

Diagnostic Preservation After ECG Digitization: From Risk Stratification to Reliability Assurance

Thai Nho Nguyen^{1,2}, Van Thang Doan^{3*}

¹Faculty of Information Technology, Industrial University of Ho Chi Minh City, Vietnam.

^{2,3}Faculty of Mechanical Engineering, Saigon Technology University, Ho Chi Minh City, Vietnam.

E-mail: ^{1,2}nhont24691@pgr.iuh.edu.vn, ^{1,2}nho.nguyenthai@stu.edu.vn, ^{3*}doanvanthang@iuh.edu.vn

Orcid ID: ^{1,2}0009-0000-5026-6593, ^{3*}0000-0002-2055-6255

Abstract

As Electrocardiogram (ECG) digitization becomes increasingly accurate, waveform similarity alone does not establish preservation of diagnostically relevant information. This study investigates whether post-digitization diagnostic instability can be identified without access to the native digital ECG and whether such reference-free observability is sufficient to support reliability assurance. PTB-XL waveforms were rendered as 3×4 ECG pages at 25 mm/s, 10 mm/mV, and 4 pixels/mm and re-digitized with a fixed image-based trace-recovery pipeline. Analyses on PTB-XL showed that unobserved lead-time regions contributed only limited incremental diagnostic information under the tested setting, indicating that missing temporal support alone could not explain the observed preservation loss. A reference-free instability score derived solely from digitized ECGs retained above-chance record-level discrimination on a patient-disjoint cohort (AUROC 0.614, unitless, for distinguishing ECGs with versus without at least one probe-defined new diagnostic error), but calibration deteriorated after population transfer. Selective acceptance reduced the occurrence of probe-defined new diagnostic errors without satisfying the multimetric preservation requirements, while class-wise analysis revealed strongly directional and diagnosis-dependent failures. Application to 99 real-world ECGs from Tam Duc Hospital further showed that the representation remained technically applicable despite structured feature-domain shifts. These findings establish a distinction between risk observability and preservation assurance: diagnostic instability can be partially stratified without an oracle reference, but successful ranking alone is insufficient to guarantee stable diagnostic decisions or calibrated reliability.

Keywords: ECG Digitization, Diagnostic Information Preservation, Reliability Assessment, Risk Stratification, Decision Stability, Calibration.

1. Introduction

Twelve-lead electrocardiography is a foundational cardiovascular measurement and an important data source for modern machine-learning systems [1, 2, 3]. Yet many historical ECG collections remain available only as paper records, scanned documents, PDF files, or photographs. Recent community benchmarks and digitization systems have made waveform

* Corresponding Author

Journal of Trends in Computer Science and Smart Technology, December 2026, Volume 8, Issue 4, Pages 978-992

DOI: <https://doi.org/10.36548/jtcsst.2026.4.004>

Received: 04.08.2026, received in revised form: 08.09.2026, accepted: 01.10.2026, published: 12.10.2026

© 2026 Inventive Research Organization. This is an open access article under the Creative Commons Attribution-NonCommercial International (CC BY-NC 4.0) License

recovery from such sources increasingly feasible [4, 5, 6, 7]. The scientific question is therefore shifting from whether an ECG can be digitized to whether diagnostically relevant information is preserved after digitization.

This distinction matters because reliability is not a single property. High waveform similarity does not guarantee preservation of diagnostic ranking; preserved ranking does not guarantee stable threshold-based decisions; and a score that separates lower- from higher-risk records may lose its probabilistic meaning under population or feature shift. These issues connect ECG digitization to broader questions of calibration, dataset shift, and deployment reliability in medical machine learning [8, 9, 10].

The present study examines post-digitization reliability at the record level. It separates the incremental diagnostic contribution of lead-time regions absent from a standard print layout from distortion introduced within the observed support, develops a reference-free representation of probe-defined diagnostic instability, evaluates that representation on a patient-disjoint cohort, and tests whether selective acceptance can satisfy explicit multimetric preservation requirements. Class-specific failure directions are then characterized, and the same representation is examined on 99 real-world hospital ECGs to distinguish technical applicability from calibrated reliability. Together, these analyses ask whether diagnostic instability can be observed without access to the native digital ECG and whether such observability is sufficient to support preservation assurance.

2. Related Work

2.1 ECG Digitization and Reconstruction

Modern ECG digitization methods have shown that waveforms can be recovered automatically from printed or image-based records with increasingly high accuracy [6, 7]. In parallel, reconstruction of a full 12-lead ECG from reduced leads and completion under missingness have advanced rapidly, from conventional deep architectures to GANs, CNN-Transformer models, diffusion models, and Mamba-based networks [11, 12, 13, 14, 15, 16]. Recent benchmarking also suggests that configurations optimized for waveform similarity need not be optimal for downstream diagnosis [17]. These developments shift the open problem from whether reconstruction is technically possible to whether the resulting signal preserves the information required for subsequent decisions. Recent paper-ECG studies also highlight that segmentation and signal-extraction errors can be amplified during digitization and that small waveform losses may propagate into downstream classification behavior [20, 21].

2.2 Diagnostic Preservation Beyond Waveform Fidelity

Metrics such as Pearson correlation, RMSE, and SNR quantify signal-level similarity but do not determine whether downstream models preserve diagnostic ranking, threshold-based decisions, or error profiles. A reconstructed waveform may remain globally similar to the reference while inducing probability shifts large enough to cross a decision threshold. Conversely, relative ordering across records may be largely retained even when absolute probabilities become miscalibrated. Reliability after digitization should therefore be examined at several distinct levels: waveform fidelity, diagnostic discriminability, decision stability, and the transportability of score meaning across populations.

Figure 1 summarizes these levels as a conceptual scaffold. The empirical contribution

of the present work begins at the point where diagnostic instability must be recognized without an oracle reference, then asks whether such recognition is sufficient for preservation assurance and whether the resulting representation remains applicable after domain transfer.

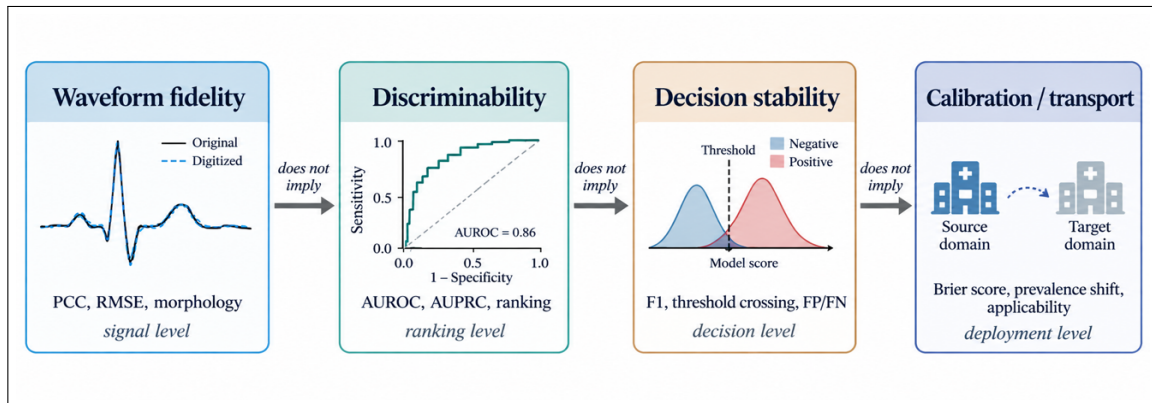


Figure 1. Reliability levels after ECG digitization

2.3 Reliability Under Domain Shift

In medical machine learning, discrimination often degrades less than calibration when outcome prevalence or feature distributions shift [8, 9, 10]. Recent work on clinical machine-learning operations and self-validating ECG systems emphasizes the need to assess reliability at the point of use [18, 19]. Nevertheless, high average diagnostic performance, or even a score that ranks higher-risk records, does not establish that diagnostic information is preserved for individual digitized ECGs. Unlike conventional confidence filtering or post-hoc calibration, which operate on a predictor's own output confidence, the present framework derives a separate reference-free score from the digitized waveform and targets digitization-associated instability of a fixed diagnostic probe rather than recalibrating that probe.

The remaining gap is therefore not signal recovery itself, but whether record-level downstream instability is observable from the digitized waveform and whether such observability is sufficient for preservation assurance. This distinction motivates evaluation of reliability beyond average reconstruction utility.

3. Materials and Methods

3.1 Data Sources and Study Cohorts

PTB-XL [1] provided native digital waveforms and labels for the five diagnostic superclasses: NORM, MI, STTC, CD, and HYP. Separate cohorts supported the missing-support analysis (21,388 ECGs), development of the reference-free stratification model (400 ECGs from 382 patients), and patient-disjoint validation (1,669 ECGs from 1,495 patients). Model and threshold choices were finalized on the development cohort before validation. A further 99 real-world ECG image/PDF sources from Tam Duc Hospital were used for domain-applicability analysis because paired native digital waveforms were unavailable. Table 1 summarizes the cohort roles.

Table 1. Data sources and analytical roles

Data source	ECGs	Patients	Analytical role
PTB-XL, missing-support analysis	21,388	-	Quantify the incremental diagnostic value of lead-time regions absent from the print layout.
PTB-XL, development cohort	400	382	Define diagnostic instability and estimate a reference-free stratification score.
PTB-XL, patient-disjoint validation	1,669	1,495	Evaluate discrimination, calibration, and preservation on patient-disjoint data.
Tam Duc Hospital	99	-	Assess cross-domain applicability where no paired native waveform is available.

3.2 Observed Temporal Support and Missing-Support Analysis

A standard 3×4 ECG print displays each short lead for only 2.5 s. All analyses, therefore, followed a non-fabrication principle: each lead was represented only over the interval directly observed in the print layout, with no repetition, concatenation, or interpolation across unobserved time. For PTB-XL, native digital ECGs were rendered into a 3×4 layout and digitized with a fixed procedure, producing paired reference-observed and digitized-observed waveforms on identical temporal support. Diagnostic differences in the preservation analyses therefore reflect digitization distortion under a common observed support, rather than a mixture of digitization error and missing-signal completion.

For the PTB-XL simulation, high-resolution native recordings were sampled onto the raster-time grid of a custom Python renderer at 25 mm/s and 10 mm/mV with 4 pixels/mm. The resulting RGB page was 1000 × 448 pixels (3 × 4 short-lead layout plus a 10-s lead-II rhythm strip), corresponding to a 100-Hz raster-time sampling rate. Rendering used Pillow/ImageDraw, and the subsequent digitization reused the fixed image-processing procedure based on border suppression, Viterbi-L1 trace recovery, and adaptive vertical recovery. Only the 12 short-lead windows were used in the preservation representation.

The incremental contribution of the unobserved lead-time regions was assessed with a nested representation. For the same ECG and label, the two input conditions were

$$O = [X_{obs}, 0, M], \quad F = [X_{obs}, X_{extra}, M]$$

where X_{obs} denotes the lead-time regions visible in the print layout, X_{extra} denotes the additional intervals available only in the native digital ECG, and M is the identical positional mask used in both representations. The two inputs have the same dimensionality, and F is a strict information superset of O . Separate models were trained for O and F under the same architecture, data partition, optimization settings, and initialization seed, with five seeds evaluated for each representation and preprocessing branch. Probabilities were averaged across seeds and decision thresholds were selected on the corresponding validation data. For the shape-normalized branch, normalization statistics were estimated exclusively from X_{obs} before constructing O and F , preventing information from X_{extra} from entering the observed-only condition. The $O - F$ contrast therefore estimates the incremental diagnostic value of unobserved lead-time regions under matched modeling conditions; it does not imply that those regions lack physiological information.

The full support grid spanned 12 leads over 10 s at 100 Hz. Accordingly, M was implemented as a binary 12 × 1000 lead-by-time matrix: entries corresponding to the actually

printed 2.5-s interval of each lead were set to 1, and unobserved lead-time positions were set to 0. The same M was provided to both O and F , so the comparison changed only the availability of X_{extra} rather than the location of observed support.

3.3 Definition of Record-Level Diagnostic Instability

For reproducibility, class-specific decision thresholds were chosen on the fold-9 validation data by exhaustive search over 37 evenly spaced values from 0.05 to 0.95 (step 0.025). For each diagnostic class, the threshold that maximized validation F1 was retained. This procedure yielded 0.425 (NORM), 0.775 (MI), 0.750 (STTC), 0.350 (CD), and 0.575 (HYP); all five thresholds were then frozen before the preservation analyses.

A fixed diagnostic probe for NORM, MI, STTC, CD, and HYP was trained on PTB-XL folds 1-8, selected on fold 9, and then held unchanged for the subsequent preservation analyses. The probe was a compact one-dimensional convolutional neural network with three convolutional blocks followed by global average pooling and was distinct from the separately trained models used in the nested O/F missing-support analysis. Its primary input representation normalized each lead by subtracting its median and dividing by its RMS. Class-specific decision thresholds were selected once on development data and then fixed at 0.425 (NORM), 0.775 (MI), 0.750 (STTC), 0.350 (CD), and 0.575 (HYP). For class c and record i , a new diagnostic error was defined as a case in which the reference waveform was classified correctly but the digitized waveform was classified incorrectly:

$$E_{ic}^{new} = \mathbb{1} [\hat{y}_{ic}^{oracle} = y_{ic}] \mathbb{1} [\hat{y}_{ic}^{digit} \neq y_{ic}] \quad (1)$$

Using the class-level error definition in Eq. (1), record-level instability was defined as

$$H_i = \mathbb{1} \left[\sum_c E_{ic}^{new} > 0 \right] \quad (2)$$

The endpoint in Eq. (2) is a property of the diagnostic probe and should not be interpreted as a clinical adverse event.

3.4 Reference-free Reliability Stratification

To remain usable when no native digital ECG is available, the reliability estimator was restricted to information derived from the digitized waveform. From each digitized 12×250 ECG tensor, the feature mapping $\phi(X_i^{digit})$ comprised 45 descriptors grouped into six families: amplitude and scale; roughness, derivative behavior, and flatness; low- and high-frequency spectral ratios; periodicity; consistency within simultaneous three-lead groups; and physiologically motivated consistency descriptors, including the Einthoven residual, augmented-lead sum residual, and correlations between neighboring precordial leads. No feature used the native waveform, NRMSE, or diagnostic labels at scoring time. The resulting score was written as

$$r_i = f_{\theta}(\phi(X_i^{digit})) \quad (3)$$

where f_{θ} was estimated only on the development cohort. Eq. (3) defines a stratification score; it is not assumed a priori to be a calibrated probability. The primary estimator was L2-regularized logistic regression ($C = 1$) after standardization. Development performance was estimated by 5-fold StratifiedGroupKFold cross-validation repeated 10 times with

patient-level grouping. The estimator was subsequently fitted to all 400 development records and evaluated on the 1,669 patient-disjoint validation records. Patient-level bootstrap resampling was used for uncertainty estimation.

Generalization was characterized jointly through discrimination, calibration, and prevalence-adjusted precision-recall performance. The assessment considered whether the lower bound of the patient-bootstrap 95% CI for AUROC exceeded 0.50, whether mean Brier skill remained positive, and whether AUPRC was at least as large as the observed instability prevalence. Brier skill was defined as $1 - \text{Brier} / \text{Brier}_{null}$. The null comparator used the development prevalence (0.4275), avoiding dependence on labels from the patient-disjoint validation cohort.

The 45 descriptors were predefined to cover complementary signal properties that can be computed without a native reference: amplitude scale, local roughness and derivative behavior, flatness, low- and high-frequency content, periodic organization, within-triplet lead consistency, and physiologically motivated inter-lead relationships. They were intended as a compact plausibility representation rather than as an optimized or causally interpretable feature set. The complete descriptor set and its six predefined feature families are listed in Table 2.

Table 2. Reference-free feature families and descriptors

Feature family (n)	Descriptors
Amplitude and scale (9)	amp_p2p_median, amp_p2p_q90, amp_p2p_max, amp_p2p_iqr; amp_rms_median, amp_rms_q90, amp_rms_max, amp_rms_iqr; amp_p2p_cv.
Roughness, derivative behavior, and flatness (12)	roughness_median, roughness_q90, roughness_max, roughness_iqr; deriv_rms_ratio_median, deriv_rms_ratio_q90, deriv_rms_ratio_max, deriv_rms_ratio_iqr; flat_fraction_median, flat_fraction_q90, flat_fraction_max, flat_fraction_iqr.
Spectral ratios (8)	lowfreq_ratio_median, lowfreq_ratio_q90, lowfreq_ratio_max, lowfreq_ratio_iqr; highfreq_ratio_median, highfreq_ratio_q90, highfreq_ratio_max, highfreq_ratio_iqr.
Periodicity (4)	periodicity_median, periodicity_q90, periodicity_max, periodicity_iqr.
Simultaneous-lead consistency (8)	triplet_abs_corr_median, triplet_abs_corr_q90, triplet_abs_corr_max, triplet_abs_corr_iqr; triplet_spread_median, triplet_spread_q90, triplet_spread_max, triplet_spread_iqr.
Physiological lead consistency (4)	einthoven_residual_norm, augmented_sum_residual_norm, precordial_neighbor_abs_corr_median, precordial_neighbor_abs_corr_min.

Family-level contribution was assessed using the FULL45 representation together with six leave-one-family-out variants and six single-family variants. All variants used the same standardization, L2 logistic regression ($C = 1$), and 5-fold StratifiedGroupKFold protocol repeated 10 times with patient-level grouping. Each variant was then fit once on all 400 development records and evaluated without tuning on the same 1,669-record patient-disjoint cohort. Redundancy was summarized by pairwise Spearman correlations among the 45 development features, and coefficient stability was summarized across the 50 grouped development fits using standardized coefficient magnitude and sign consistency.

3.5 Selective Acceptance and Preservation Assessment

Accepted fractions of 50%, 75%, and 90% were evaluated. For an accepted fraction q , the retained subset consisted of records with the lowest assigned risk scores:

$$A_q = \{i : r_i \leq Q_q(r)\} \quad (4)$$

Eq.(4) formalizes the low-score acceptance rule used at each evaluated coverage. Preservation was treated as a multimetric requirement rather than a single performance measure. For each accepted subset, the degradation vector was

$$P_q = \left(\Delta AUROC, \Delta AUPRC, \Delta F1, \max_c \Delta AUROC_c \right) \quad (5)$$

In Eq.(5), Δ denotes reference performance minus digitized performance, and both representations were evaluated on the same A_q . The predefined margins were (0.03, 0.05, 0.05, 0.08), respectively. Preservation was considered supported only when all components satisfied their margins; improvement in a single metric was not sufficient. Patient-level bootstrap confidence intervals were used to quantify uncertainty.

These values were prespecified empirical engineering/scientific preservation margins rather than clinical tolerance, equivalence, or non-inferiority thresholds. The macro-AUROC margin of 0.03 was set more stringently to constrain degradation in overall discrimination; margins of 0.05 were used for macro-AUPRC and macro-F1 because the former is more prevalence-sensitive and the latter is threshold-dependent. The 0.08 worst-class AUROC margin served as a guardrail against a large class-specific degradation being obscured by macro-averaging. The margins were inherited unchanged from the preceding digitization-preservation protocol.

3.6 Class-wise Analysis and Hospital-Domain Evaluation

Class-wise AUROC, AUPRC, F1, the direction of new errors, threshold crossings, and distance from the decision threshold were examined to characterize why ranking and decision stability diverged. These analyses were interpretive and did not alter the fixed model formulation or diagnostic thresholds.

For the Tam Duc cohort, the same representation and stratification model were applied to all 99 ECGs; the 97 records meeting the pre-existing export-quality criterion formed a sensitivity subset. Across the 1,188 short-lead segments, 435 required mild resampling to 250 samples because the image-derived sampling frequency was close to, but not exactly, 100 Hz. Resampling was confined to the observed lead interval and did not synthesize unobserved support. Because paired native digital waveforms were unavailable, the hospital analysis assessed feature-domain overlap, record-level applicability, and the distribution of assigned stratification scores rather than observed diagnostic instability.

For the feature-domain shift analysis in Figure 5, the PTB-XL reference distribution combined the 400-record development cohort and the 1,669-record patient-disjoint validation cohort ($n = 2,069$). Feature medians and IQRs used for normalization were computed from this fixed combined PTB-XL reference; the hospital data were not used to redefine the normalization statistics.

4. Results

4.1 Limited Contribution of Unobserved Lead-Time Regions

In the nested comparison, removing the lead-time regions absent from the print layout reduced macro-AUROC by only 0.00359 in the amplitude-aware branch and 0.00062 in the shape-normalized branch; changes in macro-AUPRC and macro-F1 were likewise small. These differences were nearly an order of magnitude smaller than the previously observed digitization-associated macro-AUROC degradation of approximately 0.034-0.037 under the matched observed support. The result does not imply that the unobserved intervals contain no physiological information. Under the matched model family and training design used here, however, their incremental diagnostic contribution was too small to support missing-signal completion as the primary explanation for the observed preservation loss.

4.2 Generalization of Reference-free Instability Stratification

On the development cohort, the reference-free score showed above-chance discrimination of probe-defined instability. More importantly, discrimination persisted when the fixed formulation was evaluated on 1,669 ECGs from 1,495 patients who were disjoint from the development cohort: AUROC was 0.6142, with a patient-bootstrap 95% CI of 0.5867-0.6419, and AUPRC was 0.4333 compared with an instability prevalence of 0.3409. These results indicate that diagnostic instability is not entirely latent; some of its structure is observable from the digitized waveform alone, although the level of discrimination is modest. Figure 2 summarizes the record-level risk stratification and calibration behavior on the patient-disjoint validation cohort.

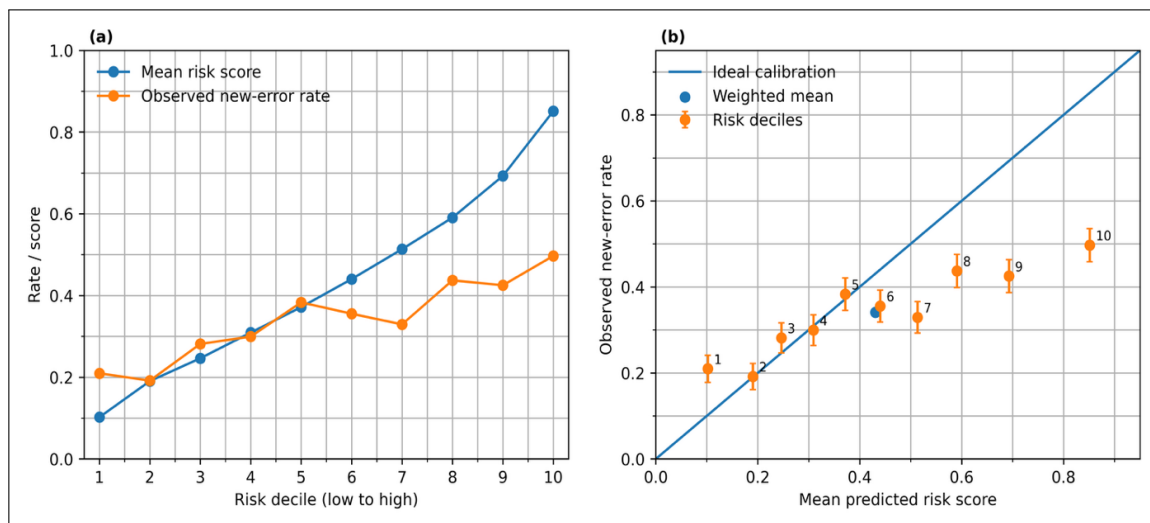


Figure 2. Risk stratification and calibration

The mean assigned score was 0.4308, close to the development prevalence (0.4275) but appreciably above the instability prevalence in the patient-disjoint cohort (0.3409). Brier skill was -0.0491 (95% CI -0.0933 to -0.0054), so the prespecified calibration criterion was not satisfied even though the AUROC and AUPRC criteria were met. Figure 2 shows both phenomena: lower and higher score deciles retained a meaningful gradient in observed instability, yet the score-to-outcome relationship departed from ideal calibration. The same score can therefore preserve relative ordering while losing absolute probabilistic meaning after population shift.

The family-level sensitivity analysis revealed heterogeneous contributions to the fixed FULL45 representation. Removing spectral descriptors produced the largest observed AUROC decrease on patient-disjoint validation (-0.0223), followed by amplitude/scale descriptors (-0.0186), whereas removing periodicity or physiological-consistency descriptors produced little change. Spectral descriptors also showed the strongest standalone discrimination (AUROC 0.6129). Pairwise redundancy was limited, with only 2.02% of feature pairs exceeding $|\rho| > 0.8$ and 0.61% exceeding $|\rho| > 0.9$. Coefficient-sign stability was highest for the amplitude/scale, roughness/derivative/flatness, and periodicity families (median 0.98) and lowest for physiological lead consistency (0.64). Table 3 summarizes the family-level sensitivity and coefficient-stability results.

Table 3. Feature-family sensitivity and coefficient stability

Feature family (n)	AUROC after removal (Δ)	Single-family AUROC	Median sign consistency
Amplitude/scale (9)	0.5956 (-0.0186)	0.5590	0.98
Roughness/derivative/flatness (12)	0.6065 (-0.0077)	0.5359	0.98
Spectral ratios (8)	0.5919 (-0.0223)	0.6129	0.92
Periodicity (4)	0.6185 (+0.0043)	0.5312	0.98
Simultaneous-lead consistency (8)	0.6104 (-0.0038)	0.5681	0.88
Physiological lead consistency (4)	0.6140 (-0.0002)	0.5647	0.64

4.3 Selective Acceptance and Preservation Performance

Selective acceptance reduced the prevalence of probe-defined new diagnostic errors, demonstrating that the score carried practically useful ranking information. At 50% coverage, the new-error rate decreased from 34.09% to 27.22%, a relative reduction of approximately 20%. Preservation, however, required a stronger condition: all diagnostic metrics had to remain within their prespecified degradation margins. Figure 3 shows the degradation profiles and corresponding uncertainty across the evaluated acceptance fractions.

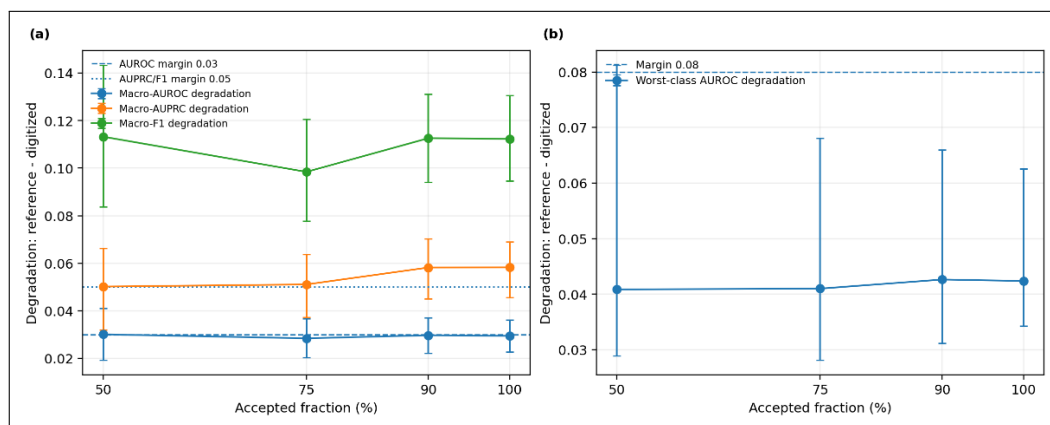


Figure 3. Preservation degradation across accepted fractions

Across all examined acceptance fractions, at least one preservation dimension exceeded its predefined degradation margin. At 75% coverage, which yielded the smallest macro-F1 degradation, macro-AUROC degradation was 0.0284 and the worst class-specific AUROC degradation was 0.0410, both within their point-estimate margins. In contrast, macro-AUPRC degradation was 0.0511 and macro-F1 degradation was 0.0984; the latter remained almost twice the 0.05 margin. Its patient-bootstrap 95% CI extended to 0.1206. Figure 3 presents the

point estimates and confidence intervals across accepted fractions, showing that the failure of preservation was not driven by a single unstable estimate. Risk ranking therefore improved case selection without establishing multimetric preservation. The non-monotonic minimum in macro-F1 degradation at 75% coverage is consistent with case-mix effects: macro-F1 depends on fixed class-specific decision thresholds and on the class composition of the accepted subset, whereas the global risk score does not rank every class-specific threshold crossing monotonically. Stricter coverage therefore need not yield a smaller macro-F1 drop.

4.4 Class-specific Decision Instability

Class-wise analysis revealed marked heterogeneity. Across the full patient-disjoint validation cohort, STTC AUROC decreased from 0.903 on the reference representation to 0.875 after digitization, a degradation of only 0.0278. In contrast, STTC F1 decreased from 0.698 to 0.397, corresponding to a degradation of 0.301. NORM was substantially more stable, with an F1 degradation of 0.0116, while MI, CD, and HYP fell between these extremes. Figure 4 summarizes the class-specific AUROC-F1 dissociation and the direction of probe-defined new errors.

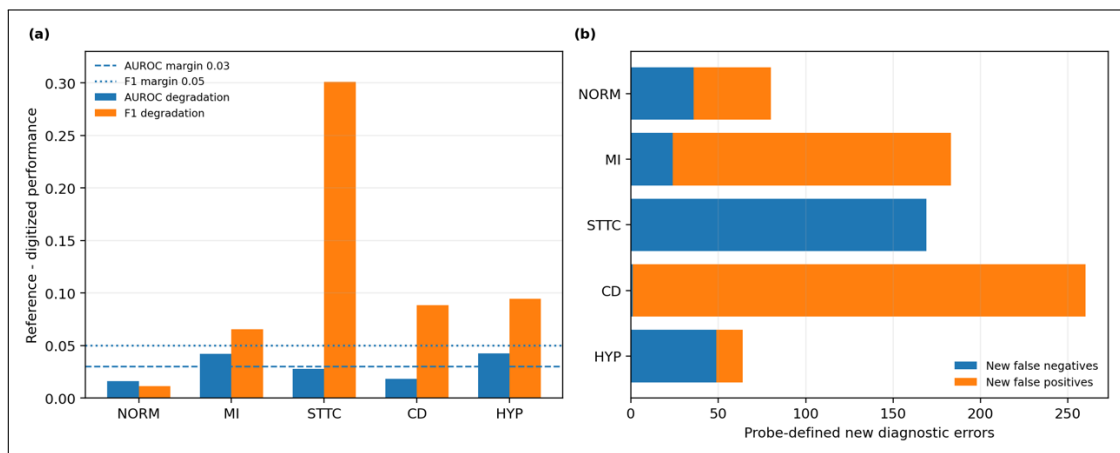


Figure 4. Class-specific degradation and direction of new errors

The direction of the new errors also differed systematically across classes: all 169 new STTC errors were false negatives, whereas 259 of 260 new CD errors were false positives. MI was predominantly shifted toward false positives, while HYP was predominantly shifted toward false negatives. Table 4 summarizes the class-specific evidence.

Table 4. Discrimination-decision dissociation

Evidence	Ref.	Digitized	Difference (95% CI)
Macro-AUROC, full patient-disjoint cohort	0.8751	0.8456	0.0294 [0.0227, 0.0361]
Macro-F1, full patient-disjoint cohort	0.6675	0.5552	0.1123 [0.0946, 0.1306]
STTC AUROC	0.903	0.875	0.0278 [0.0169, 0.0388]
STTC F1	0.698	0.397	0.3013 [0.2491, 0.3561]
STTC new errors	-	-	169/169 FN
CD new errors	-	-	259/260 FP

4.5 Hospital-domain Applicability and Feature Shift

The fixed representation and stratification model were applied to all 99 Tam Duc ECGs without retraining or recalibration. The mean and median assigned scores were 0.456 and 0.432, respectively, compared with 0.431 and 0.405 in the patient-disjoint PTB-XL cohort. The

hospital distribution therefore showed a modest rightward shift: 28.3% of Tam Duc ECGs were above the PTB-XL validation Q_{75} and 15.2% were above Q_{90} .

Feature-domain analysis did not indicate wholesale out-of-distribution behavior. The median number of features outside the full PTB-XL min-max range was zero, and no hospital record exceeded that range in more than 3 of 45 features. Nevertheless, several feature families exhibited structured shifts. The augmented-lead sum residual and Einthoven residual shifted by 1.45 and 1.09 PTB-XL IQR units, respectively; flatness-related statistics increased, whereas selected amplitude statistics decreased by approximately 0.59 IQR. The largest shifts are shown in Figure 5.

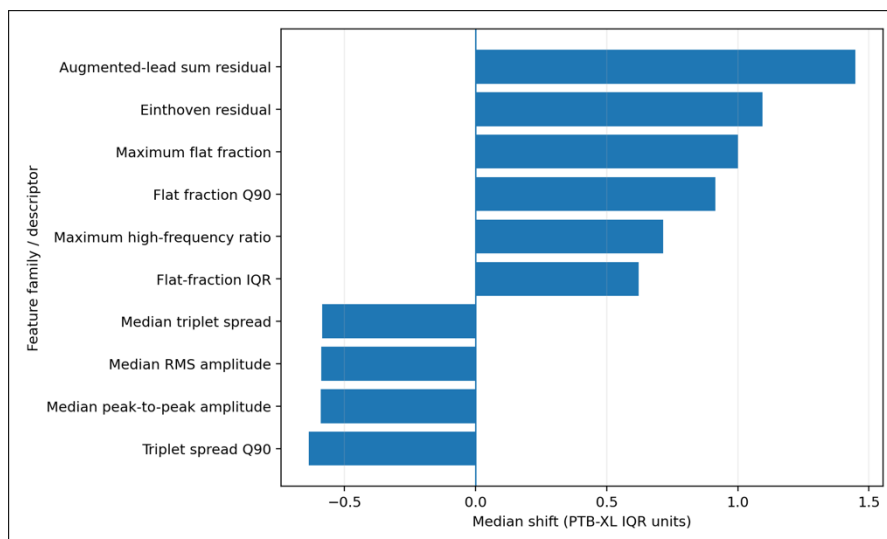


Figure 5. Feature-domain shift between Tam Duc and PTB-XL

These findings support domain applicability of the representation, but not calibrated diagnostic risk estimation, because no outcome variable corresponding to H was available for the Tam Duc cohort.

5. Discussion

5.1 Multilevel Reliability after ECG Digitization

The central finding is a systematic separation among reliability levels rather than any single AUROC value. Diagnostic instability was partially observable without a reference waveform; that signal retained discriminatory value on patient-disjoint data but did not support calibrated preservation assurance; class-wise analysis showed that failure depended on the structure of the decision boundary; and hospital data demonstrated that technical applicability can extend across domains without implying transport of calibration. Waveform similarity, discrimination, decision stability, and deployment reliability should therefore be treated as related but distinct forms of evidence.

The practical implication is that no single performance summary captures all of these levels; evidence for discrimination, decision stability, and transportability should be interpreted separately.

5.2 Discrimination-Decision Dissociation

AUROC primarily reflects relative ranking. Consequently, digitization-induced score perturbations can leave many positive-negative pairs in the same order even when individual probabilities shift appreciably. AUROC may therefore remain comparatively stable despite substantial changes in threshold-based decisions. F1, in contrast, is determined after probabilities are converted into class decisions; even a modest displacement around a decision threshold can alter the predicted label. Preservation of discrimination and preservation of decision stability are therefore related but non-equivalent properties.

The class-wise results make this distinction particularly clear. STTC retained comparatively stable discrimination while exhibiting a pronounced loss of threshold-based decision stability, with new errors occurring exclusively toward false negatives. CD exhibited the opposite directional pattern, with nearly all new errors occurring toward false positives. The asymmetry of these failures indicates that digitization-induced perturbations interact with class-specific decision boundaries rather than producing a uniform increase in prediction noise.

This result also exposes a limitation of scalar reliability scores: a single probability of instability cannot encode which diagnostic class is vulnerable, the direction of the expected error, or its relationship to the class-specific threshold. Reliability assessment after digitization may therefore benefit from class-conditional or vector-valued representations rather than a single global score.

5.3 Limits of Reference-free Reliability Stratification

The patient-disjoint evaluation shows that discrimination alone is insufficient for preservation assurance. The reference-free score retained useful ordering of records, but calibration deteriorated and selective acceptance did not satisfy the multimetric preservation requirements. Thus, a score can be informative for ranking without serving as a reliable probability of instability or a guarantee of threshold-based diagnostic preservation. This distinction narrows the claims that can be made from reference-free stratification and raises the evidentiary standard for self-assessing ECG systems.

5.4 Implications for ECG Digitization and Deployment

Recent reconstruction studies have increasingly incorporated physiological constraints, uncertainty estimation, and downstream objectives [12, 13, 14, 15, 16]. The present results complement that literature by showing that post-reconstruction evaluation should distinguish ranking from calibration, account for the direction of class-specific errors, and assess explicit preservation constraints alongside average downstream performance. The limited incremental value of unobserved lead-time regions under the tested design further suggests that completing missing temporal support is unlikely to resolve the reliability problem by itself.

A future assurance system may therefore need to combine global abstention with class-conditional reliability estimates and target-domain calibration. The Tam Duc analysis illustrates why these components should remain separate: the reference-free representation remained technically applicable in a hospital setting, yet structured feature shifts persisted and calibration transport could not be established without paired outcomes.

For practical use before target-domain outcome validation is available, the score should be treated only as a relative triage or review-prioritization signal, for example to flag

higher-scoring digitized ECGs for additional waveform-quality review or repeat digitization. It should not be interpreted as an absolute probability of diagnostic failure or used as an autonomous clinical rejection rule; any fixed action threshold would require target-domain outcome validation and, if appropriate, local recalibration.

5.5 Limitations

Several limitations define the appropriate scope of interpretation. First, H represents instability of a fixed diagnostic probe rather than a clinical outcome. Second, the reference-free score achieved only modest discrimination, and the study does not assume that it is an optimal predictor. Third, the patient-disjoint validation cohort remains within PTB-XL fold 10 and is therefore not a completely external dataset. Fourth, the Tam Duc cohort supports applicability and feature-domain transport analysis but does not provide outcome-based calibration or preservation validation. Fifth, a scalar instability score may be intrinsically limited when downstream errors are class-specific and directional. These limitations constrain the claims supported by the present data and define the evidence required before broader clinical or cross-domain inference. Computationally, the original 45-descriptor implementation required a median of 4.158 ms per 12×250 ECG (IQR 4.034–4.354 ms; 1,000 CPU trials on an Apple M2 Pro, batch size 1), and standardization plus logistic scoring required an additional median of 0.115 ms per ECG. These measurements characterize algorithmic overhead only and exclude upstream digitization and file I/O.

6. Conclusion

Diagnostic reliability after ECG digitization cannot be inferred from waveform similarity or average classification performance alone. Probe-defined instability was partially observable from the digitized signal and retained above-chance discrimination on patient-disjoint data, but calibration deteriorated and selective acceptance did not satisfy the multimetric preservation criteria. Failures were class-specific and directional, while the hospital analysis showed that technical applicability of a reference-free representation can coexist with unresolved calibration transport. These findings support evaluating ECG digitization and reconstruction through distinct evidence for discriminability, decision stability, calibration, and diagnostic information preservation rather than through reconstruction quality or a single global score alone.

Data Availability and Reproducibility

PTB-XL is publicly available [1]. The analytical definitions, feature representation, validation design, and preservation criteria are described in the Methods. The Tam Duc clinical ECG data are confidential, de-identified, and not publicly released because of patient-privacy and institutional data-protection requirements; only aggregate, non-identifying statistics are reported.

Ethics Statement

PTB-XL is a public de-identified dataset and was used under its published data conditions. Retrospective use of clinical ECG records from Tam Duc Hospital for research was approved by the Institutional Review Board of Tam Duc Hospital (Protocol ID: 24751791). Direct patient identifiers were removed before analysis, and each ECG record was assigned a non-identifying research code. Only de-identified, coded ECG data were used in this study.

Acknowledgement

The authors thank Tam Duc Hospital for supporting access to the retrospective ECG records used in the hospital-domain applicability analysis.

References

- [1] Wagner, Patrick, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. "PTB-XL, A Large Publicly Available Electrocardiography Dataset." *Scientific data* 7, no. 1 (2020): 154.
- [2] Strodthoff, Nils, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. "Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL." *IEEE journal of biomedical and health informatics* 25, no. 5 (2020): 1519-1528.
- [3] Ribeiro, Antônio H., Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton PS Ferreira et al. "Automatic Diagnosis of the 12-lead ECG Using a Deep Neural Network." *Nature communications* 11, no. 1 (2020): 1760.
- [4] Reyna, Matthew A., Weigle J. Deepanshi, Zuzana Koscova, Andoni Elola, Salman Seyedi, Kiersten Campbell, Gari D. Clifford, and Reza Sameni. "Digitization and classification of ECG Images: The George b. Moody Physionet Challenge 2024." *Computing in Cardiology* 51 (2024): 1-4.
- [5] Shivashankara, Kshama Kodthalu, Deepanshi, Afagh Mehri Shervedani, Gari D. Clifford, Matthew A. Reyna, and Reza Sameni. "ECG-Image-Kit: A Synthetic Image Generation Toolbox to Facilitate Deep Learning-Based Electrocardiogram Digitization." *Physiological measurement* 45, no. 5 (2024): 055019.
- [6] Demolder, Anthony, Viera Kresnakova, Michal Hojcka, Vladimir Boza, Andrej Iring, Adam Rafajdus, Simon Rovder et al. "High Precision ECG Digitization Using Artificial Intelligence." *medRxiv* (2024): 2024-08.
- [7] Stenhede, Elias, Agnar Martin Bjørnstad, and Arian Ranjbar. "Digitizing Paper ECGs at Scale: An Open-Source Algorithm for Clinical Research." *NPJ Digital Medicine* 9, no. 1 (2026): 145.
- [8] Sahiner, Berkman, Weijie Chen, Ravi K. Samala, and Nicholas Petrick. "Data Drift in Medical Machine Learning: Implications and Potential Remedies." *The British Journal of Radiology* 96, no. 1150 (2023): 20220878.
- [9] Liou, Lathan, Erick Scott, Prathamesh Parchure, Yuxia Ouyang, Natalia Egorova, Robert Freeman, Ira S. Hofer et al. "Assessing Calibration and Bias of a Deployed Machine Learning Malnutrition Prediction Model Within a Large Healthcare System." *NPJ Digital Medicine* 7, no. 1 (2024): 149.
- [10] dos Santos Silva, Gabriel Ferreira, Fabiano Novaes Barcellos Filho, Roberta Moreira Wichmann, Francisco Costa da Silva Junior, and Alexandre Dias Porto Chiavegatto Filho. "Strategies for Detecting and Mitigating Dataset Shift in Machine Learning for Health Predictions: A Systematic Review." *Journal of Biomedical Informatics* (2025): 104902.
- [11] Obianom, Ekenedirichukwu N., G. André Ng, and Xin Li. "Reconstruction of 12-lead ECG:

- A Review of Algorithms.” *Frontiers in Physiology* 16 (2025): 1532284.
- [12] Presacan, Oriana, Alexandru Dorobanțiu, Jonas L. Isaksen, Tobias Willi, Claus Graff, Michael A. Riegler, Arun R. Sridhar, Jørgen K. Kanters, and Vajira Thambawita. ”Evaluating the Feasibility of 12-lead Electrocardiogram Reconstruction from Limited Leads Using Deep Learning.” *Communications medicine* 5, no. 1 (2025): 139.
 - [13] Nakanishi, Ryuichi, Akimasa Hirata, and Yoshiki Kubota. ”Deep Learning-Based Multi-Lead ECG Reconstruction from Lead I with Metadata Integration and Uncertainty Estimation.” *Sensors* 26, no. 1 (2025): 212.
 - [14] Hu, Haoyang, Zekai Yu, Feiwei Qin, and Fei Yang. ”From One to Twelve: Feasibility and Clinical Utility of Deep Learning-Derived 12-lead ECGs for Remote Cardiac Monitoring.” *Frontiers in Cardiovascular Medicine* 13 (2026): 1856637.
 - [15] Chen, Xingyu, Xianyou Zeng, Quansheng Xu, Lei Dai, and Yonghong Xu. ”Arbitrary Single-lead to 12-lead ECG Reconstruction via Convolutional Mamba-UNet with Physiological Constraints.” *Neurocomputing* (2026): 134590.
 - [16] Wei, Xiaoyang, Zhiyuan Li, Yuanyuan Tian, Mengxiao Wang, Yanrui Jin, Weiping Ding, and Chengliang Liu. ”Condition-Fusion-and-Hiding Denoising Diffusion Model for Electrocardiogram Lead Reconstruction.” *Iscience* 29, no. 4 (2026).
 - [17] Zhang, Xinyu, Hailing Cai, Huilong Duan, and Xudong Lu. ”Systematic Benchmark of Reduced-Lead Configurations for 12-lead ECG Reconstruction: Multi-Model Evaluation Across All Possible Subsets.” *Frontiers in Cardiovascular Medicine* 13 (2026): 1856211.
 - [18] Rajagopal, Anjali, Shant Ayanian, Alexander J. Ryu, Ray Qian, Sean R. Legler, Eric A. Peeler, Meltiady Issa, Trevor J. Coons, and Kensaku Kawamoto. ”Machine Learning Operations in Health Care: A Scoping Review.” *Mayo Clinic Proceedings: Digital Health* 2, no. 3 (2024): 421-437.
 - [19] Rafiei, Alireza, Trisha Dwivedi, Joel Xue, Daphne E. Schlesinger, Nasim Katebi, David E. Albert, and Gari D. Clifford. ”Toward Self-validating ECG Systems: A Personalized, Uncertainty-Aware Approach for Detecting and Estimating Lead Misplacement.” *Journal of Electrocardiology* (2026): 154204.
 - [20] Yu, Xiankai, Jian Wu, Jiahao Wang, Mingjie Wang, Yi-Gang Li, and Wenjie Cai. ”SLPM: A Lightweight Deep Learning Model for End-to-End Paper ECG Digitization.” *Physiological Measurement* 47, no. 2 (2026): 025005.
 - [21] Wang, Xiang, and Jie Yang. ”A Systematic Method Combining Rotated Convolution and State Space Augmented Transformer for Digitizing and Classifying Paper ECGs.” *Symmetry* 17, no. 1 (2025): 120.