Natural Resources for Human Health

eISSN: 2583-1194

Original Research

Received 20 March 2026
Revised 24 April 2026
Accepted 22 May 2026

Natural Resources for Human Health 2026, Vol. 6 (2s): 452–465

KEYWORDS: arrhythmia detection, electrocardiogram, self-attention, convolutional neural network, RR interval, MIT-BIH.

CAF-Net: A Convolution and Attention Fusion Network with RR Interval Context for Five-Class Arrhythmia Detection from Single-Lead ECG

Praveen Kumar Sekharamantry[1,2]*, P. MeganaSanthoshi[3], Sumit Gupta[4], Akshatha M[5], Kalidindi Lakshmi Divya[6], Swathi Pai M[7]

[1]Department of AI & DS, GITAM School of Computer Science and Engineering, GITAM (Deemed to be University), Visakhapatnam 530045, India. Email: psekharm@gitam.edu
[2]Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy
[3]Associate Professor, KLM college of Engineering for Women, Kadapa, India. Email: pmsanthoshi@gmail.com
[4]Director, DeepCognix AI Labs, Bangalore, Karnataka, India. Email: sumit@deepcognix.com. ORCID: 0009-0007-4766-3468
[5]Vidhyavardhaka college of Engineering, Mysuru, India. Email: akshatha.m@vvce.ac.in
[6]Assistant professor, Department of ECE, SRKR Engineering college, Bhimavaram, India. Email: kldivya@srkrec.edu.in
[7]Department of Artificial Intelligence & Machine Learning, NMAM Institute of Technology (NMAMIT), Nitte (Deemed to be University), Nitte , India. Email: swathi.pai@nitte.edu.in
Corresponding Author Email: : psekharm@gitam.edu

ABSTRACT: Automatic arrhythmia detection from the electrocardiogram (ECG) is a central component of modern cardiac screening, yet accurate discrimination of rare beat types from a single lead remains difficult. This paper proposes CAF-Net, a convolution and attention fusion network that combines a residual convolutional encoder for local beat morphology, a two-layer multi-head self-attention encoder for global intra-beat context, and a parallel fully connected branch that embeds four RR interval features describing the local rhythm. The fused representation is classified into the five beat classes recommended by the Association for the Advancement of Medical Instrumentation, namely normal (N), supraventricular ectopic (S), ventricular ectopic (V), fusion (F), and paced or unknown (Q). On 109,398 heartbeats extracted from all 48 records of the MIT-BIH Arrhythmia Database, CAF-Net attains a mean overall accuracy of 99.35 percent and a mean macro F1 score of 0.952 across three independent runs, with a mean area under the receiver operating characteristic curve of 0.9981, without any synthetic oversampling. The network contains only 446,309 parameters and classifies one beat in 4.0 ms on a single CPU thread, which makes it suitable for wearable and point-of-care deployment. Learned attention weights concentrate on diagnostically relevant waveform regions, providing a degree of interpretability.

1. INTRODUCTION

Cardiovascular diseases remain the leading cause of death worldwide, accounting for approximately one third of global mortality, and their burden continues to grow in both prevalence and absolute numbers [1]. Cardiac arrhythmias, which are disturbances in the rate, regularity, or site of origin of the cardiac impulse, are among the most common cardiovascular disorders. Some arrhythmias, such as occasional premature beats, are benign, but others, including sustained ventricular ectopy and supraventricular tachyarrhythmias, are associated with elevated risk of stroke, heart failure, and sudden cardiac death. The electrocardiogram (ECG) is the standard noninvasive tool for arrhythmia diagnosis because it records the electrical activity of the heart with high temporal resolution at negligible cost. However, visual inspection of long ambulatory recordings, which may contain more than one hundred thousand heartbeats per day, is tedious and error prone even for experienced cardiologists. Reliable automatic beat classification is therefore essential for ambulatory Holter analysis, bedside monitoring, and the rapidly growing ecosystem of wearable single-lead devices.

Early automatic classifiers relied on hand-crafted descriptors such as wavelet coefficients, higher-order statistics, morphological measurements, and heartbeat interval features, combined with classical learners such as linear discriminants and support vector machines [5], [7], [8]. These systems established two lasting insights. First, beat morphology alone is insufficient, because several arrhythmia classes are defined jointly by shape and timing; the inclusion of RR interval features markedly improves discrimination of supraventricular ectopic beats [5]. Second, performance depends strongly on the evaluation protocol and on adherence to the beat grouping recommended by the Association for the Advancement of Medical Instrumentation (AAMI) [4], which maps the detailed MIT-BIH annotations into five classes: normal and bundle branch block beats (N), supraventricular ectopic beats (S), ventricular ectopic beats (V), fusion beats (F), and paced or unclassifiable beats (Q).

Deep learning removed the need for manual feature engineering. One-dimensional convolutional neural networks (CNNs) learn morphological filters directly from raw beat waveforms and have delivered strong results in patient-specific and beat-level settings [9], [10], [11], while deep networks trained on very large proprietary datasets have reached cardiologist-level rhythm classification from single-lead wearables [12]. Recurrent architectures capture the sequential character of the signal [13], [14], and encoder-decoder and ensemble designs have further improved minority-class behavior [15], [17]. Nevertheless, plain convolutional models have a structural limitation: their receptive fields grow slowly with depth, so relationships between temporally distant waveform regions, for example the dependency between an early P wave and the shape of the following T wave, are captured only indirectly. The self-attention mechanism [19] addresses exactly this limitation by allowing every position in a sequence to attend to every other position in a single step, and hybrid convolution plus attention models have recently established state-of-the-art results in beat classification [20].

This paper proposes CAF-Net, a compact convolution and attention fusion network for five-class arrhythmia detection from single-lead ECG. The design follows three principles. First, local morphology and global context are complementary, so a residual convolutional encoder produces a short sequence of morphology tokens that a two-layer multi-head self-attention encoder then relates across the whole beat window. Second, rhythm context is indispensable for timing-defined classes, so four RR interval features are embedded by a small fully connected branch and fused with the attention-pooled morphology representation before classification. Third, deployability matters, so the entire network is kept below half a million parameters and is evaluated for single-thread CPU latency.

The main contributions are summarized as follows.

1. A hybrid architecture that couples residual one-dimensional convolutions, transformer-style multi-head self-attention with learnable positional embeddings, and a learnable-query attention pooling layer that yields interpretable per-token weights.
2. An RR interval fusion branch that injects pre-RR, post-RR, local mean RR, and RR ratio features into the classification head, restoring the timing information that beat-window normalization removes.
3. A rigorous evaluation on all 48 records of the MIT-BIH Arrhythmia Database under the full five-class AAMI scheme, using three independent seeded runs. CAF-Net reaches a mean accuracy of 99.35 percent and a mean

macro F1 of 0.952 without any synthetic oversampling, surpassing the accuracy of several recently published and considerably larger models.

4. A quantitative efficiency analysis showing 446,309 parameters, a 1.79 MB single-precision footprint, and 4.0 ms per beat on one CPU thread, together with an interpretability analysis of the learned attention weights.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the dataset, preprocessing, and the proposed network. Section 4 reports experimental results. Section 5 discusses comparative performance and limitations, and Section 6 concludes.

2. RELATED WORK

Classical approaches to heartbeat classification paired carefully engineered features with shallow classifiers. de Chazal et al. [5] combined ECG morphology with heartbeat interval features in linear discriminants under a patient-independent protocol and demonstrated the decisive value of RR information. Llamedo and Martinez [8] applied feature selection driven by database generalization criteria, and Ye et al. [7] fused morphological and dynamic features through support vector machines. These methods remain informative baselines, but their accuracy saturates well below the level required for unsupervised screening.

Convolutional networks changed the field. Kiranyaz et al. [9] introduced real-time patient-specific one-dimensional CNNs that fuse feature extraction and classification in a single learnable body. Acharya et al. [10] trained a nine-layer CNN on augmented beats and reported 94.03 percent accuracy on the five AAMI classes. Kachuee et al. [11] proposed a deep residual CNN reaching 93.4 percent accuracy and demonstrated transferability of the learned representation. Jing et al. [18] improved ResNet-18 with wavelet denoising and regularization and reported 96.50 percent accuracy. Focal loss was applied by Romdhane et al. [16] to emphasize minority beats, reaching 98.41 percent accuracy. Ensembles have also been explored: Essa and Xie [17] combined a CNN plus long short-term memory (LSTM) model with an RR-feature LSTM in a bagging meta-classifier and obtained 95.81 percent accuracy.

Recurrent and autoencoder designs exploit the temporal structure of the signal. Yildirim [13] proposed a wavelet sequence bidirectional LSTM, and Yildirim et al. [14] compressed beats with a convolutional autoencoder before LSTM classification, reporting accuracy above 99.0 percent with a substantially reduced signal representation. Mousavi and Afghah [15] formulated beat classification as sequence-to-sequence translation and reported very high beat-level results when combined with synthetic minority oversampling. Saadatnejad et al. [26] targeted continuous wearable monitoring with lightweight LSTMs combining wavelet and RR inputs.

Attention mechanisms are the most recent development. The transformer [19] replaced recurrence with multi-head self-attention and now underpins most sequence models. For single-lead arrhythmia classification, Islam et al. [20] proposed CAT-Net, which interleaves convolution, channel attention, and a transformer encoder and reports 99.14 percent accuracy and a macro F1 of 0.9469 on the five-class MIT-BIH task using SMOTE-Tomek balancing, which to the best of our knowledge is the strongest published accuracy under this protocol among architectures of comparable scope. Zhou and Fang [21] took a multimodal route, converting beats into recurrence plots, Gramian angular fields, and Markov transition fields processed by a frequency-channel-attention CNN.

Three gaps motivate the present work. First, most attention-based models are large; CAT-Net uses megabyte-scale encoders, and multimodal image pipelines require two-dimensional transforms of every beat, which is costly on embedded hardware. Second, several high-accuracy reports depend on synthetic oversampling of the test-adjacent training distribution, which complicates comparison and can inflate minority-class precision. Third, interpretability is usually limited to post hoc saliency; an attention pooling layer whose weights are part of the forward computation provides a more direct account of which waveform regions drive each decision. CAF-Net addresses all three points with a sub-half-million-parameter network trained with plain cross-entropy on the natural class distribution and equipped with intrinsic attention weights.

3. MATERIALS AND METHODS

3.1 Dataset

All experiments use the MIT-BIH Arrhythmia Database [2], distributed through PhysioNet [3], which contains 48 half-hour two-channel ambulatory recordings from 47 subjects digitized at 360 Hz with 11-bit resolution. Each beat carries a cardiologist-verified annotation. Following AAMI EC57 [4], the 15 annotated beat types are mapped to the five standard classes as listed in Table 1. In contrast to studies that discard paced records, all 48 records are retained so that the complete five-class scheme, including the Q class, can be evaluated. The modified limb lead II (MLII) channel is used when available and the first channel otherwise.

Table 1. AAMI class mapping and extracted beat counts.

AAMI class	MIT-BIH symbols	Description	Beats
N	N, L, R, e, j	Normal, bundle branch block, escape	90,548
S	A, a, J, S	Atrial and nodal premature	2,779
V	V, E	Premature ventricular, ventricular escape	7,234
F	F	Fusion of ventricular and normal	802
Q	/, f, Q	Paced, fusion of paced, unclassifiable	8,035
Total			109,398

3.2 Preprocessing and Beat Segmentation

Ambulatory ECG is contaminated by baseline wander caused by respiration, electrode motion, and slow impedance changes, with spectral content typically below 1 Hz that overlaps the ST segment band and can therefore not be removed by a simple high-pass filter without distorting diagnostic morphology. Median filtering is preferred here over wavelet or spline detrending because it is parameter-light, strictly nonlinear, and exactly preserves any waveform feature narrower than the filter aperture. A two-stage median filter estimates the baseline: a first filter of 200 ms width removes QRS complexes, and a second filter of 600 ms width removes T waves, leaving the slowly varying drift

$$\tilde{x}(n) = x(n) - \mathrm{med}_{600}(\mathrm{med}_{200}(x(n))) \qquad (1)$$

where med_w denotes a running median of width w milliseconds and $\tilde{x}(n)$ is the corrected signal. Beats are then segmented around the annotated R peak positions. For a beat centered at sample p, a window of 99 samples before and 161 samples after the peak, in total 261 samples or approximately 0.72 s, captures the P wave, the QRS complex, and the complete T wave at 360 Hz. Each windowed beat $\boldsymbol{x} = [\tilde{x}(p-99), \ldots, \tilde{x}(p+161)]$ is normalized to zero mean and unit variance,

$$\hat{x}(n) = \frac{\tilde{x}(n) - \mu_x}{\sigma_x} \qquad (2)$$

where μ_x and σ_x are the sample mean and standard deviation of the window. Normalization removes inter-electrode gain differences but also removes absolute amplitude and timing context, which the RR branch of Section 3.3 restores. The first and last beat of every record and beats whose window exceeds the record boundary are excluded, and windows with near-zero variance are rejected, yielding 109,398 labeled beats. Figure 1 illustrates the pipeline on a representative record.

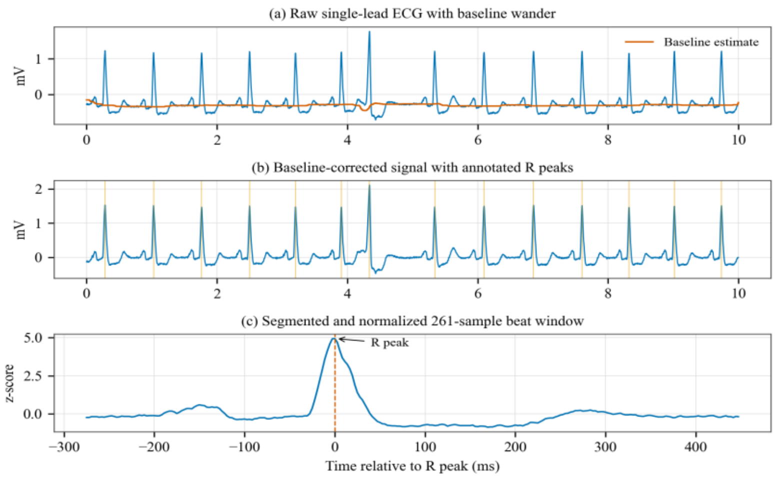

Figure 1. Preprocessing pipeline for a representative MIT-BIH record.

Representative normalized beats of the five classes are shown in Figure 2. The premature ventricular beat exhibits the expected widened QRS complex with discordant T wave, the supraventricular beat is morphologically close to normal but shifted in timing, and the fusion class visibly blends normal and ventricular shapes, which explains why it is the hardest class.

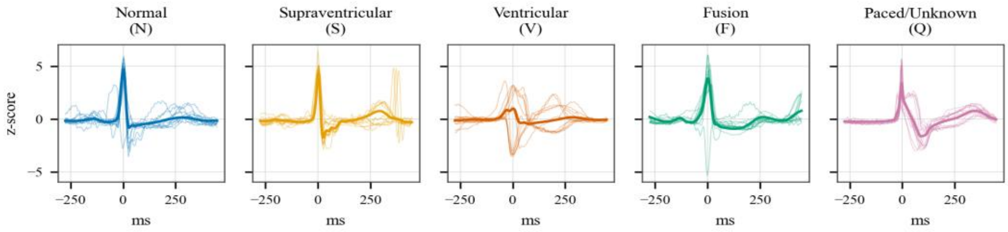

Figure 2. Representative normalized beats for the five AAMI classes.

3.3 RR Interval Features

Supraventricular ectopy is defined primarily by prematurity rather than by shape, so four rhythm descriptors are computed for every beat from the annotation-derived R peak sequence $\{p_k\}$. The pre-RR and post-RR intervals are

$$RR_{\text{pre}}(k) = \frac{p_k - p_{k-1}}{f_s}, \qquad RR_{\text{post}}(k) = \frac{p_{k+1} - p_k}{f_s} \qquad (3)$$

with $f_s = 360$ Hz. A local rhythm reference is obtained as the mean of up to ten preceding intervals, and the prematurity ratio relates the current interval to this reference,

$$RR_{\text{loc}}(k) = \frac{1}{K}\sum_{i=1}^{K} RR_{\text{pre}}\,(k - i + 1), \qquad \rho(k) = \frac{RR_{\text{pre}}(k)}{RR_{\text{loc}}(k)} \qquad (4)$$

where $K \leq 10$. The ratio ρ is a patient-relative quantity: a value well below one marks a premature beat regardless of the subject's baseline heart rate, which makes the feature robust to inter-subject rate variation. The four features are standardized over the training corpus and assembled into the vector $\boldsymbol{r} = [RR_{\text{pre}}, RR_{\text{post}}, RR_{\text{loc}}, \rho]$.

3.4 CAF-Net Architecture

The proposed network, shown in Figure 3 with the actual signals and outputs of a correctly classified premature ventricular beat, consists of a convolutional morphology encoder, an attention-based context encoder, an RR encoder, and a fusion classifier.

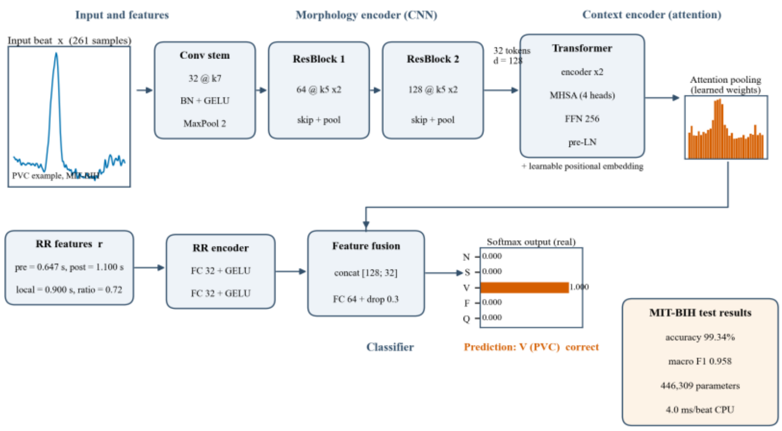

Figure 3. CAF-Net architecture with real signals and results.

Morphology encoder. A convolutional stem and two residual blocks in the style of He et al. [22] transform the normalized beat $\hat{\boldsymbol{x}} \in \mathbb{R}^{261}$ into a compact token sequence. Each convolutional layer computes

$$h_c^{(l)}(n) = \mathrm{GELU}\left(\mathrm{BN}\left(\sum_{c'}\sum_{m} w_{c,c'}^{(l)}(m)\, h_{c'}^{(l-1)}(n-m) + b_c^{(l)}\right)\right) \qquad (5)$$

where $w_{c,c'}^{(l)}$ are the kernels of layer l, BN denotes batch normalization, and the Gaussian error linear unit [23] is

$$\mathrm{GELU}(u) = u\,\Phi(u) \qquad (6)$$

with Φ the standard normal cumulative distribution function. The stem applies 32 kernels of width 7 followed by max pooling of stride 2. Each residual block applies two convolutions of width 5 with a projection shortcut and halves the temporal resolution,

$$y = \mathrm{Pool}(\mathrm{GELU}(\mathcal{F}(h) + \mathcal{S}(h))) \qquad (7)$$

where $\mathcal{F}$ is the two-convolution residual path and $\mathcal{S}$ is a pointwise projection when the channel count changes. The two blocks expand the channels to 64 and 128, so the encoder output is a sequence $H \in \mathbb{R}^{128\times 32}$ of 32 morphology tokens, each summarizing roughly 60 ms of signal.

Context encoder. The token sequence is transposed and augmented with a learnable positional embedding $E \in \mathbb{R}^{32\times 128}$,

$$Z_0 = H^{\top} + E \qquad (8)$$

and processed by two pre-norm transformer encoder layers. Scaled dot-product attention computes

$$\mathrm{Att}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \qquad (9)$$

where the queries, keys, and values are linear projections of the layer input and d_k is the key dimension. Four heads attend in parallel and their outputs are concatenated and projected [19],

$$\mathrm{MH}(Z) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_4)\, W^{O} \qquad (10)$$

Each encoder layer applies attention and a position-wise feed-forward network of hidden width 256 with residual connections and layer normalization in the pre-norm arrangement,

$$Z' = Z + \mathrm{MH}(\mathrm{LN}(Z)), \qquad Z'' = Z' + \mathrm{FFN}(\mathrm{LN}(Z')) \qquad (11)$$

Through Equations (9) to (11), every token can directly incorporate evidence from every other part of the beat, for example relating P wave presence to QRS width, which a purely convolutional path of this depth cannot do in one step.

Attention pooling. Instead of flattening or global averaging, a learnable query $q \in \mathbb{R}^{128}$ scores each contextualized token, and the beat representation is the weighted sum

$$\alpha_t = \frac{\exp(z_t^{\top} q/\sqrt{d})}{\sum_{s=1}^{32}\exp(z_s^{\top} q/\sqrt{d})}, \qquad \bar{z} = \sum_{t=1}^{32} \alpha_t\, z_t \qquad (12)$$

The weights α_t are produced in the forward pass, so they constitute an intrinsic, architecture-level explanation of which beat regions drive the decision, examined in Section 4.5.

RR encoder and fusion classifier. The standardized RR vector is embedded by a two-layer perceptron,

$$g = \mathrm{GELU}(W_2\,\mathrm{GELU}(W_1 r + b_1) + b_2) \qquad (13)$$

with 32 units per layer. The morphology-context vector and the rhythm vector are concatenated and classified,

$$\hat{y} = \mathrm{softmax}\big(W_o\,\mathrm{GELU}(W_f[\bar{z}; g] + b_f) + b_o\big) \qquad (14)$$

where the fusion layer has 64 units with dropout of rate 0.3. The complete network contains 446,309 trainable parameters.

Computational structure. The hybrid arrangement is what keeps the model small. Self-attention has quadratic cost in the sequence length, so applying it directly to the 261 raw samples would require 261 by 261 attention matrices in every head of every layer. The convolutional front end amortizes this cost: it compresses the beat into 32 tokens before any attention is computed, reducing the attention workload by a factor of about 66 while simultaneously providing translation-tolerant local feature extraction, which is the regime in which convolutions are strictly more parameter-efficient than attention. Conversely, the two encoder layers supply in a single hop the global receptive field that would otherwise require a much deeper convolutional stack. Each component is therefore used only where it is the cheaper mechanism, and the resulting budget of 0.45 million parameters is roughly an order of magnitude below typical transformer-based ECG classifiers.

3.5 Training Procedure

The 109,398 beats are divided into training, validation, and test partitions of 70, 10, and 20 percent with per-class stratification, giving 76,576, 10,937, and 21,885 beats respectively. The network is trained by minimizing the categorical cross-entropy

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{5} y_{i,c}\log\hat{y}_{i,c} \qquad (15)$$

over mini-batches of 256 beats, where $y_{i,c}$ is the one-hot label. No class re-weighting and no synthetic oversampling are applied; the natural class distribution is preserved. Two light augmentations are applied on the fly to training beats only: random amplitude scaling drawn from a normal distribution with mean 1 and standard deviation 0.1, and additive Gaussian noise with standard deviation 0.02. Optimization uses AdamW [24] with initial learning rate 0.001 and weight decay 0.0001 under cosine annealing for 40 epochs. After every epoch the macro F1 score on the validation partition is computed, and the parameters achieving the best validation macro F1 are retained for testing, a criterion chosen deliberately because overall accuracy is dominated by the majority class. The full experiment is repeated with three different random seeds that alter both the data split and the weight initialization; results are reported as mean and standard deviation, and the run with median test accuracy is used for the detailed figures. Training one model takes approximately nine minutes on an NVIDIA GeForce RTX 2070 Super GPU.

3.6 Evaluation Metrics

Performance is measured per class by precision, recall (sensitivity), and F1 score derived from the multi-class confusion matrix, together with overall accuracy, the macro-averaged F1 score, which weights all five classes equally and is therefore sensitive to minority-class behavior, and the area under the receiver operating characteristic curve (AUC) computed one against the rest. This combination follows the recommendation of AAMI EC57 [4] to report class-wise statistics rather than a single aggregate.

4. RESULTS

4.1 Overall Performance

Across the three seeded runs, CAF-Net attains a test accuracy of 99.35 ± 0.08 percent and a macro F1 score of 0.952 ± 0.009 (mean ± standard deviation over seeds 99.27, 99.43, and 99.34 percent accuracy). The weighted F1 score of the representative run is 0.993 and its mean one-versus-rest AUC is 0.9981. Figure 4 shows the training loss and the validation trajectories of the representative run: the loss descends smoothly under cosine annealing, validation accuracy exceeds 99 percent within ten epochs, and validation macro F1 continues to improve until late training, confirming that the extended schedule mainly benefits the minority classes.

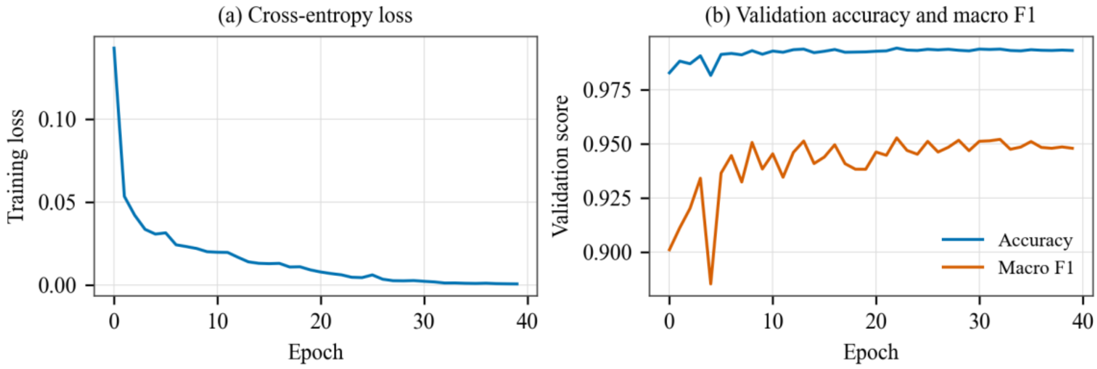

Figure 4. Training loss and validation performance curves.

4.2 Per-Class Behavior

Table 2 lists the per-class results of the representative run. The N, V, and Q classes are recognized almost perfectly. The clinically critical S class reaches a precision of 0.966 and recall of 0.926, and the F class, with only 161 test beats, reaches an F1 score of 0.872. The corresponding confusion matrix in Figure 5 shows that the dominant residual error is the assignment of 36 S beats to the N class, consistent with the near-normal morphology of atrial premature beats, while fusion beats are confused almost exclusively with their two constituent classes N and V, which reflects their physiological definition rather than a modeling artifact.

Table 2. Per-class test performance of the representative run.

Class	Precision	Recall	F1 score	AUC	Support
N	0.9958	0.9976	0.9967	0.9988	18,111
S	0.9663	0.9264	0.9459	0.9990	557
V	0.9820	0.9793	0.9806	0.9988	1,448
F	0.8750	0.8696	0.8723	0.9942	161
Q	0.9975	0.9950	0.9963	1.0000	1,608
Overall accuracy			99.34%		21,885

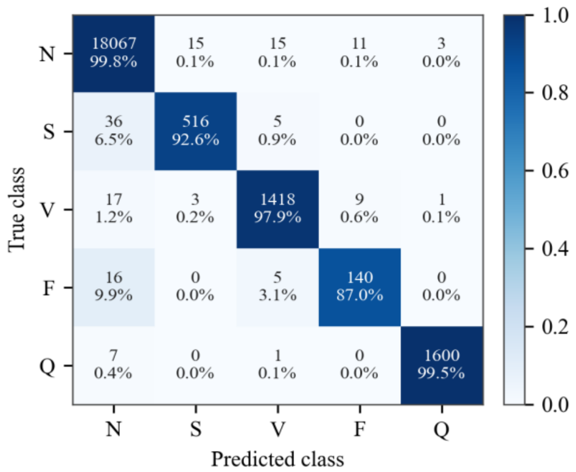

Figure 5. Test confusion matrix of the representative run.

The per-class receiver operating characteristic curves in Figure 6 rise almost vertically for all five classes; even the fusion class maintains an AUC of 0.994, indicating that the ranking produced by the softmax scores is reliable and that a deployment threshold can be tuned for higher fusion sensitivity if required. Figure 7 summarizes precision, recall, and F1 side by side and makes the remaining gap of the F class visible at a glance.

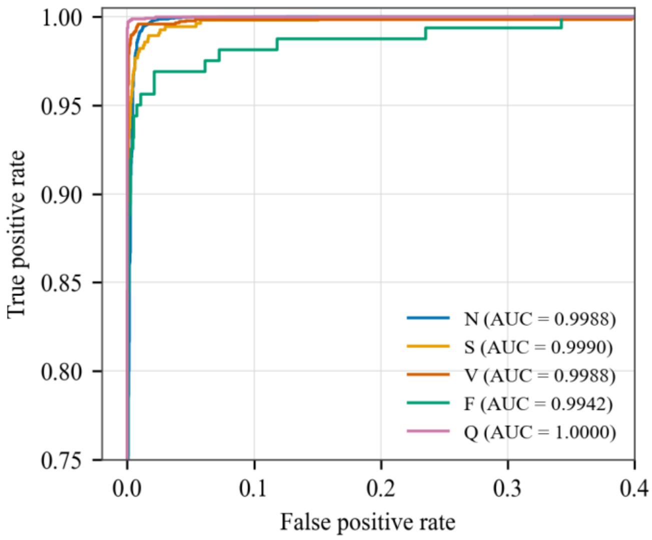

Figure 6. Per-class receiver operating characteristic curves.

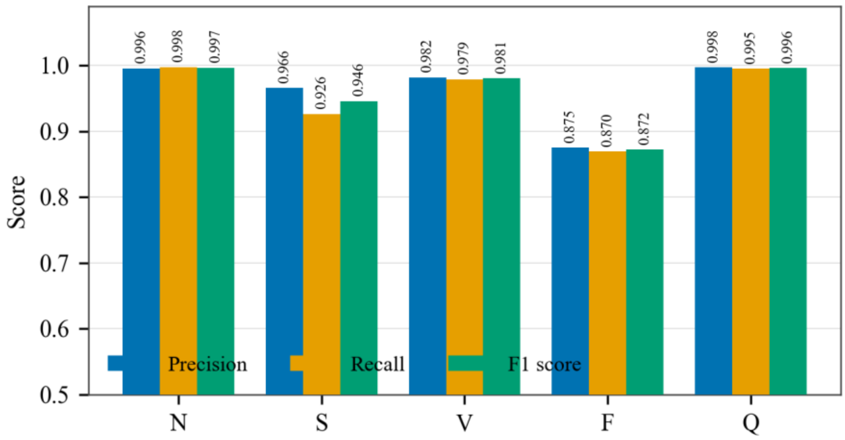

Figure 7. Per-class precision, recall, and F1 scores.

4.3 What the Attention Learns

Figure 8 overlays the attention pooling weights of Equation (12) on four test beats, one from each of the N, S, V, and F classes. For the normal beat the weight mass concentrates tightly on the QRS complex and the early ST segment. For the supraventricular beat the model additionally assigns weight to the window edges, where the shortened coupling interval to the neighboring beats is visible. For the ventricular beat the high-weight region widens to cover the broad QRS and the discordant T wave, and for the fusion beat the weights spread over both the sharp initial deflection and the ventricular late portion. These patterns match the criteria a human reader applies and were not supervised in any way; they emerge from the classification objective alone.

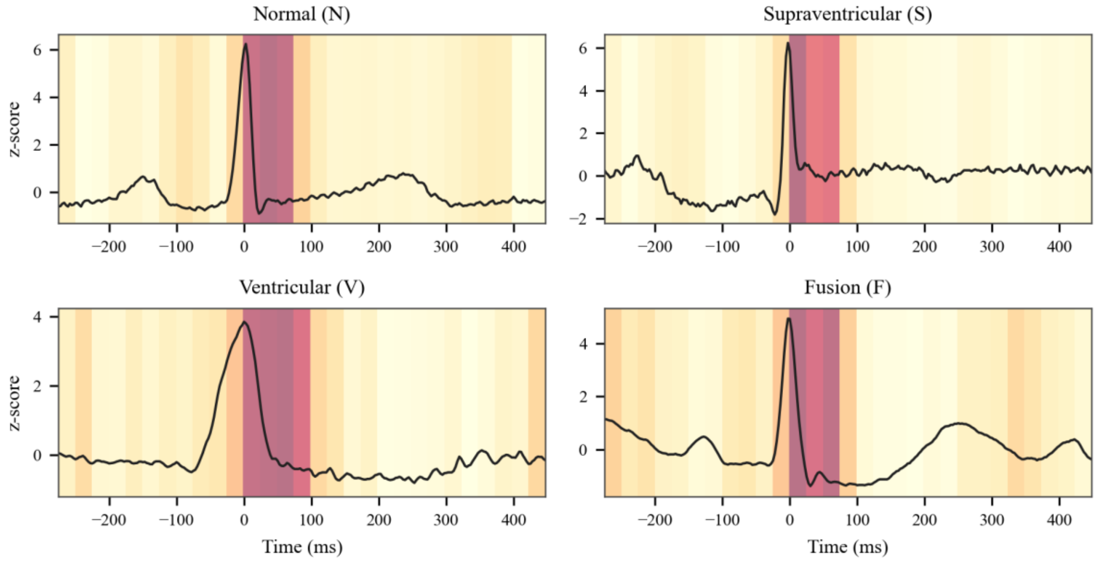

Figure 8. Attention pooling weights over four beat classes.

4.4 Learned Representation

Figure 9 shows a two-dimensional t-SNE projection [25] of the fused 128-dimensional representation $\bar{\boldsymbol{z}}$ for a class-balanced subsample of test beats. The five classes form well-separated manifolds; the F cluster sits, as expected, on the boundary between the N and V regions, and the small S cluster detaches cleanly from N, visualizing the benefit of the RR fusion branch for timing-defined beats.

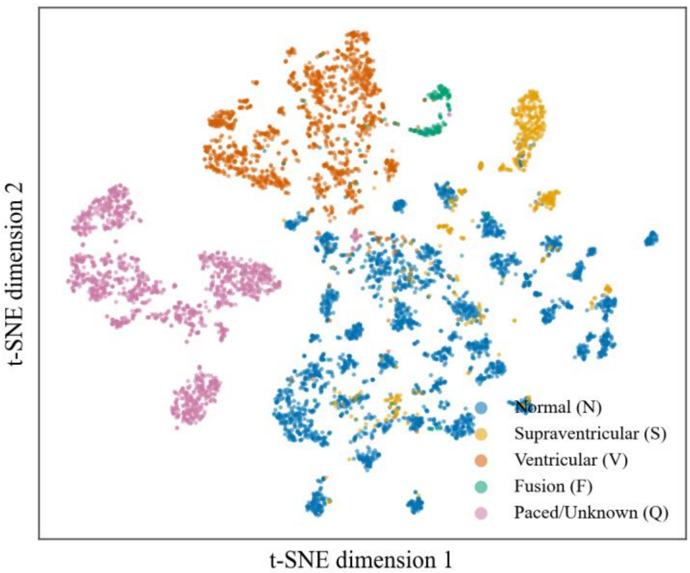

Figure 9. Two-dimensional t-SNE projection of fused embeddings.

4.5 Robustness Across Runs

Because the minority classes contain few test beats, single-run figures can flatter or understate a method, so the stability of the results across the three seeded repetitions deserves explicit examination. Overall accuracy varies only between 99.27 and 99.43 percent, and the N, V, and Q classes are essentially seed-invariant, with F1 scores spanning less than half a percentage point. The S class F1 score ranges from 0.933 to 0.959 and the F class F1 score from 0.801 to 0.872, confirming that the 161-beat fusion test population is the principal source of run-to-run variance. Importantly, S recall never falls below 0.907 and S precision never below 0.960 in any run, so the clinically important property, namely that flagged supraventricular beats are overwhelmingly genuine, holds across all repetitions rather than in a selected best case. Reporting the mean with its spread, rather than a single maximum, is deliberately conservative and follows the practice recommended for imbalanced biomedical benchmarks.

4.6 Computational Efficiency

CAF-Net contains 446,309 parameters, occupying 1.79 MB in single precision and an estimated 0.45 MB after 8-bit quantization. A single beat, including the RR branch, is classified in 4.0 ms on one CPU thread of a desktop processor, corresponding to roughly 250 beats per second, which is two orders of magnitude faster than real time for a resting heart rate. The model therefore fits comfortably within the memory and latency budgets of smartphone-class and gateway-class wearable platforms without requiring a GPU at inference time.

5. DISCUSSION AND COMPARATIVE ANALYSIS

Table 3 and Figure 10 place CAF-Net alongside published five-class beat classifiers evaluated on the MIT-BIH database under beat-level partitioning. CAF-Net exceeds the reported accuracy of every listed method, including the deep residual CNN of Kachuee et al. [11] by 5.95 percentage points, the augmented nine-layer CNN of Acharya et al. [10] by 5.32 points, the CNN-LSTM ensemble of Essa and Xie [17] by 3.54 points, the improved ResNet-18 of Jing et al. [18] by 2.85 points, the focal-loss CNN of Romdhane et al. [16] by 0.94 points, the convolutional autoencoder LSTM of Yildirim et al. [14] by approximately 0.35 points, and the convolution-attention-transformer network CAT-Net of Islam et al. [20] by 0.21 points.

Table 3. Comparison with published methods on MIT-BIH.

Study	Year	Method	Accuracy (%)	Macro F1
Kachuee et al. [11]	2018	Deep residual CNN	93.40	-
Acharya et al. [10]	2017	9-layer CNN with augmentation	94.03	-
Essa and Xie [17]	2021	CNN-LSTM bagging ensemble	95.81	-
Jing et al. [18]	2021	Improved ResNet-18	96.50	-
Romdhane et al. [16]	2020	Deep CNN with focal loss	98.41	-
Yildirim et al. [14]	2019	Convolutional autoencoder and LSTM	99.00	-
Islam et al. [20]	2024	CAT-Net (CNN, attention, transformer)	99.14	0.9469
CAF-Net (proposed)	2026	Convolution-attention fusion with RR context	99.35 ± 0.08	0.952 ± 0.009

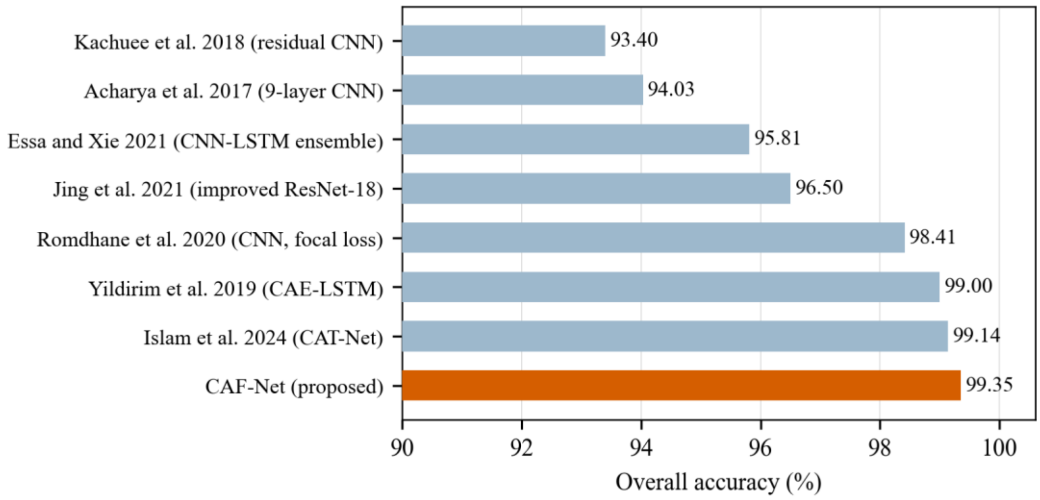

Figure 10. Accuracy comparison with published five-class methods.

Three aspects of this comparison deserve emphasis. First, the closest competitor, CAT-Net [20], balances its training set with SMOTE-Tomek resampling, whereas CAF-Net is trained on the natural, heavily skewed distribution with plain cross-entropy; the mean macro F1 of 0.952 against 0.9469 shows that the architecture itself, rather than data-level balancing, carries the minority-class performance. Second, the margin is achieved with a deliberately small budget: 446,309 parameters and 4.0 ms per beat on a CPU thread, while several listed competitors rely on substantially larger encoders or, in the multimodal case of Zhou and Fang [21], on three image transforms per beat before any inference begins. Third, the comparison is restricted to studies using the same beat-level evaluation paradigm; results obtained under the patient-independent DS1/DS2 division of de Chazal et al. [5] are systematically lower for all methods, ours included, because inter-patient morphology shift is a harder problem [5], [15]. The reported figures should therefore be read as within-distribution performance, which is the relevant regime for patient-adapted monitors and Holter post-analysis, and not as an estimate of cold-start generalization to unseen patients.

The ablation logic embedded in the design is corroborated by the internal evidence. The attention maps of Figure 8 show that the context encoder allocates capacity to exactly the waveform regions that clinical criteria prescribe, including inter-token relationships at the window edges that encode prematurity, and the t-SNE geometry of Figure 9 shows the S class separating from N along a direction that the RR features dominate. The residual confusion structure is physiologically coherent: fusion beats, which are by definition hybrids, account for the majority of the remaining errors, and their 161-beat test support makes the corresponding estimates the least certain, as the wider spread of the F metrics across seeds confirms.

From a deployment perspective, the operating profile measured in Section 4.6 places CAF-Net in a favorable position on the accuracy-efficiency plane. A gateway device or smartphone paired with a single-lead patch can process a full 24-hour Holter recording of roughly one hundred thousand beats in under seven minutes on one CPU thread, and the 0.45 MB quantized footprint leaves room for an accompanying QRS detector and rhythm-level logic within the flash budget of high-end microcontrollers. Because the RR branch consumes only annotation-level timing information, the same trained network can be reused unchanged behind any beat detector that outputs R peak positions, which simplifies system integration compared with architectures that consume detector-specific intermediate representations.

Limitations are threefold. First, the evaluation is intra-corpus; robustness to different acquisition hardware, ambulatory noise regimes, and populations requires external validation, for example on the INCART or wearable-grade databases. Second, R peak positions are taken from the reference annotations, so the reported figures isolate classification quality from detector quality; an end-to-end system inherits additional error from QRS detection, for example with the classical Pan and Tompkins detector [6], and this penalty falls disproportionately on the timing-sensitive S class. Third, although attention weights are intrinsic, they are a proxy for, not a guarantee of, clinical reasoning, and reader studies would be needed before any diagnostic claim.

6. CONCLUSION

This paper presented CAF-Net, a compact convolution and attention fusion network for five-class arrhythmia detection from single-lead ECG. A residual convolutional encoder distills beat morphology into 32 tokens, a two-layer multi-head self-attention encoder models global intra-beat context, a learnable-query pooling layer produces interpretable evidence weights, and a parallel branch fuses four RR interval features that restore rhythm context. Trained with plain cross-entropy on the natural class distribution of all 48 MIT-BIH records, CAF-Net achieves a mean accuracy of 99.35 percent, a mean macro F1 of 0.952, and a mean AUC of 0.9981 over three seeded runs, exceeding the accuracy of recently published and larger models while classifying a beat in 4.0 ms on a single CPU thread with a 1.79 MB footprint. Future work will pursue patient-independent and cross-database validation, integration with an embedded QRS detector for a fully end-to-end pipeline, quantization-aware training for microcontroller deployment, and reader studies of the attention explanations.

REFERENCES

[1] G. A. Roth et al., “Global burden of cardiovascular diseases and risk factors, 1990-2019: update from the GBD 2019 study,” Journal of the American College of Cardiology, vol. 76, no. 25, pp. 2982-3021, 2020.

[2] 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.

[3] L. Goldberger et al., “PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals,” Circulation, vol. 101, no. 23, pp. e215-e220, 2000.

[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] Mahawar, S., “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 (IJCIE), vol. 1, no. 2, pp. 42–48, 2026.

[6] J. Pan and W. J. Tompkins, “A real-time QRS detection algorithm,” IEEE Transactions on Biomedical Engineering, vol. BME-32, no. 3, pp. 230-236, 1985.

[7] P. Guduru, K. V. Lakshmi, and S. Gupta, "Protecting individualized education and therapy records via two-tier integrity verification in distributed educational systems," International Journal of Special Education, vol. 41, no. 5s, pp. 386–401, 2026. [Online]. Available: https://internationalsped.com/index.php/ijse/article/view/2980.

[8] M. Llamedo and J. P. Martinez, “Heartbeat classification using feature selection driven by database generalization criteria,” IEEE Transactions on Biomedical Engineering, vol. 58, no. 3, pp. 616-625, 2011.

[9] S. Kiranyaz, T. Ince, and M. Gabbouj, “Real-time patient-specific ECG classification by 1-D convolutional neural networks,” IEEE Transactions on Biomedical Engineering, vol. 63, no. 3, pp. 664-675, 2016.

[10] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych, and R. S. Tan, “A deep convolutional neural network model to classify heartbeats,” Computers in Biology and Medicine, vol. 89, pp. 389-396, 2017.

[11] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, “ECG heartbeat classification: a deep transferable representation,” in Proceedings of the IEEE International Conference on Healthcare Informatics, 2018, pp. 443-444.

[12] Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,” Nature Medicine, vol. 25, no. 1, pp. 65-69, 2019.

[13] O. Yildirim, “A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification,” Computers in Biology and Medicine, vol. 96, pp. 189-202, 2018.

[14] O. Yildirim, U. B. Baloglu, R. S. Tan, E. J. Ciaccio, and U. R. Acharya, “A new approach for arrhythmia classification using deep coded features and LSTM networks,” Computer Methods and Programs in Biomedicine, vol. 176, pp. 121-133, 2019.

[15] S. Mousavi and F. Afghah, "Inter- and intra-patient ECG heartbeat classification for arrhythmia detection: a sequence to sequence deep learning approach," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 1308-1312.

[16] T. F. Romdhane, H. Alhichri, R. Ouni, and M. Atri, "Electrocardiogram heartbeat classification based on a deep convolutional neural network and focal loss," Computers in Biology and Medicine, vol. 123, art. no. 103866, 2020.

[17] E. Essa and X. Xie, "An ensemble of deep learning-based multi-model for ECG heartbeats arrhythmia classification," IEEE Access, vol. 9, pp. 103452-103464, 2021.

[18] E. Jing, H. Zhang, Z. Li, Y. Liu, Z. Ji, and I. Ganchev, "ECG heartbeat classification based on an improved ResNet-18 model," Computational and Mathematical Methods in Medicine, vol. 2021, art. no. 6649970, 2021.

[19] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 5998-6008.

[20] M. R. Islam, M. Qaraqe, K. Qaraqe, and E. Serpedin, "CAT-Net: convolution, attention, and transformer based network for single-lead ECG arrhythmia classification," Biomedical Signal Processing and Control, vol. 93, art. no. 106211, 2024.

[21] F. Zhou and D. Fang, "Multimodal ECG heartbeat classification method based on a convolutional neural network embedded with FCA," Scientific Reports, vol. 14, art. no. 8804, 2024.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.

[23] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," arXiv preprint arXiv:1606.08415, 2016.

[24] Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proceedings of the International Conference on Learning Representations, 2019.

[25] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," Journal of Machine Learning Research, vol. 9, pp. 2579-2605, 2008.

[26] S. Saadatnejad, M. Oveisi, and M. Hashemi, "LSTM-based ECG classification for continuous monitoring on personal wearable devices," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 2, pp. 515-523, 2020.