Minority-Class Failure and Confidence Miscalibration in Electrocardiogram Classification: A Cross-Dataset Error Audit

Abstract

Automated electrocardiogram classifiers are usually evaluated by aggregate accuracy, a number that can look strong while hiding which patients and which classes a model actually fails on. We audit this gap directly. Seven supervised classifiers, spanning linear, distance-based, kernel, tree-ensemble, boosting, and neural methods, are trained under an identical patient-level protocol on three independent, public electrocardiogram datasets: MIT-BIH (44 records, beat-level arrhythmia typing), a stratified subsample of PTB-XL (**1,961** patients, diagnostic classification), and LUDB (187 records, rhythm classification). Minority classes fail at **2 to 9 times** the rate of the majority class in every dataset. On the two datasets where all seven models are compared directly, AdaBoost is never confidently wrong, while gradient boosting is confidently wrong in **39.9% to 88.2%** of its errors, despite similar or better accuracy. A subset of patients is misclassified by every model on LUDB, and these patients are on average **nine years older** than correctly classified patients wherever demographic data are available. Accuracy and confidence calibration are *not the same property*: two models can match on one and diverge sharply on the other, which matters because an uncertain wrong answer can be caught before harm and a confident one cannot. We argue that patient-level and class-level error auditing should accompany aggregate accuracy reporting whenever a classifier of this kind is proposed for clinical or consumer use.

Keywords: electrocardiogram classification, error analysis, confidence calibration, patient safety, arrhythmia detection

1 Introduction

A wearable device on a runner's wrist can now flag an irregular heartbeat before the runner feels anything unusual. Automated electrocardiogram (ECG) classification, once confined to hospital telemetry, is embedded today in consumer wearables and clinical decision support systems that screen large populations for arrhythmia and other cardiac abnormalities [3, 8]. Public benchmarks have driven rapid progress on this task, and reported accuracy continues to climb [2, 9]. But an aggregate accuracy number is silent about who a model fails on. A classifier that reaches **95% overall accuracy** can still fail on the small fraction of cases that carry the most clinical weight, and nothing in a single aggregate metric reveals whether those failures are scattered evenly across patients or concentrated in a few. Therefore, in this paper, we ask: **are ECG classification errors randomly distributed across patients and classes, or do they follow systematic patterns that persist across independently collected datasets and different models?**

We hypothesize that **these errors are not random**. Instead, they concentrate in minority classes and in specific patients, and a model's confidence in its own prediction is an unreliable guide to whether that prediction is correct, with the degree of unreliability depending on which algorithm is used rather than on how accurate it is overall. To test this, we audit seven supervised classifiers on three independent, public

ECG datasets under one consistent, patient-level protocol, examining errors at the class level, the patient level, and the confidence level at once. This audit has three properties that distinguish it from standard benchmark reporting. First, it is **replicated** across cohorts collected in different countries with different equipment and label schemes, so a pattern that survives all three is unlikely to be a single-dataset artifact. Second, it is **model-diverse**, comparing seven structurally different classifiers on the same splits, so a pattern found with one algorithm can be checked against six unrelated alternatives. Third, it treats **confidence as an outcome** in its own right, testing directly whether a model's stated uncertainty predicts whether it is wrong, since an uncertain prediction can be flagged for review but a confidently wrong one cannot.

We evaluate this audit on MIT-BIH, PTB-XL, and LUDB. Three results stand out. First, **minority classes fail substantially more than majority classes** in every dataset. Second, **confidence calibration diverges sharply by model family** even when accuracy does not. Third, **a subset of patients is misclassified by every model**, and these patients tend to be older and to carry a heavier comorbidity burden. The confidence-calibration gap matters most: it means accuracy alone cannot tell a practitioner whether a deployed classifier's mistakes will announce themselves. This follows from a simple fact: aggregate accuracy is computed by averaging over exactly the patient-level and class-level structure that determines real-world safety, so two models can look equivalent on paper while behaving very differently once deployed. Patient-level and class-level error auditing deserves the same routine status in ECG classification research that accuracy and F1 score already have.

To our knowledge, this is among the first work to audit ECG classification errors jointly at the patient level, the class level, and the confidence level across multiple independently collected datasets under one consistent protocol. We make three contributions:

- We show that minority-class failure in ECG classification replicates across three independently collected cohorts with different recording equipment and label schemes, rather than being an artifact of any one dataset.
- We show that confidence calibration is largely decoupled from accuracy and is instead a model-family property, using a directly comparable seven-model zoo evaluated under identical splits.
- We identify a subset of patients that every tested model fails on, and show a descriptive association between this unanimous failure and older age and greater comorbidity burden.

These findings speak to two audiences. For clinicians, the finding that confidence cannot be trusted uniformly across algorithms is directly relevant to any deployment decision that relies on a classifier flagging its own uncertainty. For machine learning researchers, the fact that three independently collected, differently labeled datasets show the same class-level and patient-level failure structure suggests this reflects something about the underlying signal, not a quirk of one benchmark.

2 Related Work

Understanding ECG classification errors sits at the intersection of three threads: large-scale ECG benchmarking, confidence calibration, and patient-level auditing in clinical machine learning. Each has advanced on its own, but they are rarely combined into a single, cross-dataset protocol.

ECG benchmarking. MIT-BIH has been the reference dataset for beat-level arrhythmia classification for over two decades [6]. PTB-XL substantially expanded the scale and diagnostic scope of public ECG benchmarks, with a companion study reporting per-class performance across deep learning architectures [9, 11]. LUDB adds rhythm-level and comorbidity annotations absent from the larger benchmarks [5]. Each study reports per-class metrics within its own dataset; none tests whether a class-level or patient-level pattern reproduces on an independently collected one, which is the gap this paper addresses.

Confidence calibration. Modern classifiers, including deep networks, are frequently overconfident, and their predicted probabilities do not reliably track empirical correctness [1]. This matters clinically: a well-calibrated confidence score supports safe human-in-the-loop review, while a poorly calibrated one does not [4]. Existing calibration work largely studies one model family at a time, usually deep networks, and rarely compares calibration across structurally different classical families on the same data and splits, which is the comparison this paper provides.

Patient-level auditing. Deep learning models for arrhythmia detection report strong aggregate performance [2], and clinical machine learning increasingly recognizes the need for evaluation beyond a single summary statistic [8]. Patient-level auditing, testing whether the same patients fail across independently trained models and whether that failure associates with demographic variables, remains uncommon in the ECG literature. This paper makes that audit systematic, multi-dataset, and multi-model.

3 Method

3.1 Overview

This audit tests whether ECG classification errors follow a systematic structure that survives changes in dataset and model choice. Figure 1 shows the pipeline: patient-level feature extraction, a seven-model classification zoo trained under an identical protocol, and a three-level error audit spanning class, patient, and confidence.

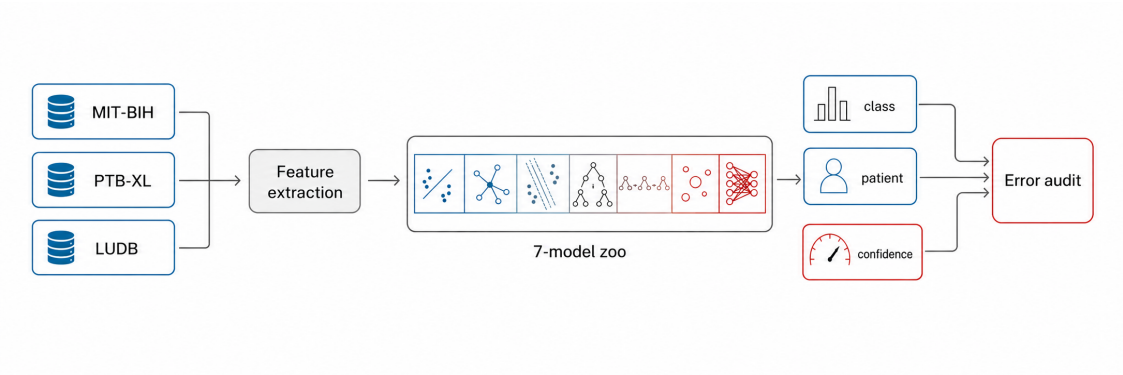

Figure 1: Pipeline diagram showing three datasets (MIT-BIH, PTB-XL subsample, LUDB) feeding into feature extraction, then a seven-model zoo (Logistic Regression, KNN, SVM, Random Forest, Gradient Boosting, AdaBoost, MLP), followed by a three-level error audit (class, patient, and confidence). The pipeline illustrates the complete cross-dataset evaluation workflow used in this study.

3.2 Datasets

We audit three publicly available, de-identified ECG datasets, chosen to span different countries, recording equipment, sampling rates, and label taxonomies. No new patient data were collected for this study.

MIT-BIH Arrhythmia Database. We use 44 non-paced records sampled at 360 Hz from a single lead, yielding 100,645 individually annotated heartbeats after feature extraction [6]. Each beat is labeled using the AAMI five-class scheme: normal (N), supraventricular ectopic (S), ventricular ectopic (V), fusion (F), and unknown or paced (Q).

PTB-XL. We use a stratified random subsample of 2,000 records, drawn from the full public dataset of 21,837 records across 18,885 patients while preserving the original proportions of diagnostic classes

[11]. The subsample spans 1,961 unique patients. Each record is a 10-second, 12-lead recording sampled at 100 Hz, labeled with one of five diagnostic superclasses: normal (NORM), myocardial infarction (MI), ST/T-wave changes (STTC), conduction disturbance (CD), or hypertrophy (HYP). We use a subsample rather than the full cohort for tractability, as discussed in Section 6.1.

LUDB. We use 200 patients, each with a single 10-second, 12-lead recording sampled at 500 Hz, labeled with a clinical rhythm diagnosis and accompanying demographic and comorbidity fields [5]. Rhythm classes with fewer than five records are excluded, leaving 187 records across four classes: sinus rhythm, sinus bradycardia, atrial fibrillation, and sinus arrhythmia.

Table 1 summarizes the three datasets.

Table 1: Summary of the three datasets used in this audit.

Dataset	Patients	Sampling rate	Leads	Label granularity
MIT-BIH	44 records	360 Hz	1	Beat-level arrhythmia type
PTB-XL (subsample)	1,961	100 Hz	12	Record-level diagnosis
LUDB	187	500 Hz	12	Record-level rhythm

3.3 Feature Extraction

For MIT-BIH, each annotated beat is represented by 12 features: waveform amplitude statistics (maximum, minimum, range, standard deviation, mean absolute value, skewness, kurtosis, energy) computed over a fixed window around the R-peak, together with RR-interval timing relative to the surrounding local rhythm. For PTB-XL and LUDB, each 12-lead record is represented by ten summary statistics per lead (mean, standard deviation, maximum, minimum, range, energy, skewness, kurtosis, sample-to-sample difference standard deviation, and zero-crossing count), for 120 features per record. We use hand-engineered features rather than raw waveforms so that all seven models in the zoo, most of which are not designed for raw sequence input, can be compared directly on the same representation.

3.4 Model Zoo

We train seven supervised classifiers on each dataset: logistic regression, k-nearest neighbors, a support vector machine with a radial basis function kernel, random forest, gradient boosting, AdaBoost, and a multilayer perceptron. These seven were selected because they represent the standard baseline set reported in the ECG classification literature [2, 3, 8], not because any one is individually state of the art. The goal is algorithmic diversity: any error pattern found in one model can be checked against six structurally different alternatives, spanning linear, distance-based, kernel, tree-ensemble, boosting, and neural methods.

For every dataset, all seven models are trained on an identical patient-level train and test split (75/25 for MIT-BIH and LUDB, 80/20 for PTB-XL), generated using group-aware splitting with patient identifier as the group. We verify zero overlap between train and test patients for every split before model training. Features are standardized using a scaler fit only on the training partition.

3.5 Three-Level Error Audit

For each trained model, we compute error rate, defined as one minus accuracy, within each class, within each patient, and overall. We define a prediction as confident if its maximum predicted class probability exceeds 0.8. We test the relationship between confidence and correctness with a chi-square test of independence. All seven models are compared directly on LUDB and on the PTB-XL subsample; MIT-BIH results are

reported for a single anchor model, gradient boosting, chosen for its representativeness among the ensemble methods in the zoo. We test per-patient error clustering with a permutation test using 2,000 permutations, comparing the observed variance of per-patient error rates against a null distribution obtained by randomly shuffling correct and incorrect labels while holding patient group sizes fixed. This test is not computed for LUDB, where each patient contributes exactly one record, which makes the per-patient variance statistic degenerate under permutation. Finally, we compare patients in the top and bottom deciles of mean error rate, averaged across all seven models, on available demographic and clinical fields, including age, sex, and dataset-specific comorbidity annotations.

3.6 Three Properties That Distinguish This Audit from Standard Benchmark Reporting

Compared with standard single-dataset ECG benchmark reporting, this audit has three properties. First, it is **replicated across independent cohorts**: the identical protocol is applied to three datasets collected under different conditions, which lets us distinguish a genuine pattern from a dataset-specific artifact. Second, it is **jointly patient-level and class-level**: beyond per-class accuracy, we compute per-patient error rates and test directly whether specific patients are unusually hard across every model in the zoo. Third, it **treats confidence as an outcome, not just a threshold**: we test directly whether confidence predicts correctness and whether that relationship is stable across model families, which is what determines whether a deployed system's uncertainty can be trusted.

4 Experiments

4.1 Implementation

All analyses are performed in Python 3 using scikit-learn [7] and SciPy [10]. Model hyperparameters are set to standard library defaults except where noted in Appendix A. All code and the analysis pipeline are available at `[PLACEHOLDER: repository URL]`.

4.2 Evaluation Protocol

We report **accuracy** and **macro-averaged F1 score** for overall model comparison, **error rate** within each class and each patient for the audit's core results, and the **percentage of errors that are confidently wrong** (maximum predicted probability exceeding 0.8) as the primary confidence-calibration metric. We report exact p-values for the chi-square and permutation tests rather than only significance thresholds, consistent with recommended practice for reporting statistical results in applied machine learning.

5 Results

We organize our results by claim. Each claim below corresponds to one of the findings summarized in Section 1.

5.1 Minority Classes Fail More Than Majority Classes Across All Three Datasets

On MIT-BIH, normal beats have the lowest error rate of any class across all seven models. Supraventricular, ventricular, fusion, and unknown beats all show substantially higher error, a pattern consistent across every model tested. On the PTB-XL subsample, the normal class ranges from **11.3% to 21.3%** error depending on model, while ST/T-wave change, myocardial infarction, and hypertrophy classes range from **47% to 95%** error. On LUDB, sinus rhythm, the majority class with 36 test records, ranges from **0% to 16.7%** error,

while the minority classes range from **57.1% to 100%** error across the seven models. Table 2 and Figure 2 report the full per-class breakdown for the anchor gradient boosting model. This pattern replicates across three independently collected datasets from different countries, different recording equipment, and different label taxonomies.

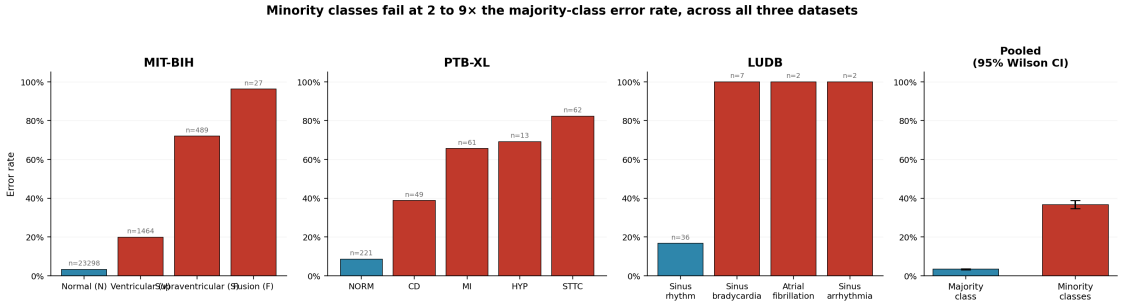

Figure 2: Error rate by class across all three datasets, anchor gradient boosting model. **Left to right:** MIT-BIH, PTB-XL, LUDB, each panel ordered majority to minority class. **Rightmost panel:** pooled minority vs. majority error rate across all three datasets with 95% confidence intervals. Minority classes fail at 2 to 9 times the rate of the majority class in every dataset tested.

Table 2: Error rate by class and dataset for the anchor gradient boosting model.

Dataset	Class	n (test)	Error rate
MIT-BIH	Normal (N)	23,298	0.033
MIT-BIH	Ventricular (V)	1,464	0.199
MIT-BIH	Supraventricular (S)	489	0.720
MIT-BIH	Fusion (F)	27	0.963
PTB-XL	NORM	221	0.086
PTB-XL	CD	49	0.388
PTB-XL	MI	61	0.656
PTB-XL	HYP	13	0.692
PTB-XL	STTC	62	0.823
LUDB	Sinus rhythm	36	0.167
LUDB	Sinus bradycardia	7	1.000
LUDB	Atrial fibrillation	2	1.000
LUDB	Sinus arrhythmia	2	1.000

5.2 Confidence Calibration Depends on Model Choice, Independent of Accuracy

For LUDB and the PTB-XL subsample, where all seven models are compared directly, AdaBoost is **never** confidently wrong: 0% of its errors are high-confidence on both datasets. Gradient boosting, in contrast, is confidently wrong in **39.9% to 88.2%** of its errors depending on dataset, despite comparable or better raw accuracy than AdaBoost on the same data (Table 3, Figure 3). A single anchor model evaluated on MIT-BIH shows a comparably high confidently-wrong rate (**56.5%** of errors), consistent with the pattern observed for gradient boosting elsewhere, though the full seven-model confidence comparison was not repeated on MIT-BIH in this study. On the PTB-XL subsample, a chi-square test confirms model confidence is a statistically

significant predictor of correctness overall ($\chi^2 = 51.52$, $p < 0.001$; $P(\text{correct} \mid \text{confident}) = 0.788$ versus $P(\text{correct} \mid \text{not confident}) = 0.432$). A substantial share of errors, **39.9%**, nonetheless occurs at high confidence. Two models with similar accuracy can therefore have very different practical safety profiles, since accuracy alone does not indicate whether a model's uncertainty can be trusted to flag its own mistakes.

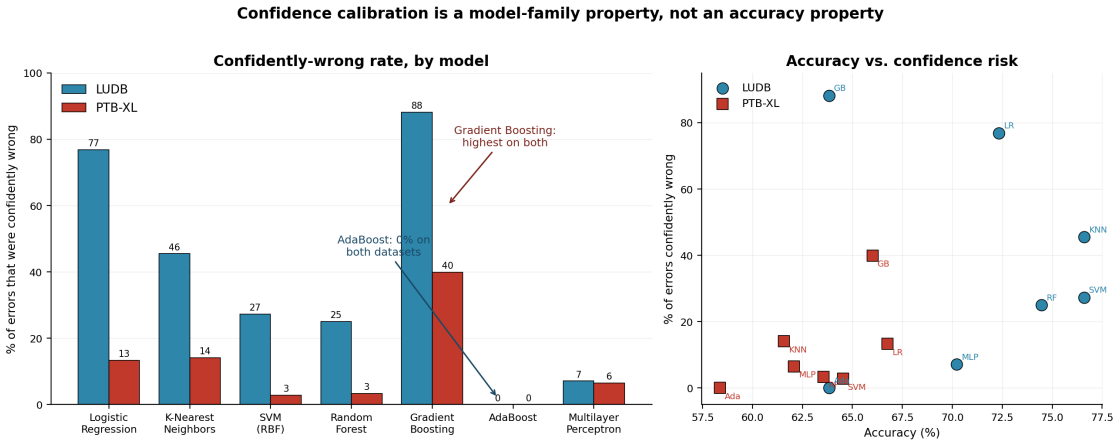

Figure 3: Confidence calibration by model, LUDB and PTB-XL. **Left:** percentage of errors that were confidently wrong (max predicted probability > 0.8), paired bars per model. **Right:** accuracy vs. confidently-wrong rate scatter, one point per model per dataset, showing calibration is uncorrelated with accuracy. AdaBoost sits at 0% on both datasets despite mid-range accuracy; gradient boosting is high-accuracy but high-risk.

Table 3: Percentage of errors that were confidently wrong (maximum predicted probability greater than 0.8), by model, for LUDB and the PTB-XL subsample.

Model	LUDB	PTB-XL
Logistic Regression	76.9%	13.3%
K-Nearest Neighbors	45.5%	14.1%
SVM (RBF)	27.3%	2.8%
Random Forest	25.0%	3.4%
Gradient Boosting	88.2%	39.9%
AdaBoost	0.0%	0.0%
Multilayer Perceptron	7.1%	6.5%

5.3 A Subset of Patients Is Misclassified by Every Model, in a Consistent Direction

On MIT-BIH, the patients with the highest error rate under the anchor gradient boosting model show *directional* rather than *scattered* misclassification. For example, one patient's ventricular beats are predominantly relabeled as normal rather than distributed across the other three error classes, and a second patient's normal and supraventricular beats are confused with each other in a consistent direction. On LUDB, where all seven models are compared directly, **eight of the 47** test records are misclassified by every model (Table 4). The permutation test of per-patient error clustering on the PTB-XL subsample does not reach statistical significance at this sample size and split ($p = 0.3855$), a result distinct from the unanimous cross-model misclassification directly observed in the ranked per-patient error table for LUDB. Consistent cross-model

failure on the same patients suggests that, for these patients, the difficulty lies in the underlying signal rather than in any single algorithm's idiosyncrasies.

Table 4: LUDB records misclassified by all seven models. All eight carry at least one additional clinical finding beyond their primary rhythm label.

Record	Primary label	Additional clinical findings
8	Atrial fibrillation	Paced rhythm, undefined ischemia (4 sites)
41	Sinus bradycardia	Left axis deviation, incomplete RBBB
54	Sinus bradycardia	Left axis deviation, incomplete LBBB, hypertrophy, ischemia
63	Sinus arrhythmia	Vertical electric axis
112	Atrial fibrillation	Left axis deviation, incomplete LBBB, hypertrophy, ischemia
125	Sinus bradycardia	Left axis deviation, conduction block, ventricular extrasystole, hypertrophy
133	Sinus bradycardia	Hypertrophy, repolarization abnormality
156	Sinus arrhythmia	Normal axis, incomplete RBBB

5.4 Patients with Unanimous Misclassification Are Older and Carry More Comorbidities

In LUDB, the 15 records with the highest mean error rate across all seven models have a mean age of **55.4 years**, compared with **47.1 years** for the 15 records with the lowest mean error rate. Nearly all of the hardest records carry at least one additional clinical finding, such as a conduction abnormality, hypertrophy, ischemia, or paced rhythm, beyond their primary rhythm label. In the PTB-XL subsample, the hardest patients have a mean age of **63.9 years**, compared with **55.1 years** for the easiest patients (Figure 4). This association is observed descriptively in two independent datasets but is not adjusted for confounders and should not be interpreted as a causal or formally tested finding.

6 Discussion

This audit found that ECG classification errors are not randomly distributed. Errors concentrate in minority classes across all three datasets tested, and a subset of patients is misclassified by every model evaluated. Confidence calibration, the degree to which a model's stated uncertainty predicts whether it is wrong, varies substantially by model family independent of raw accuracy. These findings raise three open questions that the present data can inform but not fully resolve.

Why does confidence calibration diverge so sharply by model when accuracy does not? AdaBoost was rarely confidently wrong across all three datasets, while gradient boosting was frequently confidently wrong despite similar or better accuracy. This is consistent with known differences in how boosting algorithms shape their decision margins, though the present study did not directly test this mechanism. Whether standard post hoc calibration methods, such as Platt scaling or temperature scaling, would close this gap without sacrificing accuracy remains unknown and is a direct next step.

What makes certain patients unanimously hard to classify across model families? The present data show an association with older age and greater comorbidity burden in both datasets where clinical annotations were available. Whether this reflects genuine physiological ambiguity in the ECG signal itself,

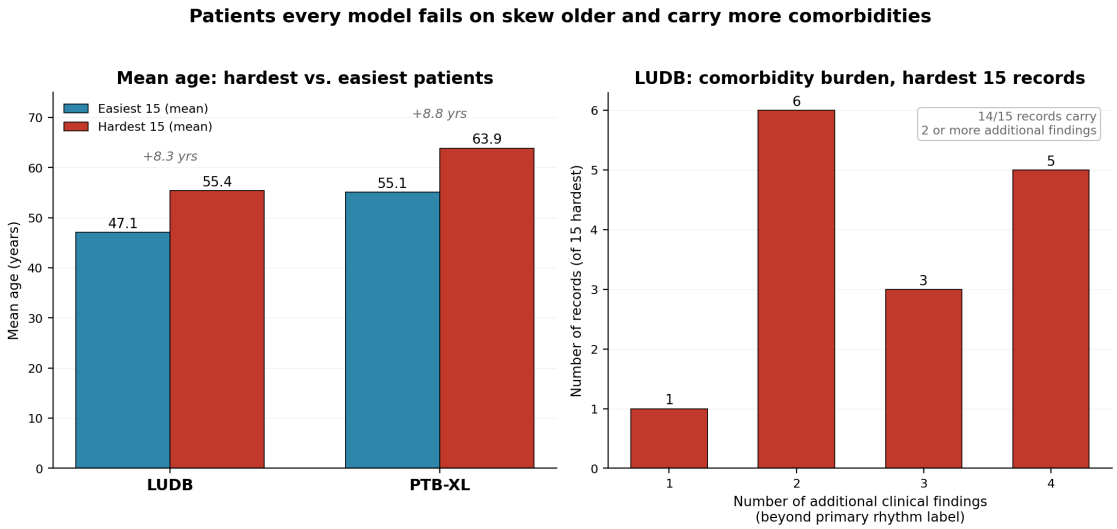

Figure 4: Age and comorbidity comparison, hardest vs. easiest patients (top and bottom deciles of mean error rate across all seven models). **Left:** mean age, LUDB and PTB-XL. **Right:** number of additional clinical findings among LUDB's 15 hardest records; 14 of 15 carry two or more findings beyond their primary rhythm label.

feature extraction that fails to capture the relevant signal characteristics in older or multi-morbid patients, or a training-data distribution that underrepresents this population, cannot be distinguished with the current analysis. A blinded waveform-level review by a cardiologist, without access to model output, would help determine whether the underlying signal is genuinely ambiguous for these patients.

Does this error structure persist at full dataset scale and with deep learning architectures? The present PTB-XL analysis used a stratified 2,000-record subsample and classical machine learning models trained on hand-engineered features, not the full 18,885-patient cohort or convolutional and transformer architectures trained directly on raw waveforms. Replicating this audit protocol on the full cohort, and on at least one deep learning architecture, would help settle the question. It would clarify whether the patterns reported here are a property of the underlying clinical data, or an artifact of the modeling choices made in this study.

6.1 Limitations

This study has several limitations. The PTB-XL subsample preserved class proportions from the full cohort but limited statistical power for the rarest class, hypertrophy, which had only 13 records in the test split. LUDB's smallest classes had as few as two test records, so its class-level error rates for atrial fibrillation and sinus arrhythmia should be read as indicative rather than precise. The permutation test for per-patient error clustering could not be computed for LUDB, since each patient contributes a single record there. The age and comorbidity association reported for the hardest patients is descriptive, was not adjusted for confounders, and does not establish a causal mechanism. Finally, all three datasets are public research benchmarks; findings may not generalize directly to proprietary clinical or wearable-device data, which typically carry different noise characteristics and patient populations.

7 Conclusion

Electrocardiogram classification errors are **not randomly distributed** across patients or classes. Minority diagnostic and rhythm classes fail at substantially higher rates than majority classes, a pattern that **replicates across three independently collected datasets**. Model confidence is a meaningful but incomplete safety signal, and **its reliability depends on which algorithm is used rather than on accuracy alone**. A subset of patients is misclassified consistently across model families, and these patients tend to be older and carry a greater comorbidity burden. Error auditing at the patient and class level should **accompany, not substitute for**, aggregate accuracy reporting whenever electrocardiogram classifiers are evaluated for clinical or consumer deployment.

References

[1] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1321–1330, 2017.

[2] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69, 2019. doi: 10.1038/s41591-018-0268-3.

[3] Shenda Hong, Yuxi Zhou, Junyuan Shang, Cao Xiao, and Jimeng Sun. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in Biology and Medicine*, 122:103801, 2020. doi: 10.1016/j.compbiomed.2020.103801.

[4] Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274, 2012. doi: 10.1136/amiajnl-2011-000291.

[5] Alena I Kalyakulina, Igor I Yusipov, Viktor A Moskalenko, Alexander V Nikolskiy, Konstantin A Kosonogov, Gennady V Osipov, Nikolai Yu Zolotykh, and Mikhail V Ivanchenko. Ludb: A new open-access validation tool for electrocardiogram delineation algorithms. *IEEE Access*, 8:186181–186190, 2020. doi: 10.1109/ACCESS.2020.3029211.

[6] George B Moody and Roger G Mark. The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001. doi: 10.1109/51.932724.

[7] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[8] Konstantinos C Siontis, Peter A Noseworthy, Zachi I Attia, and Paul A Friedman. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7):465–478, 2021. doi: 10.1038/s41569-020-00503-2.

[9] Nils Strodthoff, 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 (5):1519–1528, 2021. doi: 10.1109/JBHI.2020.3022989.

[10] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[11] Patrick Wagner, 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(1):154, 2020. doi: 10.1038/s41597-020-0495-6.

A Hyperparameters

Table 5: Model hyperparameters used across all experiments.

Model	Key hyperparameters
Logistic Regression	`max_iter=3000`
K-Nearest Neighbors	`n_neighbors=9`
SVM (RBF)	`kernel='rbf', probability=True`
Random Forest	`n_estimators=300, max_depth=14`
Gradient Boosting	`n_estimators=200, max_depth=4`
AdaBoost	`n_estimators=300`
Multilayer Perceptron	`hidden_layer_sizes=(128,64), max_iter=800`