Beyond MIT-BIH: Failure Modes of Beat-Level ECG Arrhythmia Classification at Scale

Angus Nicolson[1], Riccardo Lunelli[1], Samuel Martin Pröll[1], Axel Bauer[2], Clemens Dlaska[1]

[1] Digital Cardiology Lab, Medical University of Innsbruck, Innsbruck, Austria
[2] University Clinic of Internal Medicine III, Medical University of Innsbruck, Innsbruck, Austria

Abstract

Beat-level ECG arrhythmia classification is benchmarked almost exclusively on small datasets: most deep learning studies rely on MIT-BIH, a 48-recording, single-centre database. Whether these models generalise beyond this dataset is largely unknown. We evaluate this directly, fine-tuning xECG, an xLSTM-based foundation model, on INCART and MIT-BIH (inter-patient split, AAMI classes N, V, S, F, Q) and assessing beat-level predictions at scale on MIMIC-IV-ECG, CPSC 2018, and MIT-BIH SVDB.

This exposes two distinct failure modes. First, patch-based models process fixed-length windows, so a single beat may span consecutive patches, causing double-counting. Suppressing consecutive same-class predictions fails at elevated heart rates, when the RR interval approaches the patch duration. On MIMIC-IV-ECG, a patch-only model achieves beat-count Pearson $r = 0.686$, collapsing to $r = -0.484$ under tachycardia (HR>120 bpm); training R-peak detection jointly with classification raises these to $r = 0.987$ and $r = 0.865$. On MIT-BIH, the joint model achieves macro F1 0.550 (V-class F1 0.974) and R-peak F1 >0.999 (150 ms tolerance).

Second, and independently, distribution shift substantially reduces precision. On CPSC 2018, recording-level ventricular ectopic beat (VEB) detection yields sensitivity 0.963 and PPV 0.452 (F1=0.615). False positives are dominated by right bundle branch block and atrial fibrillation. For MIMIC-IV-ECG (pacemaker-coded recordings excluded; F1=0.716, sensitivity=0.855, PPV=0.616), false positives are dominated by conduction defects and atrial fibrillation. Joint R-peak detection does not address this failure.

These results demonstrate that standard beat-level benchmarks substantially overstate clinical utility and that robust arrhythmia classification requires both reliable beat localisation and training data that reflects real-world rhythm diversity.

1. Introduction

Most deep learning studies of beat-level ECG arrhythmia classification are benchmarked on the MIT-BIH Arrhythmia Database [1, 2]: 48 half-hour recordings from 47 subjects at a single centre. MIT-BIH has been invaluable, but it is small, decades old, and drawn from a narrow patient population. Whether models that perform well on MIT-BIH generalise to the diversity of routine clinical practice is largely untested, because independent beat-level evaluation at scale is not generally available.

In this paper we fine-tune xECG, an xLSTM-based [3] ECG foundation model [4], for beat-level arrhythmia classification and evaluate it far beyond MIT-BIH: at scale on MIMIC-IV-ECG [2, 5] (736,382 recordings), on CPSC 2018 [6], and on MIT-BIH SVDB [2, 7]. This exposes two failure modes that MIT-BIH alone would not reveal: a mechanical failure from how patch-based models tokenise the signal, corrupting beat counts under tachycardia, and a precision failure from distribution shift between training and deployment populations, independent of the first. Closing them requires beat-level data reflecting real-world patient and rhythm diversity better than MIT-BIH can.

2. Methods

2.1. Datasets

Training. xECG is fine-tuned for joint R-peak detection and beat classification on the St. Petersburg INCART 12-lead Arrhythmia Database [2] together with MIT-BIH [1, 2], using the standard inter-patient DS1/DS2 split [8] for MIT-BIH: DS1 (22 records) is split into 20 training and 2 validation records, and DS2 (22 records) is held out for testing. All 75 INCART records (32 patients) are used for training. Beats are labelled using the five AAMI superclasses: N (normal), S (supraventricular ectopic), V (ventricular ectopic), F (fusion), and Q (unknown).

We train four additional model variants, changing only the training set. Each variant and the baseline are trained with three random seeds, and Section 3.3 reports the range

across seeds; Tables 1–3 report a single baseline run. The first adds the four paced MIT-BIH records that DS1/DS2 excludes. The next two exclude one or both of the INCART patients carrying RBBB-labelled beats (2 and 4 records), leaving 8 and 0 RBBB-labelled beats in the INCART training data (3,174 in the full corpus). The fourth is a control excluding two INCART patients *without* RBBB, matched for the number of records.

Evaluation. We evaluate beat-level predictions on the held-out MIT-BIH DS2 split (22 records) and on MIT-BIH SVDB [2, 7] (78 records), a companion database from a different, supraventricular-ectopy-enriched patient population. MIMIC-IV-ECG [2, 5] is used for beat-count agreement (735,000 recordings with a valid RR-derived reference count) and, via its machine-generated report, for recording-level VEB detection: 729,313 recordings remain after excluding 7,069 flagged as poor tracing quality. The report does not complete rhythm analysis once a pacemaker is detected, so pacemaker-coded recordings have no reliable VEB ground truth; excluding these leaves 700,380 recordings, our primary MIMIC-IV-ECG cohort. CPSC 2018 [6] (6,871 recordings) provides expert-labelled recording-level diagnoses for our VEB evaluation.

2.2. Models

xECG takes all twelve leads as input, resampled to 100 Hz, and processes them as a sequence of fixed-length, non-overlapping 250 ms patches (25 samples). Leads not present in a database are zero-filled, so the same network is applied unchanged to two-lead (MIT-BIH, SVDB) and twelve-lead (INCART, CPSC 2018, MIMIC-IV-ECG) recordings. A *patch-only* model predicts one of six classes per patch: the five AAMI superclasses, plus M for a patch containing no beat (between-beat). Adjacent patches predicted with the same class are merged into a single beat by a same-class suppression heuristic, since a beat can span more than one patch. This is exact only when every true beat maps to one contiguous run of same-class patches. It breaks down at elevated heart rate, when two adjacent beats of the same class fall in adjacent patches (Figure 1).

We compare this against a *joint* model, which adds a second, independent output head trained to detect R-peaks directly, alongside the beat-classification head, so that beat localisation no longer depends on runs of same-class patch predictions. The R-peak head is trained with soft-labelled targets rather than a single hard-labelled position.

2.3. Evaluation metrics

We score predictions at three levels depending on available labels:

Beat level. Predicted beat positions are greedily matched to ground-truth beat positions by temporal proximity. We report R-peak detection F1 (presence only) and beat classification F1 (presence and correct AAMI class), both at 150 ms tolerance.

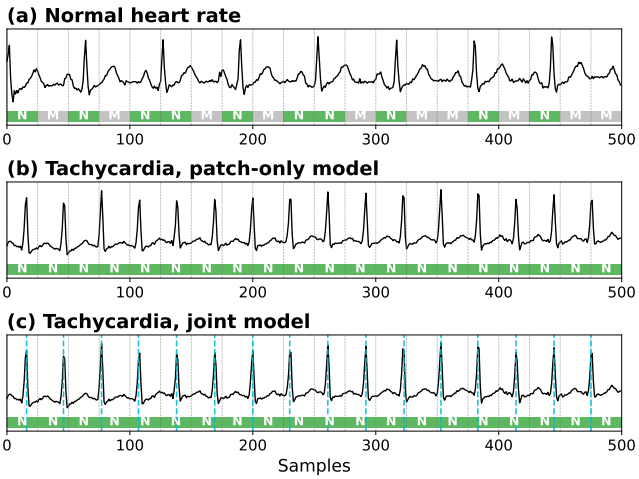

Figure 1. Model predictions illustrating the double-counting failure mode. (a) At normal heart rate, patch-level predictions alternate between N (beat) and M (between-beat). (b) Under tachycardia, the patch-only model predicts every patch as N, so consecutive beats are merged. (c) On the same segment, the joint model's R-peak head (dashed lines) correctly localises each beat.

Table 1. MIT-BIH DS2 (22 records) and SVDB (78 records), joint model (soft R-peak labels), 150 ms tolerance, 95% CI in brackets.

Metric	MIT-BIH DS2	SVDB
Macro F1	0.550 [0.427,0.648]	0.453 [0.419,0.484]
N	0.971 [0.934,0.995]	0.968 [0.954,0.978]
S	0.252 [0.137,0.658]	0.490 [0.373,0.606]
V	0.974 [0.935,0.991]	0.809 [0.733,0.868]
F	0.554 [0.018,0.795]	0.000 [0.000,0.000]
Q	0.000 [0.000,0.000]	0.000 [0.000,0.000]
R-peak	1.000 [0.999,1.000]	0.999 [0.998,1.000]

Count level. Total predicted vs. total ground-truth beat count per recording, summarised across recordings by Pearson r. This does not require R-peak labels so can be used in datasets with no positional information.

Recording-level VEB detection. Beat-level V predictions are aggregated to a per-recording binary label, positive if the recording contains at least one predicted V beat, and compared against recording-level labels (expert annotations for CPSC 2018; machine-generated report labels for MIMIC-IV-ECG). We report sensitivity, PPV and F1.

Confidence intervals. 95% CIs use a bootstrap over recordings (not beats, which are correlated within a recording), resampled with replacement 1000 times.

3. Results

3.1. Beat level

Table 1 summarises joint-model performance on MIT-BIH DS2 and SVDB. R-peak detection is close to ceiling on both. Macro F1 is held down by the rarer classes (S, F, Q; wide CIs reflect their low support). Critically, V-class F1, for which we have more support, falls from 0.974 on MIT-BIH DS2 to 0.809 on SVDB.

3.2. Double-counting under tachycardia

Table 2 reports total beat-count agreement. The patch-only model agrees reasonably overall but inverts under tachycardia: as heart rate rises, same-class suppression increasingly merges distinct adjacent beats (Figure 1b), so predicted counts fall further below true counts as true counts rise. The joint R-peak head restores agreement both overall and under tachycardia, and soft R-peak targets improve further on hard single-sample ones, most markedly under tachycardia; all remaining results use the soft-label joint model.

Table 2. Model comparison: beat-count Pearson r, MIMIC-IV-ECG ($n = 735{,}000$), 95% CI in brackets.

Model	Overall r	Tachycardia r (HR>120)
Patch-only	0.686 [0.682,0.690]	−0.484 [−0.498,−0.471]
Joint (hard labels)	0.964 [0.964,0.965]	0.816 [0.803,0.828]
Joint (soft labels)	0.987 [0.986,0.987]	0.865 [0.851,0.878]

3.3. Recording-level VEB detection

Aggregating beat-level V predictions to a recording-level VEB flag reveals a second, independent failure mode (Table 3, Figure 2). CPSC 2018 gives high sensitivity (0.963) and low PPV (0.452), with false positives dominated by right bundle branch block (47.0%) and atrial fibrillation (37.5%), and premature atrial contractions a smaller contributor (15.9%). MIMIC-IV-ECG repeats the pattern: on the primary cohort, 64.6% of false positives carry a coded conduction defect, 29.6% atrial fibrillation and 18.0% supraventricular ectopy (a recording may carry several codes, so these need not sum to 100%). Excluding recordings flagged with a conduction defect raises PPV from 0.616 to 0.743 and sensitivity from 0.855 to 0.873 (F1 0.716 to 0.803; $n = 558{,}770$), despite VEB prevalence falling from 7.9% to 6.2% in the retained cohort. The 10,427 residual false positives remain dominated by atrial fibrillation (30.8%) and supraventricular ectopy (21.7%). Figure 2 shows representative examples.

Pacing is an additional confounder: the excluded pacemaker-coded recordings from MIMIC-IV-ECG have no VEB ground truth, so we cannot validate on them, yet the model predicts at least one V beat in 59–63% (range

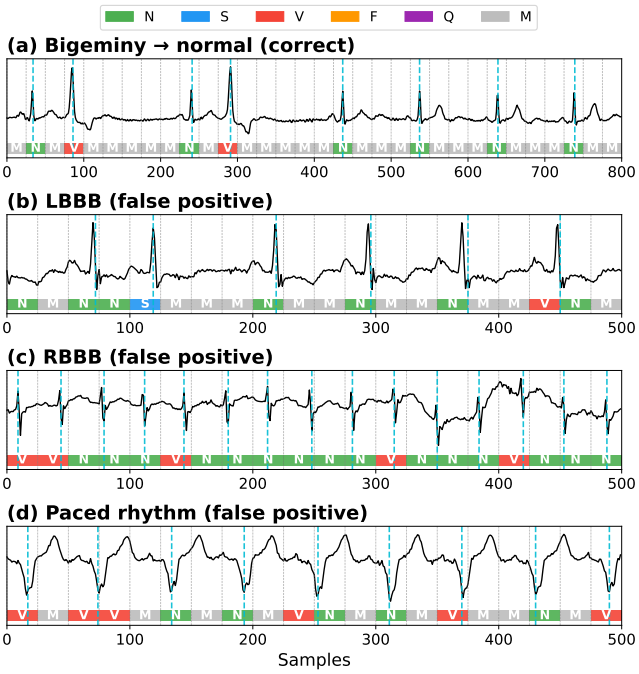

Figure 2. Joint model predictions illustrating common VEB false positives. (a) A correct example, where the model switches from bigeminy to sinus rhythm. The remaining panels are false positive examples for: (b) LBBB, (c) RBBB, and (d) paced rhythms.

is across training seeds) of them, against 10.6–10.9% of the primary cohort (Figure 2d). Training on the four paced MIT-BIH records does not fix this: the retrained model predicts the paced class in only 1.9–2.7% of paced recordings, and its paced V-prediction rate falls to 58.6–59.1% only alongside a proportionally larger fall in its overall V-prediction rate (10.0–10.3%), so the model becomes globally more conservative with no pacing-specific gain.

Excluding the INCART patients carrying RBBB-labelled beats degrades RBBB detection specifically. Across three seeds per arm, RBBB-coded recordings account for 43–49% of CPSC false positives at baseline, rising to 55–58% when one such patient is withheld and 57–63% when both are; CPSC PPV falls from 0.452–0.462 to 0.414–0.445 and 0.338–0.381. Withholding two INCART patients *without* RBBB, who contribute more beats than the two RBBB patients (7,517 vs. 6,531), leaves the RBBB share unchanged (42–47%), so the effect follows the morphology withheld rather than the reduction in training data.

Joint R-peak detection does not fix the lack of precision for VEB detection: the patch-only model reaches precision at least as high (CPSC 0.515; MIMIC primary cohort 0.681) despite its count errors (Table 2). Precision is governed by morphological confusion, not beat localisation.

Table 3. Recording-level VEB detection, joint model (soft R-peak labels), 95% CI in brackets.

Dataset	n	Sensitivity	PPV	F1
CPSC 2018	6,871	0.963 [0.948,0.976]	0.452 [0.426,0.476]	0.615 [0.591,0.638]
MIMIC-IV-ECG (pacemaker excluded)	700,380	0.855 [0.852,0.858]	0.616 [0.612,0.619]	0.716 [0.713,0.719]
MIMIC-IV-ECG (+conduction defect excluded)	558,770	0.873 [0.870,0.877]	0.743 [0.739,0.747]	0.803 [0.800,0.806]

4. Discussion

Standard benchmarking on MIT-BIH alone would have missed both failure modes. Double-counting appears only at the patch/RR-interval ratios MIT-BIH does not stress-test at scale; the precision collapse appears only once the model meets the rhythm diversity of MIMIC-IV-ECG or CPSC 2018. Held-out splits do not close the gap: the model has close to perfect V-beat prediction on MIT-BIH DS2, but performs substantially worse on SVDB and flags between a third and half of its recording-level VEB alarms falsely at scale across CPSC 2018 and MIMIC-IV-ECG.

MIT-BIH has many annotated beats but few patients (47 subjects), and most beat-level models train on it alone. Adding all of INCART does not help, because the confounders that degrade performance are represented in too few patients. Ventricular ectopic beats look markedly different under left bundle branch block than under normal conduction. INCART has no paced beats, and DS1/DS2 excludes MIT-BIH's four paced records; adding them back does not substantially improve the model: four records from one centre cannot represent paced morphology. Bundle-branch-block morphology is present but concentrated: L/R-labelled beats are 4.8% of the combined corpus ($10{,}906/226{,}895$), contributed by 6 of its 54 patients, and all but 8 of INCART's $3{,}174$ RBBB-labelled beats come from one patient. Removing the RBBB patients from training specifically increases RBBB-coded false positives in CPSC 2018, whereas removing two comparable patients without RBBB does not. Three MIT-BIH patients with RBBB remain throughout: not a complete absence of the morphology, but a shortage of patients carrying it. Patient count constrains what a corpus can teach far more tightly than beat count does.

Progress requires beat-level datasets with more *patients* carrying each confounder, not simply more beats or recordings. Icentia11k [9] reaches this patient scale, but is single-lead, outside the diagnostic 12-lead setting most clinical models target. The pattern is not new: across eight databases, Llamedo and Martinez [10] found ventricular positive predictive value ranging from 0.96 to 0.11 with no retraining, and argued that broad multi-database evaluation is necessary to characterise performance at all.

Our MIMIC-IV-ECG evaluation relies on machine-generated reports as ground truth, expert annotation being unavailable at this scale; these are imperfect and affect both the VEB metrics and the FP composition. Corroboration against expert-labelled cohorts such as CPSC 2018 mitigates this, but a large, diverse, expert-annotated beat-level dataset remains missing from the field.

References

[1] Moody G, Mark R. The impact of the MIT-BIH Arrhythmia Database. IEEE Engineering in Medicine and Biology Magazine 2001;20(3):45–50.

[2] Pollard T, Moody BE, Lehman LwH, Gow BJ, Fernandes C, Xie C, Johnson A, Mark RG, Heldt T. PhysioNet as a global platform for biomedical research. Nature Health 2026;1(8):792–795. ISSN 3005-0693.

[3] Beck M, Pöppel K, Spanring M, Auer A, Prudnikova O, Kopp M, Klambauer G, Brandstetter J, Hochreiter S. xLSTM: Extended long short-term memory. In NeurIPS. 2024; .

[4] Lunelli R, Nicolson A, Proell SM, Reinstadler SJ, Bauer A, Dlaska C. BenchECG and xECG: a benchmark and baseline for ECG foundation models. arXiv 2025;.

[5] Gow B, Pollard T, Nathanson LA, Johnson A, Moody B, Fernandes C, Greenbaum N, Waks JW, Eslami P, Carbonati T, Chaudhari A, Herbst E, Moukheiber D, Berkowitz S, Mark R, Horng S. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. PhysioNet 2023;.

[6] Liu F, Liu C, Zhao L, Zhang X, Wu X, Xu X, Liu Y, Ma C, Wei S, He Z, Li J, Kwee E. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. Journal of Medical Imaging and Health Informatics 2018;8(7):1368–1373.

[7] Greenwald S. MIT-BIH supraventricular arrhythmia database. PhysioNet, 1990.

[8] de Chazal P, O'Dwyer M, Reilly R. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Transactions on Biomedical Engineering 2004;51(7):1196–1206.

[9] Tan S, Androz G, Chamseddine A, Fecteau P, Courville A, Bengio Y, Cohen JP. Icentia11k: An unsupervised representation learning dataset for arrhythmia subtype discovery. arXiv preprint, 2019. arXiv:1910.09570.

[10] Llamedo M, Martinez JP. An automatic patient-adapted ECG heartbeat classifier allowing expert assistance. IEEE Transactions on Biomedical Engineering 2012;59(8):2312–2320.

Address for correspondence:

Angus Nicolson
angus.nicolson@i-med.ac.at