

DOI: 10.37943/WWSQ922

**Dias Konyspayev**

Master's Student, Department of Information Systems  
knspsvd@gmail.com, orcid.org/0009-0006-0332-6200  
Akhmet Baitursynuly Kostanay Regional University, Kazakhstan

**Batyrbkhan Kuzenbayev**

PhD, Department of Information Systems  
bekz@bk.ru, orcid.org/0000-0002-5680-9912  
Akhmet Baitursynuly Kostanay Regional University, Kazakhstan

**Nurlybay Abatov**

Candidate of Physical and Mathematical Sciences, Associate Professor,  
Department of Information Systems  
abatov.n57@gmail.com, orcid.org/0009-0008-0001-7693  
Akhmet Baitursynuly Kostanay Regional University, Kazakhstan

**Dinara Alipayeva**

Master's Degree, Department of Information Systems  
alippaewa@gmail.com, orcid.org/0009-0000-9024-0937  
Akhmet Baitursynuly Kostanay Regional University, Kazakhstan

**RELIABILITY-AWARE ARTIFICIAL INTELLIGENCE FOR MULTILABEL  
ELECTROCARDIOGRAM DIAGNOSIS**

**Abstract:** Artificial intelligence can classify electrocardiograms, but high discrimination alone does not show whether probabilities, decisions, and explanations remain reliable under technical signal failures. This study evaluates an integrated reliability-aware framework for multi-label electrocardiogram classification on PTB-XL with five diagnostic superclasses: normal, myocardial infarction, ST and T wave change, conduction disturbance, and hypertrophy. Patient-disjoint folds 1–8, 9, and 10 were used for training, validation, and held-out testing, respectively. XGBoost and three one-dimensional deep-learning architectures were compared, and ResNet1D was selected by validation macro area under the receiver operating characteristic curve. On the held-out test fold, ResNet1D achieved a macro area under the receiver operating characteristic curve of 0.9110, macro area under the precision-recall curve of 0.7746, and macro F1 score of 0.7051. Isotonic calibration, fitted on validation predictions only, produced a held-out macro Brier score of 0.08523 and adaptive calibration error of 0.00053. Selective prediction showed a coverage-risk trade-off: higher coverage reduced abstention but increased accepted-case risk and expected harm. Robustness tests covered Gaussian noise, baseline wander, amplitude scaling, lead dropout, and partial waveform failures; grouped limb- and precordial-lead dropout caused the largest discrimination losses. A complete-missing-lead input gate rejected zeroed or flat leads and reduced expected harm under complete lead loss, but did not detect the tested partial waveform failures. Attribution stability was high under Gaussian noise and lower under lead dropout. External analysis on Georgia and Chapman-Shaoxing data was limited to a binary abnormal-versus-normal sanity check and showed source-dependent calibration transfer. The contribution is a reproducible reliability evaluation framework, not a new architecture, full external validation, or a clinical deployment claim.

**Keywords:** artificial intelligence, electrocardiogram classification, multi-label classification, probability calibration, selective prediction, robustness, attribution stability, medical decision support

## Introduction

An electrocardiogram is one of the most widely used non-invasive tests for cardiovascular assessment. A standard twelve-lead recording captures the electrical activity of the heart over a short time interval and provides several synchronized views of the same cardiac process. From a machine learning perspective, this is not a simple single-label task. One recording may show no abnormality, one diagnostic superclass, or several coexisting abnormalities. The PTB-XL database makes this setting reproducible by providing a large collection of clinical twelve-lead electrocardiograms with diagnostic labels and recommended folds for algorithm comparison [1].

This study is situated at the intersection of medical artificial intelligence, information systems, and biomedical signal analysis. Its scientific contribution is not a new neural-network architecture or a claim of state-of-the-art electrocardiogram classification. Instead, the contribution is an integrated reliability-aware evaluation framework that separately examines clean discrimination, validation-only probability calibration, selective prediction, electrocardiogram-specific stress testing, complete-missing-lead detection, attribution stability, and a bounded external binary sanity check. This structure makes explicit which reliability properties are supported by the data and which remain limitations.

Electrocardiogram classification is suitable for this analysis because the signals are structured, multi-lead, and vulnerable to technically meaningful failures such as missing or flat leads. A high discrimination score on clean data is therefore insufficient: it does not establish that predicted probabilities are calibrated, that unreliable inputs are rejected, or that the system remains appropriate for use after signal corruption. The present study evaluates these properties separately rather than treating them as interchangeable evidence of reliability.

## Literature Review and Problem Statement

PTB-XL has become an important benchmark for electrocardiogram machine learning because it provides standardized 12-lead recordings, diagnostic labels, and patient-disjoint folds [1]. Deep-learning benchmarking on PTB-XL has reported macro area under the receiver operating characteristic curve values of approximately 0.92–0.93 for related superdiagnostic tasks [2]. Artificial intelligence-enhanced electrocardiography has also been reviewed across cardiovascular applications [3]. A recent state-of-the-art review further summarizes current artificial intelligence approaches to electrocardiogram interpretation [4]. The held-out macro area under the receiver operating characteristic curve of 0.9110 in the present study is therefore competitive but below the strongest reported benchmark values. It is not presented as a direct leaderboard comparison or a state-of-the-art claim.

Recent work has emphasized that discrimination metrics alone do not describe the reliability of medical artificial intelligence systems. Calibration assesses whether predicted probabilities correspond to observed outcomes and is an important component of predictive model evaluation [5]. Patient-level predictive uncertainty should also be communicated in clinical artificial intelligence systems [6]. Uncertainty quantification has therefore become an important area of healthcare artificial intelligence research [7]. Reliability assessment can additionally include systematic algorithmic auditing [8]. Staged evaluation has been proposed as a bridge between technical development and clinical use [9].

Conformal prediction provides a formal approach to uncertainty estimation in clinical prediction [10]. Cost-aware selective prediction has also been proposed for clinical triage under distribution shift [11]. Electrocardiogram-specific robustness remains important because technical disturbances such as baseline wander can alter the signal [12]. A systematic review of 230 studies summarizes deep-learning architectures, electrocardiogram data sources, preprocessing methods, and data-splitting strategies used across the field [13]. Attribution methods such as Integrated Gradients provide a way to inspect changes in model sensitivity under stress, but attribution stability does not demonstrate clinical reasoning or diagnostic validity. The problem addressed here is therefore how to report a transparent reliability contour around a multi-label electrocardiogram classifier without overstating its readiness for clinical deployment.

This study evaluates calibration, selective prediction, electrocardiogram-specific perturbation testing, and attribution analysis within one patient-disjoint PTB-XL protocol and reports their interaction with a complete-missing-lead abstention gate. General transparency elements were informed by CLAIM [14]. Prediction-model reporting was informed by TRIPOD+AI [15]. Early-stage clinical artificial intelligence evaluation principles were considered with reference to DECIDE-AI [16].

### **Purpose and Objectives**

The goal of this research project was to analyze a reliability-focused artificial intelligence system for multilabel electrocardiogram classification. The study hypothesized that reliability characteristics of the system, as well as its failures, would become apparent from the multilayered analysis but not from the standard evaluation framework based solely on discrimination measures.

The objectives were to: (1) define a five-superclass PTB-XL multilabel classification problem; (2) compare different model architectures using a validation-based model selection method; (3) calibrate probability predictions; (4) analyze selective operating points; (5) consider specific electrocardiogram perturbations; (6) analyze a quality-gated abstention strategy; (7) conduct a restricted external binary sanity check; and (8) analyze attribution stability of accepted predictions.

In contrast to a clinical validation trial, the project was designed as an experimental study in the applied information technology domain. It used a fixed public dataset, a validation-based model selection scheme, and a held-out test partition reserved for final analysis. There was no claim regarding the superior performance of any model, only that reliability analysis techniques could bring to light clinically relevant failure characteristics of the models.

### **Methods and Materials**

The internal dataset was PTB-XL version 1.0.3, consisting of 12-lead, 10-second electrocardiogram recordings sampled at 100 Hz [1]. Each record was represented as a  $12 \times 1000$  tensor. Records without any target superclass label were excluded, leaving 21,388 records. The five target superclasses were NORM, MI, STTC, CD, and HYP. PTB-XL folds 1–8 were used for training, fold 9 for validation, and fold 10 for held-out testing; these folds are patient-disjoint. Each lead was independently standardized using the mean and standard deviation estimated from the training data, with the same training-derived parameters applied to validation and test records.

No new interventional data were collected. Instead, the analysis used publicly available retrospective datasets and their existing annotations. This is crucial for interpreting the findings presented here: model performance on benchmark data is reproducible and can be verified, but claims about prospective clinical workflow, patient outcomes, or physician acceptance are outside the scope of this paper.

The study compared one classical baseline and three neural architectures. XGBoost did not receive raw waveform samples. Each 12-lead record was converted into 108 handcrafted features: mean, standard deviation, minimum, maximum, range, mean square, skewness, kurtosis, and zero-crossing rate for each lead [17]. The neural candidates received the normalized  $12 \times 1000$  waveform tensor directly. InceptionTime was included as one candidate architecture [18]. A temporal convolutional network was included as another [19]. ResNet1D formed the third neural candidate. XGBoost was a fixed classical baseline; the three neural models formed the model-selection pool.

Neural models were trained with AdamW optimization, an initial learning rate of 0.001, weight decay of 0.0001, batch size 64, maximum 20 epochs, gradient clipping at 1.0, and early stopping with patience 5. The selected ResNet1D used hidden dimension 128 and dropout 0.20. Training used balanced positive-class weighting, mixed-precision CUDA computation, and random seed 42. XGBoost used 300 estimators, maximum depth 6, learning rate 0.05, subsampling 0.80, column subsampling 0.80, regularization parameter lambda 1.0, and histogram tree construction.

Table 1. Validation performance of the three neural candidates used for model selection

| Model                          | Validation macro AUROC | Validation macro AUPRC |
|--------------------------------|------------------------|------------------------|
| ResNet1D                       | 0.9143                 | 0.7773                 |
| Temporal convolutional network | 0.9129                 | 0.7746                 |
| InceptionTime                  | 0.9101                 | 0.7716                 |

ResNet1D was selected because it achieved the highest validation macro area under the receiver operating characteristic curve, 0.914271. Validation macro area under the precision-recall curve was used only as the secondary selection criterion. XGBoost was evaluated as a separate classical baseline and was not omitted from the study; it was not part of the neural model-selection table.

Because the task is multi-label, each superclass was treated as a one-versus-rest output. The main internal discrimination metrics were macro AUROC, macro AUPRC, and macro F1 score. Macro F1 was calculated at a fixed probability threshold of 0.50 for every superclass; this threshold was not optimized on the test fold. Macro averaging was used because the label distribution was imbalanced, as shown in Figure 1. To compare the leading neural models, paired non-parametric bootstrap confidence intervals were calculated from matched fold-10 predictions using 2,000 resamples and random seed 42.

#### PTB-XL diagnostic superclass distribution

Task filtering retained 21,388 recordings with a visibly imbalanced label structure.

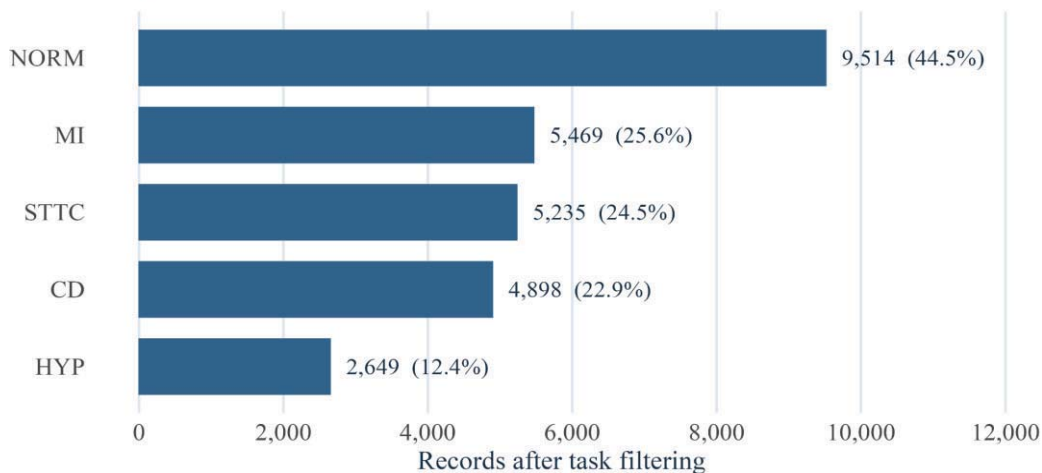


Figure 1 – PTB-XL diagnostic superclass distribution after task filtering

Scalar temperature scaling, vector temperature scaling, Platt scaling, and isotonic regression were fitted separately for each of the five classes using only fold-9 validation predictions and labels [20]. Calibration methods were selected using validation-fold calibration metrics. The selected isotonic calibrators were then applied without refitting to fold-10 predictions, where Brier score and adaptive expected calibration error were calculated. Fold-10 labels were never used for fitting or selecting a calibration method.

Decision-level evaluation used a weighted realized-harm function. False negatives were assigned weights of 1.0 for NORM, 2.5 for MI, and 2.0 for STTC, CD, and HYP. False positives were assigned a weight of 1.0 for every class. The abstention cost was fixed at 0.5. These values represent an experimental decision model that emphasizes missed abnormalities; they are not clinically validated costs:

$$\bar{H} = \frac{1}{N} \sum_{i=1}^N \left[ a_i \sum_{c=1}^5 (w_c^{FN} y_{ic} (1 - d_{ic}) + w_c^{FP} (1 - y_{ic}) d_{ic}) + (1 - a_i) C_{abs} \right] \quad (1)$$

Here,  $N$  is the number of records,  $i$  indexes records,  $c$  indexes the five classes;  $y_{ic}$  is the observed binary label,  $d_{ic}$  is the predicted binary decision,  $a_i=1$  for an accepted record and  $a_i=0$  for an abstained record,  $w_c^{FN}$  and  $w_c^{FP}$  are class-specific false-negative and false-positive weights; and  $C_{abs}$  is the abstention cost. Accepted risk was defined as the mean realized weighted misclassification loss among accepted records,  $\frac{\sum_i a_i L_i}{\sum_i a_i}$ , rather than as macro F1.

For each record, uncertainty was calculated as  $u_i = \sum_{c=1}^5 \min[(1 - p_{ic})w_c^{FP}, p_{ic}w_c^{FN}]$ , where  $p_{ic}$  is the calibrated probability for class  $c$ . A record was accepted when  $u_i$  did not exceed the validation-derived threshold for the requested coverage level and was otherwise abstained. Thresholds for target coverages of 0.90, 0.95, and 0.99 were selected using fold 9 only and applied unchanged to fold 10.

Robustness was evaluated on the full held-out test fold using fixed electrocardiogram-specific perturbations. Gaussian noise had standard deviation 0.02 in the normalized signal scale. Baseline wander was a 0.33 Hz sinusoid with amplitude 0.05. Global amplitude scaling used factors from 0.90 to 1.10. Single-lead dropout set lead I to zero; grouped limb dropout set leads I, II, III, aVR, aVL, and aVF to zero; grouped precordial dropout set V1–V6 to zero. Additional severity tests used random dropout of one, two, or three leads; a 50% flatline segment beginning at 25% of lead I; low-amplitude lead I scaled by 0.05; and a 30% segment dropout beginning at 35% of lead I. Perturbations were fixed before full-test evaluation.

The input gate was a complete-missing-lead detector, not a general electrocardiogram-quality algorithm. A record was rejected when at least one lead had a zero fraction of at least 0.98 or a lead standard deviation no greater than  $10^{-6}$ . Because the maximum permitted number of missing leads was zero, the gate reliably detected complete zeroed or flat leads but was not designed to detect all partial waveform failures, noise, baseline wander, or low-amplitude signals. Attribution analysis used Integrated Gradients for model inspection rather than clinical explanation validation [21–22].

The external analysis was a bounded binary abnormal-versus-normal sanity check using the Georgia and Chapman-Shaoxing datasets [23–24]. It did not evaluate the five PTB-XL superclasses. The primary abnormal score was  $p_{\text{abnormal}} = 1 - p_{\text{NORM}}$ , where  $p_{\text{NORM}}$  is the calibrated probability of the PTB-XL NORM class. Alternative abnormal scores based on the maximum and mean probabilities of the four non-NORM classes were reported only as sensitivity analyses. Sourcewise discrimination and calibration results were reported to expose transfer limitations rather than to claim clinical generalization.

## Results

On the held-out test fold, ResNet1D achieved macro AUROC 0.911006, macro AUPRC 0.774559, and macro F1 0.705120. The corresponding values for TCN were 0.910393, 0.775106, and 0.705802, and those for InceptionTime were 0.908244, 0.774966, and 0.704898. XGBoost was lower at 0.819511 macro AUROC, 0.623602 macro AUPRC, and 0.482811 macro F1. Relative to TCN, the ResNet1D differences were 0.000613 for macro AUROC (95% paired bootstrap interval: -0.003160 to 0.004178), -0.000548 for macro AUPRC (-0.008793 to 0.008942), and -0.000682 for macro F1 (-0.009891 to 0.008088). Relative to InceptionTime, the corresponding differences were 0.002762 (-0.001220 to 0.006635), -0.000407 (-0.009130 to 0.008957), and 0.000222 (-0.009835 to 0.010563). The intervals include zero; therefore, the leading deep models should be described as closely matched rather than statistically separated.

Per-class fold-10 AUROC/AUPRC values for ResNet1D were 0.9423/0.9184 for NORM, 0.9234/0.8226 for MI, 0.9316/0.8111 for STTC, 0.9151/0.8328 for CD, and 0.8426/0.4880 for HYP. At the fixed 0.50 threshold, class-wise F1 scores were 0.8490, 0.7436, 0.7676, 0.7280, and 0.4374, respectively. The lower HYP performance reflects its lower prevalence and should be treated as a limitation.

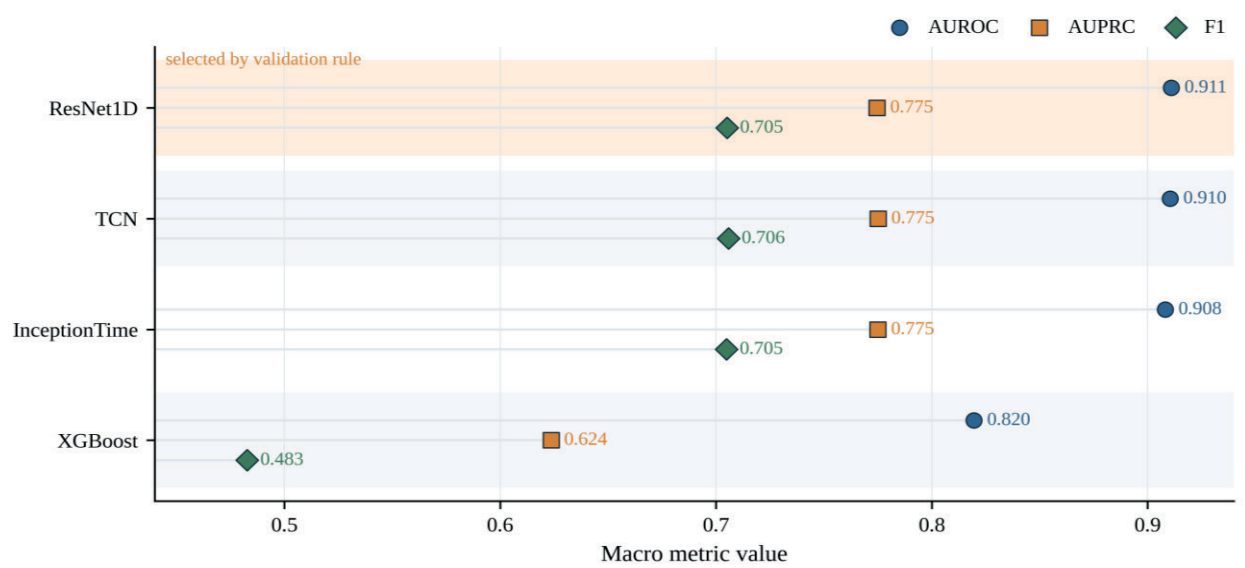


Figure 2 – Internal model comparison on the held-out PTB-XL test fold

The central contribution is the integration of reliability analyses around a conventional electrocardiogram classifier. ResNet1D was selected by the validation macro AUROC rule, but the held-out discrimination differences among the leading deep models were small and their paired bootstrap intervals included zero. The study therefore does not make a backbone-superiority claim.

Table 2. Calibration methods and held-out reliability metrics for ResNet1D

| Calibration method         | Macro Brier score | Macro adaptive calibration error |
|----------------------------|-------------------|----------------------------------|
| Isotonic calibration       | 0.08523           | 0.00053                          |
| Platt scaling              | 0.08781           | 0.01906                          |
| Scalar temperature scaling | 0.10609           | 0.09704                          |
| Vector temperature scaling | 0.10582           | 0.09354                          |

Figure 3 shows the pooled one-versus-rest calibration curve before and after isotonic calibration. The calibrated curve follows the ideal diagonal much more closely than the raw model curve. This supports a probability reliability claim, not a claim that the model became robust to all signal failures.

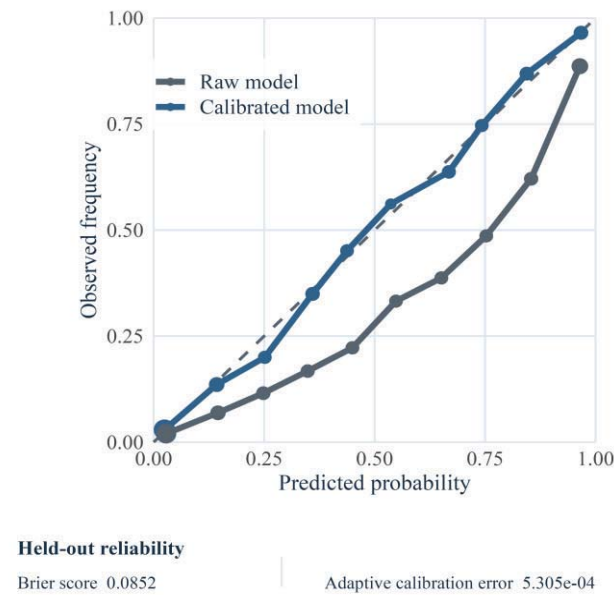


Figure 3 – Calibration curve of the selected ResNet1D model before and after isotonic calibration

Selective prediction produced a consistent coverage-risk pattern. At target coverage 0.90, achieved coverage was 0.8860, abstention was 0.1140, accepted risk was 0.7612 and expected harm was 0.7315. At target coverage 0.95, the corresponding values were 0.9402, 0.0598, 0.8110 and 0.7924. At target coverage 0.99, they were 0.9940, 0.0060, 0.8732 and 0.8709. The operating points are shown in Figure 4.

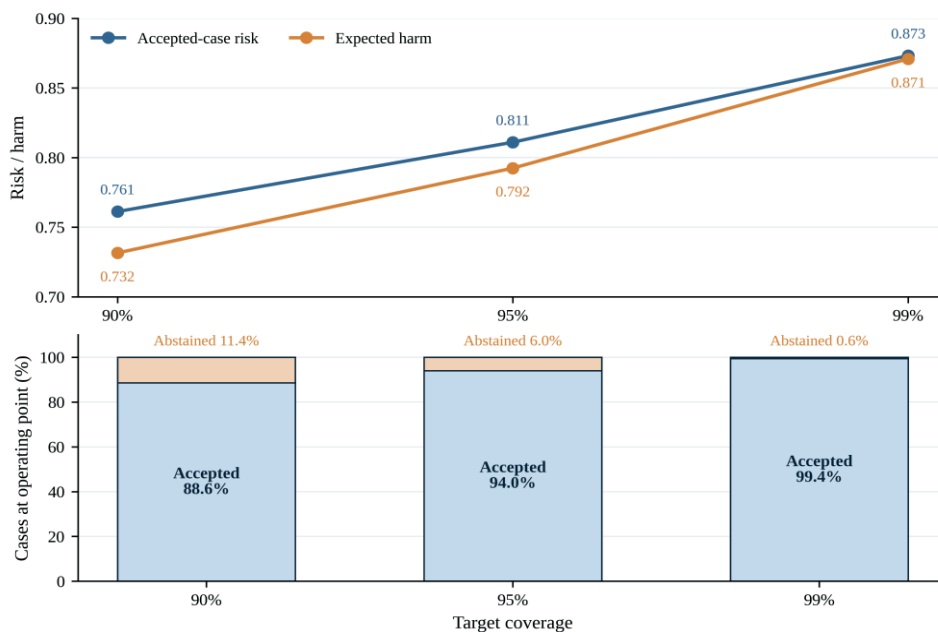


Figure 4 – Selective decision operating points across target coverage levels

Robustness evaluation showed a clear difference between mild perturbations and complete lead-loss perturbations. Gaussian noise and baseline wander were approximately neutral. Single-lead dropout reduced macro AUROC by 0.0057 and macro AUPRC by 0.0107, while increasing

macro ECE by 0.0206 and macro Brier score by 0.0087. Grouped limb-lead dropout was worse, with macro AUROC change of -0.0536 and macro AUPRC change of -0.0846. Grouped precordial-lead dropout similarly reduced macro AUROC by 0.0417 and macro AUPRC by 0.0710. These deltas are summarized in Figure 5.

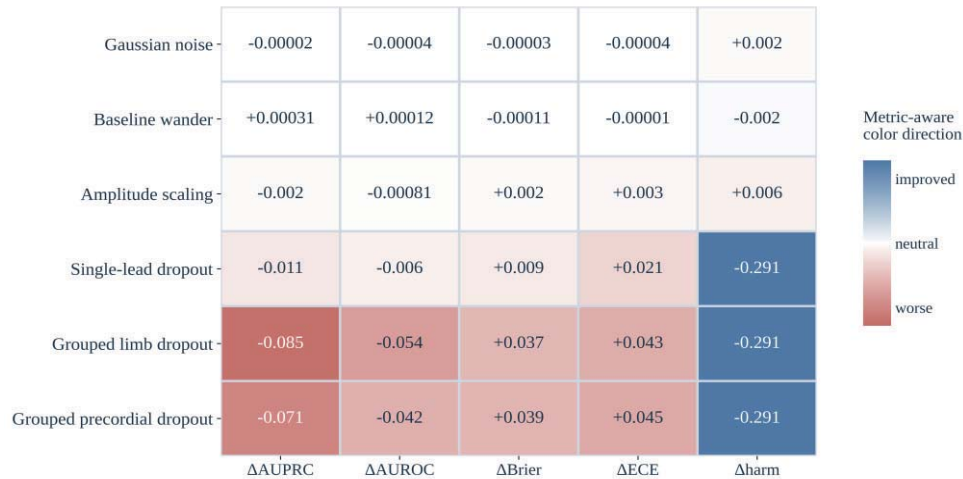


Figure 5 – Robustness under electrocardiogram-specific perturbations

The quality gate behaved as intended for complete missing-lead failures and as expected for perturbations it was not designed to detect. Its rejection rate was 0.0 for clean recordings, Gaussian noise, baseline wander and amplitude scaling. The rejection rate was 1.0 for single-lead dropout, grouped limb dropout and grouped precordial dropout. Table 3 shows the gate-versus-no-gate expected-harm differences with bootstrap intervals. Negative values indicate lower expected harm with the quality gate. The effect was supported for complete lead-loss perturbations, while partial flatline, low-amplitude lead and segment dropout had neutral intervals because the current gate did not detect them.

Table 3. Summary of quality-gated decision effect under missing-lead and partial waveform failures

| Perturbation                    | Mean harm difference | Bootstrap 95% interval | Gate rejection rate |
|---------------------------------|----------------------|------------------------|---------------------|
| Single lead dropout             | -0.3254              | [-0.3729, -0.2719]     | 1.0                 |
| Random one lead dropout         | -0.3230              | [-0.3729, -0.2715]     | 1.0                 |
| Random two lead dropout         | -0.3724              | [-0.4194, -0.3171]     | 1.0                 |
| Random three lead dropout       | -0.3909              | [-0.4348, -0.3368]     | 1.0                 |
| Grouped limb lead dropout       | -0.4298              | [-0.4860, -0.3747]     | 1.0                 |
| Grouped precordial lead dropout | -0.3266              | [-0.3657, -0.2849]     | 1.0                 |
| Partial flatline                | 0.0000               | [0.0000, 0.0000]       | 0.0                 |
| Low amplitude lead              | 0.0000               | [0.0000, 0.0000]       | 0.0                 |
| Segment dropout                 | 0.0000               | [0.0000, 0.0000]       | 0.0                 |

A qualitative electrocardiogram example was inspected to verify that the accepted-case workflow was not based only on aggregate metrics. Figure 6 shows a representative accepted abnormal PTB-XL recording with four leads. The figure is not used as diagnostic evidence by itself; it is included to show the type of waveform used by the pipeline.

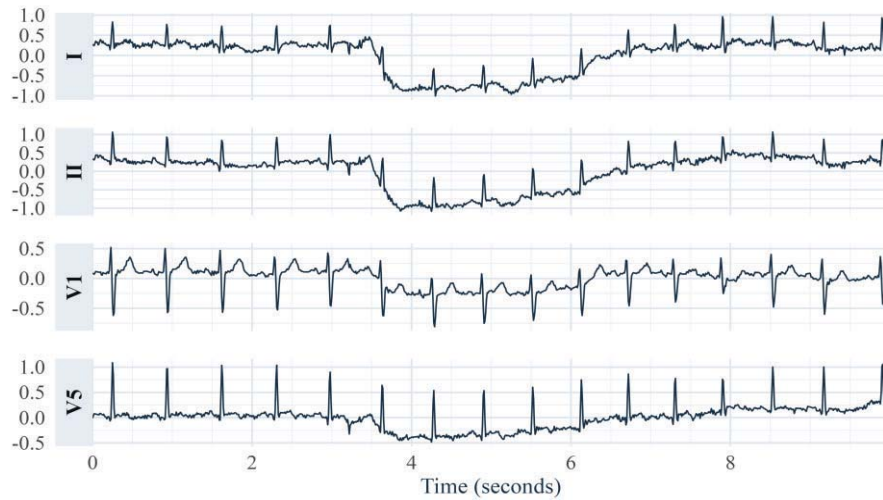


Figure 6 – Representative accepted PTB-XL electrocardiogram used for qualitative inspection

Attribution stability was high under Gaussian noise and lower under lead dropout. The mean cosine similarity between Integrated Gradients and Gradient-by-Input was 0.8509. The mean Integrated Gradients similarity under Gaussian noise was 0.9989, whereas the mean similarity under lead dropout was 0.8896. Figure 7 shows the distributions for accepted decisions. This result is a technical stability result only; it does not validate that the highlighted waveform regions correspond to a clinician's reasoning process.

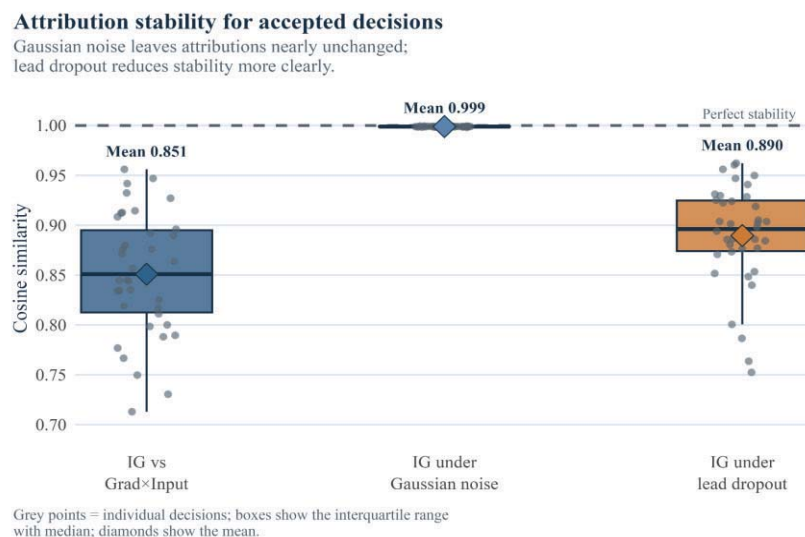


Figure 7 – Attribution stability for accepted decisions

The final analysis was the external binary sanity check. Overall abnormal-versus-normal scoring with `p_abnormal_not_norm` gave AUROC of 0.8141, AUPRC of 0.9575, Brier score of 0.2071 and calibration error of 0.2182. Source-wise results showed Georgia AUROC 0.8207, AUPRC 0.9490, Brier score 0.1948 and calibration error 0.2312. Chapman-Shaoxing showed AUROC 0.8039, AUPRC 0.9628, Brier score 0.3200 and calibration error 0.3933. Figure 8 shows these differences by source. The high AUPRC should be interpreted with caution because the abnormal class rate affects precision-recall metrics.

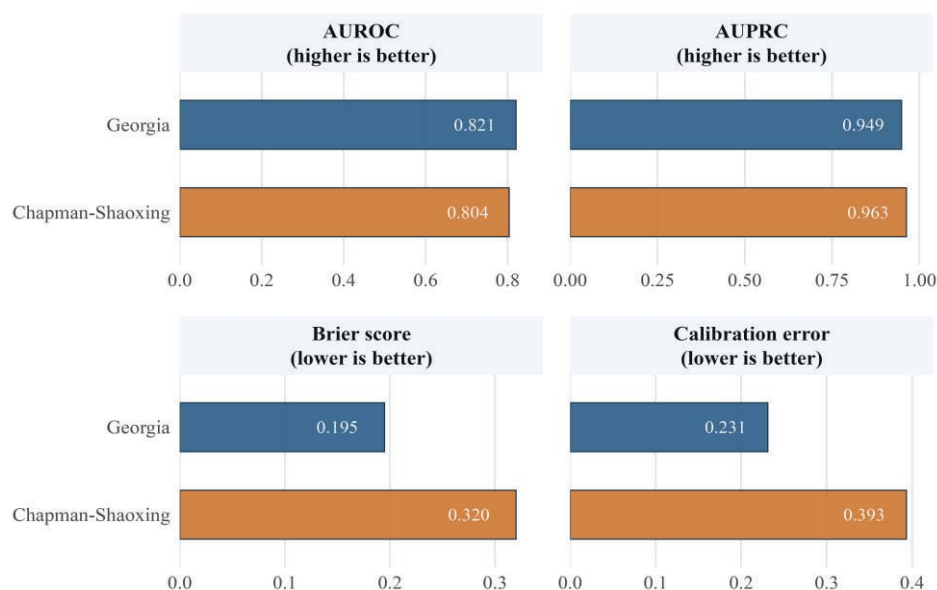


Figure 8 – External binary abnormal-versus-normal sanity check by source

## Discussion

The main finding is not that one architecture clearly dominates the others. The internal classification results show strong performance for all three deep time-series models, with only small differences between ResNet1D, the temporal convolutional network and InceptionTime. ResNet1D was a valid selected model because it followed the validation rule, but the held-out results do not justify a strong architecture-superiority claim. The contribution of the study is the reliability-aware evaluation contour around the classifier.

Calibration was the cleanest positive layer. Isotonic calibration substantially improved probability reliability on the held-out internal data. This matters because downstream decision layers use probabilities, not only ranks. However, calibration should not be confused with robustness. The grouped lead-dropout results show that calibrated probabilities still face discrimination loss and reliability drift when technically important leads are missing.

Selective prediction gave a predictable but important operating-point pattern. Higher coverage reduced abstention, but it increased accepted-case risk and expected harm. This means that a diagnostic support system cannot be described by one threshold without specifying the operating point. For safety-relevant use, the coverage level has to be reported together with the harm definition and the abstention rule.

The most safety-relevant result was the quality-gated abstention behavior under complete missing-lead failures. The quality gate rejected complete single-lead, grouped limb and grouped precordial dropout cases and reduced expected harm relative to the same selective policy without the gate. At the same time, the result is intentionally limited. The gate did not detect partial flatline, low-amplitude lead or segment dropout in the current configuration. A broad statement that all invalid electrocardiograms are handled would therefore be incorrect.

The external results should be interpreted as a bounded sourcewise binary sanity check. Discrimination and calibration differed between the Georgia and Chapman-Shaoxing sources, with poorer calibration transfer for Chapman-Shaoxing. These findings do not validate the five PTB-XL diagnostic superclasses and cannot support claims of clinical generalization or deployment readiness.

The attribution analysis adds another technical reliability view, but it does not solve interpretability. The high similarity under Gaussian noise suggests that attributions are stable under a mild waveform perturbation. The lower similarity under lead dropout is consistent with the

robustness results, because missing leads change the available signal. These observations are useful for model debugging, yet they do not prove clinical explanation quality. Clinical interpretability would require separate expert review and task-specific agreement analysis.

Several limitations follow directly from the experiment. The dataset task is public and reproducible, but it remains retrospective. The models were not tested in a prospective clinical workflow. The decision policy uses a defined harm proxy, not patient outcome data. The quality gate is narrow and detects complete missing-lead patterns, not all signal-quality defects. The external evaluation is binary and source-wise, not a full multilabel external study. These limitations do not invalidate the internal reliability results, but they define the boundary of the claims.

### Conclusion

A reliability-aware artificial intelligence pipeline was evaluated for multilabel electrocardiogram diagnosis on PTB-XL. The internal classification performance was strong, but the model comparison also showed that the top deep models were close. The strongest positive results were the held-out probability reliability after isotonic calibration and the quality-gated reduction of expected harm under complete missing-lead failures. The study also showed why discrimination metrics alone are insufficient: grouped lead dropout remained a raw-model failure mode, higher coverage increased accepted risk, calibration transfer was weak in the external binary check, and attribution stability remained a technical rather than clinical explanation measure. The system should therefore be viewed as a reproducible information technology evaluation contour for safer diagnostic decision support research, not as a deployment-ready clinical system.

### Acknowledgments

The authors declare that no external funding was received for this study.

### Data Availability

PTB-XL is publicly available through PhysioNet at <https://physionet.org/content/ptb-xl/1.0.3/> [1]. The Georgia data used for the bounded external binary sanity check are publicly available through the PhysioNet/Computing in Cardiology Challenge 2020 source [23]. The Chapman-Shaoxing data are publicly available through the corresponding database publication [24]. No private patient data were used.

### References

- [1] Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., & Schaeffter, T. (2020). PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1), 154. <https://doi.org/10.1038/s41597-020-0495-6>
- [2] Strodthoff, N., Wagner, P., Schaeffter, T., & Samek, W. (2021). Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics*, 25(5), 1519–1528. <https://doi.org/10.1109/JBHI.2020.3022989>
- [3] Siontis, K. C., Noseworthy, P. A., Attia, Z. I., & Friedman, P. A. (2021). Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7), 465–478. <https://doi.org/10.1038/s41569-020-00503-2>
- [4] Ose, B., Sattar, Z., Gupta, A., Toquica, C., Harvey, C., & Noheria, A. (2024). Artificial intelligence interpretation of the electrocardiogram: A state-of-the-art review. *Current Cardiology Reports*, 26(6), 561–580. <https://doi.org/10.1007/s11886-024-02062-1>
- [5] Van Calster, B., Collins, G. S., Vickers, A. J., Wynants, L., Kerr, K. F., Barreñada, L., Varoquaux, G., Singh, K., Moons, K. G. M., Hernandez-Boussard, T., Timmerman, D., McLernon, D. J., van Smeden, M., & Steyerberg, E. W. (2025). Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: Overview and guidance. *The Lancet Digital Health*, 7(12), 100916. <https://doi.org/10.1016/j.landig.2025.100916>
- [6] Banerji, C. R. S., Chakraborti, T., Harbron, C., & MacArthur, B. D. (2023). Clinical AI tools must convey predictive uncertainty for each individual patient. *Nature Medicine*, 29(12), 2996–2998. <https://doi.org/10.1038/s41591-023-02562-7>

- [7] Seoni, S., Jahmunah, V., Salvi, M., Datta Barua, P., Molinari, F., & Acharya, U. R. (2023). Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). *Computers in Biology and Medicine*, 165, 107441. <https://doi.org/10.1016/j.combiomed.2023.107441>
- [8] Liu, X., Glocker, B., McCradden, M. M., Ghassemi, M., Denniston, A. K., & Oakden-Rayner, L. (2022). The medical algorithmic audit. *The Lancet Digital Health*, 4(5), e384–e397. [https://doi.org/10.1016/S2589-7500\(22\)00003-6](https://doi.org/10.1016/S2589-7500(22)00003-6)
- [9] Kwong, J. C. C., Erdman, L., Khondker, A., Skreta, M., Goldenberg, A., McCradden, M. D., Lorenzo, A. J., & Rickard, M. (2022). The silent trial: The bridge between bench-to-bedside clinical AI applications. *Frontiers in Digital Health*, 4, 929508. <https://doi.org/10.3389/fdgth.2022.929508>
- [10] Vazquez, J., & Facelli, J. C. (2022). Conformal prediction in clinical medical sciences. *Journal of Healthcare Informatics Research*, 6(3), 241–252. <https://doi.org/10.1007/s41666-021-00113-8>
- [11] Kwon, H., & Kim, D.-J. (2026). Conformal selective prediction with cost aware deferral for safe clinical triage under distribution shift. *Scientific Reports*, 16, 10016. <https://doi.org/10.1038/s41598-026-40637-w>
- [12] Romero, F. P., Piñol, D. C., & Vázquez-Seisdedos, C. R. (2021). DeepFilter: An ECG baseline wander removal filter using deep learning techniques. *Biomedical Signal Processing and Control*, 70, 102992. <https://doi.org/10.1016/j.bspc.2021.102992>
- [13] Petmezas, G., Stefanopoulos, L., Kilintzis, V., Tzavelis, A., Rogers, J. A., Katsaggelos, A. K., & Maglaveras, N. (2022). State-of-the-art deep learning methods on electrocardiogram data: Systematic review. *JMIR Medical Informatics*, 10(8), e38454. <https://doi.org/10.2196/38454>
- [14] Mongan, J., Moy, L., & Kahn, C. E., Jr. (2020). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. *Radiology: Artificial Intelligence*, 2(2), e200029. <https://doi.org/10.1148/ryai.2020200029>
- [15] Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., ... Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378. <https://doi.org/10.1136/bmj-2023-078378>
- [16] Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., Denniston, A. K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B. A., Mathur, P., McCradden, M. D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D. S. W., ... Perkins, Z. B. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*, 28(5), 924–933. <https://doi.org/10.1038/s41591-022-01772-9>
- [17] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- [18] Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., & Petitjean, F. (2020). InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery*, 34(6), 1936–1962. <https://doi.org/10.1007/s10618-020-00710-y>
- [19] Lea, C., Flynn, M. D., Vidal, R., Reiter, A., & Hager, G. D. (2017). Temporal convolutional networks for action segmentation and detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 156–165. <https://doi.org/10.1109/CVPR.2017.113>

- [20] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, 70, 1321–1330. <https://proceedings.mlr.press/v70/guo17a.html>
- [21] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning*, 70, 3319–3328. <https://proceedings.mlr.press/v70/sundararajan17a.html>
- [22] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31, 9525–9536. <https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html>
- [23] Perez Alday, E. A., Gu, A., Shah, A. J., Robichaux, C., Wong, A.-K. I., Liu, C., Liu, F., Rad, A. B., Elola, A., Seyedi, S., Li, Q., Sharma, A., Clifford, G. D., & Reyna, M. A. (2021). Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020. *Physiological Measurement*, 41(12), 124003. <https://doi.org/10.1088/1361-6579/abc960>
- [24] Zheng, J., Zhang, J., Danioko, S., Yao, H., Guo, H., & Rakovski, C. (2020). A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7(1), 48. <https://doi.org/10.1038/s41597-020-0386-x>