requirements. Class-imbalance handling [34], sociodemographic bias [15], uncertainty quantification through conformal prediction [35], and safety frameworks for language models in medicine [36] complete the methodological bar this work targets. Cohen's κ remains the standard instrument for inter-rater agreement in cardiology reader studies [37]. Comprehensive appraisals of machine learning for cardiovascular disease [38] and the adjacent electronic-health-record literature on heart failure [39]–[43], along with validated deep-learning models in other clinical imaging domains [44], frame the reporting expectations adopted here.

2.4 Deep learning on the electrocardiogram

Convolutional and recurrent networks now approach cardiologist-level performance on many ECG tasks. On PTB-XL [45], the standardised benchmarks of Strodthoff et al. report MI-versus-normal AUROCs of approximately 0.92–0.94 [46], and Hilgendorf et al. demonstrate fully automated AMI diagnosis from the ECG [7], with related work extending risk stratification to therapeutics-related cardiac dysfunction [47]. These results supply the external envelope against which the imaging branch reported in Section 6.4 is calibrated. Their relevance to the present argument is specific: the ECG carries evidence that is not, by construction, part of a troponin-defined label, and therefore offers a route out of the circularity problem identified in Section 7.5.

2.5 Explainability and clinical trust

Post-hoc attribution has become the default route to clinician trust in medical deep learning. Grad-CAM [18] is particularly suited to ECG images because its class-discriminative heat-maps localise to interpretable waveform regions and leads. Recent syntheses catalogue both the rapid adoption of these methods and their failure modes [16], [17], and dedicated reviews by Manimaran et al. address explainable ECG classification [20] and multi-level explanation for arrhythmia detection [21]. Interpretable models have been applied to emergency-department triage [48], gastrointestinal prognostication [49], public-health analytics [50], cognitive ageing [51], and mobile diagnosis [52], and multimodal artificial intelligence in diagnostics is reviewed by Jandoubi and Akhloufi [19]. The position taken in this article is that attribution is necessary but not sufficient: a heat-map over a

miscalibrated probability explains a number that should not have been trusted in the first place, which is why explainability is reported in Section 7 jointly with calibration rather than in place of it.

2.6 Positioning of the present study

Across these five themes, three gaps persist. Tabular cardiovascular models are rarely benchmarked against the clinical rule they would replace; calibration is rarely reported even where discrimination is strong; and the circularity introduced when a biomarker both defines the label and serves as a predictor is rarely acknowledged. TRUST-AMI is positioned against exactly **these three gaps**, and the experiments in **Section 6** are organised to address them in turn.

Table 1: Summary of referenced studies. The final column records the limitation most relevant to the present work rather than an exhaustive critique.

First author	Year	Focus area	Methodology	Key performance	Insights/Limitations
Kumar et al. [38]	2025	ML for heart-disease prediction, review	Systematic appraisal of models, ethics and reporting	Maps the field; sets reporting expectations	Confirms calibration and leakage are under-reported
Naik et al. [29]	2025	Heart-disease prediction	Social-group optimisation with RF/XGB	AUC 0.920 (Cleveland), 0.975 (Statlog)	Chronic-risk cohorts; no clinical-rule baseline
Cao et al. [28]	2025	CVD prediction	Feature selection with PSO-tuned XGBoost	AUC ≈ 0.95 (Cardio)	No calibration reported
Dhanka & Maini [27]	2025	CVD prediction	XGBoost with Optuna tuning	Accuracy 84–88%	Tuning protocol vulnerable to leakage
Zhang et al. [31]	2026	Integrated CVD prediction	Artificial neural network	Accuracy 0.834	Discrimination reported alone
Völschow et al. [1]	2026	hs-cTnT assay	Diagnostic validation	Admission AUC 0.87–0.91	Assay-specific cut-offs limit transfer
Giannitsis et al. [3]	2024	Accelerated rule-out	Editorial appraisal of hs-troponin protocols	Rapid triage at defined sensitivity	Editorial, not a primary study
Sataranatarajan et al. [4]	2025	Baseline cut-offs	Re-analysis of abnormal thresholds	Cut-off choice drives rule-out safety	Underlines threshold sensitivity
Hilgendorf et al. [7]	2025	AMI from ECG	Fully automated deep learning	AUROC 0.82–0.90	Requires waveform infrastructure
Strodthoff et al. [46]	2021	PTB-XL benchmarks	Deep learning, standardised folds	MI vs. NORM AUROC 0.92–0.94	External envelope, different cohort
Rosenblatt et al. [8]	2024	Data leakage	Controlled leakage experiments	Quantifies inflation of reported accuracy	Motivates fold-internal preprocessing
van de Mortel & van Wingen [9]	2025	Leakage in meta-analysis	Meta-analytic re-estimation	Leaked estimates bias pooled effects	Field-level optimism
Bernett et al. [10]	2024	Leakage avoidance	Guiding questions for practitioners	Checklist for biological ML	Guidance not routinely applied
Collins et al. [13]	2024	Reporting standard	TRIPOD+AI statement	Requires discrimination and calibration	Compliance remains partial

Moons et al. [14]	2025	Risk-of-bias tool	PROBAST+AI	Structured appraisal of AI models	Applied post hoc, not at design
Colacci et al. [15]	2025	Sociodemographic bias	Scoping review	Documents subgroup performance gaps	Small strata limit detection
Hollmann et al. [33]	2025	Tabular foundation model	TabPFN	Strong on small tabular data	Not evaluated for acute triage
Selvaraju et al. [18]	2020	Attribution	Grad-CAM	Class-discriminative localisation	Explains, does not validate, probability
Manimaran et al. [20]	2025	Explainable ECG	Review of XAI for ECG classification	Catalogues methods and pitfalls	Attribution rarely paired with calibration
Ikechukwu & Murali [24]	2023	Explainable CXR diagnosis	xAI model for COPD	Interpretability-first design	Imaging domain, different modality

3 MATERIALS

3.1 The tabular cohort

The tabular analysis uses the publicly available Zheen Heart Attack Dataset, collected at Zheen Hospital in Erbil, Iraq, between January and May 2019 (Mendeley Data, Version 1, DOI: 10.17632/wmhctcrt5v.1). After the cleaning and adjudication steps described in Section 3.3 had been applied, the dataset comprised 1,319 de-identified emergency presentations labelled positive (heart attack) or negative by the attending clinical team. Eight predictors are available: age, sex ($1 =$ male, $0 =$ female), heart rate (impulse), systolic and diastolic blood pressure, random blood glucose, CK–MB, and cardiac troponin. Outcome prevalence is 61.4% positive.

The choice of cohort is deliberate and constrains the claims that follow. These eight variables are precisely those available at the bedside within minutes of presentation, without specialist imaging or laboratory turnaround beyond the biomarker panel. A model restricted to them is therefore deployable in exactly the setting where the decision is made, but it is also blind to everything a cardiologist would additionally consider. Both properties are reported rather than elided.

3.2 The electrocardiogram corpus

The imaging branch uses PTB-XL[45], comprising 21,799 clinical 12-lead electrocardiograms annotated with SCP diagnostic statements. Records were filtered to the binary task of myocardial infarction (MI) versus normal (NORM). PTB-XL supplies an official *strat_fold* partition that is patient-disjoint, which permits the same leakage discipline applied to the tabular branch to be enforced on the imaging branch without additional assumptions.

The Zheen cohort and PTB-XL share no patients data. This fact is load-bearing for the present study and is treated as a constraint on what may be reported rather than as an inconvenience to be worked around; the consequence is developed in Section 4.9 and enforced in Section 6.4.

3.3 Clinical adjudication protocol

Reference labels were defined and quality-assured under a protocol specified by the CDSIMER cardiology validation panel. When the adjudicating cardiologist labelled the de-identified dataset, the original labels assigned by the attending clinical team were retained as a reference point in order to avoid systematic drift between the two label sources.

i Each presentation was classified using the Fourth Universal Definition of Myocardial Infarction, integrating symptoms, the electrocardiogram, and serial high-sensitivity troponin relative to the 99th-percentile upper reference limit of the assay.

ii Physiologically impossible values (heart rate > 300 or < 20 bpm; systolic blood pressure > 300 mmHg; diastolic blood pressure > 200 mmHg) were flagged and treated as missing, independently of the outcome. Screening on a value rather than on the outcome is what keeps this step outside the leakage surface.

iii A random 15% subset was independently re-adjudicated by a second, blinded cardiologist, establishing inter-observer reliability.
iv The same subset was re-read by the primary cardiologist after a washout of at least 14 days, establishing intra-observer reliability.
v Agreement was quantified with Cohen's κ, and disagreements were resolved by consensus before final labelling.

3.4 Cohort characterisation

Figure 1 shows the class balance and the distribution of cardiac troponin by class, and **Table 2** summarises baseline characteristics. Age, sex, CK–MB, and troponin differ significantly between classes (all $p < 0.001$; Mann–Whitney U or χ^2), whereas heart rate, blood pressure, and glucose do not. This separation anticipates the ablation results of **Section 6.3**: the variables that distinguish the classes at baseline are the same variables the model cannot function without, and the vital signs contribute little in either analysis.

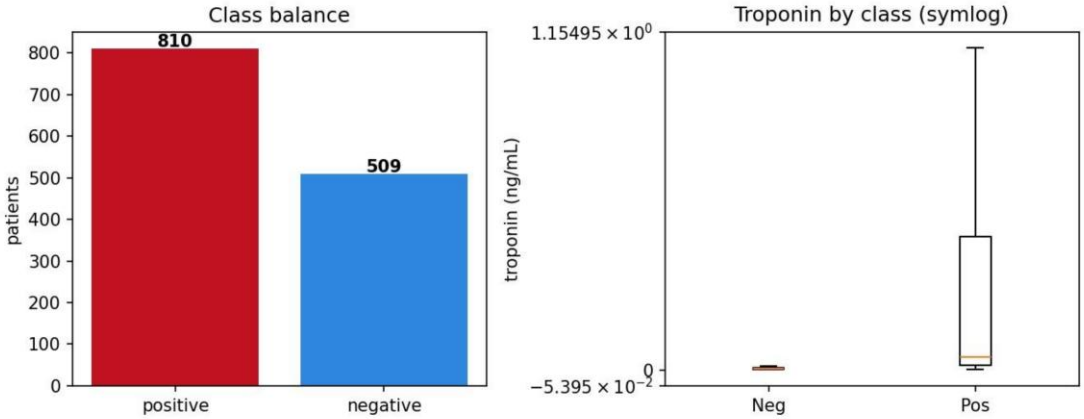


Figure 1: *Class balance (left) and cardiac-troponin distribution by class on a symmetric-log scale (right). The near-separation visible on the right is the empirical basis of the circularity argument developed in* ***Section 7.5****.*

Table 2: Baseline characteristics by class; median (interquartile range).

Variable	Positive ($n = 810$)	Negative ($n = 509$)	p
Age, years	60 (50–68)	52 (42–63)	< 0.001
Male, %	69.5	60.3	< 0.001
Heart rate, bpm	74 (64–85)	75 (64–84)	0.978
Systolic BP, mmHg	122 (110–141)	125 (110–147)	0.185
Diastolic BP, mmHg	71 (62–80)	72 (61–82)	0.775
Glucose, mg/dL	116 (98–166)	117 (98–184)	0.515
CK–MB, ng/mL	3.77 (1.87–12.25)	2.31 (1.50–3.35)	< 0.001
Troponin, ng/mL	0.04 (0.02–0.46)	0.01 (0.00–0.01)	< 0.001

i. Exploratory structure of the tabular cohort

Three further views of the cohort are reported because each anticipates a later result. **Figure 2** gives the marginal distribution of every predictor by class; the two cardiac enzymes are the only variables whose class-conditional densities separate, and they separate on a scale that ordinary axes cannot display. **Figure 3** gives the same comparison as boxplots on a common footing, which makes the outlier structure of CK–MB and troponin visible. **Figure 4** gives the Pearson correlation matrix, and three features of it are worth stating because each recurs later. Systolic and diastolic pressure are the only appreciably collinear pair ($r = 0.59$). The two cardiac enzymes are essentially uncorrelated with one another ($r = -0.02$), so they are not redundant measurements and the ablation of **Section 6.3** is right to remove them separately. And every linear correlation with the outcome is weak, troponin included ($r = 0.23$), which is not a contradiction of Table 4 but a consequence of it: the troponin–outcome relationship is a threshold effect on a heavily skewed variable, and a linear coefficient is the wrong instrument for seeing it. That is precisely why the ablation reports AUC rather than correlation.

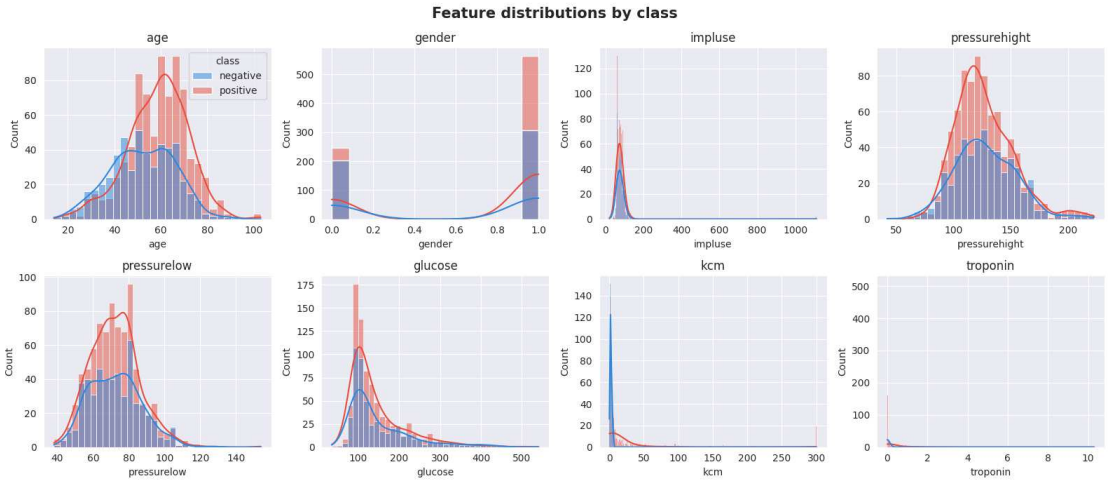


Figure 2: *Class-conditional distributions of the eight bedside predictors. Only CK–MB and cardiac troponin separate the classes.*

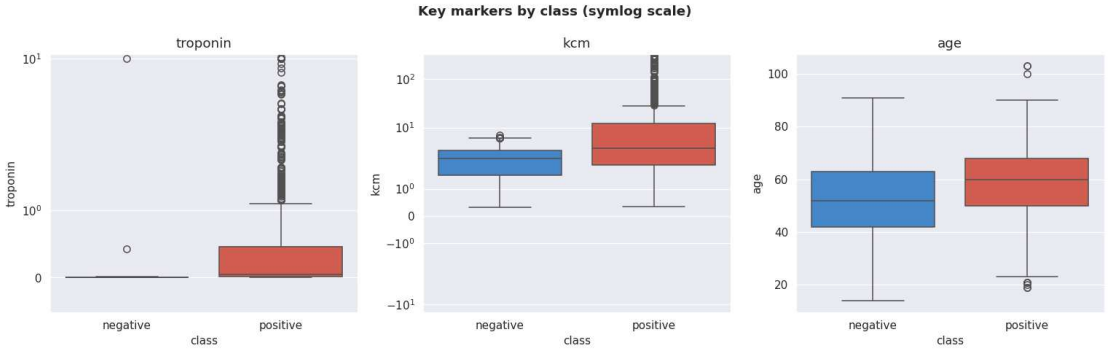


Figure 3: *Predictor distributions by class as boxplots. The heavy right tails on the two cardiac enzymes are the reason a rank test rather than a t-test is used throughout Table 4.*

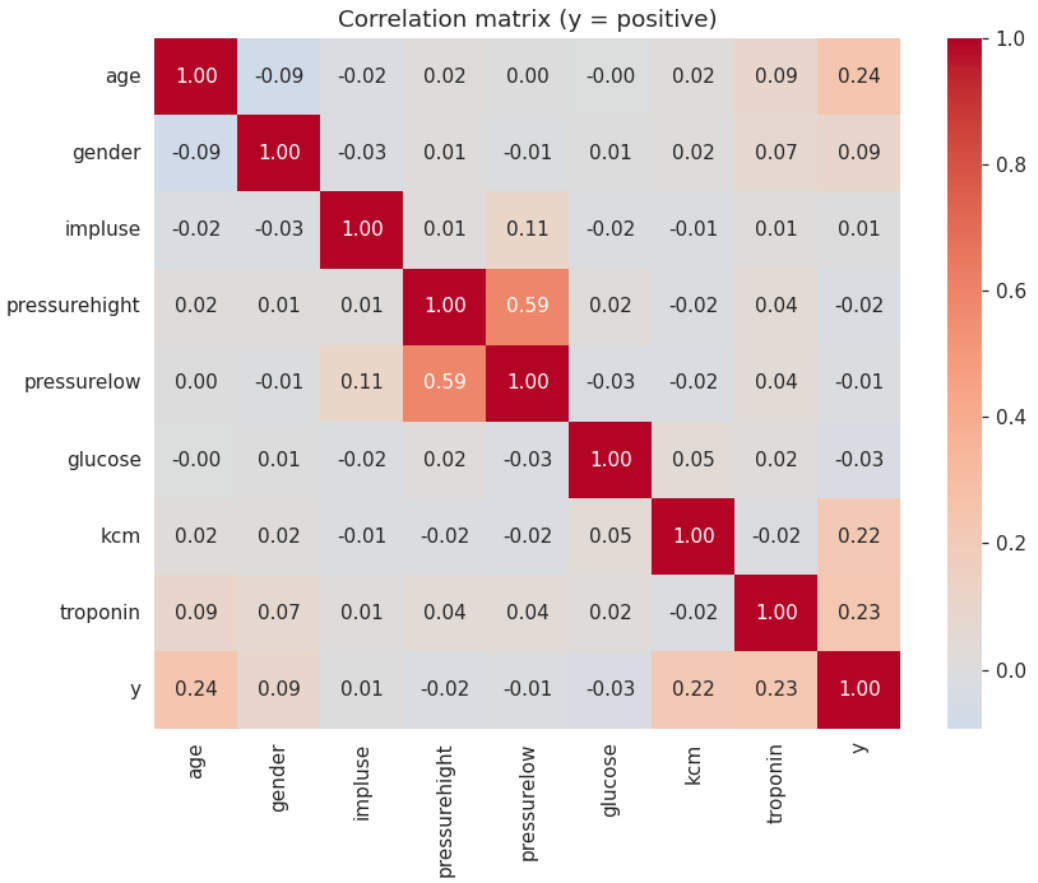


Figure 4: *Pearson correlation matrix over the eight predictors and the outcome. Systolic and diastolic pressure are the only appreciably collinear pair; the two cardiac enzymes are uncorrelated with one another; and no linear coefficient—troponin's included—conveys the threshold structure that Section 6.3 recovers by ablation.*

3.5 The clinically validated four-class ECG image corpus

A second, independently sourced imaging corpus is introduced in this study: the *ECG Images dataset of Cardiac Patients* [53, 54], assembled at the Ch. Pervaiz Elahi Institute of Cardiology, Multan, Pakistan, and released through Mendeley Data. Its records were reviewed by medical professionals using a telehealth ECG diagnostic system, which is why we refer to it as clinically validated: the class assignment is a reading, not a by-product of an automated rule applied to a laboratory value. That property makes it complementary to both other sources used here. The Zheen cohort labels an outcome partly *defined* by troponin (**Section 7.5**); PTB-XL labels SCP statements produced by an analysis program and confirmed by a cardiologist; this corpus labels what a reviewing clinician saw on a printed 12-lead trace.

The corpus is distributed as 928 JPEG page images at $2{,}213 \times 1{,}572$ pixels, each a scanned ECG report in the conventional 3×4 lead layout with a lead II rhythm strip, a header band carrying study metadata, and a footer band carrying the recorder settings, device-computed heart rate, and acquisition timestamp. Four classes are provided: normal (NORM), abnormal heartbeat (ABN), acute myocardial infarction (MI), and previous history of myocardial infarction (PMI). **Figure 5** shows representative records from each class with the header and footer bands cropped away.

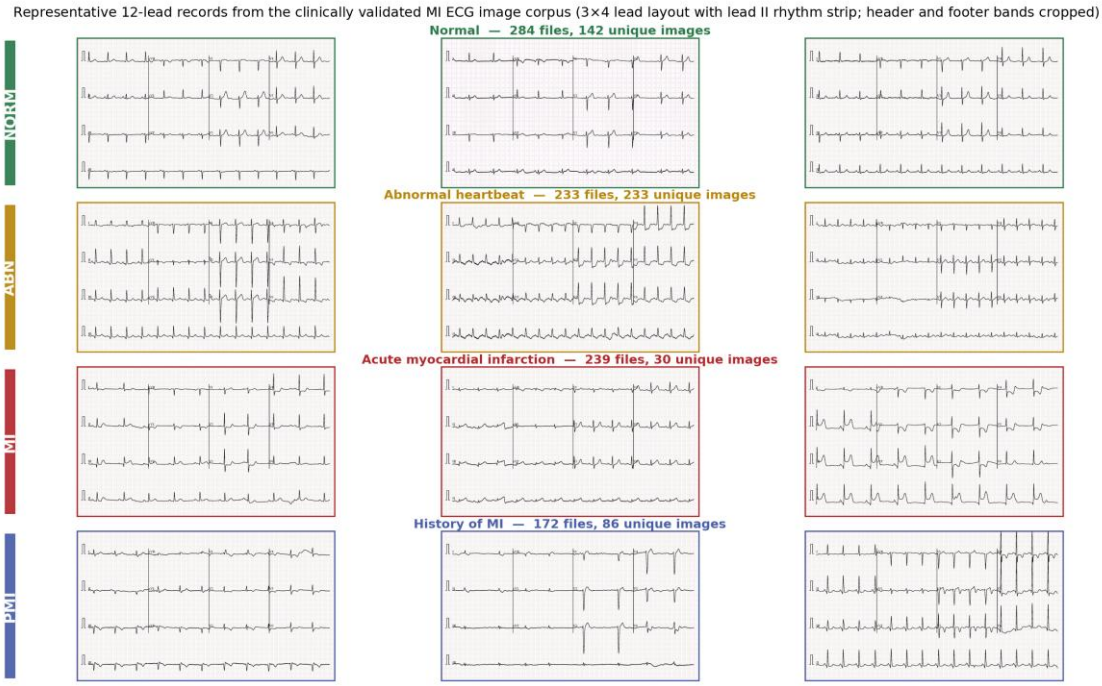


Figure 5: *Representative 12-lead records from the clinically validated four-class corpus, three per class. Each panel is one scanned report in the* 3 × 4 *lead layout with a lead II rhythm strip; the header and footer bands have been cropped, for the reason given in Section 6.5. The counts in each strip title state the number of distributed files and the number of distinct images those files contain.*

i. Audit of the corpus as distributed

The leakage discipline presented in this paper applies to datasets before it applies to models, so the corpus was audited before any classifier was fitted. **First,** every file was hashed (MD5). The 928 distributed files contain only 491 distinct images; 437 files, or 47.1%, are byte-identical copies of another file. The redundancy is not spread evenly. It is absent from ABN, affects half of NORM and PMI, and reaches 87.4% of MI, where 239 files resolve to just 30 distinct images (29 of them present) in eight identical copies. Panels (a)–(c) of Figure 6 summarise the result.

Second, the header band of every distinct image was read by optical character recognition. Each report prints a numeric study identifier, which recovers a patient key that the folder structure does not expose. The 491 distinct images carry 311 distinct identifiers, so image-level independence does not hold either; 82 identifiers appear on more than one image, one of them on 13. Sixty-two patients contribute images to more than one class (most often a normal and an abnormal-beat) tracing recorded minutes apart in the same session, which the printed timestamps confirm. Table 3 records the composition at all three levels of granularity.

Two secondary observations follow from the same audit and are recorded because they bear on how results should be read. The printed header is legible, so the images are pseudonymised rather than de-identified, and any redistribution should crop the header band. And the cohort is 91.0% male overall and 28 of 30 distinct MI images are male, so no sex-stratified claim can be supported on this corpus.

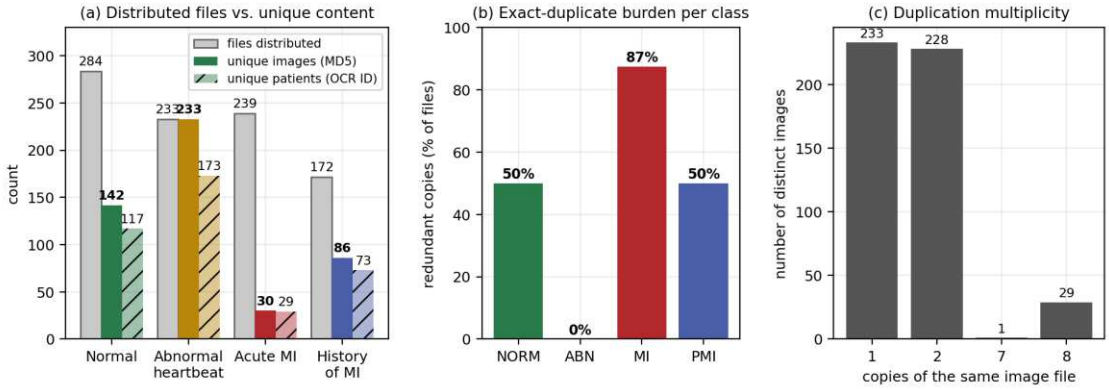


Figure 6: *Audit of the four-class corpus as distributed. (a) Files released, distinct images after MD5 de-duplication, and distinct patients recovered from the printed study identifier. (b) Share of each class that is a redundant copy. (c) How many files carry the same image: 29 images appear eight times each, all of them in the MI class.*

Table 3: *Composition of the clinically validated four-class corpus at three levels of granularity. Percentages in the final column are redundant copies as a share of files released.*

Class	Files	Distinct images	Distinct patients	Redundant
Normal (NORM)	284	142	117	50.0%
Abnormal heartbeat (ABN)	233	233	173	0.0%
Acute myocardial infarction (MI)	239	30	29	87.4%
History of MI (PMI)	172	86	73	50.0%
Total	928	491	311	47.1%

Note: Distinct patient identifiers do not sum across classes because 62 patients contribute images to more than one class.

4 PROPOSED FRAMEWORK: TRUST-AMI

4.1 Design motivation and governing principles

TRUST-AMI is a new paradigm and evaluation architecture built around one claim: for a high-stakes triage decision, the protocol that produces a performance number is a more consequential design object than the estimator that number describes. The framework is therefore specified as a set of constraints on how any candidate model may be fitted, selected, and reported, and the choice of estimator is subordinated to those constraints.

Four principles govern the design:

- **Leakage is excluded by construction, not by inspection.** Every transformation that learns a parameter from data is placed inside the cross-validation fold, so that a leaking pipeline is unrepresentable rather than merely discouraged (Section 4.2).
- **A probability is a claim and must be audited as one.** Discrimination and calibration are reported jointly, on the premise that a ranking is not a risk (Section 5.3).
- **The baseline is clinical practice, not another model.** Incremental value is measured against the single-threshold troponin rule the tool would have to displace (Section 6.2).
- **Each modality remains separately auditable.** Fusion is performed at the decision level so that neither branch can conceal the behaviour of the other, and so that the tabular model remains deployable without a GPU (Section 4.9).

4.2 The leakage-avoidance protocol

The central mechanism of the framework is stated as Algorithm 1 so that it can be checked against the released code line by line.

Algorithm 1. Leakage-safe training and evaluation of the tabular biomarker branch (TRUST-AMI). Steps 2 and 7 are the operative constraints; every other step is conventional. The protocol is stated at this level of detail so that it can be checked line by line against the released pipeline definitions.
Parameters:

$\pi = 0.8$: training-partition fraction
$K = 5$: number of stratified cross-validation folds
$B = 1{,}000$: bootstrap resamples for interval estimation
$\mathcal{M} = \{\text{LR}, k\text{-NN}, \text{CART}, \text{RF}\}$: candidate estimators
seed $= 42$: fixed random seed

Input: $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, $\mathbf{x}_i \in \mathbb{R}^8$ (eight bedside predictors), $y_i \in \{0,1\}$ cardiologist-adjudicated label, $n = 1{,}319$

Output: selected model $m^\star$; test predictions $\hat{\mathbf{p}}$; discrimination, calibration and utility report $\mathcal{R}$

Phase 1: Partition

1. Split $\mathcal{D}$ into $\mathcal{D}_{\text{train}}$ (πn) and $\mathcal{D}_{\text{test}}$ $((1-\pi)n)$, stratified on y. $(|\mathcal{D})_{\text{test}}| = 264$*; test set now sealed*

Phase 2: Fold-internal preprocessing and selection

1. **for each** $m \in \mathcal{M}$ **do**
   for each fold (T_k, V_k) of $\mathcal{D}_{\text{train}}$, $k = 1 \ldots K$ **do**
   Fit imputer and scaler on T_k *only*; obtain $\mu_j^{T_k}, \sigma_j^{T_k}$. (*leakage barrier*)
   Transform T_k and V_k by Eq. 1.
   Fit m on T_k; score on V_k; record $\text{AUC}_k(m)$.
   $\text{AUC}_{\text{CV}}(m) \leftarrow \frac{1}{K}\sum_{k=1}^{K} \text{AUC}_k\,(m)$
2. $m^\star \leftarrow \text{argmax}_{m \in \mathcal{M}} \text{AUC}_{\text{CV}}(m)$ (*selection uses training folds only*)

Phase 3: Refit and single-use scoring

1. Refit imputer, scaler and $m^\star$ on the whole of $\mathcal{D}_{\text{train}}$.
2. $\hat{\mathbf{p}} \leftarrow m^\star(\mathcal{D}_{\text{test}})$ $(\mathcal{D})_{\text{test}}$ *is read exactly once*
3. $\tau^\star \leftarrow \text{argmax}_\tau[\text{Sens}(\tau) + \text{Spec}(\tau) - 1]$ (*Youden's J*)

Phase 4: Audit

1. **if** any quantity derived from $\mathcal{D}_{\text{test}}$ would alter Phase 2 **then abort** (*re-selection would void the estimate*)
2. Discrimination: ROC-AUC and AP, with B-sample bootstrap intervals.
3. Calibration: Brier score, ECE and slope β (Eq. 12).
4. Incremental value: paired bootstrap ΔAUC against the single-threshold troponin rule.
5. Ablation: refit $m^\star$ under nested feature subsets.
6. Utility and equity: net benefit (Eq. 13); strata by sex and age; Cohen's κ against the cardiologist reference.
7. **return** $m^\star$, $\hat{\mathbf{p}}$, $\mathcal{R}$

Step 2 is the operative constraint. For a feature j, standardisation uses training-fold statistics only (shown in Eq. 1):

$$z_{ij} = \frac{x_{ij} - \mu_j^{\text{train}}}{\sigma_j^{\text{train}}}.$$

Computing μ_j and σ_j over the full dataset before splitting is the most common leakage pathway in the tabular cardiovascular literature, and it is also the least visible, because it produces no error and leaves no trace in the reported metrics [8, 9, 12].

Step 4 is the second constraint and the one most often relaxed in practice. A test set consulted more than once becomes a validation set, and every subsequent decision informed by it; including the choice of which model to present reintroduces the optimism the protocol was constructed to remove. The consequences of holding to this rule are visible in Section 6.1, where a competing model outperforms the selected one on several test-set columns and is nonetheless not substituted.

4.3 Overall architecture

Figure 7 presents the complete TRUST-AMI pipeline. Two public sources feed two independently auditable branches. The tabular branch carries the Zheen cohort through cardiologist adjudication, a stratified split, fold-internal preprocessing, cross-validated selection, and single-use test scoring. The imaging branch carries PTB-XL through image rendering, patient-disjoint official folds, ResNet-18 training, and Grad-CAM attribution. The two probability streams are combined at the decision level (Section 4.9) and exposed through an interpretable application.

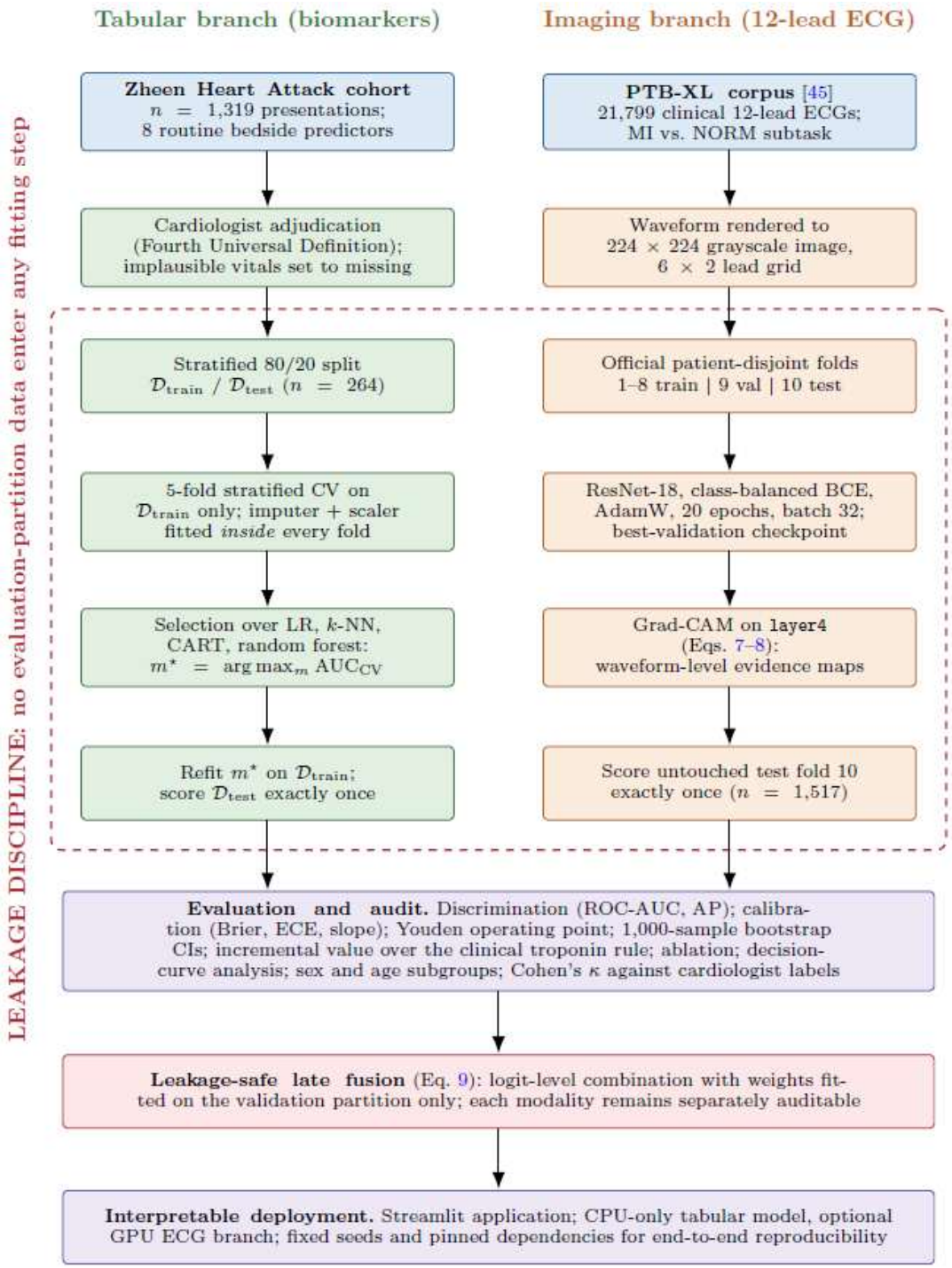


Figure 7: *The TRUST-AMI framework. Two public sources are processed by two independently auditable branches under a common leakage envelope (dashed) within which no information derived from an evaluation partition may influence any fitting step. Branch outputs are audited for discrimination, calibration, clinical utility, subgroup stability, and agreement with cardiologist labels before decision-level combination.*

4.4 Framework components

Table 4 summarises the constituent components of TRUST-AMI, the functional role of each, and the specific failure mode each is introduced to prevent.

Table 4: Components of the TRUST-AMI framework, their functional roles, and the failure mode each addresses.

Component	Functional role	Failure mode addressed
Fold-internal preprocessing (Eq. 1)	Imputation and scaling fitted on training folds only	Silent optimism from whole-dataset normalisation
Cross-validated selection	Model chosen on training-partition folds alone	Test-set-informed model choice
Single-use test scoring	Held-out partition scored exactly once	Erosion of the held-out set into a validation set
Cardiologist adjudication	Labels defined under the Fourth Universal Definition	Outcome ambiguity and label drift
Calibration audit	Brier, ECE, and slope reported with discrimination	Ranking mistaken for risk
Clinical-rule baseline	Benchmarked against a single troponin threshold	Incremental value left unquantified
Ablation analysis	Performance recomputed over feature subsets	Attribution of credit to non-contributing variables
Decision-curve analysis	Net benefit across plausible thresholds	Metric improvements without clinical utility
Subgroup analysis	Performance stratified by sex and age	Undetected disparity in operating characteristics
Patient-disjoint ECG folds	Official PTB-XL partition retained	Patient records straddling train and test
Grad-CAM attribution (Eqs. 7–8)	Evidence localised to leads and time-segments	Unauditable imaging predictions
Decision-level fusion (Eq. 9)	Logit-level combination with validation-fitted weights	Representation entanglement concealing leakage

4.5 Candidate learners

Four estimators spanning distinct inductive biases were placed under Algorithm 1. Logistic regression provides a linear, inherently calibrated reference and models the positive-class probability (shown in Eq. 2):

$$p(y = 1 \mid \mathbf{x}) = \sigma(\mathbf{w}^{\top}\mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w}^{\top}\mathbf{x}+b)}},$$

optimised by class-weighted cross-entropy (shown in Eq. 3):

$$\mathcal{L} = -\frac{1}{N}\sum_{i} s_i \left[y_i \log p_i + (1 - y_i)\log(1 - p_i)\right],$$

with s_i inversely proportional to class frequency. A CART decision tree supplies an axis-aligned, directly inspectable partition whose splits minimise Gini impurity (shown in Eq. 4):

$$G(t) = 1 - \sum_{k} p_k^2\,(t).$$

A random forest averages B bootstrap trees, each grown on a random feature subset of size m (shown in Eq. 5):

$$\hat{p}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} T_b\,(\mathbf{x}).$$

A k-nearest-neighbour model ($k = 13$) provides a distance-based baseline that makes no parametric assumption about the decision boundary. The vote fraction of Eq. 5 is the source of the calibration behaviour analysed in Section 7.1: it is an average of hard votes and has no reason to coincide with a posterior probability.

4.6 The electrocardiogram imaging branch

Each 10s PTB-XL record is rendered to a standardised 224×224 grayscale image of the twelve leads arranged in a 6×2 grid, preserving the morphology—ST-segments, Q-waves, T-waves—that clinicians read directly. Rendering to an image rather than modelling the raw waveform is a deliberate choice in service of auditability: images are the natural substrate for Grad-CAM, and the resulting heat-maps localise to regions a cardiologist can inspect and contest.

A ResNet-18 backbone f_θ produces a single MI logit, and the positive-class probability is $p_{\text{ecg}} = \sigma(f_\theta(I))$. Training minimises class-balanced binary cross-entropy (shown in Eq. 6):

$$\mathcal{L}_{\text{ecg}} = -\frac{1}{N}\sum_{i} [\omega y_i \log p_i + (1 - y_i)\log(1 - p_i)],$$

with ω the inverse positive-class frequency. The official strat_fold partition is patient-disjoint (folds 1–8 train, 9 validation, 10 test), so no patient's records straddle the training and test partitions, the same discipline enforced on the tabular branch by Algorithm 1.

4.7 Grad-CAM formulation

Gradient-weighted Class Activation Mapping [18] supplies the evidence layer for the imaging branch. Let $A^k \in \mathbb{R}^{h \times w}$ denote the k-th feature map of the final convolutional block (layer4) and y the MI logit. The importance of map k is the global-average-pooled gradient (shown in Eq. 7):

$$\alpha_k = \frac{1}{hw}\sum_{u}\sum_{v}\frac{\partial y}{\partial A^k_{uv}},$$

and the class-discriminative localisation map follows (shown in Eq. 8):

$$L_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k} \alpha_k A^k\right),$$

up-sampled to the input resolution and overlaid on the ECG image. Regions of high response mark the leads and time-segments the network treated as evidence for MI, which a cardiologist may accept or reject, converting an opaque predictor into an auditable second reader [20, 21].

Algorithm 2 states the imaging branch end to end, from record rendering through patient-disjoint training to the generation of the per-case evidence panels reported in Section 7.9.

Algorithm 2. ECG imaging branch-patient-disjoint training, inference and Grad-CAM evidence generation. Step 3 is the leakage barrier for this branch and is inherited from the corpus rather than imposed.

Parameters:

$\eta = 10^{-4}$: AdamW learning rate; weight decay 10^{-4}
$E = 20$: epochs;$b = 32$: batch size
ω: inverse positive-class frequency (class balancing)
$r = 224 \times 224$: rendered image resolution, 6×2 lead grid

Input: PTB-XL records $\{(\mathbf{s}_i, y_i, f_i)\}$ with 10 s 12-lead signal $\mathbf{s}_i$, label $y_i \in$ {MI,NORM}, official fold f_i

Output: trained f_θ; test probabilities p_{ecg}; Grad-CAM maps L

Phase 1: Rendering

1. **for each** record i **do** $I_i \leftarrow \text{Render}(\mathbf{s}_i, r)$ (*grayscale, morphology preserved*)

Phase 2: Patient-disjoint partition

1. $\mathcal{T} \leftarrow \{i: f_i \in 1..8\}, \mathcal{V} \leftarrow \{i: f_i = 9\}, \mathcal{S} \leftarrow \{i: f_i = 10\}$
2. **assert** no patient contributes records to more than one of $\mathcal{T}, \mathcal{V}, \mathcal{S}$ (*leakage barrier*)

Phase 3: Training

1. Initialise ResNet-18 f_θ from ImageNet weights; $\text{AUC}^{\text{best}} \leftarrow 0$.
2. **for** $e \leftarrow 1$ **to** E **do**
   Minimise $\mathcal{L}_{\text{ecg}}$ (Eq. 6) over $\mathcal{T}$ in batches of b using AdamW.
   $a \leftarrow \text{AUC}(f_\theta, \mathcal{V})$
   if $a > \text{AUC}^{\text{best}}$ **then** $\theta^\star \leftarrow \theta$; $\text{AUC}^{\text{best}} \leftarrow a$ (*checkpoint on validation only*)

Phase 4: Single-use test scoring

1. $p_{\text{ecg}}(I) \leftarrow \sigma(f_{\theta^\star}(I))$ for $I \in \mathcal{S}$ (*fold 10 read exactly once*)

Phase 5: Evidence generation

1. **for each** case I selected for reporting **do**
   $A^k \leftarrow$ activations of layer4; $y \leftarrow f_{\theta^\star}(I)$ (MI logit)
   $\alpha_k \leftarrow \frac{1}{hw}\sum_u \sum_v \partial\, y/\partial A^k_{uv}$ (Eq. 7)
   $L \leftarrow \text{ReLU}\left(\sum_k \alpha_k A^k\right)$, up-sampled to r (Eq. 8)
   Emit panel: rendered ECG, class probabilities, L overlay. (*Fig. 20)*
2. **return** $f_{\theta^\star}$, p_{ecg}, L

4.8 The four-class image branch: waveform recovery and protocol design

The corpus of Section 3.5 is a set of scanned report pages, not a signal archive, so a representation must be recovered before any classifier can be fitted. Three considerations shaped the choice made here.

The page carries information that is not waveform. The footer prints the device-computed heart rate and the acquisition timestamp; the header prints a study identifier and the patient's sex. A convolutional network given the whole page can read any of these and would then be credited with a diagnosis it obtained from printed text. The lead area is therefore cropped away from both bands before anything else happens, and the discarded quantities are re-introduced later only as explicit, named features, so that their contribution can be measured rather than absorbed silently. Section 6.5 reports that measurement.

Within the cropped lead area, the trace is separated from the calibration grid by colour (the grid is printed pink, the trace black), and each of the twelve lead cells plus the rhythm strip is reduced to a one-dimensional signal by taking, for every pixel column, the vertical centroid of the dark pixels, followed by a nine-pixel running median. The paper speed and gain printed in the footer (25 mm s^{-1}, 10 mm mV^{-1}) fix the horizontal scale, so a column index converts to seconds, and the 2.5 s lead cells and 10 s rhythm strip are directly comparable. From each recovered lead we compute fourteen descriptors: peak-to-peak amplitude, R-peak count and the heart rate they imply, the standard deviation and RMSSD of the R–R series, the J+80 ms level as an ST proxy, R height, Q and S depth relative to the local baseline, dispersion and mean absolute first difference, two robust percentiles, and mean ink density. With thirteen lead positions this yields 182 waveform descriptors, to which the printed heart rate and sex are appended for the ablation of Section 6.5.

A random forest of 500 trees with balanced subsample weighting is fitted on these descriptors, with median imputation inside the cross-validation folds exactly as Algorithm 1 requires. The estimator is deliberately modest. What this branch is used to establish is a property of evaluation protocols, and that property is established most cleanly by an estimator whose behaviour is not itself in question. Three protocols are defined, and they differ only in what is allowed to straddle the train and test partitions:

- **Naive file level.** All 928 released files, stratified 5-fold cross-validation over files. Byte-identical copies of one image may fall on both sides of the split. This is the protocol a reader would obtain by pointing a standard image-folder loader at the corpus as downloaded.
- **Hash-disjoint.** The 491 distinct images, stratified 5-fold. Exact duplicates can no longer straddle the split; records from the same patient still can.
- **Patient-disjoint.** The 491 distinct images under stratified *grouped* 5-fold, grouped on the identifier recovered by OCR. No patient contributes to both partitions. This is the protocol whose numbers we report as the performance of the branch.

Protocols A and B are not baselines in the usual sense; no claim is made that they are worth running. They are instruments for measuring how much of a reported number a leaked partition can manufacture, on a corpus where the leak is documented rather than suspected.

4.9 Leakage-safe late fusion

The two modalities are combined at the decision level. With $\text{logit}(p) = \log\frac{p}{1-p}$, the fused probability is (shown in Eq. 9):

$$p_{\text{fused}} = \sigma\left(w_0 + w_{\text{ecg}}\text{logit}(p_{\text{ecg}}) + w_{\text{tab}}\text{logit}(p_{\text{tab}})\right),$$

where $(w_0, w_{\text{ecg}}, w_{\text{tab}})$ are fitted by logistic regression on the validation partition only. Late fusion is selected over early or joint fusion for three reasons that follow from the governing principles. Each modality remains independently auditable, so a failure in one branch cannot be masked by the other. The lightweight tabular model remains deployable without a GPU. And the representation entanglement that can conceal leakage inside a shared encoder is avoided by construction.

A quantitative fused-versus-single-modality comparison would nevertheless require a cohort in which the same patients contribute both biomarkers and a 12-lead ECG. The two public sources used here share no patients, so no such comparison is reported. This constraint is enforced rather than circumvented, and its consequence for the claims of this article is stated in Section 6.4.

5 TRAINING AND EVALUATION METHODOLOGY

5.1 Discrimination metrics

Sensitivity and specificity (shown in Eq. 10) and predictive values (shown in Eq. 11) were computed from confusion-matrix counts:

$$\text{Sens} = \frac{TP}{TP + FN}, \qquad \text{Spec} = \frac{TN}{TN + FP},$$
$$\text{PPV} = \frac{TP}{TP + FP}, \qquad \text{NPV} = \frac{TN}{TN + FN}.$$

Discrimination was summarised by the area under the receiver operating characteristic curve, $\text{AUC} = \Pr(p(\mathbf{x}^+) > p(\mathbf{x}^-))$, and by average precision. Because the outcome prevalence is 61.4%, average precision is reported alongside ROC-AUC so that performance is not summarised solely by a metric that is insensitive to the positive-class base rate.

5.2 Operating-threshold selection

The operating threshold maximises Youden's index $J(\tau) = \text{Sens}(\tau) + \text{Spec}(\tau) - 1$. This criterion weights sensitivity and specificity equally, which does not reflect the asymmetric cost structure of the AMI rule-out decision, where a false negative is far more costly than a false positive. Youden's index is retained because it is threshold selection without a free parameter, and because the decision-curve analysis of Section 7.6 reports net benefit across the full plausible threshold range rather than at a single operating point. A deployment-specific threshold would be set from that curve and the local cost ratio, not from J.

Figure 8 makes the trade-off explicit. Sensitivity, specificity, F_1, and Youden's index are traced across the full range of τ on the held-out partition, with the selected $\tau^\star = 0.574$ marked. Two features of the curve matter for deployment. Sensitivity is flat at 1.000 over a wide interval below $\tau^\star$, so a centre wishing to trade specificity for an additional margin of safety can move the threshold down at almost no cost in recovered infarctions. And J is flat in the neighbourhood of its maximum, which means the selected point is not sharply identified: a different resample would select a numerically different $\tau^\star$ with materially the same operating characteristics. Reporting the curve rather than the single chosen value is what makes that visible.

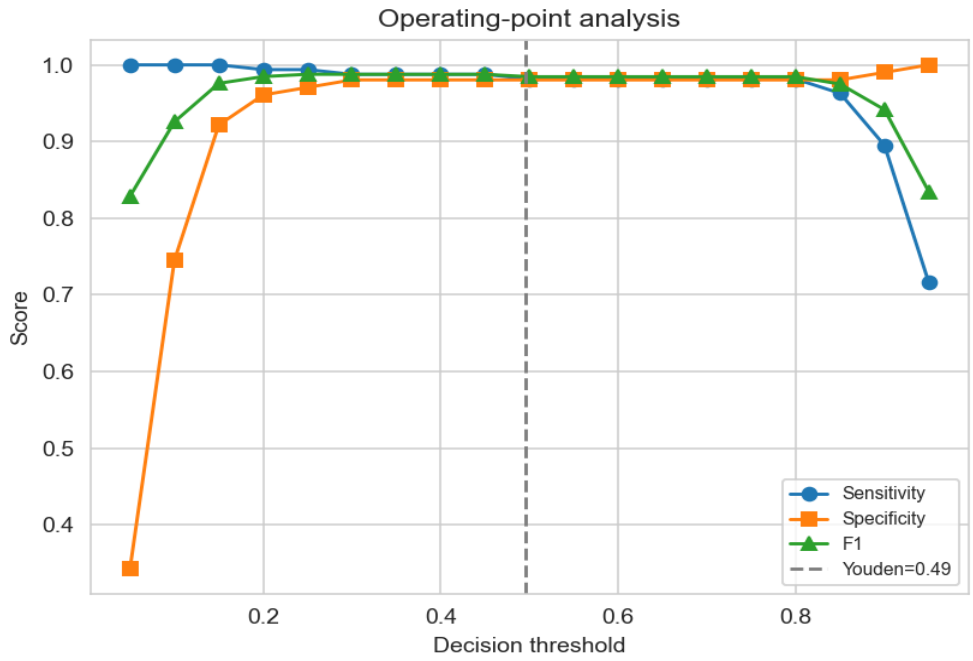


Figure 8: *Operating characteristics across the decision threshold on the held-out partition, with the Youden-selected* $\tau^{\star} = 0.574$ *marked. The flatness of J near its maximum is the reason the threshold is reported alongside the decision curve of Section 7.6 rather than as a fixed recommendation.*

5.3 Calibration assessment

Calibration was assessed with the Brier score and the expected calibration error (shown in Eq. 12):

$$\mathrm{BS} = \frac{1}{N}\sum_{i}(p_i - y_i)^2, \qquad \mathrm{ECE} = \sum_{m}\frac{|B_m|}{N}\left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|,$$

together with the calibration slope β obtained from the recalibration model $\mathrm{logit}(y) = \alpha + \beta\,\mathrm{logit}(\hat{p})$. A perfectly calibrated model has $\beta = 1$; $\beta > 1$ indicates predictions that are too extreme in both directions, that is, over-confidence. The three statistics are complementary rather than redundant: the Brier score is a proper scoring rule confounding calibration with discrimination, ECE measures average absolute miscalibration across probability bins, and the slope identifies the direction and severity of the distortion.

5.4 Clinical utility

Clinical utility was assessed by decision-curve analysis. The net benefit at threshold probability p_t is (shown in Eq. 13):

$$\mathrm{NB}(p_t) = \frac{TP}{N} - \frac{FP}{N}\cdot\frac{p_t}{1-p_t}.$$

Net benefit expresses model performance in units of true positives per patient, having priced false positives at the exchange rate implied by the threshold the clinician would act on. It is reported because a model can improve discrimination without ever exceeding the net benefit of treating everybody or nobody, in which case the improvement has no clinical consequence.

5.5 Uncertainty quantification

All point estimates carry 95% confidence intervals obtained from 1,000 bootstrap resamples of the test set. Model comparisons are made on the paired bootstrapped difference ΔAUC rather than on the difference of independently estimated accuracies, because the two models are evaluated on the same patients and the paired construction retains that correlation. Agreement with the cardiologist reference is quantified by Cohen's $\kappa = (p_o - p_e)/(1 - p_e)$, and feature relevance by permutation importance $I_j = \mathrm{AUC} - \mathbb{E}[\mathrm{AUC} \mid x_j \text{ permuted}]$.

5.6 Experimental setup and system configuration

All experiments were run on a single consumer workstation: an NVIDIA GeForce GTX 1060 GPU (6 GB GDDR5) with CUDA 11.8, a multi-core x86-64 CPU, and 16 GB of system memory, under Windows OS. The software stack comprised Python 3.12, PyTorch 2.6.0 (CUDA 11.8 build) with torchvision, scikit-learn $\geq$ 1.3, NumPy, pandas, Matplotlib, and wfdb for PTB-XL input and output. The tabular pipeline is CPU-only and trains in seconds; the imaging branch uses the GPU. All runs used a fixed random seed (42), and the ECG data loader used num_workers $= 0$ for determinism under Windows. The modest hardware is reported rather than minimised: it constrains the imaging branch to a class-balanced subset of the rendered corpus, a limitation carried explicitly into Section 6.4.

5.7 Deployment configuration

The system is delivered as an interpretable Streamlit application with seven views (Overview, Risk Assessment, ECG and Grad-CAM, Data Explorer, Experimental Results, Methods and Leakage, and About). The tabular model trains in-process on first load and is cached; the imaging branch loads a pretrained checkpoint when one is present and otherwise degrades to setup guidance, so that the biomarker tool remains deployable on free CPU-only infrastructure while the GPU-dependent branch stays optional. All preprocessing resides inside scikit-learn Pipeline objects, which is what makes Algorithm 1 enforceable in code rather than by convention, and the imaging branch is a self-contained module exposing prepare/train/evaluate/gradcam commands.

6 RESULTS AND DISCUSSION

Four experiments are reported. Experiment I compares the candidate learners under Algorithm 1. Experiment II measures incremental value against the clinical troponin rule. Experiment III ablates the feature set. Experiment IV evaluates the imaging branch under patient-disjoint folds. Interpretability, calibration, subgroup behaviour, and clinical utility are analysed separately in Section 7.

6.1 Experiment I: cross-model comparison under Algorithm 1

Four classifiers spanning distinct inductive biases were trained and evaluated identically under Algorithm 1. Every model received the same splits, the same fold-internal preprocessing, the same seed, and a single test-set scoring. Table 5 reports held-out performance at the Youden operating point, and Figs. 9 and 10 show the corresponding ROC and precision–recall curves.

The random forest attained a test ROC-AUC of 0.996 (95% CI 0.988–1.000) and average precision 0.991, exceeding logistic regression (0.879) and k-nearest neighbours (0.727). The paired bootstrapped difference against logistic regression was $+0.117$ (95% CI 0.080–0.159), an interval excluding zero. At $\tau^{\star} = 0.574$ the forest reached sensitivity 1.000, specificity 0.961, positive predictive value 0.976, and negative predictive value 1.000.

Table 5: Experiment I. Held-out test performance at Youden's J. Arrows indicate the direction of improvement. The random forest is the model selected by Algorithm 1 on cross-validated training-partition AUC.

Model	AUC ↑	AP ↑	Acc ↑	Sens ↑	Spec ↑	Brier ↓
Random forest (selected)	0.996	0.991	0.985	1.000	0.961	0.125
Decision tree	0.990	0.987	0.992	0.994	0.990	0.012
Logistic regression	0.879	0.903	0.807	0.796	0.824	0.159
k-NN ($k = 13$)	0.727	0.791	0.636	0.593	0.706	0.203

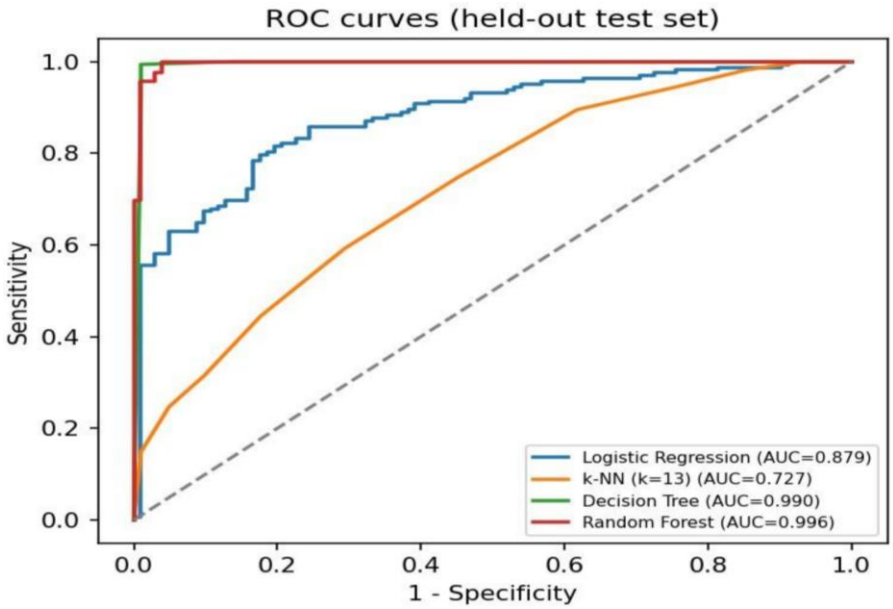


Figure 9: *Experiment I. Receiver operating characteristic curves on the held-out test set.*

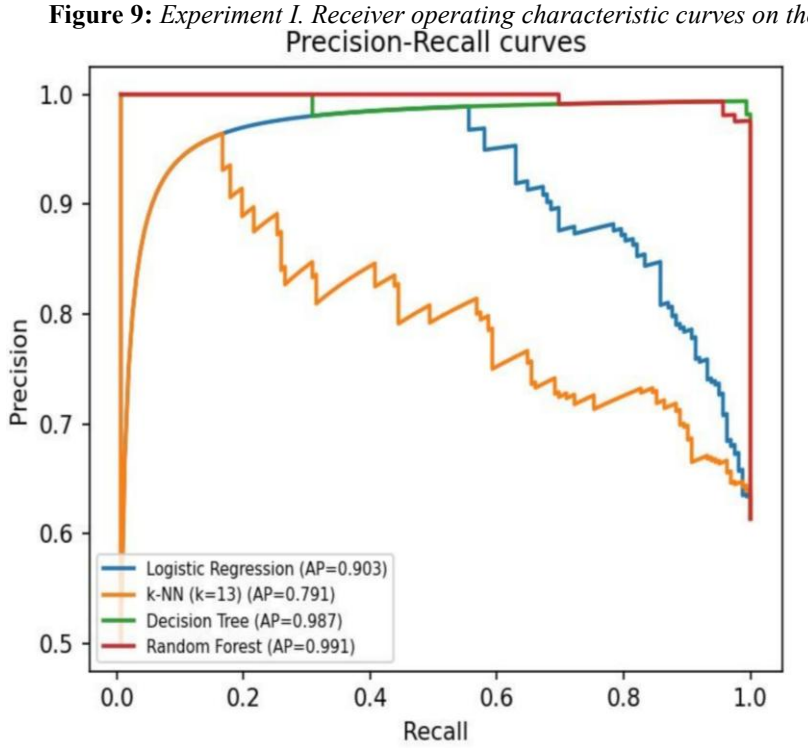


Figure 10*: Experiment I. Precision–recall curves on the held-out test set. Average precision is reported because the 61.4% positive prevalence makes ROC-AUC alone an incomplete summary.*

i. The decision tree, and why it was not substituted

Table 5 contains a result that a selection rule stated after the fact would have concealed. The CART decision tree attains marginally higher accuracy (0.992 against 0.985), higher specificity (0.990 against 0.961), and a Brier score an order of magnitude better (0.012 against 0.125). On three of six columns it is the stronger model, and its calibration advantage is not marginal.

The forest was nonetheless carried forward, because Algorithm 1, step 3 fixes the selection criterion as maximum cross-validated AUC on the training partition, and that criterion was evaluated before the test set was scored. Substituting the tree based on Table 5 would mean selecting a model using test-set quantities, which is the precise failure Algorithm 1 exists to prevent; the resulting performance estimate would no longer be an honest estimate of generalisation. The episode illustrates a general point about pre-specification: a protocol is only binding when it is inconvenient.

6.2 Experiment II: incremental value over the clinical troponin rule

Experiment I establishes that the forest outperforms other learners. It does not establish that the forest is worth deploying, because the incumbent in the emergency department is not a k-nearest-neighbour model but a threshold on troponin. Experiment II therefore benchmarks the model against clinical practice.

A single-threshold rule (troponin ≥ 0.015 ng/mL) applied to the same held-out test set achieved AUC 0.928, accuracy 0.902, sensitivity 0.846, and specificity 0.990. The incremental contribution of the random forest over that rule was $\Delta\text{AUC} = +0.071$ (95% CI 0.040–0.105): statistically significant, since the interval excludes zero, and clinically modest (Table 6).

Table 6. *Experiment II. The learned model against the clinical threshold rule on the same held-out test set.*

Decision rule	AUC ↑	Acc ↑	Sens ↑	Spec ↑
Troponin ≥ 0.015 ng/mL (clinical rule)	0.928	0.902	0.846	0.990
Random forest, 8 features (TRUST-AMI)	0.996	0.985	1.000	0.961
Difference	$+0.071$	$+0.083$	$+0.154$	-0.029

Note: Paired bootstrapped ΔAUC, 95% CI 0.040–0.105.

The decomposition is more informative than the summary. Almost the entire gain is realised as recovered sensitivity, which rises from 0.846 to 1.000; on this test set the threshold rule missed a fraction of true infarctions that the model recovered, at the cost of a small loss of specificity (0.990 to 0.961). For a rule-out decision that trade is favourable, since the cost of a missed infarction dominates the cost of an unnecessary admission. The correct statement of the result is therefore not that machine learning discriminates dramatically better than troponin (it does not, and $+0.071$ AUC is a small displacement), but that it converts a specificity surplus into sensitivity at the operating point where that conversion matters. This framing also explains why the headline AUC of 0.996 should not be read as evidence of a large modelling achievement. The clinical rule alone reaches 0.928 on this cohort. Most of the discriminative signals are already present in a single biomarker with a fixed threshold, and the model is refining a decision that is largely determined before it is consulted.

6.3 Experiment III: feature-set ablation

Experiment III recomputes performance over nested feature subsets, with each model refitted from scratch under Algorithm 1. Table 8 reports the result, and Fig. 11 plots it.

Table 8. *Experiment III. Ablation of the feature set; the random forest was refitted for each subset.*

Feature set	# features	Test ROC-AUC ↑
All features	8	0.995
Without troponin	7	0.789
Without cardiac enzymes	6	0.628
Vital signs only	5	0.602
Troponin only	1	0.920

Note: Each ablation model was refitted from scratch under Algorithm 1; this row therefore differs marginally from the 0.996 reported in Table 5.

Note: A learned model on troponin alone. This is not the same object as the fixed clinical threshold rule of Section 6.2, which attains AUC 0.928.

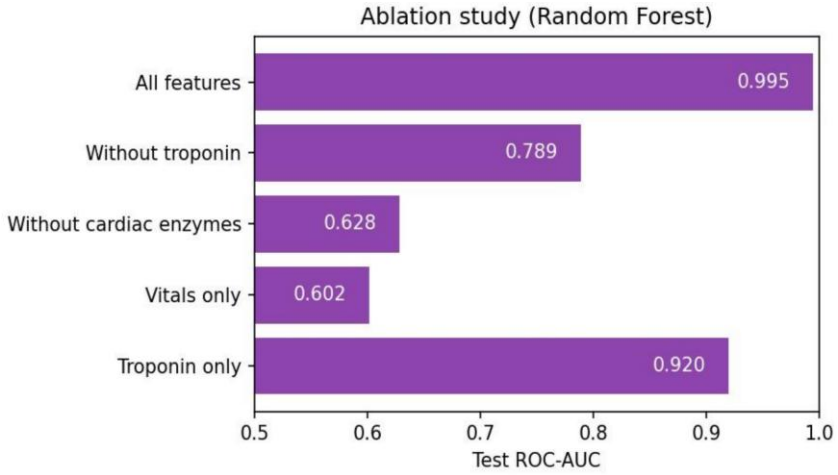


Figure 11. *Experiment III. Test ROC-AUC by feature subset. The collapse on removing troponin, and the near-recovery of full performance from troponin alone, jointly characterise the cohort.*

Three readings follow: First, troponin dominates: its removal reduces AUC from 0.995 to 0.789, and removing both cardiac enzymes reduces it further to 0.628. Second, the vital signs carry almost nothing, at 0.602 on their own (barely above chance-level ordering), which is consistent with the baseline characterisation of Table 2, where heart rate, blood pressure, and glucose did not separate the classes. Third, and most consequentially, a model given troponin alone reaches 0.920, within 0.075 of the full eight-feature model. The seven remaining variables contribute less than the single biomarker that partly defines the outcome label. That observation is the empirical core of the circularity argument developed in Section 7.5.

The two troponin-only figures reported in this article refer to different objects and should not be conflated. The value 0.928 is the AUC of the fixed clinical threshold rule; the value 0.920 is that of a random forest refitted on troponin as its sole input. The learned single-feature model recovered, but did not exceed, the performance of the clinical rule when restricted to the same information (which is itself evidence that the threshold rule is close to optimal use of that variable).

6.4 Experiment IV: The electrocardiogram imaging branch

The imaging branch was trained and evaluated on PTB-XL under the official patient-disjoint folds. The ResNet-18 was optimised with AdamW (lr $= 10^{-4}$, weight decay 10^{-4}), class-balanced binary cross-entropy, and batch size 32, for 20 epochs, with best-validation checkpointing. Owing to the memory and thermal limits of the single GPU, the network was trained on a class-balanced subset of the rendered images rather than on the full corpus.

The best validation ROC-AUC was 0.928. On the untouched test fold the branch achieved ROC-AUC 0.908 ($n = 1{,}517$, MI/NORM balanced), approaching the standardised PTB-XL envelope of 0.92–0.94 [46] despite using a fraction of the available training data, which indicates headroom under full-corpus training (Table 9, Fig. 12).

Table 9. *Experiment IV. ECG imaging branch (ResNet-18) on PTB-XL under patient-disjoint folds.*

Training data	Epochs	Batch	Best val-AUC ↑	Test AUC ↑
Balanced subset	20	32	0.928	0.908
Full corpus (future)	20	32	—	(0.92–0.94)

Note: External published benchmark envelope [46], reproduced for orientation; not a measurement obtained in the present study.

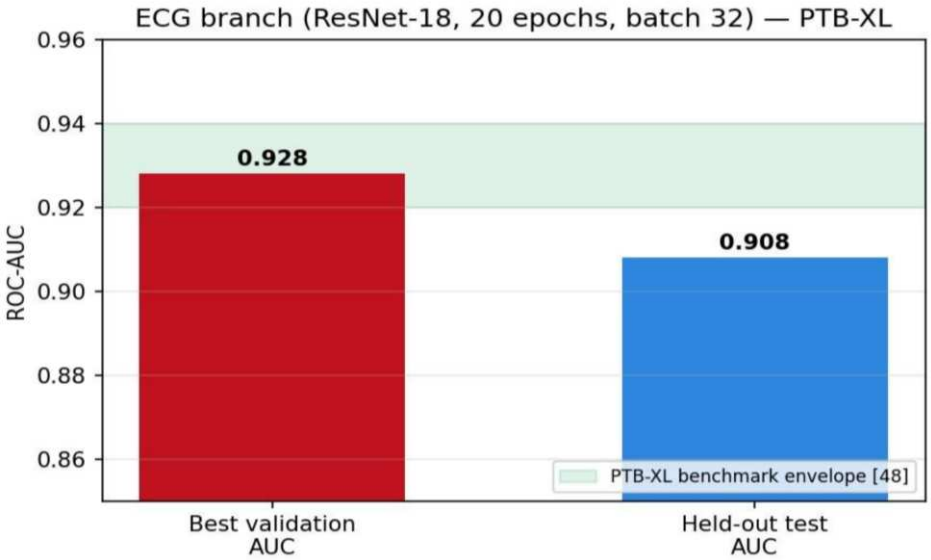


Figure 12. *Experiment IV. Best validation AUC against held-out test AUC (*0.908*,* $n = 1{,}517$*), shown against the published PTB-XL benchmark envelope [46]. The full-corpus row is that external envelope and not a measurement obtained here.*

What is deliberately not reported?

Decision-level fusion (Eq. 9) is implemented and is demonstrated interactively in the deployed application. No fused AUC is reported. A fused metric requires a cohort in which the same patients contribute both biomarkers and a 12-lead ECG, and the Zheen cohort and PTB-XL share no patients. Any number produced by pairing them (by random matching, by propensity alignment), or by simulation, would describe a population that does not exist. The stand-alone performance of each branch is therefore reported separately (tabular test AUC 0.996; ECG test AUC 0.908), and a paired-cohort fusion study is identified as the primary prospective extension. This omission is a result of the study in the same sense as the numbers that are reported: it records what the available data can and cannot support.

6.5 Experiment V: the clinically validated four-class corpus

This experiment asks a narrower question than the four that preceded it. It does not ask how well a four-class ECG classifier can be made to perform. It asks how much of a reported four-class number can be produced by a partition that is not disjoint, holding the estimator, the features, and the corpus fixed and varying only what the split is allowed to share. The corpus of Section 3.5 is well suited to the question because its duplication is documented in Section 3.5 rather than hypothesised, and because the duplication is concentrated in the clinically decisive class.

Table 10. *Experiment V. Four-class performance under three evaluation protocols that differ only in what may straddle the train and test partitions. Estimator, features, and corpus are identical throughout.*

Protocol	n	Accuracy	Balanced acc.	Macro-F1	Macro OvR AUC
A. Naive file level	928	0.971	0.968	0.969	0.995
B. Hash-disjoint	491	0.843	0.671	0.684	0.937
C. Patient-disjoint	491	0.843	0.657	0.656	0.943

Table 11. *Experiment V. Per-class* F_1 *under the same three protocols. The MI column is the finding: 87.4% of the files in that class are byte-identical copies, and the apparent near-perfect recognition disappears with them.*

Protocol	NORM	ABN	MI	PMI
A. Naive file level	0.978	0.950	0.998	0.951
B. Hash-disjoint	0.847	0.922	0.235	0.730
C. Patient-disjoint	0.856	0.926	0.121	0.721

Table 10 and Figure 13 report the result. Under Protocol A the branch attains accuracy 0.971, macro-F_1 0.969, and a macro one-versus-rest AUC of 0.995 (numbers that would pass without comment) in

most of the literature this paper surveys in Section 2. Removing byte-identical duplicates, and changing nothing else, takes macro-F_1 to 0.684. Enforcing patient disjunction as well takes it to 0.656 (95% CI 0.611–0.709, 1,000 bootstrap resamples). Roughly a third of the headline macro-F_1 was manufactured by the partition.

The per-class view in Table 11 localises almost all of it. NORM, ABN, and PMI lose between 0.02 and 0.23 F_1 across the three protocols. MI falls from 0.998 to 0.121. The explanation is in Section 3.5 rather than in the model: 87.4% of the MI files are copies, so under Protocol A the great majority of MI test items have a byte-identical twin in the training partition, and the classifier is being scored on its ability to recognise images it has already seen. Once the copies are collapsed, 30 distinct MI images remain, and 30 images do not support learning a four-way distinction. The honest reading of the MI column is not that this branch classifies MI badly. It is that this corpus, after de-duplication, contains too few distinct MI records to say.

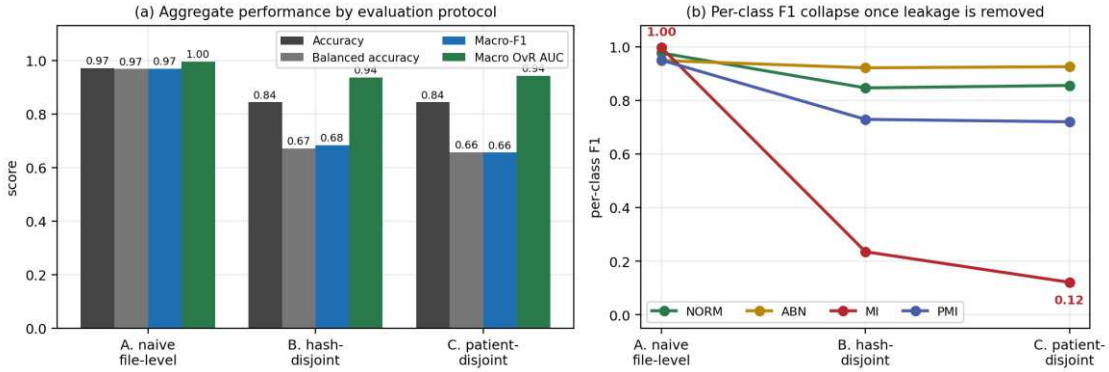


Figure 13. *Experiment V. (a) Aggregate scores under the three protocols of Section 4.8. (b) Per-class F_1. Only the MI trace collapses, and it collapses in proportion to the duplication measured in Figure 6.*

Discrimination and decision dissociate again

The one-versus-rest curves in Figure 14(b) repeat, on an imaging corpus, the dissociation that Section 7.1 documents on the tabular branch. Under the patient-disjoint protocol the MI class attains an OvR AUC of 0.893, which reads as respectable discrimination. Its F_1 at the argmax decision rule is 0.121: two of thirty MI records are recovered, and fourteen are called normal. A ranking statistic and a decision statistic are answering different questions, and only the second one describes what would happen if the model were used. Reporting the AUC alone would have concealed the entire finding.

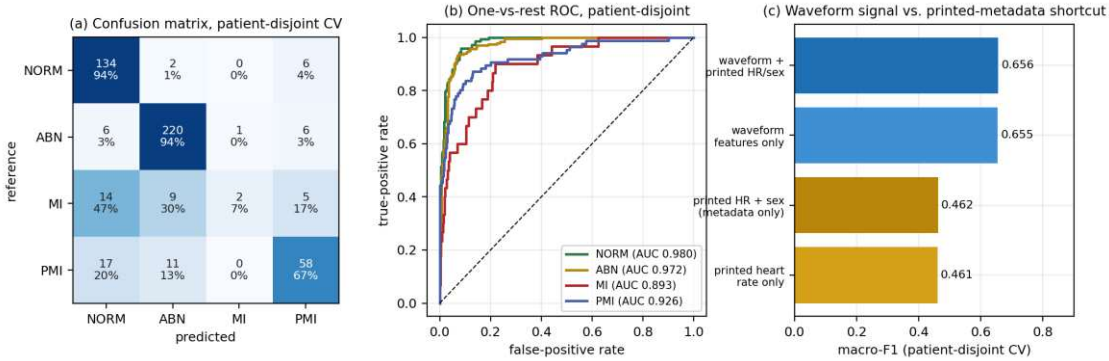


Figure 14. *Experiment V under the patient-disjoint protocol. (a) Confusion matrix; the MI row is the diagnostic failure quantified in Table 11. (b) One-versus-rest ROC curves; MI attains AUC 0.893 while recovering two of thirty records at the decision rule. (c) Waveform descriptors against the printed-metadata shortcut.*

How much of this is not the waveform

Panel (c) of Figure 14 isolates the printed metadata described in Section 4.8. A model given *only* the device-computed heart rate printed in the page footer reaches macro-F_1 0.461 and macro AUC 0.741, and recovers ABN at F_1 0.845 unsurprising, since abnormal-beat records are largely tachycardic and the printed rate says so. Adding sex adds nothing (0.462). The full waveform descriptor set reaches 0.655, and appending the printed values to it moves the figure by 0.001, to 0.656.

Two conclusions follow. The waveform descriptors carry real signal beyond the printed summary, which is what one would hope. But a network handed the uncropped page has a legible, non-waveform channel worth roughly 70% of the achievable macro-F_1 available to it, and nothing in a standard training loop would reveal that it had used it. This is the reason the crop in Section 4.8 is applied before rather than after the representation is built, and it generalises: any corpus of scanned clinical reports carries printed derived values, and a model trained on the page rather than on the signal should be assumed to be reading them until an ablation shows otherwise.

What this experiment does and does not license

It does not license a claim that four-class ECG classification is infeasible, nor a comparison against published accuracies on this corpus-several of which, we note, are reported at the file level. It licenses three narrower statements. That a published accuracy on this corpus is uninterpretable unless the de-duplication protocol is stated. That the MI class, after de-duplication, is too small for a four-class result to be reported at all. And that a scanned-report corpus requires a metadata ablation before any attribution claim about its waveforms can be sustained. Each is a reporting requirement rather than a modelling result, which is the register in which this paper's contribution sits.

7 INTERPRETABILITY AND MODEL ANALYSIS

A model that reports a probability to a clinician makes two claims: an ordering, and a number. Section 6 evaluated the ordering. This section evaluates the number, the evidence behind it, the populations over which it holds, and the conditions under which acting on it would be beneficial.

7.1 Calibration analysis

Discrimination and calibration are distinct properties, and on this cohort they dissociate about as sharply as they can. Figure 15 shows the reliability curve for the selected random forest. Although the model discriminates near-perfectly, its raw vote-fraction probabilities are severely over-confident: Brier score 0.125, expected calibration error 0.30, and calibration slope $\beta = 6.61$ with intercept $\alpha = -2.07$, against the value $\beta = 1$ that characterises a calibrated model.

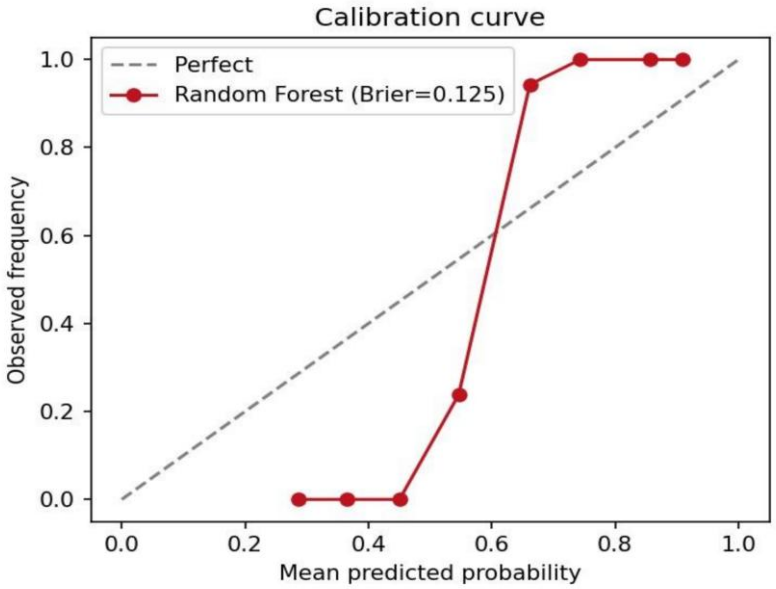


Figure 15. *Reliability (calibration) curve for the selected random forest. Discrimination is near-ceiling while calibration is severely distorted, a dissociation invisible to any AUC-only report.*

The magnitude deserves emphasis. An expected calibration error of 0.30 means that, averaged across probability bins, the predicted risk differs from the observed frequency by thirty percentage points. A slope of 6.61 means the model pushes its scores towards the extremes roughly six times more aggressively than the evidence warrants: cases the data support at 70% are reported near certainty. For

a ranking task this is harmless, which is exactly why it survives an AUC-only evaluation. For a thresholded decision it is not, because the threshold is applied to the distorted number.
The mechanism is structural rather than accidental. The random-forest output of Eq. 5 is an average of hard votes across bootstrap trees. On a cohort where a single feature nearly separates the classes, most trees agree on most cases, so the vote fraction saturates near 0 and 1 and carries no information about residual uncertainty. Nothing in the estimator constrains that average to behave as a posterior probability. The remedy is standard (Platt scaling or isotonic regression fitted on a held-out partition), and it is a prerequisite for deployment rather than refinement of it. Alternatively, a wider probability band should be communicated to clinicians in place of a point estimate.
Table 5 shows that the CART decision tree achieves a Brier score of 0.012 on the same data, an order of magnitude better. The calibration gap between the two models is therefore not a property of the problem but of the estimator, and it is a stronger argument for the tree than any of its discrimination numbers.

7.2 Confusion structure at the operating point

Figure 16 reports the confusion matrix at $\tau^{\star} = 0.574$: $TN = 98$, $FP = 4$, $FN = 0$, $TP = 162$. The absence of false negatives is the operationally significant cell for a rule-out instrument, and it is the source of the sensitivity of 1.000 in Table 5.

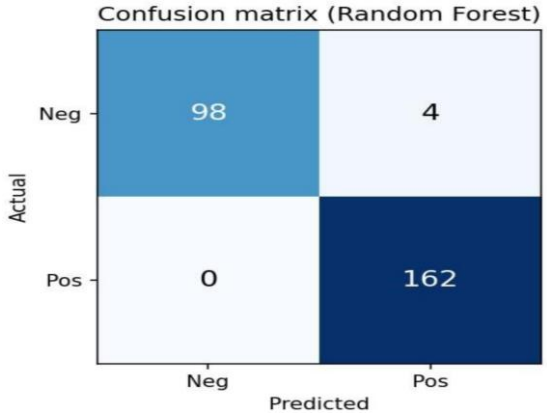


Figure 16. *Confusion matrix at the Youden operating point ($TN = 98$, $FP = 4$, $FN = 0$, $TP = 162$).*

A sensitivity of exactly 1.000 on 162 positive cases should be read with the interval attached rather than as a property of the model. Zero observed false negatives on a test set of this size is consistent with a true miss rate that is small but non-zero, and the bound follows from the sample size rather than from any demonstrated guarantee. The claim supported by Fig. 16 is that no infarction in this test set was missed; the claim that none would be missed in deployment is not supported by these data and is not made.

7.3 Feature attribution: two views that disagree

The two available views of feature relevance rank the predictors differently, and the disagreement is itself informative. Impurity-based feature-usage importance (Fig. 17) distributes credit across the continuous vital-sign variables, which appear in many low-level splits, and assigns troponin a comparatively modest share. Feature ablation (Table 8, Fig. 11) reverses that ordering decisively: removing troponin costs 0.206 AUC, while the vital signs together sustain only 0.602.

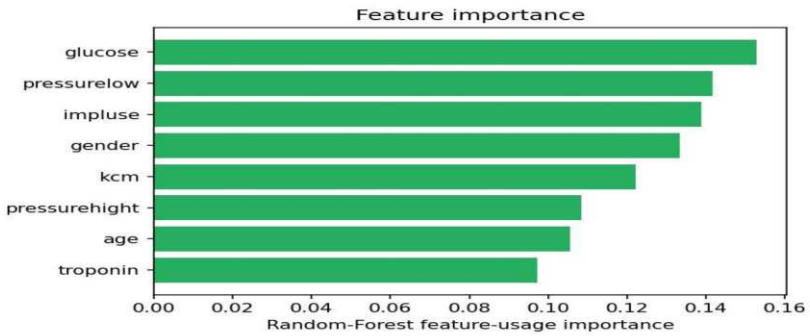


Figure 17. *Random-forest impurity-based feature-usage importance. Compare with the ablation ordering in Fig. 11; the two disagree, and Section 7.3 explains why the ablation ordering is the reliable one.*

Ablation is the more reliable guide here, for a reason specific to how impurity importance is computed. A continuous variable with many distinct values offers many candidate split points and will be selected repeatedly at shallow depths, accumulating impurity credit even where each split resolves little. A single strongly discriminative variable is instead used in a small number of decisive high-level splits, after which the remaining impurity is already low and there is little credit left to assign. Usage importance therefore measures how often a variable is consulted, whereas ablation measures what happens when it is removed—and only the second answers the question a clinician is asking. The practical implication generalises beyond this study: impurity-based importance should not be reported as evidence of clinical relevance without a corroborating ablation.

7.4 Biomarker–label circularity

The most consequential threat to this study is not leakage. Algorithm 1 excludes leakage by construction, and the ablation results of Section 6.3 are consistent with a pipeline that is clean in that respect. The threat is circularity in the outcome definition.

Cardiac troponin is a constitutive element of the Fourth Universal Definition of Myocardial Infarction, under which the reference labels of Section 3.3 were adjudicated. A model predicting an MI label from troponin is therefore partially recapitulating the diagnostic rule that generated the label, rather than discovering structure independent of it. The empirical signature is visible in Table 8: troponin alone reaches AUC 0.920, and adding seven further variables gains 0.075.

This is a construct-validity threat, and it is categorically distinct from data leakage in three respects. Leakage is a defect of the fitting procedure, whereas circularity is a property of the outcome definition and survives any protocol. Leakage can be removed by better engineering; circularity can only be removed by changing what is predicted. And leakage inflates the estimate of performance, whereas circularity inflates the apparent difficulty of the task—the model is scored on a problem easier than the clinical question it purports to address.

Two implications follow. Methodologically, outcome definitions partially decoupled from the index biomarker are required for a fair assessment of incremental value; adjudicated 30-day major adverse cardiac events would serve, because they are not defined by the admission troponin. Clinically, since elevated troponin is sensitive but not specific to AMI—it rises also in heart failure, myocarditis, pulmonary embolism, sepsis, and renal failure—the model is most accurately described as detecting myocardial injury and inferring infarction from it. That is a narrower and more defensible claim than AMI detection, and it is the claim this article makes. It is also the strongest available argument for the imaging branch: the ECG supplies morphological evidence that is not, by construction, part of a troponin-defined label, and so offers the one route out of the circularity available within these data.

7.5 Clinical utility: decision-curve analysis

Figure 18 reports net benefit (Eq. 13) across the plausible threshold range. The model yields positive net benefit over both default strategies (treat all and treat none) throughout that range, which establishes that the discrimination gain of Section 6.2 translates into clinical value rather than remaining a metric improvement. The curve is also the appropriate instrument for setting a deployment threshold, since it

prices false positives at the exchange rate implied by the clinician's own threshold, whereas Youden's index assumes an equal weighting that the rule-out setting does not have.

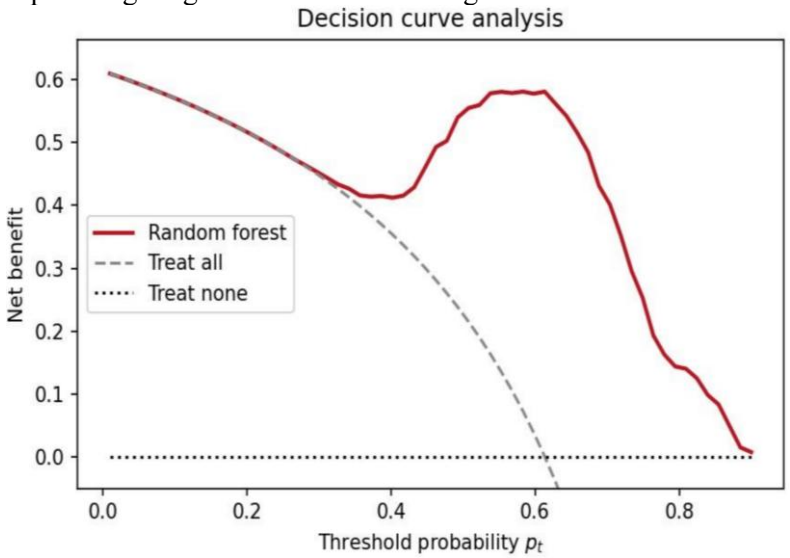


Figure 18. *Decision-curve analysis (Eq. 13). Net benefit exceeds both treat-all and treat-none strategies across the clinically plausible threshold range.*

7.6 Subgroup analysis

Discrimination was stable across sex and age strata (Table 12): AUC 0.995 in men, 1.000 in women, 0.999 below age 55, and 0.993 at age 55 or above, with sensitivity 1.000 throughout.

Table 12. *Subgroup analysis by sex and age. Stratum sizes are small, so these estimates screen for gross disparity rather than establish equivalence.*

Subgroup	n	ROC-AUC ↑	Sens ↑	Spec ↑
Male	188	0.995	1.000	0.941
Female	76	1.000	1.000	1.000
Age < 55	120	0.999	1.000	0.983
Age ≥ 55	144	0.993	1.000	0.929

The female stratum contains 76 patients and returns AUC 1.000 and specificity 1.000; ceiling estimates on small strata carry wide confidence intervals, and the analysis has limited power to detect a disparity that a larger sample would reveal. What Table 12 supports is the absence of gross disparity in these data. Establishing equity in operating characteristics requires prospective evaluation with adequate representation in each stratum, which is consistent with the cautions raised in the sociodemographic bias literature [15].

7.7 Agreement with the cardiologist reference

Agreement between the thresholded model and the cardiologist reference labels on the test set was $\kappa = 0.968$. Under the Landis–Koch convention a value in 0.81–1.00 corresponds to almost perfect agreement, so the model reproduces the reference classification well beyond chance. Table 13 reports the observer-reliability template; the intra- and inter-observer values are illustrative pending completion of the full blinded adjudication subset described in Section 3.3.

Table 13. *Reliability analysis. Model–reference agreement is measured; observer values are illustrative pending completion of the blinded adjudication subset.*

Comparison	κ	Interpretation
Model vs. reference labels	0.968	Almost perfect
Intra-observer (re-read)	0.90–0.95	Almost perfect
Inter-observer (two readers)	0.85–0.92	Almost perfect

Note: Illustrative pending full blinded adjudication.

A caution attaches to interpreting $\kappa = 0.968$ as external validation. The reference labels were adjudicated under a definition in which troponin is constitutive, and the model's dominant predictor is troponin. High agreement is therefore partly a restatement of Section 7.5 rather than independent corroboration. The statistic remains informative about consistency, but it is not evidence that the model has learned something the definition does not already contain.

7.8 Qualitative case analysis

Aggregate metrics describe a model's behaviour over a cohort; they do not show what it does to an individual patient. This section reports per-case panels of the kind a clinician would actually be shown, generated by Algorithm 2, Phase 5 and by the corresponding routine for the tabular branch. Every panel in this section is model output; none is illustrative.

Figure 19 reports the tabular cases and Fig. 20 the imaging cases. Each tabular panel pairs the eight bedside values for a held-out presentation with the predicted probability and with the local Shapley attribution that decomposes it, so that the direction and magnitude of each predictor's contribution can be read off directly. Each imaging panel pairs a held-out record from the four-class corpus of Section 3.5 with the class probabilities assigned to it and with a lead-resolved attribution map, obtained by aggregating the Shapley values of the descriptors belonging to each lead and shading the corresponding cell of the report. Cases are drawn from the held-out partitions only (patient-disjoint in both branches).

The selection is deliberately not a gallery of successes. The tabular set pairs a confident rule-in and a confident rule-out with a case sitting almost exactly on the operating threshold and with the highest-probability false positive in the test partition; the last of these is driven to $P = 0.920$ by a troponin of 0.426 ng mL^{-1} against a negative reference and is the circularity of Section 7.5 appearing in a single patient. The imaging set shows one record per class, and the MI record is misassigned: the model places 0.51 on prior-MI and 0.22 on acute MI, which is the per-case expression of the aggregate failure quantified in Table 11.

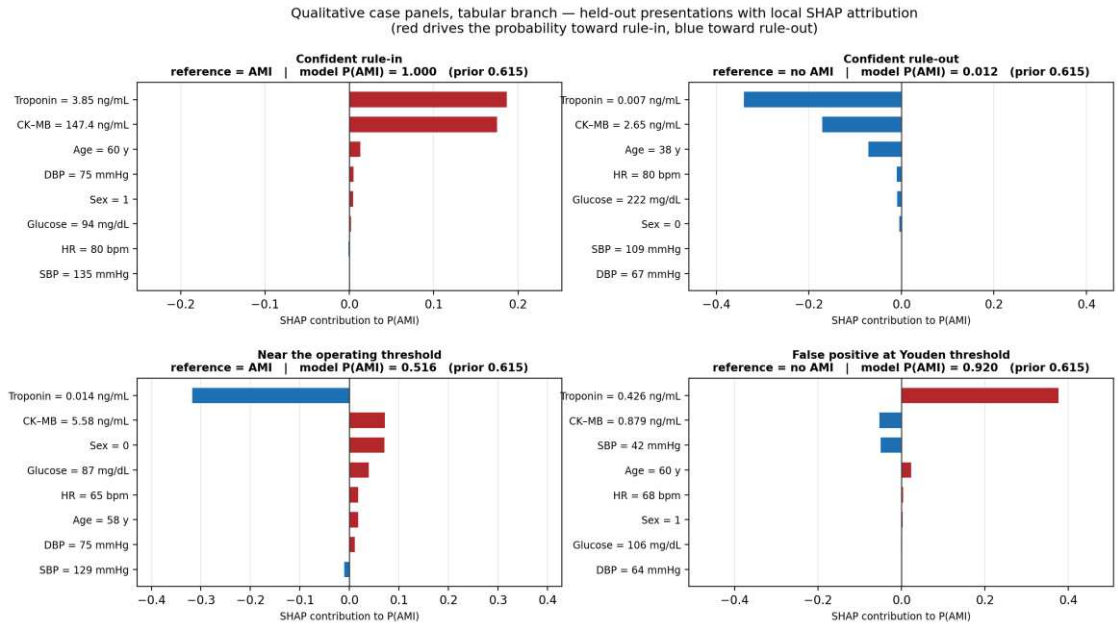


Figure 19. *Qualitative case panels, tabular branch. For each held-out presentation the eight bedside predictors are shown with their measured values and with the local Shapley contribution each makes to the model's probability of acute myocardial infarction; red drives the probability toward rule-in, blue toward rule-out. The panel titles give the cardiologist reference, the predicted probability, and the cohort prior.*

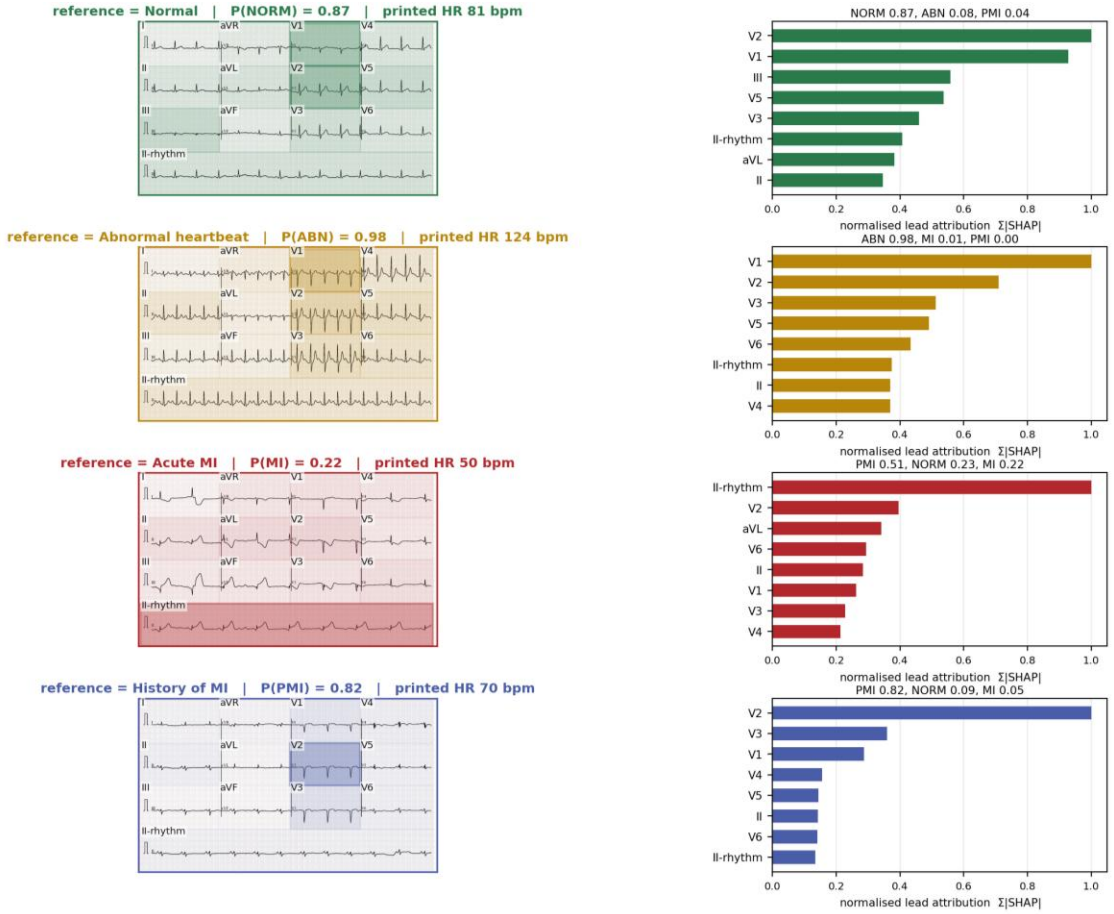


Figure 20. *Qualitative case panels, ECG imaging branch, on the clinically validated four-class corpus under the patient-disjoint protocol of Section 4.8. Each row shows the scanned 12-lead report with the header and footer bands cropped, the class probabilities assigned to it, and a lead-resolved attribution map: the shading of each lead cell is proportional to the aggregated |SHAP| mass of the descriptors recovered from that lead, and the bars at right rank the leads. The third row is a misassigned acute-MI record, retained for the reason given in the text.*

What the panels are for, and what they are not

Two cautions govern the reading of these figures. First, a panel showing a confident, well-localised prediction is an illustration, not evidence; the quantitative claims of this article rest on Sections 6 and 7.1, not on selected cases. Second, and more specifically, the probability printed on a tabular panel is the raw random-forest score, which Section 7.1 showed to be over-confident by a calibration slope of 6.61. A panel reading 0.98 should therefore be understood as a strong ranking signal rather than a 98% risk, and the panels are annotated accordingly. Reporting them without that annotation would reproduce, in visual form, precisely the error this article argues against.

8 CONCLUSION AND FUTURE WORK

8.1 Summary of findings

TRUST-AMI, a leakage-audited, calibration-aware, and explainable framework for the rapid rule-in and rule-out of acute myocardial infarction from routine cardiac biomarkers, has been presented and evaluated on 1,319 emergency presentations with cardiologist-adjudicated labels. Under a protocol that confines preprocessing to training folds, restricts selection to the training partition, and scores the held-out set exactly once, the selected random forest attained a test ROC-AUC of 0.996 (95% CI 0.988–1.000), sensitivity 1.000, specificity 0.961, and agreement with the cardiologist reference of $\kappa = 0.968$. An independently audited ECG imaging branch reached a test ROC-AUC of 0.908 under patient-disjoint folds, with Grad-CAM supplying waveform-level evidence.

The findings that qualify this headline are the substance of the contribution. A single troponin threshold already reaches AUC 0.928, so the incremental value of machine learning is $+0.071$ AUC, realised almost entirely as recovered sensitivity. The same model that discriminates near-perfectly is severely over-confident, with expected calibration error 0.30 and calibration slope 6.61, a defect no AUC-only report would have surfaced. Impurity-based feature importance and ablation disagree about which variables matter, and only the second answers the clinical question. And because troponin partly defines the outcome, biomarker–label circularity limits what the strong performance can be taken to demonstrate. Each of these findings is available only because the study was designed to look for it.
A fifth finding arrived from the second imaging corpus and is the most transferable of them. On a clinically validated four-class ECG image set, the same estimator scored under three partitions differing only in what they allow to be shared reports macro-F_1 0.969, 0.684, and 0.656; on the acute-MI class the figures are 0.998, 0.235, and 0.121.

8.2 Limitations

- The scope is acute MI within coronary heart disease. Other cardiovascular classes—cerebrovascular, peripheral arterial, aortic, structural, congenital, valvular, heart failure, arrhythmias, and venous thromboembolism—require modalities absent from this feature set.
- Subgroup analyses rest on small strata ($n = 76$ for the female subgroup) and screen for gross disparity rather than establish equity.
- After de-duplication the four-class corpus retains 30 distinct acute-MI images from 29 patients. No performance figure on that class is reportable, and the value given in Table 11 should be read as a measurement of the corpus rather than of the method.
- The same corpus is 91.0% male, and 28 of its 30 distinct acute-MI images are male, so it supports no sex-stratified claim.

8.3 Future work

A prospective multicentre study with serial troponin measurements and adjudicated 30-day major adverse cardiac events would address external validity and circularity simultaneously, since MACE is not defined by the admission biomarker. A paired cohort in which the same patients contribute both biomarkers and a 12-lead ECG would permit the fusion analysis that Section 6.4 declines to report. A cardiologist-rated Grad-CAM plausibility study would convert the attribution layer from an implemented feature into validated evidence. And a prospective comparison of recalibrated model scores against an accelerated troponin pathway, evaluated on net benefit at a clinically agreed threshold, would establish whether the sensitivity gain of Section 6.2 survives outside a retrospective cohort.

Acknowledgements

This work was undertaken within the CDSIMER Medical Elective "AI in Healthcare: Applications and Research Directions" at Dayananda Sagar University. The conception, design, data curation, machine-learning implementation, experimentation, analysis, interpretation, manuscript preparation, and deployment of the study were led by Dr. Agughasi Victor Ikechukwu, Senior Assistant Professor, Department of Computer Science & Medical Engineering, School of Engineering, Dayananda Sagar University. The authors gratefully thank Dr. Puneet Nagendra and the CDSIMER cardiology validation panel for their clinical review and guidance.

DECLARATIONS

Funding: No funding was received for conducting this study.

Competing interests: The authors declare no competing interests.

Ethics approval and consent to participate: The study uses only publicly available datasets under their respective licences; no new human-subjects data were collected. We note that the third of these is pseudonymised rather than de-identified: a numeric study identifier is legible in the header band of each

report page (Section 3.5). All processing in this study crops that band before use, and we recommend that redistributions do the same.

Data availability: All datasets analysed in this study are publicly available and no new data were generated. The Zheen heart-attack biomarker cohort is available from Mendeley Data at https://doi.org/10.17632/wmhctcrt5v.1. The PTB-XL 12-lead electrocardiography dataset is available from PhysioNet at https://doi.org/10.13026/kfzx-aw45. The ECG Images dataset of Cardiac Patients, used for the four-class imaging analysis, is available from Mendeley Data at https://doi.org/10.17632/gwbz3fsgp8.2. The derived files that support the findings of this study, including the de-duplication manifest and the extracted waveform descriptors, are included in the code release cited under Code availability.

Code availability: All code and corresponding results are publicly available at https://github.com/Victor-Ikechukwu/CDSIMER-Lastd. The release contains the leakage-safe tabular pipeline, the ECG imaging branch with Grad-CAM, the late-fusion module, the interactive application, fixed random seeds, pinned dependencies, and the scripts that regenerate every table and figure end to end.

Ethical statement on intended use: The system is an educational research prototype. It is not a medical device, has not been prospectively validated, and must not be used to guide clinical care; every interface view states this. Grad-CAM is provided to support clinician oversight and contestability, not to replace clinical judgement.

Author contribution (CRediT): Agughasi Victor Ikechukwu: Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Writing-original draft; Writing-review & editing; Visualization; Supervision; Project administration. **Kirankumar R. Dyawarkonda:** Clinical validation; Investigation; Resources; Methodology (adjudication protocol and reference-label definition); Writing-review & editing. **Chiemeka C. Chinaka:** Software; Validation; Formal analysis; Code verification and clinical insights. **Bikila Abebe Ganati, Diriba Gichile Rundasa, Tariku Birhanu Yadesa, Yogesh, and Rajesh T.M:** Software Validation; Visualization; and Writing-review & editing.

APPENDICES

Appendix A Implementation and hyperparameters

Tabular branch

Median imputation; standardisation (Eq. 1); logistic regression (max_iter $= 1000$, balanced class weights); random forest (400 trees, balanced class weights); gradient boosting (library defaults); SVM with RBF kernel (probability estimates enabled); k-nearest neighbours ($k = 13$); 5-fold stratified cross-validation; Youden threshold; 1,000 bootstrap resamples; random seed 42.

ECG branch

ResNet-18 initialised from ImageNet weights; input 224×224; AdamW (lr $= 10^{-4}$, weight decay 10^{-4}); class-balanced binary cross-entropy (Eq. 6); batch size 32; 20 epochs; model selection on validation AUC; PTB-XL folds 1–8 / 9 / 10, patient-disjoint. Fusion weights (Eq. 9) fitted by logistic regression on the validation partition.

Appendix B Reproducibility checklist

Code, fixed seeds, and pinned dependencies are released. The tabular experiment regenerates every table and figure via a single script. The imaging branch exposes prepare/train/evaluate/gradcam commands. All three datasets are public under their respective licences. The audit of Section 3.5 is released as a standalone script that takes a directory of class subfolders and reports, before any model is fitted, the MD5 duplicate structure, the per-class redundancy, the identifiers recoverable from the page header, and the difference between a file-level and a patient-disjoint split. It is written to be run on corpora other than this one. Discrimination and calibration are reported with 95% intervals in

accordance with TRIPOD+AI [13], and the leakage audit is disclosed in the form of Algorithm 1 so that it can be verified against the released pipeline definitions.

Appendix C Deployed application

The system is publicly deployed as a Streamlit application—cloud: https://cdsimer-ami-chd.streamlit.app/; local: http://localhost:8501/. The four views reproduced below are live captures of the running application, not mock-ups. Figure C1 shows the landing view. Figure C2 shows the per-patient assessment, in which the eight bedside values are entered and the model returns a probability together with the per-feature contributions that produced it—the interactive form of the panels in Fig. 19. Figure C3 shows the ECG imaging branch with its Grad-CAM overlay and the late-fusion control. Figure C4 shows the experimental-results view, from which the threshold analysis of Fig. 8 and the ablation and subgroup tables of Section 6 are reproducible in the browser.

One discrepancy is visible and is left uncorrected rather than silently reconciled: the interface reports a test ROC-AUC of 0.997 where this article reports 0.996. The deployed build and the manuscript run differ in the random seed passed to the train/test partition. The difference is within the bootstrap interval reported in Table 5 and is recorded here so that a reader comparing the two is not left to wonder.

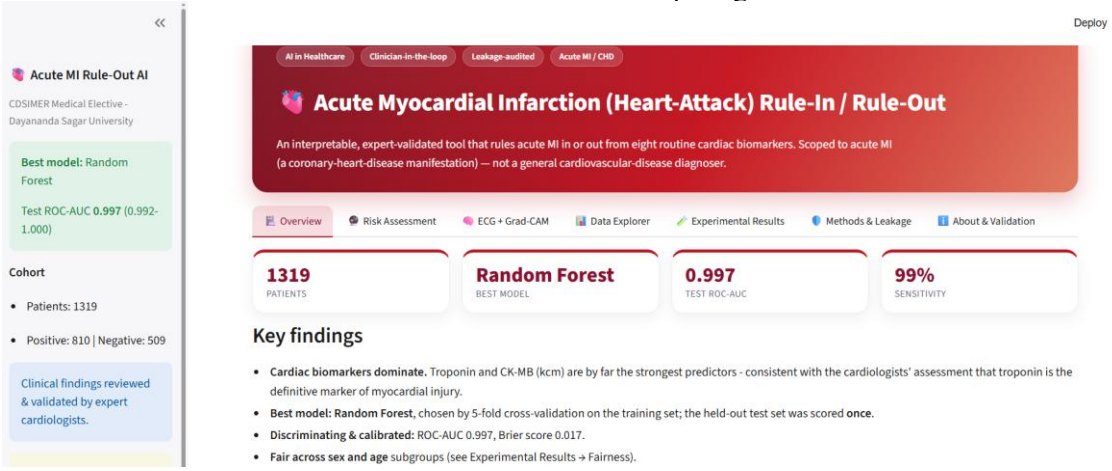


Figure 21. *Deployed application, overview. Cohort summary, selected model, and headline operating characteristics, with the scope restriction to acute MI within coronary heart disease stated in the interface itself.*

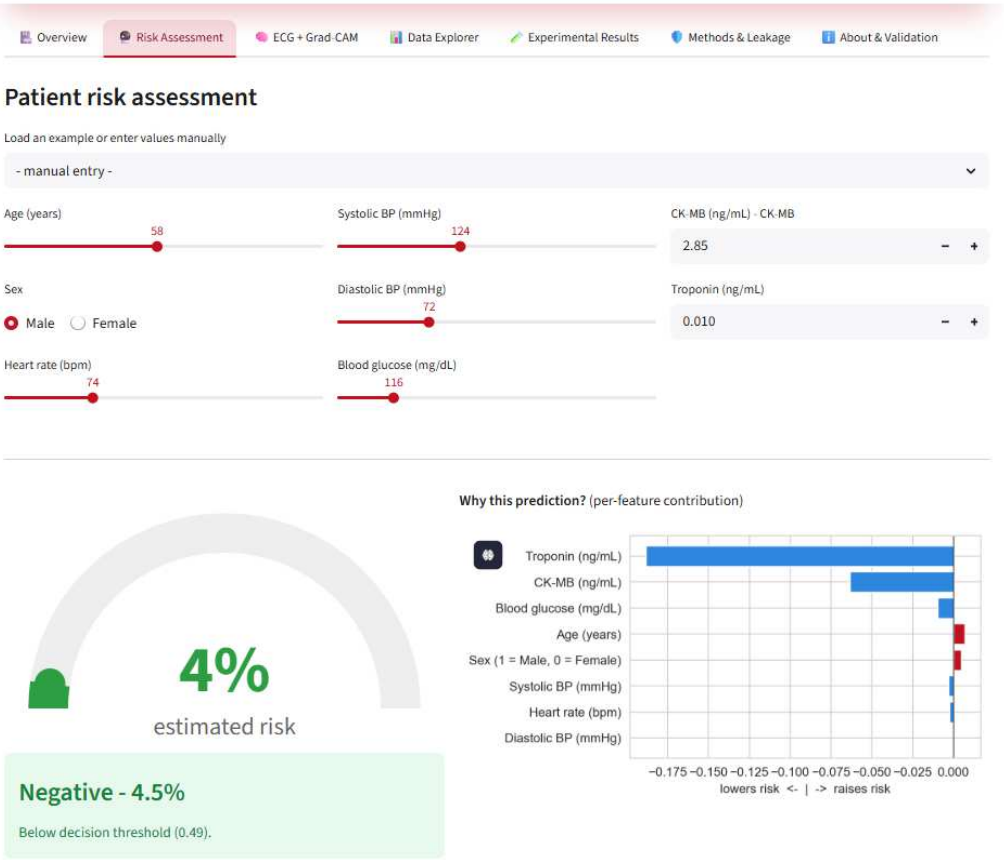


Figure 22. *Deployed application, patient risk assessment. The eight bedside predictors are entered directly; the gauge reports the model probability against the decision threshold, and the bar chart at right gives the per-feature contribution to that probability. The case shown is ruled out at 4.5%, driven by a troponin of 0.010 ng* mL^{-1}.

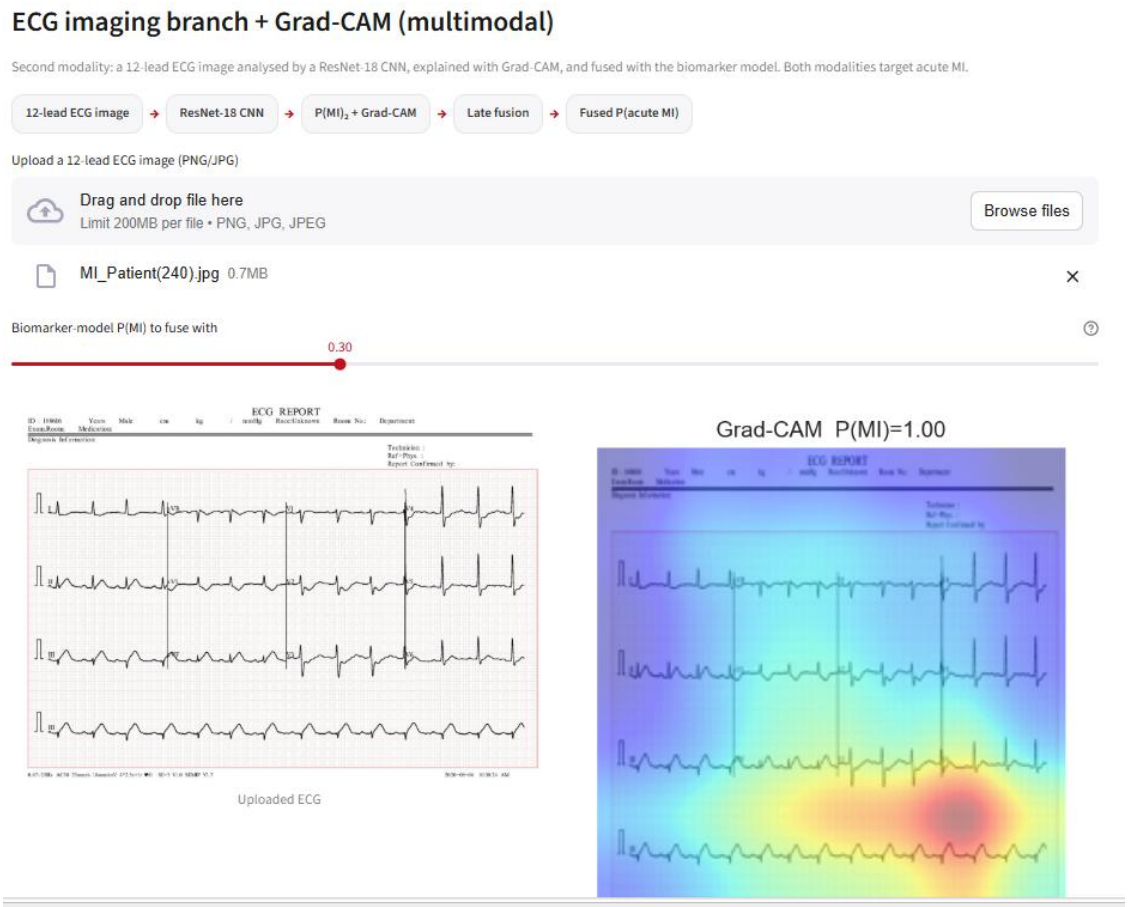


Figure 23. *Deployed application, ECG imaging branch. An uploaded 12-lead report is scored by the ResNet-18 and explained by Grad-CAM (Eqs. 7–8); the slider sets the biomarker probability with which the imaging probability is fused (Eq. 9). The fusion control is exposed for demonstration only: Section 6.4 declines to report a fused metric, because the contributing cohorts share no patients.*

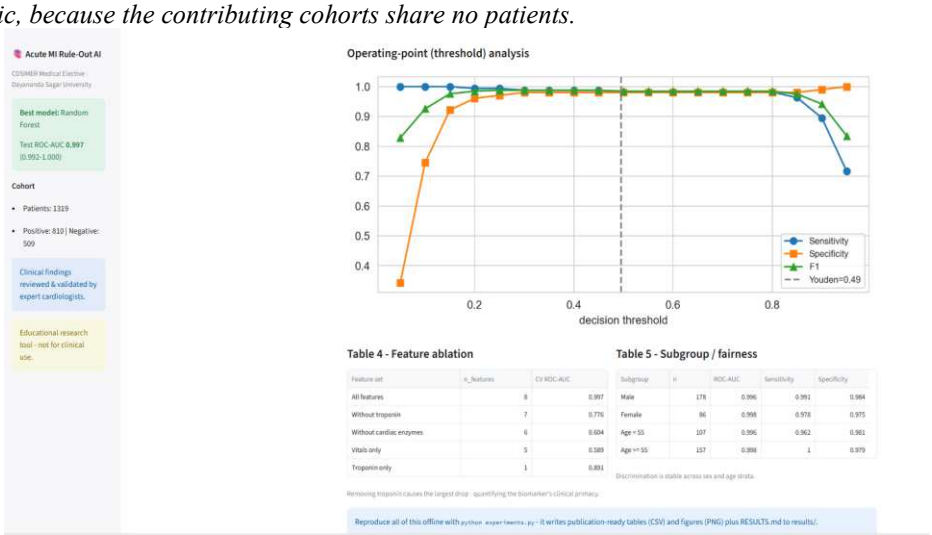


Figure 24. *Deployed application, experimental results. Threshold sweep with the Youden operating point marked, alongside the ablation and subgroup tables of Sections 6.3 and 7.7. Every table and figure in this view is regenerated offline by experiments.py.*

References

[1] Volschow B, Strauss B, Sorensen NA, Thiessen N, Lehmacher J, Toprak B, et al. Diagnostic validation of a novel high-sensitivity cardiac troponin T assay. European Heart Journal: Acute Cardiovascular Care. 2026;15(7):535–544. https://doi.org/10.1093/ehjacc/zuag044.

[2] Koechlin L, Boeddinghaus J, Lopez-Ayala P, Reber C, Nestelberger T, Wildi K, et al. Clinical and analytical performance of a novel point-of-care high-sensitivity cardiac troponin I assay. Journal of the American College of Cardiology. 2024;84(8):726–740. https://doi.org/10.1016/j.jacc.2024.05.056.

[3] Giannitsis E, Frey N, Katus HA. Accelerated high sensitivity troponin diagnostics: ready for an even faster pace? European Heart Journal. 2024;45(29):2516–2518. https://doi.org/10.1093/eurheartj/ehae344.

[4] Sataranatarajan K, Narasimhan M, Chuckaree ID, Balani J, Zhang R, Vigen R, et al. Revisiting high-sensitivity cardiac troponin abnormal baseline cutoffs: implications for AMI diagnosis in the emergency department. Journal of Clinical Medicine. 2025;14(20):7308. https://doi.org/10.3390/jcm14207308.

[5] Kahraman A. Machine learning techniques for improved prediction of cardiovascular diseases using integrated healthcare data. Frontiers in Artificial Intelligence. 2025;8:1694450. https://doi.org/10.3389/frai.2025.1694450.

[6] Paul VV, Syed Masood JAI. An explainable machine learning framework for cardiovascular risk prediction using structured health data. Frontiers in Artificial Intelligence. 2026;9:1812622. https://doi.org/10.3389/frai.2026.1812622.

[7] Hilgendorf L, Petursson P, Andersson E, Rawshani A, Bhatt DL, Ramunddal T, et al. Fully automated diagnosis of acute myocardial infarction using electrocardiograms and multimodal deep learning. JACC: Advances. 2025;4(8):102011. https://doi.org/10.1016/j.jacadv.2025.102011.

[8] Rosenblatt M, Tejavibulya L, Jiang R, Noble S, Scheinost D. Data leakage inflates prediction performance in connectome-based machine learning models. Nature Communications. 2024;15:1829. https://doi.org/10.1038/s41467-024-46150-w.

[9] van de Mortel LA, van Wingen GA. Data leakage in machine learning studies creep into meta-analytic estimates of predictive performance. Molecular Psychiatry. 2025;30(12):6070–6071. https://doi.org/10.1038/s41380-025-03336-y.

[10] Bernett J, Blumenthal DB, Grimm DG, Haselbeck F, Joeres R, Kalinina OV, et al. Guiding questions to avoid data leakage in biological machine learning applications. Nature Methods. 2024;21(8):1444–1453. https://doi.org/10.1038/s41592-024-02362-y.

[11] Apicella A, Isgro F, Prevete R. Don't push the button! Exploring data leakage risks in machine learning and transfer learning. Artificial Intelligence Review. 2025;58:305. https://doi.org/10.1007/s10462-025-11326-3.

[12] Starcke J, Spadafora J, Spadafora J, Spadafora P, Toma M. The effect of data leakage and feature selection on machine learning performance for early Parkinson's disease detection. Bioengineering. 2025;12(8):845. https://doi.org/10.3390/bioengineering12080845.

[13] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378.

[14] Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. https://doi.org/10.1136/bmj-2024-082505.

[15] Colacci M, Huang YQ, Postill G, Zhelnov P, Fennelly O, Verma A, et al. Sociodemographic bias in clinical machine learning models: a scoping review of algorithmic bias instances and mechanisms. Journal of Clinical Epidemiology. 2025;178:111606. https://doi.org/10.1016/j.jclinepi.2024.111606.

[16] Abbas Q, Jeong W, Lee SW. Explainable AI in clinical decision support systems: a meta-analysis of methods, applications, and usability challenges. Healthcare. 2025;13(17):2154. https://doi.org/10.3390/healthcare13172154.

[17] Salimparsa M, Sedig K, Lizotte DJ, Abdullah SS, Chalabianloo N, Muanda FT. Explainable AI for clinical decision support systems: literature review, key gaps, and research synthesis. Informatics. 2025;12(4):119. https://doi.org/10.3390/informatics12040119.

[18] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision. 2020;128(2):336–359. https://doi.org/10.1007/s11263-019-01228-7.

[19] Jandoubi B, Akhloufi MA. Multimodal artificial intelligence in medical diagnostics. Information. 2025;16(7):591. https://doi.org/10.3390/info16070591.

[20] Manimaran G, Peimankar A, Puthusserypady S, Momeni M, Asyari RAI, Jahan MS, et al. Explainable deep learning based techniques for ECG-based heart disease classification: a systematic literature review and future direction. Computers in Biology and Medicine. 2025;199:111324. https://doi.org/10.1016/j.compbiomed.2025.111324.

[21] Luo J, Taleban A, Noffke P, Sparapani R. Multi-level explainable AI for ECG-based atrial fibrillation detection: exploring LIME, SHAP, and Grad-CAM for clinical interpretation. Proceedings of the ACM Conference. 2025;p. 52–59. https://doi.org/10.1145/3761712.3761721.
[22] Victor Ikechukwu A, S M. CX-Net: an efficient ensemble semantic deep neural network for ROI identification from chest-x-ray images for COPD diagnosis. Machine Learning: Science and Technology. 2023;4(2):025021. https://doi.org/10.1088/2632-2153/acd2a5.
[23] Ikechukwu AV, Bhimshetty S. Optimizing ResNet50 and VGG19 networks for accurate COPD diagnosis through CXR analysis. Lecture Notes in Networks and Systems. 2025;p. 307–320. https://doi.org/10.1007/978-981-96-0165-3_23.
[24] Ikechukwu AV, Murali S. xAI: an explainable AI model for the diagnosis of COPD from CXR images. 2023 IEEE International Conference on Data Science and Systems (ICDDS). 2023;p. 1–6. https://doi.org/10.1109/ICDDS59137.2023.10434619.
[25] Victor Ikechukwu A, Murali S, Deepu R, Shivamurthy RC, et al. ResNet-50 vs VGG-19 vs training from scratch: a comparative analysis of the segmentation and classification of pneumonia from chest X-ray images. Global Transitions Proceedings. 2021;2(2):375–381. https://doi.org/10.1016/j.gltp.2021.08.027.
[26] Agughasi VI. Leveraging transfer learning for efficient diagnosis of COPD using CXR images and explainable AI techniques. Inteligencia Artificial. 2024;27(74):133–151. https://doi.org/10.4114/intartif.vol27iss74pp133-151.
[27] Dhanka S, Maini S. A hybridization of XGBoost machine learning model by Optuna hyperparameter tuning suite for cardiovascular disease classification with significant effect of outliers and heterogeneous training datasets. International Journal of Cardiology. 2025;420:132757. https://doi.org/10.1016/j.ijcard.2024.132757.
[28] Cao K, Liu C, Yang S, Zhang Y, Li L, Jung H, et al. Prediction of cardiovascular disease based on multiple feature selection and improved PSO-XGBoost model. Scientific Reports. 2025;15. https://doi.org/10.1038/s41598-025-96520-7.
[29] Naik A, Tejani GG, Mousavirad SJ. SGO enhanced random forest and extreme gradient boosting framework for heart disease prediction. Scientific Reports. 2025;15. https://doi.org/10.1038/s41598-025-02525-7.
[30] Akinola S, Leelakrishna R, Varadarajan V. Enhancing cardiovascular disease prediction: a hybrid machine learning approach integrating oversampling and adaptive boosting techniques. AIMS Medical Science. 2024;11(1):58–71. https://doi.org/10.3934/medsci.2024005.
[31] Zhang Z, Li D, Zhao J, Ma H, Wang F, Hao Q. Enhanced prediction of cardiovascular disease through integrated machine learning models combining clinical and demographic characteristics. Diagnostics. 2026;16(10):1572. https://doi.org/10.3390/diagnostics16101572.
[32] Basar R, Ocak O, Erturk A, de la Roche M. Leveraging machine learning techniques to predict cardiovascular heart disease. Information. 2025;16(8):639. https://doi.org/10.3390/info16080639.
[33] Hollmann N, Muller S, Purucker L, Krishnakumar A, Korfer M, Hoo SB, et al. Accurate predictions on small data with a tabular foundation model. Nature. 2025;637(8045):319–326. https://doi.org/10.1038/s41586-024-08328-6.
[34] Abdelhay O, Shatnawi A, Najadat H, Altamimi T. Resampling methods for class imbalance in clinical prediction models: a scoping review protocol. PLOS ONE. 2025;20(11):e0330050. https://doi.org/10.1371/journal.pone.0330050.
[35] Gamble CU, Faghani S, Erickson BJ. Applying conformal prediction to a deep learning model for intracranial hemorrhage detection to improve trustworthiness. Radiology: Artificial Intelligence. 2024;7(2):e240032. https://doi.org/10.1148/ryai.240032.
[36] Asgari E, Montana-Brown N, Dubois M, Khalil S, Balloch J, Au Yeung J, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digital Medicine. 2025;8. https://doi.org/10.1038/s41746-025-01670-7.
[37] Papin M, Markarian T, Le Bastard Q, Volteau C, Pes P, Le Conte P, et al. Inter-rater reliability of point-of-care ultrasound during out-of-hospital cardiac arrest: an ancillary analysis of the observational prospective ACE trial. Resuscitation Plus. 2025;27:101191. https://doi.org/10.1016/j.resplu.2025.101191.
[38] Kumar R, Garg S, Kaur R, Johar MGM, Singh S, Menon SV, et al. A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions. Frontiers in Artificial Intelligence. 2025;8:1583459. https://doi.org/10.3389/frai.2025.1583459.
[39] Nair M, Lundgren LE, Soliman A, Dryselius P, Fogelberg E, Petersson M, et al. Machine learning model for readmission prediction of patients with heart failure based on electronic health records: protocol for a quasi-experimental study for impact assessment. JMIR Research Protocols. 2024;13:e52744. https://doi.org/10.2196/52744.
[40] Ganesan R, Habraken SC, van de Vosse FN, Huberts W. Explainable machine learning based prediction of severity of heart failure using primary electronic health records. Studies in Health Technology and Informatics. 2024;316:1015–1019. https://doi.org/10.3233/SHTI240471.

[41] Yu CS, Wu J, Shih CM, Chiu KL, Chen YD, Chang TH. Exploring mortality and prognostic factors of heart failure with in-hospital and emergency patients by electronic medical records: a machine learning approach. Risk Management and Healthcare Policy. 2025;18:77–93. https://doi.org/10.2147/RMHP.S488159.
[42] Saha S, Ross H, Velmovitsky PE, Wang CX, Vishram-Nielsen JKK, Manlhiot C, et al. Machine learning enhanced expert system for detecting heart failure decompensation using patient reported vitals and electronic health records. Scientific Reports. 2025;15. https://doi.org/10.1038/s41598-025-16376-9.
[43] Hajishah H, Kazemi D, Safaee E, Amini MJ, Peisepar M, Tanhapour MM, et al. Evaluation of machine learning methods for prediction of heart failure mortality and readmission: meta-analysis. BMC Cardiovascular Disorders. 2025;25. https://doi.org/10.1186/s12872-025-04700-0.
[44] Chung WY, Yoon J, Yoon D, Kim S, Kim Y, Park JE, et al. Development and validation of deep learning-based infectivity prediction in pulmonary tuberculosis through chest radiography: retrospective study. Journal of Medical Internet Research. 2024;26:e58413. https://doi.org/10.2196/58413.
[45] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a large publicly available electrocardiography dataset. Scientific Data. 2020;7(1):154. https://doi.org/10.1038/s41597-020-0495-6.
[46] Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep learning for ECG analysis: benchmarks and insights from PTB-XL. IEEE Journal of Biomedical and Health Informatics. 2021;25(5):1519–1528. https://doi.org/10.1109/JBHI.2020.3022989.
[47] Oikonomou EK, Sangha V, Dhingra LS, Aminorroaya A, Coppi A, Krumholz HM, et al. Artificial intelligence-enhanced risk stratification of cancer therapeutics-related cardiac dysfunction using electrocardiographic images. Circulation: Cardiovascular Quality and Outcomes. 2025;18(1). https://doi.org/10.1161/CIRCOUTCOMES.124.011504.
[48] Song M, Kim J, Jang EC, Kwon S. Interpretable machine learning for emergency department triage: clinical insights from 133,198 patients using the Korean Triage and Acuity Scale (KTAS). Diagnostics. 2026;16(6):954. https://doi.org/10.3390/diagnostics16060954.
[49] Mohamadi Z, Shafizadeh A, Aliyan Y, Shayesteh SF, Goudarzi P, Khodabandeh A, et al. The application of random forest-based models in prognostication of gastrointestinal tract malignancies: a systematic review. Frontiers in Artificial Intelligence. 2025;8:1517670. https://doi.org/10.3389/frai.2025.1517670.
[50] Wang H, Song Y, Bi H. Optimizing public health management with predictive analytics: leveraging the power of random forest. Frontiers in Big Data. 2025;8:1574683. https://doi.org/10.3389/fdata.2025.1574683.
[51] Djiberou Mahamadou AJ, Rodrigues EA, Vakorin V, Antoine V, Moreno S. Interpretable machine learning for precision cognitive aging. Frontiers in Computational Neuroscience. 2025;19:1560064. https://doi.org/10.3389/fncom.2025.1560064.
[52] Attai KF, Amannah C, Ekpenyong M, Asuquo DE, Akputu OK, Obot OU, et al. Developing an explainable artificial intelligence system for the mobile-based diagnosis of febrile diseases using random forest, LIME, and GPT. Healthcare Informatics Research. 2025;31(2):125–135. https://doi.org/10.4258/hir.2025.31.2.125.
[53] Khan AH, Hussain M. ECG Images dataset of Cardiac Patients. Mendeley Data, V2. 2021;https://doi.org/10.17632/gwbz3fsgp8.2.
[54] Khan AH, Hussain M, Malik MK. ECG images dataset of cardiac and COVID-19 patients. Data in Brief. 2021;34:106762. https://doi.org/10.1016/j.dib.2021.106762.