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.

Explainable machine learning for leakage-aware risk stratification of electrocardiograph verification records

Adin Neslanović

adin.neslanovic@ffsa.unsa.ba

Cantonal Hospital Zenica

Irhad Neslanović

University of Zenica

Berina Hasanefendić

Clinical Center University of Sarajevo

Ermin Begović

Clinical Center University of Sarajevo

Almir Badnjević

Agency for Identification Documents, Registers and Data Exchange of Bosnia and Herzegovina

Research Article

Keywords: electrocardiograph, verification, medical device, machine learning, LightGBM, label leakage, calibration, explainable artificial intelligence

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10844417/v1

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

Additional Declarations: No competing interests reported.

Explainable machine learning for leakage-aware risk stratification of electrocardiograph verification records

Adin Neslanović[1], Irhad Neslanović[2], Berina Hasanefendić[3,4], Ermin Begović[3], Almir Badnjević[5]

[1]*Department of Medical Biochemistry and Immunological Diagnostics, Cantonal Hospital Zenica, Zenica, Bosnia and Herzegovina,*

[2]*Faculty of Polytechnic Engineering, Department of Software Engineering, University of Zenica, Zenica, Bosnia and Herzegovina,*

[3]*Clinical Biochemistry and Laboratory Medicine, Clinical Center, University of Sarajevo, Sarajevo, Bosnia and Herzegovina,*

[4]*Department for Laboratory Technologies, Faculty of Health Studies, University of Sarajevo, Sarajevo, Bosnia and Herzegovina,*

[5]*Agency for Identification Documents, Registers and Data Exchange of Bosnia and Herzegovina, Banja Luka, Bosnia and Herzegovina,*

Corresponding author: Adin Neslanović (ORCID: 0009-0000-4560-0964)
adin.neslanovic@ffsa.unsa.ba

Abstract: Metrological verification of electrocardiographs (ECGs) performance and safety generates structured technical and measurement data that may support the identification of devices requiring additional expert review. Machine-learning models may assist verification workflows by devices at elevated risk of failure. This study investigates the predictive value and limitations of routinely collected metrological verification data for identifying electrocardiographs at increased risk of verification failure. A retrospective analysis of 3,956 electrocardiograph verification records, including 3,870 passing and 86 failing records, was conducted. Technical and measurement variables available before the final verification decision were selected, while variables associated with potential data leakage were excluded. Three machine-learning algorithms, LightGBM, Balanced Random Forest, and SMOTE Random Forest, were compared using repeated stratified five-fold cross-validation, with PR-AUC as the primary model selection criterion. The selected model was further evaluated using out-of-fold predictions, bootstrap confidence intervals, calibration analysis, feature importance, and group-based validation to assess performance on previously unseen manufacturers and device models. Among the evaluated algorithms, LightGBM achieved the highest precision-recall performance, demonstrating substantial improvement over the baseline failure prevalence and revealing meaningful predictive information in routinely collected metrological data. Explainability analyses identified measurement-related characteristics and device-specific information as important predictors, while ablation analysis demonstrated that predictive discrimination was retained even after removing device identity. The study further demonstrated the value of combining leakage-controlled modelling, explainability, and robustness assessment to critically evaluate machine-learning performance in real-world medical device verification. Nevertheless, limited precision, poor probability calibration, and reduced performance on unseen manufacturers and device models highlighted important limitations, while decision curve analysis provided insufficient evidence of consistent operational benefit. This study demonstrates the potential of routinely collected metrological verification data to support AI-driven risk stratification, contributing to the broader transition toward data-driven medical device management advocated by the World Health Organization. While the findings establish a basis for further development of intelligent quality assurance systems, limitations in precision, calibration, and generalizability highlight the need for independent validation and demonstrated operational benefit before implementation in routine verification practice.

Keywords: *electrocardiograph; verification; medical device; machine learning; LightGBM; label leakage; calibration; explainable artificial intelligence*

1 Introduction

Medical devices are an important part of modern healthcare, and their safe and reliable operation depends on appropriate inventory management, maintenance, inspection, testing, and verification. The World Health Organization has emphasised the importance of structured inventory and maintenance management of medical devices as part of healthcare technology management [1]. This is especially relevant for devices that measure physiological parameters, because inaccurate measurements may affect clinical interpretation and further medical decisions. Electrocardiographs are among these devices, and their performance can therefore be regularly assessed according to defined technical and metrological requirements [2,3]. Verification and calibration procedures are therefore important components of quality assurance for not only ECG devices, but generally for electronic medical devices as the device that does not meet required performance characteristics may introduce errors into recorded signals and consequently affect diagnostic confidence.

Medical device inspection and verification processes generate a large amount of structured technical data. [1] These data include information about the device, manufacturer, verification procedure, assigned measurement values, measured values, and measurement errors, preventive and corrrective maintenance, calibration results, etc. Previous work has shown that digital solutions can improve the organisation and traceability of medical device inspection processes [4]. Standardised inspection and conformity-assessment data can also create a basis for further analysis of medical device performance and post-market behaviour.

The use of artificial intelligence and machine learning in medical device management has increased during recent years. Several studies have investigated whether historical inspection and conformity-assessment data can be used for prediction of medical device performance. Machine-learning approaches have been applied to defibrillators and infant incubators, as well as to infusion pumps, patient monitors, mechanical ventilators, dialysis machines, and other medical devices. Hundur et al. developed machine-learning models for predicting the operational status of ECG devices using performance and safety parameters collected during accredited inspection procedures [2]. Badnjević et al. proposed a conformity-assessment approach for electrocardiographs to support post-market surveillance and the digital transformation of medical-device management [3]. These studies demonstrate the relevance of structured verification data for medical-device quality assurance, but they also emphasize the need for rigorous validation and transparent feature selection when machine learning is applied to real-world inspection records. These studies also indicate that technical and inspection data may contain information that can support maintenance planning, performance assessment, and post-market surveillance. In contrast to conventional diagnostic machine learning, where the primary input is often a physiological signal or image, verification-oriented modeling uses technical measurements, device metadata, calibration deviations, and quality-control parameters to estimate the probability of equipment nonconformity. This approach is also relevant for electrocardiographs. Previous research proposed a standardised conformity-assessment method for electrocardiographs for post-market surveillance purposes [3].

The use of machine learning in this context also creates challenges that can be addressed in order to improve the explainability and transparency of proposed models. Explainability is especially important in this setting because maintenance and quality-control personnel must understand why a record is prioritized and must retain responsibility for final technical decisions [5]. The application of artificial intelligence in safety-critical medical-device contexts also requires careful attention to data quality, traceability, validation, and governance. AI-based outputs can not replace statutory verification procedures; rather, they can provide transparent decision support within an established quality-management process [6]. In this respect, one of the challenges is the appropriate selection of predictors in the databases of verification results. Medical device verification databases may contain technical measurements together with administrative information generated at different stages of the verification process. Some variables are available before the final verification decision, while others are generated during or after that decision. If variables that directly or indirectly reflect the final outcome are included in model development, the reported predictive performance may be overestimated and the interpretation of the model may become misleading. This problem is commonly described as label leakage. [7–9] In

retrospective datasets, where information from the complete verification workflow is available in the same database, the timing, availability, and role of each predictor can therefore be clearly defined [7]. Transparency also requires a clear definition of the point at which prediction is intended to take place. Another challenge is the strong imbalance between passing and failing verification records in the dataset for model training. In that case, a model may show good discrimination while still producing a considerable number of false-positive predictions. For this reason, model evaluation can include complementary measures of discrimination, precision, recall, calibration, and classification error. The stability of model performance can also be examined when the model is applied to previously unseen manufacturers or device models. Such analyses provide additional information on how the model behaves, where its limitations are, and how its outputs can be interpreted.

The aim of this study was to develop and internally evaluate an explainable machine-learning model for risk stratification of electrocardiograph verification records using routinely collected technical and metrological data. Several classification approaches were compared under a leakage-controlled design. The selected model was evaluated in terms of discrimination, calibration, threshold behaviour, feature importance, and robustness when previously unseen manufacturers and device models were considered. Particular attention was given to the transparency of predictor selection, the interpretation of model outputs, and the identification of limitations that may affect model generalisability. The study therefore examined whether routinely collected verification data can provide a transparent and interpretable basis for identifying records with increased risk of verification failure. Rather than introducing a new leakage-prevention method, this study examines the reliability and limitations of machine learning applied to routinely collected ECG verification records. Its contribution lies in evaluating whether these data provide predictive information that remains meaningful after controlling for outcome-related variables and assessing discrimination, calibration, explainability, and robustness.

2 Results

The distribution of verification outcomes for devices subject to study is presented in Table 1. The study design, dana explination and algorithm selection is explained in details in section 5 Materials and Methods.

Table 1. Dataset composition and outcome distribution.

Outcome	**Number of records**
PASS - prevalence	3.870 (97.83%)
FAIL - prevalence	86 (2.17%)
Total	**3.956**

2.2 Repeated cross-validation model comparison

LightGBM achieved the highest mean PR-AUC and was selected as the final model. LightGBM achieved a mean PR-AUC of 0.297, compared with 0.269 for Balanced Random Forest and 0.257 for SMOTE Random Forest.

The mean ROC-AUC was similar for LightGBM and Balanced Random Forest, at 0.840 and 0.841, respectively. However, LightGBM achieved substantially higher mean recall for failing records, at 0.716, compared with 0.476 for Balanced Random Forest and 0.456 for SMOTE Random Forest.

Balanced Random Forest and SMOTE Random Forest achieved greater mean specificity and MCC values under the fixed 0.50 threshold. Nevertheless, PR-AUC was chosen as the primary selection criterion because the intended purpose was to prioritize rare failing records.

Although Balanced Random Forest showed slightly greater mean ROC-AUC and MCC, LightGBM was retained as the final model because it achieved the highest mean PR-AUC and higher sensitivity for failing records. In a failure-detection context with only 86 positive records, PR-AUC provides the more relevant ranking-based criterion. Comparative repeated cross-validation performance of the candidate classifiers is presented in Table 2.

Table 2. Repeated stratified 5-fold cross-validation comparison of candidate classifiers.

Model	PR-AUC, mean	PR-AUC, SD	ROC-AUC, mean	ROC-AUC, SD	Recall FAIL, mean	Recall FAIL, SD	Specificity, mean	Specificity, SD	MCC, mean	MCC, SD	Brier score, mean	Brier score, SD
LightGBM	0.297	0.092	0.840	0.047	0.716	0.136	0.783	0.037	0.174	0.036	0.123	0.008
Balanced Random Forest	0.269	0.110	0.841	0.050	0.476	0.113	0.937	0.012	0.233	0.053	0.157	0.006
SMOTE Random Forest	0.257	0.092	0.814	0.060	0.456	0.093	0.944	0.018	0.241	0.054	0.051	0.006

2.3 Out-of-fold performance of the final model

The selected LightGBM model was evaluated using out-of-fold predictions, achieving a PR-AUC of 0.271 and ROC-AUC of 0.811. At the prespecified classification threshold of 0.50, the model correctly identified 57 of 86 failing verification records (recall: 0.663), while generating 791 false-positive alerts, resulting in a precision of 0.067. These findings indicate a substantial difference between the model's ability to discriminate verification outcomes and its precision at the selected threshold. At this threshold, 848 of 3,956 verification records (21.4%) were classified as high risk and would require additional review under the proposed workflow. Of these alerts, 791 (93.3%) corresponded to passing records, illustrating the substantial false-positive burden associated with the model.

The complete performance metrics are summarised in Table 3. Figure 1 presents the receiver operating characteristic and precision–recall curves, while Figure 2 illustrates the corresponding confusion matrix.

Table 3. Out-of-fold performance of the selected LightGBM model.

Metric	Value
PR-AUC	0.271
ROC-AUC	0.811
Recall for FAIL	0.663

Specificity	0.796
Precision for FAIL	0.067
Balanced accuracy	0.729
Matthews correlation coefficient	0.163
F2 score	0.239
Brier score	0.122
Expected calibration error, 10 bins	0.200
True positives	57
False positives	791
True negatives	3,079
False negatives	29

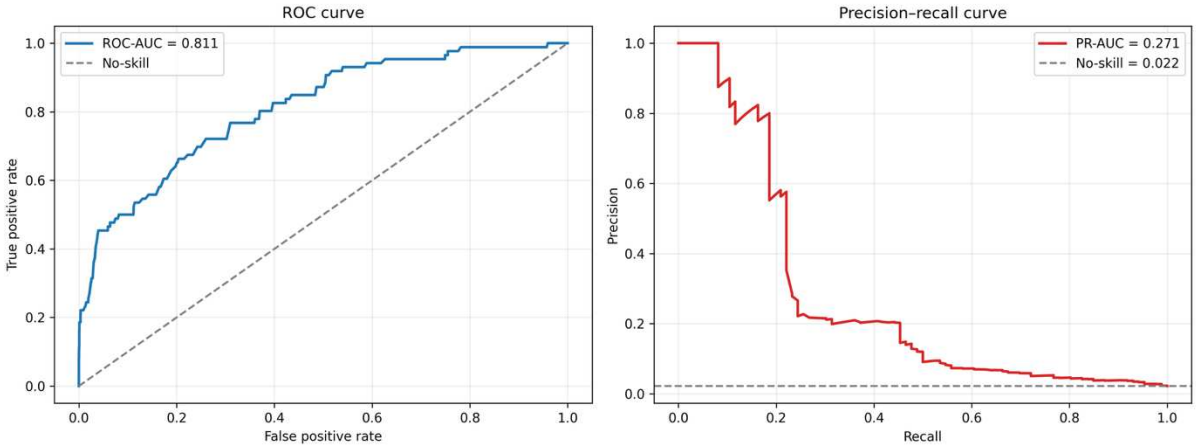


Figure 1. Receiver operating characteristic and precision-recall curves for the final LightGBM C-strict model. *The out-of-fold ROC-AUC was 0.811 and PR-AUC was 0.271. The horizontal reference line in the precision-recall plot represents the baseline failure prevalence of approximately 0.022.*

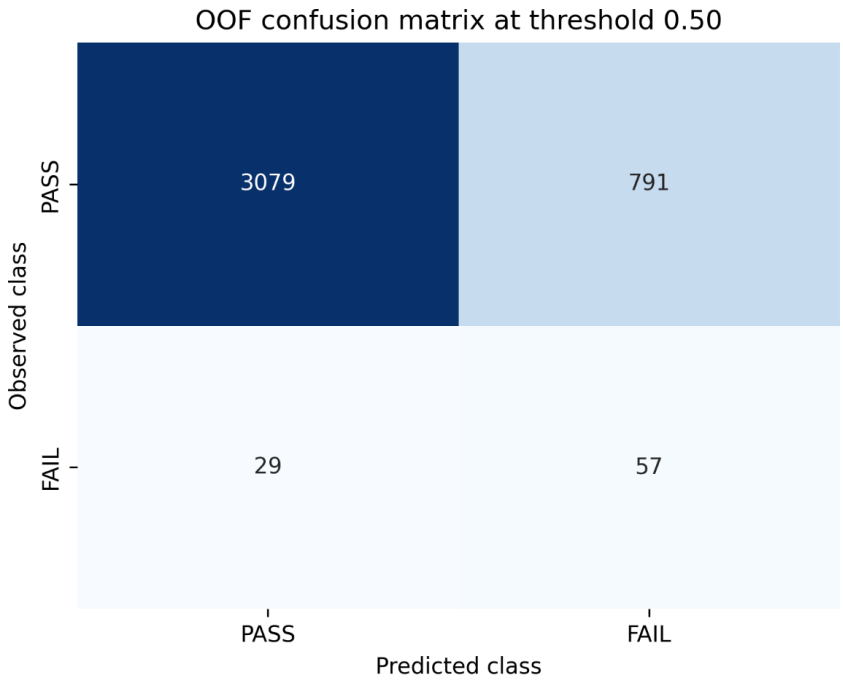


Figure 2. Out-of-fold confusion matrix for the final LightGBM model at a probability threshold of 0.50. *The model correctly identified 57 failing records and 3,079 passing records, while producing 29 false negatives and 791 false positives.*

2.4 Bootstrap confidence intervals

Bootstrap analysis of the out-of-fold predictions showed varying levels of uncertainty across performance metrics. The 95% confidence intervals ranged from 0.179 to 0.364 for PR-AUC, 0.763 to 0.854 for ROC-AUC, and 0.561 to 0.758 for recall of failing records. The complete bootstrap confidence intervals are presented in Table 4, while the corresponding distributions of selected performance metrics are provided in Supplementary Figure S1 and Supplementary Figure 1.

Table 4. Bootstrap 95% confidence intervals for out-of-fold performance metrics.

Metric	Point estimate	95% CI, lower bound	95% CI, upper bound
PR-AUC	0.271	0.179	0.364
ROC-AUC	0.811	0.763	0.854
Recall for FAIL	0.663	0.561	0.758
Specificity	0.796	0.783	0.808
Precision for FAIL	0.067	0.051	0.084
Matthews correlation coefficient	0.163	0.123	0.200
F2 score	0.239	0.189	0.286
Brier score	0.122	0.115	0.128

The bootstrap distributions demonstrate that the model's ability to identify failing records remained above random performance across resampled datasets. However, the width of the PR-AUC and recall distributions also shows the uncertainty associated with the limited number of failures.

2.5 Calibration analysis

Calibration analysis revealed substantial discrepancies between predicted probabilities and observed verification-failure rates, with a Brier score of 0.122 and an expected calibration error of 0.200. The reliability diagram (Fig. 3) showed systematic overestimation of failure probabilities across the evaluated calibration bins. Detailed calibration results are provided in Supplementary Table S1.

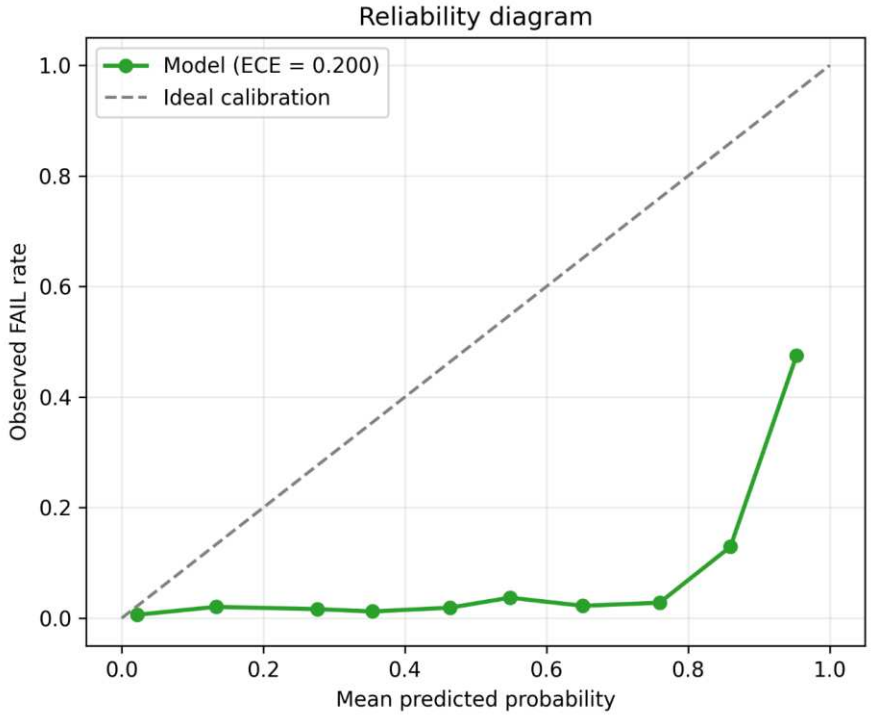


Figure 3. Reliability diagram for the final LightGBM C-strict model. *The expected calibration error was 0.200. Deviations between the model curve and the ideal calibration line show that predicted probabilities were not perfectly calibrated.*

2.6 Calibration slope and exploratory threshold optimization

Exploratory threshold analysis revealed substantial trade-offs between failure detection and false-positive alert burden (Table 5). Compared with the prespecified threshold of 0.50, the F2-optimal threshold of 0.79 reduced false-positive alerts from 791 to 154, but also decreased the number of correctly identified failing records from 57 to 39. The Youden-optimal threshold increased failure detection but generated more false-positive alerts, whereas the cost-sensitive threshold of 0.92 substantially reduced false positives at the expense of recall. All alternative thresholds were derived from the same out-of-fold predictions and were therefore considered exploratory.

Table 5. Threshold sensitivity analysis for the final LightGBM C-strict model.

Selection rule	Threshold	Recall FAIL	Specificity	Precision FAIL	Balanced accuracy	MCC	F2 score	TP	FP	TN	FN
Prespecified reporting threshold	0.50	0.663	0.796	0.067	0.729	0.163	0.239	57	791	3,079	29
Exploratory F2-optimal threshold	0.79	0.454	0.960	0.202	0.707	0.280	0.363	39	154	3,716	47

Selection rule	Threshold	Recall FAIL	Specificity	Precision FAIL	Balanced accuracy	MCC	F2 score	TP	FP	TN	FN
Exploratory Youden-optimal threshold	0.36	0.721	0.740	0.058	0.731	0.151	0.220	62	1,006	2,864	24
Exploratory cost-sensitive threshold, FN:FP = 5:1	0.92	0.209	0.996	0.563	0.603	0.335	0.239	18	14	3,856	68

Note: Thresholds other than 0.50 were derived exploratorily from the same out-of-fold predictions and require confirmation in an independent validation dataset. FN false negative; FP false positive; TN true negative; TP true positive.

2.7 Decision curve analysis

Decision curve analysis showed that the LightGBM model achieved greater net benefit than the review-all strategy across the evaluated threshold range (Fig. 4). Compared with the review-none strategy, the model demonstrated positive net benefit only at low threshold probabilities (0.01–0.03), while its net benefit was negative at higher thresholds. These findings indicate that the potential decision benefit was restricted to a narrow threshold range and requires further evaluation in relation to clinically or operationally relevant review decisions.

In this analysis, the review-all strategy represents additional expert review of every verification record, whereas the review-none strategy represents routine verification without any additional model-triggered review. Neither strategy implies omission of mandatory metrological verification procedures.

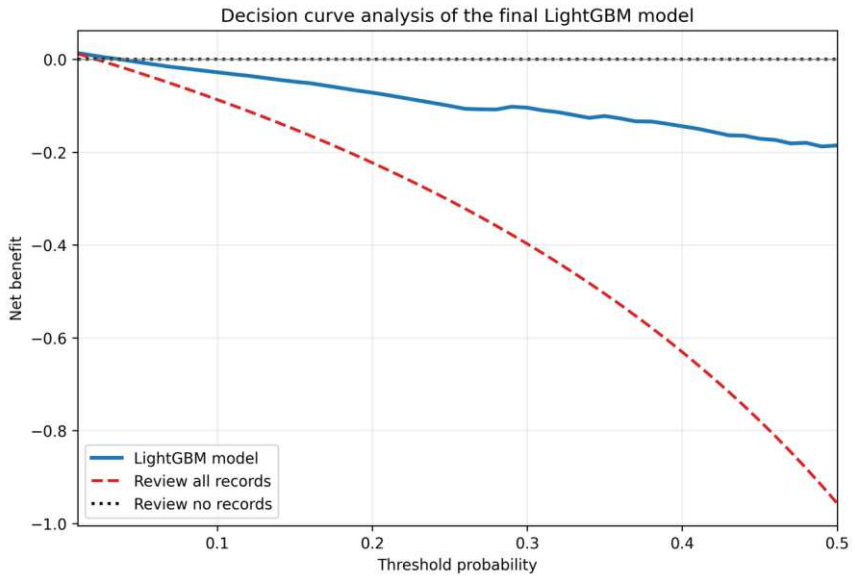


Figure 4. Decision curve analysis for the final LightGBM C-strict model. *Net benefit is shown across threshold probabilities from 0.01 to 0.50. The model was compared with strategies of reviewing all records and reviewing no records. The model provided greater net benefit than indiscriminate review of all records but did not exceed the review-none strategy, indicating limited current utility for probability-threshold-based implementation.*

2.8 Permutation feature importance

Permutation importance analysis identified the first measured heart-rate value as the most influential predictor, followed by device identity, the standard deviation of absolute measurement error, and manufacturer. The remaining predictors showed smaller contributions to model performance, while verification method had negligible permutation importance. The ranking and variability of predictor importance are illustrated in Fig. 5, with complete numerical results provided in Supplementary Table S2.

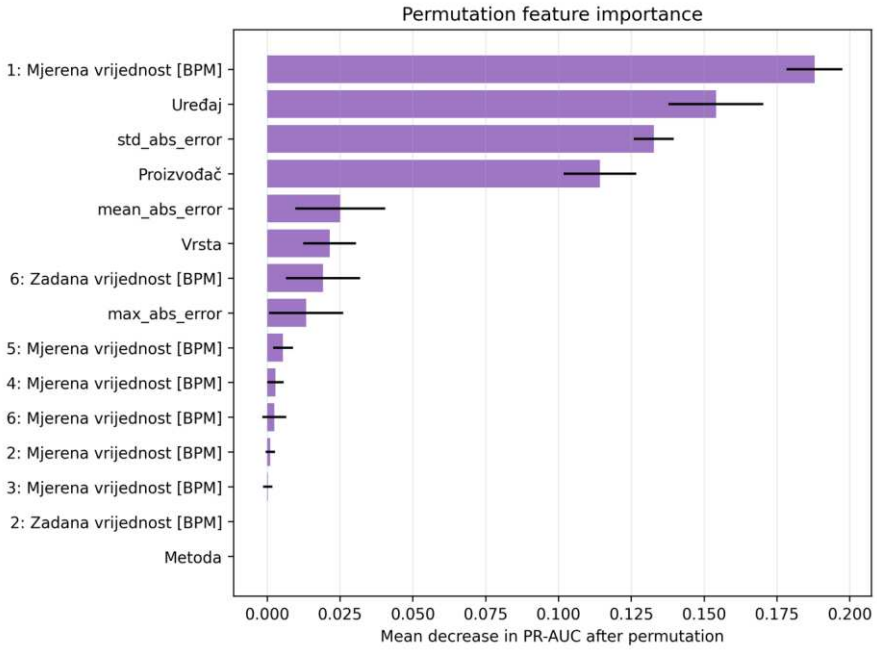


Figure 5. Permutation importance of the final LightGBM C-strict model. *Importance is expressed as the mean decrease in PR-AUC after random permutation of each feature. Error bars show the standard deviation across permutations.*

The importance analysis supports the technical plausibility of the model. The strongest features were not post-decision administrative variables but measurement-related and device-related properties. This finding is consistent with the leakage-controlled design of the final model.

2.9 Ablation analysis of device identity

An ablation analysis was performed to assess the contribution of device identity to model performance (Table 6). Removing this predictor reduced PR-AUC from 0.271 to 0.252 and ROC-AUC from 0.811 to 0.769. At the classification threshold of 0.50, recall increased from 0.663 to 0.721, while specificity decreased from 0.796 to 0.653 and precision from 0.067 to 0.044. These results indicate that device identity contributed to predictive performance, although the model retained discriminatory ability after its removal.

Table 6. Ablation analysis of the contribution of device identity to model performance.

Model	Features	PR-AUC	ROC-AUC	Recall FAIL	Specificity	Precision FAIL	Balanced accuracy	MCC	F2 score	Brier score
Primary C-strict LightGBM, with Device	**40**	**0.271**	**0.811**	**0.663**	**0.796**	**0.067**	**0.729**	**0.163**	**0.239**	**0.122**
Ablation LightGBM, without Device	**39**	**0.252**	**0.769**	**0.721**	**0.653**	**0.044**	**0.687**	**0.114**	**0.177**	**0.157**

Note: Both models were evaluated with the same stratified 5-fold out-of-fold prediction procedure and a classification threshold of 0.50. Delta values are calculated relative to the primary C-strict model.

2.10 SHAP-based explanation of the final model

SHAP analysis identified mean absolute measurement error as the most influential transformed predictor (mean absolute SHAP value: 0.670). After aggregating contributions from one-hot encoded categorical variables, manufacturer (25.59%) and device identity (22.65%) showed the highest grouped importance, followed by mean absolute measurement error (17.23%) and verification type (9.35%). These findings indicate that both device-related characteristics and metrological measurements contributed to model predictions. The SHAP beeswarm and global feature importance plots are presented in Figures 6, while detailed grouped importance results are provided in Supplementary Table S8.

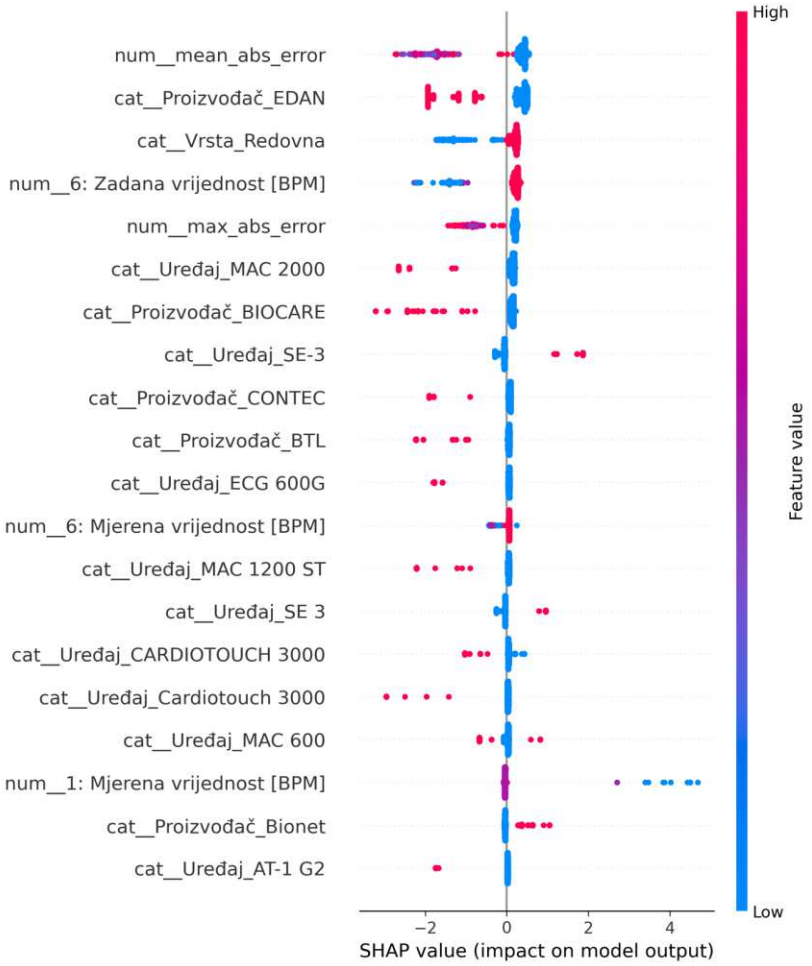


Figure 6. SHAP beeswarm plot for the final LightGBM C-strict model. Each point represents one verification record. Horizontal position shows the SHAP contribution to predicted verification-failure risk, while colour represents the corresponding transformed predictor value. The figure displays the 20 transformed features with the greatest mean absolute SHAP values.

2.11 Group-based robustness analysis

Group-based robustness analyses assessed model performance on previously unseen manufacturers and device models. Compared with standard stratified out-of-fold evaluation (PR-AUC: 0.271; ROC-AUC: 0.811), predictive performance declined under both validation schemes. Leave-one-manufacturer-out validation yielded a pooled PR-AUC of 0.153 and ROC-AUC of 0.733, while leave-one-device-out validation achieved 0.204 and 0.734, respectively. At the predefined threshold of 0.50, recall was particularly low for unseen manufacturers (0.349), while precision remained limited under both validation schemes (0.047 and 0.034, respectively). These findings indicate reduced model transportability across previously unseen device groups. Group-specific performance estimates were further constrained by small sample sizes and the absence of failing records in some categories. Detailed results are provided in Supplementary Tables S9 and S10.

3 Discussion

3.1 Predictive value of routine metrological verification data

Metrological verification of electrocardiographs is primarily intended to establish whether a device meets defined technical and performance requirements. However, the measurements collected during

these procedures may also provide information about patterns associated with device nonconformity. This possibility is relevant to the broader development of structured medical device information systems promoted by the World Health Organization [1] and builds on previous research concerning standardised conformity assessment of electrocardiographs [3].

The present study found that routinely collected verification data contain measurable predictive information. Among the three evaluated algorithms, LightGBM achieved the highest mean PR-AUC during repeated cross-validation. Its out-of-fold PR-AUC of 0.271 was considerably higher than the baseline failure prevalence of 2.17%, while the ROC-AUC of 0.811 indicated that the model could distinguish between passing and failing records with reasonable discriminatory ability. These findings suggest that technical measurements and device characteristics contain information associated with verification outcomes.

An important distinction, however, is that the model operates after technical measurements have been completed but before the final verification decision. It therefore estimates the likelihood of an unsuccessful verification based on measurements that are already available, rather than predicting device failure before inspection. Since these measurements also contribute to the established conformity-assessment process, predictive discrimination alone does not demonstrate an advantage over existing acceptance criteria [2,3].

The contribution of this study is therefore not the introduction of a new machine-learning algorithm or leakage-prevention method. Instead, it examines the extent to which routinely collected verification data retain predictive value when outcome-related information is excluded, and whether this information is sufficiently reliable to support further risk-based assessment. Defining the prediction time point and excluding variables that could reveal the final decision were essential to this evaluation, consistent with established concerns about label leakage in predictive modelling [7].

3.2 Predictive performance and its practical significance

Although the model demonstrated discriminatory ability, its performance at the prespecified classification threshold revealed a substantial limitation. It correctly identified 57 of 86 failing records but generated 791 false-positive alerts, resulting in a precision of only 6.7%. Consequently, approximately 93% of records identified as high risk had actually passed verification.

This difference between discrimination and precision is particularly important when unsuccessful verifications are rare. ROC-AUC measures how well a model ranks positive observations above negative observations, whereas precision describes how often a positive prediction is correct. In highly imbalanced datasets, relatively good discrimination can therefore coexist with a substantial number of incorrect alerts [12].

For a verification laboratory, these results have direct practical consequences. If every high-risk prediction resulted in additional technical review, 848 of the 3,956 records would require further attention. This represents more than one-fifth of the analysed workload and approximately fourteen false-positive alerts for every correctly identified failing record. Conversely, 29 failing records would not receive additional model-triggered review. These records would still undergo the mandatory verification procedure, but the model would not provide additional prioritisation for them.

The threshold sensitivity analysis further demonstrated this trade-off. Increasing the threshold to 0.79 reduced false-positive alerts from 791 to 154 but also reduced the number of correctly identified failing records from 57 to 39. Lowering the threshold to 0.36 identified 62 failing records but increased false-positive alerts to 1,006. These findings show that improved detection and reduced workload cannot be assumed to occur simultaneously.

The practical value of risk stratification therefore depends on whether a laboratory can identify a useful number of additional failing records within an acceptable review capacity. The present study did not establish such a balance. In particular, the proportion of failures identified among the highest-ranked 5%, 10%, or 20% of records remains to be evaluated. Until this analysis is completed, the model's usefulness as a prioritisation tool cannot be established from discrimination metrics alone.

3.3 Probability calibration and evidence of decision benefit

A second important finding concerns the reliability of the predicted probabilities. A model may correctly rank records according to relative risk while providing inaccurate estimates of their actual probability of failure. This distinction is fundamental when predicted probabilities are intended to guide operational decisions [14–16].

In the present study, the Brier score of 0.122 and expected calibration error of 0.200 indicated substantial discrepancies between predicted and observed failure probabilities. The Brier score was also considerably higher than the approximately 0.0213 obtained from a constant-prevalence reference prediction. Furthermore, the negative calibration-in-the-large intercept (−3.549) indicated systematic overestimation of failure risk, while the calibration slope of 0.623 suggested overly extreme probability estimates.

These findings indicate that the model's outputs cannot currently be interpreted as reliable absolute probabilities of verification failure. A predicted probability of 50%, for example, does not necessarily correspond to a 50% observed failure rate. This limits the suitability of using fixed probability thresholds to determine which records require additional review.

Decision curve analysis provided further evidence of this limitation. The model achieved positive net benefit relative to the review-none strategy only at low threshold probabilities between 0.01 and 0.03. At higher thresholds, the model did not demonstrate positive net benefit compared with reviewing no additional records. Although the analysis suggests some potential benefit within a narrow threshold range, it does not establish that these thresholds are appropriate or operationally useful in routine verification practice.

Taken together, the calibration and decision curve findings demonstrate that discriminatory performance does not necessarily translate into effective decision support. Probability recalibration may improve the reliability of estimated risks, but recalibration alone cannot establish operational usefulness. Both probability estimates and decision thresholds require independent validation against predefined laboratory objectives.

3.4 Interpretation of predictors and model robustness

Understanding which variables contribute to predictions is important in medical device quality assurance, where technical decisions must remain transparent and traceable. The present study examined feature importance using permutation importance and SHAP analysis, which provide complementary perspectives on model behaviour [17,18].

Permutation importance identified the first measured heart-rate value, device designation, variability in absolute measurement error, and manufacturer as the most influential predictors. Grouped SHAP analysis similarly highlighted manufacturer and device-related information, while mean and maximum absolute measurement errors contributed to individual predictions. These findings suggest that the model draws information from both metrological measurements and device characteristics.

Nevertheless, feature importance does not establish a causal relationship between individual characteristics and verification failure. Manufacturer and device categories may reflect differences in

equipment design, technical characteristics, verification histories, or the composition of the dataset. Their importance must therefore be interpreted as a model-specific predictive association rather than evidence that particular manufacturers or devices are intrinsically less reliable.

The ablation analysis provided further insight into this dependence. Removing device identity reduced PR-AUC from 0.271 to 0.252, indicating that the model retained some discriminatory ability without this predictor. However, the reduction in specificity and precision suggests that device-related information contributed to overall model performance. Since manufacturer remained among the predictors, the ablation results do not establish that the remaining predictive information was independent of device characteristics.

The group-based robustness analyses revealed a more substantial limitation. When previously unseen manufacturers were evaluated, pooled PR-AUC decreased to 0.153. Evaluation on previously unseen device models produced a PR-AUC of 0.204. Both results were below the standard out-of-fold estimate of 0.271, indicating reduced predictive performance outside the device groups represented during training.

These findings are important because verification laboratories may encounter different equipment portfolios over time. A model that performs adequately on familiar device categories may not provide equally reliable predictions for new manufacturers or models. The group-based results therefore reinforce the need to distinguish internal predictive performance from generalisability. Although these analyses provide evidence about robustness, they do not replace independent external validation.

3.5 Methodological limitations and implications for further research

Several methodological limitations affect the interpretation of these findings. Most importantly, the dataset contained only 86 failing records among 3,956 observations. This limited number of positive outcomes creates uncertainty in performance estimates and increases the risk of overfitting, particularly when categorical encoding substantially expands the number of model inputs. Small samples and limited positive-event counts can also make cross-validation estimates unstable [11]. Although the study used repeated cross-validation, out-of-fold predictions, bootstrap confidence intervals, and group-based robustness assessments, all analyses were based on a single retrospective dataset. The bootstrap intervals were calculated from fixed out-of-fold predictions and did not fully account for uncertainty introduced by model selection and fitting. Moreover, independent temporal and external validation were not performed. Consequently, the reported estimates may not accurately reflect performance in future verification records or different laboratory environments.

Further limitations include the absence of a simple statistical baseline, such as logistic regression, and the lack of formal sensitivity analyses addressing missing-data handling and correlated predictors. These issues are particularly relevant when interpreting the apparent contribution of individual measurement variables and device characteristics. Despite these limitations, the study provides a structured assessment of the opportunities and challenges associated with applying machine learning to routine metrological verification data. Its main value lies in demonstrating that predictive information is present, while identifying the conditions under which this information may become unreliable or operationally insufficient.

Future research should first establish whether ranking records by predicted risk offers an advantage over existing verification procedures. This requires evaluating the proportion of failing records captured within predefined review capacities, comparing the model with simpler prediction methods and established acceptance rules, and quantifying the additional workload required. Subsequent studies should assess performance using temporally separated and independent multicentre datasets, followed by probability recalibration and prospective evaluation of decision thresholds. Any future implementation would also require monitoring of prediction performance, calibration, changes in device populations, and false-positive alert burden. Such an approach is consistent with broader

expectations concerning transparency, validation, and human oversight in medical device management [6,8]. The present findings therefore support continued investigation of data-driven risk stratification in metrological verification, but do not yet justify routine operational deployment. Establishing practical benefit will require evidence that predictive information can improve the allocation of expert review resources without compromising established verification requirements.

4 Conclusion

This study demonstrates that routinely collected metrological verification data contain information that can support the development of data-driven approaches to medical device quality assurance, consistent with the World Health Organization's emphasis on structured information systems and evidence-informed healthcare technology management [1]. The leakage-controlled LightGBM model identified patterns associated with electrocardiograph verification failure, showing that existing technical records have analytical value beyond documenting individual conformity decisions.

However, predictive discrimination did not translate into demonstrated operational benefit. Low precision, poor probability calibration, substantial false-positive burden, and reduced performance on previously unseen manufacturers and device models limit the model's current applicability. These findings highlight the importance of evaluating machine-learning approaches not only through predictive performance, but also through their reliability, generalizability, and practical consequences.

The principal contribution of this research is therefore to establish both the potential and the limitations of applying machine learning to routine metrological verification data. Future research should focus on independent validation, probability recalibration, and evidence of improved risk-based prioritization within laboratory workflows. Such developments could contribute to more intelligent and data-informed medical device management systems while preserving established verification requirements and professional responsibility for conformity decisions.

5 Materials and Methods

5.1 Study design and dataset

This retrospective modeling study used a dataset containing 3,956 electrocardiograph verification records issued between 4 January 2021 and 12 March 2025. The dataset included 3,870 records classified as passing and 86 records classified as failing.

No a priori sample-size calculation was performed, as this retrospective study was based on an existing dataset of verification records. The number of observations was therefore determined by the available data rather than by a predefined statistical precision or power requirement.

The outcome distribution was strongly imbalanced: failing records represented 2.17% of the dataset. Consequently, accuracy was not treated as a primary measure of model quality. Instead, PR-AUC, recall for failing records, specificity, balanced accuracy, Matthews correlation coefficient, F2 score, and calibration metrics were used to provide a more informative assessment.

The binary outcome variable was **Verification pass**, where:

- YES - PASS (DA) represented a passing verification;
- NO - FAIL (NE) represented a failing verification.

For modeling purposes, pass was encoded as 0 and fail as 1. The unit of analysis was an individual electrocardiograph verification record rather than a patient. The dataset contained technical

measurements and device-related information, without patient demographics, clinical diagnoses, or patient-level ECG recordings.

5.2 Prediction time point

The intended prediction setting was defined as:

Post-measurement, pre-decision risk stratification.

This means that the model may use technical metadata and measurement results obtained during the verification procedure, but it may not use variables that are generated after, or are directly involved in, the final pass/fail verification decision.

This prediction time point was selected to avoid label leakage and to preserve potential operational value. The model is intended to help prioritize review after technical measurement data are available but before final administrative confirmation of the verification result.

5.3 Leakage-control strategy

All variables were reviewed according to their temporal availability and potential relationship to the target outcome. Variables were excluded when they met one or more of the following criteria:

1. They directly represented the target outcome.
2. They were created after the verification process or after the pass/fail decision.
3. They were administrative identifiers with limited generalizability.
4. They represented visual inspection or explicit conformity outcomes.
5. They encoded permissible deviation rules that were directly tied to final acceptance criteria.
6. They consisted of unstructured free-text documentation that could contain outcome-related information.

The following variables were excluded from the C-strict model:

- Number of record
- Date of issue
- Verification status (Pass / Fail)
- Request for verification
- Serial number
- Verification pass / fail
- Visual inspection feature 1
- Visual inspection feature 2
- Visual inspection feature 3
- Visual inspection feature 4

- 1: Compliant
- Note
- Number of work order
- DEVICE ERROR
- 1: Permissible error [± %]

The exclusion of administrative and post-decision variables was based on the principle that predictors must be available at the intended time of model use. Failure to respect predictor timing may create unrealistic estimates of performance and poor prospective transportability [7–9].

The leakage-control procedure focused on excluding outcome-related variables and ensuring that retained predictors were available at the intended prediction time point. However, feature-level exclusion alone does not eliminate the possibility of temporal dependence or repeated-device observations across training and validation subsets.

5.4 Final predictor set

The final C-strict predictor set contained 40 original variables:

- Four categorical technical metadata variables;
- Ten assigned-value variables;
- Ten measured-value variables;
- Ten measurement-error variables;
- Six derived summary features.

The final leakage-controlled predictor set is summarised in Table 7.

The categorical variables were:

- Method
- Type
- Manufacturer
- Device

The raw numerical variables consisted of the assigned, measured, and error values across ten verification points. Points 1–6 represented heart-rate measurements in BPM, whereas points 7–10 represented amplitude-related measurements in mV.

The following derived features were calculated from the ten error columns:

$$mean_abs_error = \frac{1}{n}\sum_{i=1}^{n} \quad | \, error_i \, |$$

$$max_abs_error = max(| \cdots | \, error_{10} \, |)$$

$$sum_abs_error = \sum_{i=1}^{10} \; | \; error_i \; |$$

$$std_abs_error = SD(| \cdots | \; error_{10} \; |)$$

$$median_abs_error = median(| \cdots | \; error_{10} \; |)$$

The variable n_zero_measured_points represented the number of measurement points with a recorded measured value equal to zero.

Table 7. Final leakage-controlled predictor set used in the C-strict model.

Predictor group	Variables included	Number of variables
Categorical metadata	Method, Type, Manufacturer, Device	4
Assigned values	1–10: Set value	10
Measured values	1–10: Measured value	10
Measurement errors	1–10: Measurement error	10
Derived error features	mean_abs_error, max_abs_error, sum_abs_error, std_abs_error, median_abs_error, n_zero_measured_points	6
	Total	**40**

Although the model included 40 original predictors, categorical encoding increased the dimensionality of the input data. The dataset contained 61 manufacturer categories and 334 device categories, resulting in a high-dimensional representation relative to the 86 observed failures. This imbalance between model complexity and the number of positive outcomes increases the potential risk of overfitting.

5.5 Data preprocessing

Numerical features were converted to numeric format. Missing numerical values were imputed using the median calculated within the training folds, and missingness indicators were added. Categorical variables were converted to string format, missing values were assigned a dedicated category, and categories were one-hot encoded.

All preprocessing operations were embedded within the cross-validation pipeline. This prevented information from validation folds from influencing preprocessing parameters estimated in training folds.

5.6 Candidate machine-learning models

Three candidate classifiers were compared:

1. **LightGBM**
2. **Balanced Random Forest**
3. **SMOTE Random Forest**

The candidate models were selected to compare a gradient-boosting approach with two random forest-based methods designed to address class imbalance. This comparison was intended to evaluate their relative performance within the present dataset rather than to establish the superiority of these algorithms over all possible classification approaches.

LightGBM is a gradient-boosted decision-tree algorithm designed for efficient learning from structured data [10]. Balanced Random Forest modifies the sampling process of random forests to improve

learning from minority-class observations. SMOTE Random Forest combines synthetic minority oversampling with random forest classification.

The LightGBM model used class balancing and the following principal settings:

- objective: binary classification;
- 300 estimators;
- learning rate: 0.03;
- 15 leaves;
- subsampling: 0.80;
- column subsampling: 0.80;
- random state: 42.

5.7 Internal validation and model selection

Model comparison used repeated stratified 5-fold cross-validation with 5 repeats, producing 25 validation folds per candidate model. Stratification ensured that each validation fold retained an approximately comparable proportion of failing records; repeated cross-validation was used to reduce dependence on a single data partition [11].

The primary selection criterion was mean PR-AUC. PR-AUC was selected because the failure class was rare and because precision–recall analysis provides a more informative evaluation than ROC-AUC when the positive class is strongly underrepresented [12] .

Secondary metrics included:

- ROC-AUC;
- recall for FAIL;
- specificity;
- precision for FAIL;
- balanced accuracy;
- Matthews correlation coefficient;
- F2 score;
- Brier score.
- MCC was included because it provides a balanced summary of binary classification performance and is particularly informative in imbalanced settings [13].

5.8 Out-of-fold evaluation

After selecting the best model, a single stratified 5-fold cross-validation procedure was used to generate one out-of-fold predicted probability for every record. These predictions were used to calculate final internal performance estimates.

Model selection and final out-of-fold evaluation were performed using records from the same retrospective dataset. As the evaluation did not employ a fully nested cross-validation framework, the reported performance estimates may retain optimism associated with the model-selection process.

A threshold of 0.50 was used for the primary confusion matrix. This threshold was not claimed to be clinically or operationally optimal; it was used as a transparent default threshold for reporting. Future implementation can select a decision threshold according to the acceptable balance between missed failures and additional manual reviews.

5.9 Bootstrap confidence intervals

A total of 2,000 bootstrap samples of the out-of-fold predictions were generated. For each bootstrap sample, PR-AUC, ROC-AUC, recall, specificity, precision, MCC, F2 score, and Brier score were recalculated. The 2.5th and 97.5th percentiles were used to derive 95% confidence intervals.

Bootstrap resampling was performed on the fixed out-of-fold predictions rather than by repeating the complete model-development procedure. Consequently, the resulting confidence intervals do not fully capture uncertainty associated with model fitting, model selection, and potential variations in the training data.

5.10 Calibration analysis

Calibration was evaluated using the Brier score, a 10-bin reliability diagram, and expected calibration error (ECE) [14-16]. The Brier score measures the mean squared difference between predicted probabilities and observed outcomes . A lower score indicates better probabilistic accuracy.

ECE was calculated as the weighted average absolute difference between the mean predicted probability and observed event frequency across probability bins. The model's calibration was considered moderate because ECE was 0.200 rather than close to zero.

5.11 Model explainability

Global feature importance was assessed using permutation importance [17]. Each predictor was randomly permuted multiple times, and the resulting mean decrease in PR-AUC was calculated. A greater decrease in PR-AUC indicates a stronger contribution of the predictor to model performance.

Permutation importance was selected because it evaluates the importance of original input features rather than only tree split counts. However, importance estimates can be interpreted cautiously when features are correlated, because correlated predictors can share or redistribute apparent importance. Global model explainability was assessed using both permutation importance and SHAP (SHapley Additive exPlanations) analysis. Permutation importance was calculated as the mean decrease in PR-AUC after repeated random permutation of each original predictor. SHAP values were additionally calculated to quantify the contribution of transformed predictors to individual model outputs [18]. For categorical variables represented by multiple one-hot encoded columns, SHAP contributions were aggregated to the original-predictor level by summing mean absolute SHAP values across the corresponding encoded columns. Agreement between grouped SHAP and permutation-importance rankings was assessed using Spearman and Kendall rank correlations. Both approaches were interpreted as measures of predictive association rather than causal effect, particularly for correlated measurement and device-identity variables. The complete leakage-controlled modelling workflow is summarised in Fig. 7.

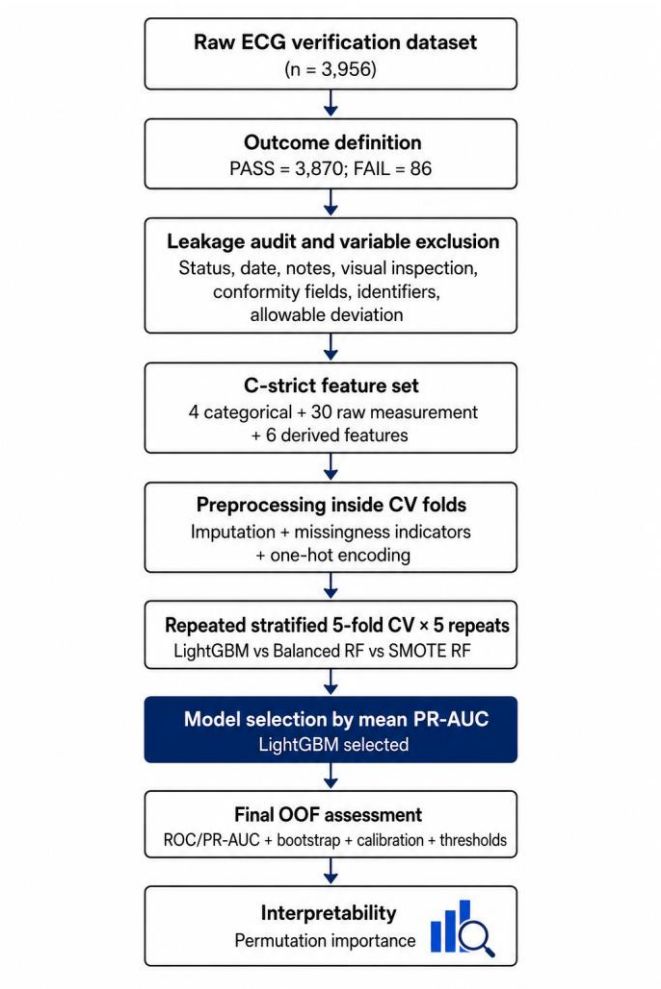


Figure 7. Workflow of the leakage-controlled machine-learning framework for electrocardiograph verification risk stratification.

Raw verification records were screened to remove target, post-decision, administrative, inspection, conformity-related, and free-text variables. The remaining post-measurement, pre-decision predictors were preprocessed within cross-validation folds and evaluated using repeated stratified cross-validation. Candidate classifiers were compared using PR-AUC as the primary selection criterion. The selected LightGBM model was subsequently evaluated using out-of-fold performance estimation, bootstrap confidence intervals, calibration assessment, threshold optimization, decision curve analysis, permutation importance, SHAP analysis, device-identity ablation, and group-based robustness validation.

Declarations

Ethics approval and consent to participate

This retrospective study analyzed technical verification records and did not involve direct patient contact, intervention, or identifiable patient information. According to institutional policy, formal ethics approval and informed consent were not required.

Consent for publication

Not applicable.

Availability of data and materials

The dataset analyzed during the current study is not publicly available because it contains institution-controlled verification records. De-identified data may be available from the corresponding author upon reasonable request and subject to institutional approval.

Code availability statemnt

The code used for data preprocessing, model development, validation, and statistical analysis is available from the corresponding author upon reasonable request, subject to institutional approval and applicable data governance requirements.

Competing interests

The authors declare that they have no competing interests.

Funding

No external funding was received for this study.

Authors' contributions

Adin Neslanović: Conceptualization, data curation, methodology, software, formal analysis, validation, visualization, writing—original draft, and writing—review and editing.

Irhad Neslanović: Conceptualization, methodology, software, formal analysis, interpretation of results, visualization, and writing—review and editing.

Berina Hasanefendić: Clinical and technical expertise regarding electrocardiograph verification practice, interpretation of the practical relevance of the findings, critical revision of the manuscript, and approval of the final version.

Ermin Begović: Clinical and technical expertise regarding electrocardiograph verification practice, interpretation of the practical relevance of the findings, critical revision of the manuscript, and approval of the final version.

Almir Badnjević: Scientific supervision and mentorship, methodological refinement, interpretation of results, critical revision of the manuscript for important intellectual content, and final preparation of the manuscript for publication.

All authors reviewed and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

References

1. World Health Organization. (2025). Inventory and maintenance management information system for medical devices. https://www.who.int/publications/i/item/9789240111257
2. International Electrotechnical Commission. IEC 60601-2-25:2011. Medical electrical equipment—Part 2-25: Particular requirements for the basic safety and essential performance of electrocardiographs. Geneva: IEC; 2011.

3. Badnjević A, Magjarević R, Mrdjanović E, Gurbeta Pokvić L. A novel method for conformity assessment testing of electrocardiographs for post-market surveillance purposes. Technology and Health Care. 2023.
4. Badnjević A, Magjarević R, Mrdjanović E, Gurbeta Pokvić L. A novel method for conformity assessment testing of electrocardiographs for post-market surveillance purposes. Technology and Health Care. 2023.
5. Cummins L, Sommers A, Bakhtiari Ramezani S, Mittal S, Jabour J, Seale M, Rahimi S. Explainable predictive maintenance: a survey of current methods, challenges and opportunities. 2024.
6. European Parliament and Council. Regulation (EU) 2017/745 on medical devices. Official Journal of the European Union. 2017; L117:1–175.
7. Davis SE, Matheny ME, Balu S, Sendak MP. A framework for understanding label leakage in machine learning for health care. Journal of the American Medical Informatics Association. 2024;31:274–280. doi:10.1093/jamia/ocad178.
8. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.
9. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Annals of Internal Medicine. 2015;162:55–63. doi:10.7326/M14-0697.
10. Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. In: Guyon I, von Luxburg U, Bengio S, et al., editors. *Advances in Neural Information Processing Systems 30*. Red Hook: Curran Associates; 2017. p. 3146–3154.

11. Varoquaux G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*. 2018;180:68–77. doi:10.1016/j.neuroimage.2017.06.061.

12. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10:e0118432. doi:10.1371/journal.pone.0118432.

13. Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*. 2021; 14:13. doi:10.1186/s13040-021-00244-z.

14. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*. 1950;78:1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

15. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*. 2010;21:128–138. doi:10.1097/EDE.0b013e3181c30fb2.

16. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Medicine*. 2019; 17:230. doi:10.1186/s12916-019-1466-7.

17. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*. 2019;20:1–81.

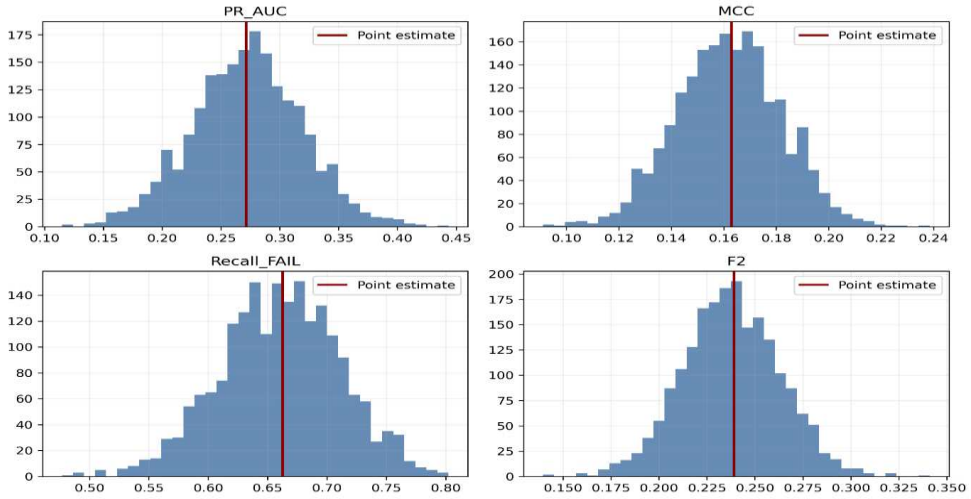
18. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, et al., editors. *Advances in Neural Information Processing Systems 30*. Red Hook: Curran Associates; 2017. p. 4765–4774.

Electronic Supplementary Material

Supplementary Table S1. Ten-bin calibration results for the final LightGBM C-strict model.

Bin	N	Mean_predicted_probability	Observed_FAIL_rate
1	2240	0.02185635101	0.00625
2	243	0.1333637217	0.02057613169
3	304	0.2762388514	0.01644736842
4	163	0.3540947064	0.01226993865
5	158	0.4633984029	0.01898734177
6	242	0.5482573688	0.03719008264
7	312	0.6512435789	0.02243589744
8	107	0.7596369849	0.02803738318
9	147	0.8596810928	0.1292517007
10	40	0.9530426086	0.475

Note: Predicted probabilities were divided into 10 equally spaced bins. Mean predicted probability represents the average model-estimated probability of verification failure within each bin, whereas observed FAIL rate represents the empirical proportion of failing records. Bins with no records are reported as not applicable. Expected calibration error was 0.200 and the Brier score was 0.122.



Supplementary figure 1. Bootstrap distributions of selected out-of-fold performance measures for the final LightGBM model. The vertical lines indicate point estimates for PR-AUC, MCC, recall for failing records, and F2 score

Supplementary Table S2. Full permutation feature-importance results for the final LightGBM C-strict model.

Feature	Permutation_importance_mean	Permutation_importance_SD	Normalized_importance_percent
1: Measured value [BPM]	0.1879537913	0.009565309631	27.62223097
Device	0.1541139699	0.01627076055	22.64903327
std_abs_error	0.1327553813	0.006861143264	19.51011352
Manufacturer	0.1143193316	0.01247473349	16.80069851
mean_abs_error	0.02506673356	0.01544484535	3.683879423
Type	0.02146831762	0.009089755608	3.155045843
6: Set value [BPM]	0.01920793122	0.01271503019	2.822852942
max_abs_error	0.0134251814	0.01270833262	1.973003359
5: Measured value [BPM]	0.005443985369	0.003394512566	0.8000637829
4: Measured value [BPM]	0.002893189873	0.002820553964	0.4251915238
6: Measured value [BPM]	0.002468169858	0.004127106646	0.3627293571
2: Measured value [BPM]	0.001067885152	0.001613618478	0.1569394803
3: Measured value [BPM]	0.0002600522061	0.001620595558	0.03821802184
Method	0	0	0

2: Set value [BPM]	0	0	0
1: Set value [BPM]	0	0	0
4: Set value [BPM]	0	0	0
3: Set value [BPM]	0	0	0
7: Measured value [mV]	0	0	0
10: Set value [mV]	0	0	0
9: Measured value [mV]	0	0	0
9: Set value [mV]	0	0	0
8: Set value [mV]	0	0	0
7: Set value [mV]	0	0	0
8: Measurement error [%]	0	0	0
7: Measurement error [%]	0	0	0
8: Measured value [mV]	0	0	0
1: Measurement error [%]	0	0	0
5: Measurement error [%]	0	0	0
2: Measurement error [%]	0	0	0
3: Measurement error [%]	0	0	0
4: Measurement error [%]	0	0	0
9: Measurement error [%]	0	0	0
median_abs_error	0	0	0
6: Measurement error [%]	0	0	0
10: Measurement error [%]	0	0	0
5: Set value [BPM]	-7.40E-18	1.89E-17	0
10: Measured value [mV]	-6.81E-06	3.67E-05	0
n_zero_measured_points	-0.0002528345153	0.0002541254999	0
sum_abs_error	-0.002237770923	0.003354259068	0

Note: Permutation importance was calculated as the mean decrease in precision–recall area under the curve after random permutation of each predictor. Standard deviations represent variability across 30 permutations. Features with zero or negative importance were retained in the supplementary table for transparency.

Supplementary Table S3. Calibration-in-the-large and calibration slope of out-of-fold predictions for the final LightGBM C-strict model.

Metric	Estimate	Ideal_value	Interpretation
Calibration-in-the-large (intercept)	-3.548464665	0	0 is ideal. Negative values indicate average risk overprediction; positive values indicate underprediction.
Calibration slope	0.6227752992	1	1 is ideal. Values below 1 indicate predictions that are too extreme; values above 1 indicate insufficient spread.

Note: Calibration-in-the-large was estimated as the intercept of a logistic calibration model with the logit of the predicted probability included as an offset. The calibration slope was estimated by logistic regression of the observed outcome on the logit of the predicted probability. Ideal values are 0 for the intercept and 1 for the slope. A negative intercept indicates average overprediction of verification-failure risk, whereas a slope below 1 indicates overly extreme predicted probabilities.

Supplementary Table S4. Full threshold sensitivity analysis for the final LightGBM C-strict model across thresholds from 0.05 to 0.95.

Threshold	Recall_FAIL	Specificity	Precision_FAIL	Balanced_accuracy	MCC	F2	Youden_J	TP	FP	TN	FN	Cost_FN5_FP1
0.01	0.9534883721	0.2599483204	0.02783435166	0.6067183463	0.07138331826	0.1246200608	0.2134366925	82	2864	1006	4	2884
0.02	0.9534883721	0.3651162791	0.0322961796	0.6593023256	0.09690374116	0.1422129726	0.3186046512	82	2457	1413	4	2477
0.03	0.9302325581	0.430749354	0.03504161191	0.6804909561	0.1065587759	0.152264941	0.3609819121	80	2203	1667	6	2233
0.04	0.9186046512	0.4759689922	0.03749406739	0.6972868217	0.115327282	0.1611587107	0.3945736434	79	2028	1842	7	2063
0.05	0.9186046512	0.4821705426	0.03792606817	0.7003875969	0.117055552	0.1627523692	0.4007751938	79	2004	1866	7	2039
0.06	0.8720930233	0.5136950904	0.03832396525	0.6928940568	0.1125257046	0.1629726206	0.3857881137	75	1882	1988	11	1937
0.07	0.8488372093	0.5279069767	0.03842105263	0.688372093	0.1099671353	0.1626559715	0.376744186	73	1827	2043	13	1892
0.08	0.8488372093	0.5475452196	0.04002192982	0.6981912145	0.1159613027	0.1683579336	0.3963824289	73	1751	2119	13	1816
0.09	0.8488372093	0.5602067183	0.04112676056	0.7045219638	0.1199354714	0.1722510618	0.4090439276	73	1702	2168	13	1767
0.1	0.8372093023	0.5751937984	0.04195804196	0.7062015504	0.1213511732	0.1747572816	0.4124031008	72	1644	2226	14	1714
0.11	0.8255813953	0.5873385013	0.04256594724	0.7064599483	0.121939521	0.1764413519	0.4129198966	71	1597	2273	15	1672
0.12	0.8255813953	0.5994832041	0.04380012338	0.7125322997	0.1260447309	0.1806615776	0.4250645995	71	1550	2320	15	1625
0.13	0.8023255814	0.607751938	0.04347826087	0.7050387597	0.1220111937	0.1786639047	0.4100775194	69	1518	2352	17	1603
0.14	0.8023255814	0.6126614987	0.04400510204	0.7074935401	0.1237226158	0.1804393305	0.4149870801	69	1499	2371	17	1584
0.15	0.8023255814	0.619379845	0.04474708171	0.7108527132	0.1260965132	0.1829268293	0.4217054264	69	1473	2397	17	1558
0.16	0.8023255814	0.6299741602	0.04596935376	0.7161498708	0.1299192754	0.1869918699	0.4322997416	69	1432	2438	17	1517
0.17	0.7790697674	0.6320413437	0.04493628437	0.7055555556	0.1237133705	0.1825613079	0.4111111111	67	1424	2446	19	1519
0.18	0.7790697674	0.6328165375	0.04502688172	0.7059431525	0.1239960975	0.182860262	0.4118863049	67	1421	2449	19	1516
0.19	0.7790697674	0.6328165375	0.04502688172	0.7059431525	0.1239960975	0.182860262	0.4118863049	67	1421	2449	19	1516
0.2	0.7790697674	0.6366925065	0.04548540394	0.707881137	0.1254180529	0.1843698404	0.4157622739	67	1406	2464	19	1501
0.21	0.7790697674	0.6379844961	0.04564032698	0.7085271318	0.1258951668	0.1848785872	0.4170542636	67	1401	2469	19	1496
0.22	0.7790697674	0.6379844961	0.04564032698	0.7085271318	0.1258951668	0.1848785872	0.4170542636	67	1401	2469	19	1496
0.23	0.7790697674	0.638501292	0.04570259209	0.7087855297	0.1260864569	0.1850828729	0.4175710594	67	1399	2471	19	1494
0.24	0.7790697674	0.6392764858	0.04579630895	0.7091731266	0.1263738712	0.1853901494	0.4183462532	67	1396	2474	19	1491
0.25	0.7674418605	0.6410852713	0.04536082474	0.7042635659	0.1235484413	0.1834352418	0.4085271318	66	1389	2481	20	1489
0.26	0.7674418605	0.6410852713	0.04536082474	0.7042635659	0.1235484413	0.1834352418	0.4085271318	66	1389	2481	20	1489
0.27	0.7674418605	0.6560723514	0.04724409449	0.7117571059	0.1292228606	0.1895462378	0.4235142119	66	1331	2539	20	1431
0.28	0.7674418605	0.6718346253	0.0494011976	0.7196382429	0.1354530504	0.1964285714	0.4392764858	66	1270	2600	20	1370
0.29	0.7209302326	0.7051679587	0.05153782211	0.7130490956	0.1350760896	0.2003878474	0.4260981912	62	1141	2729	24	1261
0.3	0.7209302326	0.7139534884	0.05303678358	0.7174418605	0.1389959414	0.2048909451	0.4348837209	62	1107	2763	24	1227

0.31	0.7209302326	0.7139534884	0.05303678358	0.7174418605	0.1389959414	0.2048909451	0.4348837209	62	1107	2763	24	1227
0.32	0.7209302326	0.7183462532	0.05381944444	0.7196382429	0.1410026972	0.2072192513	0.4392764858	62	1090	2780	24	1210
0.33	0.7209302326	0.7183462532	0.05381944444	0.7196382429	0.1410026972	0.2072192513	0.4392764858	62	1090	2780	24	1210
0.34	0.7209302326	0.7183462532	0.05381944444	0.7196382429	0.1410026972	0.2072192513	0.4392764858	62	1090	2780	24	1210
0.35	0.7209302326	0.7382428941	0.0576744186	0.7295865633	0.1505238471	0.2184637068	0.4591731266	62	1013	2857	24	1133
0.36	0.7209302326	0.7400516796	0.05805243446	0.7304909561	0.1514273675	0.2195467422	0.4609819121	62	1006	2864	24	1126
0.37	0.7209302326	0.7400516796	0.05805243446	0.7304909561	0.1514273675	0.2195467422	0.4609819121	62	1006	2864	24	1126
0.38	0.6976744186	0.7511627907	0.05865102639	0.7244186047	0.1494857219	0.2194586686	0.4488372093	60	963	2907	26	1093
0.39	0.6976744186	0.753746 77	0.05923000987	0.7257105943	0.1508296747	0.2210759027	0.4514211886	60	953	2917	26	1083
0.4	0.6976744186	0.7555555556	0.05964214712	0.7266149871	0.1517795755	0.2222222222	0.4532299742	60	946	2924	26	1076
0.41	0.6976744186	0.757622739	0.06012024048	0.7276485788	0.1528745852	0.2235469449	0.4552971576	60	938	2932	26	1068
0.42	0.6976744186	0.757622739	0.06012024048	0.7276485788	0.1528745852	0.2235469449	0.4552971576	60	938	2932	26	1068
0.43	0.6976744186	0.757622739	0.06012024048	0.7276485788	0.1528745852	0.2235469449	0.4552971576	60	938	2932	26	1068
0.44	0.6744186047	0.7666666667	0.06035379813	0.7205426357	0.1499916757	0.2222222222	0.4410852713	58	903	2967	28	1043
0.45	0.6744186047	0.7679586563	0.06066945607	0.7211886305	0.1506981532	0.2230769231	0.4423772 61	58	898	2972	28	1038
0.46	0.6744186047	0.7739018088	0.06216505895	0.7241602067	0.1540044701	0.2270947533	0.4483204134	58	875	2995	28	1015
0.47	0.6744186047	0.7741602067	0.06223175966	0.7242894057	0.1541503852	0.2272727273	0.4485788114	58	874	2996	28	1014
0.48	0.6627906977	0.784754522	0.06404494382	0.7237726098	0.1563005665	0.2309562399	0.4475452196	57	833	3037	29	978
0.49	0.6627906977	0.784754522	0.06404494382	0.7237726098	0.1563005665	0.2309562399	0.4475452196	57	833	3037	29	978
0.5	0.6627906977	0.7956072351	0.06721698113	0.7291989664	0.1628954401	0.2390939597	0.4583979328	57	791	3079	29	936
0.51	0.6511627907	0.7997416021	0.06738868833	0.7254521964	0.1614223449	0.2382978723	0.4509043928	56	775	3095	30	925
0.52	0.6395348837	0.8023255814	0.06707317073	0.7209302326	0.1589625791	0.2362542955	0.4418604651	55	765	3105	31	920
0.53	0.6046511628	0.8198966408	0.0694259012	0.7122739018	0.1580305013	0.2378774016	0.4245478036	52	697	3173	34	867
0.54	0.6046511628	0.823255814	0.07065217391	0.7139534884	0.1603567451	0.2407407407	0.4279069767	52	684	3186	34	854
0.55	0.6046511628	0.8253229974	0.07142857143	0.7149870801	0.1618134481	0.2425373134	0.4299741602	52	676	3194	34	846
0.56	0.5813953488	0.8325581395	0.07163323782	0.7069767442	0.1583627386	0.2399232246	0.4139534884	50	648	3222	36	828
0.57	0.5813953488	0.8354005168	0.07278020378	0.7083979328	0.1604509459	0.2424830262	0.4167958656	50	637	3233	36	817
0.58	0.5581395349	0.8511627907	0.07692307692	0.7046511628	0.1637586047	0.2479338843	0.4093023256	48	576	3294	38	766
0.59	0.5581395349	0.8542635659	0.07843137255	0.7062015504	0.1663097783	0.2510460251	0.4124031008	48	564	3306	38	754
0.6	0.5581395349	0.8558139535	0.07920792079	0.7069767442	0.1676090788	0.2526315789	0.4139534884	48	558	3312	38	748
0.61	0.5581395349	0.8573643411	0.08	0.707751938	0.1689247116	0.2542372881	0.4155038876	48	552	3318	38	742
0.62	0.5465116279	0.8682170543	0.08438061041	0.7073643411	0.1738863935	0.2608213097	0.4147286822	47	510	3360	39	705

0.63	0.5348837209	0.8842377261	0.09311740891	0.7095607235	0.1848917729	0.2744630072	0.419121447	46	448	3422	40	648
0.64	0.5348837209	0.8855297158	0.09406952965	0.7102067183	0.1862730157	0.2761104442	0.4204134367	46	443	3427	40	643
0.65	0.5348837209	0.8857881137	0.09426229508	0.7103359173	0.1865514755	0.2764423077	0.4206718346	46	442	3428	40	642
0.66	0.5	0.8888888889	0.09090909091	0.6944444444	0.1747925569	0.2631578947	0.3888888889	43	430	3440	43	645
0.67	0.5	0.9180878553	0.1194444444	0.7090439276	0.2119879468	0.3053977273	0.4180878553	43	317	3553	43	532
0.68	0.476744186	0.9279069767	0.128125	0.7023255814	0.2164205889	0.3087349398	0.4046511628	41	279	3591	45	504
0.69	0.476744186	0.9320413437	0.1348684211	0.7043927649	0.2238195643	0.3163580247	0.4087855297	41	263	3607	45	488
0.7	0.476744186	0.934625323	0.1394557823	0.7056847545	0.2287199099	0.3213166144	0.411369509	41	253	3617	45	478
0.71	0.4651162791	0.9361757106	0.1393728223	0.7006459948	0.2256058696	0.3169572108	0.4012919897	40	247	3623	46	477
0.72	0.4651162791	0.9361757106	0.1393728223	0.7006459948	0.2256058696	0.3169572108	0.4012919897	40	247	3623	46	477
0.73	0.4651162791	0.9382428941	0.1433691756	0.7016795866	0.2297458787	0.3210272873	0.4033591731	40	239	3631	46	469
0.74	0.4651162791	0.9405684755	0.1481481481	0.7028423773	0.2346031548	0.325732899	0.4056847545	40	230	3640	46	460
0.75	0.4534883721	0.9434108527	0.1511627907	0.6984496124	0.2344183757	0.3239202658	0.3968992248	39	219	3651	47	454
0.76	0.4534883721	0.9439276486	0.15234375	0.6987080103	0.2355750185	0.325	0.3974160207	39	217	3653	47	452
0.77	0.4534883721	0.9475452196	0.1611570248	0.7005167959	0.2440376461	0.3327645051	0.4010335917	39	203	3667	47	438
0.78	0.4534883721	0.9547803618	0.1822429907	0.7041343669	0.2632037374	0.3494623656	0.4082687339	39	175	3695	47	410
0.79	0.4534883721	0.9602067183	0.2020725389	0.7068475452	0.2800524081	0.3631284916	0.4136950904	39	154	3716	47	389
0.8	0.4418604651	0.961498708	0.2032085561	0.7016795866	0.2771810564	0.3578154426	0.4033591731	38	149	3721	48	389
0.81	0.4302325581	0.9625322997	0.2032967033	0.6963824289	0.2734013717	0.3517110266	0.3927648579	37	145	3725	49	390
0.82	0.4069767442	0.965374677	0.2071005917	0.6861757106	0.2685113867	0.3411306043	0.3723514212	35	134	3736	51	389
0.83	0.3720930233	0.9677002584	0.2038216561	0.6698966408	0.2538253671	0.3193612774	0.3397932817	32	125	3745	54	395
0.84	0.3372093023	0.9708010336	0.2042253521	0.654005168	0.2414545991	0.2983539095	0.3080103359	29	113	3757	57	398
0.85	0.3139534884	0.9718346253	0.1985294118	0.6428940568	0.2287429588	0.28125	0.2857881137	27	109	3761	59	404
0.86	0.3139534884	0.9741602067	0.2125984252	0.6440568475	0.2383548904	0.2866242038	0.2881136951	27	100	3770	59	395
0.87	0.2674418605	0.9785529716	0.2169811321	0.622997416	0.2221505507	0.2555555556	0.245994832	23	83	3787	63	398
0.88	0.2441860465	0.9821705426	0.2333333333	0.6131782946	0.2213837144	0.2419354839	0.2263565891	21	69	3801	65	394
0.89	0.2325581395	0.9870801034	0.2857142857	0.6098191214	0.2429473972	0.2415458937	0.2196382429	20	50	3820	66	380
0.9	0.2209302326	0.9945736434	0.475	0.607751938	0.3141293072	0.2473958333	0.215503876	19	21	3849	67	356
0.91	0.2209302326	0.9948320413	0.4871794872	0.607881137	0.3184719076	0.2480417755	0.2157622739	19	20	3850	67	355
0.92	0.2093023256	0.9963824289	0.5625	0.6028423773	0.3348631223	0.2393617021	0.2056847545	18	14	3856	68	354
0.93	0.1976744186	0.9966408269	0.5666666667	0.5971576227	0.326644845	0.2272727273	0.1943152455	17	13	3857	69	358
0.94	0.1860465116	0.9976744186	0.64	0.5918604651	0.3380973896	0.216802168	0.1837209302	16	9	3861	70	359

0.95	0.18604 65116	0.99896 64083	0.8	0.5925064 599	0.38042 07721	0.21978 02198	0.18501 29199	1 6	4	38 66	7 0	354

Supplementary Table S5. Feature-level missingness in the candidate pre-decision predictor set.

Threshold_probability	Net_benefit_model	Net_benefit_review_all	Net_benefit_review_none
0.01	0.01341524446	0.01185770751	0
0.02	0.008052867254	0.001774622893	0
0.03	0.002999489227	-0.008516360377	0
0.04	-0.001390293225	-0.01902173913	0
0.05	-0.00669203342	-0.02974828375	0
0.06	-0.01140739625	-0.04070305273	0
0.07	-0.01630842493	-0.05189340813	0
0.08	-0.02003560909	-0.06332703214	0
0.09	-0.02409748997	-0.07501194458	0
0.1	-0.0279743849	-0.08695652174	0
0.11	-0.03194692176	-0.09916951637	0
0.12	-0.03548120232	-0.1116600791	0
0.13	-0.0398957498	-0.1244377811	0
0.14	-0.04424248125	-0.137512639	0
0.15	-0.04826622257	-0.1508951407	0
0.16	-0.05150705378	-0.1645962733	0
0.17	-0.05679035657	-0.1786275537	0
0.18	-0.06191274754	-0.1930010604	0
0.19	-0.06732077544	-0.2077294686	0
0.2	-0.07191607685	-0.222826087	0
0.21	-0.07720367076	-0.2383048982	0
0.22	-0.08295092168	-0.254180602	0
0.23	-0.08869644006	-0.2704686618	0
0.24	-0.09450002661	-0.2871853547	0
0.25	-0.1003538928	-0.3043478261	0
0.26	-0.1066802394	-0.3219741481	0
0.27	-0.1077572475	-0.340083383	0
0.28	-0.1081620043	-0.3586956522	0
0.29	-0.1021340378	-0.3778322107	0
0.3	-0.1042539362	-0.397515528	0
0.31	-0.1100474788	-0.4177693762	0
0.32	-0.113989175	-0.4386189258	0
0.33	-0.120036823	-0.4600908501	0
0.34	-0.1262677329	-0.4822134387	0
0.35	-0.1222096912	-0.5050167224	0
0.36	-0.127369818	-0.5285326087	0
0.37	-0.1336767939	-0.5527950311	0
0.38	-0.1340307903	-0.5778401122	0
0.39	-0.138851133	-0.6037063435	0
0.4	-0.1442534547	-0.6304347826	0
0.41	-0.149603263	-0.6580692704	0
0.42	-0.156532199	-0.6866566717	0
0.43	-0.1637042556	-0.7162471396	0
0.44	-0.1646865521	-0.7468944099	0
0.45	-0.1710635169	-0.7786561265	0
0.46	-0.1737538853	-0.8115942029	0
0.47	-0.1812579888	-0.8457752256	0
0.48	-0.1799603329	-0.881270903	0
0.49	-0.1879002359	-0.9181585678	0
0.5	-0.1855409505	-0.9565217391	0

Supplementary Table S6. Hyperparameters and preprocessing configuration of the candidate machine-learning models.

Component	Configuration
Cross-validation	RepeatedStratifiedKFold; 5 folds × 5 repeats; random state 42

Component	Configuration
Numerical imputation	Median imputation within training folds, with missingness indicators
Categorical preprocessing	Most-frequent imputation; one-hot encoding; unknown categories ignored
LightGBM	Binary objective; 300 trees; learning rate 0.03; num_leaves 15; subsample 0.80; colsample_bytree 0.80; class_weight balanced; random state 42
Balanced Random Forest	400 trees; max_features square root; min_samples_leaf 2; sampling_strategy all; replacement true; random state 42
SMOTE	k_neighbors 5; random state 42; applied inside training folds only
SMOTE Random Forest	400 trees; max_features square root; min_samples_leaf 2; random state 42
Model-selection metric	Mean PR-AUC
Primary reporting threshold	0.50
Exploratory threshold rule	Maximum F2 score based on OOF predictions

Supplementary Table S7. Full leakage audit: included and excluded variables, rationale, and availability at the intended prediction time.

Variable group	Decision	Rationale
Verification pass	Excluded	Target outcome
Verification status	Excluded	Administrative process/status variable; may be post-decision
Date of issue	Excluded	Post-procedural temporal metadata
Request for verification	Excluded	Contract/order/request identifier with high cardinality
No of report, No of work order, Serial number	Excluded	Administrative or device identifiers; risk of memorization and poor transportability
Visual inspection 1–4	Excluded	Inspection outcomes likely tied to final decision
Compliant 1–10	Excluded	Direct conformity outcome indicators
Permissible error 1–10	Excluded	Decision-rule fields linked to acceptance criteria
Note	Excluded	Free text; potentially outcome-proximal information
Method, Type, Manufacturer, Device	Included	Technical/device metadata available before final decision
Set value, measured value and measurement error	Included	Post-measurement, pre-decision technical measurements
Derived error summaries	Included	Calculated solely from included measurement-error features

Supplementary Table S8. Grouped SHAP importance by original predictor for the final LightGBM C-strict model.

Original_predictor	Mean_abs_grouped_SHAP	N_transformed_columns	Normalized_importance_percent
Manufacturer	0.9949238074	61	25.59210516
Device	0.8806459649	334	22.65257296
mean_abs_error	0.6698705863	1	17.23086568
Type	0.3635147119	3	9.350572039
6: Set value [BPM]	0.3460856782	1	8.902250608
max_abs_error	0.3436337005	1	8.839179174
6: Measured value [BPM]	0.09988223139	1	2.569238519
1: Measured value [BPM]	0.07213181323	1	1.855423436
5: Measured value [BPM]	0.04101967723	1	1.055135967
std_abs_error	0.02665486515	1	0.6856345248
sum_abs_error	0.02492313222	1	0.6410897155
4: Measured value [BPM]	0.01544336231	1	0.3972446425
2: Measured value [BPM]	0.004062304261	1	0.1044933462
3: Measured value [BPM]	0.003250745274	1	0.08361787533
n_zero_measured_points	0.001424020429	1	0.03662961956
10: Measured value [mV]	0.0001074678194	1	0.00276436016
5: Set value [BPM]	4.60E-05	1	0.001182373027
10: Measurement error [%]	0	1	0
3: Measurement error [%]	0	1	0
2: Set value [BPM]	0	1	0
2: Measurement error [%]	0	1	0
1: Set value [BPM]	0	1	0
10: Set value [mV]	0	1	0
1: Measurement error [%]	0	1	0
4: Set value [BPM]	0	1	0
6: Measurement error [%]	0	1	0
8: Measured value [mV]	0	1	0
8: Measurement error [%]	0	1	0
7: Set value [mV]	0	1	0
7: Measured value [mV]	0	1	0
7: Measurement error [%]	0	1	0
9: Measured value [mV]	0	1	0
9: Set value [mV]	0	1	0
8: Set value [mV]	0	1	0
Method	0	1	0
9: Measurement error [%]	0	1	0
median_abs_error	0	1	0
missingindicator_10: Measurement error [%]	0	1	0
missingindicator_10: Measured value [mV]	0	1	0
5: Measurement error [%]	0	1	0
4: Measurement error [%]	0	1	0
3: Set value [BPM]	0	1	0
missingindicator_1: Measured value [BPM]	0	1	0
missingindicator_1: Measurement error [%]	0	1	0
missingindicator_10: Set value [mV]	0	1	0
missingindicator_1: Set value [BPM]	0	1	0
missingindicator_3: Measurement error [%]	0	1	0
missingindicator_2: Measurement error [%]	0	1	0
missingindicator_2: Measured value [BPM]	0	1	0
missingindicator_2: Set value [BPM]	0	1	0

missingindicator_4: Measured value [BPM]	0	1	0
missingindicator_4: Set value [BPM]	0	1	0
missingindicator_5: Measurement error [%]	0	1	0
missingindicator_5: Measured value [BPM]	0	1	0
missingindicator_5: Set value [BPM]	0	1	0
missingindicator_3: Measured value [BPM]	0	1	0
missingindicator_3: Set value [BPM]	0	1	0
missingindicator_4: Measurement error [%]	0	1	0
missingindicator_6: Set value [BPM]	0	1	0
missingindicator_6: Measured value [BPM]	0	1	0
missingindicator_6: Measurement error [%]	0	1	0
missingindicator_7: Measurement error [%]	0	1	0
missingindicator_8: Measured value [mV]	0	1	0
missingindicator_7: Measured value [mV]	0	1	0
missingindicator_7: Set value [mV]	0	1	0
missingindicator_8: Measurement error [%]	0	1	0
missingindicator_9: Measured value [mV]	0	1	0
missingindicator_9: Measurement error [%]	0	1	0
missingindicator_8: Set value [mV]	0	1	0
missingindicator_9: Set value [mV]	0	1	0
missingindicator_median_abs_error	0	1	0
missingindicator_mean_abs_error	0	1	0
missingindicator_max_abs_error	0	1	0
missingindicator_std_abs_error	0	1	0

Note: For one-hot encoded categorical predictors, signed SHAP values were summed across all derived category columns for each record before calculation of the mean absolute grouped SHAP value. Normalized importance percentages are relative to the sum of mean absolute grouped SHAP values across all predictors. SHAP values indicate model-specific contributions to predicted verification-failure risk and do not establish causal effects.

Supplementary Table S9. Group-level performance results from leave-one-manufacturer-out validation of the final LightGBM C-strict model.

Manufacturer	N_records	N_PASS	N_FAIL	Mean_predicted_risk	PR_AUC	ROC_AUC
GE - General Electric	716	697	19	0.3741455353	0.2376296	0.7018424828
MINDRAY	59	57	2	0.3674128296	1	1
CONTEC	147	145	2	0.4005065886	0.2	0.9724137931
Schiller	715	698	17	0.3029779584	0.09324344755	0.7718692061
EDAN	720	714	6	0.3858447371	0.01023890785	0.593837535
Bionet	236	230	6	0.3919394353	0.2997557998	0.743115942
WELCH ALLYN	53	48	5	0.2466801157	0.5304347826	0.8875
Cardioline	188	182	6	0.2785970163	0.3673450678	0.777014652

ASPEL	59	59	0	0.3573885349		
MAQUET	2	2	0	0.07041058326		
BIOCARE	206	205	1	0.379313326	0.007352941176	0.6707317073
Medical Econet	72	71	1	0.3450821458	0.02040816327	0.323943662
Fukuda Denshi	34	34	0	0.3400972069		
HEALTH PLUS	8	8	0	0.3843067947		
COMEN	4	4	0	0.4071099478		
ESAOTE	94	90	4	0.2386955113	0.09090909091	0.7777777778
Siemens	16	16	0	0.1945451998		
CARDIOSPY	15	15	0	0.3399451373		
LABTECH ltd	14	14	0	0.3599205426		
SOMO	24	24	0	0.3901722407		
MWZ	3	2	1	0.3333200413	0.5	0.75
MORTARA	31	31	0	0.3107989405		
GIMA	31	30	1	0.2679479275	0.05263157895	0.4333333333
NIHON KOHDEN	76	74	2	0.3388698086	0.03921568627	0.6621621622
Fazzini	52	50	2	0.3673004378	0.05714285714	0.67
LUMED	16	16	0	0.4092102001		
DRAGER	1	1	0	0.5076481612		
Forlong	68	67	1	0.3858189758	1	1
Bioset	1	1	0	0.6259211549		
Innomed	7	6	1	0.4333711486	1	1
BTL	131	130	1	0.2751032243	1	1
MAC	1	1	0	0.02568313001		
Trismed	31	29	2	0.2955343453	0.5588235294	0.8706896552
Hellige	5	4	1	0.5479827684	1	1
CAREWELL	4	4	0	0.3856369929		
Marquette	14	14	0	0.2674399963		
Fukuda Cardiosuny	3	3	0	0.006845019607		
PHILIPS	4	4	0	0.2299993435		
EKONA	3	1	2	0.4949281854	0.6666666667	0.5
MEDWELT	1	1	0	0.09186120629		
PROGETTI	4	4	0	0.3832248281		
SHENZHEN	11	11	0	0.1940230527		
HORMANN	14	13	1	0.166797381	0.2	0.8461538462
HP	21	20	1	0.3383112123	0.07142857143	0.675
PHYSIO CONTROL	1	1	0	0.5076481612		
ARI MEDICAL	2	2	0	0.302964961		
BURDICK	5	5	0	0.4097774716		
Cardiette	7	7	0	0.2231777488		
ZON CARE	3	2	1	0.3068807812	0.5	0.75
CARDIORAPID	3	3	0	0.1719397058		
LOHMEIER	1	1	0	0.09186120629		
SEMCARES	3	3	0	0.5112110186		
ARX	1	1	0	0.5076481612		
Enraf Nonius	2	2	0	0.2536979564		
IMED	2	2	0	0.2535096293		
Spacelabs	1	1	0	0.001879105951		
DELMAR REYNOLDS	2	2	0	0.04907959206		
PROMED	3	3	0	0.3397938262		
plusMED	2	2	0	0.2577872023		
HEILAND	2	2	0	0.004571593644		
WZHAN ZONACARE	1	1	0	0.005996009136		

Supplementary Table S10. Group-level performance results from leave-one-device-out validation of the final LightGBM C-strict model.

Device	N_records	N_PASS	N_FAIL	Mean_predicted_risk	PR_AUC	ROC_AUC

T2100	5	5	0	0.2466270887		
BENEHEART R3	29	29	0	0.1916328963		
ECG 600G	129	128	1	0.398483055	0.1111111111	0.96875
CARDIOVIT AT 101	12	12	0	0.2802771119		
SE-301	97	97	0	0.03241377745		
CARDIOTOUCH 3000	152	151	1	0.6608753716	0.008771929825	0.6225165563
CARDIOVIT AT 102 PLUS	3	3	0	0.1843266883		
CP 100	46	41	5	0.2563263734	0.5428571429	0.8829268293
MAC 400	62	56	6	0.4919703019	0.5517241379	0.7678571429
MAC 1200 ST	90	89	1	0.4568607862	1	1
MAC 600	212	209	3	0.4111178988	0.01139130785	0.341307815
SE-3	163	158	5	0.01030187744	0.03571428571	0.5727848101
SE3	11	11	0	0.03176170101		
MAC 1200ST	32	31	1	0.382429916	0.04166666667	0.6129032258
DELTA 3+	77	74	3	0.183333129	0.6893939394	0.9076576577
MAC 800	11	11	0	0.5001495717		
AS CARD	1	1	0	0.1186232896		
Cardiotouch 3000	54	52	2	0.7715351131	1	1
AT-102 G2	7	7	0	0.3499992249		
AT 102	65	64	1	0.295679324	0.5	0.9921875
AT-1	60	56	4	0.3750159704	0.1025641026	0.6696428571
AT-102	140	139	1	0.2649089926	0.01538461538	0.7661870504
iE 12A	5	5	0	0.02901699738		
CARDIO-M	6	5	1	0.07362528558	0.1666666667	0
AT 101	18	16	2	0.2491717659	0.625	0.90625
IE 6	22	22	0	0.02872509953		
BENEHEART R12	24	23	1	0.1023589549	1	1
IE6	30	30	0	0.02236191658		
CARDIMAX	3	3	0	0.3676413289		
300 C	4	4	0	0.3831821704		
ELECTROCARDIOGRAPH	1	1	0	0.09186120629		
AT-1 G2	59	59	0	0.4364621065		
SE 3	152	151	1	0.01610086951	0.007299270073	0.5463576159
P80	47	44	3	0.256942038	0.1578947368	0.8181818182
CARDIOSTAT 31	8	8	0	0.2548654384		
EC-3H	4	4	0	0.4071099478		
EC-12S	1	1	0	0.09186120629		
IE 3	29	29	0	0.02630091525		
BENEHEART D3	5	5	0	0.09006989054		
CARDIOVIT AT-1 G2	44	44	0	0.3340036245		
E6	23	23	0	0.3887690817		
MAC 2000	211	211	0	0.366443958		
AT 2 plus	4	4	0	0.2768700103		
SE-601	120	120	0	0.03628179834		
SE-1200 EXPRESS	50	50	0	0.03702696925		
BIOSET 3500	4	4	0	0.1502108555		
IE 12A	44	43	1	0.03143241763	0.03333333333	0.6511627907
ELI 150	14	14	0	0.2566094683		
SE-12	13	13	0	0.06425117721		
ASCARD GREEN	17	17	0	0.08939120397		
CARDIOGIMA 12M	8	8	0	0.3623317307		

CARDIOFAX GEM	18	18	0	0.4590737757		
CARDIOFAX S	9	9	0	0.3861536813		
ECG 600A	8	7	1	0.3110341552	0.25	0.7857142857
Cardiofax S	6	6	0	0.5496831161		
ECG 1200	8	8	0	0.664542239		
ECG 12	10	10	0	0.3834853042		
C500	1	1	0	0.5076481612		
ESAOTE P 80	2	2	0	0.6547483676		
CS-200	2	2	0	0.550545967		
AT-2 PLUS	23	23	0	0.3745410149		
AT 1	111	108	3	0.3391581409	0.1965174129	0.7916666667
EC-12R	3	3	0	0.3426840033		
CM 300	70	69	1	0.3833199304	1	1
MAC 500	70	63	7	0.3736953103	0.137254902	0.6507936508
IE12A	39	37	2	0.06850113097	0.6666666667	0.9864864865
HEARTSCREEN HS 60G	3	2	1	0.6695378647	1	1
ECG 100G	2	1	1	0.4376511932	1	1
ECG 300A	6	6	0	0.5356489312		
IE 12 A	11	11	0	0.01861939384		
CARDIOSPY EC-12H	1	1	0	0.09186120629		
SE 301	7	7	0	0.02932101764		
CARDIO M	28	28	0	0.07887296118		
CARDIOMAX FX-4010	10	10	0	0.3548882553		
CARDIOVIT MS-2010	3	3	0	0.3895467359		
SE 1200	12	12	0	0.036842418		
P 80	33	33	0	0.2926140478		
08SD	10	10	0	0.01296250406		
08-SD	4	4	0	0.0384873298		
SE-2012	2	2	0	0.03298254864		
SE-601B	7	7	0	0.03548578376		
Cardiovit AT-102	2	2	0	0.3076101247		
F3	1	1	0	0.0341060145		
08 MT	13	13	0	0.02551275551		
CARDIOVIT AT-102 G2	5	5	0	0.5563559905		
CARDIOVIT AT-1 G2	1	1	0	0.07512875413		
Delta 3+	17	17	0	0.2095070263		
9022K	4	4	0	0.4791103656		
08-MT	8	8	0	0.03337806909		
SCHILLER AT-1	3	3	0	0.3895467359		
CARDIO M+	16	16	0	0.1018467713		
CARDIO-M+	4	4	0	0.06653005793		
P8000	9	8	1	0.308013321	0.1428571429	0.625
BTL-08 LC ECG	2	2	0	0.02695237401		
B5 ASCARD	2	2	0	0.1219569537		
CARDIOVIT AT-101	19	18	1	0.2091929603	0.125	0.8055555556
CARDIOVIT AT-1	4	4	0	0.5546302456		
Cardiovit AT-1 G2	15	15	0	0.326503792		
MERCURIAL	10	10	0	0.122973027		
Cardiovit AT-101	4	4	0	0.6577314537		
AT-101	18	18	0	0.348051209		
CARDIOVIT AT-1	23	21	2	0.4477535302	0.1111111111	0.5952380952
ECG600G	5	5	0	0.02484227311		
ECG 1200A	16	15	1	0.4216071847	0.1	0.7
ECG600A	5	5	0	0.4568143267		
MT Plus ECG	2	2	0	0.05098964376		
ASCARD GRAY	1	1	0	0.1186232896		
08 MT PLUS	3	2	1	0.3419534309	1	1
ASCARD GREY	11	11	0	0.1001014611		

CORINA	2	2	0	0.5150250216		
CARDIOSOFT	1	1	0	0.5161681456		
Delta 3 Plus	4	3	1	0.1244845674	0.25	0.1666666667
BTL 08 MT PLUS ECG	2	2	0	0.05098964376		
BTL-08 SD ECG	13	13	0	0.02557379359		
mac 600	6	5	1	0.2362093446	0.2	0.6
AT 102 PLUS	9	9	0	0.2481585629		
AT-2	3	3	0	0.3895467359		
CARDIOCARE	9	8	1	0.5772447065	0.1666666667	0.6875
CARDIPIA 400	16	16	0	0.2693502862		
EK 53	3	3	0	0.5112110186		
ELI 250	14	14	0	0.3732862182		
AT-1	3	2	1	0.4905505646	0.3333333333	0.5
CM300	1	1	0	0.08252818656		
1112	1	1	0	0.5076481612		
EC-12H	3	3	0	0.2322222619		
BTL-08 MT PLUS ECG	3	3	0	0.03529800303		
CARDIOSMART	6	5	1	0.3463020368	1	1
MICROSMART	6	6	0	0.3567944336		
Cardiocare	2	2	0	0.5945129969		
DELTA 3 PLUS	61	59	2	0.1550739898	0.05882352941	0.7033898305
AT-102	1	1	0	0.5478738715		
GREY	8	8	0	0.1033744291		
iE 6	1	1	0	0.005426676834		
08 LT	8	8	0	0.02556829265		
HOLTER RECORDER	4	4	0	0.3016628807		
BIOCARE IE12A	4	4	0	0.02466937037		
BTL-08 LT	6	6	0	0.03663665129		
3-Channel ECH Holter Cardiospy	2	2	0	0.507843175		
3-Channel ECG Holter	10	10	0	0.4650922296		
SE-1200 Express	1	1	0	0.0341060145		
CARIOCARE	2	1	1	0.7794548422	0.5	0.5
ASCARD ORANGE	2	2	0	0.06707580804		
Cardio M Plus	15	15	0	0.2236949329		
SE 12	21	21	0	0.05875576532		
IE 300	2	2	0	0.01904363689		
PERSONAL 120	2	2	0	0.6547483676		
DELTA 1+	12	12	0	0.3333279284		
SCHILLER CARDIOVIT FT1	2	2	0	0.550545967		
CARDIMAX FX 4010	2	2	0	0.3284066305		
AT - 10 PLUS	2	2	0	0.550545967		
CARDIOMAX FX 4010	3	3	0	0.2150260304		
ECG 300G	5	5	0	0.1182423492		
CM 1200	2	2	0	0.507843175		
12A	10	10	0	0.0322052207		
IE3	3	3	0	0.02287340551		
Cardiocare	4	3	1	0.2301567381	1	1
IE-3	4	4	0	0.02747618254		
CARDISUNY ALFA 1000	4	4	0	0.005867885694		
MODEL 200	2	2	0	0.2577872023		
MAC 1600	2	2	0	0.04661374193		
SE 601	50	50	0	0.04328984713		
CARDIOFAX M	3	3	0	0.4230000778		
CARDIOSPY	2	2	0	0.302964961		
B5	6	6	0	0.1106416462		
1200A	6	6	0	0.6103875395		
08 SD	7	7	0	0.01589997036		
ASCARD DRAY	1	1	0	0.1186232896		
08-MD	2	2	0	0.05098964376		
ECG 1200 A	4	4	0	0.7218129771		

08 LC	6	6	0	0.03596339826		
MAC 1200	2	2	0	0.5150250216		
SEER 1000	1	1	0	0.09197132556		
AT 2+	2	1	1	0.770511277	1	1
TLC 5000	1	1	0	0.09186120629		
DELTA 3 +	2	2	0	0.3602965566		
Cardiofax C	4	4	0	0.4802514096		
CARDIOGIMA 12 M	1	1	0	0.4516639533		
Cardiofax	5	4	1	0.3963148312	0.25	0.625
E12A	1	1	0	0.4353709117		
SE-1202	1	1	0	0.005222385098		
ECG 300H	2	2	0	0.149970909		
ECG80A	2	1	1	0.06257891214	0.5	0.5
ELI 150C	1	1	0	0.1082630924		
ELI 250 C	1	1	0	0.06412308928		
EPG View	4	4	0	0.3832248281		
Mac 500	3	3	0	0.5183865276		
AT1	7	6	1	0.2824175491	0.25	0.75
AT-102	3	3	0	0.367908709		
PM-ECG-300G	3	3	0	0.2012560547		
600G	2	2	0	0.08912496844		
CARDIOVIT AT-10	2	2	0	0.550545967		
Cardipia 400	6	6	0	0.5160931568		
SE 1200 EXPRESS	9	9	0	0.03755201609		
AR600ADV	4	4	0	0.5880514726		
08 MT PLUS ECG	2	2	0	0.02934162008		
AT-2 Plus	1	1	0	0.5478738715		
CARDIPIA 800	7	6	1	0.4697434791	1	1
ECG 301	2	2	0	0.507843175		
Bioset 3500	4	2	2	0.1876364155	1	1
ECG300A	2	2	0	0.07170991866		
BTL-08LC	3	3	0	0.02064473129		
08-MT PLUS	3	3	0	0.03349220687		
CARDIO TOUCH 3000	2	2	0	0.5945129969		
SE 3E	3	3	0	0.02155029445		
CARDIOFAX-C 2150K	2	2	0	0.3151899696		
GREEN	9	9	0	0.09319294015		
CM 1200A	1	1	0	0.5076481612		
FX 2111	7	7	0	0.1601752244		
CARDIOMAX 2111	1	1	0	0.2991645282		
E30	2	2	0	0.02934162008		
Page Writer	1	0	1	0.485566622		
PAGE WRITER 2OOi	1	1	0	0.7591832387		
AT-102 plus	2	2	0	0.550545967		
PAGE WRITER 100	6	6	0	0.271310969		
Cardio M	5	5	0	0.05131210459		
MAC 200	3	3	0	0.3483161713		
AT-102 PLUS	2	2	0	0.276086935		
Delta 1	3	3	0	0.220269337		
ESAOTE	2	2	0	0.3467234049		
08 HOLTER	3	3	0	0.05106038897		
BIOSET 3700	6	6	0	0.1772176715		
CARDIOFAX	7	7	0	0.658094622		
CARDIOGIMA 12	16	16	0	0.2667736823		
MAC 5000	2	2	0	0.302219079		
CARDIPIA 200	1	1	0	0.2275205741		
ERGOLINE	1	1	0	0.09186120629		
CARDIOFAX 1250 K	3	3	0	0.6373410918		
CS-200 EXCELLENCE	2	2	0	0.3076101247		
LIFEPAK 1000	1	1	0	0.5076481612		
FX-3010	1	1	0	0.2991645282		
ECG-C06B	2	2	0	0.302964961		
ATRIA 6100	4	4	0	0.3831821704		
ECG 12 VIEW	6	6	0	0.4404661775		

BTL 08-LT ECG	1	1	0	0.05114571519		
AR 600	3	3	0	0.3397354912		
BIOSET 9000	3	3	0	0.3504598817		
AT - 1	3	3	0	0.1842090404		
CARDIOFAX C	1	1	0	0.1708624571		
ECG MODEL 1112L 250	3	3	0	0.3426840033		
Cardiofax Gem	1	1	0	0.01008978173		
CARDIOTOUCH, 3000	2	2	0	0.4361737464		
ZQ-1203G	2	1	1	0.2329341696	1	1
BTL 08 LC	3	3	0	0.03380160188		
IE-12A	3	3	0	0.03362803189		
CARDIOSTAT	5	5	0	0.002646103822		
Cadiotouch 3000	1	1	0	0.8411269568		
FX 4010	2	2	0	0.3284066305		
Microsmart	3	3	0	0.3427049418		
CardioSmart ST	1	1	0	0.5076481612		
EXCEL 103	3	3	0	0.1714312892		
BTL-08 MT PLUS	1	1	0	0.007314266589		
K130B	3	3	0	0.1719397058		
ECG - 9022K	5	4	1	0.1990037816	0.5	0.875
MAC600	1	1	0	0.007314266589		
PLUSMED	1	1	0	0.0771734371		
1200 ST	1	1	0	0.5161681456		
M607	1	1	0	0.09186120629		
AT 2 PLUS	3	3	0	0.1843266883		
CARDIOVIT AT-2 PLUS	3	3	0	0.3669603362		
SE-1200	1	1	0	0.0341060145		
CP 50	2	2	0	0.002638322129		
ATIG2	1	1	0	0.07512875413		
Cardimax FX 2111	3	3	0	0.006044322564		
3-Channel ECK Holter	1	1	0	0.5076481612		
CP100	4	4	0	0.3547066275		
ECG 80A	2	2	0	0.6614401274		
CARDIOTOUCH3000	1	1	0	0.8411269568		
PAGE WRITER 200i	9	9	0	0.2984462488		
PAGE WRITET 100	3	3	0	0.5318820306		
CUSTO NORM	1	1	0	0.5076481612		
ECG 600 G	1	1	0	0.0771734371		
CARDIOMAX - FX3010	1	1	0	0.2991645282		
08LT	1	1	0	0.001612015512		
08MT	2	2	0	0.001930020584		
AT - 102	3	3	0	0.1841205786		
600 B	1	1	0	0.001364258722		
Cardiofax gem	3	3	0	0.4277360598		
ZQ-1201G	2	2	0	0.2543171199		
EKG 503D	2	2	0	0.2536979564		
CARDIAX PC ECG	2	2	0	0.2535096293		
HELLIGE	2	2	0	0.005514788251		
VOYAGER V12C	1	1	0	0.001879105951		
Atria 6100	1	1	0	0.5076481612		
NIHON KOHDEN	1	1	0	0.008385116834		
ELI 320	1	1	0	0.4349333763		
VOAGER 12	2	2	0	0.04907959206		
DPS VISIO	1	0	1	0.9658622653		
ECG300G	1	1	0	0.001189766329		
T12	3	3	0	0.3397938262		
CARDIOFAX ECG 9022K	2	2	0	0.3198037486		
DELTA 3	2	2	0	0.1836357658		
HEARTSCREEN	3	3	0	0.339927869		
Delta 1 plus	3	3	0	0.1218053638		
IE12E	2	2	0	0.01761186841		
AT 101	2	2	0	0.276086935		
CARDIOVIT CS-2000	2	2	0	0.550545967		
08LC	1	1	0	0.001079034611		

mac 800	1	1	0	0.5161681456		
AT 1	1	1	0	0.5478738715		
AT-10 PLUS	1	1	0	0.5478738715		
MAC 2000	1	1	0	0.09197132556		
BTL-08 SD3	1	1	0	0.007314266589		
AT 2	2	2	0	0.550545967		
BIOMEDICA	2	2	0	0.3429684743		
CARDIOVIT CS 200 TOUCH	2	2	0	0.2779857219		
CARDIOVIT AT-1	3	3	0	0.5532198028		
CARDIOTOUCH 2000	1	1	0	0.8411269568		
CARDIOVIT AT-102	2	2	0	0.276086935		
CARDIO M-PLUS	1	1	0	0.08602329574		
M1770A	1	1	0	0.7591832387		
AT-104PC	1	1	0	0.07512875413		
CARDIOFAX Q	1	1	0	0.002223730122		
TC 30	1	1	0	0.00636218524		
EM-601	5	5	0	0.1740489748		
EM-1201	1	1	0	0.09186120629		
EM-301	3	3	0	0.08809906879		
Megacart	1	1	0	0.001879105951		
CARDIOTOUCH	2	2	0	0.4287260707		
DELTA 1 PLUS	1	1	0	0.000675887885		
SE 301	1	1	0	0.0341060145		
SICARD 600	1	1	0	0.5076481612		
ELI250	1	1	0	0.003325772233		
SICARD 460	1	1	0	0.5076481612		
EDAN SE 3	1	1	0	0.0341060145		
3 M	1	1	0	0.4516639533		
COMPACT 9	1	1	0	0.003129044452		
Heart Mirror	1	1	0	0.001214772454		
Deadalus	1	1	0	0.03112522837		
AT 10+	1	1	0	0.001137312565		
CARDIO M +	1	1	0	0.08602329574		
AT 102+	1	1	0	0.001351507197		
Cardipia 200	1	0	1	0.4840435558		
ASCARD	1	1	0	0.1186232896		
AR1200ADU	1	1	0	0.009024255378		

Supplementary Table S1. Ten-bin calibration results for the final LightGBM C-strict model.

Bin	N	Mean_predicted_probability	Observed_FAIL_rate
1	2240	0.02185635101	0.00625
2	243	0.1333637217	0.02057613169
3	304	0.2762388514	0.01644736842
4	163	0.3540947064	0.01226993865
5	158	0.4633984029	0.01898734177
6	242	0.5482573688	0.03719008264
7	312	0.6512435789	0.02243589744
8	107	0.7596369849	0.02803738318
9	147	0.8596810928	0.1292517007
10	40	0.9530426086	0.475

Note: Predicted probabilities were divided into 10 equally spaced bins. Mean predicted probability represents the average model-estimated probability of verification failure within each bin, whereas observed FAIL rate represents the empirical proportion of failing records. Bins with no records are reported as not applicable. Expected calibration error was 0.200 and the Brier score was 0.122.

The reliability diagram confirms moderate calibration. The discrepancy between predicted and observed failure rates supports use of the model as a triage tool rather than a stand-alone probabilistic decision rule.

Supplementary Table S1. Calibration summary for the final LightGBM model.

Calibration metric	**Value**
Brier score	0.122
Expected calibration error, 10 bins	0.200
Interpretation	Moderate calibration; probabilities should be interpreted as relative risk scores