Cross-Dataset Generalization of ECG Arrhythmia Classification: How Much Does Accuracy Degrade from MIT-BIH to PTB-XL, and Which Arrhythmia Subtypes Are Least Robust?

Sanjeet Jayaseelan

Homestead High School

Abstract

Most published ECG arrhythmia classifiers are trained and tested on a single database and report >95% accuracy, but this design cannot show whether a model has learned transferable cardiac physiology or dataset-specific artifacts. We trained a deliberately transparent random-forest classifier on hand-engineered R-R interval and QRS-morphology features from the MIT-BIH Arrhythmia Database (inter-patient DS1/DS2 split) and evaluated it both in-dataset (held-out MIT-BIH) and cross-dataset on an independently collected database (PTB-XL). Macro-averaged F1 fell from 0.61 to 0.35 across databases. The degradation was large and uneven, and — against naive intuition — the ventricular-ectopic (V) class, the second-strongest class in-dataset (F1 0.77), collapsed cross-dataset (F1 0.23, recall 0.26), while an apparent gain in the supraventricular (S) class was traced to a label-composition confound rather than genuine robustness. A controlled experiment that matched the training sampling rate to PTB-XL (downsampling MIT-BIH to 100 Hz) left cross-dataset performance essentially unchanged (macro-F1 0.3546 to 0.3562, a difference of 0.0016), and V recall moved slightly in the wrong direction for a resampling-artifact explanation (0.257 to 0.236, a decline rather than the recovery that hypothesis would predict). This is evidence, though not proof, that the collapse reflects population/label-distribution shift rather than an acquisition artifact, since the control does not equalize beat-detection method between the two databases. Per-class, mechanism-level error analysis with an interpretable model gives a more informative picture of ECG domain shift than a single black-box accuracy.

1. Introduction

Deep and classical machine-learning models routinely exceed 95% accuracy on single-database ECG arrhythmia benchmarks. Whether that performance reflects transferable cardiac knowledge or overfitting to one hospital's recording equipment and patient population is an active clinical-ML concern known as domain shift. A structured map of the recent (2021-2026) generalization/domain-shift ECG subfield — 85 papers screened from Europe PMC and arXiv — shows the field expanding rapidly (>75% of papers from 2025-2026) but with two consistent gaps: it rarely reports per-class degradation across a full arrhythmia taxonomy, and it rarely resolves the label-harmonization problem between databases explicitly. Reported cross-dataset failures range from graceful to catastrophic, motivating a careful, interpretable per-class study.

Hypothesis. Cross-dataset performance will drop substantially (not a few points), and the drop will be uneven — some arrhythmia subtypes will degrade far more than normal sinus rhythm.

2. Data and Methods

2.1 Datasets

MIT-BIH Arrhythmia Database — 48 records, 360 Hz, 2-lead (lead MLII used); beat-level annotations. Training source.

PTB-XL — 21,799 records, 100 Hz, 12-lead (lead II used); strip-level SCP-ECG diagnostic/rhythm statements. Independent test source (different hospital, population, equipment, and era).

2.2 Label Harmonization

The databases are annotated at different granularity: MIT-BIH labels every heartbeat; PTB-XL labels each 10-second strip. We adopted the AAMI beat-class grouping (N/S/V/F/Q) for MIT-BIH and built an explicit crosswalk from PTB-XL SCP codes to AAMI classes, shown in Table 1.

Table 1. AAMI class crosswalk between MIT-BIH beat symbols and PTB-XL SCP codes.

AAMI class	MIT-BIH beat symbols	PTB-XL SCP codes
N (normal)	N, L, R, e, j	SR, SBRAD, SARRH
S (supraventricular)	A, a, J, S	PAC, AFIB, AFLT, SVARR, SVTAC, PSVT, STACH
V (ventricular)	V, E	PVC, BIGU, TRIGU

Paced (PACE) records were excluded as non-comparable. Each PTB-XL strip received a single AAMI label by presence of these codes, prioritizing ventricular over supraventricular over normal. Analysis was restricted to the three classes (N, S, V) shared and well-populated across both databases; F and Q were near-absent and non-comparable. On PTB-XL, QRS complexes were detected on lead II with the WFDB XQRS detector, and each detected beat inherited its strip's AAMI label. This inheritance is deliberately noisy — a strip labeled "PVC" contains mostly normal beats — and is treated as a stated limitation (Section 6), not a hidden one.

2.3 Features and Model

For every beat we computed seven features from a 0.5-40 Hz band-passed signal: preceding R-R interval, following R-R interval, local average R-R (±5 beats), R-R ratio (pre/local), pre/post R-R ratio, QRS amplitude, and an energy-based QRS-width proxy. The classifier was a random forest (300 trees, balanced class weights, minimum leaf 5). MIT-BIH training used the inter-patient DS1 (train) / DS2 (test) split of de Chazal et al. [1], so that no patient appears in both training and the in-dataset test set.

2.4 Sampling

All PTB-XL V (14,868 beats) and S (38,084 beats) beats were used; N was sub-sampled (2,500 strips) for download tractability, yielding 85,181 test beats from 6,212 records.

2.5 Control Experiment

To separate acquisition shift (360 Hz vs 100 Hz) from population/label shift, we repeated the full pipeline with MIT-BIH signals downsampled to 100 Hz before feature extraction and training, matching PTB-XL's sampling rate, and re-evaluated cross-dataset.

3. Results

In-dataset (held-out MIT-BIH DS2): macro-F1 0.6064 (→0.61), balanced accuracy 0.63. N and V classified well (F1 0.961 and 0.765); S was already weak inter-patient (F1 0.093), consistent with the known difficulty of the supraventricular class.

Cross-dataset (PTB-XL): macro-F1 fell to 0.3546 (→0.35) and balanced accuracy to 0.380 — a drop of roughly 0.25 macro-F1, confirming that degradation is substantial, not cosmetic.

Table 2. Per-class F1 degradation from in-dataset to cross-dataset evaluation.

AAMI class	In-dataset F1	Cross-dataset F1	Absolute drop	% drop
N (normal)	0.96	0.50	0.46	48%
S (supraventricular)	0.09	0.34	-0.25	-266%
V (ventricular)	0.77	0.23	0.54	70%

Table 3. Cross-dataset confusion matrix (360 Hz-trained model, rows = true, columns = predicted).

	pred N	pred S	pred V
true N	20,854	3,578	7,797
true S	21,782	8,980	7,322
true V	8,913	2,141	3,814

4. Error Analysis: Which Classes Collapse, and Why

V (ventricular) is the least robust class — the counter-intuitive headline. V was the model's second-best class in-dataset (F1 0.765) but collapsed cross-dataset (F1 0.226, recall 0.257); 60% of true PTB-XL V beats (8,913 of 14,868) were misclassified as normal. Because the QRS-width proxy and amplitude are among the top morphology features, an obvious candidate mechanism was the sampling-rate difference (a wide ventricular QRS is represented by ~3.6× fewer samples at 100 Hz than at 360 Hz). Section 5 addresses this directly.

Figure 1 makes the mechanism concrete at the feature level. PTB-XL's strip-inherited V beats have a median QRS-width proxy of 50.0 ms and a median RR ratio of 1.00 — both consistent with normal morphology and non-premature timing — versus MIT-BIH's expert-annotated V beats at a median QRS width of 97.2 ms and RR ratio of 0.72. In other words, the beats inheriting a "ventricular" label from their parent PTB-XL strip mostly do not exhibit the wide, premature morphology that defines ventricular ectopy in the training data. This is a direct, feature-level consequence of the strip-to-beat label inheritance described in Section 2.2: it explains the collapse more precisely than population shift stated in general terms.

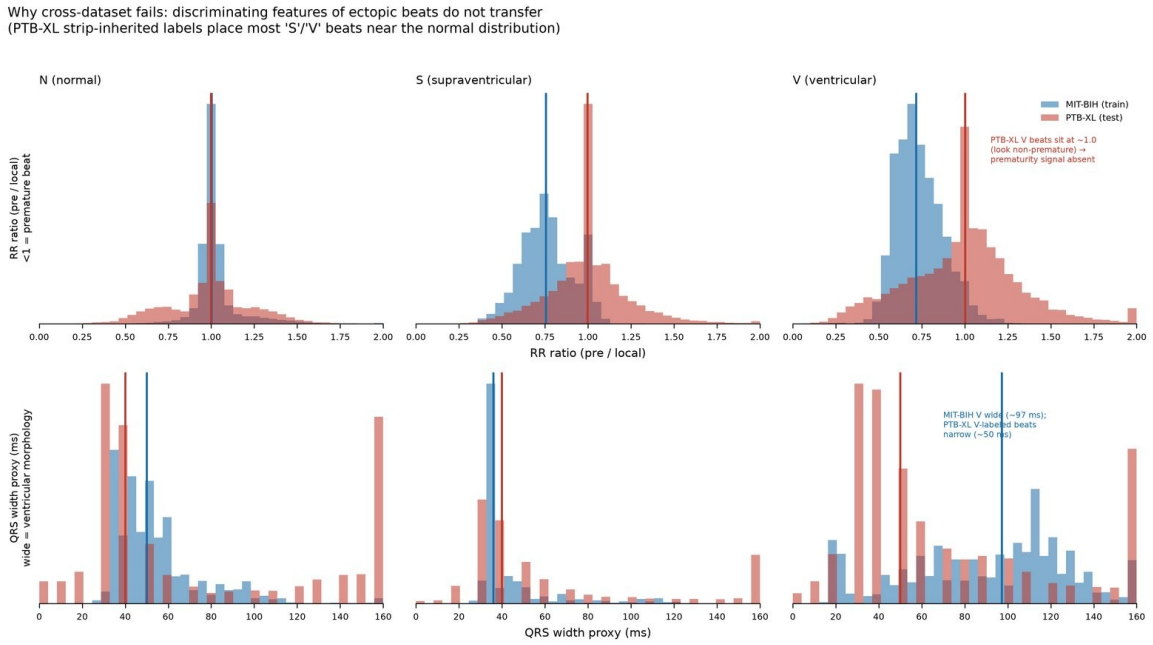

Figure 1. Distribution of RR ratio and QRS-width proxy by AAMI class, MIT-BIH (train) vs. PTB-XL (test). Vertical lines mark class medians.

S (supraventricular) "improves" — but this is a label-composition artifact, not robustness. Cross-dataset S-F1 rose (0.093 → 0.340), which is surprising at face value. The cause: PTB-XL's S-labeled records are dominated by atrial fibrillation/flutter, whose grossly irregular R-R intervals are easy for R-R-based features to catch, whereas MIT-BIH's S class is dominated by isolated premature atrial contractions (PACs) — subtle single-beat events. The two databases' "S" classes are physiologically different populations. This is a textbook domain-shift confound: the label name is stable across databases but its underlying distribution is not, and it would silently inflate a headline metric if per-class composition were not examined.

N (normal) degrades as expected. F1 fell 0.961 → 0.498, with normal beats increasingly misclassified as S or V under PTB-XL's noisier acquisition and beat-detection.

5. Control: Is V's Collapse a Resampling Artifact or Genuine Population Shift?

Downsampling MIT-BIH to 100 Hz before training (matching PTB-XL) left cross-dataset performance essentially unchanged: macro-F1 moved from 0.3546 to 0.3562 (Δ0.0016) — a difference small enough to be noise, not a meaningful rise. Per-class cross-dataset F1 under 100 Hz-matched training was N 0.4946 (Δ-0.0032), S 0.3524 (Δ0.0121), V 0.2215 (Δ-0.0042) — every class within 0.013 of the 360 Hz-trained model.

Critically, V recall did not improve when the sampling-rate mismatch was removed — it declined slightly, from 0.257 (360 Hz-trained) to 0.236 (100 Hz-trained). If the ventricular collapse were caused by the 360 → 100 Hz acquisition mismatch, training the model natively at 100 Hz should have recovered V performance substantially. It did not move in that direction at all. This is evidence against the resampling-artifact hypothesis and for genuine population/label-distribution

shift between the two databases — though as with any single control, it narrows the explanation rather than proving it (see limitation below).

6. Limitations

1. Strip-to-beat label inheritance injects label noise on PTB-XL; cross-dataset numbers are a lower bound on achievable per-beat performance and are best read as relative degradation.

2. The 100 Hz control isolates sampling rate but not beat-detection method. MIT-BIH beats come from expert .atr annotations in both the main experiment and the control; PTB-XL beats in both cases come from xqrs-detected positions with strip-inherited labels. The control therefore cleanly isolates the sampling-rate variable, but it does not rule out that detection-method or label-inheritance differences (rather than sampling rate) drive part of the cross-dataset gap — that remains a separate, uncontrolled source of shift.

3. N sub-sampling on PTB-XL means absolute prevalence-weighted metrics would differ; class-wise F1 and recall (reported here) are unaffected.

4. Only three AAMI classes were comparable; F and Q were excluded.

5. A single interpretable feature model was used by design; a CNN might shift absolute numbers but the per-class relative pattern — the object of study — is the contribution.

7. Conclusion

Training-database performance overstates real-world ECG arrhythmia performance: macro-F1 fell from 0.61 to 0.35 across independently collected databases, confirming the hypothesis that cross-dataset degradation is substantial and uneven. The pattern is partly counter-intuitive — the morphologically defined ventricular class, strong in-dataset, was the least robust, while an apparent gain in the supraventricular class was an artifact of differing label composition. A controlled resampling experiment found no evidence that the ventricular collapse is a sampling-rate artifact — if anything, matching the rate nudged V recall slightly lower — and feature-level analysis shows PTB-XL's inherited V labels are attached to beats that do not exhibit the wide, premature morphology defining true ventricular ectopy, pointing toward genuine population/label shift. Reporting a single cross-dataset accuracy would have hidden all of these findings. Per-class, mechanism-level error analysis with an interpretable model is not a lesser alternative to a black-box benchmark but a more informative one — and identifying which subtypes fail, and why, is the actionable output for anyone deploying an ECG classifier across sites.

References

1. de Chazal P, O'Dwyer M, Reilly RB. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Trans Biomed Eng. 2004;51(7):1196-1206.

2. Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol Mag. 2001;20(3):45-50.

3. Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, Schaeffter T. PTB-XL, a large publicly available electrocardiography dataset. Sci Data. 2020;7(1):154.