Reconstructing Missing Precordial Leads: An 11-Method Benchmark of Morphology, Clinical Utility and Latency

Mehdi Akeddar, Guillaume Chacun, Bruno Da Rocha Carvalho, Marina Zapater

REDS Institute, HEIG-VD, HES-SO, Yverdon-les-Bains, Switzerland

Abstract

Continuous ECG monitoring depends on the precordial leads, yet these electrodes are the hardest to place reliably and are prone to poor skin contact. We ask how much an arrhythmia detector loses when V2, V4 and V5 are missing, and whether reconstructing them restores that information. On 8,017 test recordings from six PhysioNet cohorts, we mask leads V2, V4 and V5 and reconstruct them from the nine remaining leads using eleven methods across three families: encoder–decoder models (1D U-Net, MCMA UNet3+ variants), classical baselines (ridge, gradient boosting, PCA/SVD, interpolation, k-NN), and generative models (GAN, Mamba diffusion). Each method is evaluated on three axes: waveform morphology (RMSE, Pearson CC), clinical utility through a 27-class SE-ResNet detector (PhysioNet/CinC Challenge score, per-class F1), and inference latency and memory. Removing the three leads lowers the Challenge score from 0.63 to 0.54; reconstruction recovers it to 0.61. The aggregate understates the effect: on the four classes that depend on V2/V4/V5, losing the leads reduces PVC F1 sixteen-fold, and reconstruction restores it. The 1D U-Net gives the best morphology (RMSE 0.239) at 2.7 ms per record; models trained to reconstruct outperform those trained to generate.

1. Introduction

Continuous and ambulatory ECG monitoring is increasingly used to detect transient arrhythmias, but it depends on reliable signal from all twelve leads. The six precordial (chest) electrodes require precise anatomical placement and are a frequent source of positional error [1]; misplacement can mimic or mask anterior ischaemia and infarction, altering interpretation in a substantial fraction of affected records [2]. Chest electrodes are also susceptible to contact loss and motion artefact during prolonged recording. When a precordial lead degrades or is not acquired, the leads that reflect anterior and lateral myocardial activity are lost, and automated interpretation degrades.

One response is to acquire fewer electrodes at the edge and reconstruct the missing leads when needed. Our target deployment acquires nine leads from seven electrodes and reconstructs V2, V4 and V5 on the same 10-second, 500 Hz window when an anomaly gate triggers, before passing the twelve-lead record to a 27-class arrhythmia detector and then a clinician. The question is whether reconstruction restores the information an automated detector needs, not how closely the waveform is reproduced.

Lead reconstruction is an established problem, with dozens of published algorithms from linear transforms to deep networks [3]. Prior work recovers a full twelve-lead ECG from one or two leads [4] or from a reduced five-electrode set [5], and recent U-Net models quantify how much reconstructable information each lead carries [6]. These studies report waveform fidelity, which does not measure whether diagnostic content is preserved, and they use different datasets, splits and metrics, so results are not comparable.

We address this with a single controlled benchmark. On one split of a six-cohort, arrhythmia-rich PhysioNet dataset, we fix the reconstruction task (V2, V4, V5 from the nine remaining leads) and evaluate eleven methods from three families on three axes: waveform morphology, downstream clinical utility through a fixed arrhythmia detector, and inference latency and memory. The nearest prior benchmark compares four reconstruction methods on a single cohort using morphology and downstream classification [7]; we extend this to eleven methods across three families, add deployment cost as a third axis, and use a larger, more arrhythmia-diverse dataset. This quantifies the detection cost of losing V2, V4 and V5 and identifies which methods recover it, at what computational cost.

2. Methods

2.1. Data and preprocessing

We use the PhysioNet/CinC Challenge 2021 multi-source corpus [8, 9], drawing on CPSC and CPSC-Extra [10], PTB-XL [11], the Chapman–Shaoxing and Ningbo databases [12], and the Georgia 12-lead set. The St. Petersburg INCART database is excluded because its label schema is incompatible with the others. All

records are resampled to 500 Hz and cropped to 10-second windows (5,000 samples), band-pass filtered (0.5–40 Hz, 4th-order Butterworth) and per-lead z-score normalised. Records are split 8:1:1 by fold into train (folds 0–7), validation (fold 8) and test (fold 9); all results below are reported on the $n = 8{,}017$ test recordings. Labels follow the 27 scored SNOMED-CT classes of the Challenge; three pairs of diagnostically equivalent classes (complete RBBB/RBBB, PAC/SVPB, PVC/VPB) are merged by logical OR before scoring, giving 24 scored classes.

2.2. Reconstruction task

The task is fixed throughout: reconstruct leads V2, V4 and V5 from the nine remaining leads (I, II, III, aVR, aVL, aVF, V1, V3, V6) on the same 10-second window. The three target leads are zeroed in the input; the deep models output all twelve leads with the reconstruction loss computed only on the three masked targets, and the classical methods predict the three targets directly. Every method is trained and evaluated on the same split with this identical masking, so differences reflect the method, not the protocol.

2.3. Reconstruction methods

We benchmark eleven methods in three families (Table 1). The *encoder–decoder* family contains a vanilla 1D U-Net (plain skip concatenation, channels 32–256, kernel 7, MSE loss) and three variants of a multi-channel masked autoencoder (MCMA, a 1D UNet 3+ [13, 14]) that differ only in their training objective: a Soft-DTW [15] plus spectral loss, a QRS-window plus Pearson-correlation loss, and a combination adding Soft-DTW at low weight. The *classical* family comprises per-timepoint ridge regression ($\alpha = 1$), histogram gradient boosting (depth 6, 100 iterations, learning rate 0.1), an 8-component PCA/SVD lead-subspace projection with ridge-mapped scores, training-free anatomical lead interpolation, and k-NN gallery retrieval ($k = 5$, Euclidean, 5,000-record gallery, unweighted neighbour mean). The *generative* family contains a GRU encoder–decoder GAN (bidirectional encoder, hidden 128, hinge discriminator loss, generator adversarial $+ 0.01\, L_1$, seed 23) and a Mamba state-space diffusion model (SSSD [16, 17], $d_{\text{model}} = 192$, 8 layers, 200-step linear schedule), trained from scratch on this task (early stop at epoch 52) rather than fine-tuned from a published checkpoint. Encoder–decoder and GAN models are trained with learning rate 10^{-3} and the diffusion model with 10^{-4}, for up to 100 epochs with early stopping (seed 42 unless stated); classical methods are fit on the same training partition.

Table 1. The eleven reconstruction methods.

Method	Family	Training signal / detail
1D U-Net	Enc–dec	MSE, plain skip
MCMA SoftDTW	Enc–dec	Soft-DTW + spectral
MCMA QRS+P	Enc–dec	QRS-window + Pearson
MCMA QRS+P+S	Enc–dec	+ Soft-DTW (γ 0.1)
Ridge	Classical	per-timepoint linear
Grad. boosting	Classical	histogram GB
PCA/SVD	Classical	lead-subspace proj.
Lead interp.	Classical	anatomical, no training
k-NN	Classical	gallery retrieval
GAN	Generative	adversarial
Mamba diff.	Generative	SSSD, 200-step

2.4. Downstream arrhythmia detector

Clinical utility is measured with a fixed 27-class detector: a 1D SE-ResNet 18 [18], the SE-ResNet entry in the CinC Challenge 2021. It emits 27 logits, and its 27 decision thresholds were tuned on its own validation fold to maximise the Challenge metric and then frozen; no threshold is retuned per reconstruction method. The same frozen detector scores every reconstruction, so any change in detection is attributable to the reconstructed leads alone. We evaluate it in three conditions: on the clean twelve-lead record, on the record with V2/V4/V5 zeroed (*removed*), and on the record with those leads *reconstructed*.

2.5. Evaluation and statistics

Methods are compared on three axes. *Morphology*: masked RMSE, per-lead RMSE and Pearson correlation on the $n = 8{,}017$ test records (8,014–8,017 per lead; a few records lack one target lead). *Clinical utility*: the official CinC Challenge score (a weighted-confusion-matrix metric), AUROC and per-class F1 from the detector. *Cost*: inference latency per 10-second record (batch size 1, 3 warm-up and 10 timed repeats) and peak GPU memory; deep and generative models are timed on an RTX 3090 (24 GB) and classical methods on CPU, so the two are not directly comparable. Confidence intervals are 95% bootstrap (2,000 resamples over 251 batches of 32). Significance uses the paired Wilcoxon signed-rank test against the MCMA Soft-DTW baseline with Holm–Bonferroni correction. Detector scores are means over per-batch scores at batch size 32; memory limits forced batch size 4 for the Mamba model, so its batch-mean scores are not comparable and are omitted from the figures. Where the generative family is compared against the reconstruction-trained methods we instead use record-level Challenge scores, computed once over the concatenated detector predictions of all $n = 8{,}017$ test records, which removes the

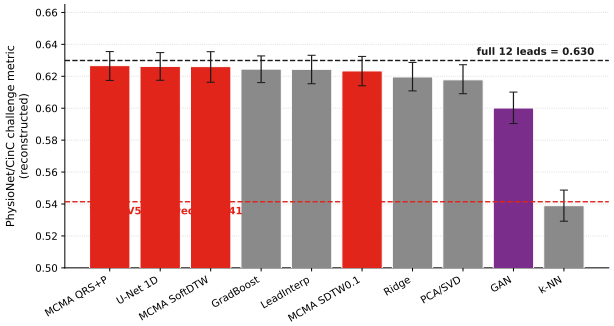

Figure 1. Challenge score after reconstruction (95% bootstrap CI). Dashed lines: clean twelve-lead and leads-removed baselines, averaged over methods. Mamba diffusion is omitted (incomparable batch size).

batching dependence.

3. Results

3.1. Reconstruction restores detection

Removing V2, V4 and V5 lowers the aggregate Challenge score from 0.630 on the clean twelve-lead record to 0.541, a loss of roughly nine points. The mean score across the ten comparable methods returns to 0.613, i.e. 7.1 of those points. Detector AUROC, averaged over the ten methods for which it is defined, follows the same pattern ($0.965 \rightarrow 0.948 \rightarrow 0.961$). Figure 1 shows per-method scores: nine of ten methods fall within 0.60–0.63, and the top eight are statistically indistinguishable from the MCMA Soft-DTW baseline (paired Wilcoxon, Holm-corrected). Only the GAN (0.600) and k-NN are significantly worse ($p < 0.005$); k-NN does not recover, remaining at the removed baseline (0.539 vs. 0.541).

3.2. Where the missing leads matter

The aggregate score is insensitive to V2/V4/V5 because most of the 27 classes do not depend on them. Table 2 isolates the four classes whose definition does. Removing the leads reduces PVC F1 sixteen-fold (0.160 to 0.010) and Q-wave-abnormality F1 nearly threefold (0.261 to 0.091), with non-overlapping 95% CIs (clean [0.195, 0.327] vs. removed [0.044, 0.144]). The 1D U-Net and the QRS+Pearson MCMA variant restore PVC F1 to the clean twelve-lead level (0.168 and 0.167), whereas lead interpolation reaches 0.114 and the GAN 0.030, 18% of its own clean 0.169. The U-Net and interpolation CIs overlap, so the ranking within the reconstruction methods is not resolved at 193 PVC-positive records; the gap to the GAN is. Methods with equal aggregate scores therefore differ on the classes that depend on the leads. Absolute F1 is low for PVC because the detector reaches only 0.09 sensitivity on clean twelve-lead data; the detector, not the reconstruction, is the ceiling.

3.3. The accuracy–speed trade-off

The 1D U-Net gives the best morphology (RMSE 0.239) at 2.7 ms per record and 152 MB, faster and smaller than the three MCMA variants (RMSE ≈0.245, 6.3–6.4 ms, 246 MB). Classical methods are cheaper but less accurate: interpolation and ridge run in 0.2–0.5 ms on CPU at RMSE 0.35–0.37. The generative models have the highest error (GAN RMSE 0.806, Mamba diffusion 0.912) after k-NN (0.998); the Mamba model is also the slowest at 619 ms. All methods run within the 10-second real-time budget.

4. Discussion

When a GPU is available and the precordial-dependent classes matter for the indication, the 1D U-Net is preferred: it gives the best morphology, the fastest GPU inference and the smallest footprint, and it restores PVC F1 to the clean twelve-lead level. MCMA matches it on the clinical classes at higher cost. On CPU-only edge hardware, interpolation and ridge recover most of the aggregate score but only 71% of clean PVC F1, so they are suitable only where the precordial-dependent classes are not the target.

On this task, models trained to reconstruct outperform models trained to generate. Both generative models are the least accurate on morphology after k-NN, and both are last on record-level Challenge score: Mamba diffusion 0.571 and the GAN 0.601, against 0.619–0.627 for every reconstruction-trained method (leads-removed baseline 0.542 at record level, 0.541 as batch mean; clean 0.630). The GAN also retains only 18% of its clean PVC F1. Reconstruction is a conditional, deterministic mapping from nine leads to three, and optimising that mapping directly appears more effective than sampling from a learned distribution. The claim rests on two generative models, one of which (Mamba) was trained from scratch here and may

Table 2. Per-class F1 (test n = 8,017) on the four precordial-dependent classes: full twelve-lead, leads removed, and reconstructed by four methods (Q+P: MCMA QRS+Pearson). Reconstructed PVC F1 95% CIs: U-Net [0.100, 0.234], Q+P [0.098, 0.235], interpolation [0.058, 0.177], GAN [0.000, 0.067].

Class (n)	Full	Rem.	U-Net	Q+P	Interp	GAN
PVC (193)	0.160	0.010	0.168	0.167	0.114	0.030
Q-wave (214)	0.261	0.091	0.226	0.234	0.219	0.161
T-abn. (1165)	0.604	0.463	0.591	0.596	0.581	0.558
T-inv. (394)	0.276	0.209	0.240	0.252	0.200	0.268

be under-trained, so it is a result for this task rather than a general ordering of the families.

Several limitations apply. The masking is fixed to V2/V4/V5; other lead combinations were not tested, so this study bounds the cost of losing these three leads and not the general information content of each lead. The detector is the ceiling on the precordial-dependent classes, so the reconstruction differences we measure are lower bounds on what a stronger detector might reveal. Labels are record-level, so we do not assess beat-level PVC morphology or ST-segment accuracy. Finally, we use a single split; although it spans six cohorts across three continents, there is no external validation cohort. Transformer baselines, per-patient adaptation, an ST-aware objective and external validation are next steps.

5. Conclusion

Missing precordial leads cost roughly nine Challenge-score points, and most reconstruction methods recover most of them. The aggregate hides the effect, which is concentrated in the four classes that depend on V2, V4 and V5, where PVC F1 falls sixteen-fold and is restored by reconstruction. Across morphology, clinical utility and cost, a 1D U-Net is the best all-round method, and methods trained to reconstruct outperform methods trained to generate. For a reduced-electrode monitor, reconstructing the missing precordial leads on an anomaly trigger preserves automated diagnostic accuracy at low cost.

Code and data availability

The data are the public PhysioNet/CinC 2021 WFDB records. Code for all eleven methods and the evaluation is available from the authors.

References

[1] Medani S, Hensey M, Caples N, Owens P. Accuracy in precordial ECG lead placement: improving performance through a peer-led educational intervention. J Electrocardiol 2018;51(1):50–54.

[2] Rehman M, Rehman N. Precordial ECG lead mispositioning: its incidence and estimated cost to healthcare. Cureus 2020;12(7):e9040.

[3] Obianom E, Ng G, Li X. Reconstruction of 12-lead ECG: a review of algorithms. Front Physiol 2025;16:1532284.

[4] Presacan O, Dorobantiu A, Isaksen J, Willi T, Graff C, Riegler M, Sridhar A, Kanters J, Thambawita V. Evaluating the feasibility of 12-lead electrocardiogram reconstruction from limited leads using deep learning. Commun Med 2025;5(1):139.

[5] Maheshwari S, Acharyya A, Rajalakshmi P, Puddu P, Schiariti M. Accurate and reliable 3-lead to 12-lead ECG reconstruction methodology for remote health monitoring applications. IRBM 2014;35(6):341–350.

[6] Gradowski T, Buchner T. Deep learning model for ECG reconstruction reveals the information content of ECG leads. Bio Algorithms Med Syst 2025;21(1):13–23.

[7] Zhang X, Cai H, Duan H, Lu X. Systematic benchmark of reduced-lead configurations for 12-lead ECG reconstruction. Front Cardiovasc Med 2026;13:1856211.

[8] Reyna M, Sadr N, Alday EP, Gu A, Shah A, Robichaux C, Rad AB, Elola A, Seyedi S, Ansari S, Ghanbari H, Li Q, Sharma A, Clifford G. Will two do? Varying dimensions in electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021. In Proc. Comput. Cardiol. (CinC), volume 48. 2021; 1–4.

[9] Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov P, Mark R, Mietus J, Moody G, Peng CK, Stanley H. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[10] 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. J Med Imaging Health Inform 2018;8(7):1368–1373.

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

[12] Zheng J, Zhang J, Danioko S, Yao H, Guo H, Rakovski C. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. Sci Data 2020;7(1):48.

[13] Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, Han X, Chen YW, Wu J. UNet 3+: a full-scale connected UNet for medical image segmentation. In Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP). 2020; 1055–1059.

[14] Chen J, Wu W, Liu T, Hong S. Multi-channel masked autoencoder and comprehensive evaluations for reconstructing 12-lead ECG from arbitrary single-lead ECG. npj Cardiovasc Health 2024;1(1):31.

[15] Cuturi M, Blondel M. Soft-DTW: a differentiable loss function for time-series. In Proc. 34th Int. Conf. Mach. Learn. (ICML), volume 70. 2017; 894–903.

[16] Alcaraz JL, Strodthoff N. Diffusion-based time series imputation and forecasting with structured state space models. Trans Mach Learn Res 2023;ArXiv:2208.09399.

[17] Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces, 2023. ArXiv:2312.00752.

[18] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). 2018; 7132–7141.

Address for correspondence:

Mehdi Akeddar
REDS Institute, HEIG-VD, HES-SO
mehdi.akeddar@heig-vd.ch