Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SAMVT-Net: Source-Aware Multi-View Token Learning for Heterogeneous ECG Classification

Runchuan Li
Zhongyuan University of Technology
Shangwen Zhang
Zhongyuan University of Technology
Shengya Shen
Shengya_Shen@163.com

Zhengzhou University of Economics and Business
Junjie Bian
Xizang University of Tibetan Medicine
Jingbin Zhang
Henan Luyun Software Co., Ltd.
Xinchao Han
Zhongyuan University of Technology
Hua-Nong Ting
Universiti Malaya

Research Article

Keywords: electrocardiogram, physiological multi-view tokens, Transformer fusion, prototype classification, long-tail learning, multi-label diagnosis

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11202938/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SAMVT-Net: Source-Aware Multi-View Token Learning for Heterogeneous ECG Classification

Runchuan Li[a,b,c], Shangwen Zhang[a,b], Shengya Shen[d,e,*], Junjie Bian[f], Jingbin Zhang[g], Xinchao Han[a], Hua-Nong Ting[h]

[a]School of Computer Science, Zhongyuan University of Technology, Zhengzhou, 451191, China
[b]International Joint Laboratory for AI Interpretability and Reasoning Applications, Zhongyuan University of Technology, Zhengzhou, 451191, China
[c]Henan Institute of Metrology, Zhengzhou, 450008, China
[d]School of Foreign Languages, Zhengzhou University of Economics and Business, Zhengzhou, 451191, China
[e]Henan Province Engineering Research Center of Multimodal Perception and Intelligent Interaction Technology, Zhengzhou, 451191, China
[f]Institute of Tibetan Medicine, Xizang University of Tibetan Medicine, Lhasa, 850000, China
[g]Henan Luyun Software Co., Ltd., China
[h]Department of Biomedical Engineering, Faculty of Engineering, Universiti Malaya, Jalan Pantai Baharu, 50603, Kuala Lumpur, Malaysia

Abstract

ECG classification requires representations that accommodate differences in signal duration, lead configuration, and label structure. We present SAMVT-Net, a compact supervised model built around physiological tokens and explicit source identity. The representation organizes the physiological information available to each task: heartbeat inputs combine waveform morphology with measured inter-beat timing and auxiliary spectral summaries, while record-level inputs use morphology and spectrum. Learnable lead identifiers inform attention-based fusion of these summaries. Task-specific readouts use cosine-similarity prototypes for single-label prediction and sigmoid outputs for co-occurring diagnoses, while retaining the same tokenization and fusion interface. Evaluation covers heartbeat classification on the MIT-BIH Arrhythmia Database and record-level diagnosis on PTB-XL, including its official patient-separated folds. Controlled comparisons associate the main multi-view gain with morphology–rhythm complementarity and show the clearest effect of lead identity in Transformer fusion. A separate decision-head study identifies the contribution of prototype classification within the chosen imbalance-aware training configuration. An auxiliary-token control supports omitting the zero-input token from the record-level fusion sequence. Matched-split reimplementations, representation geometry, graded input perturbations, and inference profiling further characterize predictive performance and execution cost. The results support source-aware physiological token learning as a compact approach to the evaluated ECG tasks, with component effects that depend on the input, backbone, and prediction objective.

Keywords: electrocardiogram, physiological multi-view tokens, Transformer fusion, prototype classification, long-tail learning, multi-label diagnosis

1. Introduction

Electrocardiogram (ECG) classification links electrical recordings to several distinct prediction problems. A heartbeat model assigns a class to a short segment, whereas a record-level model can assign several diagnoses to a longer recording. The MIT-BIH Arrhythmia Database and PTB-XL make this distinction concrete through different signal durations, lead configurations, and label structures [1, 2]. Here, heterogeneous ECG classification refers to these three differences. Recent ECG-based studies extend the prediction target to structural heart disease and sudden cardiac death [3, 4]. Their clinical scope also emphasizes why performance must be interpreted against a defined endpoint: detecting a beat abnormality, assigning record-level diagnoses, and predicting a future event require different labels and validation designs.

Deep networks can learn useful waveform features directly from ECG recordings [5]. Yet local morphology, the spacing of neighboring beats, and the spatial information carried by different leads are not interchangeable. A short R-centered segment can resolve beat shape without explicitly encoding adjacent RR intervals. Similarly, features extracted with the same encoder from two leads have the same vector format but different acquisition meanings. A compact model must decide which information to retain before fusion and how to identify its origin.

Existing methods address these questions through temporal encoders, lead graphs, or structured tokens. Heterogeneous graph attention represents intra-lead and inter-lead relationships explicitly [6]. ST-MEM identifies leads within a spatiotemporal masked-reconstruction model [7]. These studies establish source identity as a useful design principle. Its value after each lead has been compressed into a few physiological tokens by shared encoders is less clear, as is its dependence on the fusion operator. This distinction matters because pooling changes the information available to fusion, while different operators can exploit token content and ordering differently.

A further distinction concerns the decision rule. Heartbeat

*Corresponding author: Shengya Shen.
Email address: shengyashen@foxmail.com (Shengya Shen)

categories are mutually exclusive and unevenly represented, whereas diagnostic labels may co-occur. Sampling, loss weighting, and prototype classification modify different parts of this problem. When these choices change together, a performance gain cannot be assigned to sample exposure, the training objective, or decision geometry. Controlled comparisons under the same upstream architecture and data partition are needed to establish what decision geometry adds beyond the loss and sample exposure.

SAMVT-Net addresses these questions through separately trained task models that share a physiological token interface. Heartbeat inputs combine morphology, measured RR timing, and auxiliary spectral summaries. Record-level inputs use multi-lead morphology and spectrum, with no surrogate RR token in the adopted configuration. Lead and view identifiers precede fusion, and the readout follows the label structure. The contributions are:

1. **Physiology-structured multi-view representation.** The model organizes available physiological information into task-dependent compact views. Heartbeat inputs use morphology and measured beat timing with an auxiliary spectral view; record-level inputs use morphology and spectral summaries. Controlled comparisons identify morphology–rhythm complementarity in heartbeat classification and test whether a constant auxiliary token is needed for record-level diagnosis.
2. **Source-aware token fusion.** Explicit lead identifiers condition interactions among physiological tokens produced by shared encoders. Separate lead/view controls and CNN, TCN, and Transformer comparisons test where this design is useful, with the clearest observed effect in attention-based fusion.
3. **Task-adapted prototype decisions.** A cosine-similarity prototype readout accommodates single-label tasks within the common architecture; sigmoid readouts accommodate co-occurring diagnoses. A controlled loss–sampler–head sequence distinguishes the prototype contribution from established training strategies, and a separate output-head comparison examines label interaction. Input views and readouts adapt to the available information and label structure while preserving the common fusion interface.

Section 2 positions these choices within existing work. Sections 3–5 describe the tasks, architecture, and evidence; Sections 6–7 interpret the results and present the conclusions.

2. Related work

2.1. From waveform features to physiological views

Feature-based ECG methods make the information entering a classifier explicit. Morphological descriptors paired with support vector machines focus on waveform shape [8]. Wavelet-packet statistics and scattering transforms instead describe variation across scales [9, 10]. Multifeature stacking with DWKNN extends the descriptor set used for heartbeat classification [11]. Feature-based random forests have also been applied to premature ventricular contraction detection, while attributable features have been paired with an AdaBoost–random-forest classifier [12, 13]. Their results depend jointly on descriptor selection and classifier choice. A broader feature set alone does not establish which physiological views add information beyond those already available.

Learned temporal encoders shift this choice toward data-driven feature extraction. CNN–BiLSTM models combine local convolutions with sequence context [14]. Transformer models learn content-dependent interactions over continuous ECG inputs [15]. ECG-Mamba uses convolutional reduction and bidirectional state-space modeling, together with a task-specific augmentation scheme [16]. These methods differ in how they retain and mix temporal information. BiLSTM-Treg preserves relationships between consecutive beats through heartbeat sequences [17], whereas explicit RR descriptors supply beat spacing directly [18]. For an R-centered waveform, the relevant question is whether measured beat intervals add information beyond the morphology learned within the segment; sequence-model capacity alone does not resolve that comparison.

Time–frequency representations provide another route to multiple views. Chen et al. apply Swin Transformer classification to wavelet time–frequency maps [19]. This retains a two-dimensional transform representation before hierarchical attention. SAMVT-Net instead reduces each lead to a pooled morphology token and a coarse spectral token, then adds RR timing for heartbeat tasks. This choice reduces the sequence presented to fusion. The remaining issue is the conditional value of a view: a spectral summary may help a morphology-only model yet add little once beat timing is available. The view-combination study therefore measures spectrum both with and without rhythm, distinguishing complementary information from an increase in the number of branches.

2.2. Source identity and token fusion

Multi-view fusion has two separate requirements: representing complementary signal properties and preserving their source. Multimodal spatiotemporal ECG models combine features at different representational levels [20]. MFCT, for example, fuses ConvNeXt features from recurrence-plot images with Transformer features from the original waveform [21]. Recent arrhythmia models extend this direction in different ways: MVFF-SAT combines raw-heartbeat, morphological, and pathological views with temporal fusion, while BiSAM-ECGNet aligns recurrence-plot, raw-signal, and pathological-sequence branches in a lightweight cross-modal architecture [22, 23]. These approaches enlarge the information presented to the classifier. Lead graphs address a different question by encoding relationships between acquisition channels; heterogeneous graph attention distinguishes intra-lead and inter-lead interactions [6]. Neither representational diversity nor an explicit graph isolates the value of source identification itself. The question for compact fusion is whether an additive lead identifier provides useful acquisition information to pooled tokens without requiring a lead graph.

ST-MEM incorporates lead and temporal structure into masked ECG reconstruction [7]. Its temporal patches preserve locations to support reconstruction; SAMVT-Net's tokens summarize physiological views before supervised fusion. Lead em-

beddings are thus an established tool, but their use does not establish equal roles for lead and view identity after physiological encoding. In SAMVT-Net, one morphology encoder processes every lead, while different encoders process morphology and spectrum. Lead tags can identify an acquisition source that the shared encoder does not label explicitly; view tags may duplicate distinctions already introduced by the separate encoders. These are different hypotheses and require separate controls.

We therefore test lead and view identifiers separately and compare the effect of lead identification across CNN, TCN, and Transformer fusion.

2.3. Pretraining and architectural reuse

ECG-BERT studies temporal-semantic modeling for arrhythmia classification [24]. ECG-FM pursues reusable ECG representations through masked and contrastive pretraining [25]. The cardiac sensing foundation model extends this direction to heterogeneous cardiac signals and associated text, with adaptation across sensing configurations [26]. The scale and supervision of pretraining also differ. ECG-LFM combines contrastive and masked learning across more than ten million ECGs before downstream adaptation [27]. ECG-CLIP aligns ECGs with clinical text, enabling diagnostic queries without condition-specific training [28]. These approaches expand the information available during representation learning, but their performance depends on pretraining exposure, the downstream endpoint, and the adaptation protocol. They do not isolate the value of physiological pooling or source tags in a compact model trained from scratch.

These pretraining studies leave open whether a compact supervised model can retain one tokenization–fusion interface across heartbeat-level single-label and record-level multi-label tasks without shared pretrained weights. SAMVT-Net addresses this question through separately trained task models, changing the available inputs and readout while preserving the common interface. The auxiliary-token control tests whether record-level inputs need an additional token when no record-specific RR descriptor is supplied. The main comparisons consequently use models trained from scratch on the same task partitions. The shifted-window baseline adapts the Swin attention pattern to one-dimensional inputs [29]. The original source method and its local architecture-level implementation are distinguished throughout the experimental comparison. ECG-FM is discussed as a foundation-model contextual reference, with its results confined to the Supplementary Material.

2.4. Decision geometry under imbalance and co-occurring labels

Loss design and sample exposure act at different stages of learning. Focal loss emphasizes hard examples [30]. LDAM changes class-dependent margins, while logit adjustment uses class priors to alter decision scores [31, 32]. These objectives do not imply that stacking them with a sampler will improve every operating point. In particular, a change in minority recall can coincide with a change in precision. The relevant comparison therefore holds the representation fixed and examines the loss, sampler, and decision head in sequence.

Prototypical networks classify examples through distances to class representatives [33]. ProtoECGNet develops case-based, multi-label ECG interpretation with separate rhythm, local-morphology, and global-spatial branches [34]. Its prototypes connect predictions to representative ECG patterns. SAMVT-Net's learned prototypes instead serve as cosine-similarity directions in a fused feature space; they are decision parameters, not retrieved waveform explanations. The open question within this representation is the incremental value of prototype geometry over a linear decision rule under the same loss and sampling scheme. That question is distinct from retrieving explanatory cases, and the comparison with a simple cross-entropy classifier is also needed to interpret the cost and benefit of the complete training configuration.

For co-occurring diagnoses, one output per label is required, but explicit interaction between output tokens remains a design choice. Zhu et al. evaluated automated multilabel diagnosis of rhythm and conduction abnormalities in a clinical cohort [35]. That formulation establishes the relevance of concurrent predictions; it does not by itself determine whether an interacting label head improves a fixed upstream representation. Deep metric learning with KNN provides one alternative formulation for multi-labelled ECG classification [36]. Whether explicit label-token interaction improves aggregate prediction over independent sigmoid outputs remains to be tested under the same upstream architecture and training protocol. The output-head comparison tests that increment and reports HYP sensitivity separately, because a shift in one label's recall need not improve overall prediction. These controls assess the incremental value of the readout within each task.

3. Data preprocessing and task construction

Both datasets are publicly available PhysioNet resources: MIT-BIH Arrhythmia Database is used for heartbeat-level single-lead experiments [1], whereas PTB-XL is used for ten-second twelve-lead record-level experiments [2]. The following subsections specify signal processing, label construction, and partitioning. Each matched comparison reuses the partition files specified for its reference model. The task-level results, matched output-head comparison, and later shared-reference studies are reported as separate experiment series. We use MIT5 and MIT9 to denote the five- and nine-class tasks constructed from the MIT-BIH Arrhythmia Database.

3.1. MIT-BIH Arrhythmia Database heartbeat tasks

MIT-BIH Arrhythmia Database signals are sampled at 360 Hz. For each record, the implementation selects MLII when available, followed by lead II or ECG1, and otherwise uses the first lead. A third-order Butterworth band-pass filter with a 0.5–45 Hz passband is applied using forward–backward filtering. Heartbeats are centered on the annotated R peaks, retaining 100 samples before and 160 samples (approximately 0.44 s) after each peak, which produces a 260-sample segment. Each segment is normalized using the median and interquartile range,

$$\widehat{x_t} = \frac{x_t - \mathrm{median}(x)}{Q_{0.75}(x) - Q_{0.25}(x) + \epsilon}. \tag{1}$$

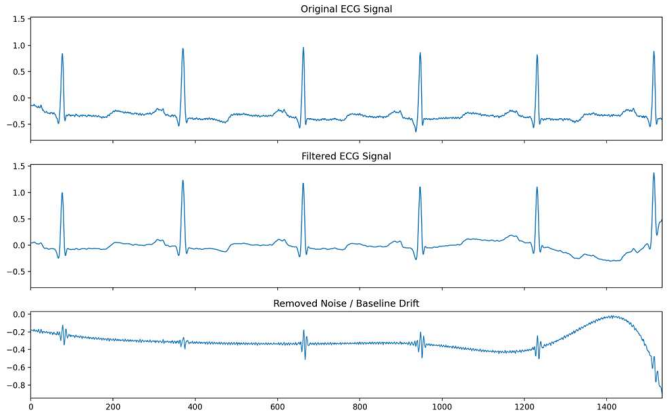


Figure 1: Representative MIT-BIH Arrhythmia Database signal preprocessing. The implemented pipeline uses a 0.5–45 Hz Butterworth filter followed by heartbeat-level robust normalization.

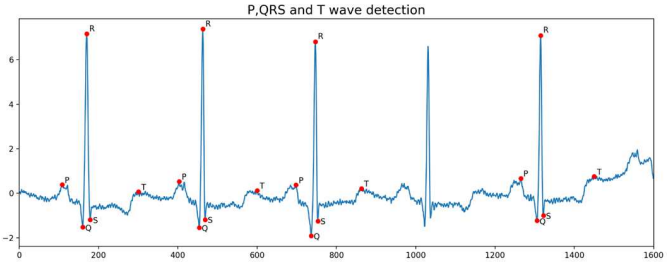


Figure 2: Representative P-QRS-T morphology around an annotated R peak; the model input contains 100 pre-R and 160 post-R samples.

The five-class task uses the groups N, S, V, F, and Q. The nine-class task uses the grouped labels N, L, R, A, V, E, F, P, and f in the same order as the implementation. Their annotation mappings are specified in Section 5.3.2. Labeled heartbeats are pooled and divided into class-stratified training, validation, and test subsets in a 7:2:1 ratio with seed 42. The resulting partition files are reused throughout the MIT-BIH Arrhythmia Database experiments. The partitioning unit is the individual heartbeat.

For heartbeat i, the rhythm descriptor contains the preceding interval, the following interval, and a normalized interval ratio:

$$\boldsymbol{r}_i = \left[RR_i^-,\ RR_i^+,\ \frac{RR_i^-}{\mathrm{mean}^+(RR_i^-, RR_i^+) + 10^{-6}} \right], \tag{2}$$

where RR_i^- and RR_i^+ denote the preceding and following RR intervals, and mean^+ averages the available positive intervals. Use of RR_i^+ requires the next R peak, making this an offline heartbeat-classification setting. The optional DWT branch used in the input-view comparison interpolates the fourth-level db4 detail coefficients to 260 samples and applies the same robust normalization. Figure 1 illustrates the filtering and normalization pipeline, and Fig. 2 shows the heartbeat window around an annotated R peak.

3.2. PTB-XL record-level tasks

PTB-XL inputs are ten-second, twelve-lead records sampled at 100 Hz [2]. The supplied preprocessors offer lead-wise median/IQR normalization and clipped z-score normalization. For binary classification, a record is labeled ABNORMAL if it contains at least one non-NORM superclass. The five-label task predicts normal ECG (NORM), myocardial infarction (MI), conduction disturbance (CD), ST/T change (STTC), and hypertrophy (HYP). A record can carry more than one label; Table 1 reports the earlier custom-split test supports.

The earlier experiments use fixed training, validation, and test partitions in a 7:2:1 ratio, with the ECG record as the partitioning unit. Their preprocessing implementations provide binary-label stratification and multilabel stratification with fallback procedures. Archived predictions, thresholds, and run arguments define those results. The earlier five-label configurations include a constant auxiliary token; their results remain identified with that configuration.

The additional five-label evaluation follows the recommended PTB-XL v1.0.3 folds [37]: folds 1–8 for training, fold 9 for validation, and fold 10 exclusively for testing. Diagnostic statements are mapped to the five diagnostic superclasses; 411 records with none of these labels are excluded. The retained subsets contain 17,084/2,146/2,158 records from 14,823/1,917/1,877 patients, respectively. The archived split-validation report passes all checks and records zero patient overlap for every pair of subsets. Supplementary Table S39 gives the label counts. Both PTB-XL tasks use complete ten-second records. The adopted record-level fusion sequence omits the constant auxiliary token; the official-fold experiment compares it directly with the earlier 26-token configuration (Section 5.13).

4. Proposed SAMVT-Net

4.1. Overall architecture

The main symbols are collected in Supplementary Table S1; local definitions accompany the corresponding equations.

For an input $X \in \mathbb{R}^{C\times T}$ with C leads and T samples, the network produces one morphology token and one spectral token per lead. MIT-BIH Arrhythmia Database inputs also supply a rhythm token. Lead and view embeddings identify the physiological tokens before they enter the Transformer with a classification token. The classification-token output feeds a prototype head for the three single-label tasks or a label-interaction head for PTB-XL five-label prediction (Fig. 3). Including CLS, the sequence contains four tokens for MIT-BIH Arrhythmia Database and 25 for both adopted PTB-XL configurations: twelve morphology tokens and twelve spectral tokens. The earlier five-label configuration adds a zero-input auxiliary token, giving 26 tokens. It is retained as a named reference in the original studies and in the official-fold control, rather than being treated as a measured rhythm view.

4.2. Physiological multi-view token encoder

4.2.1. Multi-scale morphology token

Morphology varies over short deflections and wider waveform segments. Parallel kernel widths provide several receptive fields before temporal pooling, without requiring delineation of individual waves. For each lead X_c, five parallel branches use kernel sizes $\mathcal{K} = \{5, 9, 17, 33, 65\}$. Each branch consists of a standard convolution, batch normalization, GELU, a depthwise

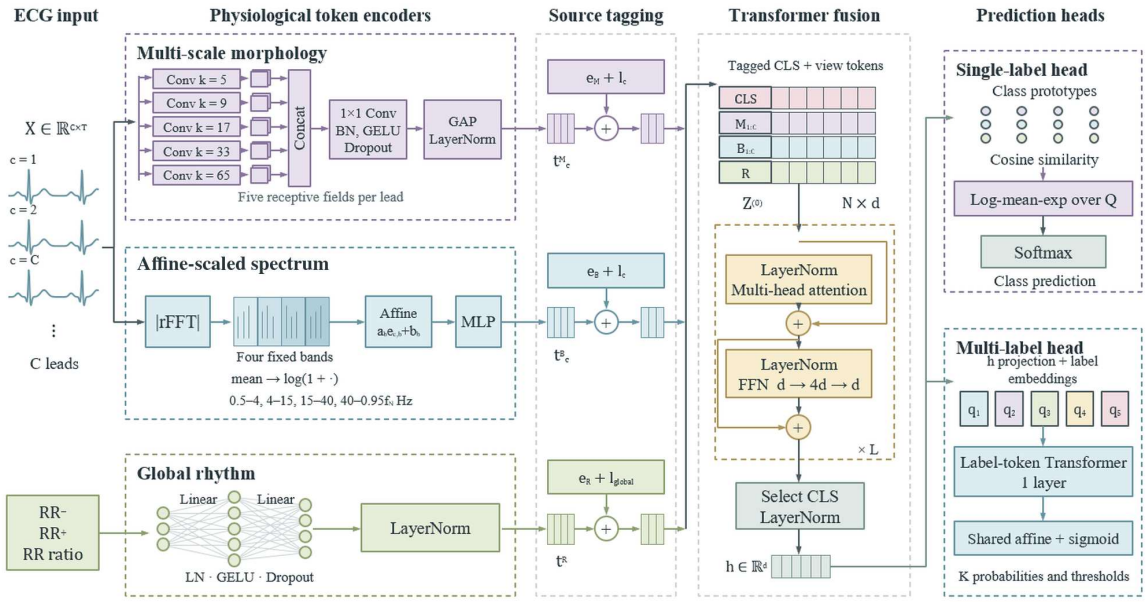


Figure 3: SAMVT-Net data flow. Each lead contributes morphology (M) and spectral (B) tokens; RR timing contributes a global rhythm (R) token for heartbeat inputs. Additive lead/view identifiers precede Transformer fusion and the CLS readout. The output path is selected for the task; weights are trained separately. Heartbeat models use measured RR timing and prototype outputs. The rhythm path shown is used only when measured RR inputs are supplied. Both adopted PTB-XL configurations omit that path from fusion; binary classification uses prototype outputs and five-label classification uses sigmoid label outputs. The earlier 26-token five-label control is described separately in the text. CLS denotes the classification token.

convolution with kernel size 3, batch normalization, and GELU. Let

$$H_{c,k} = \phi(\mathrm{BN}(\mathrm{DWConv}_3(\phi(\mathrm{BN}(\mathrm{Conv}_k(X_c)))))) \,. \tag{3}$$

The branch outputs are concatenated along channels. A 1×1 convolution, batch normalization, GELU, and dropout mix the resulting features before global average pooling and layer normalization:

$$\begin{aligned} \widetilde{H}_c &= \mathcal{D}(\phi(\mathrm{BN}(\mathrm{Conv}_1([H_{c,k}]_{k\in\mathcal{K}})))) \,, \\ \boldsymbol{t}_c^M &= \mathrm{LN}(\mathrm{GAP}(\widetilde{H}_c)) \in \mathbb{R}^d. \end{aligned} \tag{4}$$

Here ϕ denotes GELU, $\mathcal{D}$ denotes dropout (inactive at evaluation), BN denotes batch normalization, LN denotes layer normalization, and GAP averages over the temporal axis. The same tokenizer weights are applied to each lead. Branch widths sum to d; for $d = 128$, they are 25, 25, 25, 25, and 28 channels. The concatenation $[H_{c,k}]_{k\in\mathcal{K}}$ is along channels. For stride-one branches, the two convolutions have an effective receptive field of $k + 2$ samples, spanning 7–67 samples across branches. This corresponds to approximately 19–186 ms at 360 Hz and 70–670 ms at 100 Hz. The widths cover progressively broader local patterns; they are not assigned to fixed P, QRS, or T boundaries. Global pooling produces one morphology token per lead. Temporal interactions within a lead are therefore summarized before the token-fusion stage.

4.2.2. *Affine-scaled spectral-band token*

The spectral branch exposes coarse magnitude information as a separate view, allowing its contribution to be compared with the convolutional representation. Four broad bands keep the descriptor small; they are fixed design choices rather than learned physiological boundaries. The spectral tokenizer summarizes the real-FFT magnitude $|\mathcal{F}(X_c)|$ in four fixed bands: 0.5–4 Hz, 4–15 Hz, 15–40 Hz, and 40 Hz to 0.95 times the Nyquist frequency. For the FFT bins assigned to band Ω_b, the descriptor is

$$e_{c,b} = \log\left(1 + \frac{1}{\max(1, |\Omega_b|)} \sum_{f\in\Omega_b} |\mathcal{F}(X_c)_f|\right). \tag{5}$$

Each log-magnitude descriptor is rescaled by a trainable coefficient a_b and offset b_b. An MLP projects the resulting four values to dimension d:

$$\boldsymbol{t}_c^B = \mathrm{MLP}\Big([a_b e_{c,b} + b_b]_{b=1}^B\Big) \in \mathbb{R}^d. \tag{6}$$

Here $B = 4$, f indexes FFT bins, and $|\Omega_b|$ is the bin count. The implementation includes the lower band edge and excludes the upper edge; an empty band yields zero. The MLP follows Linear–LayerNorm–GELU–Dropout–Linear–LayerNorm. For non-four-band sensitivity runs, the edges are equally spaced between 0.5 Hz and 0.95 times the Nyquist frequency; those runs vary the partition as well as the descriptor count. The band limits remain fixed during training; only the scale and offset of each spectral descriptor are learned. The descriptor averages magnitude, not spectral power, and the logarithm compresses its numerical range. Frequency resolution is approximately 1.38 Hz for a 260-sample heartbeat at 360 Hz and 0.1 Hz for a ten-second record. In the heartbeat pipeline, the 0.5–45 Hz filter also attenuates the upper part of the fourth band. Thus, band values depend on both sampling and preprocessing; the four bands are compact engineering summaries, not disease-specific physiological intervals.

4.2.3. Rhythm token

A 260-sample R-centered window does not explicitly contain both neighboring beat intervals. The rhythm branch supplies this timing information separately. Because the descriptor uses the following interval RR_i^+, it is available only after the next R peak and is used here for offline classification. For MIT-BIH Arrhythmia Database, the three RR descriptors are mapped by a two-layer MLP with LayerNorm, GELU, and dropout:

$$\boldsymbol{t}^R = \mathrm{LN}(W_2\mathcal{D}(\phi(\mathrm{LN}(W_1\boldsymbol{r} + \boldsymbol{b}_1))) + \boldsymbol{b}_2)\,. \tag{7}$$

The matrices W_1, W_2 and biases $\boldsymbol{b}_1, \boldsymbol{b}_2$ are learned, and $\mathcal{D}$ is the dropout operation. The rhythm branch produces one global token per heartbeat. It supplies no token to the adopted PTB-XL fusion sequence. In the earlier five-label configuration, its fixed input $[0, 0, 0]$ produces the same auxiliary vector for every record at evaluation; dropout remains active during training. This vector is a learned constant, not record-specific rhythm information. The official-fold control keeps the same allocated modules in both variants and changes only whether this vector enters fusion. Figure 4 summarizes the encoders; the zero-input note refers to that earlier auxiliary control.

4.3. Lead- and view-aware Transformer fusion

Pooling maps all views to the same feature dimension but does not encode their lead or view names. Additive source embeddings provide those identifiers to attention without specifying a graph or changing the attention operator. Let $v \in \{M, B, R\}$ denote the view type and c the lead index. Every token is tagged by a learnable view embedding $\boldsymbol{e}_v$ and lead embedding $\boldsymbol{l}_c$:

$$\widetilde{\boldsymbol{t}}_{v,c} = \boldsymbol{t}_{v,c} + \boldsymbol{e}_v + \boldsymbol{l}_c. \tag{8}$$

The rhythm token uses a designated global-lead embedding. The classification token is also given a dedicated view embedding, $\widetilde{\boldsymbol{t}}_{\mathrm{cls}} = \boldsymbol{t}_{\mathrm{cls}} + \boldsymbol{e}_{\mathrm{cls}}$, without a lead embedding. Prepending it produces

$$Z^{(0)} = [\widetilde{\boldsymbol{t}}_{\mathrm{cls}}; \widetilde{\boldsymbol{t}}_{M,1:C}; \widetilde{\boldsymbol{t}}_{B,1:C}; \widetilde{\boldsymbol{t}}_R], \tag{9}$$

where the rhythm term is included for measured heartbeat RR descriptors and omitted for the adopted PTB-XL configurations. The earlier five-label reference substitutes its constant auxiliary vector at that position. A pre-normalized Transformer encoder with L layers and R heads yields

$$Z^{(L)} = \mathrm{TransformerEncoder}(Z^{(0)}), \qquad \boldsymbol{h} = \mathrm{LN}(Z_0^{(L)}). \tag{10}$$

The feed-forward sublayer expands from d to $4d$ and uses GELU and dropout. Self-attention updates the source-tagged tokens jointly without a predefined graph. Its queries and keys are linear projections of the normalized tagged features, so attention weights depend on both signal content and learned source identity. This makes source identity available to fusion without imposing a physical calibration, an orthogonality constraint, or a predefined lead relation. Source-tagging ablations bypass the relevant additions; the lead-masking study excludes unavailable morphology and spectral tokens using a key-padding mask. With Transformer fusion disabled, the implementation instead averages the source-tagged physiological tokens and applies LayerNorm; the CLS token is not used in that ablation. The optional DWT comparison branch uses a separate morphology tokenizer and its own view embedding and is disabled in the main models. Figure 5 shows the resulting sequence of token concatenation, source embedding, Transformer fusion, and classification-token readout.

4.4. Task-specific classification heads

4.4.1. Multi-prototype single-label head

The prototype head compares a fused feature with learned class representatives using normalized similarity. More than one representative allows several directions within a class, although the benefit of multiplicity is a separate empirical question. For each class k, the single-label head maintains Q prototypes $\boldsymbol{p}_{k,q}$. After ℓ_2 normalization, the prototype similarities are

$$s_{i,k,q} = \frac{\boldsymbol{h}_i^\top \boldsymbol{p}_{k,q}}{\|\boldsymbol{h}_i\|_2 \|\boldsymbol{p}_{k,q}\|_2}. \tag{11}$$

The head aggregates prototypes using a log-mean-exp operation with a learnable positive scale $\rho = \exp(\zeta)$ initialized at 20:

$$z_{i,k} = \log \sum_{q=1}^{Q} \exp(\rho s_{i,k,q}) - \log Q. \tag{12}$$

This head is used for MIT-BIH Arrhythmia Database five-class, MIT-BIH Arrhythmia Database nine-class, and PTB-XL binary classification. The nine-class model additionally contains a five-group auxiliary linear head whose targets are derived from the grouped nine-class labels, as specified in Section 5.3.2. During inference, the softmax prediction uses the unadjusted prototype logits $z_{i,k}$. Training-time logit adjustment is applied only inside the LDAM-Focal objective defined below.

4.4.2. PTB-XL five-label interaction head

A single softmax cannot represent co-occurring diagnoses. The five-label formulation therefore assigns one sigmoid probability to each label. For the multi-label task, the prototype head is replaced by K trainable label embeddings $\boldsymbol{q}_k$. Each label token receives the projected global feature,

$$Q_k^{(0)} = \boldsymbol{q}_k + W_h\boldsymbol{h} + \boldsymbol{b}_h, \tag{13}$$

Here $\boldsymbol{b}_h$ is the bias of the context projection. A single pre-normalized Transformer layer with four attention heads and feed-forward width $4d$ updates the five label tokens jointly in the reported configuration. A shared affine projection (including a scalar bias) produces one logit per label, and sigmoid functions yield the label probabilities. The label embedding makes each output token distinct despite their shared input context, and self-attention allows their context-dependent responses to interact. The resulting scores are learned predictive associations; no clinical dependency graph is supplied. Section 5.6.3 compares this complete output formulation with an independent affine readout applied to the fused feature in a matched training series. Figure 6 contrasts the single-label and multi-label output paths and distinguishes the training-only logit adjustment from inference.

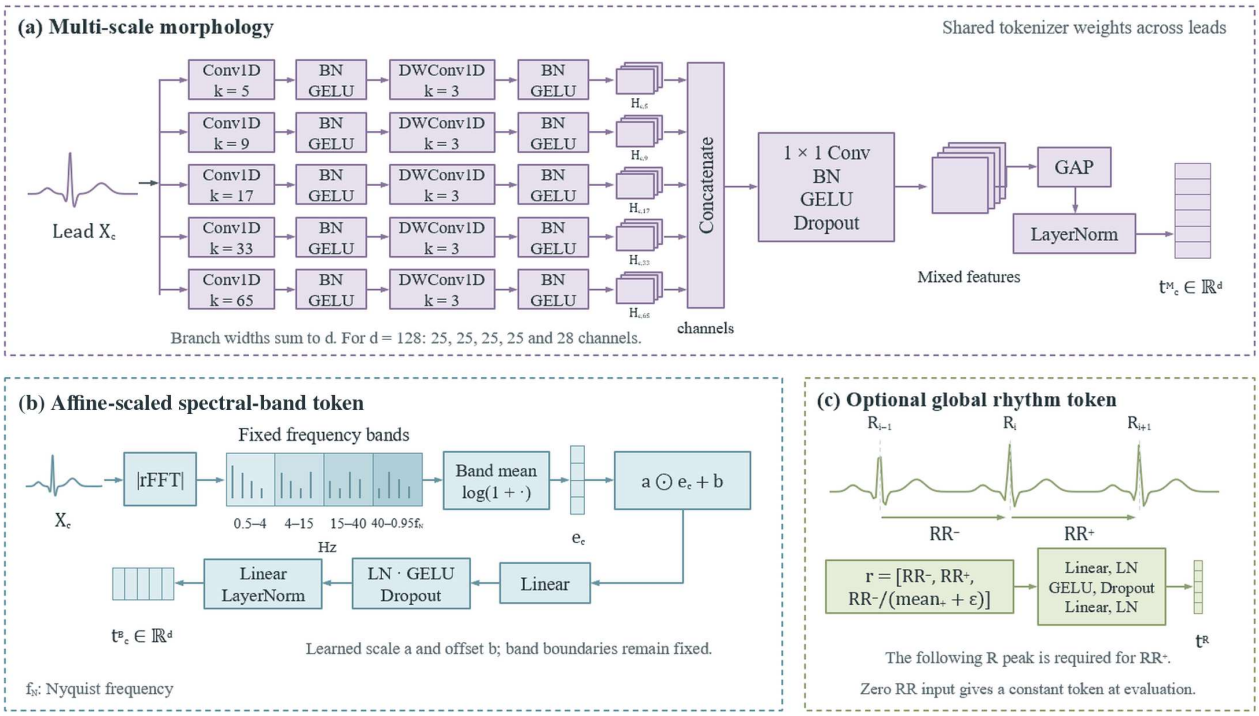


Figure 4: Physiological token encoders. (a) Five convolution branches, channel mixing, and global average pooling produce one morphology token per lead. (b) Fixed-band real-FFT magnitudes undergo learned affine scaling and projection; f_N denotes the Nyquist frequency. (c) RR inputs produce one global timing token and require the following R peak. BN denotes batch normalization, LN layer normalization, DWConv depthwise convolution, GAP global average pooling, and MLP multilayer perceptron. Panel (c) supplies a fusion token only for measured RR inputs in the adopted architecture. Its zero-input annotation describes the earlier auxiliary control. PTB-XL branch settings are specified in Table 2.

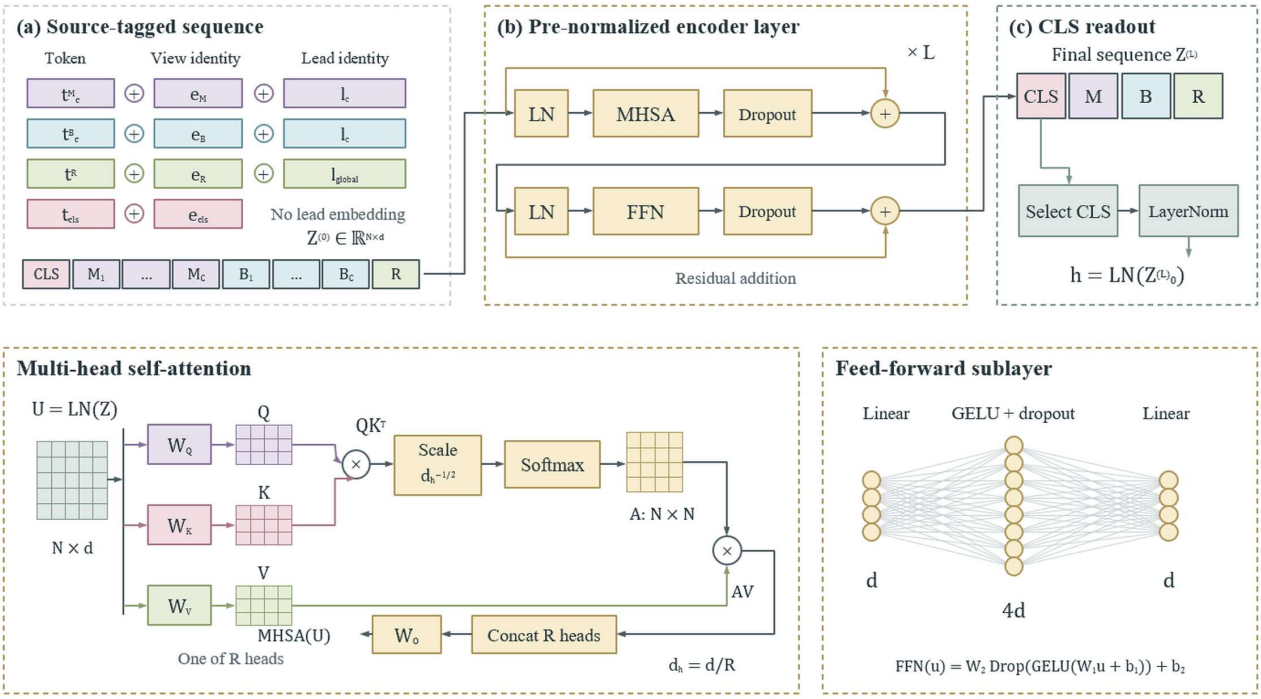


Figure 5: Source-tagged fusion. (a) View and lead embeddings identify the physiological tokens; CLS has its own view embedding. (b) Each pre-normalized layer applies multi-head self-attention (MHSA), a feed-forward network (FFN), and residual additions. Here N is the sequence length, d the token width, R the head count, and $d_h = d/R$. (c) The normalized CLS output is the fused feature $\boldsymbol{h}$. The R position is optional and absent in the adopted PTB-XL sequence. No predefined graph is used.

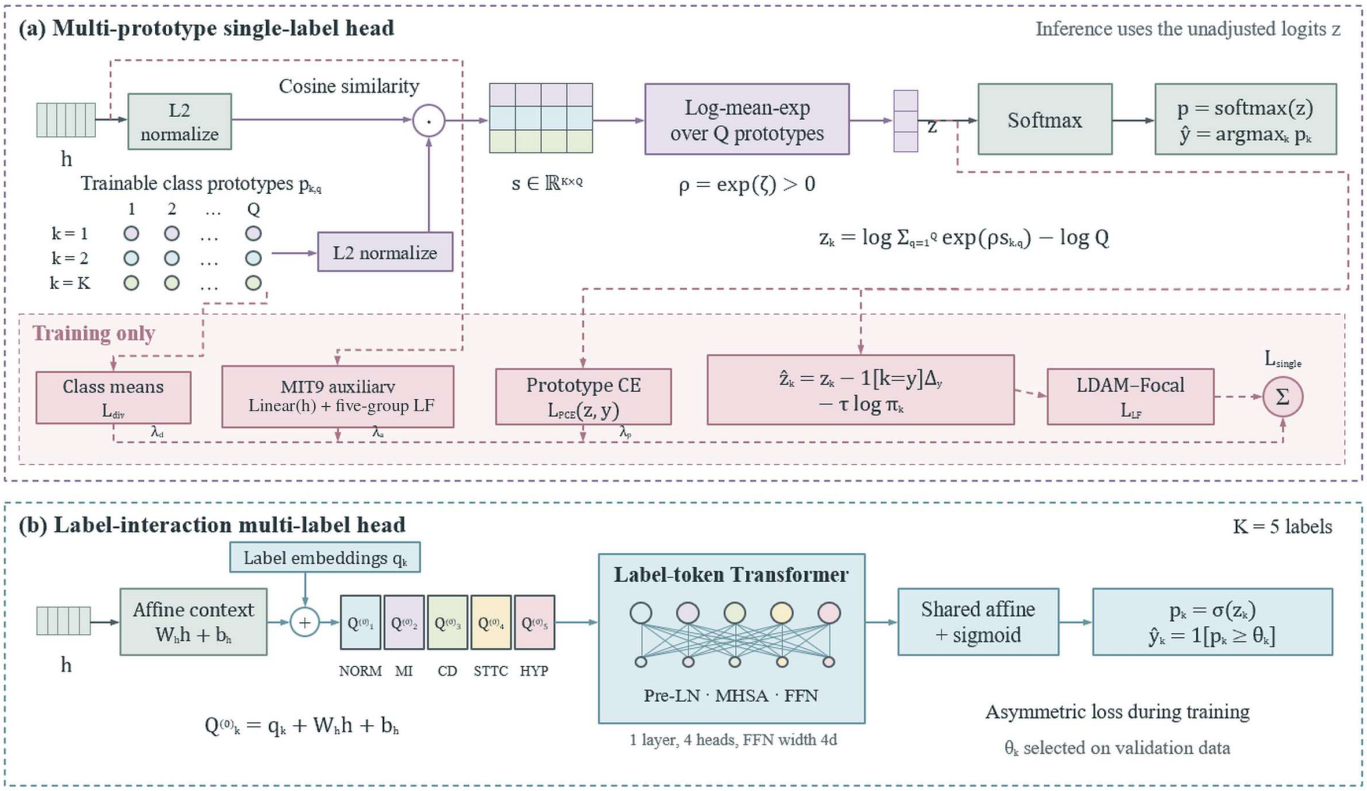


Figure 6: Task-specific output and training paths. (a) Cosine similarities to Q prototypes per class are combined by log-mean-exp before softmax. Dashed connections denote training-only losses; logit adjustment is not applied at inference. (b) Five label tokens share the projected context $\boldsymbol{h}$, pass through one label-token Transformer layer, and produce sigmoid probabilities. Thresholds θ_k are selected on validation data.

4.5. Single-label long-tail objective

Let K denote the number of output classes and n_k the training count for class k. The LDAM margin is

$$\Delta_k = \frac{n_k^{-1/4}}{\max_c n_c^{-1/4}} m_{\max}, \tag{14}$$

where $m_{\max} = 0.5$. During training, the target-class margin and class prior are applied when computing the LDAM-Focal objective:

$$\widehat{z}_{i,k} = z_{i,k} - \mathbb{I}(k = y_i)\Delta_{y_i} - \tau \log \pi_k, \qquad \pi_k = \frac{n_k}{\sum_c n_c}. \tag{15}$$

Inference uses the unadjusted prototype logits $z_{i,k}$. Effective-number class weights use $\beta = 0.9999$ and are normalized to unit mean before clipping to $[0.5, 8]$. Let ℓ_i^{CE} be the weighted cross-entropy evaluated on $\widehat{z}_i$. The LDAM-Focal term is

$$\mathcal{L}_{\mathrm{LF}} = \frac{1}{N} \sum_i \left(1 - \exp(-\,\mathrm{stopgrad}(\ell_i^{\mathrm{CE}}))\right)^{\gamma_f} \ell_i^{\mathrm{CE}}, \tag{16}$$

with $\gamma_f = 2$. Because ℓ_i^{CE} includes a class weight, its exponential is not generally the unweighted target-class probability used in standard focal loss. A further cross-entropy term, $\mathcal{L}_{\mathrm{PCE}}$, is applied to the prototype logits. An inter-class penalty constrains the similarity of normalized mean prototypes $\bar{\boldsymbol{p}}_k$:

$$\mathcal{L}_{\mathrm{div}} = \frac{1}{K(K-1)} \sum_{k \neq c} \max\left(0, \bar{\boldsymbol{p}}_k^{\top} \bar{\boldsymbol{p}}_c - m_{\mathrm{div}}\right), \tag{17}$$

where $m_{\mathrm{div}} = 0.1$. The single-label objective is

$$\mathcal{L}_{\mathrm{single}} = \mathcal{L}_{\mathrm{LF}} + \lambda_p \mathcal{L}_{\mathrm{PCE}} + \lambda_d \mathcal{L}_{\mathrm{div}} + \lambda_a \mathcal{L}_{\mathrm{aux}}, \tag{18}$$

Here $\lambda_p = 0.1$ and $\lambda_d = 0.02$. The auxiliary term has weight $\lambda_a = 0.3$ for MIT-BIH Arrhythmia Database nine-class classification and is absent in the other tasks. It applies a separate LDAM-Focal criterion to the five-group auxiliary logits, using auxiliary class counts and the same margin, focal, and logit-adjustment settings; it is not a plain auxiliary cross-entropy. The prototype CE term, in contrast, is unweighted cross-entropy on the unadjusted main logits. The diversity penalty acts between class means; it does not explicitly separate the prototypes within an individual class. These terms address distinct training operations: margins modify class decisions, weighting changes example contributions, and the prototype penalty regularizes class-representative geometry. Their combination defines the training configuration, rather than a newly derived loss family.

4.6. Balanced sampling and multi-label objective

Loss weighting acts only on the examples presented in an update. Sampling additionally changes how often each class is seen, which matters when rare classes occupy a small fraction of the training set. The sampler interpolates between a square-root class distribution, $s_k = \sqrt{n_k}/\sum_c \sqrt{n_c}$, and a uniform distribution, $u_k = 1/K$:

$$\omega_k = \alpha u_k + (1-\alpha)s_k, \qquad w_i = \frac{\omega_{y_i}}{n_{y_i}}. \tag{19}$$

A weighted random sampler draws N training samples with replacement using w_i. Thus, $\alpha = 0$ is pure square-root class sampling, $\alpha = 1$ is uniform class sampling, and disabling the sampler restores natural shuffled sampling. The final MIT configuration uses $\alpha = 0.5$.

For PTB-XL five-label classification, the sampler and prototype losses are disabled. The asymmetric loss uses $\gamma_- = 4$, $\gamma_+ = 1$, and negative-probability clipping of 0.05. Per-label thresholds are selected on the validation set by maximizing the label-wise F1 over $\{0.05, 0.06, \ldots, 0.95\}$ and are then fixed for test evaluation.

4.7. Optimization procedure

The single-label configurations use a compact AdEMAMix-style implementation [38], which combines short- and long-term gradient averages. This choice retains gradient history while allowing recent updates to influence the step. The PTB-XL five-label runs use AdamW. No optimizer comparison was performed. For AdEMAMix, the gradient g_t updates the averages as follows:

$$\begin{aligned} m_t^{(s)} &= \beta_1 m_{t-1}^{(s)} + (1-\beta_1)g_t, \\ m_t^{(l)} &= \beta_3 m_{t-1}^{(l)} + (1-\beta_3)g_t, \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2)g_t^2. \end{aligned} \tag{20}$$

After bias correction, the parameter update is

$$\theta_t = (1-\eta\lambda)\theta_{t-1} - \eta\frac{\widehat{m}_t^{(s)} + \alpha_A \widehat{m}_t^{(l)}}{\sqrt{\widehat{v}_t} + \epsilon}. \tag{21}$$

$\widehat{m}_t^{(s)} = m_t^{(s)}/(1-\beta_1^t)$, $\widehat{m}_t^{(l)} = m_t^{(l)}/(1-\beta_3^t)$, and $\widehat{v}_t = v_t/(1-\beta_2^t)$ denote bias correction; η is the current learning rate. The optimizer uses $(\beta_1, \beta_2, \beta_3) = (0.9, 0.999, 0.9999)$, $\alpha_A = 0.5$, $\epsilon = 10^{-8}$, and decoupled weight decay $\lambda = 10^{-4}$. A cosine-annealing schedule decreases the learning rate to 1% of its initial value. The single-label defaults enable mixed precision on CUDA and clip gradients to norm 1.0. The five-label runs disable mixed precision and clip gradients to norm 0.5. Early-stopping patience is 20.

For the completed single-label runs, checkpoint selection used the composite validation score implemented in the training code,

$$S_{\text{single}} = 0.45F1_{\text{macro}} + 0.25Sen_{\text{macro}} + 0.10G_{\text{sel}} + 0.20F1_F, \tag{22}$$

where the final term is replaced by Macro-F1 if class F is absent. The implementation-specific selection term is $G_{\text{sel}} = \left(\prod_{k=1}^{K} \text{Sen}_k\right)^{1/(2K)}$. For reporting, the standard G-mean definition is $G = \left(\prod_{k=1}^{K} \text{Sen}_k\right)^{1/K}$. The trained checkpoints were not reselected after this reporting correction. For PTB-XL five-label classification,

$$S_{\text{multi}} = 0.6AUC_{\text{macro}} + 0.4F1_{\text{macro}}. \tag{23}$$

For multi-label checkpoint selection, Macro-F1 is computed at a fixed threshold of 0.5. After loading the selected checkpoint, label-wise F1 thresholds are tuned on validation predictions and fixed for testing. The four-task experiments and one-factor sensitivity runs use seed 42; the repeated-run study adds seeds 7 and 21 and reports the mean and sample standard deviation. Each sensitivity run uses validation-based checkpoint selection followed by evaluation on the fixed test set. Supplementary Table S2 distinguishes the experiment series. The official-fold comparison uses the same multi-label checkpoint rule, with label-wise thresholds selected only on fold 9 and then fixed before fold-10 evaluation.

5. Experimental results and analysis

5.1. Evaluation protocol and experiment series

Table 1 combines the MIT5 sample distribution and the earlier custom-split PTB-XL five-label test supports; official-fold supports are reported separately in Supplementary Table S39. Single-label metrics are derived from one-versus-rest confusion counts; the reported G-mean is $G = (\prod_k \text{Sen}_k)^{1/K}$. Multi-label AUC uses continuous scores, whereas F1 and subset accuracy use validation-selected thresholds. Macro averaging gives each class equal weight. Three-seed summaries use seeds 7, 21, and 42 and report sample standard deviations. These summarize initialization variability with the data partition held fixed.

The task-level results, matched output-head comparison, and later shared-reference studies are separate experiment series. Each ablation or control is compared with its own reference, and each series is summarized separately. The original MIT5 three-seed $\tau = 1.0$ result records the original task-level evaluation; the new matched series anchors the controlled architecture comparison. The later shared references anchor source-tagging and lead-masking comparisons. All custom-split PTB5 results, including the architecture comparisons, source/head controls, perturbations, and profiles, retain their earlier 26-token reference. The official-fold study evaluates both the 26-token reference and the adopted 25-token configuration. Scores and component effects are not transferred between these configurations or partitions. Supplementary Table S2 records these roles.

Same-split neural baselines provide architectural references under the reported input, capacity, and training settings. Published scores give context under their respective evaluation protocols. For the source-tagging study, Holm correction covers 18 paired tests: three variants and six metrics. Each contrast uses three paired training seeds. Additional paired analyses use separate multiplicity families for the six output-head metrics and the 19 lead-masking Macro-F1 contrasts. The official-fold study adds a six-metric auxiliary-token family and a four-metric comparison of the adopted configuration with CNN. We report two-sided paired t tests, exact Wilcoxon signed-rank tests, and

Table 1: MIT5 class supports and earlier custom-split PTB-XL positive test supports. PTB-XL labels can co-occur; official-fold counts are in Supplementary Table S39.

MIT5 class	Train	Validation	Test
N	63,414	18,118	9,059
S	1,947	556	278
V	5,065	1,446	724
F	561	161	80
Q	5,628	1,608	804
PTB-XL: 2,139 test records			
NORM: 950; MI: 548; CD: 522; STTC: 486; HYP: 269			

unadjusted 95% confidence intervals for mean differences; Holm correction is applied within each family. With three seed pairs, the smallest attainable two-sided exact Wilcoxon p value is 0.25. These tests concern training variability on a fixed partition. Bold entries identify numerical maxima within the stated comparison, not statistical significance. The new controlled series uses seeds 7, 21, and 42. Its full-model Macro-F1 is 95.78 ± 0.71% on MIT5 and 72.40 ± 0.20% on PTB-XL five-label prediction (PTB5). These references are used for the matched architecture comparisons and the new perturbation tests; they are kept distinct from the original task results and the earlier source-tagging series. Across-seed SD describes initialization variability on fixed partitions. Paired contrasts use matched seed labels, with confidence intervals and multiplicity corrections reported for the stated comparison families in Supplementary Tables S29 and S32–S34. Three seed pairs give coarse exact-test resolution; observed mean differences are therefore distinguished from statistical significance.

5.2. Implementation settings

Table 2 lists the principal settings. MIT-BIH Arrhythmia Database configurations use up to 120 epochs and batch size 128. The five-label runs use up to 100 epochs, training/evaluation batches of 64/128, AdamW, and seeds 7, 21, and 42 in repeated comparisons. The matched MIT-BIH Arrhythmia Database component-ablation script specifies evaluation batches of 256.

The reference width $d = 128$ is divisible across four attention heads; two encoder layers act on the short physiological-token sequence. Four spectral descriptors and two prototypes per class keep these branches compact. These define the reference configurations; Section 5.11 reports the saved post hoc parameter responses. The morphology kernels have widths 5, 9, 17, 33, and 65 samples. Dropout is 0.2 in the single-label configurations and 0.15 in the five-label runs. The single-label heads use two prototypes per class; weight decay is 10^{-4}. Initialization follows the archived implementation. Spectral scales start at one and offsets at zero. Prototype and label-token entries are drawn from a zero-mean normal distribution with standard deviation 0.02; the CLS token uses truncated-normal initialization with standard deviation parameter 0.02. Lead and view tables use the default normal initialization of `nn.Embedding`, and convolutional and affine layers retain their PyTorch defaults. Source embeddings and spectral affine parameters are unconstrained. Prototype normalization is applied in the forward computation, while the positive similarity scale is parameterized by an exponential. Configuration switches follow the available inputs and target structure: measured RR timing for heartbeat tasks and no rhythm token in either adopted record-level fusion sequence.

Table 2 describes the adopted configuration; earlier PTB5 results use the 26-token reference defined in Section 4.2.3. The official-fold comparison uses AdamW, learning rate 10^{-3}, weight decay 10^{-4}, gradient clipping at 0.5, and early-stopping patience 20, with a maximum of 100 epochs. A cosine schedule reaches 10^{-5} at the full budget. All five configurations use the same asymmetric loss, validation criterion, and label-wise threshold grid from 0.05 to 0.95 in steps of 0.01. The available run histories record the selected epoch and validation score. Common kernel widths correspond to different durations at 360 and 100 Hz, so common architecture does not imply identical temporal scales.

5.3. Task-level classification results

5.3.1. MIT-BIH Arrhythmia Database five-class classification

The primary MIT-BIH Arrhythmia Database five-class result uses $\tau = 1.0$ and seeds 7, 21, and 42 (Table 3). Accuracy is 99.28 ± 0.04%, Macro-F1 is 95.55 ± 0.17%, and F-class F1 is 86.97 ± 0.96%. The $\tau = 1.5$ run has the highest Macro-F1 within the logit-adjustment sweep and is reported separately as a selected single-run result.

Supplementary Table S3 and Fig. 7a detail that selected single run. Of the 80 F-class test beats, 73 are correctly classified, giving 91.25% sensitivity and 90.68% F1. With this support, one additional error changes sensitivity by 1.25 percentage points.

5.3.2. MIT-BIH Arrhythmia Database nine-class classification

The nine-class results are summarized in Supplementary Table S4 and Fig. 7b. Accuracy is 98.96% and Macro-F1 is 94.62%. The small E and F groups have F1 values of 80.00% and 87.90%, respectively. The display order is N, L, R, A, V, E, F, P, and f, matching the internal class order.

These are grouped annotation categories, rather than nine individual diagnostic labels. L and R denote left and right bundle branch block beats. A groups annotations A, a, J, and S; V groups V and r; and E groups ventricular, atrial, and junctional escape annotations E, e, and j. F denotes fusion of ventricular and normal beats. P represents paced beats (annotation "/"), and f represents fusion of paced and normal beats. The underlying annotation meanings follow the WFDB definitions [39].

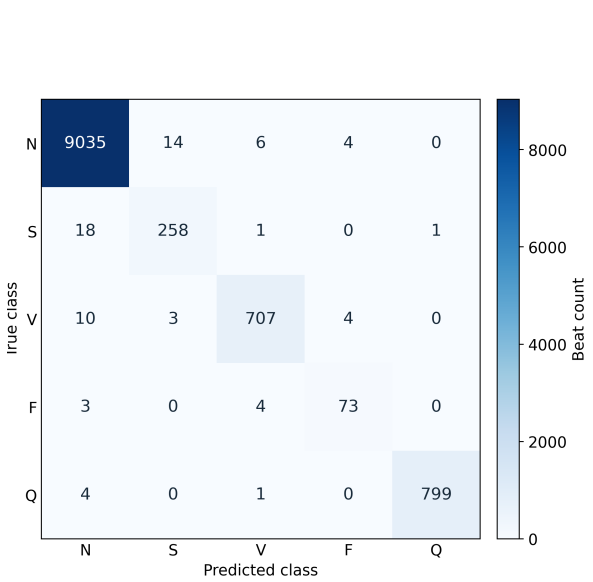
Training includes a five-group auxiliary loss with weight 0.3. In the supplied implementation, the auxiliary targets are derived from the grouped nine-class labels: N/L/R map to N, A to S, V/E to V, F to F, and P/f to Q. This grouped mapping assigns all E-group beats to auxiliary V, including e and j. It therefore differs from applying the five-class task's annotation-level mapping directly, where e and j belong to N.

5.3.3. PTB-XL binary and five-label classification

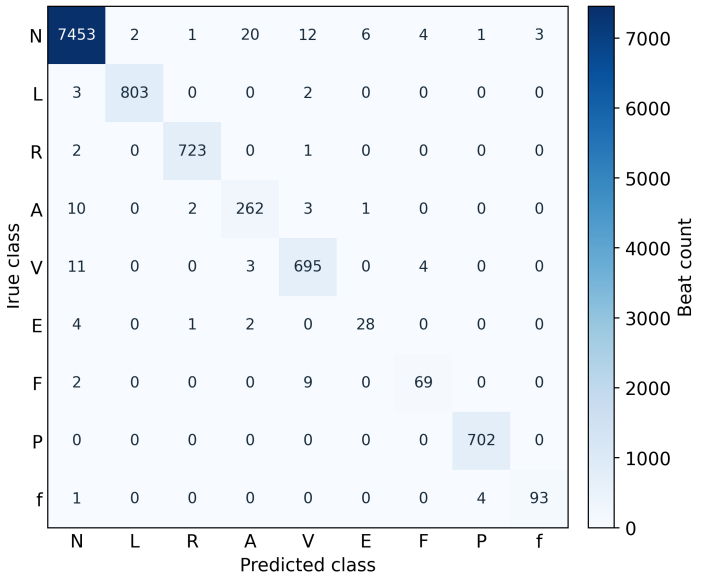
For PTB-XL binary classification, accuracy is 86.30% and Macro-F1 is 86.12% (Supplementary Table S5). The confusion matrix contains 108 normal records predicted as abnormal and

Table 2: Main implementation settings

Setting	MIT5	MIT9	PTB binary	PTB five-label
Input	1×260	1×260	12×1000	12×1000
Batch size	128	128	64	64
Initial learning rate	10^{-3}	10^{-3}	3×10^{-4}	10^{-3}
d, bands, layers, heads	128, 4, 2, 4	128, 4, 2, 4	128, 4, 2, 4	128, 4, 2, 4
Rhythm token in fusion	Measured	Measured	No	No
Total tokens, including CLS	4	4	25	25
Output head	Prototype	Prototype + auxiliary five-group	Prototype	Label interaction
Primary loss	LDAM-Focal	LDAM-Focal + auxiliary	LDAM-Focal	Asymmetric loss
Sampler	Balanced, $\alpha = 0.5$	Balanced, $\alpha = 0.5$	Balanced, $\alpha = 0.5$	Disabled



(a) MIT5, selected $\tau = 1.5$ run.



(b) MIT9, implementation class order.

Figure 7: SAMVT-Net heartbeat test confusion counts. Rows denote true classes and columns predictions. Cells and colors show unnormalized beat counts, with one prediction per test beat. MIT5 contains 10,945 test beats and represents the selected single run; neither panel averages over seeds.

185 abnormal records predicted as normal (Fig. 8). For the earlier 26-token five-label configuration on the custom split, the selected checkpoint is at epoch 39, with a validation selection score of 0.8285. Test Macro-AUC is 91.84%, Macro-F1 is 72.59%, Micro-F1 is 76.90%, and subset accuracy is 60.12%. Supplementary Table S6 lists the fixed, validation-selected thresholds and per-label results. Performance varies substantially by label: NORM F1 is 86.46%, whereas HYP F1 is 50.19%. This five-label experiment evaluates adaptation to twelve-lead inputs with co-occurring diagnostic outputs. The saved prediction audit and per-label counts are retained with the reproducibility materials.

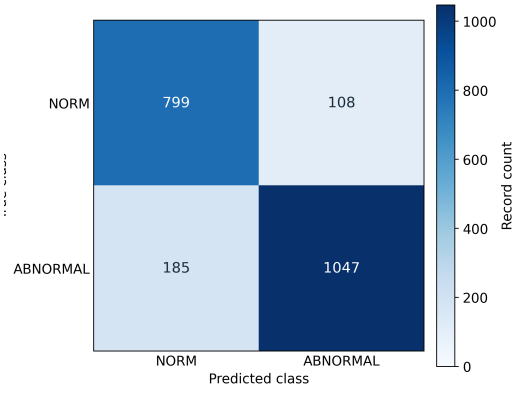


Figure 8: PTB-XL binary confusion matrix for 2,139 test records. Rows denote true classes and columns predictions; cells and the linear color scale show unnormalized record counts from a single run.

The later shared PTB-XL reference gives 91.94 ± 0.15% Macro-AUC, 72.70±0.61% Macro-F1, 77.18±0.30% Micro-F1, and 60.18 ± 1.46% subset accuracy. HYP-F1 is 49.43 ± 2.11%.

The corresponding shared MIT5 reference gives 95.44 ± 0.21% Macro-F1 and 86.61 ± 1.89% F-class F1. These later runs provide the reference for their associated comparisons. The adopted 25-token PTB5 configuration is evaluated on the official folds in Section 5.13; its results do not replace the earlier custom-split scores.

5.4. Matched-split architecture comparisons

Table 4 compares SAMVT-Net with four architecture-level reimplementations under the same task partitions and seeds. The PTB5 comparison uses the custom split and the earlier 26-token SAMVT-Net configuration. The supplied protocol specifies training from scratch, with a 120-epoch MIT5 budget and a 100-epoch PTB5 budget. ECG-Mamba uses a pure-PyTorch selective state-space implementation based on the corresponding architecture family [16]. ProtoECGNet uses a compact one-dimensional case/prototype branch [34]. ECG-BERT combines CNN tokenization with a BERT-style Transformer encoder [24]. Swin-ECG adapts shifted-window attention to a one-dimensional ECG sequence [29]. These are matched-split reimplementations, rather than executions of the complete original training and pretraining recipes.

SAMVT-Net has the highest mean Macro-F1 among these implementations on both tasks: 95.78 ± 0.71% on MIT5 and 72.40 ± 0.20% on PTB5. The corresponding margins over the strongest baseline mean are 2.81 points over ECG-BERT and 0.90 points over ProtoECGNet. On MIT5, Macro-Sen is 96.07% and F-class F1 is 87.97%; on PTB5, Macro-AUC is 91.53% and Micro-F1 is 76.71%. These results compare the supplied implementations, including their input handling and task-specific training choices. Equal partitions and seed labels do not isolate the architecture from every training choice or equate model capacity. Complete scores, seed-wise results, and paired Macro-F1 contrasts are provided in the Supplementary Material.

5.5. Physiological views and task-adapted decisions

5.5.1. Morphology–rhythm complementarity

The controlled view study holds the full-model training configuration fixed while changing the active physiological branches (Table 5a). Morphology alone gives 94.82% Macro-F1. Adding spectrum gives 94.93%, whereas adding rhythm gives 95.66%. The full M+S+R model gives 95.78%, a 0.96-point difference from morphology alone, with F-class F1 increasing from 86.13% to 87.97%. The main observed complementarity is therefore between morphology and rhythm.

Conditioned on rhythm, the spectral increment is 0.12 points. Its paired 95% confidence interval is $[-1.88, 2.12]$ points. The full-model versus morphology-only interval is $[-1.89, 3.80]$ points. The small sample of initializations supports a description of these observed effects, rather than a claim that all three views contribute independently or significantly. Spectrum is retained as an auxiliary summary of waveform content. Earlier input-view and component batches remain in Supplementary Tables S7–S9 with their own reference settings.

5.5.2. Loss, sampling, and prototype readout

The controlled decision sequence in Table 5b distinguishes the head from its training context. CE with a linear readout gives 95.44% Macro-F1; focal loss gives 95.24%. LDAM-Focal with the linear readout gives 81.59%, and adding square-root sampling gives 90.17%. Replacing that readout with the prototype head gives 95.78% Macro-F1 and 96.07% Macro-Sen. Relative to A3, these are increases of 5.61 and 10.97 points, respectively; F-class F1 increases by 4.15 points.

The A3–A4 comparison measures the prototype configuration within the retained margin-loss and sampling scheme. It does not represent a 5.61-point gain over ordinary cross-entropy: the full-model difference from A0 is 0.34 points in Macro-F1 and 1.53 points in F-class F1. The A3–A4 Macro-F1 difference has an unadjusted paired interval of $[0.75, 10.48]$ points, while its Holm-adjusted t-test p value is 0.459 across the 15 decision-study contrasts. The results support a conditional contribution of prototype decisions, rather than uniform improvement at every step of the training sequence.

5.5.3. Earlier fusion and component comparisons

The original repeated-run study associates prototype replacement with a 2.64-point Macro-F1 decrease and Transformer removal with a 0.35-point decrease (Supplementary Table S7). In the separate single-run fusion study, Transformer fusion gives 96.38% Macro-F1, compared with 95.74% for weighted averaging, 95.54% for concatenation, and 95.15% for cross-attention (Supplementary Table S10). The earlier CNN, BiLSTM, TCN, Transformer, and CNN–Transformer comparisons remain in Supplementary Table S11. These series provide additional component context without pooling their reference scores with the new controlled study.

5.6. Source identity and output controls

5.6.1. Lead and view identity controls

The earlier 26-token PTB-XL source-tagging study evaluates the lead and view identifiers individually and jointly, using the same encoder and training configuration. Disabling lead tags removes both the per-lead additions and the global tag on the constant auxiliary token. Disabling view tags also removes the dedicated CLS view tag. The allocated embedding tables remain in the implementation; their additions are bypassed. This tests the source-tagging pathways rather than a reduction in the allocated parameter budget.

Removing lead tags lowers mean Macro-F1 from 72.70% to 70.78%, Micro-F1 by 2.22 percentage points, and subset accuracy by 4.00 points (Table 6). Macro-F1 falls in all three paired seeds. Removing both tag types yields 70.67% Macro-F1. Removing view tags alone instead raises Macro-F1 by 0.18 points but lowers Micro-F1 and subset accuracy by 0.42 and 0.65 points. Lead tagging has the clearer observed association with performance, while the effect of view tags varies by metric. Supplementary Tables S16–S18 report all runs and the 18 paired contrasts; Holm-adjusted p values range from 0.052 to 1.000.

Table 3: Task-level results with partitions and configurations kept separate (%). PTB5 denotes five-label diagnosis. F-F1 applies only to MIT5/MIT9; dashes denote metrics not applicable or not included here.

Task / series	Runs	Acc	Macro-AUC	Macro-F1	F-F1
MIT5, primary (τ = 1.0)	3 seeds	99.28 ± 0.04	–	95.55 ± 0.17	86.97 ± 0.96
MIT5, selected (τ = 1.5)	Seed 42	99.33	–	96.26	90.68
MIT5, follow-up shared reference	3 seeds	99.28 ± 0.02	–	95.44 ± 0.21	86.61 ± 1.89
MIT9	Seed 42	98.96	–	94.62	87.90
PTB-XL binary	Seed 42	86.30	–	86.12	–
PTB5 custom, 26-token single run	Seed 42	–	91.84	72.59	–
PTB5 custom, 26-token reference	3 seeds	–	91.94 ± 0.15	72.70 ± 0.61	–
PTB5 official folds, 25-token adopted	3 seeds	–	90.06 ± 0.18	70.72 ± 0.50	–

Table 4: Matched-split reimplementations trained from scratch (three seeds; mean±sample SD, %).

Model	Acc / AUC	Macro-F1	Sen / Micro-F1	Minority F1
(a) MIT5: accuracy, Macro-F1, Macro-Sen, and F-class F1				
SAMVT-Net	99.31 ± 0.08	95.78 ± 0.71	96.07 ± 0.63	87.97 ± 2.62
ECG-Mamba	98.33 ± 0.14	91.15 ± 0.32	86.81 ± 0.45	79.59 ± 1.42
ProtoECGNet	98.24 ± 0.21	91.03 ± 0.74	86.25 ± 0.89	79.80 ± 1.74
Swin-ECG	98.50 ± 0.14	92.49 ± 0.50	90.24 ± 0.51	83.18 ± 1.22
ECG-BERT	98.81 ± 0.14	92.97 ± 0.83	89.30 ± 1.77	80.74 ± 1.67
(b) PTB5 custom split; 26-token SAMVT reference				
SAMVT-Net	91.53 ± 0.40	72.40 ± 0.20	76.71 ± 0.31	48.98 ± 1.66
ECG-Mamba	90.69 ± 0.19	70.38 ± 0.23	74.68 ± 0.12	46.83 ± 1.05
ProtoECGNet	91.12 ± 0.50	71.50 ± 0.59	75.88 ± 1.12	48.18 ± 1.51
Swin-ECG	90.54 ± 0.23	70.34 ± 0.32	74.96 ± 0.60	45.65 ± 2.69
ECG-BERT	90.86 ± 0.22	70.38 ± 0.39	74.95 ± 0.69	44.54 ± 1.08

All rows use seeds 7, 21, and 42 and their fixed task partitions. The PTB5 panel uses the earlier 26-token configuration on the custom split. Baselines are architecture-level reimplementations: selective state-space (ECG-Mamba), shifted-window 1-D attention (Swin-ECG), CNN tokens with a BERT-style encoder (ECG-BERT), and a compact case/prototype branch (ProtoECGNet). The names identify the architecture families; these scores are not the source papers' original or pretrained-checkpoint results.

Table 5: Controlled MIT5 view and decision comparisons (three seeds; mean±sample SD, %).

Configuration	Macro-F1	Macro-Sen	F-class F1
(a) Physiological views			
Morphology (M)	94.82 ± 1.29	94.95 ± 0.63	86.13 ± 4.70
M + spectrum (S)	94.93 ± 0.45	95.15 ± 0.68	86.08 ± 2.43
M + rhythm (R)	95.66 ± 0.19	96.05 ± 0.32	87.22 ± 1.22
M + S + R	95.78 ± 0.71	96.07 ± 0.63	87.97 ± 2.62
(b) Training objective and decision head			
A0: CE + linear	95.44 ± 0.35	94.46 ± 0.59	86.44 ± 1.66
A1: Focal + linear	95.24 ± 0.71	94.38 ± 0.80	85.88 ± 2.64
A2: LDAM-Focal + linear	81.59 ± 3.76	74.43 ± 3.87	61.58 ± 19.39
A3: A2 + square-root sampler	90.17 ± 1.35	85.10 ± 2.04	83.82 ± 3.36
A4: A3 + prototype head	95.78 ± 0.71	96.07 ± 0.63	87.97 ± 2.62

M+S+R and A4 share the same full-model result; they are not independent replications. CE denotes cross-entropy. Adjacent A rows change the reported training or classifier configuration; the A3–A4 contrast is conditional on LDAM-Focal and square-root sampling.

5.6.2. Dependence on the fusion backbone

The lead-identity comparison is repeated with CNN, TCN, and Transformer fusion using seeds 7, 21, and 42. Transformer Macro-F1 changes from 70.80% without lead identification to 73.17% with it, a 2.37-point difference. The corresponding changes are +0.51 points for CNN and −0.37 points for TCN. Table 6 keeps all three backbones, so the direction can be compared directly.

Explicit lead names provide the clearest observed gain in attention-based fusion. A plausible explanation is that attention over shared-encoder tokens otherwise lacks a direct lead identifier, whereas convolutional fusion can also use a fixed token layout. This is an interpretation of the architecture and results, not a measurement of attention's internal mechanism. The source-backbone comparison therefore supports a backbone-dependent effect rather than universal improvement. Full metrics and the 15 paired contrasts are retained in the Supplementary Material.

5.6.3. Label-specific effect of output interaction

The matched PTB-XL five-label control replaces the complete interacting-label output formulation with an independent affine readout while retaining the same upstream architecture, data partition, optimization protocol, and threshold-selection procedure. Both variants are trained end to end for seeds 7, 21, and 42. Mean Macro-F1 is 72.41 ± 0.48% with interaction and 72.50 ± 0.83% with the linear readout; the paired difference changes direction across seeds. Mean HYP sensitivity is 57.62% and 53.41%, respectively, whereas HYP-F1 is 48.91% and 49.07% (Table 6). The 4.21-point sensitivity difference has a paired 95% confidence interval of [−13.65, 22.08] points and an unadjusted t-test p value of 0.417. All six Holm-adjusted t-test p values equal 1.000. The observed recall shift therefore describes this training series, rather than an established aggregate advantage. This control compares complete readouts, including their different parameterizations; it does not isolate attention from head capacity. Supplementary Table S24 reports all metrics and paired statistics.

5.7. Feature visualization and quantitative separation

The new view-combination reports provide a second, across-model perspective on representation quality. Mean silhouette is 0.707 for morphology alone, 0.737 for M+S, 0.753 for M+R, and 0.741 for the full model. Mean within-class distance changes from 2.921 for morphology alone to 2.535 for M+R and 2.688 for the full model; mean inter-centroid distance changes from 15.718 to 16.218 and 16.045, respectively. Thus the full model improves these descriptors over morphology alone, while M+R has the strongest reported cluster separation. This is consistent with the predictive view comparison and does not establish an extra clustering benefit from spectrum. The supplementary view-geometry table retains all six supplied metrics. These report-level descriptors are kept separate from the explicitly standardized six-stage analysis below because their extraction and normalization details are not included in the new summaries.

The representation analysis uses t-SNE [40] alongside quantitative scores computed from saved features. The MIT-BIH Arrhythmia Database five-class run with $\tau = 1.0$ and seed 42 supplies 10,945 test observations at six representation stages; its saved predictions reproduce 95.64% Macro-F1. Morphology, spectrum, rhythm, the mean of source-tagged tokens before fusion, and the fused CLS feature each have 128 dimensions; classifier logits have five. The token mean is extracted from the same trained model before its Transformer stage.

For high-dimensional analysis, each feature coordinate is standardized over all test observations, solely to define a common distance convention for this descriptive analysis. Ground-truth classes define the groups. The silhouette coefficient [41] compares a sample's mean distance to its own class with its nearest other class; its class-macro average gives each class equal weight. The Davies–Bouldin index [42] summarizes within-class scatter relative to between-class centroid distances. Higher silhouette and lower Davies–Bouldin values indicate stronger separation under this standardized Euclidean geometry.

Table 7 reports a class-macro silhouette of 0.178 before fusion and 0.697 for the fused feature, while the Davies–Bouldin index decreases from 2.209 to 0.491. Dimension-normalized within-class dispersion decreases from 0.793 for the tagged token mean to 0.382 for the fused feature, while between-class centroid dispersion increases from 0.441 to 2.186. These equal-class-weighted squared-distance summaries complement silhouette and centroid-scatter scores (Supplementary Table S25). The spectral token has a macro silhouette of −0.069, indicating overlapping classes under this distance criterion. The scores describe successive representations learned jointly; they do not isolate a causal effect of source tags or establish statistical independence between views. The five-dimensional logits are the task-aligned readout, whereas the preceding stages represent 128-dimensional features.

Figure 9 projects the same 1,858 observations at every stage: at most 500 per class, sampled without replacement with seed 42 (N/S/V/F/Q: 500/278/500/80/500). Standardized features are reduced to 50 principal components when needed, then projected with perplexity 30, PCA initialization, seed 42, and 2,000 iterations. Trustworthiness at ten neighbors [43], evaluated against the standardized features before PCA, ranges from 0.993 to 0.999. This quantifies local neighborhood preservation through the dimensionality-reduction pipeline. The two-dimensional silhouette values are reported separately because the projection uses a capped subset and changes the geometry. Supplementary Figures S2–S3 provide the earlier qualitative raw-input and ablation views. The quantitative analysis here uses the matched saved representations and fixed projection settings.

5.8. Model size and inference cost

The profiled MIT5 and earlier 26-token PTB5 configurations contain 457,061 and 671,333 parameters, respectively, requiring 1.74 and 2.56 MiB for FP32 weights. Table 8 reports their CPU forward latency beside the matched implementations. On PTB5, SAMVT-Net has 0.0162 G profiler-counted FLOPs and 0.0081 G approximate MACs. The counted arithmetic is lower than that of the Mamba, BERT, and Swin implementations, but above that of the compact prototype baseline. The profiler excludes FFT and some custom or elementwise operations, so these are partial operator counts rather than exact total costs.

Arithmetic count and wall-clock time rank the implementations differently. SAMVT-Net requires 24.32 ± 0.20 ms on PTB5 in the one-thread Xeon test, compared with 3.30–8.60 ms for the other implementations. Small parameter and counted-operation budgets therefore do not imply the fastest CPU execution. The current branch and token-processing structure is a possible source of overhead; the available profile does not attribute time to individual operators.

The separately measured RTX 4090 profile uses PyTorch 2.8.0, CUDA 12.8, Python 3.12.3, and Linux 5.15, with 30 warm-up and 100 synchronized timed iterations in FP32 evaluation mode. PTB5 mean forward latency is 3.27 ms at batch one and 6.76 ms at batch 32; the latter corresponds to 4,730 records/s. Peak allocated GPU memory is 34.20 and 764.66 MiB, respectively. MIT5 yields 3.29 and 3.59 ms at the same batch sizes. The full GPU profile remains in Supplementary Table S20. These CPU/GPU profiles describe the earlier 26-token PTB5 implementation, not a new latency or memory measurement

Table 6: PTB-XL source-identity and output-head controls for the earlier custom-split configuration (mean±sample SD across three seeds, %). Panels use separate 26-token reference series. AUC is the area under the ROC curve; Subset-Acc is exact agreement across all labels.

Configuration	Macro-AUC	Macro-F1	Micro-F1	Subset-Acc	HYP-F1
Full	91.94 ± 0.15	72.70 ± 0.61	77.18 ± 0.30	60.18 ± 1.46	49.43 ± 2.11
w/o Lead tags	90.63 ± 0.34	70.78 ± 0.23	74.96 ± 0.18	56.18 ± 1.11	48.68 ± 2.22
w/o View tags	91.76 ± 0.31	72.88 ± 0.25	76.76 ± 0.34	59.53 ± 1.54	50.48 ± 0.74
w/o Both tags	90.45 ± 0.18	70.67 ± 0.17	75.25 ± 0.10	57.72 ± 0.28	46.61 ± 1.71
Lead identity across backbones (separate reference series)					
CNN w/o lead ID	92.18 ± 0.12	72.97 ± 0.36	76.81 ± 0.36	57.86 ± 0.38	51.38 ± 1.58
CNN + lead ID	92.02 ± 0.19	73.48 ± 0.17	76.99 ± 0.23	58.25 ± 1.17	53.24 ± 1.18
TCN w/o lead ID	92.22 ± 0.05	73.32 ± 0.42	77.11 ± 0.46	58.95 ± 0.94	51.87 ± 0.31
TCN + lead ID	92.16 ± 0.51	72.95 ± 0.39	77.02 ± 0.18	58.36 ± 0.73	50.55 ± 1.50
Transformer w/o lead ID	91.02 ± 0.20	70.80 ± 0.51	74.99 ± 0.29	56.16 ± 0.55	47.83 ± 3.17
Transformer + lead ID	92.00 ± 0.11	73.17 ± 0.82	77.21 ± 0.65	59.73 ± 0.93	51.49 ± 1.52
Output-head series (separate reference)			Macro-F1	HYP-F1	HYP sensitivity
Interacting label tokens			72.41 ± 0.48	48.91 ± 0.98	57.62 ± 3.04
Independent linear readout			72.50 ± 0.83	49.07 ± 2.15	53.41 ± 5.27

Table 7: Representation geometry for the saved MIT-BIH Arrhythmia Database five-class run (τ = 1.0, seed 42)

Representation	Dimensions	$S^{\mathrm{HD}}_{\mathrm{macro}} \uparrow$	$\mathrm{DB}^{\mathrm{HD}} \downarrow$	$S^{\mathrm{2D}}_{\mathrm{macro}} \uparrow$	$T(10) \uparrow$
Morphology token	128	0.141	2.610	0.220	0.993
Spectral token	128	-0.069	2.999	-0.015	0.998
Rhythm token	128	0.086	2.321	0.130	0.999
Tagged token mean	128	0.178	2.209	0.251	0.994
Fused CLS feature	128	0.697	0.491	0.638	0.994
Classifier logits	5	0.739	0.354	0.629	0.996

HD scores use all 10,945 standardized test features without PCA. 2D scores and trustworthiness use the same capped subset of 1,858 observations in every panel. S_{macro} is the class-macro silhouette; DB is the Davies–Bouldin index. Arrows indicate the favorable direction. Scores provide descriptive geometry for one saved model.

Table 8: Model size, profiler-counted arithmetic, and CPU forward latency for the matched implementations.

Model	Params (M)	FLOPs (G)	MACs (G)	CPU ms
MIT5				
ECG-BERT	0.918	0.0781	0.0391	2.34 ± 0.02
ECG-Mamba	0.152	0.0098	0.0049	2.81 ± 0.09
ProtoECGNet	0.269	0.0007	0.0003	1.25 ± 0.00
SAMVT-Net	0.457	0.0022	0.0011	2.70 ± 0.05
Swin-ECG	0.449	0.0728	0.0364	1.84 ± 0.05
PTB5				
ECG-BERT	0.924	0.3923	0.1961	6.71 ± 0.04
ECG-Mamba	0.158	0.0371	0.0186	8.60 ± 0.11
ProtoECGNet	0.275	0.0026	0.0013	3.30 ± 0.02
SAMVT-Net	0.671	0.0162	0.0081	24.32 ± 0.20
Swin-ECG	0.459	0.2330	0.1165	3.60 ± 0.04

Intel Xeon Platinum 8352V (2.10 GHz), PyTorch 2.8.0, Python 3.12.3, one CPU thread, batch one, 20 warm-up and 100 timed iterations. Latency is mean±sample SD across three seed-specific model profiles. FLOPs use the operator formulas in torch.profiler; FFT and some custom/elementwise operators are uncounted. MACs approximate one-half of counted FLOPs. PTB5 uses the earlier 26-token SAMVT implementation; no 25-token timing is inferred. Data loading is outside the timing region.

of the adopted 25-token sequence. Acquisition, preprocessing, loading, and host–device transfers are outside the model timing region; the GPU measurements and the new CPU comparison are separate profiling series.

5.9. Input perturbations and available lead information

5.9.1. Graded lead exclusion

The custom-split PTB5 lead-loss study evaluates the earlier 26-token matched-series models at every severity from one to eleven unavailable leads. The supplied summary has three training seeds and three mask patterns per seed at each nonzero severity. These nested mask evaluations are not independent model-training runs. The earlier source-series experiment uses ten masks per seed at selected severities and is retained separately in Supplementary Tables S19 and S26.

Macro-F1 changes from 72.40% with all leads to 71.78%, 71.22%, and 70.14% after one, two, and three lead exclusions

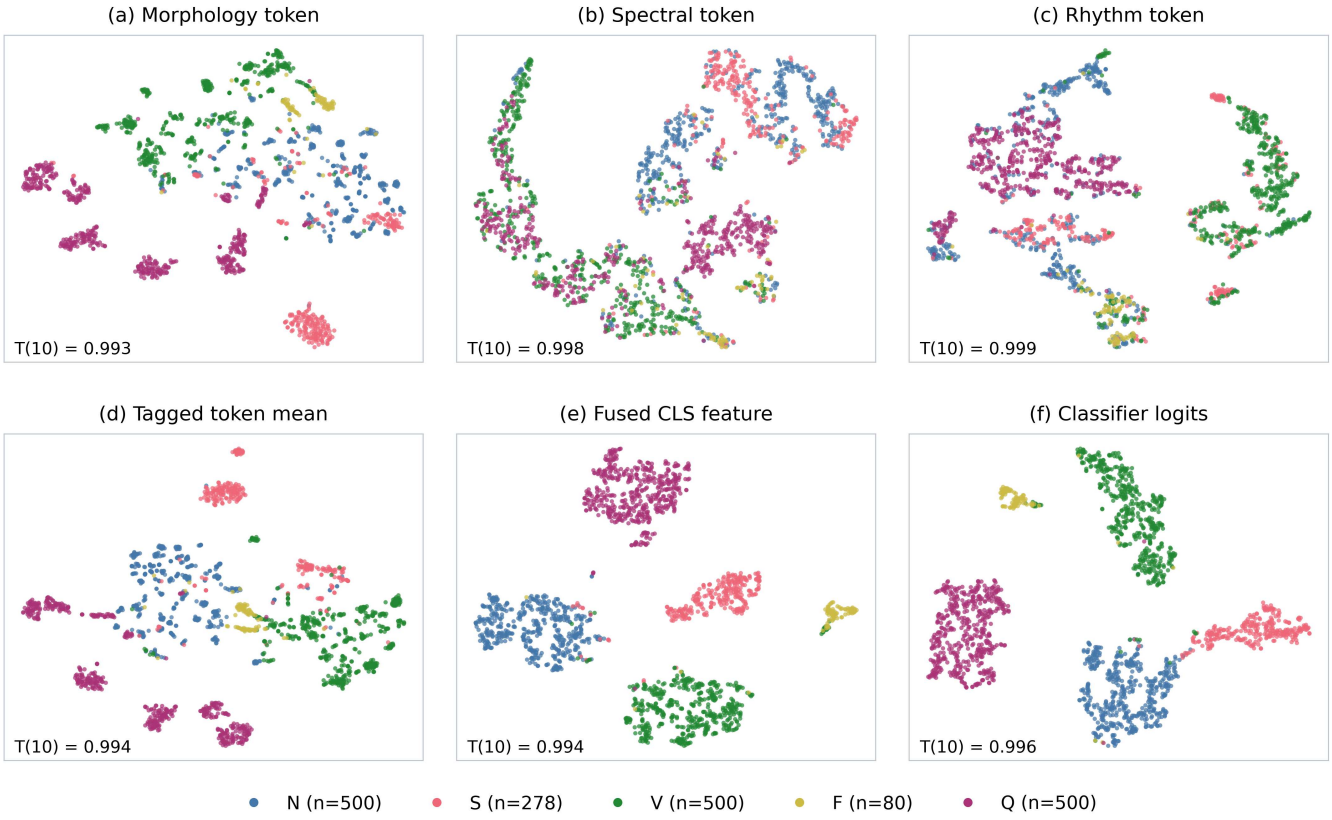


Figure 9: Matched t-SNE views of six saved representation stages from the $\tau = 1.0$, seed-42 run. Every panel uses the same 1,858 test observations sampled without replacement, capped at 500 per class with seed 42. Features are coordinate-standardized, reduced to 50 principal components when applicable, then projected with perplexity 30, PCA initialization, seed 42, and 2,000 iterations. Axes are independently fitted projection coordinates. $T(10)$ denotes trustworthiness at ten neighbors; class counts are shown in the shared legend. Table 7 separates high-dimensional class separation from projection quality.

(Fig. 10). The losses are 0.62, 1.18, and 2.26 points. At six exclusions, Macro-F1 is 64.04%; with one retained lead, it is 38.85%. The curve is non-monotonic at several adjacent severities because it also depends on which leads remain; the three sampled masks do not enumerate all possible lead combinations. These results characterize tolerance to mild synthetic lead loss and the larger changes under severe input reduction. They evaluate the complete model, rather than isolating source tagging as the cause of robustness.

The earlier individual-lead study resolves lead identity at fixed thresholds: deleting aVR, aVL, or V4 gives mean Macro-F1 losses of 0.11, 0.31, and 0.19 points, respectively, while deleting lead I gives a 1.43-point loss. Its full reference is 72.70%, and its ten-mask random single-lead average is 71.99%. Those values describe the earlier models and are not substituted into the new 72.40% reference curve.

5.9.2. Response to the reported SNR perturbations

Table 9 separates the two tasks at 10, 15, 20, and 25 dB; the PTB5 series uses the earlier 26-token custom-split reference. PTB5 Macro-F1 changes from 72.40% on clean inputs to 72.17% at 25 dB and 71.62% at 20 dB. At 15 and 10 dB it is 69.63% and 61.16%, respectively. The moderate-noise conditions therefore preserve the aggregate PTB5 score more closely than the stronger perturbations.

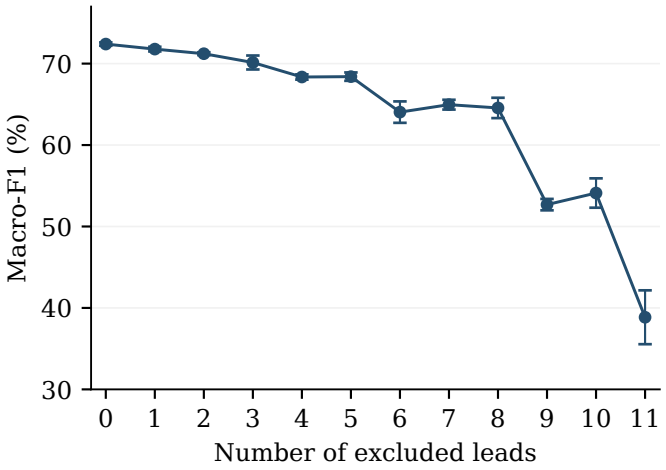


Figure 10: Complete PTB5 lead-exclusion gradient for the earlier 26-token custom-split matched series. Points are the reported Macro-F1 means; error bars show the supplied across-seed SD after mask aggregation. Nonzero severities use three mask patterns per seed and three training seeds. Clean uses all twelve leads. The full metrics appear in the Supplementary Material.

MIT5 responds differently: Macro-F1 is 82.31% at 25 dB and 56.73% at 20 dB, compared with 95.78% on clean inputs. At 15 and 10 dB it is 35.04% and 30.35%, with zero F-class F1 in both conditions. Thus tolerance observed for record-level PTB5 does not extend to the reported heartbeat-level corruption test. Differences in duration, lead redundancy, normalization, and

signal content could all affect this response; the summary-level evidence does not isolate their individual roles.

5.10. Causal timing variants

Two additional MIT-BIH Arrhythmia Database five-class variants replace the RR descriptor with intervals available at the current R peak. One retains the original waveform preprocessing; the other also applies a one-pass causal band-pass filter. The three-seed Macro-F1 values are 96.06 ± 0.39% and 95.84 ± 0.52%, respectively; the corresponding F-class F1 values are 88.57 ± 1.26% and 86.76 ± 2.49% (Supplementary Table S23). These separately trained variants are reported as descriptive pipeline evaluations. Both retain 160 post-R samples in their beat windows, corresponding to a 0.44-s decision delay at 360 Hz.

5.11. Configuration sensitivity

The new three-seed study changes kernel set, spectral-band count, prototype count, dropout, and learning rate one factor at a time. Across the tested structural settings, mean Macro-F1 spans 95.19–95.96%. Band counts of three to five give 95.78–95.95%, and one to four prototypes give 95.68–95.96%. Dropout values 0.10, 0.20, and 0.30 give 95.24%, 95.78%, and 95.66%, respectively. These ranges describe local sensitivity around the reference configuration, not invariance to arbitrary parameter changes.

Learning rate has the larger observed effect: 5×10^{-4}, 10^{-3}, and 2×10^{-3} give 95.31%, 95.78%, and 94.69% Macro-F1. At the highest rate, F-class F1 is 84.48%, compared with 87.97% at the reference rate. The complete new sweeps and paired contrasts are provided in the Supplementary Material. The earlier single-run sweeps remain in Supplementary Tables S12–S13. These are post hoc test responses of the reported configurations; the manuscript retains the reference setting rather than selecting a new setting from test performance.

5.12. PTB-XL initialization for MIT5

Table 10 reports the PTB-XL-to-MIT5 initialization direction, followed by target-task training. Mean Macro-F1 is 95.47% after PTB-XL initialization and 95.44% from scratch; F-class F1 is 87.02% and 86.61%, respectively. The 0.03-point Macro-F1 difference describes similar target-task performance in this three-seed comparison. It does not establish a transfer advantage or evaluate a frozen model on a new database.

5.13. Official-fold PTB-XL evaluation and auxiliary-token control

The official patient-separated folds provide an additional evaluation under a common protocol for CNN, TCN, Transformer, and two SAMVT-Net configurations. All five use the same 100-Hz twelve-lead records, preprocessing, training budget, validation procedure, and seeds 7, 21, and 42. The two SAMVT-Net variants share the same morphology/spectral encoders, source embeddings, fusion layers, label head, and allocated parameter structure. The only architectural switch is inclusion of the zero-input auxiliary vector in the fusion sequence: 26 tokens with it and 25 without it.

The 25-token configuration achieves 90.06 ± 0.18% Macro-AUC, 70.72 ± 0.50% Macro-F1, 75.03 ± 0.51% Micro-F1, and 57.34 ± 1.47% subset accuracy (Table 11). These are the highest mean aggregate scores among the five matched configurations. The margins over CNN are 0.01, 0.13, 0.32, and 2.27 percentage points, respectively. None is statistically established in the three-seed paired comparison: all four unadjusted 95% intervals span zero, with paired t-test p values of 0.950, 0.769, 0.369, and 0.107. The mean ranking does not establish superiority over CNN.

Removing the auxiliary token changes mean Macro-F1 from 70.34% to 70.72%, Micro-F1 from 74.53% to 75.03%, and subset accuracy from 55.05% to 57.34%. The paired Macro-F1 difference is +0.38 points, with an unadjusted 95% interval of $[-0.59, 1.35]$. For subset accuracy, the difference is +2.29 points with interval $[1.47, 3.10]$ and Holm-adjusted paired-t p value 0.041 across the six auxiliary-token metrics. Its exact two-sided Wilcoxon p value is 0.25, reflecting the limited resolution of three seed pairs. These results favor omission in the evaluated configuration without establishing an improvement in every metric or under every statistical test.

The per-label changes are also mixed. Without the token, mean F1 increases for NORM, MI, STTC, and HYP by 0.11, 1.02, 0.45, and 0.95 points, while CD-F1 decreases by 0.63 points. HYP sensitivity falls from 57.63% to 54.83%, whereas its precision rises from 39.81% to 42.78% and specificity from 87.94% to 89.84%; HYP-F1 increases from 47.06% to 48.01%. Removing the token therefore changes the precision–recall balance; its label-wise effects are not uniformly positive. Supplementary Tables S40–S42 give the individual runs, complete per-label metrics, and paired statistics.

The fixed zero input carries no record-specific timing information. Since omitting its token preserves or improves all four mean aggregate scores, we adopt the 25-token record-level fusion sequence. The official-fold control supports this simpler fusion sequence. The implementation retains and computes the auxiliary branch in both variants, excluding its output from attention in the 25-token variant; both therefore allocate 671,333 parameters. Parameter pruning and a new runtime gain are not inferred. Earlier custom-split results remain attached to their original 26-token configurations.

6. Discussion

6.1. What the physiological views contribute

The multi-view evidence is most informative when the branches are assigned different roles according to the task. The adopted PTB-XL configuration uses morphology and spectrum, whereas measured beat timing is supplied for heartbeat classification. Morphology describes the waveform within a beat window, whereas RR intervals describe its temporal context. In the new controlled study, adding rhythm to morphology gives a mean Macro-F1 change of 0.83 points, while adding spectrum gives

Table 9: Earlier configurations: Reported SNR stress-test responses (three seeds; mean±sample SD, %). Clean results are the matched-series references.

Condition	MIT5 Macro-F1	MIT5 F-F1	PTB5 Macro-F1	PTB5 HYP-F1
Clean	95.78 ± 0.71	87.97 ± 2.62	72.40 ± 0.20	48.98 ± 1.66
25 dB	82.31 ± 9.08	37.89 ± 33.71	72.17 ± 0.23	48.93 ± 1.85
20 dB	56.73 ± 15.43	0.74 ± 1.28	71.62 ± 0.36	47.31 ± 3.56
15 dB	35.04 ± 6.80	0.00 ± 0.00	69.63 ± 0.46	41.54 ± 4.93
10 dB	30.35 ± 4.90	0.00 ± 0.00	61.16 ± 1.58	19.42 ± 13.67

SNR denotes signal-to-noise ratio. The supplied result summaries specify SNR levels but do not identify the noise distribution or its insertion point in preprocessing; interpretation is confined to this reported corruption test.

Table 10: PTB-XL initialization followed by MIT5 target fine-tuning (three seeds; mean±sample SD, %).

Target / initialization	Macro-F1	F-class F1
MIT5 / scratch	95.44 ± 0.21	86.61 ± 1.89
MIT5 / PTB-XL	95.47 ± 0.36	87.02 ± 0.83

Both rows are evaluated after MIT5 target training and belong to the earlier transfer series.

0.11 points. With rhythm present, spectrum adds 0.12 points. The learned convolutional features and spectral summaries originate from the same waveform, so some overlap is plausible. The data support morphology–rhythm complementarity as the central representation result and spectrum as an auxiliary view.

The feature analyses are consistent with this interpretation without proving a unique mechanism. The new reports give the strongest silhouette and centroid-separation ratio for M+R, even though the full model has the highest mean predictive Macro-F1. Cluster geometry and classification therefore need not rank configurations identically. The earlier six-stage analysis also shows greater class separation after fusion than in the mean of the source-tagged tokens. Because those stages are jointly trained, their geometry describes the resulting representation rather than the isolated causal effect of one module.

Clinical interpretability requires a different kind of evidence. Gliner et al. compared gradient-based ECG image explanations with electrophysiologist interpretations [44]. VAE-SCAN links learned ECG representations to clinical factors and examines the associated performance–explainability trade-off [45]. The present cluster descriptors assess separation in feature space; they neither identify a diagnostic waveform feature nor validate an explanation against clinical judgment. Physiologically named input branches should therefore not be equated with an intrinsically interpretable classifier.

6.2. Why lead identity matters differently across backbones

The shared per-lead encoder maps each lead into the same feature space. Attention can compare token content, but the feature format alone does not specify whether a token came from lead I, aVR, or V4. Adding a lead embedding supplies that information directly to the query, key, and value computations. This provides a concrete rationale for the source-aware interface.

Two complementary comparison settings support this design choice within their own references. In the earlier source study, bypassing lead tags lowers Macro-F1 by 1.92 points and subset accuracy by 4.00 points. In the new backbone study, lead identification has a 2.37-point mean Macro-F1 effect with Transformer fusion, compared with +0.51 for CNN and −0.37 for TCN. The evidence therefore favors a backbone-dependent benefit, with the clearest numerical effect in attention-based fusion. It does not support a claim that every backbone improves. Removing the auxiliary token leaves the lead-identity pathway intact, but these component-effect estimates retain their earlier 26-token references; the official-fold control measures the auxiliary token rather than re-estimating a lead-embedding ablation.

View identity has a different empirical role. Separate morphology and spectral encoders can already generate distinguishable feature distributions, and removing view tags produces mixed changes across metrics. This suggests possible redundancy, although the experiments do not directly quantify what information the embeddings encode. The source-aware contribution is consequently centered on lead identity and its interaction with the fusion design, rather than treating all identifiers as equally necessary.

6.3. Prototype decisions and the scope of task adaptation

The decision study separates the effectiveness of a configuration from the originality of its ingredients. Focal loss, class-dependent margins, and balanced sampling are established training tools. Their combination with a linear head does not yield monotonic improvement in the present experiment. The prototype configuration has a clear numerical effect relative to A3, especially in Macro-Sen, but its Macro-F1 advantage over CE-linear is 0.34 points. This anchors the task-adaptation claim to the chosen decision geometry and training context rather than a general superiority of the complete imbalance scheme.

The original prototype comparison also changes the precision–recall balance: the full model gives 95.54% Macro-Sen and 95.58% Macro-PPV, while prototype replacement gives 89.59% and 97.12%, respectively. A normalized similarity head with multiple class directions can represent several regions for one class. The experiments are consistent with a useful decision formulation for the evaluated single-label tasks; they do not identify prototype count or any one regularizer as the sole cause.

The auxiliary-token control shows why task adaptation should allow the input views to differ. Measured RR intervals add timing context to a short heartbeat window; a fixed zero-input vector adds no record-specific timing to PTB-XL. The official-fold results support removing that vector from record-level fusion, while the preserved precision–recall trade-off prevents a claim of universal label-wise improvement. The adopted configuration uses the available physiological inputs; the earlier token set remains an explicit experimental reference.

Table 11: Official-fold PTB-XL patient-separated evaluation (three seeds; mean±sample SD, %). Bold indicates the highest mean in each column.

Configuration	Macro-AUC	Macro-F1	Micro-F1	Subset-Acc
CNN	90.05 ± 0.18	70.59 ± 0.27	74.71 ± 0.30	55.07 ± 0.86
TCN	89.46 ± 0.34	69.21 ± 0.27	73.46 ± 0.24	53.04 ± 1.94
Transformer	88.26 ± 0.36	67.78 ± 0.47	72.06 ± 0.84	53.94 ± 1.49
SAMVT-Net + constant (26 tokens)	89.98 ± 0.23	70.34 ± 0.12	74.53 ± 0.24	55.05 ± 1.74
SAMVT-Net adopted (25 tokens)	**90.06 ± 0.18**	**70.72 ± 0.50**	**75.03 ± 0.51**	**57.34 ± 1.47**

Training: folds 1–8; validation: fold 9; test: fold 10. Seeds: 7, 21, 42. All models share preprocessing, training budget, loss, checkpoint selection, and validation-only threshold tuning. Parameter counts are 270,853 (CNN), 279,173 (TCN), 445,061 (Transformer), and 671,333 for each SAMVT variant. The auxiliary modules remain allocated and evaluated in both SAMVT variants; only fusion-sequence inclusion changes. Scores are separate from the earlier custom-split studies.

For multi-label diagnosis, sigmoid outputs represent co-occurring classes, whereas interaction between the label tokens is an additional choice. The output-head control gives a mean HYP sensitivity difference of +4.21 points but a Macro-F1 difference of −0.09 points. Its paired intervals span zero. Label interaction is therefore retained as an output formulation, with the aggregate evidence described separately from the label-specific operating point.

6.4. Recording conditions and execution cost

Lead masking and waveform corruption test different properties. Mild lead exclusion removes a small part of the spatial input while leaving other recorded leads intact. In the new PTB5 series, one to three exclusions cost 0.62–2.26 Macro-F1 points. Stronger lead reduction produces larger changes and a non-monotonic curve across sampled lead sets. The result supports tolerance to limited synthetic lead loss for the evaluated model, rather than arbitrary single-lead substitution. UniCardio addresses incomplete or degraded cardiovascular recordings through generative reconstruction and cross-modal translation [46]. Our lead-exclusion experiment measures prediction from the remaining input and does not reconstruct the missing signal. Comparing those strategies would require a separate control for reconstruction error and its effect on diagnosis.

The SNR results further separate record-level and heartbeat-level behavior. PTB5 retains Macro-F1 within 0.78 points of clean performance at 20–25 dB, whereas MIT5 changes substantially at those levels. This distinction matters because clean classification performance cannot establish noise tolerance. The reported perturbations characterize the two task pipelines and do not identify the physiological or preprocessing cause of their different responses. The starfish-inspired wearable system of Chen et al. combines device mechanics, signal compensation, and multimodal sensing to support cardiac monitoring during movement [47]. Such acquisition-level motion control is outside the present synthetic-noise evaluation, so these results cannot establish reliability during ambulatory use.

Execution cost also requires more than one measure. The compact parameter budget and profiler-counted arithmetic describe storage and a subset of operations. The one-thread CPU comparison shows that these quantities do not predict latency ordering. GPU batching yields 4,730 PTB5 records/s under the stated forward-only profile, while acquisition still requires the task-defined recording window. The reported timings measure forward execution on the tested hardware, excluding acquisition and preprocessing. The PTB5 perturbation and runtime measurements apply to the earlier 26-token configuration and do not provide a robustness or speed estimate for the adopted 25-token variant.

6.5. Relation to recent 2025–2026 ECG classifiers and evaluation scope

SAMVT-Net reuses an architecture across separately trained tasks. The PTB-XL-to-MIT5 experiment additionally tests initialization followed by target fine-tuning, with similar mean Macro-F1 to training from scratch. It is therefore the common token interface, rather than demonstrated weight transfer, that links the task configurations.

To place the MIT5 result against recent literature, Table 12 follows the comparison style commonly used in ECG classification studies and reports a focused set of 2025–2026 MIT-BIH five-class methods. The selected studies all report headline accuracy below the SAMVT-Net three-seed result; where a macro-averaged F1 score is explicitly available, it is also below the SAMVT-Net value. Zheng et al. combine dual-scale STFT representations with an SE-enhanced multi-input CNN and report 99.13% accuracy and 94.46% Macro-F1 [48]. Lamba et al. use stationary-wavelet/Hilbert features with ACO-optimized FCN classification and report 99.10% accuracy for the AAMI five-class task [49]. In 2026, Adeleye and Elbattah report 98.79% accuracy and 92.93% Macro-F1 for DWT features with a Cascade Forest classifier, while Umarani et al. report 95.14% accuracy and 94.80% Macro-F1 using DWT and morphological descriptors with ABC-optimized XGBoost [50, 51].

Against this selected recent set, SAMVT-Net is numerically higher by 0.15 percentage points in accuracy and 1.09 points in Macro-F1 than the dual-scale SE-CNN of Zheng et al.; by 0.49 and 2.62 points, respectively, than the DWT–Cascade Forest result of Adeleye and Elbattah; and by 4.14 and 0.75 points, respectively, than the ABC-XGBoost result of Umarani et al. The ACoFCN result of Lamba et al. is 0.18 points lower in accuracy, while its publication does not provide a directly comparable macro-averaged F1 headline for this table. These differences are descriptive because the studies use different splits and training procedures. They nevertheless show that the SAMVT-Net MIT5 result is competitive not only in overall accuracy but also in class-balanced Macro-F1 among the recent methods selected here.

The numerical comparison is accompanied by a different representation trade-off. The selected image/time–frequency and

Table 12: Comparison with selected recent MIT-BIH five-class studies. Values are those reported by the cited publications; "–" indicates that a directly comparable macro-averaged F1 value was not explicitly reported. Cross-paper partitions, augmentation, and model-selection protocols differ, so the table provides literature context rather than a matched statistical ranking.

Method	Main representation / classifier	Reported MIT-BIH setting	Acc (%)	Macro-F1 (%)
Zheng et al. (2025) [48]	Dual-scale STFT + SE multi-input CNN	AAMI N/S/V/F/Q; 90/10 beat split with Mixup	99.13	94.46
Lamba et al. (2025) [49]	SWT–Hilbert features + ACO-FCN	AAMI N/S/V/F/Q; optimized five-class classifier	99.10	–
Adeleye and Elbattah (2026) [50]	DWT features + Cascade Forest	AAMI N/S/V/F/Q; 80/20 stratified split	98.79	92.93
Umarani et al. (2026) [51]	DWT + morphology + ABC-XGBoost	AAMI N/S/V/F/Q; five-class evaluation	95.14	94.80
SAMVT-Net (ours)	**Morphology + spectrum + measured RR; source-tagged Transformer**	**N/S/V/F/Q; fixed 7:2:1 beat split, three seeds**	**99.28±0.04**	**95.55±0.17**

engineered-feature methods rely on STFT maps, wavelet/Hilbert transforms, or explicit DWT/morphological feature banks. SAMVT-Net instead compresses the available signal into morphology, spectrum, and measured-RR tokens before source-aware fusion. The matched MIT5 implementation contains 0.457 million parameters, and the same token–fusion interface is retained for MIT9 and PTB-XL with task-dependent inputs and readouts. The advantage supported by the present experiments is therefore a combination of competitive recent-study performance, compact representation, explicit source identity, and component-level experimental traceability rather than a claim that unmatched literature protocols constitute a single leaderboard.

The official-fold PTB-XL study extends evaluation beyond the custom partition using patient-disjoint training, validation, and test subsets. The adopted configuration has the highest mean aggregate scores among the five matched configurations, with small margins over CNN and uncertainty that precludes a general superiority claim. It is a within-database patient-separated evaluation, not a cross-database or clinical validation. All predictive comparisons refer to their specified partitions in Section 3. Repeated seeds measure training variability on those partitions, and each experiment series retains its own reference. The 25-token configuration is adopted on the basis of the reported auxiliary control; its official-fold scores and the earlier 26-token custom-split scores are distinct observations, not a paired estimate of a split or architecture effect.

7. Conclusion

SAMVT-Net provides a compact physiological token interface for heartbeat and record-level ECG classification. Its central representation result is the complementary use of morphology and rhythm, while its central fusion result is the observed value of explicit lead identity, particularly with Transformer fusion. Spectral summaries provide an auxiliary view for heartbeat classification, and prototype decisions supply a task-specific readout whose contribution is separated from the effects of loss design and sampling. The official-fold auxiliary control further supports a record-level sequence built from morphology and spectral tokens with lead identity, omitting the zero-input token.

The experimental evidence combines three-seed MIT5 results, matched reimplementations, official patient-separated PTB-XL evaluation, backbone controls, quantitative feature descriptors, perturbation studies, and hardware profiling. On MIT5, SAMVT-Net achieves 99.28±0.04% accuracy and 95.55±0.17% Macro-F1 while retaining a compact physiological-token representation. The recent-study comparison in the Discussion places these values above the selected 2025–2026 five-class methods on the reported headline metrics, while the controlled experiments show where the gains arise inside the proposed architecture. Across tasks, the combination of compact physiological views, explicit source tagging, and task-adapted readouts provides a consistent and experimentally traceable design. Future work should focus on lead-availability-aware training, broader external validation, and more efficient execution of the physiological branches.

Data availability statement

MIT-BIH Arrhythmia Database and PTB-XL are available from PhysioNet at `https://dx.doi.org/10.13026/C2F305` and `https://dx.doi.org/10.13026/kfzx-aw45`. Reproducibility materials retain model and training code, historical five-label predictions and thresholds, saved MIT5 features, and follow-up run arguments and metric reports. The official-fold materials additionally include the five-configuration training implementation, split-validation report, fifteen seed-level metric histories, and validation-selected thresholds. Scripts reproduce summary arithmetic, paired official-fold comparisons, nested lead-mask aggregation, and feature-geometry analysis. Section 6.5 and the Supplementary Material describe the analysis supported by each artifact group.

CRediT authorship contribution statement

Runchuan Li: Writing – original draft, Writing – review & editing, Data curation, Methodology, Project administration.
Shangwen Zhang: Writing – original draft, Conceptualization, Methodology, Software, Validation, Resources.

Shengya Shen: Writing – original draft, Writing – review & editing, Project administration, Methodology, Conceptualization, Resources.
Junjie Bian: Writing – review & editing, Funding acquisition, Methodology, Software.
Jingbin Zhang: Writing – review & editing, Funding acquisition, Methodology, Validation.
Xinchao Han: Writing – review & editing, Data curation, Supervision.
Hua-Nong Ting: Writing – review & editing, Data curation, Supervision.
All authors have read and approved the final manuscript.

Acknowledgments

This work was supported by the Key Scientific Research Project of Colleges and Universities in Henan Province (No. 25A520039), the Natural Science Foundation of Zhongyuan University of Technology (No. K2025YB011), the Postgraduate Research Innovation Project of Zhongyuan University of Technology (No. YKY2026ZK23), and the Science and Technology Guidance Project of China National Textile and Apparel Council (No. 2026030).

References

[1] G. B. Moody, R. G. Mark, The impact of the MIT-BIH arrhythmia database, IEEE Engineering in Medicine and Biology Magazine 20 (3) (2001) 45–50. doi:10.1109/51.932724.

[2] P. Wagner, N. Strodthoff, R.-D. Bousseljot, et al., PTB-XL, a large publicly available electrocardiography dataset, Scientific Data 7 (1) (2020) 154. doi:10.1038/s41597-020-0495-6.

[3] T. J. Poterucha, L. Jing, R. P. Ricart, M. Adjei-Mosi, J. Finer, D. Hartzel, et al., Detecting structural heart disease from electrocardiograms using AI, Nature 644 (2025) 221–230. doi:10.1038/s41586-025-09227-0.

[4] Z. Obermeyer, A. Schubert, J. Ross, S. Mullainathan, M. Lingman, An ECG biomarker for sudden cardiac death discovered with deep learning, Nature 655 (2026) 210–218. doi:10.1038/s41586-026-10674-6.

[5] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, A. Y. Ng, Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network, Nature Medicine 25 (1) (2019) 65–69. doi:10.1038/s41591-018-0268-3.

[6] M. Zhong, F. Li, W. Chen, Automatic arrhythmia detection with multi-lead ECG signals based on heterogeneous graph attention networks, Mathematical Biosciences and Engineering 19 (12) (2022) 12448–12471. doi:10.3934/mbe.2022581.

[7] Y. Na, M. Park, Y. Tae, S. Joo, Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram, in: International Conference on Learning Representations, 2024.
URL https://openreview.net/forum?id=WcOohbsF4H

[8] H. Li, Z. An, S. Zuo, W. Zhu, L. Cao, Y. Mu, W. Song, Q. Mao, Z. Zhang, E. Li, et al., Classification of electrocardiogram signals with waveform morphological analysis and support vector machines, Medical & Biological Engineering & Computing 60 (1) (2022) 109–119. doi:10.1007/s11517-021-02461-4.

[9] Y. Kutlu, D. Kuntalp, Feature extraction for ECG heartbeats using higher order statistics of WPD coefficients, Computer Methods and Programs in Biomedicine 105 (3) (2012) 257–267. doi:10.1016/j.cmpb.2011.10.002.

[10] S. Nahak, A. Pathak, G. Saha, Fragment-level classification of ECG arrhythmia using wavelet scattering transform, Expert Systems with Applications 224 (2023) 120019. doi:10.1016/j.eswa.2023.120019.

[11] S. Ji, R. Li, S. Shen, B. Li, B. Zhou, Z. Wang, Heartbeat classification based on multi-feature combination and stacking-DWKNN algorithm, Journal of Healthcare Engineering 2021 (2021) 8811837. doi:10.1155/2021/8811837.

[12] T. Xie, R. Li, S. Shen, X. Zhang, B. Zhou, Z. Wang, Intelligent analysis of premature ventricular contraction based on features and random forest, Journal of Healthcare Engineering 2019 (2019) 5787582. doi:10.1155/2019/5787582.

[13] R. Li, W. Zhang, S. Shen, J. Yao, B. Li, B. Zhou, G. Chen, Z. Wang, An intelligent heartbeat classification system based on attributable features with AdaBoost+ random forest algorithm, Journal of Healthcare Engineering 2021 (2021) 9913127. doi:10.1155/2021/9913127.

[14] J. Cheng, Q. Zou, Y. Zhao, ECG signal classification based on deep CNN and BiLSTM, BMC Medical Informatics and Decision Making 21 (1) (2021) 365. doi:10.1186/s12911-021-01736-y.

[15] R. Hu, J. Chen, L. Zhou, A transformer-based deep neural network for arrhythmia detection using continuous ECG signals, Computers in Biology and Medicine 144 (2022) 105325. doi:10.1016/j.compbiomed.2022.105325.

[16] H. Jiang, H. Mutahira, S. Wei, M. S. Muhammad, ECG-Mamba: Cardiac abnormality classification with non-uniform-mix augmentation on 12-lead ECGs, IEEE Journal of Translational Engineering in Health and Medicine 13 (2025) 461–470. doi:10.1109/JTEHM.2025.3613609.

[17] J. Yao, R. Li, S. Shen, W. Zhang, Y. Peng, G. Chen, Z. Wang, Combining rhythm information between heartbeats and BiLSTM-Treg algorithm for intelligent beat classification of arrhythmia, Journal of Healthcare Engineering 2021 (2021) 8642576. doi:10.1155/2021/8642576.

[18] J. Duan, Q. Wang, B. Zhang, C. Liu, C. Li, L. Wang, Accurate detection of atrial fibrillation events with RR intervals from ECG signals, PLOS ONE 17 (8) (2022) e0271596. doi:10.1371/journal.pone.0271596.

[19] S. Chen, H. Wang, H. Zhang, C. Peng, Y. Li, B. Wang, A novel method of swin transformer with time-frequency characteristics for ECG-based arrhythmia detection, Frontiers in Cardiovascular Medicine 11 (2024) 1401143. doi:10.3389/fcvm.2024.1401143.

[20] X. Han, A. Zhang, R. Li, S. Shen, D. Zhang, B. Jin, L. Mao, L. Yang, S. Zhang, An arrhythmia intelligent recognition method based on a multimodal information and spatio-temporal hybrid neural network model, Computers, Materials & Continua 82 (2) (2025) 3443–3465. doi:10.32604/cmc.2024.059403.

[21] R. Li, S. Lu, X. Han, S. Shen, X. Zhang, L. Han, Z. Yu, Method for cardiac abnormal event recognition based on multimodal spatiotemporal feature fusion model, International Journal of Machine Learning and Cybernetics 17 (5) (2026) 217. doi:10.1007/s13042-026-03050-3.

[22] R. Li, D. Su, S. Shen, S. Lu, Q. Zhang, S. Zhang, MVFF-SAT: A classification model for ECG signals integrating multi perspective features, Measurement 256 (2025) 118149. doi:10.1016/j.measurement.2025.118149.

[23] R. Li, L. Mao, M. Guo, P. Wang, H. Zhang, S. Shen, J. Bian, BiSAM-ECGNet: lightweight cross-modal alignment network for arrhythmia classification, International Journal of Machine Learning and Cybernetics 17 (7) (2026) 354. doi:10.1007/s13042-026-03194-2.

[24] G. Zhang, S. Shen, Research on arrhythmia classification algorithm based on ECG-BERT model with temporal semantic enhancement, in: Proceedings of the 4th International Conference on Biomedical and Intelligent Systems (IC-BIS 2025), 2025, pp. 117–124. doi:10.1145/3745034.3745054.

[25] K. McKeen, S. Masood, A. Toma, B. Rubin, B. Wang, ECG-FM: an open electrocardiogram foundation model, JAMIA Open 8 (5) (2025) ooaf122. doi:10.1093/jamiaopen/ooaf122.

[26] X. Gu, W. Tang, J. Han, V. Sangha, F. Liu, S. N. Gowda, A. H. Ribeiro, P. Schwab, K. Branson, L. Clifton, A. L. P. Ribeiro, Z. Liu, D. A. Clifton, Cardiac health assessment across scenarios and devices using a multimodal foundation model pretrained on data from 1.7 million individuals, Nature Machine Intelligence 8 (2) (2026) 220–233. doi:10.1038/s42256-026-01180-5.

[27] S. Lin, Z. Li, Q. Wu, Y. Chen, Y. Yang, H. Zhao, A self-supervised electrocardiogram foundation model for empowering cardiovascular disease prediction and genetic factor discovery, Nature Communications 17 (2026) 5771. doi:10.1038/s41467-026-72436-2.

[28] X. Zhou, T. Li, H. Hayama, K. Nakamura, S.-H. Liu, W. Chen, et al., Diagnosis of cardiac conditions from 12-lead electrocardiogram through natural language supervision, npj Digital Medicine 8 (2025) 697. doi:10.1038/s41746-025-02074-3.

[29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
URL https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_

Transformer_Using_Shifted_Windows_ICCV_2021_paper.html
[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[31] K. Cao, C. Wei, A. Gaidon, N. Arechiga, T. Ma, Learning imbalanced datasets with label-distribution-aware margin loss, in: Advances in Neural Information Processing Systems, Vol. 32, 2019.
[32] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, S. Kumar, Long-tail learning via logit adjustment, in: International Conference on Learning Representations, 2021.
[33] J. Snell, K. Swersky, R. S. Zemel, Prototypical networks for few-shot learning, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.
[34] S. Sethi, D. Chen, T. Statchen, M. C. Burkhart, N. Bhandari, B. Ramadan, B. Beaulieu-Jones, ProtoECGNet: Case-based interpretable deep learning for multi-label ECG classification with contrastive learning, in: Proceedings of the 10th Machine Learning for Healthcare Conference, Vol. 298 of Proceedings of Machine Learning Research, PMLR, 2025.
URL https://proceedings.mlr.press/v298/sethi25a.html
[35] H. Zhu, C. Cheng, H. Yin, X. Li, P. Zuo, J. Ding, et al., Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study, The Lancet Digital Health 2 (2020) e348–e357. doi:10.1016/s2589-7500(20)30107-2.
[36] Z. Li, H. Zhang, Fusing deep metric learning with KNN for 12-lead multi-labelled ECG classification, Biomedical Signal Processing and Control 85 (2023) 104849. doi:10.1016/j.bspc.2023.104849.
[37] P. Wagner, N. Strodthoff, R.-D. Bousseljot, W. Samek, T. Schaeffter, PTB-XL, a large publicly available electrocardiography dataset, PhysioNetVersion 1.0.3 (2022). doi:10.13026/kfzx-aw45.
[38] M. Pagliardini, P. Ablin, D. Grangier, The AdEMAMix optimizer: better, faster, older, arXiv preprint arXiv:2409.03137 (2024). doi:10.48550/arXiv.2409.03137.
[39] PhysioNet, WFDB programmer's guide: annotation codes, https://physionet.org/physiotools/wpg/wpg_36.htm, accessed 19 September 2026.
[40] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579–2605.
URL https://www.jmlr.org/papers/v9/vandermaaten08a.html
[41] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65. doi:10.1016/0377-0427(87)90125-7.
[42] D. L. Davies, D. W. Bouldin, A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1 (2) (1979) 224–227. doi:10.1109/TPAMI.1979.4766909.
[43] J. Venna, S. Kaski, Neighborhood preservation in nonlinear projection methods: An experimental study, in: Artificial Neural Networks—ICANN 2001, 2001, pp. 485–491. doi:10.1007/3-540-44668-0_68.
[44] V. Gliner, I. Levy, K. Tsutsui, M. R. Acha, J. Schliamser, A. Schuster, et al., Clinically meaningful interpretability of an AI model for ECG classification, npj Digital Medicine 8 (2025) 109. doi:10.1038/s41746-025-01467-8.
[45] K. Patlatzoglou, L. Pastika, J. Barker, E. Sieliwonczyk, G. R. Khattak, B. Zeidaabadi, et al., The cost of explainability in artificial intelligence-enhanced electrocardiogram models, npj Digital Medicine 8 (2025) 747. doi:10.1038/s41746-025-02122-y.
[46] Z. Chen, Y. Miao, L. Wang, L. Fan, D. P. Mandic, J. Zhu, Versatile cardiovascular signal generation with a unified diffusion transformer, Nature Machine Intelligence 8 (2026) 6–19. doi:10.1038/s42256-025-01147-y.
[47] S. Chen, Q. Ouyang, X. Meng, Y. Yang, C. Li, X. Miao, et al., Starfish-inspired wearable bioelectronic systems for physiological signal monitoring during motion and real-time heart disease diagnosis, Science Advances 11 (2025) eadv2406. doi:10.1126/sciadv.adv2406.
[48] B. Zheng, W. Luo, M. Zhang, H. Jin, Arrhythmia classification based on multi-input convolutional neural network with attention mechanism, PLOS ONE 20 (6) (2025) e0326079. doi:10.1371/journal.pone.0326079.
[49] S. Lamba, S. Kumar, M. Diwakar, FADLEC: feature extraction and arrhythmia classification using deep learning from electrocardiograph signals, Discover Artificial Intelligence 5 (2025) 82. doi:10.1007/s44163-025-00290-0.
[50] V. Adeleye, M. Elbattah, Comparative analysis of feature extraction methods for ECG arrhythmia classification using ensemble learning, BioMedInformatics 6 (3) (2026) 33. doi:10.3390/biomedinformatics6030033.
[51] S. Umarani, V. Kavitha, M. S. S. Sasikumar, S. Prabhu, ECG arrhythmia classification via wavelet-driven feature extraction and swarm-optimised gradient boosting, Computers in Biology and Medicine 213 (2026) 111839. doi:10.1016/j.compbiomed.2026.111839.