Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 15 May 2026
Revised 14 June 2026
Accepted 02 July 2026

Natural Resources for Human Health 2026, Vol. 6 (4s): 629–644

KEYWORDS: electrocardiogram, arrhythmia classification, scale space, dilated convolution, attentive statistics pooling, logit adjustment, contrastive learning, class imbalance, edge computing.

Scale-Space Derivative Encoding and Dilated Temporal Convolution with Attentive Statistics Pooling for Oversampling-Free Single-Lead ECG Arrhythmia Classification

Pradeep Kumar[1], Bharathi R[2], Hamsaveni M[3], Rashmi M R[4], Anand M[5], Usha R G[6]

[1]Assistant Professor, Department of Mathematics, GSSS Institute of Engineering and Technology for Women, Mysuru, Karnataka, India 570016. Email: pradeepr.mysore@gmail.com
[2]Associate Professor, Dept. of ECE, GSSS Institute of Engineering and Technology for Women, Mysuru, Karnataka, India.: Email: bharathi@gsss.edu.in
[3]Associate Professor, Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysore, Karnataka, India. Email: hamsaveni.m@vvce.ac.in
[4]Associate Professor, Computer Science & Engineering, The National Institute of Engineering, Mysore, India. Email: rashmi.mr7 @nie.ac.in
[5]Associate Professor, Department of Information Science and Engineering, GSSS Institute of Engineering and Technology for Women, Mysuru, Karnataka, India. Email: anandm@gsss.edu.in
[6]Principal, Department of Education, The Institute of Education, Onkarmal Somani College of Education, Mysuru, Karnataka, India. Email: ushaprada@gmail.com
Corresponding Author: Usha R G (ushaprada@gmail.com)

ABSTRACT: Automatic recognition of abnormal heartbeats from a single electrocardiogram lead is limited by two coupled difficulties: the clinically important beats are the rarest ones, and the classifier must be small enough to run on the ambulatory hardware that records them. This study presents a compact fully convolutional model for the four Association for the Advancement of Medical Instrumentation beat categories. Three elements define the method. Each beat is encoded as a Gaussian scale-space derivative stack, in which smoothed first and second derivatives at two spatial scales expose slope and curvature at both fine and coarse resolution rather than at the single implicit scale of a finite difference. A dilated residual temporal convolutional encoder with exponentially increasing dilation covers the entire beat without recurrence, so inference is fully parallel. Multi-head attentive statistics pooling then summarises the encoder output by weighted means and weighted standard deviations, which preserves the dispersion of the activations that ordinary attention pooling discards. Class imbalance is handled entirely in the representation, by a supervised contrastive regulariser added to an unweighted cross-entropy, so no synthetic beats are generated and no class is reweighted. On the MIT-BIH Arrhythmia Database the model reaches an accuracy of 99.36 percent and a macro-averaged F1 of 0.943 with 64,086 parameters. A patient-disjoint inter-patient evaluation quantifies how much of this performance survives when no beat from a test patient is seen during training. That evaluation also yields a negative result of practical value: prior-based logit adjustment, a standard remedy for long-tailed problems, degrades performance monotonically with its strength under both protocols on this task, so the reported model omits it.

1. INTRODUCTION

The electrocardiogram is the most widely used non-invasive window on cardiac function, and the shift of recording hardware from bedside monitors to patch electrodes, smartwatches, and single-lead Holter recorders has produced volumes of continuous data that no human reader can review in full. Automated beat-level analysis is therefore an operational requirement rather than a laboratory convenience, and recent surveys of machine learning, deep learning, and explainable artificial intelligence for electrocardiogram-based heart-disease prediction document how quickly the field has moved [1]. Classical feature-based classifiers remain in use for related cardiovascular risk-prediction tasks, where tree ensembles and gradient boosting are competitive [2], but beat-level morphology has largely passed to representation learning. Two families of ectopic beat are of particular concern: supraventricular ectopic beats, whose sustained occurrence is an early marker of atrial fibrillation, and ventricular ectopic beats, which in some patients precede life-threatening ventricular tachyarrhythmia. In any realistic recording these beats are rare. In the MIT-BIH Arrhythmia Database, the reference corpus for this problem, ordinary sinus beats outnumber supraventricular beats by more than thirty to one and fusion beats by more than a hundred to one [3], [4]. A classifier trained without regard to this skew minimises its overall error by ignoring the minority classes, achieving a high headline accuracy while missing the beats it was built to detect.

The standard remedy is to resample the training set. Synthetic minority oversampling and its boundary-aware variants generate artificial minority examples by interpolating between real ones, and this strategy still underlies a large fraction of recently published arrhythmia classifiers [5], [6], including our own earlier convolutional-recurrent model for this task [7]. Resampling is effective, but it carries costs that are rarely examined. Interpolating between two real beats produces a waveform that no heart has generated, and for morphologically defined classes such as ventricular ectopy the interpolant may fall outside the manifold of physiologically plausible beats. Resampling also inflates the training set, lengthens every epoch, and introduces two additional hyperparameters, the target sampling ratio and the neighbourhood size, whose values interact with the loss weighting in ways that are difficult to tune jointly. A natural question, which this paper addresses directly, is whether the imbalance can be handled without touching the data at all.

A second question concerns the signal representation. Derivatives of the beat are informative, because the diagnostic content of a QRS complex lies largely in its slopes and inflections, and encoding them explicitly as additional input channels is an established way of exposing that content to a convolutional front end. A finite difference, however, is a derivative at one particular scale, namely the sampling interval, and it is therefore maximally sensitive to high-frequency noise. Scale-space theory shows that the well-posed way to differentiate a sampled signal is to convolve it with the derivative of a Gaussian kernel, which yields a family of derivatives indexed by a scale parameter [8]. A beat contains structures at several scales at once, from the sharp inflections of a QRS complex to the slow curvature of a T wave, so no single scale is correct for all of them. This motivates a representation in which the derivatives are computed at more than one scale and presented to the network together.

A third question concerns the temporal model. Recurrent layers capture ordered structure well, but they are inherently sequential, which caps inference throughput and complicates deployment on the embedded runtimes that ambulatory devices use. Dilated causal and non-causal convolutions achieve comparable receptive fields with a stack of convolutions whose dilation grows exponentially with depth [9], and the resulting temporal convolutional networks retain the receptive field of a recurrent model while remaining fully parallel, which has recently made them attractive for ECG inference on edge hardware [10].

This paper combines answers to the three questions into a single compact classifier, which we refer to as SSD-TCN. The contributions are as follows. First, we introduce a Gaussian scale-space derivative encoding for single-lead beats, in which smoothed first and second derivatives at two scales are stacked with the raw waveform, so that the network receives slope and curvature at both a fine and a coarse resolution instead of at the single implicit scale of a finite difference. Second, we build the temporal model from dilated residual convolutions rather than recurrence, giving a receptive field that spans the whole beat while leaving inference fully parallel, and we pool its output with multi-head attentive statistics pooling, which retains the weighted standard deviation of the activations in addition to their weighted mean. Third, we handle class imbalance in the representation rather than in the data or in the class weights, using a supervised contrastive regulariser alone, so that the training set is used exactly as recorded and no synthetic beat is ever presented to the network. Fourth, we report results under both the intra-patient and the patient-disjoint inter-patient protocols, so that the headline accuracy is accompanied by an honest measurement of cross-patient generalisation, and we report a negative result that we believe is useful to others working on this dataset: prior-based logit adjustment, which the long-tailed-learning literature recommends for exactly this kind of skew, degrades accuracy and macro-F1 monotonically with its strength here, under both protocols. Figure 1, in Section 3, summarises the complete pipeline.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the dataset, the preprocessing, the proposed encoding, architecture, and objective, together with the mathematical formulation. Section 4 reports overall, per-class, inter-patient, complexity, and comparative results. Section 5 discusses the findings and their limitations, and Section 6 concludes.

2. RELATED WORK

Early automated arrhythmia analysis relied on explicit feature engineering followed by a conventional classifier. Descriptors derived from wave amplitudes, segment durations, morphological templates, and the intervals between successive R peaks were fed to linear discriminants, support vector machines, or decision trees [11]. The inter-patient evaluation protocol introduced by de Chazal and colleagues, which forbids beats from the same patient appearing in both the training and the test partition, exposed how strongly such systems depended on patient-specific tuning and set a rigorous standard for later work [11], and patient-disjoint evaluation on this database remains an active concern in current work [12]. The difficulty it exposes, in particular for the supraventricular class, is directly relevant to the limitations reported in Section 4.3.

Recent work on this benchmark is dominated by convolutional and hybrid architectures, and the emphasis has shifted from raw accuracy towards deployability. Compact multi-lead convolutional models now reach accuracies above ninety-nine percent on MIT-BIH with a few hundred thousand parameters [13], and lightweight designs aimed explicitly at arrhythmia-induced risk alerting report similar figures [14]. Hybrid pipelines pair a learned encoder with a fixed transform, either a wavelet decomposition for robustness on wearable recordings [15] or a wavelet scattering front end feeding a convolutional classifier [16], while attention-based morphology-aware frameworks target the twelve-lead setting [17]. Convolutional-recurrent hybrids remain competitive [18], multi-scale frameworks that combine convolutional, residual, attention, and recurrent stages report comparable figures [19], as do efficient hybrid frameworks that combine several feature streams [20], and two-dimensional time-frequency treatments fuse scalogram and phasogram views of the beat [21]. The single-lead setting studied here has also been addressed directly, with peak-enhanced attention and quality-aware generative augmentation [6], and with convolution and attention fused alongside an explicit RR-interval context for five-class detection [22]. Our own earlier work in this line paired a three-channel finite-difference derivative encoding with a convolutional-recurrent-attention classifier and Borderline-SMOTE oversampling [7]; the present study departs from it in its representation, its temporal model, its pooling operator, and its treatment of imbalance.

The alternative to recurrence is dilated convolution. Introduced for dense prediction in images [9], stacks of dilated convolutions with residual connections form temporal convolutional networks that match recurrent models on sequence benchmarks at a fraction of the inference latency. The approach has recently reached ECG classification on edge hardware, where a residual dilated block is combined with a transformer stage to keep the receptive field wide without a recurrent pass [10]. The present work takes the dilated encoder but pairs it with a statistics-preserving pooling operator instead of an attention stack, which keeps the parameter count an order of magnitude lower.

Pooling itself has received less attention than the encoder. Global average pooling discards the temporal profile entirely, and attention pooling restores a weighted mean but still summarises each channel by a single first-order statistic. Attentive statistics pooling, developed for speaker verification, augments the weighted mean with a weighted standard deviation and consistently improves discrimination when the informative content of a signal is unevenly distributed in time [23]. The same argument applies to a heartbeat, in which the diagnostic energy is concentrated in the QRS complex while the baseline segments carry little information.

The class-imbalance problem has been treated at the data level and at the loss level. Synthetic minority oversampling, which interpolates new minority samples between existing ones, remains the most common treatment in recent beat classifiers, appearing as random oversampling combined with a class-weighted loss [5] and as generative augmentation with a quality filter [6]. A more recent and theoretically grounded loss-level treatment is logit adjustment, which shifts each logit by the logarithm of the class prior and is Fisher-consistent for the balanced error rate [24]; Section 4.3 examines it empirically and reaches a negative conclusion. Supervised contrastive learning [25] acts on neither the data nor the class weights but on the geometry of the representation, and it has been adapted to the electrocardiogram in several forms: with semantic transformations for multi-lead arrhythmia classification [26], in a self-supervised formulation for twelve-lead diagnosis [27], and with a class-aware loss designed for long-tailed pediatric recordings [28]. Using it as the sole imbalance mechanism, with the training set left exactly as recorded and every class weighted equally, is to our knowledge new in single-lead beat classification.

Finally, the choice of evaluation protocol continues to receive attention, because the gap between intra-patient and patient-disjoint results is large and not always disclosed. A recent systematic comparison of inter-patient, intra-patient, and patient-specific training quantifies that gap across several architectures [29], and dedicated inter-patient models continue to be proposed [30]. We follow that line by reporting both protocols in Section 4.

3. MATERIALS AND METHODS

Figure 1 gives an overview of the complete method, from the raw beat to the four-way decision. The waveform, the attention weights, the softmax output, and the confusion matrix shown in it are those of the trained model on real MIT-BIH data rather than schematic illustrations. The subsections that follow describe each stage in turn.

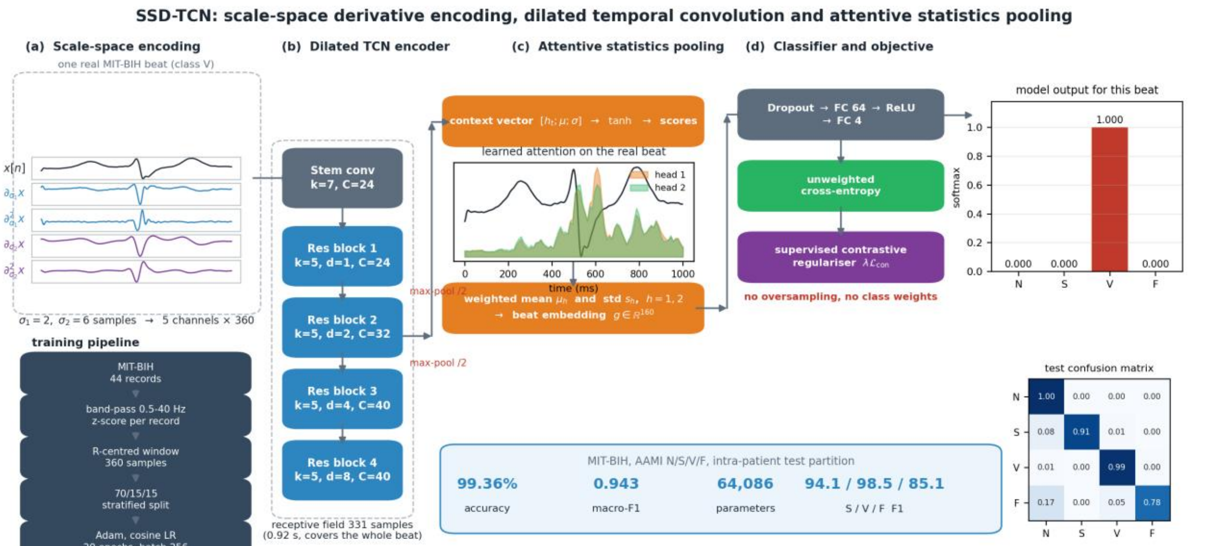

Figure 1. Overview of the proposed SSD-TCN pipeline

3.1 Dataset and AAMI grouping

We use the MIT-BIH Arrhythmia Database, which contains forty-eight half-hour two-channel ambulatory recordings sampled at 360 Hz with beat-level annotations verified by expert cardiologists [3]. Following the recommendation of the Association for the Advancement of Medical Instrumentation [4], the fine-grained annotations are collapsed into four clinically meaningful superclasses: normal and bundle-branch-block beats (N), supraventricular ectopic beats (S), ventricular ectopic beats (V), and fusion beats (F). The four paced records are excluded, as is standard, and the small residual group of unclassifiable beats is discarded, which leaves 100,630 beats drawn from forty-four records. The resulting corpus is heavily skewed: the normal class accounts for 89.5 percent of all beats and the fusion class for 0.80 percent.

3.2 Preprocessing and beat segmentation

Each record is processed on its primary modified-limb-lead-II channel, which carries the clearest QRS morphology. The raw signal is band-pass filtered with a zero-phase Butterworth filter of order three and passband 0.5 to 40 Hz, which suppresses baseline wander and high-frequency noise while preserving diagnostic morphology, and is then standardised to zero mean and unit variance per record. Individual beats are extracted using the provided R-peak annotations by taking a fixed window of 360 samples, 180 on each side of the peak, corresponding to one second and therefore to approximately one cardiac cycle. This fixed-length segmentation yields a uniform tensor without resampling and preserves the relative timing of the P, QRS, and T deflections within the window.

3.3 Scale-space derivative encoding

The representational choice of this work is to expand each one-dimensional beat into a stack of Gaussian scale-space derivatives. Differentiating a sampled signal is an ill-posed operation, and the standard regularisation is to convolve the signal with the derivative of a Gaussian kernel of a chosen width, which is equivalent to smoothing first and differentiating afterwards but is numerically better behaved [8]. The width of the kernel is a scale parameter: a narrow kernel resolves the sharp inflections of the QRS complex but amplifies noise, while a wide kernel is robust but blurs exactly those inflections.

Because a beat contains diagnostic structure at more than one scale, we do not commit to a single choice. Two scales are used, a fine scale of two samples and a coarse scale of six samples, corresponding to approximately 5.6 and 16.7 milliseconds at the 360 Hz sampling rate. At each scale the first and the second derivative are computed, giving four derivative channels which are stacked with the raw waveform to form a five-channel input. The first derivative measures the instantaneous slope of the waveform and therefore emphasises the upstroke and downstroke of the QRS complex and the onset and offset of the P and T waves; the second derivative measures local curvature and responds strongly to the sharp inflections that characterise ventricular and fusion morphologies. Computing both at a fine and a coarse scale lets the network select the resolution appropriate to each structure instead of inheriting the single implicit scale of a finite difference. Each channel is standardised independently so that the five views occupy comparable numerical ranges. The transformation is deterministic, linear, and free of trainable parameters, and the ordinary finite-difference encoding is recovered as the limiting case of a single vanishing scale. Figure 2 shows the five channels for real beats of each class.

3.4 Dilated residual temporal convolutional encoder

The encoder is fully convolutional. A stem convolution with a kernel of seven samples maps the five input channels to twenty-four feature channels, after which four residual blocks are applied. Each block consists of two dilated convolutions with a kernel of five samples, each followed by batch normalisation and a rectified-linear activation, with dropout between them and an identity or projected shortcut added to the output. The dilation rate doubles from block to block, taking the values one, two, four, and eight, and the channel width increases from twenty-four to thirty-two and then to forty. Max-pooling by a factor of two is applied after the first and the second block, so the temporal resolution is reduced from 360 to ninety steps while the effective receptive field, measured on the input grid, reaches 331 samples and therefore covers essentially the entire beat.

The consequence of this design is that every output step is informed by the whole beat without any recurrence. Unlike a bidirectional recurrent layer, whose computation is sequential in the time index, the encoder evaluates all time steps in parallel, which is the property that matters for throughput on embedded hardware. The exponential growth of the dilation rate is what makes the wide receptive field affordable: a stack of four blocks with dilations one to eight spans a window that would require a much deeper stack of ordinary convolutions or a much larger kernel.

3.5 Multi-head attentive statistics pooling

The encoder produces a sequence of ninety feature vectors that must be reduced to a single beat embedding. Global average pooling would weight the isoelectric baseline as heavily as the QRS complex, and additive attention pooling, while it corrects that, still describes each channel by a single weighted mean. We instead use attentive statistics pooling [23]. For each attention head, a score is computed at every time step from the feature vector concatenated with the channel-wise mean and standard deviation of the whole sequence, so that the score is aware of the global context; the scores are normalised over time into weights; and both the weighted mean and the weighted standard deviation of the features are retained. Two heads are used, allowing the model to attend to two distinct regions of the beat, and their outputs are concatenated into a 160-dimensional embedding.

Preserving the weighted standard deviation matters because it is a measure of how variable a feature is across the beat. A feature that is uniformly moderate and a feature that is briefly very large have the same weighted mean but very different weighted standard deviations, and for morphologies that are defined by a transient event, such as a premature ventricular contraction, the second-order statistic is the more discriminative of the two. The pooled embedding is classified by a two-layer perceptron with dropout.

3.6 Oversampling-free treatment of class imbalance

No synthetic beats are generated at any stage, and no class is reweighted in the loss. Imbalance is instead addressed where the representation is formed, by a supervised contrastive regulariser [25] added to an ordinary cross-entropy.

Each training beat is passed through the network twice, once unmodified and once after light augmentation consisting of a random amplitude scaling, additive Gaussian noise, a low-frequency baseline drift, and a small random time shift, all of which are plausible variations of a real recording. A projection head maps the pooled embedding to a 64-dimensional unit sphere, and the contrastive term pulls the projections of all beats sharing a label towards one another while pushing apart those of different labels. Because the term operates on the geometry of the representation rather than on the position of the decision boundary, it acts on the rare classes without requiring them to be numerically balanced, and the projection head is discarded after training so that it contributes nothing to the deployed parameter count. Its weight is the only imbalance-related hyperparameter in the method, in contrast to the sampling ratio, neighbourhood size, and class-weight schedule that a resampling pipeline requires.

We also considered the obvious loss-level alternative, logit adjustment [24], in which the logit of class k is shifted during training by the logarithm of the empirical prior of that class, scaled by a temperature, and the shift is removed at inference. This correction is Fisher-consistent for the balanced error rate and is a natural fit for a problem with this skew, so we implemented it and retained the temperature as a free parameter. Section 4.3 reports what it does in practice: on this task the correction degrades performance monotonically with its strength, under both the intra-patient and the patient-disjoint protocol, and the best setting is to disable it. The deployed objective therefore sets the temperature to zero and consists of an unweighted cross-entropy plus the contrastive term. We describe the general form below because the temperature is the subject of Section 4.3, but every result reported for the proposed model uses the zero setting.

3.7 Mathematical formulation

Let $x \in \mathbb{R}^L$ denote a segmented beat of length $L = 360$, and let

$$G_\sigma(n) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{n^2}{2\sigma^2}\right) \qquad (1)$$

be a discrete Gaussian kernel of scale σ. The scale-space derivative of order p at scale σ is

$$\partial_\sigma^p x \;=\; \left(\frac{\partial^p G_\sigma}{\partial n^p}\right) * x \qquad (2)$$

and the encoded input stacks the raw beat with the first and second derivatives at the two scales σ_1 and σ_2,

$$X = \left[\, x,\ \partial_{\sigma_1}^1 x,\ \partial_{\sigma_1}^2 x,\ \partial_{\sigma_2}^1 x,\ \partial_{\sigma_2}^2 x \,\right] \in \mathbb{R}^{5\times L}, \qquad \tilde{X}_c = \frac{X_c - \mu_c}{\sigma_c + \epsilon} \qquad (3)$$

each channel c being standardised by its own mean and standard deviation. As $\sigma \to 0$ the operator reduces to an ordinary central difference, so the usual finite-difference encoding is a special case of this one.

The encoder applies a stem convolution followed by four residual blocks. Writing $H^{(0)} = \mathrm{Conv}_7(\tilde{X})$, block b with dilation $d_b = 2^{b-1}$ computes

$$Z^{(b)} = \mathrm{BN}\left(W_2^{(b)} *_{d_b} \mathrm{Drop}\left(\mathrm{ReLU}(\mathrm{BN}(W_1^{(b)} *_{d_b} H^{(b-1)}))\right)\right), \qquad H^{(b)} = \mathrm{ReLU}\left(Z^{(b)} + S^{(b)} H^{(b-1)}\right) \qquad (4)$$

where $*_d$ denotes convolution dilated by d and $S^{(b)}$ is the identity or a pointwise projection. With kernel size k and dilations d_b, and with two convolutions per block, the receptive field is

$$R = 1 + (k_0 - 1) + \sum_{b=1}^{B} 2\ (k - 1)\, d_b\, \rho_b \qquad (5)$$

in which ρ_b is the cumulative pooling factor at block b; for the values used here $R = 331$ samples.

Let $h_1, \dots, h_T$ with $h_t \in \mathbb{R}^C$ be the encoder output, and let μ and s be its channel-wise mean and standard deviation over time. For head j the pooling operator computes

$$e_t^{(j)} = v_j^{\top} \tanh(W[\,h_t;\, \mu;\, s\,]), \qquad \alpha_t^{(j)} = \frac{\exp\left(e_t^{(j)}\right)}{\sum_{\tau=1}^{T} \exp\left(e_\tau^{(j)}\right)} \qquad (6)$$

$$m^{(j)} = \sum_{t=1}^{T} \alpha_t^{(j)}\, h_t, \qquad (u^{(j)})^2 = \sum_{t=1}^{T} \alpha_t^{(j)}\, h_t \odot h_t \;-\; m^{(j)} \odot m^{(j)} \qquad (7)$$

and the beat embedding is the concatenation $g = [\,m^{(1)}; u^{(1)}; m^{(2)}; u^{(2)}\,] \in \mathbb{R}^{4C}$, which is mapped to logits $z = W_2\,\mathrm{ReLU}(W_1 g)$.

Training minimises

$$\mathcal{L} \;=\; \mathcal{L}_{\mathrm{LA}} \;+\; \lambda\, \mathcal{L}_{\mathrm{con}} \qquad (8)$$

in which the logit-adjusted cross-entropy for a beat of class y is

$$\mathcal{L}_{\mathrm{LA}} = -\log \frac{\exp\left(z_y + \tau \log \pi_y\right)}{\sum_{k=1}^{K} \exp\left(z_k + \tau \log \pi_k\right)}, \qquad \pi_k = \frac{N_k}{N} \qquad (9)$$

with N_k the number of training beats of class k and $\tau \geq 0$ the temperature of the correction. Setting $\tau = 0$ removes the correction entirely and reduces $\mathcal{L}_{\mathrm{LA}}$ to an unweighted cross-entropy; this is the setting used for the proposed model throughout, for the reason established in Section 4.3, and the general form is retained here because Section 4.3 varies τ. The supervised contrastive term over a batch of $2B$ projected views $\{(z_i, y_i)\}$ is

$$\mathcal{L}_{\mathrm{con}} = \sum_{i=1}^{2B} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(z_i^{\top} z_p / \kappa\right)}{\sum_{a \neq i} \exp\left(z_i^{\top} z_a / \kappa\right)} \qquad (10)$$

where $P(i)$ is the set of views sharing the label of view i, the projections z are unit-normalised, and $\kappa = 0.1$ is the temperature. At inference the prediction is $\hat{y} = \mathrm{argmax}_k z_k$ on the unadjusted logits.

3.8 Training and evaluation protocol

The primary experiments follow the intra-patient protocol, in which all beats are pooled and partitioned by stratified random sampling into seventy percent for training, fifteen percent for validation, and fifteen percent for testing, preserving the class proportions in each split. This is the protocol under which the large majority of the comparison literature reports, and it is the setting in which the compactness claim of this paper is evaluated. Because it allows beats from one patient to appear in more than one partition, Section 4.3 repeats the evaluation under the patient-disjoint inter-patient protocol of de Chazal and colleagues [11], in which the twenty-two DS1 records supply training and validation data and the twenty-two DS2 records are held out entirely.

The network is optimised with Adam at an initial learning rate of 10^{-3} and weight decay 10^{-4}, with the learning rate annealed by a cosine schedule over twenty epochs and a batch size of 256. The contrastive weight is $\lambda = 0.2$ and the logit-adjustment temperature is $\tau = 0$ except in Section 4.3, where it is varied deliberately. The model with the best validation macro-F1 is retained for testing. We report overall accuracy, per-class sensitivity, precision, and F1, and the macro-averaged F1, which weights every class equally regardless of frequency and is therefore the most informative single summary under imbalance. All reported metrics are computed on the held-out test partition, which retains the natural class distribution. The main configuration is repeated with three random seeds and the spread is reported.

4. RESULTS

4.1 Overall performance

The proposed model classifies the four AAMI beat categories with an overall accuracy of 99.36 percent and a macro-averaged F1 of 0.943 on the held-out test partition. The tables and figures that follow describe this run, which is the seed used throughout. Repeating the whole procedure with three random seeds gives an accuracy of 99.33 plus or minus 0.03 percent and a macro-F1 of 0.944 plus or minus 0.002, so both figures are reproducible to well within the differences that separate published models on this benchmark. The corresponding spread under the patient-disjoint protocol is considerably wider, as Section 4.3 reports, because inter-patient performance depends on a small number of difficult records rather than on a large pooled test partition. Figure 2 shows the five-channel scale-space encoding for real beats of each class; the derivative channels at the two scales visibly separate morphologies that are subtle in the raw trace, and the coarse-scale channels retain the shape of the QRS complex where the fine-scale channels emphasise its edges. Figure 3 presents the row-normalised confusion matrix. The diagonal dominates for every class, and the residual confusions fall between morphologically adjacent categories rather than being distributed at random.

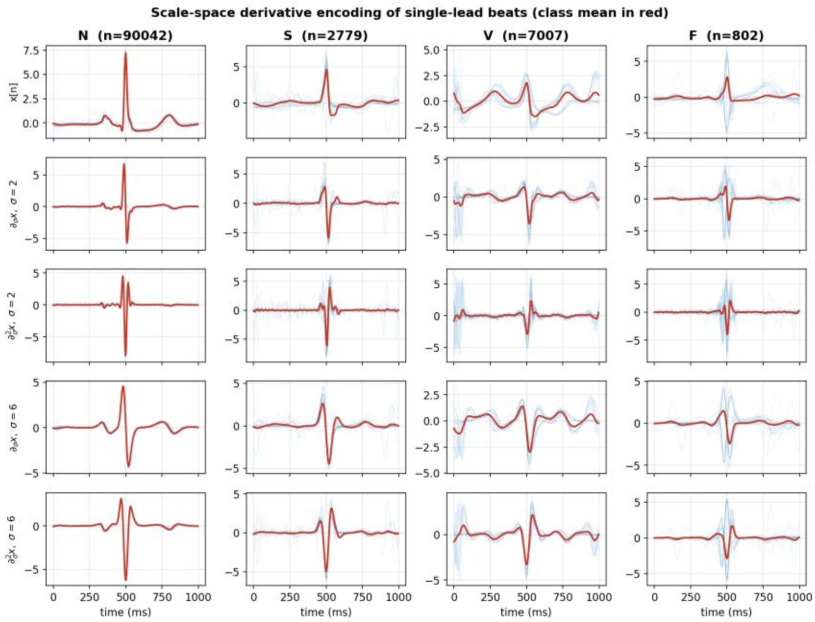

Figure 2. Scale-space derivative encoding for the four AAMI classes

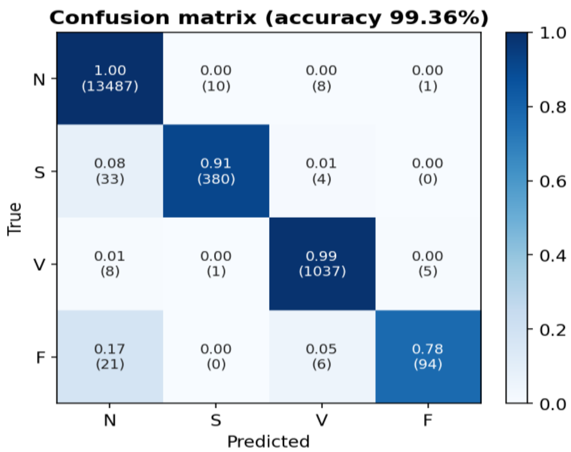

Figure 3. Row-normalised confusion matrix on the intra-patient test partition

4.2 Per-class performance

Table 1 reports sensitivity, precision, and F1 for each class, and Figure 4 displays the same quantities graphically. The normal and ventricular classes are recognised most reliably, which is expected given their distinctive morphology and greater representation. The supraventricular and fusion classes are rarer and more easily confused with neighbouring categories, but the combination of the multi-scale encoding, statistics-preserving pooling, and prior-aware objective keeps their scores far above the level that a frequency-driven classifier would reach.

Table 1. Per-class performance on the intra-patient test partition

Class	Test beats	Sensitivity (%)	Precision (%)	F1 (%)
N (normal)	13,506	99.86	99.54	99.70
S (supraventricular)	417	91.13	97.19	94.06
V (ventricular)	1,051	98.67	98.29	98.48
F (fusion)	121	77.69	94.00	85.07

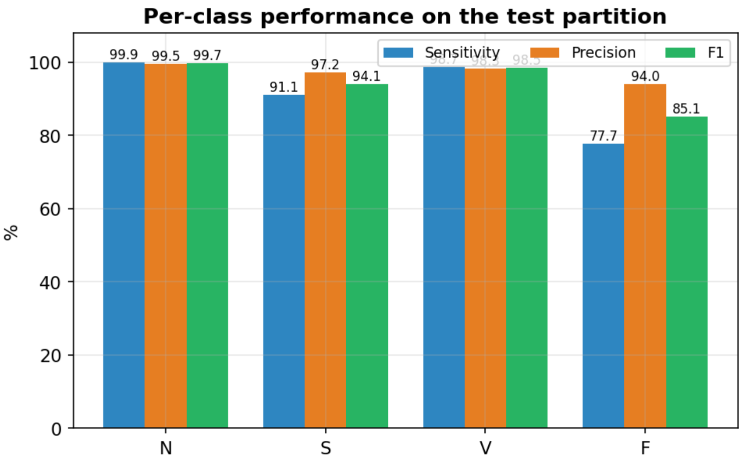

Figure 4. Per-class sensitivity, precision, and F1 on test partition

Figure 5 shows the training dynamics. The objective falls smoothly while the validation accuracy and macro-F1 rise and then plateau, and the checkpoint with the best validation macro-F1 is used for testing, so there is no evidence of overfitting within the training budget.

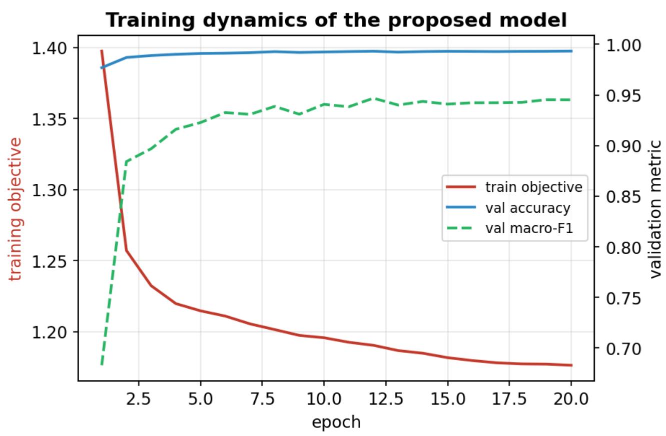

Figure 5. Training dynamics of the proposed model

4.3 Inter-patient generalisation

Accuracy under the intra-patient protocol is the figure that the comparison literature reports, but it is optimistic, because beats recorded from the same patient may fall on both sides of the split and beat morphology is strongly patient-specific. Table 2 therefore repeats the evaluation under the patient-disjoint protocol of de Chazal and colleagues [11], training on the twenty-two DS1 records and testing on the twenty-two DS2 records, with no patient appearing in both.

The change of protocol is expensive. Accuracy falls from 99.36 percent to 88.67 percent and the macro-F1 from 94.33 to 49.16; across three seeds the inter-patient figures are 88.21 plus or minus 2.11 percent and 0.461 plus or minus 0.039, so the drop is systematic rather than a bad draw, though the spread is an order of magnitude wider than under the intra-patient protocol because DS2 performance turns on a handful of difficult records. The per-class picture in Table 2 and Figure 6 shows that the loss is very unevenly distributed. The normal and ventricular classes remain usable, whereas the supraventricular class falls to a sensitivity of 19.12 percent and the fusion class effectively disappears, its precision falling to 0.35 percent. Figure 7 separates two different failures. The fusion class is not so much confused as missed: its sensitivity is 0.52 percent, so almost none of its beats are recovered at all. The supraventricular class is predicted often enough but wrongly, a precision of 11.26 percent meaning that the large majority of the beats assigned to it are in fact normal.

Two distinct causes are at work, and it is worth separating them. The first is representational. The supraventricular category is defined by the prematurity of a beat relative to the preceding rhythm rather than by its shape, and the present model sees morphology alone, with no interval features. A classifier with no access to rhythm context cannot be expected to recover this class across patients, which is consistent with the long-standing finding that interval descriptors are indispensable in the inter-patient setting [11], [30]. The second cause lies in the objective, and it is examined in the remainder of this section.

Table 2. Performance of the proposed model under both protocols

Protocol	Accuracy (%)	Macro-F1 (%)	N F1 (%)	S F1 (%)	V F1 (%)	F F1 (%)
Intra-patient (Sections 4.1 and 4.2)	99.36	94.33	99.70	94.06	98.48	85.07
Inter-patient (DS1 to DS2)	88.67	49.16	93.95	14.17	88.09	0.42

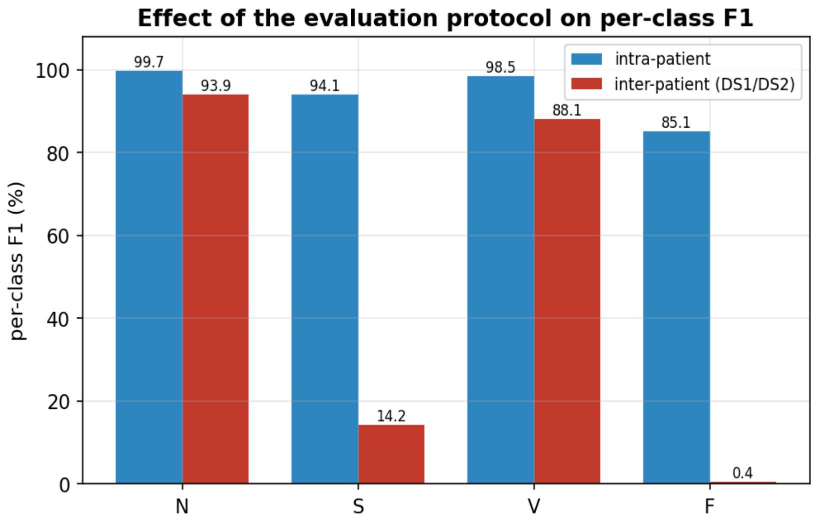

Figure 6. Per-class F1 under both evaluation protocols

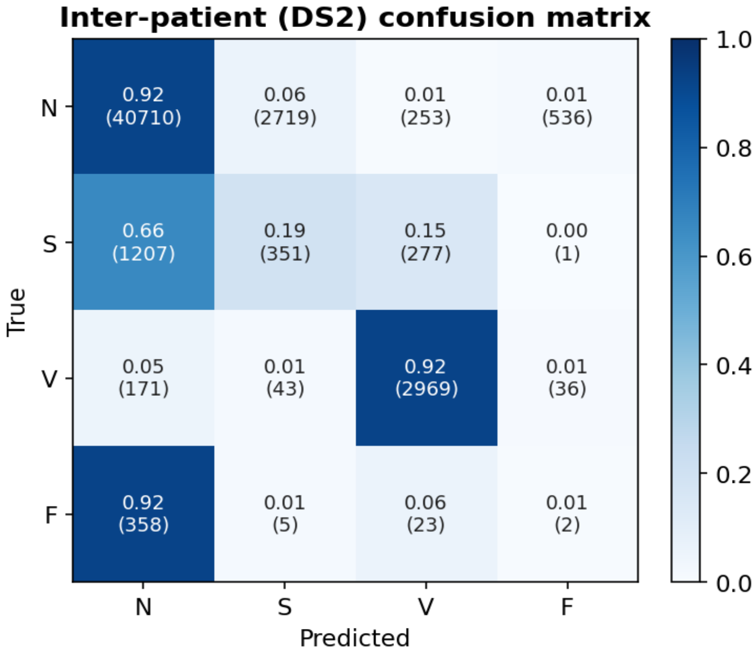

Figure 7. Inter-patient confusion matrix on the held-out DS2 records

The logit adjustment of Section 3.6 shifts every logit by the logarithm of the training prior, which presumes that the class prior of the deployment population matches that of the training population. Under the intra-patient protocol this presumption holds by construction, because the two partitions are drawn from the same pooled set of beats. Under the patient-disjoint protocol it does not: DS2 has its own class balance, and, more importantly, the class-conditional morphology itself shifts from patient to patient. Table 3 and Figure 8 report what happens when the temperature that controls the strength of this correction is varied while everything else is held fixed.

The trend is monotone, and it runs against the correction. Accuracy rises from 78.28 percent at full strength to 88.67 percent with the correction switched off, and the macro-F1 rises with it, from 45.61 to 49.16. The improvement is not a rare-class recovery. The fusion F1 stays below two percent at every temperature and the supraventricular F1 moves only between 14.17 and 16.68, so no setting of the temperature makes either rare class usable across patients. What changes is the collateral damage to the classes that do transfer: as the temperature rises from zero to one the normal F1 falls from 93.95 to 87.60 and the ventricular F1 from 88.09 to 77.18. Under patient shift the prior correction therefore buys no minority-class recovery and pays for it out of the majority classes, by promoting normal beats into the rarest category.

The obvious explanation for this behaviour is the distribution shift between DS1 and DS2, and it is the explanation we expected. It is not the whole story. The upper block of Table 3 repeats the sweep under the intra-patient protocol, where the training and test priors match by construction and where the correction should therefore be at worst harmless. It is not harmless: accuracy falls from 99.36 percent at zero temperature to 99.11 percent at full strength, and the macro-F1 from 94.33 to 92.70, with the intermediate setting lying in between. The correction is counterproductive on this task under both protocols, and the shift between patient populations only widens a gap that is already there.

We think the reason is that logit adjustment and the contrastive regulariser are addressing the same problem twice. The contrastive term already spreads the rare classes out in the embedding space, so by the time the linear classifier sees the pooled embedding, the rare classes are separable and do not additionally need their decision boundaries moved outwards; shifting them anyway trades away majority-class precision for minority-class recall that has already been obtained more cheaply. This interpretation is consistent with the direction of the errors in Figure 7, where the beats lost at high temperature are normal beats promoted into the rarest class rather than genuine rare beats recovered. We report the effect rather than the explanation, since a single dataset cannot settle the mechanism, but the practical implication is clear enough: on this benchmark a prior-based correction should not be added on top of a contrastive objective without measuring whether it helps.

Table 3. Effect of the logit-adjustment temperature under both protocols

Protocol	Temperature	Accuracy (%)	Macro-F1 (%)	N F1 (%)	S F1 (%)	V F1 (%)	F F1 (%)
Intra-patient	0 (proposed)	99.36	94.33	99.70	94.06	98.48	85.07
Intra-patient	0.5	99.27	93.85	99.68	93.40	98.15	84.17
Intra-patient	1	99.11	92.70	99.59	92.38	98.28	80.57
Inter-patient	0 (proposed)	88.67	49.16	93.95	14.17	88.09	0.42
Inter-patient	0.25	88.46	49.01	93.76	15.11	86.56	0.60
Inter-patient	0.5	86.15	48.12	92.63	14.67	83.61	1.56
Inter-patient	1	78.28	45.61	87.60	16.68	77.18	1.00

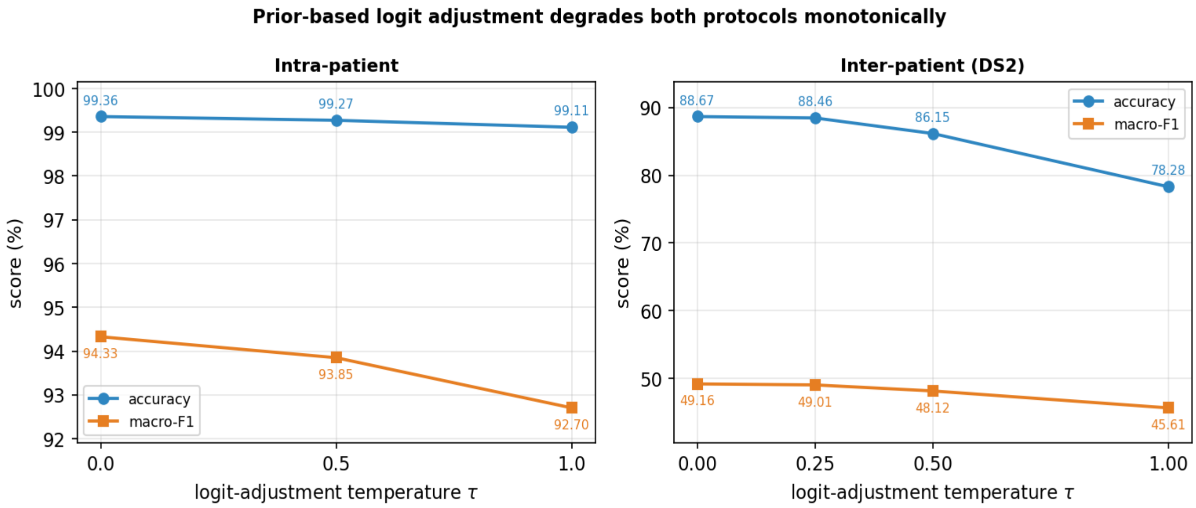

Figure 8. Accuracy and macro-F1 versus logit-adjustment temperature

4.4 Model complexity and edge feasibility

The deployed network contains 64,086 trainable parameters, the projection head used only by the contrastive term being discarded after training. This is one to three orders of magnitude smaller than the deep residual and image-based models that report comparable accuracy, and the parameter set occupies a few hundred kilobytes in single-precision form. Every operation in the network, namely dilated one-dimensional convolution, batch normalisation, rectified-linear activation, max-pooling, a softmax over the time axis, and small dense layers, is supported by embedded inference runtimes such as TensorFlow Lite Micro; dilated convolution in particular requires no operator that ordinary convolution does not. The scale-space encoding is a pair of fixed finite-impulse-response filters per scale and can be folded into the stem convolution or precomputed at negligible cost. Because the encoder has no recurrence, all ninety output positions are computed independently, which is the property that determines throughput on hardware without efficient sequential primitives. The model is therefore a realistic candidate for on-device continuous monitoring rather than a workstation-only solution.

4.5 Comparison with recent work

Table 4 places the proposed model against representative methods evaluated on the MIT-BIH Arrhythmia Database under the intra-patient protocol. The comparison is necessarily indirect, because published studies differ in their class grouping, preprocessing, and split ratios, and the entries are quoted as reported in the respective papers. It nonetheless establishes that the present model reaches accuracy at least as high as recent convolutional, recurrent, and time-frequency approaches while using a markedly smaller parameter budget, and that it does so without generating a single synthetic training beat and without reweighting any class.

Table 4. Comparison with representative methods on MIT-BIH

Study	Year	Method	Leads / classes	Accuracy (%)
Scarpiniti [21]	2024	Scalogram-phasogram fusion CNN	1 lead, 5 classes	98.50
Ye et al. [18]	2025	Hybrid CNN-BiLSTM	1 lead, 5 classes	98.87
Elsheikhy et al. [13]	2025	Lightweight multi-lead CNN, 307,669 parameters	2 leads, 5 classes	99.18
Proposed	2026	SSD-TCN, 64,086 parameters	1 lead, 4 classes	**99.36**

5. DISCUSSION

The experiments support three conclusions and one caution.

First, the scale-space derivative encoding is a cheap way to give a convolutional front end an explicitly multi-resolution view of the beat. It costs no trainable parameters, it is a pair of fixed filters per scale, and it generalises the finite-difference encoding common in this literature by making the differentiation scale explicit rather than leaving it fixed at the sampling interval. Because the transform is linear and untrained, it can be adopted independently of the rest of the architecture.

Second, recurrence is not necessary for this task. A stack of four dilated residual blocks reaches a receptive field of 331 samples, which covers the whole beat, and the resulting classifier reaches an accuracy of 99.36 percent with 64,086 parameters. Every output position is computed independently, so throughput on hardware without efficient sequential primitives is limited by arithmetic rather than by the length of the beat. For a model intended to run on an ambulatory device, that is the more useful property.

Third, class imbalance can be handled without touching the data. The training partition is used exactly as recorded, no synthetic beat is generated, and no class is reweighted; the only imbalance-specific machinery is a contrastive term whose projection head is discarded before deployment. This removes the sampling ratio, the neighbourhood size, and the class-weight schedule from the hyperparameter budget, and it removes the awkwardness of training a clinical classifier partly on interpolated waveforms that no heart produced.

The caution concerns the fourth element we expected to need. Prior-based logit adjustment [24] is the treatment that current long-tailed-learning theory recommends for a skew of this magnitude, and it is Fisher-consistent for the balanced error rate, so we implemented it and expected it to help. Section 4.3 shows that on this task it does the opposite, monotonically in its strength, under both the intra-patient and the patient-disjoint protocol. We report this because the result is easy to miss: had we fixed the temperature at the value the literature suggests and never swept it, we would have published a model roughly two macro-F1 points worse than the one reported here and would have attributed part of its performance to the very component that was costing it. Practitioners combining a contrastive objective with a prior correction on skewed physiological data should measure the combination rather than assume it.

Several limitations should be noted. The primary evaluation follows the intra-patient protocol, which is the setting of the comparison literature but which allows beats from a given patient to appear in both the training and the test partition; Section 4.3 quantifies what is lost when that assumption is removed, and the gap should be read as the honest measure of what a deployed system would face on an unseen patient. The fusion class remains the most difficult under either protocol, reflecting both its scarcity and its intrinsic ambiguity as a hybrid of normal and ventricular activation, and under the patient-disjoint protocol neither it nor the supraventricular class is usable at any setting we tried. That limitation is structural rather than incidental: the supraventricular category is defined by the timing of a beat within the surrounding rhythm, and a classifier restricted to the morphology of an isolated window has no access to that information. Adding interval features to the model is the obvious remedy and the first item of future work, but it would change the input contract of the method and is therefore left outside the present study. The model classifies pre-segmented beats and assumes reliable R-peak detection, so integrating a lightweight on-line peak detector into an end-to-end pipeline is a natural next step, as is validation on independent databases and on continuous wearable recordings.

6. CONCLUSION

We have presented SSD-TCN, a compact classifier for single-lead heartbeats that encodes each beat as a Gaussian scale-space derivative stack, processes it with a dilated residual temporal convolutional encoder whose receptive field spans the entire beat, and pools the result with multi-head attentive statistics pooling. Class imbalance is handled in the representation, by a supervised contrastive regulariser added to an unweighted cross-entropy, so that no synthetic beats are generated and no class is reweighted. On the MIT-BIH Arrhythmia Database the model reaches an accuracy of 99.36 percent and a macro-F1 of 0.943 with 64,086 parameters, and an inter-patient evaluation reports what survives a patient-disjoint split. A temperature sweep of the prior-based correction that the long-tailed-learning literature would recommend for this skew shows it to be counterproductive here under both protocols, which is a result worth recording for others working on the same benchmark. The findings indicate that multi-scale signal encoding, a recurrence-free temporal model, and a contrastively regularised objective that leaves the data and the class weights untouched together allow a small network to match much larger ones, which is encouraging for the practical goal of accurate arrhythmia screening on wearable and ambulatory devices. Future work will integrate on-line beat detection, extend the evaluation to independent databases, and study the scale parameters as learnable quantities.

REFERENCES

[1] S. Mahawar, "Recent advances in heart disease prediction from ECG signals: a survey of machine learning, deep learning, and explainable AI approaches," International Journal of Computational Intelligence in Engineering, vol. 1, no. 2, pp. 42-48, 2026.

[2] L. M. A. Khan and U. N. Ranjitha, "Enhancing cardiovascular disease prediction through optimization algorithms: a comparative study of Extra Trees classifier and XGBoost models," in Proceedings of the 2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC), 2024, pp. 1-10.

[3] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH Arrhythmia Database," IEEE Engineering in Medicine and Biology Magazine, vol. 20, no. 3, pp. 45-50, 2001, doi: 10.1109/51.932724.

[4] Association for the Advancement of Medical Instrumentation, "Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms," ANSI/AAMI EC57, 2012.

[5] H. Elgazzar, "Automated ECG arrhythmia classification using convolutional neural networks with effective class imbalance handling," Applied Sciences, vol. 16, no. 11, art. 5321, 2026, doi: 10.3390/app16115321.

[6] Y. Zhang and Y. Xia, "Single-lead ECG arrhythmia classification based on peak-enhanced attention network and quality-aware GAN data augmentation framework," Sensors, vol. 26, no. 12, art. 3852, 2026, doi: 10.3390/s26123852.

[7] K. T. Rajgopal, A. F. J. Alshammari, K. Rashmi, R. Puneeth, S. Gupta, and Y. Ambika Naik, "An automated healthcare framework for single-lead ECG arrhythmia classification using lightweight CNN-BiLSTM and temporal attention," Natural Resources for Human Health, vol. 6, no. 1s, pp. 255-266, 2026.

[8] [8] T. Lindeberg, “Scale-space theory: a basic tool for analysing structures at different scales,” Journal of Applied Statistics, vol. 21, no. 2, pp. 225-270, 1994, doi: 10.1080/757582976.

[9] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” in Proceedings of the International Conference on Learning Representations, 2016.

[10] G. A. Gracy and S. C. Pravin, “A lightweight residual dilated temporal transformer block for ECG classification on edge devices,” Scientific Reports, vol. 16, no. 1, art. 8834, 2026, doi: 10.1038/s41598-026-35531-4.

[11] P. de Chazal, M. O’Dwyer, and R. B. Reilly, “Automatic classification of heartbeats using ECG morphology and heartbeat interval features,” IEEE Transactions on Biomedical Engineering, vol. 51, no. 7, pp. 1196-1206, 2004, doi: 10.1109/TBME.2004.827359.

[12] H. Zakaria, E. S. H. Nurdiniyah, A. M. Kurniawati, D. Naufal, and N. Sutisna, “Morphological arrhythmia classification based on inter-patient and two leads ECG using machine learning,” IEEE Access, vol. 12, pp. 147372-147386, 2024, doi: 10.1109/ACCESS.2024.3469640.

[13] D. H. Elsheikhy, A. S. Hassan, N. M. Yhiea, A. M. Fareed, and E. A. Rashed, “Lightweight deep learning architecture for multi-lead ECG arrhythmia detection,” Sensors, vol. 25, no. 17, art. 5542, 2025, doi: 10.3390/s25175542.

[14] C. Deepti and A. Dammur, “SG-DSN: a lightweight network deep learning model for arrhythmia classification and arrhythmia-induced cardiomyopathy risk alerting using multi-lead ECG,” Signals, vol. 7, no. 4, art. 72, 2026, doi: 10.3390/signals7040072.

[15] U. Thapa, B. M. Pati, A. Taparugssanagorn, and L. Mucchi, “Generalizable hybrid wavelet-deep learning architecture for robust arrhythmia detection in wearable ECG monitoring,” Sensors, vol. 25, no. 21, art. 6590, 2025, doi: 10.3390/s25216590.

[16] M. E. Ait Bourkha, A. Hatim, D. Nasir, and S. El Beid, “Optimized wavelet scattering network and CNN for ECG heartbeat classification from MIT-BIH arrhythmia database,” International Journal of Advanced Computer Science and Applications, vol. 16, no. 2, 2025, doi: 10.14569/IJACSA.2025.01602116.

[17] M. Kolhar and R. N. A. Kazi, “ABMA: an attention-based morphology-aware framework for automated 12-lead ECG arrhythmia classification,” Diagnostics, vol. 16, no. 14, art. 2274, 2026, doi: 10.3390/diagnostics16142274.

[18] Y. Ye, K. Chipusu, M. A. Ashraf et al., “Hybrid CNN-BLSTM architecture for classification and detection of arrhythmia in ECG signals,” Scientific Reports, vol. 15, art. 34510, 2025, doi: 10.1038/s41598-025-17671-1.

[19] Manisha, K. A. Kotegar, and S. Bilakeri, “CaReS-BiNet: a multi-scale deep learning framework for ECG arrhythmia classification,” Scientific Reports, vol. 16, art. 58810, 2026, doi: 10.1038/s41598-026-58810-6.

[20] R. Bahrami and A. M. Fotouhi, “A novel efficient hybrid deep learning framework for ECG-based heartbeat arrhythmia classification,” Neural Computing and Applications, vol. 37, no. 21, pp. 16409-16425, 2025, doi: 10.1007/s00521-025-11319-x.

[21] M. Scarpiniti, “Arrhythmia detection by data fusion of ECG scalograms and phasograms,” Sensors, vol. 24, no. 24, art. 8043, 2024, doi: 10.3390/s24248043.

[22] P. K. Sekharamantry, P. MeganaSanthoshi, S. Gupta, A. M, K. L. Divya, and S. P. M., “CAF-Net: a convolution and attention fusion network with RR interval context for five-class arrhythmia detection from single-lead ECG,” Natural Resources for Human Health, vol. 6, no. 2s, pp. 452-465, 2026.

[23] K. Okabe, T. Koshinaka, and K. Shinoda, “Attentive statistics pooling for deep speaker embedding,” in Proceedings of Interspeech, 2018, pp. 2252-2256, doi: 10.21437/Interspeech.2018-993.

A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, “Long-tail learning via logit adjustment,” in Proceedings of the International Conference on Learning Representations, 2021.

[24] P. Khosla et al., “Supervised contrastive learning,” in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 18661-18673.

[25] D. Le, S. Truong, P. Brijesh, D. A. Adjeroh, and N. Le, “sCL-ST: supervised contrastive learning with semantic transformations for multiple lead ECG arrhythmia classification,” IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 6, pp. 2818-2828, 2023, doi: 10.1109/JBHI.2023.3246241.

[26] K. Ma, T. Zhang, H. Zhang, and W. Huang, “Self-supervised contrastive learning achieves 12-lead ECG classification,” Biomedical Signal Processing and Control, vol. 112, art. 108420, 2026, doi: 10.1016/j.bspc.2025.108420.

[27] Y. Chen, Z. Huang, and Z. Feng, “Advancing long-tailed pediatric arrhythmia classification with a novel contrastive loss and multimodal learning,” Biomedical Signal Processing and Control, vol. 127, art. 111077, 2026, doi: 10.1016/j.bspc.2026.111077.

[28] R. Bahrami and A. M. Fotouhi, “Investigation of inter-patient, intra-patient, and patient-specific based training in deep learning for classification of heartbeat arrhythmia,” Cardiovascular Engineering and Technology, vol. 16, no. 3, pp. 333-342, 2025, doi: 10.1007/s13239-025-00777-y.

[29] C. Zhou et al., “Inter-patient ECG heartbeat classification for arrhythmia classification: a new approach of multi-layer perceptron with weight capsule and sequence-to-sequence combination,” Frontiers in Physiology, vol. 14, art. 1247587, 2023, doi: 10.3389/fphys.2023.1247587.