IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 31, No. 10, October, 2026, pp. 3369-3378
Journal homepage: http://iieta.org/journals/isi

EchoFuseNet: A Lightweight Multimodal Fusion Network with Recurrence-Based Representations for Electrocardiogram Arrhythmia Classification

Bharathi Radarapu[1*], Eadha Neelima[2]

[1] GITAM Institute of Technology, GITAM Deemed to be University, Visakhapatnam 530001, India
[2] Department of CSE, GITAM Institute of Technology, GITAM Deemed to be University, Visakhapatnam 530001, India

Corresponding Author Email: bradarap@gitam.in

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/isi.311011

Received: 27 May 2026
Revised: 7 August 2026
Accepted: 21 August 2026
Available online: 31 October 2026

Keywords:
electrocardiogram, arrhythmia classification, multimodal fusion, lightweight deep learning, Recurrence Plot, Gramian Angular Field, Markov Transition Field, explainable artificial intelligence

ABSTRACT

Cardiac arrhythmia classification from electrocardiogram (ECG) signals remains challenging due to variations in signal morphology, class imbalance, and the computational cost of deep learning models. Existing high-performing approaches often require complex architectures, limiting their suitability for computationally constrained healthcare applications. This study proposes EchoFuseNet, a lightweight multimodal fusion network for ECG arrhythmia classification. The proposed framework transforms ECG heartbeat segments into three complementary representations, including Recurrence Plot (RP), Gramian Angular Field (GAF), and Markov Transition Field (MTF), which are processed through parallel lightweight convolutional branches and integrated using a compact feature fusion strategy. Experimental evaluation on the primary ECG heartbeat dataset demonstrates that EchoFuseNet achieves an accuracy of 99.86% with only 688,822 trainable parameters, a model size of 2.76 MB, and 0.057 Giga Floating Point Operations (GFLOPs) per inference. Ablation experiments confirm that combining multiple ECG representations provides complementary information compared with single-modality inputs. Gradient-weighted Class Activation Mapping (Grad-CAM) analysis is further conducted to examine the regions that contribute to the model's predictions. External evaluation on the Massachusetts Institute of Technology–Beth Israel Hospital (MIT-BIH) Arrhythmia Database achieves 91.74% accuracy and 93.09% weighted F1-score, indicating the impact of dataset variation on model generalization. The proposed framework provides a compact and interpretable solution for ECG arrhythmia classification, with computational characteristics suitable for further investigation in resource-constrained healthcare environments.

1. INTRODUCTION

1.1 Background and motivation

Cardiovascular disease is a significant public health problem, and timely identification of cardiac arrhythmia is crucial for diagnosis, ongoing surveillance, and clinical decision-making [1, 2]. Electrocardiography is a widely used method for assessing cardiac electrical activity because it is non-invasive, inexpensive, and applicable in both clinical and ambulatory settings [3]. As more and more wearable sensors, mobile health platforms, and telemedicine systems are employed, there is a need for automated electrocardiogram (ECG) analysis methods that can provide reliable prognoses while operating with limited computational capacity [4, 5].

1.2 Challenges in existing approaches

Convolutional neural networks (CNNs) and hybrid neural architectures have enhanced the classification accuracy of ECG arrhythmia in deep learning [6]. Many successful models, however, are based on computationally expensive designs, such as attention mechanisms, recurrent layers, or transformer-based components [7]. These methods yield high accuracy but can incur heavy memory usage, high inference costs, and hardware requirements that may not be feasible for wearable and edge-device applications [8]. Thus, it is a significant challenge to design ECG classification models with high predictive accuracy and lightweight architecture for real-time, resource-constrained deployment [8].

Another important challenge is generalization. ECG datasets may vary in patient populations, recording conditions, sampling parameters, noise levels, and class distributions. This means that a model that performs well on one dataset may not perform as well on an external dataset [9]. This becomes especially critical for minority classes of arrhythmia, where the number of samples is small and morphological variability is high, making classification more challenging [9].

1.3 Role of multimodal electrocardiogram representations

Waveform-level information can be obtained by directly modelling one-dimensional ECG signals, while transformed two-dimensional representations can generate complementary

temporal and structural patterns [10]. Recurrence Plots (RPs) detect nonlinear recurrence behavior, Gramian Angular Fields (GAFs) are based on temporal correlations encoded by angles, and Markov Transition Fields (MTFs) characterize the dynamics of transitions between signal states. These representations can be fused to enable the model to learn richer ECG features than a single representation alone [11].

On the other hand, if the model architecture is too complex, fusing multimodal ECG representations may incur higher computational cost. The research challenge is then not only how to merge multiple ECG representations, but how to do this efficiently while keeping the architecture compact enough for edge-oriented inference [12].

1.4 Research gap

Although recent deep learning-based approaches have substantially improved ECG arrhythmia classification and enabled richer representations of ECG signals, several important research gaps remain. Existing studies have explored CNN, CNN-Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BLSTM), attention-based, and transformer-based architectures to improve ECG classification performance [13]. However, these approaches primarily emphasize predictive performance, while computational aspects such as model size, Floating Point Operations (FLOPs), inference latency, memory requirements, and energy consumption are not consistently evaluated. This limitation is particularly important for real-time healthcare applications on resource-constrained wearable and edge devices, where lightweight architectures are required [14].

Another important gap concerns multimodal ECG representation and fusion. Recent studies have demonstrated that combining complementary ECG representations or modalities can improve feature learning and classification performance [15]. However, integrating multiple modalities can introduce additional computational and memory overhead, particularly when attention mechanisms, transformers, or complex fusion architectures are employed [15]. Therefore, there remains a need for a lightweight multimodal framework that can exploit complementary temporal and structural information while maintaining low computational complexity and real-time inference capability (Table 1).

Table 1. Computational complexity of the proposed model

Metric	Value
Number of parameters	688,822
Model size	2.76 MB
FLOPs per inference	0.057 GFLOPs

Note: GFLOPs = Giga Floating Point Operations.

Furthermore, cross-dataset generalization remains a significant unresolved challenge. ECG datasets can differ in patient populations, acquisition devices, recording conditions, signal quality, noise characteristics, and class distributions. Consequently, a model that performs well on one dataset may not maintain the same level of performance when applied to another dataset. Cross-database analysis has highlighted the importance of evaluating ECG models under distributional differences between datasets. However, many existing deep learning studies continue to focus primarily on performance on individual datasets [16], thereby limiting evidence regarding the robustness and transferability of these models.

In addition, interpretability and deployment efficiency are not consistently addressed together with classification performance. Although recent research has explored interpretable ECG classification [16], most studies still emphasize predictive metrics rather than jointly evaluating model complexity, computational efficiency, interpretability, and generalization. This creates a gap between achieving high classification accuracy and developing models that are practically suitable for real-world edge-based healthcare applications.

Therefore, these limitations highlight the need for a lightweight, computationally efficient, interpretable, and generalizable multimodal ECG classification framework. Motivated by these research gaps, the proposed EchoFuseNet integrates complementary ECG representations while controlling computational complexity. The framework aims to strike an effective balance among classification accuracy, model compactness, inference efficiency, interpretability, and cross-dataset robustness, thereby improving its suitability for real-time deployment on resource-constrained wearable and edge healthcare devices.

1.5 Proposed contribution

To meet these requirements, this study proposes a lightweight multimodal fusion framework for ECG arrhythmia classification, called EchoFuseNet. It converts ECG beat segments into RP, GAF, and MTF images, processes these images with lightweight convolutional branches in parallel, and fuses the resulting features using a compact fusion strategy.

The main contributions of this study are as follows:

1. We propose a lightweight multimodal ECG arrhythmia classification framework using RP, GAF, and MTF representations (Table 2).

Table 2. Parameters used for multimodal electrocardiogram (ECG) image transformations

Modality	Parameter	Value
RP	Embedding dimension	2
RP	Time delay	1
RP	Threshold	10% recurrence
GAF	Normalization	[−1, 1]
GAF	Field type	Summation GASF
MTF	Quantile bins	10
MTF	Markov order	First-order

Note: RP = Recurrence Plot, GAF = Gramian Angular Field, MTF = Markov Transition Field, GASF = Gramian Angular Summation Field.

2. We evaluate the model using predictive performance metrics and computational indicators, including accuracy, F1-score, parameter count, model size, FLOPs, latency, and energy consumption.

3. We analyze model interpretability using Gradient-weighted Class Activation Mapping (Grad-CAM) to examine whether the network focuses on relevant ECG regions.

4. We examine cross-dataset generalization to assess the practical transferability of the proposed framework under dataset shift.

1.6 Paper organization

This paper is organized as follows: Section 2 summarizes related studies on ECG classification, multimodal representation, lightweight architecture, and generalization.

Section 3 presents the proposed methodology, consisting of preprocessing, ECG image transformation, model architecture, and training setup. Section 4 reports the experimental results, including classification performance, ablation analysis, efficiency evaluation, robustness testing, and cross-dataset validation. Section 5 compares the proposed model with existing models. Section 6 discusses the results, limitations, and implications for practice, and Section 7 concludes the study and outlines future research directions.

2. RELATED WORKS

Advancements in automated ECG arrhythmia classification have shifted from standard feature-based techniques to deep learning algorithms that can learn discriminative patterns directly from ECG signals. Previous methods typically rely on manually extracted features, such as RR interval, wavelet coefficients, and statistical descriptors [17]. These were helpful in controlled environments, but their performance was limited by feature design and by decreased robustness to patient variability, recording noise, and dataset shift.

In recent years, CNNs, recurrent neural networks (RNNs), attention mechanisms, and transformer-based models have been increasingly used for ECG classification. These models can capture complex temporal and morphological features in ECG signals and have achieved strong classification performance [18]. Many of these methods, however, have high parameter counts, high memory consumption, or large computational expense [18], making them impractical for wearable sensor deployment and edge devices. There have been attempts to develop lightweight architectures, such as depthwise-separable convolutional networks, to reduce computational complexity [14], but many of these models rely on a single ECG representation and may not fully leverage complementary temporal information.

Another line of research transforms one-dimensional ECG beats into two-dimensional images that can be analyzed by image-based neural networks. RP represents nonlinear recurrence patterns [19], GAF represents temporal correlation, and MTF represents the dynamics of transitions between signal states. These representations can provide structural information that may not be adequately captured by direct one-dimensional signal modelling. Prior work on multimodal ECG classification based on transformed representations often adds architectural complexity via attention or strong fusion modules.

Efficiency is a significant issue for real-time ECG monitoring. To be viable for clinical and wearable applications, models must be not only accurate, but also compact, fast, and energy-efficient. Thus, parameter count, FLOPs, latency, memory size, and energy consumption are key evaluation metrics. Numerous previous studies have reported accuracy and F1-score but offered little convincing evidence regarding deployment feasibility, particularly under central processing unit (CPU)- or edge-oriented inference settings. This also leaves a divide between high-performing research models and those that could actually run on low-power devices.

Cross-dataset generalization is another challenge in ECG arrhythmia classification. ECG datasets may vary according to patient age, sensor quality, sampling frequency, noise characteristics, and class proportions. This means that a model with good performance on a held-out test set from the same distribution as the training data can perform worse on a different dataset. This is especially relevant for minority arrhythmia classes, which have fewer training examples and greater morphological variation, often leading to lower recall and macro-level performance.

Overall, the current literature indicates the potential of deep learning and multimodal ECG representations for enhancing arrhythmia classification, while highlighting several limitations [15]. Many models are too costly, are not rigorously tested for edge deployment, or are only weakly tested for external dataset shift. To address this gap, the present study proposes a lightweight multimodal fusion network, called EchoFuseNet, that fuses RP, GAF, and MTF representations using compact convolutional branches. The goal is to achieve a trade-off between predictive performance, computational efficiency, interpretability, and transferability to other datasets.

3. METHODOLOGY

3.1 Dataset description

The proposed EchoFuseNet framework Figure 1 was applied to a primary ECG heartbeat dataset to evaluate ECG arrhythmia classification, and the Massachusetts Institute of Technology–Beth Israel Hospital (MIT-BIH) Arrhythmia Dataset was used to test the proposed method across datasets. There were six classes of heartbeat, namely Fusion beat (F), Multiform beat (M), Normal beat (N), Unknown beat (Q), Supraventricular beat (S), and Ventricular beat (V) (Table 3).

The ECG Heartbeat Categorization Dataset was used to evaluate the proposed EchoFuseNet framework. Initially, an 80:20 split was used to separate the dataset into training and testing subsets. The original class distribution following this split is shown in Table 4, highlighting the dataset's intrinsic class imbalance. Only the training set was balanced by oversampling the minority classes to lessen the negative effects of class imbalance during model training; the test set was kept in its original distribution to guarantee an objective assessment and prevent information leakage.

The primary ECG dataset was partitioned at the individual heartbeat/sample level using an 80:20 random split. For each of the six heartbeat classes, the available samples were randomly divided into training and validation subsets using a test size of 0.20 and random_state = 42. The partition was therefore performed on individual heartbeat samples rather than at the patient or recording level. No patient identifier or recording identifier was used during this partitioning. Consequently, heartbeat samples from the same patient or recording may appear in both subsets, and patient-related information leakage cannot be completely ruled out. The reported primary-dataset results should therefore be interpreted as sample-level evaluation rather than strict patient-independent generalization. Class balancing was applied only to the training subset, while the validation/test data were retained without balancing.

To explore external generalization, we investigated an ECG dataset that differs in recording conditions, patient characteristics, noise levels, and class distribution, namely the MIT-BIH Arrhythmia Dataset. To evaluate the generalization ability of the proposed framework under dataset shift, we performed external validation on the MIT-BIH Arrhythmia Database, a publicly available benchmark dataset from the

PhysioNet repository. The database was created at the Massachusetts Institute of Technology (MIT) and Beth Israel Hospital (BIH) and comprises 48 half-hour two-channel ambulatory ECG recordings obtained from 47 subjects. The ECG signals were sampled at 360 Hz and manually annotated by clinical experts.

The dataset is a standard benchmark for automated arrhythmia detection. The MIT-BIH Arrhythmia Database is a good candidate to evaluate cross-dataset generalization, as it differs from the primary dataset in patient demographics, recording conditions, noise characteristics, and class distribution.

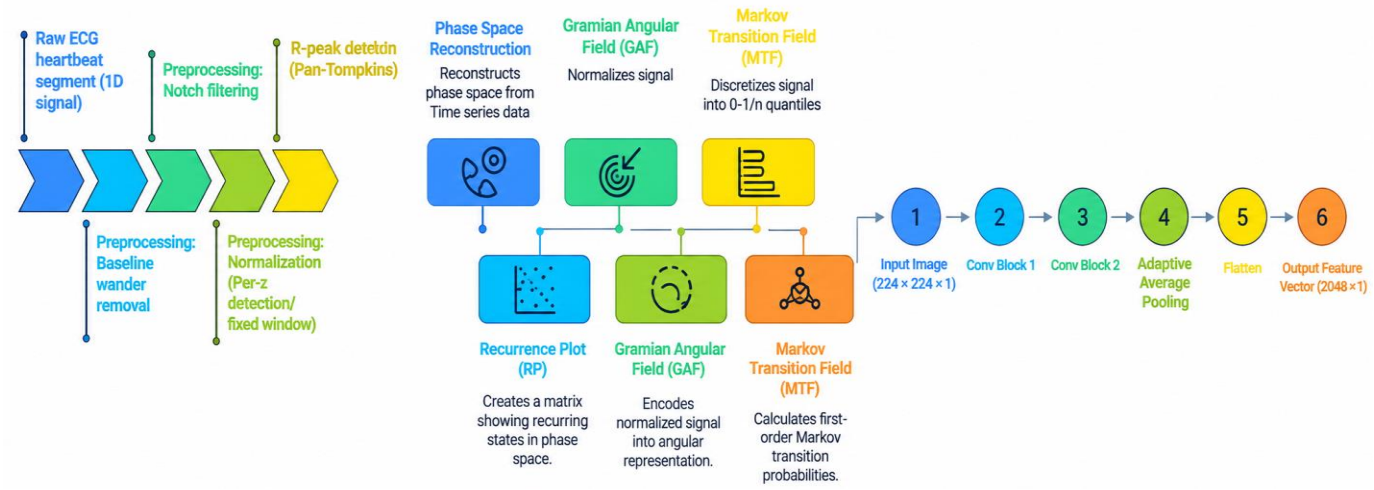


Figure 1. Overall architecture of the proposed EchoFuseNet framework

Table 3. Class mapping used for Massachusetts Institute of Technology–Beth Israel Hospital (MIT-BIH) arrhythmia cross-dataset evaluation

MIT-BIH Class	Label	Description
0	N	Normal beat
1	S	Supraventricular beat
2	V	Ventricular beat
3	F	Fusion beat
4	Q	Unknown/Another beat

Table 4. Class-wise distribution of the primary electrocardiogram (ECG) dataset

Class	Label	Train Count	Test Count	Test %
F	Fusion beat	642	161	0.65%
M	Multiform	8,405	2,101	8.47%
N	Normal	75,709	18,926	76.32%
Q	Unknown	6,431	1,608	6.48%
S	Supraventricular	2,223	556	2.24%
V	Ventricular	5,789	1,447	5.83%
Total	–	99,199	24,799	100%

3.2 Electrocardiogram preprocessing

An ECG preprocessing pipeline was used to produce clean, consistent heartbeat segments for multimodal transformation. Filtering was first applied to eliminate baseline wander and power-line interference. R-peaks were then detected using the Pan–Tompkins algorithm. Once R-peak detection was complete, fixed-length ECG beat segments were cut around each detected R-peak. This ensured that every sample represented one heartbeat and had the same structure when transformed and classified.

3.3 Multimodal transformation

Each one-dimensional ECG beat was converted into three two-dimensional image representations—RP, GAF, and MTF—to capture complementary temporal and spatial information from different aspects of the ECG beat. These representations capture different facets of ECG morphology and temporal behavior.

3.3.1 Recurrence Plot

The recurrence matrix is computed as:

$$RP_{i,j} = \Theta(\varepsilon - \|x_i - x_j\|_2) \quad (1)$$

where, x_i denotes delay-embedded vectors, $\Theta(\cdot)$ is the Heaviside step function, and ε is selected empirically to achieve approximately 10% recurrence density.

3.3.2 Gramian Angular Field

The ECG signal is normalized to [−1, 1] and mapped to angular space:

$$\varphi_t = arccos(x_t) \quad (2)$$

The Gramian Angular Summation Field (GASF) is computed as:

$$G_{i,j} = cos(\varphi_i + \varphi_j) \quad (3)$$

which encodes temporal correlations between time points [19].

3.3.3 Markov Transition Field

The signal is discretized into $Q = 10$ quantile bins to form a first-order Markov transition matrix $P \in \mathbb{R}^{Q \times Q}$. The MTF image is defined as:

$$M_{i,j} = P_{si,sj} \quad (4)$$

where, s_i represents the quantile state of sample i. This representation captures temporal transition dynamics [20].

3.4 Proposed model architecture

EchoFuseNet consists of three lightweight parallel convolutional branches. Each ECG image representation (RP, GAF, or MTF) is processed by its own branch in Algorithm 1. All input images are resized to 64 × 64 grayscale images. The features obtained from the three branches are concatenated to form a fused feature vector and fed to fully connected layers for the final classification of the six arrhythmia classes.

Each branch uses the following lightweight convolutional structure: Conv(1→16, 3×3) → BatchNorm → ReLU → MaxPool; Conv(16→32, 3×3) → BatchNorm → ReLU → MaxPool; Conv(32→48, 3×3) → BatchNorm → ReLU → MaxPool; followed by adaptive average pooling to (4×4). The three branches are concatenated into a feature vector of size 2304, which is passed through fully connected layers: 2304 → 256 → 128 → 64 → Output (Softmax).

EchoFuseNet contains 688,822 parameters, has a model size of 2.76 MB, and requires 0.057 Giga Floating Point Operations (GFLOPs) per inference, indicating a compact computational profile for resource-constrained ECG inference.

Algorithm 1. EchoFuseNet workflow

Require: ECG signal x(t)
Ensure: Predicted arrhythmia class ŷ
1: xf(t) ← Fbp(x(t)) ▷ Baseline and notch filtering
2: {rk} ← R(xf(t)) ▷ R-peak detection
3: for each detected peak rk do
4: sk ← xf(t)[rk − w : rk + w] ▷ Beat segmentation
5: IRP ← ΦRP(sk)
6: IGAF ← ΦGAF(sk)
7: IMTF ← ΦMTF(sk)
8: fRP ← CRP(IRP)
9: fGAF ← CGAF(IGAF)
10: fMTF ← CMTF(IMTF)
11: f ← [fRP; fGAF; fMTF]
12: ŷ ← Softmax(Fc(f))
13: end for
14: return ŷ

3.5 Loss function and optimization

Categorical cross-entropy loss for multi-class classification was used to train the network. The Adam optimizer was used with a learning rate of 0.001. The batch size was 64, and the maximum number of training epochs was 50. Overfitting was mitigated using early stopping when validation performance plateaued. All experiments were carried out in PyTorch.

4. RESULTS

4.1 Classification performance

EchoFuseNet's classification performance on the primary test set was excellent. The proposed multimodal fusion strategy effectively differentiated the six ECG arrhythmia classes, achieving 99.86% accuracy. Class-wise results demonstrated high precision, recall, and F1-score across most classes.

The confusion matrix also demonstrates the model's classification accuracy. The high number of correctly classified samples indicates that the model learned discriminative ECG features from the fused RP, GAF, and MTF representations. Performance was comparatively weaker for minority classes, such as fusion beats, due to limited sample size and morphological variability, as expected.

These results should be interpreted with some caution: strong performance on the primary test set does not guarantee universal generalization. The cross-dataset evaluation reported later in this section, therefore, assesses the model's transferability to a different ECG distribution.

4.2 Ablation analysis

The ablation study demonstrates that each ECG representation contributes information useful for classification. The individual performance of RP, GAF, and MTF was good, but the best overall result was obtained with the multimodal fusion model. This indicates that the three representations are complementary and do not capture the same ECG features. The improvement from single-modality variants to the complete fusion model supports the design of EchoFuseNet: recurrence behavior is captured by RP, temporal correlation is captured by GAF (Table 5), and transition dynamics are captured by MTF. Fusing these representations enables the model to acquire a more detailed representation of ECG beat morphology.

Table 5. Ablation comparison of single-modality variants and full multimodal fusion

Model Variant	Test Accuracy	Weighted F1-Score
RP + GAF + MTF (Full fusion)	99.86%	99.86%
GAF-only	99.72%	99.72%
MTF-only	99.74%	99.74%
RP-only	99.75%	99.54%

Note: RP = Recurrence Plot, GAF = Gramian Angular Field, MTF = Markov Transition Field.

The model captures the majority class, Normal, well, as seen in Figure 2, and is also effective for other arrhythmia types. The normalized confusion matrix in Figure 2 shows that most classes have a recall above 0.85, though Fusion and Unknown perform less well due to the limited number of examples in those classes. Overall, the model performs well across different ECG distributions, handling both normal and other arrhythmia types effectively.

Confidence for supraventricular beats is comparatively low, owing to their known similarity to normal beats. Notably, fusion beats—which are relatively rare and morphologically distinct—are identified with high confidence. This indicates that the system effectively fuses information to distinguish beat types, even in the face of real-world class-frequency imbalances and morphological variability. The framework handles supraventricular and fusion beats reasonably well despite these challenges.

4.3 Efficiency and deployment-oriented evaluation

All experiments were conducted using the PyTorch framework on a desktop workstation with CPU-based inference. The model was trained using the Adam optimizer (learning rate = 0.001), a batch size of 64, and early stopping based on validation performance. The computational

complexity of EchoFuseNet was evaluated by measuring the number of parameters, model size, FLOPs, and inference latency in Table 6. Latency was measured directly by averaging inference times across multiple runs using both native PyTorch CPU and Open Neural Network Exchange (ONNX) Runtime CPU execution.

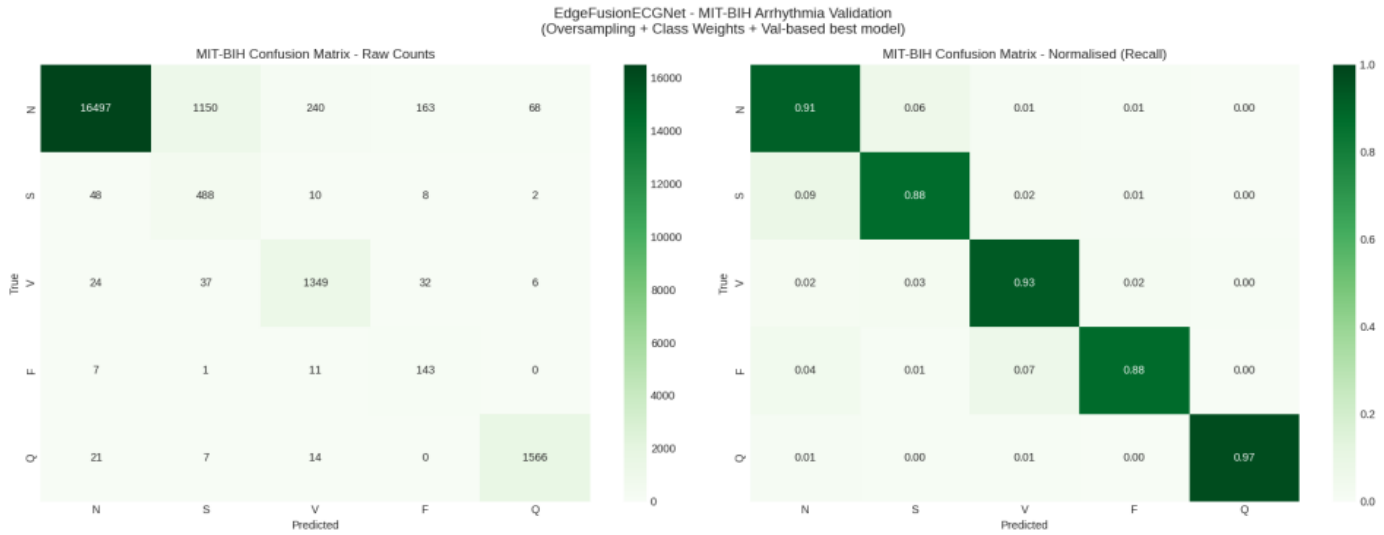


Figure 2. Confusion matrix of the proposed model on the Massachusetts Institute of Technology–Beth Israel Hospital (MIT-BIH) Dataset: (left) raw classification counts and (right) normalized recall values

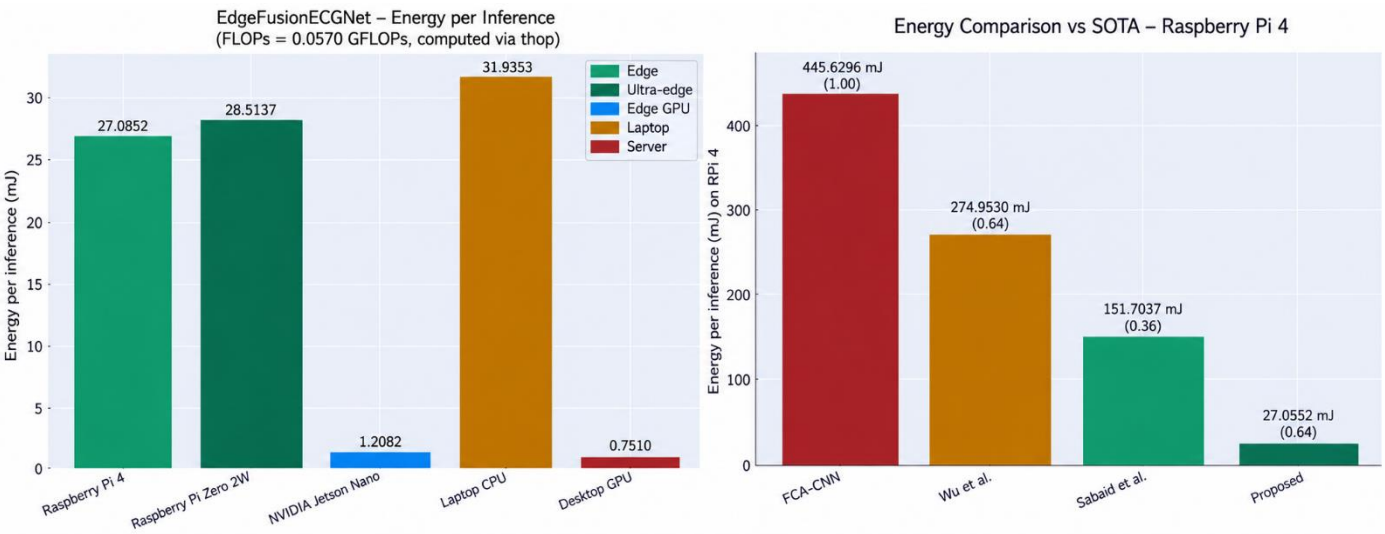


Figure 3. Energy consumption of the proposed model: (left) across hardware platforms and (right) comparison with state-of-the-art models on Raspberry Pi 4

Table 6. Inference latency comparison across deployment frameworks

Framework	Mean Latency	Standard Deviation	Frames per Second
PyTorch CPU (all threads)	20.5 ms	±1.3 ms	~49
ONNX Runtime CPU (all threads)	6.4 ms	±0.3 ms	155.1

Note: CPU = central processing unit; ONNX = Open Neural Network Exchange.

To assess potential deployment on edge devices, energy consumption was not measured on physical hardware. Instead, it was estimated using the measured inference latency and the published Thermal Design Power (TDP) specifications for representative edge platforms (Raspberry Pi 4, Raspberry Pi Zero 2W, and NVIDIA Jetson Nano). Therefore, the reported energy values should be interpreted as analytical estimates of deployment feasibility rather than direct device measurements. Actual energy consumption may vary depending on hardware configuration, operating frequency, workload, and runtime conditions.

Table 7. Estimated energy consumption per inference across edge hardware platforms

Device	Category	TDP (W)	E/Inf (mJ)	Inferences/ Wh
Raspberry Pi 4 (ARM Cortex A-72)	Edge	6.4	27.04	133,160
Raspberry Pi Zero 2W	Ultra Edge	2.0	28.51	126,255
NVIDIA Jetson Nano	Edge GPU	10.0	1.21	2,979,624

Note: TDP = Thermal Design Power.

Table 7 shows the estimated energy required for each prediction on the devices considered. Energy consumption is substantially lower than for heavier models, indicating that the proposed model is more efficient (Figure 3).

4.4 Noise robustness

Noise robustness was evaluated by adding Gaussian noise at different sigma values as shown in Table 8. The results show that EchoFuseNet maintained strong performance under low and moderate noise levels. Performance degraded substantially under higher-noise conditions, especially at extreme sigma values, indicating that the model is robust to moderate perturbations but not fully reliable under severe signal corruption.

Table 8. Noise robustness evaluation

Noise Sigma	Test Accuracy	Weighted F1-Score	Assessment
0.00 (clean)	99.86%	99.86%	Baseline
0.01	99.86%	99.86%	No degradation
0.05	99.83%	99.83%	Minimal drop
0.10	99.83%	99.83%	Minimal drop
0.15 (training max)	99.72%	99.72%	Minimal drop
0.20	99.24%	99.25%	Still strong
0.30	82.20%	82.60%	Significant degradation
0.50	8.48%	1.40%	Model fails (extreme noise)

4.5 Cross-dataset generalization

To evaluate transferability under dataset shift, cross-dataset evaluation was performed using a different dataset (MIT-BIH Arrhythmia). The accuracy and weighted F1-score for this evaluation were 91.74% and 93.09%, respectively. The macro F1-score of 73.89% is considerably lower, indicating that generalization for minority classes remains a challenge.

This result is significant because it demonstrates both the promise and the current limitations of the proposed framework. The model performs well on the primary dataset but less effectively on the external dataset, indicating that further testing is needed before deployment claims can be made.

The class-wise results show that EchoFuseNet achieved strong recognition of the Normal (N), Ventricular (V), and Unknown/Other (Q) classes, with F1-scores of 0.9516, 0.8758, and 0.9506, respectively. The Supraventricular (S) and Fusion (F) classes achieved lower F1-scores of 0.4901 and 0.4263, respectively as shown in Table 9. The lower F1-scores observed for the Supraventricular and Fusion classes indicate that minority-class discrimination remains challenging under cross-dataset distribution shift. Although these minority classes achieved relatively high recall, their precision was lower, indicating more false-positive predictions. Overall, the model achieved 91.74% test accuracy, 0.9309 weighted F1-score, and 0.7389 macro F1-score, demonstrating good overall cross-dataset performance while highlighting the remaining challenge of minority-class discrimination.

Table 9. Class-wise performance of EchoFuseNet on the Massachusetts Institute of Technology–Beth Israel Hospital Arrhythmia Dataset

Class	Description	Precision	Recall	F1-Score
N	Normal beat	0.9921	0.9143	0.9516
S	Supraventricular beat	0.3416	0.8669	0.4901
V	Ventricular beat	0.8475	0.9061	0.8758
F	Fusion beat	0.2792	0.8800	0.4263
Q	Unknown/other beat	0.9217	0.9700	0.9506
Macro average		0.6764	0.9140	0.7389
Weighted average		0.9556	0.9174	0.9309

4.6 Prediction confidence and interpretability

Prediction-confidence analysis in Figure 4 showed that the model's output was high-confidence for most correctly classified samples. However, confidence decreased for some minority or morphologically similar classes, indicating that class imbalance and inter-class similarity remain significant issues [9].

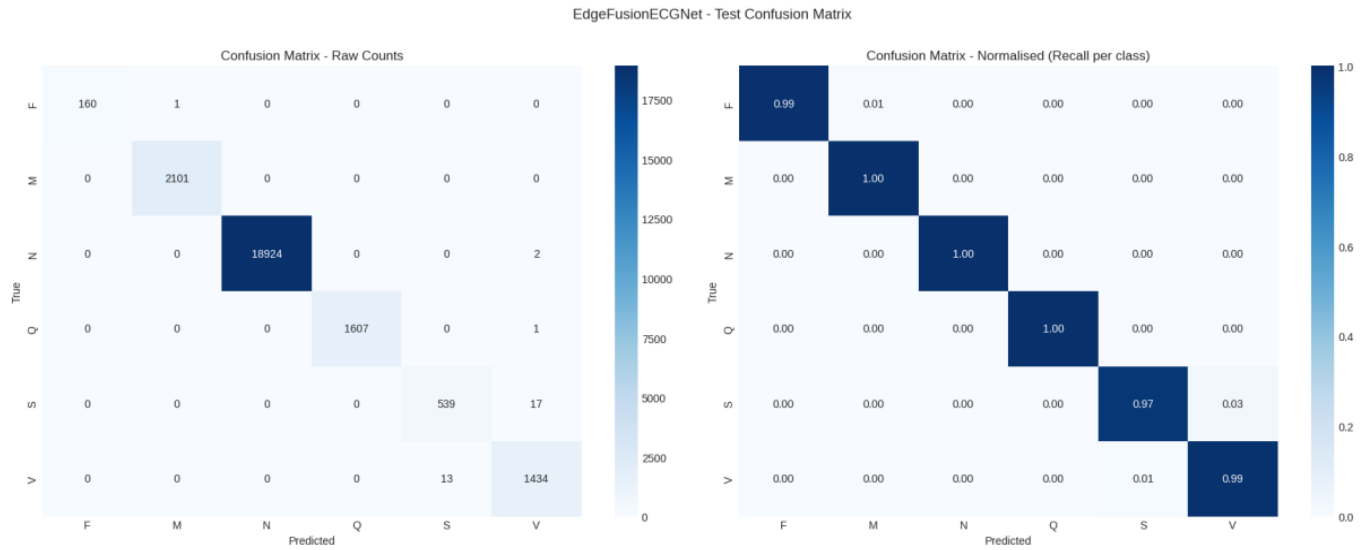


Figure 4. Prediction confidence distribution of EchoFuseNet across electrocardiogram (ECG) arrhythmia classes

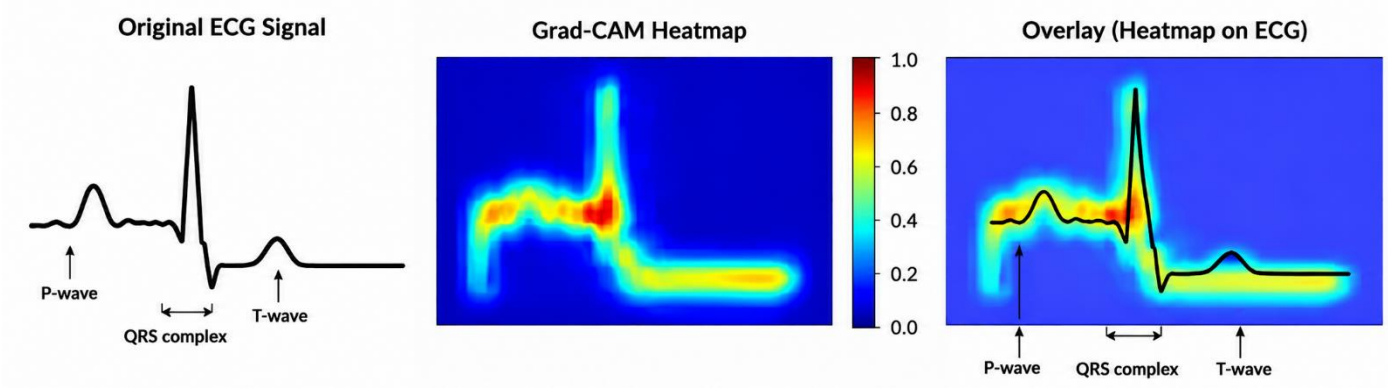


Figure 5. Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations of EchoFuseNet predictions for representative arrhythmia classes across the Recurrence Plot (RP), Gramian Angular Field (GAF) and Markov Transition Field (MTF) modalities

Note: The highlighted regions indicate areas contributing more strongly to the model's prediction.

Grad-CAM was used to analyze which regions of the ECG influenced the model's predictions in Figure 5. The visualizations indicate that EchoFuseNet focuses on relevant regions of the ECG, such as QRS complexes and rhythm changes. This supports interpretability, although Grad-CAM results should be treated as explanatory evidence rather than clinical validation.

5. COMPARATIVE ANALYSIS WITH EXISTING MODELS

To assess its effectiveness, the proposed model was compared with prevalent approaches for ECG arrhythmia classification in terms of predictive performance and computational efficiency. Accuracy, F1-score, parameter count, FLOPs, and edge-device suitability were considered. This is important because a model developed for resource-limited ECG monitoring must also be evaluated for inference cost and deployment feasibility.

Table 10 compares the proposed model with several ECG classification models. The proposed model achieved competitive classification performance while requiring fewer parameters and less computation than the representative models included in the comparison; however, these results should be interpreted in the context of the differing datasets and evaluation protocols used across studies.

The comparison in Table 11 provides an indicative assessment of EchoFuseNet's classification performance and computational characteristics relative to representative ECG classification models. However, the reported results should not be interpreted as a direct head-to-head comparison because the studies compared used different datasets, preprocessing procedures, data partitioning strategies, input representations, and evaluation protocols. Therefore, differences in accuracy, F1-score, parameter count, and FLOPs may partly reflect variations in experimental settings rather than differences in model architecture alone. Consequently, the comparison is used primarily to highlight the potential accuracy–computational efficiency trade-off of EchoFuseNet. In particular, the relatively low parameter count and FLOPs of the proposed model indicate computational compactness within the reported experimental setting, but do not establish universal superiority over existing ECG classification approaches. This comparison should be interpreted with caution, as published models may use different datasets, preprocessing techniques, split strategies, and evaluation protocols. The comparative results are best used to appreciate the efficiency–performance trade-off rather than to assert overall superiority over all existing ECG models.

Table 10. Cross-dataset performance on the Massachusetts Institute of Technology–Beth Israel Hospital (MIT-BIH) Arrhythmia Dataset

Metric	Value
Test accuracy	91.74%
Weighted F1-score	93.09%
Macro F1-score	73.89%
Best accuracy value	94.30%

Table 11. Comparison of EchoFuseNet with representative ECG classification models

Model	Modalities	Accuracy (%)	F1-Score	Parameters (M)	FLOPs (G)	Edge Suitability
Zhou and Fang [11]	STFT + CNN	98.71	0.986	8.40	0.94	Low
Sadad et al. (IoT-CNN) [12]	1D ECG	97.89	0.978	2.10	0.32	Medium
Wong et al. [20]	1D + 2D	99.12	0.991	5.20	0.74	Low
Proposed EchoFuseNet	RP + GAF + MTF	99.86	0.9986	0.69	0.057	High

Note: CNN = Convolutional neural network, ECG = electrocardiogram, RP = Recurrence Plot, GAF = Gramian Angular Field, MTF = Markov Transition Field.

6. DISCUSSION

The results show that EchoFuseNet offers a favorable trade-off between classification performance and computational efficiency. Performance on the main test set was excellent, while the architecture remained compact, with 688,822

trainable parameters, a model size of 2.76 MB, and 0.057 GFLOPs per inference. These properties indicate that the proposed lightweight multimodal design can effectively exploit complementary ECG representations while maintaining a low computational cost.

The ablation study supports the benefit of multimodal fusion. Each of the RP, GAF, and MTF representations captures a distinct aspect of ECG morphology and temporal behavior: RP models recurrence structure, GAF models temporal correlations, and MTF models transition dynamics. The individual representations demonstrated competitive performance, while the full RP + GAF + MTF fusion model achieved the strongest overall performance, indicating that these representations provide complementary information for arrhythmia classification.

The efficiency analysis reinforces the motivation for the proposed architecture. The small parameter count, low model size, low FLOPs, and favorable ONNX Runtime latency indicate the potential for efficient inference in resource-constrained environments. However, these results are based on computational analysis performed on a desktop CPU and do not constitute validation on physical edge or wearable hardware. Therefore, the reported efficiency should be interpreted as evidence of computational feasibility rather than confirmation of hardware deployment capability. Real-world deployment would require evaluation on representative target hardware, real-time data streams, and broader patient populations.

The robustness experiments show that EchoFuseNet performs well under low and moderate levels of Gaussian noise, although performance degrades significantly at high noise levels. This suggests that the model is not fully robust to extreme signal corruption. Further investigation using more realistic noise sources, such as motion artifacts, electrode displacement, and baseline drift under wearable monitoring, would therefore be beneficial.

One of the main limitations of this study is cross-dataset generalization. EchoFuseNet performed well on the primary dataset but showed reduced performance on the external MIT-BIH Arrhythmia Database, achieving 91.74% accuracy and a weighted F1-score of 93.09%, with a lower macro F1-score. This decrease in accuracy and macro F1-score on the external MIT-BIH Arrhythmia Database suggests that the model remains sensitive to dataset shift and class imbalance, particularly for minority classes. The external evaluation, therefore, provides an initial indication of the model's generalization capability rather than definitive evidence of universal applicability across diverse ECG datasets.

From an application perspective, EchoFuseNet has a compact computational profile, indicating potential applicability to intelligent healthcare information systems requiring efficient automated ECG analysis, including remote patient monitoring, wearable cardiac monitoring, telemedicine, and Internet of Medical Things (IoMT) environments. However, these represent potential use cases rather than demonstrated deployment scenarios, as the present study did not evaluate the model on physical edge or wearable hardware. Hardware-level validation is therefore required before practical deployment.

The overall contribution of EchoFuseNet lies not only in its high classification performance on the primary dataset but also in its computational efficiency and interpretability. Grad-CAM visualizations further indicate that the model focuses on relevant ECG morphological regions, providing an additional perspective on its decision-making process. Future work should focus on broader external validation, patient-independent testing, improved sensitivity for minority classes, realistic assessment of wearable noise, and evaluation on additional hospital and wearable ECG datasets. These efforts would help establish EchoFuseNet's robustness and practical applicability across diverse patient populations, acquisition conditions, and recording devices.

7. CONCLUSIONS

This paper presented a lightweight multimodal framework, EchoFuseNet, for ECG arrhythmia classification. The proposed model is based on parallel lightweight convolutional branches combined with a compact fusion strategy operating on RP, GAF, and MTF representations. EchoFuseNet achieved 99.86% accuracy on the primary evaluation dataset with a low parameter count of 688,822, a model size of 2.76 MB, and 0.057 GFLOPs per inference.

The experimental results demonstrate that multimodal ECG representations can enhance classification performance without significantly increasing inference cost. The ablation study confirms the contributions of RP, GAF, and MTF fusion, while the latency and energy analysis support resource-constrained ECG inference. The model further focuses on relevant ECG regions, as indicated by Grad-CAM visualizations, which supports its interpretability.

However, further validation is needed. Cross-dataset evaluation revealed lower accuracy (91.74%) and weighted F1-score (93.09%), showing that generalization is still limited. Future work should include more robust external validation, patient-independent testing, improved sensitivity for minority classes, and evaluation on more diverse clinical and wearable ECG datasets.

REFERENCES

[1] Sabor, N., Gendy, G., Mohammed, H., Wang, G., Lian, Y. (2022). Robust arrhythmia classification based on QRS detection and a compact 1D-CNN for wearable ECG devices. IEEE Journal of Biomedical and Health Informatics, 26(12): 5918-5929. https://doi.org/10.1109/JBHI.2022.3207456

[2] Zhang, T., Lian, C., Xu, B.R., Su, Y.X., Zeng, Z.G. (2024). Cardiac signal classification via an optional multimodal, multiscale receptive-field CNN-enhanced Transformer. Knowledge-Based Systems, 300: 112175. https://doi.org/10.1016/j.knosys.2024.112175

[3] Bishop, C.M. (2006). Pattern Recognition and Machine Learning. Springer, New York.

[4] Campanharo, A.S., Sirer, M.I., Malmgren, R.D., Ramos, F.M., Amaral, L.A.N. (2011). Duality between time series and networks. PLoS One, 6(8): e23378. https://doi.org/10.1371/journal.pone.0023378

[5] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 1800-1807. https://doi.org/10.1109/CVPR.2017.195

[6] Ye, Y., Chipusu, K., Ashraf, M.A., Ding, B., Huang, Y., Huang, J. (2025). Hybrid CNN-BLSTM architecture for classification and detection of arrhythmia in ECG

signals. Scientific Reports, 15(1): 34510. https://doi.org/10.1038/s41598-025-17671-1

[7] Soares, Q.B., Andrade, D.A., Ribeiro, E., Verardino, R.G.S., Reis, T.C., Samesima, N. (2024). Clinical assessment of a lightweight CNN model for real-time atrial fibrillation prediction in continuous wearable monitoring. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, pp. 1-4. https://doi.org/10.1109/EMBC53108.2024.10781856

[8] Wang, Y., Rani, S. (2025). Hybrid CNN-LSTM model for ECG-based arrhythmia detection in internet of medical things. Discover Internet of Things, 6(1): 7. https://doi.org/10.1007/s43926-025-00131-7

[9] Kim, S., Chon, S., Kim, J.K., Kim, J., Gil, Y., Jung, S. (2022). Lightweight convolutional neural network for real-time arrhythmia classification on low-power wearable electrocardiograph. In 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, United Kingdom, pp. 1915-1918. https://doi.org/10.1109/EMBC48229.2022.9871156

[10] Malleswari, P.N., Odugu, V.K., Rao, T.S., Aswini, T.V.N.L. (2024). Deep learning-assisted arrhythmia classification using 2-D ECG spectrograms. EURASIP Journal on Advances in Signal Processing, 2024(1): 104. https://doi.org/10.1186/s13634-024-01197-1

[11] Zhou, F., Fang, D. (2024). Multimodal ECG heartbeat classification method based on a convolutional neural network embedded with FCA. Scientific Reports, 14(1): 8804. https://doi.org/10.1038/s41598-024-59311-0

[12] Sadad, T., Safran, M., Khan, I., Alfarhood, S., Khan, R., Ashraf, I. (2023). Efficient classification of ECG images using a lightweight CNN with attention module and IoT. Sensors, 23(18): 7697. https://doi.org/10.3390/s23187697

[13] Majid, M.D., Anwar, M., Bilal, S.F., et al. (2026). A hybrid learning framework for automated multiclass electrocardiogram classification with SimCardioNet. Scientific Reports, 16(1): 7621. https://doi.org/10.1038/s41598-026-36932-1

[14] An, X., Shi, S., Wang, Q., Yu, Y., Liu, Q. (2024). Research on a lightweight arrhythmia classification model based on knowledge distillation for wearable single-lead ECG monitoring systems. Sensors, 24(24): 7896. https://doi.org/10.3390/s24247896

[15] Ahmad, Z., Tabassum, A., Guan, L., Khan, N.M. (2021). ECG heartbeat classification using multimodal fusion. IEEE Access, 9: 100615-100626. https://doi.org/10.1109/ACCESS.2021.3097614

[16] Kim, M., Shin, M. (2026). Enhancing ECG classification generalization through unified multi-dataset training. Sensors, 26(6): 1830. https://doi.org/10.3390/s26061830

[17] Clifford, G.D., Liu, C., Moody, B., et al. (2017). AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017. In 2017 Computing in Cardiology (CinC), Rennes, France, pp. 1-4. https://doi.org/10.22489/CinC.2017.065-469

[18] Reddy, C.K.R., Daduvy, A., Kaza, V.S., Shuaib, M., Mohzary, M., Alam, S., Sheneamer, A. (2025). A multi-scale convolutional LSTM-dense network for robust cardiac arrhythmia classification from ECG signals. Computers in Biology and Medicine, 91: 110121. https://doi.org/10.1016/j.compbiomed.2025.110121

[19] Eckmann, J.P., Kamphorst, S.O., Ruelle, D. (1987). Recurrence plots of dynamical systems. Europhysics Letters, 4(9): 973-977. https://doi.org/10.1209/0295-5075/4/9/004

[20] Wong, D.L.T., Li, Y., Deepu, J., Ho, W.K., Heng, C.H. (2022). An energy efficient ECG ventricular ectopic beat classifier using binarized CNN for edge AI devices. IEEE Transactions on Biomedical Circuits and Systems, 16(2): 222-232. https://doi.org/10.1109/TBCAS.2022.3152623