

Automated identification of post-infarction ventricular tachycardia ablation targets with spatio-temporal graph convolutional networks on 3-D electroanatomical manifolds

Amine Bougarne

`amine.bougarne@um5r.ac.ma`

Mohammed V University

Article

Keywords: Ventricular tachycardia, Abnormal ventricular potentials, Intracardiac electrogram, Electroanatomical mapping, Graph convolutional network

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11035410/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Abnormal ventricular potentials (AVPs) recorded during substrate mapping guide catheter ablation in post-infarction ventricular tachycardia, and annotating them by hand is slow and operator dependent. We asked whether a spatio-temporal graph convolutional network (ST-GCN) defined on the reconstructed left-ventricular surface detects AVPs better than established alternatives. Using the ARGO dataset (nine patients, 1962 annotated bipolar electrograms, three independent electrophysiologists plus a consensus), we built one graph per patient with nodes at mapping sites and edges weighted by geodesic distance along the triangulated endocardial mesh. The network combined a dilated temporal encoder, Chebyshev graph convolutions, multi-head graph attention, and an auxiliary head regressing the distance to delivered radiofrequency lesions. Six comparators were trained under identical leave-one-patient-out cross-validation, among them the conventional bipolar-voltage rule (below 1.5 mV) and gradient boosting on 38 expert descriptors. Decision thresholds were estimated on pooled inner-validation patients only. Gradient boosting reached the highest AUC (0.973 $\pm$ 0.033), ahead of the voltage rule (0.952 $\pm$ 0.053) and the ST-GCN (0.942 $\pm$ 0.052); no pairwise comparison favoured the ST-GCN (Holm-adjusted $p > 0.06$). A 13-variant ablation showed that removing spatial message passing (0.949) or substituting Euclidean for geodesic edge weights (0.954) left performance unchanged, whereas removing the wideband amplitude channel lowered it to 0.921. Stratifying by voltage explains the pattern: below 0.5 mV every annotated electrogram was an AVP, above 1.5 mV 95.6% were physiological. In ARGO, expert AVP annotation is largely determined by bipolar amplitude, which limits how well the dataset can separate competing modelling strategies.

1 Introduction

Scar-related ventricular tachycardia (VT) after myocardial infarction arises from reentry through surviving myocardial bundles embedded in fibrotic tissue [1, 2]. These bundles conduct slowly and asynchronously with the surrounding myocardium, and the intracardiac bipolar electrograms (EGMs) recorded above them show fractionation, low amplitude and late components [3]. Such signals, variously called abnormal ventricular potentials (AVPs) or local abnormal ventricular activities (LAVAs), are the standard target of substrate-guided ablation [30]. Jais and colleagues showed in 2012 that eliminating LAVAs is a reproducible procedural endpoint [4], and long-term follow-up has since associated complete elimination with better arrhythmia-free survival [5, 6].

Identifying these potentials remains a visual task. During electroanatomical mapping the operator inspects each acquired EGM and decides whether it carries an abnormal component, often under time pressure and with hundreds of sites to review. Agreement between experts is imperfect, and the decision depends on catheter orientation, wavefront direction and filter settings [7, 8]. Automating the decision would shorten procedures and make the annotation of research datasets reproducible, which is why several groups have applied signal processing and machine learning to the problem. Baldazzi et al. characterised the spectral signatures separating AVPs from physiological potentials in 450 EGMs from seven patients [9], then trained supervised classifiers on 1584 EGMs using time, time-scale, frequency

and spatial descriptors [10, 32]. Related work has explored peak-frequency metrics for near-field detection during VT circuits [11, 12].

Two things have limited progress. First, until recently no annotated benchmark was openly available, so results could not be compared across groups. The ARGO dataset, released on PhysioNet in 2026, closes that gap: 1962 bipolar EGMs from nine post-infarction VT patients, each annotated independently by three electrophysiologists with a final consensus, together with unipolar EGMs, 12-lead surface ECG, and the CARTO 3 voltage and local activation time maps [13, 14]. Second, published models treat each EGM in isolation. The arrhythmogenic substrate is a spatially organised structure, and neighbouring mapping sites are not independent. Graph neural networks are the natural formalism for data sampled irregularly on a surface, and they have been applied to cardiac electrophysiology for activation prediction and personalised modelling [15, 16], but not, to our knowledge, to AVP detection on clinical electroanatomical maps.

We therefore built a spatio-temporal graph convolutional network in which each mapping site is a node of a patient-specific graph and edges are weighted by geodesic distance along the reconstructed endocardial surface. The model was designed to exploit three sources of information that single-signal classifiers discard: the waveform itself, the substrate attributes attached to the mapping site, and the labels of neighbouring sites. We then did something that is often skipped. Rather than comparing the network only against other neural architectures, we benchmarked it against the bipolar-voltage rule that electrophysiologists already use, against gradient boosting on expert descriptors, and against three deep temporal baselines, all under the same leave-one-patient-out protocol. We also quantified the agreement between the three human annotators, which sets the ceiling any automated method can reasonably approach.

The result was not the one we expected, and we report it as we found it. The graph did not help. Neither did the geodesic metric, the attention mechanism, or the auxiliary spatial task. What the ablation isolated instead was a single dependency on signal amplitude, and stratifying the cohort by bipolar voltage showed why: in ARGO the expert label is nearly a deterministic function of amplitude. This observation concerns the reference standard rather than any particular architecture, and it has consequences for anyone planning to benchmark on this dataset.

2 Materials and methods

2.1 Dataset and reference standard

ARGO version 1.0.0 was downloaded from PhysioNet under the CC BY-NC-SA 4.0 licence [13, 14, 31]. It contains 1962 recordings from nine patients with post-ischaemic VT mapped with CARTO 3 (Biosense Webster) [29]. Each recording is a 2.5 s, 15-channel segment sampled at 1 kHz: one bipolar EGM (band-limited 16–500 Hz by the mapping system), the two unipolar EGMs of the same electrode pair (2-240

Hz), and the 12-lead surface ECG (0.5–120 Hz). The gain is 333.3333 a.u./mV. Only the last beat of each segment is documented as reliable, and all analyses respect that constraint.

Each bipolar EGM carries four annotation files: three independent annotations by expert electrophysiologists and a consensus. Every annotation assigns one of three classes (Physiological, AVP, Unknown) and, for AVPs, the onset and end of the abnormal component. Per patient, the dataset also provides the triangulated left-ventricular mesh (vertex coordinates and connectivity), the bipolar voltage and local activation time attached to each mesh vertex, the 3-D coordinates of every mapping site, and the coordinates of the radiofrequency lesions actually delivered during the procedure. The original acquisition was approved by the local ethics committee (Prot. n. 351/2021/CE) as reported by the dataset authors [13]; the present work is a secondary analysis of fully anonymised public data and required no additional approval.

The class annotations are stored as WFDB binary annotation files using labels outside the standard WFDB code table, and the reader available in wfdb-python returned empty results on every file. We therefore implemented the WFDB binary annotation format directly and recovered the class letter from the auxiliary string attached to each marker. Five competing decoding strategies were evaluated and arbitrated against the class counts published with the dataset. The retained strategy reproduced them exactly (776 Physiological, 940 AVP, 246 Unknown) and recovered 940 delineated intervals, one per AVP. This verification is reported because a silent decoding failure would invalidate every downstream number.

2.2 Signal preprocessing and beat selection

The bipolar EGM was processed in two bands with zero-phase Butterworth filters (order 4, filtfilt). A wideband version (1-450 Hz) preserves peak-to-peak amplitude, which is the quantity underlying the clinical voltage criterion; a high-frequency version (30–400 Hz) emphasises the fractionated components associated with AVPs. Both were preceded by a 50 Hz notch ($Q = 30$). The two unipolar channels were filtered in the wideband setting and the surface ECG was left in its native band. Restricting the network to the 30–400 Hz band alone, as an earlier version of this pipeline did, removes 31% of the median peak-to-peak amplitude and measurably degrades performance (Section 3.4).

The last ventricular depolarisation was located on the quadratic norm of the 12 surface leads using an energy detector of the Pan-Tompkins family, then refined to the local energy maximum. A fixed 500 ms window spanning 100 ms before to 400 ms after that peak was extracted. The window was defined a priori from physiological reasoning and never optimised against the labels, which rules out the most common form of leakage in this setting. As a quality check we verified afterwards that 96.8% of the 940 expert-delineated AVP intervals fall entirely inside it. Median AVP duration was 61 ms (IQR 38–83).

Channel-wise means and standard deviations were computed on the training patients of each fold and applied unchanged to the test patient. Normalising per recording would have destroyed the relative amplitude information that distinguishes scar from healthy myocardium, so we deliberately avoided it.

2.3 Electroanatomical graph construction

For each patient we built an undirected graph whose nodes are the mapping sites. Mapping sites in ARGO coincide with mesh vertices (median projection error 0.00 mm), so each node inherits the bipolar voltage and local activation time of its vertex, averaged over a 3 mm geodesic neighbourhood.

Edge weights follow a Gaussian kernel of the geodesic distance measured along the triangulated endocardium rather than the straight-line distance in R3. The mesh edge graph was weighted by Euclidean edge length and all pairwise geodesic distances between mapping sites were obtained by Dijkstra shortest paths. Two sites on opposite walls can be close in R3 while being far apart along the myocardium, and the geodesic metric was intended to prevent the graph from connecting them. Each node was linked to its $k = 8$ nearest neighbours with weight $\exp(-d^2/\sigma^2)$, $\sigma = 12$ mm, edges above 40 mm removed, self-loops added, and the result symmetrised. When the k -nearest-neighbour rule left the graph disconnected, components were joined by their shortest geodesic edge so that messages can travel across the whole chamber.

The adjacency matrix was normalised symmetrically [27] and the rescaled Laplacian used as the basis for Chebyshev polynomial convolutions [26]. Graphs are small (46 to 839 nodes), so all operators were implemented densely in PyTorch. The dense formulation is exact, faster than sparse routing at this scale, and removes the dependency on compiled scatter extensions, which matters for reproducibility on hosted notebooks.

2.4 Expert descriptors

Thirty-eight descriptors were extracted per node for the gradient-boosting comparator and for the univariate analysis. They cover amplitude (peak-to-peak, RMS, absolute maximum, measured on the wideband channel), fractionation (number of deflections, zero crossings, waveform length, Katz fractal dimension, maximum slope), late activity (Hilbert envelope energy after the QRS, delay of the last energy peak, EGM duration, activation delay), spectral content (relative band powers 30–100, 100–250 and 250–400 Hz, dominant frequency, spectral centroid, spectral entropy, 95% spectral edge), multiscale energies from a five-level dyadic decomposition, unipolar descriptors (maximum negative slope and its timing, peak-to-peak), and the two substrate attributes taken from the CARTO map.

2.5 Proposed model and comparators

The proposed network has three stages. A temporal encoder applies three dilated convolutional blocks (dilations 1, 2, 4; kernel 7; 32, 64 and 96 channels) with batch normalisation, GELU activation and squeeze-and-excitation recalibration, followed by an attention-weighted temporal pooling that produces one 96-dimensional embedding per node. The substrate attributes are projected to 16 dimensions and concatenated. A spatial stage then applies two residual Chebyshev convolutions of order 3 on the rescaled Laplacian, followed by four-head graph attention restricted to the support of the adjacency matrix [28]. Two heads read the resulting node representation: a classification head predicting the probability of AVP, and a regression head predicting the distance from the node to the nearest delivered

radiofrequency lesion, clipped at 40 mm. The loss is α times the binary cross-entropy plus $(1 - \alpha)$ times a Huber loss on the normalised distance, with $\alpha = 0.7$. The classification term is weighted by the training-fold class ratio and smoothed by 0.03. The model has 306,443 trainable parameters.

Six comparators were trained under the same protocol. The bipolar-voltage rule scores each site by a monotone decreasing function of its CARTO voltage and thresholds it at 1.5 mV, following the criterion in routine use since Marchlinski et al. [17]. Gradient boosting (XGBoost, 400 trees, depth 4, learning rate 0.05) was fitted on the 38 descriptors [18]. Three deep temporal baselines share the input tensor but discard the graph: the temporal encoder alone (1-D CNN), a two-layer bidirectional LSTM, and a two-layer Transformer encoder. Finally, a variant of the proposed network deprived of the substrate attributes tests whether waveform and spatial context alone can match the voltage map.

2.6 Validation protocol, metrics and statistics

Evaluation used leave-one-patient-out cross-validation. For each of the nine outer folds, one patient was held out entirely, one of the remaining eight was set aside as an inner validation patient for early stopping (rotated deterministically with the seed), and the other seven were used for training. Deep models were trained for up to 150 epochs with AdamW (learning rate $1.2e-3$, weight decay $1e-4$, cosine annealing with warm restarts, gradient clipping at 2.0) and stopped on inner-validation PR-AUC with a patience of 30. Every experiment was repeated with three seeds (13, 42, 2024).

Three design choices deserve emphasis because they are the usual sources of optimistic bias. Normalisation statistics, early stopping and threshold selection all use training-fold information only. The decision threshold was not taken from a single validation patient, which we found unstable, but from the pooled predictions of all inner-validation patients across seeds, still strictly outside the test patient. The 246 Unknown electrograms participate in message passing on the graph, so the model sees the full spatial context, but they are masked in the loss and excluded from every metric; this is a specific advantage of the transductive graph formulation over per-signal classifiers.

The learning unit is the node; the unit of statistical inference is the patient. Discrimination was summarised by AUC-ROC and PR-AUC, with 95% confidence intervals from a cluster bootstrap resampling patients rather than nodes (2000 replicates). Threshold-dependent metrics (F1, sensitivity, specificity, Matthews correlation, balanced accuracy) and the Brier score are reported at the pooled threshold. Models were compared by one-sided Wilcoxon signed-rank tests on the nine per-patient AUCs with Holm correction. With nine pairs the smallest attainable one-sided p value is 0.00195, and no difference can be declared significant below that floor; we state this explicitly because it constrains what this cohort size can demonstrate. Agreement between annotators was measured by Fleiss' kappa [19] and by the proportion of unanimous triplets.

Clinical localisation was quantified by delta-d3D, the Euclidean distance from a predicted site to the nearest delivered lesion, evaluated for the top-1, top-5 and top-10 ranked sites and against two references: an oracle taking the centroid of the truly annotated AVPs, and uniform random selection.

Calibration was assessed by the Brier score, a 10-bin expected calibration error, a reliability diagram and decision-curve analysis [20]. Temporal attribution used integrated gradients [21]; for each delineated AVP we computed the fraction of saliency mass falling inside the expert interval and compared it with the fraction expected from the interval length alone.

Analyses ran on Kaggle Notebooks (Python 3.12, PyTorch 2.10, NVIDIA T4) in 81 min. Reporting follows TRIPOD + AI [22] where applicable.

3 Results

3.1 Cohort, reference standard and expert agreement

The analysis pipeline is summarised in Fig. 1. All 1962 recordings were read without failure and the per-patient counts matched the dataset descriptor exactly (Table 1, Fig. 2a). The cohort was predominantly male (7/9), median age 68 years (range 50–83), with a median left-ventricular ejection fraction of 30% (range 21–42). Mapping density varied by more than an order of magnitude, from 46 sites in patient 5 to 839 in patient 6.

Table 1

Cohort characteristics and composition of the annotated electrogram set. Class counts are from the expert consensus. AVP prevalence is computed on labelled electrograms only, excluding the Unknown class. LVEF left-ventricular ejection fraction, EGM electrogram, AVP abnormal ventricular potential.

Patient	Sex	Age (y)	LVEF (%)	EGM (n)	AVP (n)	Physiol. (n)	Unknown (n)	AVP (%)
1	M	68	32	157	85	54	18	61.2
2	F	68	30	104	52	36	16	59.1
3	M	83	42	90	67	17	6	79.8
4	M	50	30	471	74	386	11	16.1
5	M	78	27	46	21	12	13	63.6
6	M	61	31	839	508	168	163	75.1
7	M	60	21	76	34	35	7	49.3
8	F	57	28	129	66	54	9	55.0
9	M	69	24	50	33	14	3	70.2
Cohort (n = 9)	7 M / 2 F	68 (50–83)	30 (21–42)	1962	940	776	246	54.8

AVP prevalence was 54.8% overall but ranged from 16.1% (patient 4) to 79.8% (patient 3). That spread is the reason a random-split evaluation would be misleading here, and it justifies both the leave-one-patient-out design and the systematic reporting of PR-AUC alongside AUC-ROC.

Agreement between the three annotators was moderate (Fig. 2b; Fleiss' kappa 0.641 for AVP versus non-AVP, 0.607 across the three classes). Individual experts agreed with the consensus on 88.2% of electrograms (mean kappa 0.764). Agreement was not uniform across annotators: the second and third agreed with each other at kappa 0.768, whereas the first agreed with them at 0.580 and 0.579 (Table 2). Whatever an automated method achieves has to be read against these figures rather than against perfect agreement.

Table 2

Agreement between the three independent expert annotators and with the consensus, for the three-class task and for the binary AVP versus non-AVP task (n = 1962 electrograms). Cohen's kappa for pairwise comparisons, Fleiss' kappa for the three raters jointly.

Comparison	Three classes	AVP vs non-AVP	Raw agreement
	kappa	kappa	(%) binary
Annotator 1 vs 2	0.543	0.580	78.8
Annotator 1 vs 3	0.561	0.579	79.1
Annotator 2 vs 3	0.722	0.768	88.3
Annotator 1 vs consensus	0.584	0.615	80.7
Annotator 2 vs consensus	0.762	0.847	92.4
Annotator 3 vs consensus	0.778	0.830	91.5
Fleiss' kappa (3 raters)	0.607	0.641	-
Model vs consensus	-	0.566	79.3

3.2 Graph geometry and substrate characterisation

Graphs contained 9809 undirected edges in total, with a mean node degree of 9.97 and one connected component per patient after bridging (Supplementary Table S1). The geodesic-to-Euclidean distance ratio over retained edges had a median of 1.005 and a 95th percentile of 1.219, with only 5.9% of edges above 1.20. On this dataset the two metrics are therefore almost interchangeable at the scale of a k-nearest-neighbour neighbourhood. Mapping points sit exactly on mesh vertices and their immediate neighbours are adjacent on the surface, so surface curvature has little room to express itself over such short distances. We report this because it directly contradicts the premise that motivated the geodesic construction, and the ablation confirms it quantitatively.

The spatial ground truth behaved as physiology predicts. Sites annotated AVP lay closer to delivered lesions than physiological sites (median 3.9 vs 6.5 mm; Mann-Whitney $p = 4.5e-23$, Cliff's delta = 0.275) and had much lower local voltage (0.57 vs 2.10 mV, $p = 6.0e-245$).

Univariate screening ranked amplitude descriptors first. Bipolar RMS reached an AUC of 0.973, peak-to-peak amplitude 0.970, and the CARTO voltage itself 0.968; all 38 descriptors were significant after Benjamini-Hochberg correction (Supplementary Table S2). A single variable therefore separates the two classes about as well as any multivariate model we trained, which is the first sign that the task is easier than its clinical framing suggests.

3.3 Detection performance

Table 3 and Fig. 3 give the leave-one-patient-out results. Gradient boosting on expert descriptors was the best model (AUC 0.973 +/- 0.033, PR-AUC 0.984, MCC 0.827). The clinical voltage rule came second (0.952 +/- 0.053). The proposed ST-GCN reached 0.942 +/- 0.052 with a 95% cluster-bootstrap interval of 0.888–0.977. The three deep temporal baselines were lower: 1-D CNN 0.915, Transformer 0.901, BiLSTM 0.856. Removing the voltage map from the ST-GCN dropped it to 0.883.

Table 3

Leave-one-patient-out performance. Values are mean +/- standard deviation across the nine held-out patients, averaged over three seeds. Confidence intervals come from a cluster bootstrap resampling patients (2000 replicates). Threshold-dependent metrics use the pooled inner-validation threshold. p values are one-sided Wilcoxon signed-rank tests against the proposed model with Holm correction; the smallest attainable value at n = 9 is 0.00195.

Model	AUC-ROC	95% CI	PR-AUC	F1	Se	Sp	MCC	p
Bipolar voltage < 1.5 mV	0.952 +/- 0.053	0.941–0.987	0.960 +/- 0.046	0.892	0.946	0.779	0.744	1.000
XGBoost (expert features)	0.973 +/- 0.033	0.949–0.993	0.984 +/- 0.016	0.923	0.887	0.938	0.827	1.000
1-D CNN (temporal only)	0.915 +/- 0.064	0.774–0.973	0.915 +/- 0.063	0.823	0.856	0.730	0.618	0.193
BiLSTM	0.856 +/- 0.128	0.675–0.971	0.891 +/- 0.056	0.805	0.832	0.649	0.511	0.059
Transformer	0.901 +/- 0.064	0.796–0.974	0.928 +/- 0.032	0.809	0.941	0.499	0.485	0.059
ST-GCN without voltage map	0.883 +/- 0.106	0.734–0.963	0.888 +/- 0.113	0.760	0.814	0.814	0.630	0.059
ST-GCN (proposed)	0.942 +/- 0.052	0.888–0.977	0.943 +/- 0.060	0.849	0.929	0.622	0.574	ref.

Paired comparisons did not favour the proposed model against either the voltage rule or gradient boosting (Holm-adjusted $p = 1.0000$ in both cases). Against the weaker deep baselines the differences pointed in the expected direction but did not reach significance ($p = 0.0586$ against the BiLSTM, the Transformer and the voltage-free variant; $p = 0.1934$ against the CNN). Per-patient AUCs (Supplementary Table S3) show where the models diverge: patient 6, the most densely mapped and the one with the highest AVP prevalence, was hardest for every deep model (ST-GCN 0.858, BiLSTM 0.553) while the voltage rule handled it well (0.940).

Threshold-dependent metrics expose a second problem. The ST-GCN operated at high sensitivity and poor specificity (0.929 and 0.622 $\pm$ 0.294), whereas gradient boosting was balanced (0.887 and 0.938). The large specificity variance across patients persisted even after pooling the threshold over all inner-validation patients, which suggests it reflects genuine between-patient heterogeneity in the score distribution rather than an artefact of threshold estimation.

Agreement between the model and the consensus was 79.3% (kappa 0.566), below the 88.2% and kappa 0.764 achieved by individual human experts. On this measure the network does not reach the human ceiling.

3.4 Ablation study

The ablation was designed to isolate the contribution of each component, and it isolated almost nothing (Table 4, Fig. 4). Removing spatial message passing entirely left performance unchanged (0.949 vs 0.947, $p = 0.82$). Replacing geodesic with Euclidean edge weights left it unchanged as well (0.954, $p = 0.43$), consistent with the near-unity distance ratio reported above. Removing graph attention (0.942), removing the auxiliary regression head (0.940), and varying k between 4 and 16 or σ between 6 and 24 mm all produced differences within the seed-to-seed noise.

Table 4

Ablation study. Each variant modifies one element of the proposed model, all else being equal, with a reduced budget of 90 epochs and one seed. AUC values are mean +/- standard deviation across the nine held-out patients. p values are two-sided Wilcoxon signed-rank tests against the full model.

Variant	AUC-ROC	PR-AUC	F1	delta AUC	p
ST-GCN, full model	0.947 +/- 0.054	0.957 +/- 0.053	0.871	ref.	ref.
No spatial message passing	0.949 +/- 0.051	0.960 +/- 0.044	0.869	+ 0.002	0.820
Euclidean instead of geodesic weights	0.954 +/- 0.040	0.954 +/- 0.050	0.882	+ 0.007	0.426
No substrate attributes	0.913 +/- 0.084	0.913 +/- 0.088	0.856	-0.034	0.129
No graph attention	0.942 +/- 0.049	0.953 +/- 0.040	0.889	-0.004	0.496
No auxiliary regression head	0.940 +/- 0.062	0.939 +/- 0.088	0.883	-0.007	0.301
Expert features instead of waveform	0.932 +/- 0.053	0.942 +/- 0.052	0.891	-0.015	0.313
No wideband amplitude channel	0.921 +/- 0.083	0.933 +/- 0.066	0.874	-0.026	0.148
Intracardiac channels only	0.951 +/- 0.056	0.958 +/- 0.046	0.888	+ 0.004	0.461
k = 4	0.948 +/- 0.062	0.957 +/- 0.064	0.890	+ 0.001	0.813
k = 16	0.954 +/- 0.036	0.949 +/- 0.069	0.807	+ 0.007	0.820
sigma = 6 mm	0.945 +/- 0.058	0.949 +/- 0.075	0.876	-0.002	0.820
sigma = 24 mm	0.947 +/- 0.054	0.951 +/- 0.067	0.887	-0.000	0.652

Two variants did move the needle, and both concern amplitude. Removing the substrate attributes cost 0.034 AUC (0.913), and removing the wideband channel, leaving only the 30–400 Hz band, cost 0.026 (0.921). Neither reached significance at $n = 9$, but they are the two largest effects in the table and they point at the same underlying quantity. Everything the architecture contributes beyond amplitude is, on this dataset, indistinguishable from noise.

3.5 Stratification by bipolar voltage

Stratifying nodes by bipolar voltage explains the whole pattern (Table 5, Fig. 5). Below 0.5 mV, 359 of 359 labelled electrograms were annotated AVP and not a single one was annotated physiological. Above 1.5 mV, 627 of 656 labelled electrograms (95.6%) were physiological. Only the border zone between 0.5 and 1.5 mV contains both classes in appreciable numbers (552 AVP, 149 physiological), and even there the prevalence reaches 78.7%.

Table 5 Composition, expert agreement and model performance by bipolar voltage stratum. Class counts are from the consensus. Unanimity is the proportion of electrograms on which all three annotators agree. AUC values are mean +/- standard deviation across evaluable patients; the dense-scar stratum contains no physiological electrogram and therefore admits no discrimination metric. The pre-specified hypothesis concerned the border zone.

Stratum	Nodes (n)	AVP (n)	Physiol. (n)	Prevalence (%)	Unanimity (%)	Fleiss kappa
Dense scar (< 0.5 mV)	405	359	0	100.0	75.1	0.375
Border zone (0.5-1.5 mV)	850	552	149	78.7	68.2	0.522
Normal (> 1.5 mV)	707	29	627	4.4	77.8	0.303

Discrimination within strata (AUC-ROC, mean +/- SD across evaluable patients)

Model	Border zone (n = 9)	Normal (n = 6)	Dense scar
Bipolar voltage < 1.5 mV	0.823 +/- 0.120	0.854 +/- 0.134	not evaluable
XGBoost (expert features)	0.952 +/- 0.065	0.870 +/- 0.074	not evaluable
1-D CNN	0.795 +/- 0.182	0.864 +/- 0.109	not evaluable
Transformer	0.820 +/- 0.181	0.797 +/- 0.172	not evaluable
ST-GCN without voltage map	0.787 +/- 0.163	0.831 +/- 0.161	not evaluable
ST-GCN (proposed)	0.810 +/- 0.170	0.864 +/- 0.096	not evaluable

Expert agreement was lowest in that border zone: 68.2% of triplets were unanimous, against 75.1% in dense scar and 77.8% in normal myocardium. Fleiss' kappa gives the opposite ranking (0.522 in the border zone against 0.375 and 0.303 elsewhere), but kappa depends on the marginal distribution and is not interpretable when prevalence is 100% or 4.4%. Unanimity is the appropriate statistic here, and it confirms that clinical ambiguity concentrates where voltage is intermediate.

We had pre-specified a test in that zone: the ST-GCN should outperform the voltage rule where the rule loses its discriminative power. It did not. Restricted to the border zone the ST-GCN reached 0.810 +/- 0.170 against 0.823 +/- 0.120 for the voltage rule (difference - 0.014, Wilcoxon $p = 0.68$), while gradient boosting reached 0.952 +/- 0.065. The hypothesis is rejected, and the margin by which gradient boosting wins in the hardest stratum is larger than anywhere else.

3.6 Localisation of ablation targets

Ranking sites by predicted probability and measuring the distance from the ten highest-ranked to the nearest delivered lesion (Fig. 6) gave 9.7 +/- 11.6 mm for the ST-GCN, against 11.0 +/- 4.7 mm for random selection and 8.2 +/- 5.4 mm for the oracle centroid of truly annotated AVPs (Table 6). The entire exploitable range between chance and oracle is 2.8 mm, and the model recovers about half of it. The voltage rule did better (7.7 +/- 7.3 mm), as did gradient boosting (8.5 +/- 10.5 mm).

Table 6 Localisation of ablation targets. delta-d3D is the Euclidean distance from a predicted site to the nearest radiofrequency lesion actually delivered, in millimetres, averaged over the highest-ranked sites. Hit rate is the proportion of the ten highest-ranked sites lying within the stated radius. Two references bound the achievable range: the oracle centroid of truly annotated AVPs and uniform random selection.

Model	top-1 (mm)	top-10 (mm)	Hit at 5 mm (%)	Hit at 10 mm (%)	rho	MAE (mm)
Bipolar voltage < 1.5 mV	5.5 +/- 3.2	7.7 +/- 7.3	54.4	80.0	-	-
XGBoost (expert features)	9.0 +/- 10.2	8.5 +/- 10.5	58.5	74.4	-	-
1-D CNN	9.6 +/- 11.9	10.1 +/- 10.9	51.9	73.3	0.10	9.7
Transformer	8.9 +/- 11.6	8.9 +/- 11.6	55.2	77.4	0.15	10.3
ST-GCN without voltage map	12.3 +/- 11.6	9.1 +/- 9.7	60.7	75.2	0.24	8.8
ST-GCN (proposed)	9.3 +/- 9.6	9.7 +/- 11.6	45.6	72.2	0.26	8.8
Oracle (true AVP centroid)	-	8.2 +/- 5.4	-	-	-	-
Random selection	-	11.0 +/- 4.7	-	-	-	-

The auxiliary regression head was weakly informative at best: Spearman correlation 0.26 between predicted and true distance, mean absolute error 8.8 mm on a range clipped at 40 mm. We would not act

on such a prediction.

3.7 Calibration, interpretability and sensitivity analyses

Two findings argue that the negative result is not a training failure. First, the ST-GCN was well calibrated: Brier score 0.095, expected calibration error 0.060, calibration slope 1.07. Second, integrated-gradient attribution concentrated 50.7% of the temporal saliency mass inside the expert-delineated AVP interval, against 12.7% expected from the interval length alone, a median enrichment of 3.86-fold (Wilcoxon $p = 2.0e-18$, $n = 126$ delineated AVPs across all nine held-out patients). The delineation was never supplied during training. The network therefore learns to look where the expert looks, produces trustworthy probabilities, and still fails to beat a single-variable rule, which locates the difficulty in the task rather than in the model.

A sensitivity analysis reclassifying the 246 Unknown electrograms as non-AVP lowered the ST-GCN AUC to 0.884, a drop of 0.058. Restricting the analysis to the five patients with at least 100 mapping sites left it essentially unchanged (0.944). Across the three seeds the mean AUC varied by 0.012.

4 Discussion

The central result of this study is a property of the reference standard, not of any architecture. In ARGO, whether an electrogram is annotated as an AVP is almost entirely predictable from its bipolar amplitude. Below 0.5 mV the annotation is invariably AVP; above 1.5 mV it is almost invariably physiological. A single descriptor reaches an AUC of 0.973, the clinical rule in routine use since 2000 reaches 0.952, and a 306,000-parameter spatio-temporal network with graph convolutions, attention and an auxiliary spatial task reaches 0.942. Thirteen ablation variants failed to attribute any measurable contribution to the graph, the geodesic metric, the attention mechanism or the multi-task objective.

This is coherent with how the annotations were produced. The experts annotated by visually inspecting bipolar EGMs, and amplitude is the most salient visual property of such a trace. The CARTO voltage attached to each mesh vertex is itself derived from the peak-to-peak amplitude of the same signal, so the substrate feature and the waveform are not independent sources of evidence. What we measure is therefore the extent to which expert judgement is driven by amplitude, and the answer is: to a very large extent.

The consequence for method development is direct. A model evaluated on ARGO can reach an AUC above 0.94 while learning nothing beyond amplitude, and its reported performance will look excellent. Comparisons that omit a voltage baseline cannot distinguish a model that captures fractionation and late activity from one that has rediscovered peak-to-peak amplitude. We would argue that any future work on this dataset should report the 1.5 mV rule as a mandatory comparator, in the same way that a majority-class baseline is expected in imbalanced classification. This is a specific instance of the shortcut-learning problem described by Geirhos et al. [23] and of the evaluation failures catalogued for medical imaging by Roberts et al. [24] and Varoquaux and Cheplygina [25].

It would be wrong to read this as evidence that spatial context is useless in electroanatomical mapping. Two more mundane explanations come first. The geodesic and Euclidean metrics were numerically almost identical here (median ratio 1.005), so the comparison between them never had the opportunity to matter; a dataset with sparser sampling, or with epicardial and endocardial points on the same mesh, would separate them. And with nine patients the smallest attainable one-sided Wilcoxon p is 0.00195, so effects smaller than roughly 0.03 AUC cannot be resolved. Our ablation can exclude a large contribution from the graph; it cannot exclude a small one.

The interpretability analysis is, to us, the most encouraging part of the study. Attribution mass concentrated almost four-fold inside the expert-delineated interval without the delineation ever being used as a training signal. That the network independently converges on the segment the electrophysiologist marks suggests the temporal encoder does capture the morphology of interest. The limitation is that this morphology adds little once amplitude is known, at least for reproducing the consensus label.

Our results sit alongside rather than against the earlier work of Baldazzi and colleagues, who reported strong classification performance on an overlapping cohort using engineered descriptors [10]. Our gradient-boosting comparator, built on a similar descriptor family, is also the best model here. The difference is one of framing: we ask whether the learned model improves on the clinical rule, and on this dataset it does not.

Several limitations bound these conclusions. Nine patients from a single centre and a single mapping platform limit both statistical power and generalisability, although the leave-one-patient-out design keeps the evaluation honest with respect to between-patient transfer. The reference standard is expert consensus rather than clinical outcome, so a model that disagrees with the experts is not necessarily wrong; the distance to delivered lesions provides a second, independent spatial anchor, but delivered lesions reflect operator strategy rather than an absolute ground truth. Only endocardial left-ventricular data are available, and AVPs of epicardial or intramural origin are not represented. Finally, the dataset descriptor itself restricts the material to algorithm development, and we draw no clinical recommendation from it.

Three directions follow naturally. The first is to define the task where the clinical rule fails, for instance by predicting annotator disagreement rather than the consensus, since the 246 Unknown electrograms and the border-zone triplets are precisely the cases where an automated second opinion would be useful. The second is to anchor the label on something other than visual annotation, such as procedural outcome or LAVA elimination during ablation. The third is to test the graph formulation on cohorts where geodesic and Euclidean neighbourhoods actually diverge, which is where its inductive bias should finally pay off.

5 Conclusion

We benchmarked a spatio-temporal graph convolutional network against six comparators for the detection of abnormal ventricular potentials in the ARGO dataset, under leave-one-patient-out cross-validation with no information leakage across folds. The network did not outperform gradient boosting on expert descriptors, nor the conventional bipolar-voltage rule, and a 13-variant ablation attributed no measurable contribution to the graph, the geodesic metric, the attention mechanism or the auxiliary spatial task. Stratification by voltage shows why: expert annotation of AVPs in this dataset is largely determined by bipolar amplitude, with no physiological electrogram annotated below 0.5 mV and 95.6% of electrograms above 1.5 mV annotated physiological. Reporting the 1.5 mV rule as a comparator should become standard practice for work on this benchmark. The pipeline, including the native WFDB annotation decoder that was required to read the dataset correctly, is released openly.

Declarations

Ethics approval This is a secondary analysis of the openly available ARGO dataset. The original acquisition was approved by the local ethics committee (Prot. n. 351/2021/CE) and all patients gave written informed consent, as reported by the dataset authors. No additional approval was required.

Data availability The ARGO dataset is openly available on PhysioNet (<https://doi.org/10.13026/8gh2-e660>) under the CC BY-NC-SA 4.0 licence. No new patient data were generated.

Consent to participate and for publication Not applicable. The analysis uses fully anonymised data that the dataset authors released publicly; informed consent for the original acquisition was obtained by them.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests The author declares no competing interests.

Author contributions A. Bougarne is the sole author and was responsible for conceptualisation, methodology, software, formal analysis, investigation, data curation, visualisation, writing of the original draft, and review and editing.

Use of generative artificial intelligence A large language model was used to assist with code development for the analysis pipeline and with language editing of the manuscript. The author designed the study, verified every analytical step, checked all reported figures against the released notebook output, and takes full responsibility for the content of this work. No text was published without author verification.

Acknowledgements The author thanks the investigators of the ARGO study for releasing an annotated benchmark under an open licence, without which this analysis would not have been possible.

References

1. de Bakker, J. M. T. et al. Reentry as a cause of ventricular tachycardia in patients with chronic ischemic heart disease: electrophysiologic and anatomic correlation. *Circulation* **77** (3), 589–606 (1988).
2. Gardner, P. I., Ursell, P. C., Fenoglio, J. J. & Wit, A. L. Electrophysiologic and anatomic basis for fractionated electrograms recorded from healed myocardial infarcts. *Circulation* **72** (3), 596–611 (1985).
3. Stevenson, W. G. et al. Fractionated endocardial electrograms are associated with slow conduction in humans: evidence from pace-mapping. *J. Am. Coll. Cardiol.* **13** (2), 369–376 (1989).
4. Jais, P. et al. Elimination of local abnormal ventricular activities: a new end point for substrate modification in patients with scar-related ventricular tachycardia. *Circulation* **125** (18), 2184–2196 (2012).
5. Wolf, M. et al. Long-term outcome of substrate modification in ablation of post-myocardial infarction ventricular tachycardia. *Circ. Arrhythm. Electrophysiol.* **11** (2), e005635 (2018).
6. Sacher, F. et al. Substrate mapping and ablation for ventricular tachycardia: the LAVA approach. *J. Cardiovasc. Electrophysiol.* **26** (4), 464–471 (2015).
7. Cronin, E. M. et al. 2019 HRS/EHRA/APHRS/LAHRS expert consensus statement on catheter ablation of ventricular arrhythmias. *Europace* **21** (8), 1143–1144 (2019).
8. Anter, E., Tschabrunn, C. M., Buxton, A. E. & Josephson, M. E. High-resolution mapping of postinfarction reentrant ventricular tachycardia. *Circulation* **134** (4), 314–327 (2016).
9. Baldazzi, G. et al. Spectral characterisation of ventricular intracardiac potentials in human post-ischaemic bipolar electrograms. *Sci. Rep.* **12**, 4782 (2022).
10. Baldazzi, G., Orru, M., Viola, G. & Pani, D. Computer-aided detection of arrhythmogenic sites in post-ischemic ventricular tachycardia. *Sci. Rep.* **13** <https://doi.org/10.1038/s41598-023-33866-w> (2023).
11. Cauti, F. M. et al. Analysis of electrogram peak frequency during ventricular tachycardia ablation: insights into human tridimensional ventricular tachycardia circuits. *Heart Rhythm.* **22** (1), 128–134 (2025).
12. Tonko, J. B. et al. Near-field detection and peak frequency metric for substrate and activation mapping of ventricular tachycardias in two- and three-dimensional circuits. *Europace* **26**, euae154 (2024).
13. Orru, M. et al. The ARGO dataset: annotated and delineated intracardiac electrograms of post-ischemic ventricular tachycardia. *PLoS One.* **21** (6), e0350993 (2026).
14. Orru, M. et al. Annotated dataset of post-ischemic ventricular tachycardia electrograms (ARGO), version 1.0.0. PhysioNet. (2026). <https://doi.org/10.13026/8gh2-e660>
15. Morier, M., Rodriguez Padilla, J., Gallinari, P. & Sermesant, M. Learning cardiac electrophysiology with graph neural networks for fast data-driven personalised predictions. In: Functional Imaging and

- Modeling of the Heart (FIMH 2025). Lecture Notes in Computer Science, vol 15672. Cham: Springer; https://doi.org/10.1007/978-3-031-94559-5_13 (2025).
16. Ruiz Herrera, C. et al. Physics-informed neural networks to learn cardiac fiber orientation from multiple electroanatomical maps. *Eng. Comput.* **38**, 3957–3973 (2022).
 17. Marchlinski, F. E., Callans, D. J., Gottlieb, C. D. & Zado, E. Linear ablation lesions for control of unmappable ventricular tachycardia in patients with ischemic and nonischemic cardiomyopathy. *Circulation* **101** (11), 1288–1296 (2000).
 18. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; pp. 785 – 94. (2016).
 19. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76** (5), 378–382 (1971).
 20. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Mak.* **26** (6), 565–574 (2006).
 21. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML). PMLR; pp. 3319-28. (2017).
 22. Collins, G. S. et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).
 23. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
 24. Roberts, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).
 25. Varoquaux, G. & Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit. Med.* **5**, 48 (2022).
 26. Defferrard, M., Bresson, X. & Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In: Advances in Neural Information Processing Systems 29 (NeurIPS). pp. 3844-52. (2016).
 27. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (ICLR). (2017).
 28. Velickovic, P. et al. Graph attention networks. In: International Conference on Learning Representations (ICLR). (2018).
 29. Gepstein, L., Hayam, G. & Ben-Haim, S. A. A novel method for nonfluoroscopic catheter-based electroanatomical mapping of the heart: in vitro and in vivo accuracy results. *Circulation* **95** (6), 1611–1622 (1997).
 30. Di Cori, A. et al. Approaching ventricular tachycardia ablation in 2024: an update on mapping and ablation strategies, timing, and future directions. *J. Clin. Med.* **13** (17), 5017 (2024).

31. Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101** (23), e215–e220 (2000).

32. Baldazzi, G., Orru, M., Matraxia, M., Viola, G. & Pani, D. Supervised classification of ventricular abnormal potentials in intracardiac electrograms. *Comput. Cardiol. (CinC)*. **47**, 1–4. <https://doi.org/10.22489/CinC.2020.397> (2020).

Figures

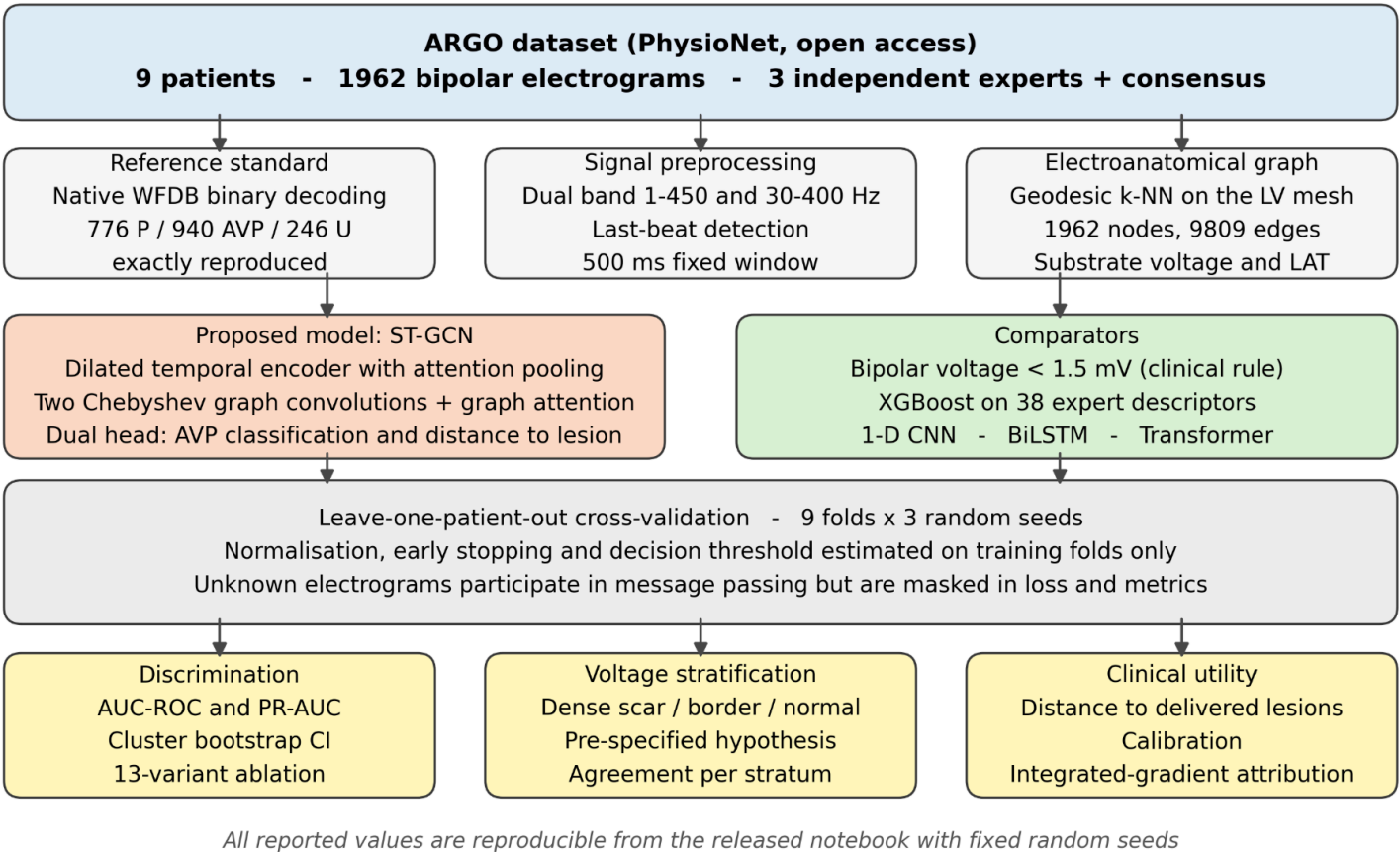


Figure 1

Study design and analysis pipeline. The reference standard was recovered with a native decoder for the WFDB binary annotation format and verified against the class counts published with the dataset. Each bipolar electrogram was processed in two frequency bands and windowed on the last ventricular depolarisation. One graph per patient was built on the reconstructed left-ventricular surface, with mapping sites as nodes and edges weighted by geodesic distance. Normalisation, early stopping and the decision threshold were estimated on training folds only.

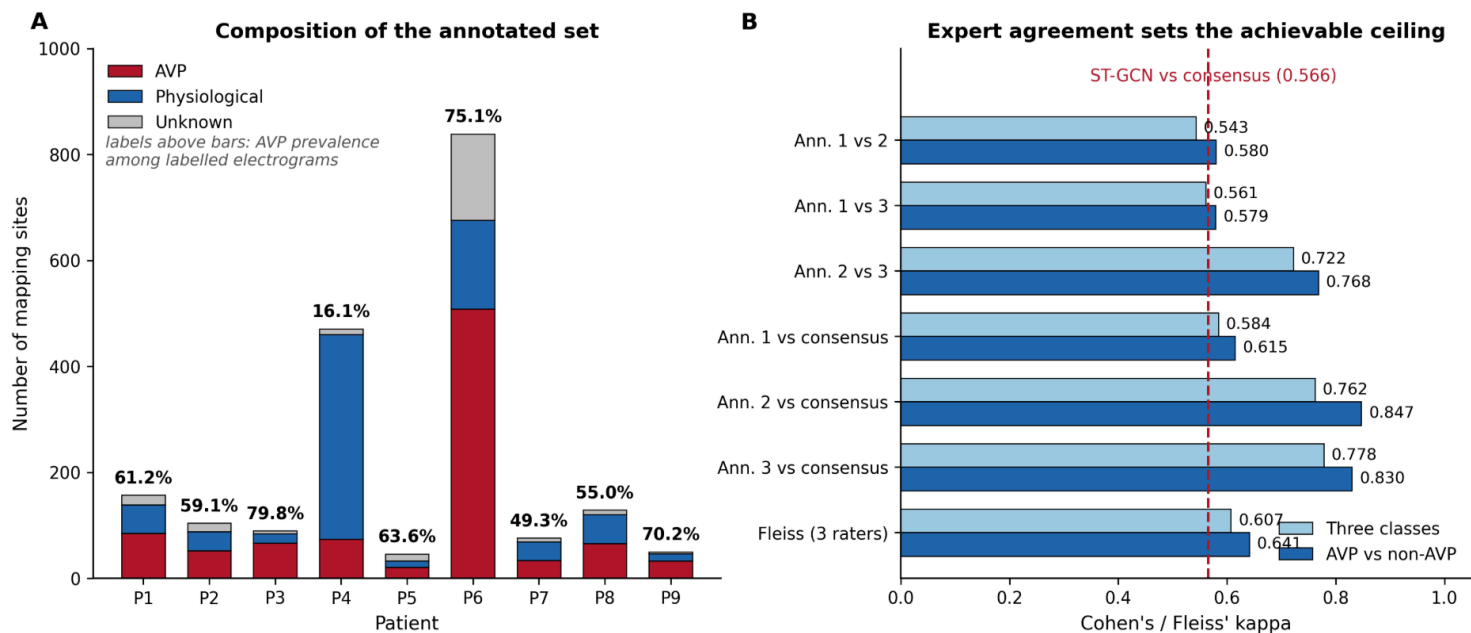


Figure 2

Cohort composition and expert agreement. a Number of mapping sites per patient by consensus class, with the abnormal-potential prevalence among labelled electrograms shown above each bar; prevalence ranges from 16.1% to 79.8% across the cohort. b Pairwise Cohen's kappa between annotators and against the consensus, together with Fleiss' kappa for the three raters jointly, for the three-class task and for the binary task. The dashed line marks the agreement reached by the proposed model.

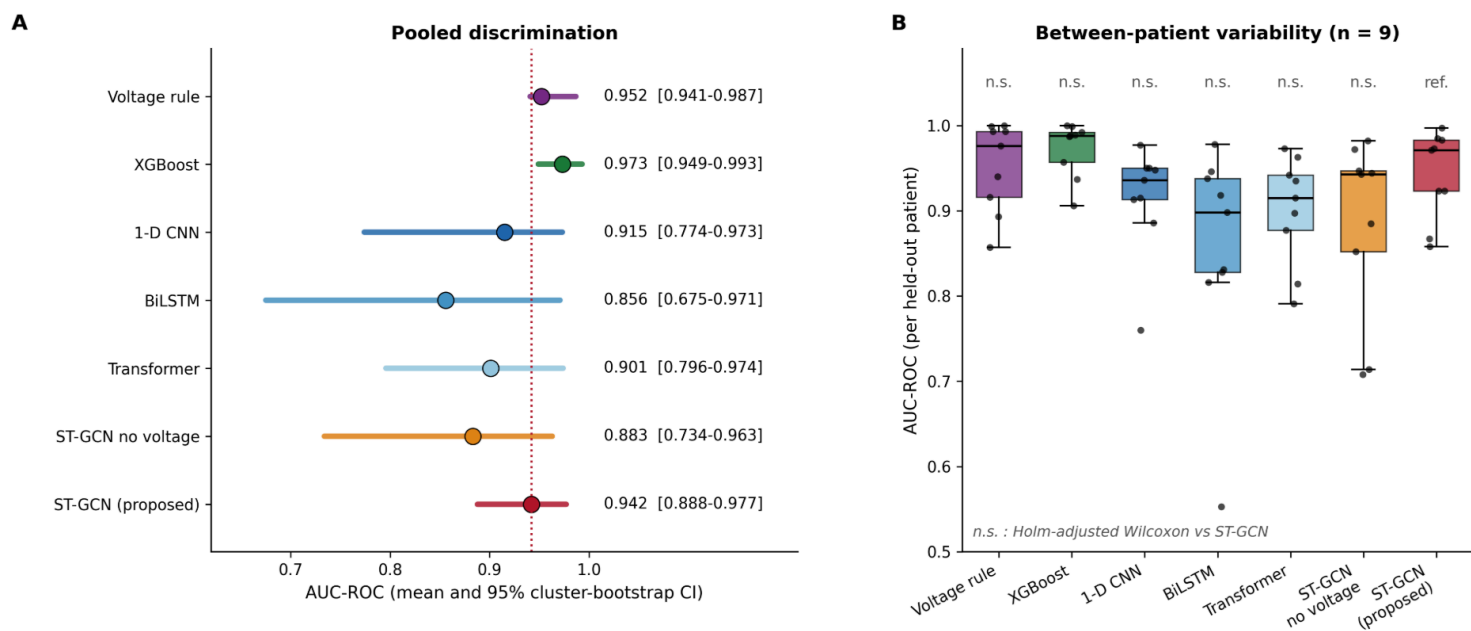


Figure 3

Discrimination under leave-one-patient-out cross-validation. a Mean AUC-ROC with 95% confidence intervals obtained by cluster bootstrap resampling patients rather than nodes (2000 replicates); the

dotted line marks the proposed model. b Distribution of per-patient AUC, one point per held-out patient. n.s. denotes a Holm-adjusted one-sided Wilcoxon signed-rank test against the proposed model that did not reach significance; at n = 9 the smallest attainable p value is 0.00195.

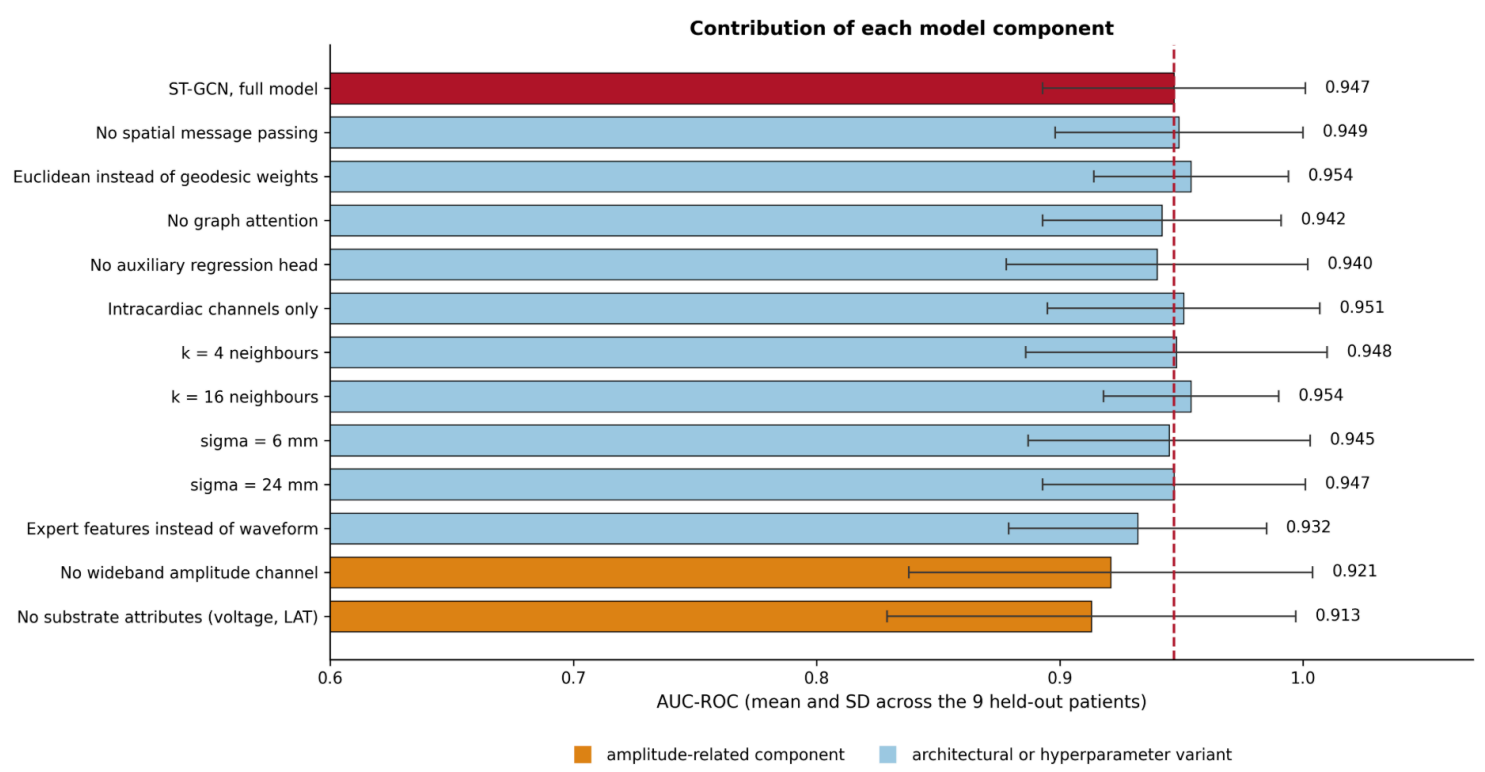


Figure 4

Ablation study. Each bar removes or modifies one element of the proposed model, all else being equal, with a reduced budget of 90 epochs and one seed. The dashed line marks the full model. Removing spatial message passing or replacing geodesic with Euclidean edge weights leaves performance unchanged. The two variants that lower the mean, shown in orange, both remove amplitude information.

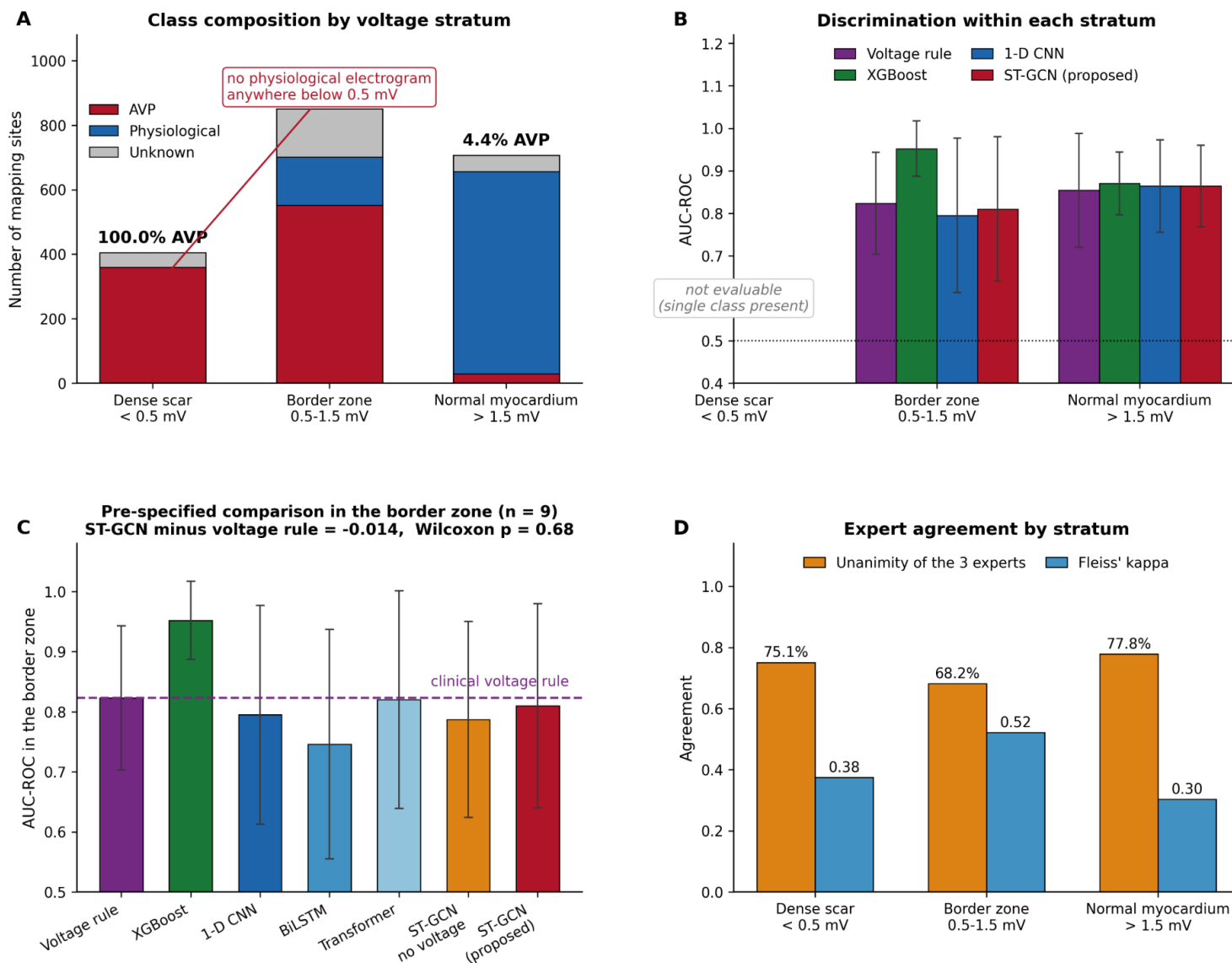


Figure 5

Stratification by bipolar voltage. a Class composition of each stratum; no electrogram below 0.5 mV was ever annotated physiological, and 95.6% of those above 1.5 mV were. b Discrimination within each stratum; the dense-scar stratum contains a single class and admits no metric. c Per-patient AUC restricted to the border zone, where the pre-specified hypothesis was tested. d Expert agreement by stratum: unanimity is lowest in the border zone, whereas Fleiss' kappa gives the opposite ranking because it is confounded by prevalence at the extremes.

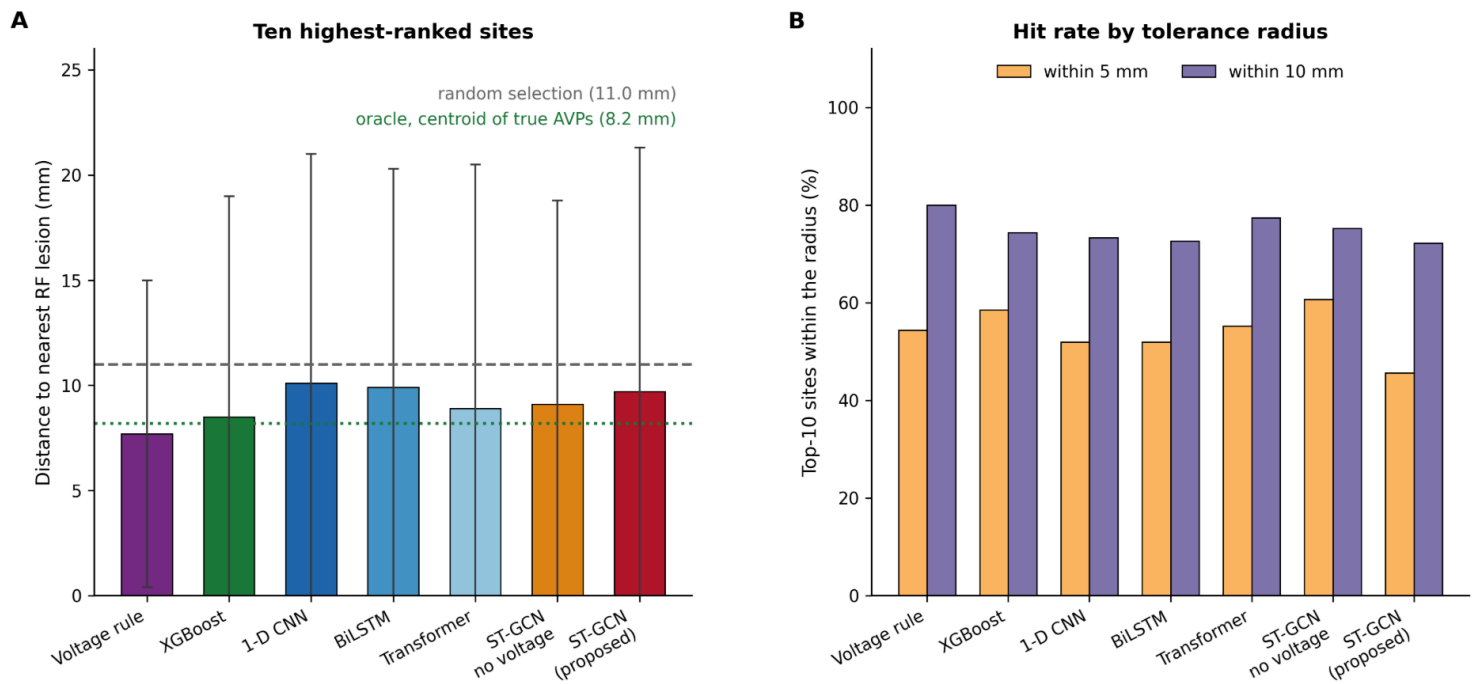


Figure 6

Localisation of ablation targets. a Mean distance from the ten highest-ranked sites to the nearest radiofrequency lesion actually delivered, bounded by uniform random selection above and by the oracle centroid of truly annotated abnormal potentials below. The whole exploitable range spans 2.8 mm. b Proportion of the ten highest-ranked sites lying within 5 and 10 mm of a delivered lesion.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterial.docx](#)