Review of Computer Engineering Research
2026 Vol. 13, No. 3, pp. 156-172
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/76.v13i3.5144

A convolutional attention transformer network for ECG beat classification

Shanmukha Rao Narsupalli[1,2+]
Rajesh Kumar Pullagura[2]
Rajeswara Rao Gangula[3]

[1]*Department of Electronics and Communication, MVGR College of Engineering, Vizianagaram, India.*
[2,3]*Department of Electronics and Communication, College of Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India.*
[1,2]*Email: shanmukh@mvgrce.edu.in*
[2]*Email: rajeshauce@gmail.com*
[3]*Email: grajesh197@gmail.com*

(+ Corresponding author)

Article History
Received: 2 January 2026
Revised: 7 April 2026
Accepted: 11 May 2026
Published: 3 September 2026

Keywords
AAMI
Arrhythmia
Attention mechanism
Classification
ECG Beat
Ensemble learning
MIT-BIH dataset.

ABSTRACT

This work aims to improve automated ECG beat classification in accordance with the Association for the Advancement of Medical Instrumentation (AAMI)-recommended classes by learning both local morphological patterns and longer-range temporal dependencies from ECG segments. We develop a complete analysis pipeline using the Massachusetts Institute of Technology–Beth Israel Hospital (MIT-BIH) Arrhythmia Database, including beat segmentation around detected R-peaks, standard pre-processing, and evaluation under a consistent train/validation/test protocol. The proposed Convolutional–Attention–Transformer (CAT) Network integrates convolutional feature extraction for local waveform morphology, an attention mechanism to emphasize diagnostically salient regions, and transformer-based sequence modeling to capture contextual dependencies. It is benchmarked against representative machine learning and ensemble baselines such as k-nearest neighbor (k-NN), support vector machine (SVM), Random Forest (RF), XGBoost, and CatBoost. Experimental results demonstrate that the CAT Network achieves superior or competitive performance across standard metrics, particularly improved macro-averaged performance and minority-class sensitivity, indicating better balance under class imbalance. We also provide a confusion matrix and class-wise analyses to highlight error patterns across AAMI classes and explain the practical impact of the proposed design choices. These findings suggest that the CAT Network can serve as an effective and reproducible framework for AAMI-compliant ECG beat classification, supporting downstream decision support and large-scale screening, while motivating future work on cross-database generalization, computational optimization for edge deployment, and enhanced interpretability for clinical adoption.

Contribution/Originality: This study contributes to ECG beat-classification literature by integrating a novel combination of convolution, attention, and transformer blocks within a unified American Association for the Advancement of Medical Instrumentation (AAMI)-compliant pipeline and benchmarking them against strong machine-learning and ensemble baselines. It documents improved balanced performance and minority-class sensitivity under a consistent, reproducible evaluation protocol.

1. INTRODUCTION

The heart's electrical activity is checked with electrocardiography, or ECG, in a way that does not hurt. This is shown in Figure 1. When we look at an ECG, we see the P wave, the QRS complex, and the T wave. These parts show what is happening in the heart. The P wave is about depolarization. The QRS complex is about depolarization.

The T wave is about the repolarisation, as we can see in the work by AbdElminaam, et al. [1] and Sharma, et al. [2]. If the shape or timing of these parts is not right, it can mean there is a problem with the heart. That is why doctors still need to look at the ECG results when they are checking patients and trying to figure out what is wrong with them. Electrocardiography is very important for this. Cardiac arrhythmias are a concern when we talk about abnormalities. This is because cardiac arrhythmias involve heart rhythms that are too fast, too slow, or just not regular. If we do not catch arrhythmias on time, it can increase the risk of serious complications like stroke, heart failure, or even sudden cardiac death from cardiac arrhythmias. Normally, experts review these things, which is the standard way of doing things in clinics. When we have to keep watching cardiac arrhythmias all the time, and the recordings are very long, it becomes difficult for experts to keep up with manual assessment, especially when screening many people for cardiac arrhythmias. This practical limitation has driven growing interest in automated beat classification, where both conventional machine learning models and more recent deep learning approaches are used to support ECG arrhythmia analysis [3-6].

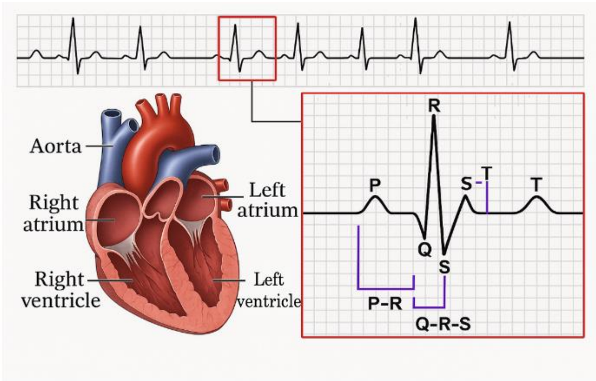

Figure 1. Basic ECG signal with different characteristic waves.

Table 1. Comparative synthesis of representative ECG arrhythmia classification studies and how the proposed cat network addresses prior limitations.

Literature	Model type	Key idea	Limitation(s)	How the Proposed CAT Network addresses the gap
Zhang, et al. [7]	Multi-lead DL fusion	Multi-branch fusion for 12-lead ECG	Lead-specific setting; protocol differs from single-lead AAMI mapping	Single-lead AAMI-focused design with consistent benchmarking
Zou, et al. [8]	RF + context feature	Temporal context feature improves RF	Feature-engineering dependent; limited deep dependency modeling	Learns features end-to-end; transformer captures dependencies
Islam, et al. [9]	CNN + attention + transformer	Hybrid blocks for imbalanced ECG	Architecture and evaluation settings vary across studies	Unified design evaluated under a consistent protocol with baselines
Bhattacharyya, et al. [10]	Ensemble ML (RF+SVM)	DWT features + ensemble	Performance depends on handcrafted features	End-to-end morphology learning + attention emphasis
Chen, et al. [6]	Multiscale CNN ensemble	Wearable-oriented CNN ensemble	Limited long-range modeling without an explicit transformer	Adds transformer blocks for longer-range dependency modeling
Fu, et al. [11]	GPT-based ECG interpretation	Text generation for ECG explanation	Not optimised for beat-wise AAMI classification; higher complexity	Beat-wise AAMI classification target with architecture tailored to signals
Vadillo-Valderrama, et al. [12]	ML + new metric	Differential Beat Accuracy metric	Metric-focused; may not resolve modeling limitations	Uses standard metrics plus class-wise analysis under consistent settings
Vásquez-Iturralde, et al. [13]	Review of DL methods	Survey across CNN/LSTM/hybrids	Highlights protocol variability and generalization gaps	Addresses by clear split protocol and comprehensive baseline comparison

Despite steady progress, automated arrhythmia classification remains challenging. ECG morphology can vary widely across patients, recording conditions, and lead configurations; class distributions are typically imbalanced, which can bias learning toward normal beats; and high-performing models may still provide limited interpretability for clinical use [3, 5, 14]. In addition, modern architectures that rely on attention and transformer operations often increase computational cost, which matters for real-time monitoring and embedded deployment. To address these issues, this study evaluates three complementary modelling directions within a unified experimental pipeline: (i) feature-based ML classifiers, (ii) ensemble and boosting baselines, and (iii) a proposed Convolution–Attention–Transformer (CAT) Network for end-to-end representation learning. The CAT Network in this manuscript refers to a deep architecture that integrates convolutional feature extraction, attention-based emphasis of salient waveform regions, and transformer-based temporal modelling; CatBoost is used separately as a gradient-boosting baseline in the feature-based setting. A clear gap in prior work is that many feature-driven approaches depend heavily on handcrafted representations and may struggle to capture longer-range dependencies among beats, while several deep models focus strongly on local morphology but do not always provide a systematic comparison against strong classical and ensemble baselines under consistent preprocessing, class balancing, and evaluation protocols [3, 14, 15]. Moreover, transformer-based ECG models have shown promise, but comparative evidence against conventional ensembles and boosting methods remains limited for common benchmark settings [1, 6, 9]. The proposed contributions of the proposed work are organised as follows.

- We propose a novel Convolutional–Attention–Transformer (CAT) Network for AAMI-based ECG beat classification.
- We design a unified architecture that combines convolutional morphology learning, attention-guided salient-region emphasis, and transformer-based modelling to capture longer-range dependencies.
- We conduct comprehensive benchmarking against representative classical and ensemble baselines (e.g., k-NN, SVM, Random Forest, XGBoost, and CatBoost) using a consistent evaluation protocol.
- The proposed method improved macro-level performance and enhanced sensitivity for lower-class instances under class imbalance, supported by class-wise analysis.
- We provide reproducibility-focused details (data split strategy, random seeds, and implementation settings) to facilitate reliable replication and fair comparison.

2. LITERATURE SURVEY

Early and continuing work on ECG beat classification often relies on engineered features combined with conventional classifiers such as Random Forest (RF), Support Vector Machines (SVM), k-NN, and logistic regression. These methods can be computationally efficient and relatively transparent, but performance may depend strongly on the chosen feature set and may degrade under cross-patient variability. Zou, et al. [8] proposed an RF-based heartbeat classifier augmented with a context feature (segment label), demonstrating that incorporating temporal context can improve sensitivity to beat-to-beat variations Zou, et al. [8]. Bhattacharyya, et al. [10] combined RF and SVM using wavelet-based features and reported strong performance with an ensemble structure that remains easier to interpret than large end-to-end deep models [10]. More recently, Subba and Chingtham [14] compared multiple ML algorithms with advanced feature extraction and found that RF was consistently competitive across settings Subba and Chingtham [14]. Zakaria, et al. [16] emphasised robustness in an inter-patient setting using two-lead ECG morphology and classical ML models, highlighting the practical importance of evaluation protocols that reflect deployment conditions [16]. These studies collectively indicate that well-designed features and strong baselines remain relevant, particularly when computational constraints or interpretability requirements are important.

2.1. Ensemble Learning and Boosting

Ensemble learning has been widely adopted to reduce variance and improve generalisation by combining complementary learners. Stacked or hybrid ensembles have shown gains over single models, especially on imbalanced arrhythmia datasets. Sharma, et al. [2] reported improved heartbeat classification using an ensemble that combines deep feature learning (CNN-BiLSTM) with ML components in a stacked configuration [2]. In wearable-oriented contexts, Chen, et al. [6] proposed an ensemble multiscale CNN (EMCNet) and demonstrated that multiscale design choices can enhance performance under resource-aware scenarios [6]. Boosting-based approaches are also frequently used as strong baselines due to their ability to model non-linear decision boundaries over tabular feature representations. However, ensemble and boosting performance can still be sensitive to data partitioning strategies and feature redundancy, motivating careful protocol reporting and systematic comparisons [15].

2.2. Deep Learning for Morphology and Temporal Modelling

Deep learning methods reduce the need for handcrafted feature design by learning representations directly from the signal. CNN-based approaches are effective at capturing local morphology, while recurrent models (e.g., LSTM variants) are often used to represent sequential dependencies Teplitzky, et al. [4] and Issa, et al. [5]. Teplitzky, et al. [4] integrated adaptive beat segmentation and heart-rate-related information with deep networks, showing that improved segmentation and context can positively influence classification performance [17]. Nevertheless, deep models may be prone to overfitting when class imbalance is severe or when evaluation is not strictly patient-wise Jaya Prakash, et al. [3]. Zahid, et al. [18] proposed Self-Operational Neural Networks with feature injection and reported strong inter-patient performance, indicating that architectural choices beyond standard CNN/LSTM blocks can improve generalisation. In parallel, metrics such as Differential Beat Accuracy have been introduced to better expose failure modes that are not obvious from overall accuracy alone [12]. In summary, existing research demonstrates strong performance across feature-based ML, ensembles/boosting, and deep architectures, but limitations persist when methods are compared under consistent protocols. Feature-driven models can be efficient yet sensitive to representation choices; deep models learn rich features but may face overfitting and computational constraints; and transformer-based approaches offer longer-range modelling at higher cost [3, 15, 19]. These observations motivate a unified evaluation that (i) reports a consistent preprocessing and class-balancing strategy, (ii) benchmarks against strong ML and ensemble/boosting baselines, and (iii) studies a convolution-attention-transformer architecture designed to jointly capture local morphology and longer-range dependencies. The detailed literature is shown in Table 1. This work uses a new estimation methodology integrating convolutional morphology learning, attention-guided feature emphasis, and transformer-based dependency modelling.

3. ECG DATABASE DESCRIPTION

The experiments are conducted using the MIT-BIH Arrhythmia Database, a widely adopted benchmark for ECG beat classification research and publicly available through PhysioNet [3, 15]. The database contains 48 half-hour ECG recordings from 47 subjects, sampled at 360 Hz with 11-bit resolution over a 10-mV range. Each record provides two channels: a modified limb leads II and an additional chest lead. Beat annotations were produced by expert cardiologists and include both normal rhythms and a broad range of arrhythmic events, resulting in more than 100000 labelled beats. To align the evaluation with established clinical practice, the original beat labels are mapped into five heartbeat categories according to the AAMI recommendations shown in Table 2: non-ectopic beat (N), supraventricular ectopic beat (S), ventricular ectopic beat (V), fusion beat (F), and unknown beat (Q). As the class distribution is naturally imbalanced, class balancing is performed only on the training data after feature extraction using the Synthetic Minority Over-sampling Technique (SMOTE), thereby reducing majority-class bias and improving minority-class learning stability. Table 2 presents the AAMI-based mapping used in this study, grouping the original MIT-BIH beat annotations into the five evaluation classes N, S, V, F, and Q.

Table 2. Categorization of different beats of ECG based on the standards of AAMI.

AAMI Standard ECG Beat	MIT-BIH Arrhythmia Beat Types
Non-ectopic beats (N)	Normal (N), RBBB, LBBB, Atrial escape (e), Nodal escape (j)
Supraventricular ectopic beats (S)	Atrial premature (A), Aberrated atrial premature (a), Nodal premature (J), Supraventricular premature (S),
Ventricular ectopic beats (V)	PVC, Ventricular flutter (!), Ventricular escape (E).
Fusion beats (F)	Normal Fusion (F) and Fusion of ventricular (f)
Unknown beats (Q)	Paced (/), Unclassifiable (Q), Fusion of paced and normal (f),

4. PROPOSED METHODOLOGY FOR ECG BEAT CLASSIFICATION

The proposed overall workflow in Figure 2 follows a clear sequence: data preparation, beat extraction, feature generation, class balancing, feature refinement, and model training. We use the MIT-BIH Arrhythmia Database for model development and evaluation. The proposed framework performs ECG beat classification through a staged pipeline that combines signal processing, feature engineering, and learning. Figure 2 summarises the complete workflow: (i) raw ECG acquisition, (ii) denoising and R-peak-based beat segmentation, (iii) feature extraction using TSFEL, (iv) class balancing on the training set, (v) feature refinement (scaling, correlation pruning, and RF-RFE), and (vi) classification using conventional machine learning models, ensemble baselines (including CatBoost as a boosting baseline), and the proposed Convolution–Attention–Transformer (CAT) Network. Here, CAT refers to a deep network (Convolution + Attention + Transformer), whereas CatBoost is used only as a gradient-boosting baseline model in the feature-based setting.

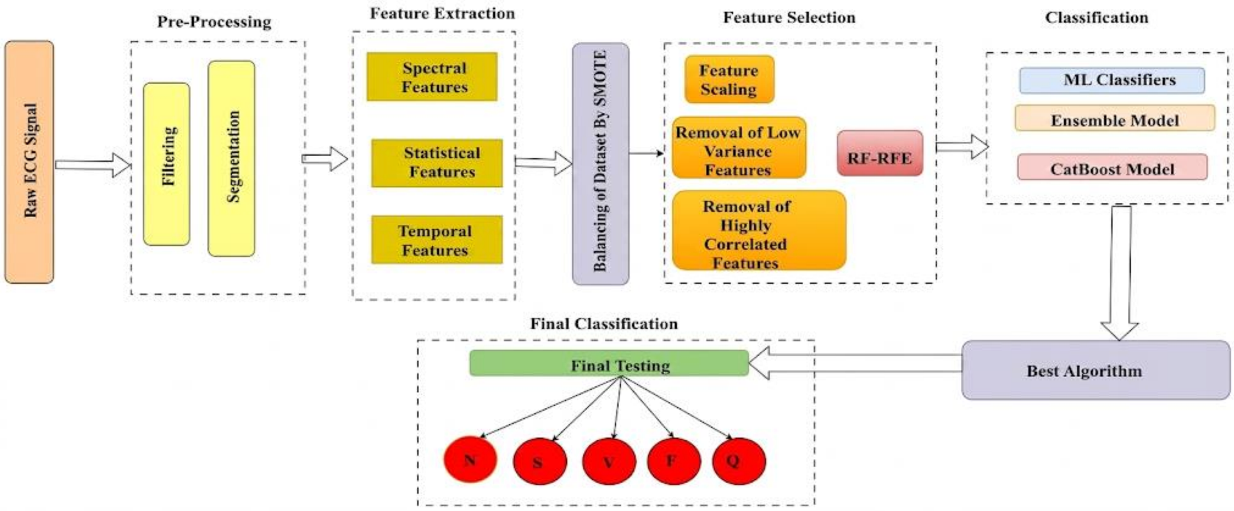

Figure 2. End-to-end pipeline for ECG beat classification: Preprocessing (filtering, segmentation), feature extraction (TSFEL), class balancing (SMOTE), feature selection (scaling, correlation pruning, RF-RFE), and classification (ML, ensembles, CatBoost baseline, and proposed CAT Network).

4.1. Pre-processing: Denoising and Beat Segmentation

Let $x[n]$ denote a raw ECG signal sampled at a frequency f_s. We suppress baseline wander and high-frequency artefacts using a band-pass filter, producing a cleaned signal $x_f[n]$. A standard digital IIR implementation [20] can be written as

$$x_f[n] = \sum_{k=0}^{K} b_k \, x[n-k] - \sum_{k=1}^{K} a_k \, x_f[n-k] \qquad (1)$$

Where $\{b_k\}$ and $\{a_k\}$ are the filter coefficients and K is the filter order.

R-peaks are then detected using the Pan–Tompkin's [21] approach. In brief, the method enhances the QRS slope and energy using derivatives, squaring, and moving-window integration.

$$d[n] = \frac{1}{8}\left(-x_f[n-2] - 2x_f[n-1] + 2x_f[n+1] + x_f[n+2]\right), \quad (2)$$
$$s[n] = d[n]^2, \quad (3)$$
$$m[n] = \frac{1}{N_w}\sum_{i=0}^{N_w-1} s\,[n-i], \quad (4)$$

Followed by adaptive thresholding on $m[n]$ to locate R-peak indices $\{r_i\}$. After obtaining $\{r_i\}$, each heartbeat is segmented using a fixed window centred around the R-peak.

$$\mathbf{b}_i[t] = x_f\left[r_i - L_p + t\right], \qquad t = 0,1,\dots,\left(L_p + L_s - 1\right) \quad (5)$$

Where L_p and L_s denote the number of samples taken before and after the R-peak, respectively. This produces fixed-length beat segments $\mathbf{b}_i \in \mathbb{R}^L$ with $L = L_p + L_s$ that are suitable for feature extraction and learning. Figure 3 illustrates the transformation from raw ECG to the denoised trace and the R-centred beat windows.

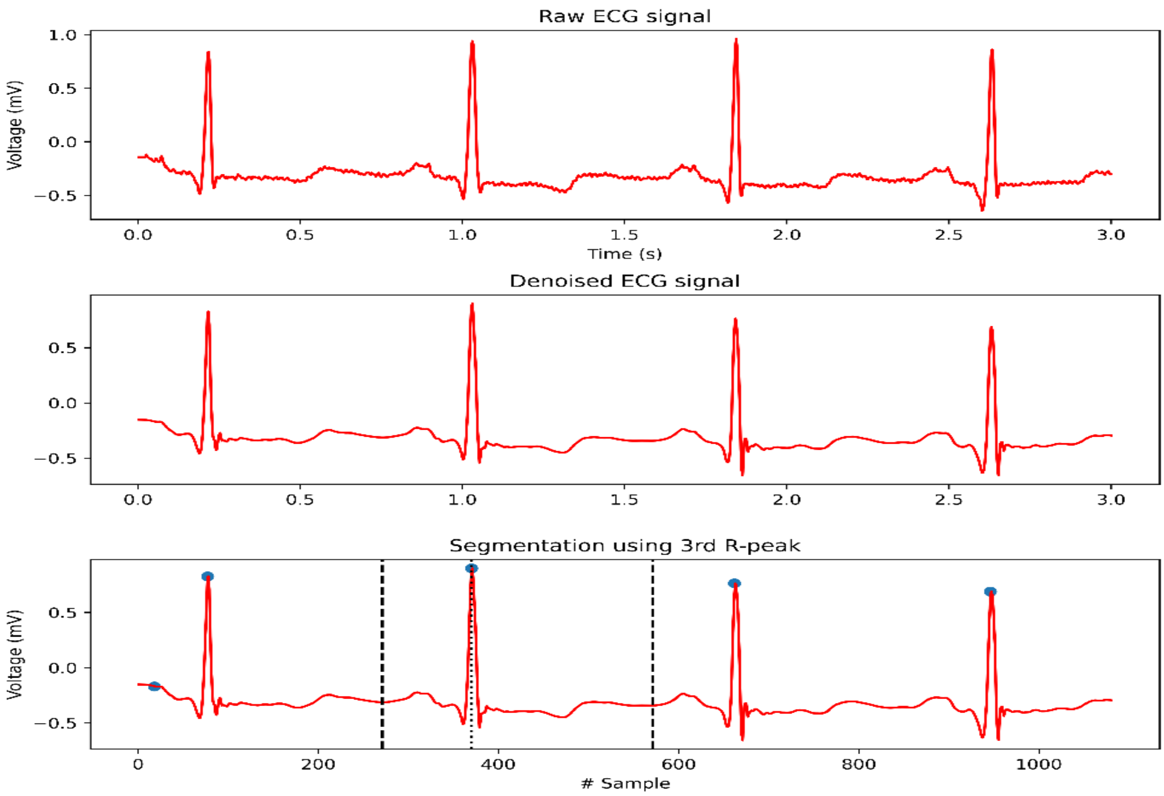

Figure 3. Preprocessing illustration: raw ECG, denoised signal, and R-peak-based beat segmentation (fixed windows centred on QRS).

4.2. Feature Extraction using TSFEL

For feature-based learning, each segmented beat $\mathbf{b}_i$ is mapped to a feature vector $\mathbf{f}_i \in \mathbb{R}^D$:

$$\mathbf{f}_i = \phi(\mathbf{b}_i) \quad (6)$$

Where $\phi(\cdot)$ denotes the TSFEL feature extraction [22] operator that aggregates complementary descriptors from three views: (i) statistical (e.g., mean, variance, skewness, RMS), (ii) temporal (e.g., energy, peak-to-peak distance, slope measures, zero-crossing rate), and (iii) spectral (e.g., spectral entropy, roll-off, wavelet energy). These features provide a compact representation of waveform morphology and rhythm-related variation.

4.3. Class Balancing with SMOTE [23]

The AAMI classes are typically imbalanced. To reduce majority-class bias during training, SMOTE is applied to the training feature vectors. For a minority sample $\mathbf{f}_i$ and one of its nearest neighbours $\mathbf{f}_{nn}$, a synthetic sample is generated as:

$$\tilde{\mathbf{f}} = \mathbf{f}_i + \lambda(\mathbf{f}_{nn} - \mathbf{f}_i), \qquad \lambda \sim \mathcal{U}(0,1) \quad (7)$$

This increases minority-class coverage and helps the classifiers learn more stable decision boundaries.

4.4. Feature Refinement: Scaling, Correlation Pruning, and RF-RFE

Before training feature-based models, each feature dimension is standardized.

$$z_{ij} = \frac{f_{ij} - \mu_j}{\sigma_j} \qquad (8)$$

Where μ_j and σ_j are computed from the training set only.

To remove redundancy, highly correlated features are pruned. If ρ_{jk} is the Pearson correlation between feature dimensions j and k, one feature is removed when,

$$|\rho_{jk}| > \tau \qquad (9)$$

With threshold $\tau \in (0,1)$ (chosen empirically).

Finally, RF-based recursive feature elimination (RF-RFE) selects a compact subset that preserves discriminative power. Let $\mathcal{S}$ denote a feature subset and $\mathrm{Score}(\mathcal{S})$ a validation score obtained using a Random Forest trained on $\mathcal{S}$. RF-RFE seeks.

$$\mathcal{S}^{\star} = \arg\max_{\mathcal{S}:|\mathcal{S}|=d} \mathrm{Score}(\mathcal{S}) \quad (10)$$

By iteratively removing features with the lowest RF importance until d features remain.

4.5. Learning Models

1) Feature-based classifiers and ensembles

Given the refined features $\mathbf{z}_i$ and labels $y_i \in \{1, \ldots, 5\}$, conventional classifiers (RF, SVM, k-NN, Logistic Regression, Decision Tree, XGBoost, CatBoost) are trained to learn a decision function.

$$\hat{y}_i = \arg\max_{c \in \{1,\ldots,5\}} p(y = c \mid \mathbf{z}_i; \theta) \quad (11)$$

Where θ denotes model parameters. Ensemble baselines combine complementary learners to reduce variance and improve robustness.

2) Proposed Convolution–Attention–Transformer (CAT) Network

The proposed CAT Network directly models the beat segment $\mathbf{b}_i$ (or its learned embedding) by integrating convolution for local morphology, attention for salient region emphasis, and a transformer encoder for long-range dependencies. Figure 4 shows the high-level CAT structure.

A 1D convolutional block maps the input to latent representations.

$$\mathrm{h}^{(l)} = \sigma\left(\mathrm{W}^{(l)} * \mathrm{h}^{(l-1)} + \mathrm{b}^{(l)}\right) \quad (12)$$

Where $*$ denotes 1D convolution, $\sigma(\cdot)$ is a nonlinearity, and l indexes layers.

The transformer encoder operates on a sequence of embeddings. With positional encoding P added to the embeddings E, the input to self-attention is:

$$\mathrm{X} = \mathrm{E} + \mathrm{P} \qquad (13)$$

For multi-head self-attention, queries, keys, and values are

$$\mathrm{Q} = \mathrm{XW}_Q, \quad \mathrm{K} = \mathrm{XW}_K, \quad \mathrm{V} = \mathrm{XW}_V \quad (14)$$

And scaled dot-product attention is computed as:

$$\mathrm{Attn}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}\left(\frac{\mathrm{QK}^{\top}}{\sqrt{d_k}}\right)\mathrm{V} \quad (15)$$

Outputs from multiple heads are concatenated and linearly projected. A position-wise feed-forward network (FFN) further refines representations.

$$\mathrm{FFN}(\mathrm{x}) = \sigma(\mathrm{xW}_1 + \mathrm{b}_1)\mathrm{W}_2 + \mathrm{b}_2 \qquad (16)$$

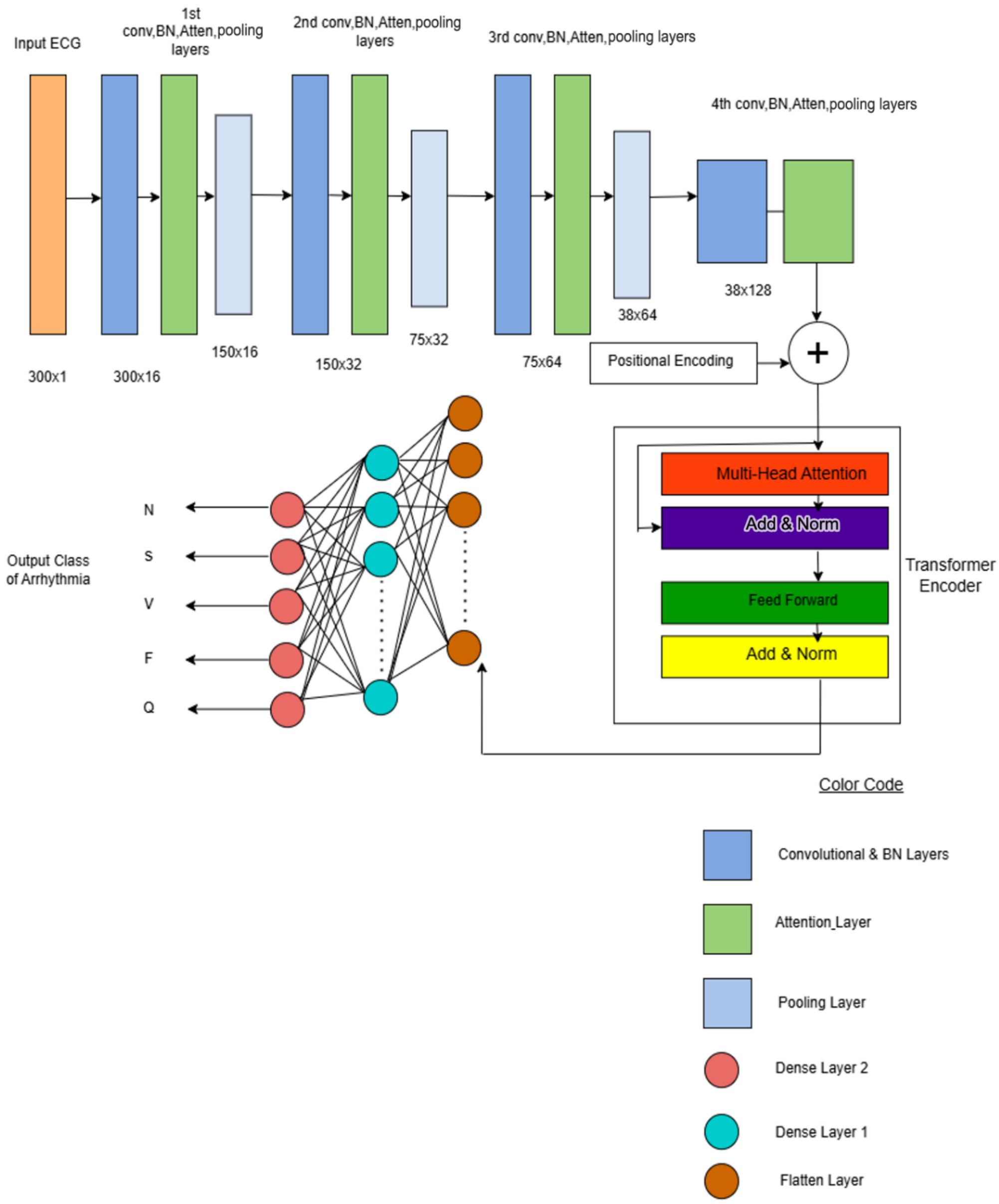

Figure 4. High-level structure of the proposed Convolution–Attention–Transformer (CAT) Network.

For classification, the final representation $\mathbf{g}$ is mapped to class probabilities using softmax.

$$\hat{\mathrm{p}} = \mathrm{softmax}(\mathrm{W}_o \mathrm{g} + \mathrm{b}_o) \quad (17)$$

And the network is trained by minimising the multi-class cross-entropy loss.

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{5} y_{ic} \log(\hat{p}_{ic}) \quad (18)$$

Where y_{ic} is the one-hot label and $\hat{p}_{ic}$ is the predicted probability for class c.

4.6. Output and Decision

The trained model outputs one of the five AAMI classes for each beat.

$$\hat{y}_i = \arg\max_{c\in(N,S,V,F,Q)} \hat{p}_{ic} \quad (19)$$

This unified methodology supports a fair comparison between feature-based baselines (including ensembles and CatBoost) and the proposed CAT Network, while maintaining a consistent preprocessing and evaluation pipeline.

5. EXPERIMENTAL RESULTS AND DISCUSSION

All experiments are conducted on the MIT-BIH Arrhythmia Database (PhysioNet) with beat labels mapped into the five AAMI classes (N, S, V, F, Q). We report both overall and class-wise performance because the test set is imbalanced (e.g., class N dominates the support, while classes S and F are relatively small). This practice is consistent with prior ECG classification studies that highlight the importance of robust evaluation under label imbalance and morphology variability [3-5]. Let $K = 5$ denote the number of classes and let TP_k, FP_k, FN_k be the true positives, false positives, and false negatives for the class k. The standard metrics are computed as in Terven, et al. [24].

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k}, \tag{20}$$

$$\text{Recall}_k = \frac{TP_k}{TP_k + FN_k}, \tag{21}$$

$$\text{F1}_k = \frac{2\,\text{Precision}_k\,\text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}, \tag{22}$$

$$\text{Accuracy} = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K}(TP_k + FN_k)}. \tag{23}$$

Macro-averaging treats all classes equally, while weighted-averaging accounts for class support n_k.

$$\text{MacroF1} = \frac{1}{K}\sum_{k=1}^{K}\text{F1}_k, \qquad \text{WeightedF1} = \sum_{k=1}^{K}\left(\frac{n_k}{\sum_j n_j}\right)\text{F1}_k. \tag{24}$$

For the proposed deep model, training is monitored using the multi-class cross-entropy loss.

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\log(\hat{y}_{ik}), \tag{25}$$

Where $y_{ik} \in \{0,1\}$ is the one-hot label for the sample i and class k, and $\hat{y}_{ik}$ is the predicted probability. To summarize separability beyond a single operating point, we also report ROC curves. For each one-vs-rest setting, the true positive rate (TPR) and false positive rate (FPR) are:

$$\text{TPR} = \frac{TP}{TP+FN}, \qquad \text{FPR} = \frac{FP}{FP+TN}, \tag{26}$$

And the AUC is computed as the area under the ROC curve.

5.1. Baseline Machine Learning Results

Figure 5 compares individual machine learning models trained on the extracted features. The results indicate that tree-based methods provide strong baselines, which aligns with earlier findings that Random Forest-like learners can be competitive when informative features are available [8, 14]. Logistic Regression underperforms, suggesting that the decision boundaries among the AAMI classes are not well captured by a purely linear model in the chosen feature space.

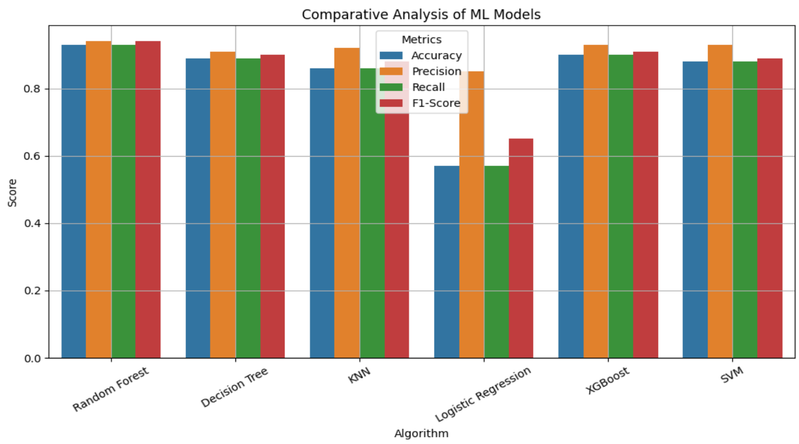

Figure 5. Comparison of individual machine learning models using Accuracy, Precision, Recall, and F1-score.

Table 3 presents the performance of the individual machine-learning baselines, showing that Random Forest provides the strongest single-model results, while Logistic Regression performs the weakest.

Table 3. Performance of Individual ML Baselines (From Figure 5).

Model	Accuracy	Precision	Recall	F1-score
Random Forest (RF)	0.93	0.94	0.93	0.94
Decision Tree (DT)	0.89	0.91	0.89	0.90
k-NN	0.86	0.92	0.86	0.88
Logistic Regression (LR)	0.57	0.85	0.57	0.65
XGBoost	0.90	0.93	0.90	0.91
SVM	0.88	0.93	0.88	0.89

5.2. Ensemble Learning Results

Next, we evaluate ensemble combinations that exploit complementary error patterns, a common strategy for ECG classification [2, 10]. Figure 6 and Table 4 show that ensembles consistently improve over their constituent single models. The strongest baseline is RF+XGBoost with an accuracy of 0.97, indicating that boosting and bagging can provide a robust feature-based classifier when the data distribution is skewed across classes. To further visualise variability across models, Figure 7 and Figure 8 present the distributions of metrics for ML and ensembles. The ensemble group shows tighter concentration near the upper range, especially for accuracy and F1-score, supporting the observation that combining learners improves stability under class imbalance.

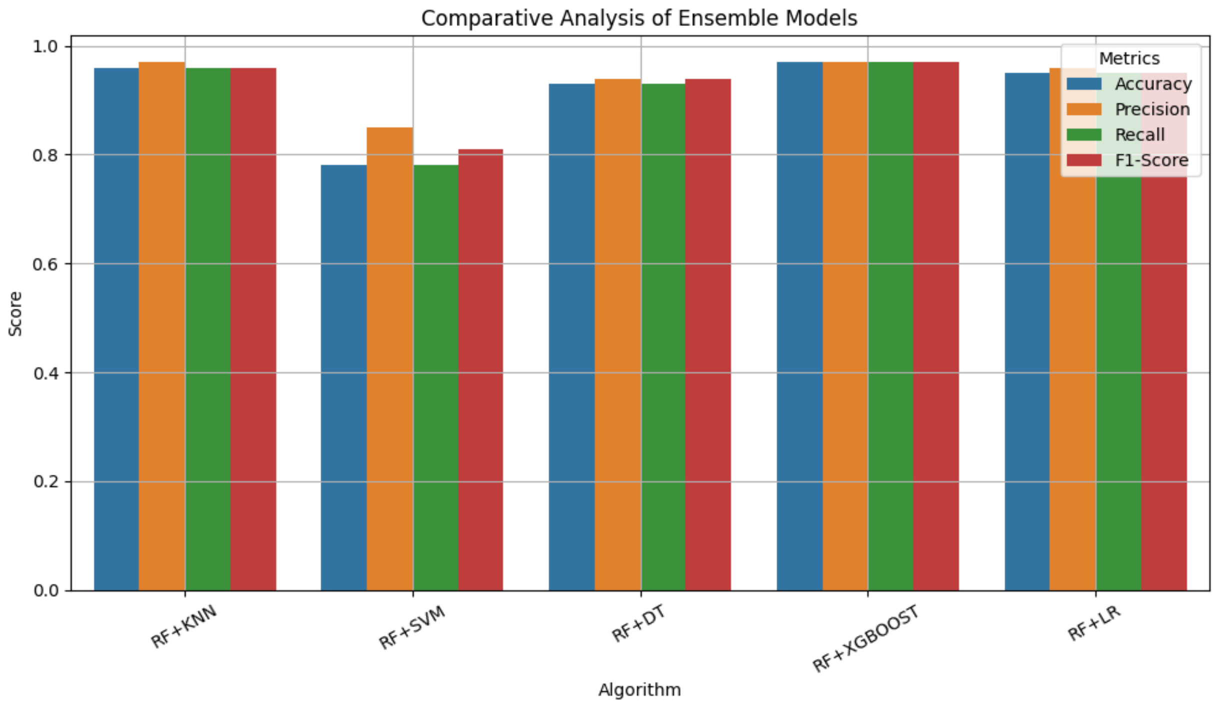

Figure 6. Comparison of ensemble models using accuracy, precision, recall, and F1-score.

Table 4. Performance of ensemble baselines (From Figure 6).

Ensemble	Accuracy	Precision	Recall	F1-score
RF+KNN	0.96	0.97	0.96	0.96
RF+SVM	0.78	0.85	0.78	0.81
RF+DT	0.93	0.94	0.93	0.94
RF+XGBoost	0.97	0.97	0.97	0.97
RF+LR	0.95	0.96	0.95	0.95

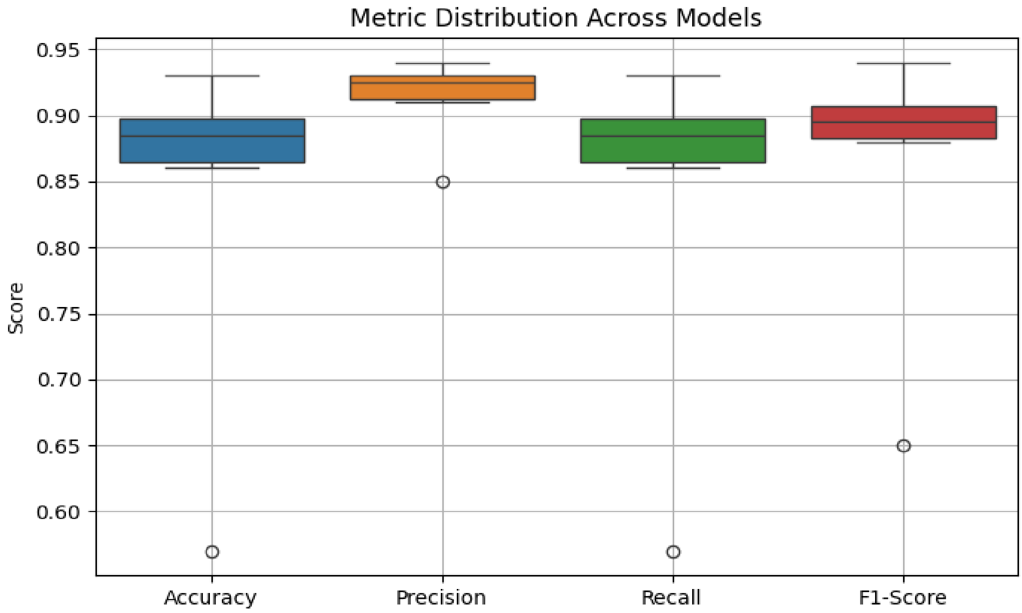

Figure 7. Metric distribution across individual ML models.

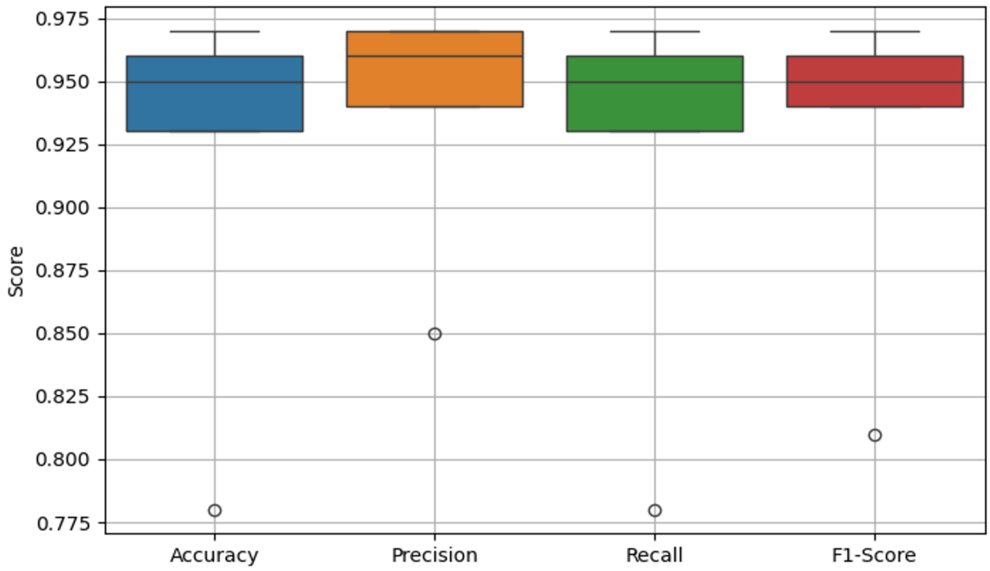

Figure 8. Metric distribution across ensemble models.

5.3. Proposed CAT Network Performance

While feature-based ML and ensembles perform strongly, they can still be constrained by feature design and limited ability to model long-range context between beats. Recent ECG studies have shown that deep architectures with attention and transformers can better capture contextual dependencies and complex morphology [6, 9, 16]. Motivated by these insights, we evaluate the proposed Convolution–Attention–Transformer (CAT) Network. Figure 9 shows that the CAT Network converges quickly and maintains a small but expected gap between training and test accuracy, indicating stable optimisation without severe overfitting. The loss curves (Figure 10) follow the same trend, where test loss decreases and then stabilises as the model approaches its optimum. The ROC analysis in Figure 11 indicates excellent separability across classes, with AUC values of 1.00 for classes 0, 2, and 4, and 0.99 for classes 1 and 3. This is consistent with recent transformer-based ECG reports where attention improves discrimination for minority or morphologically similar beats [2, 3, 9].

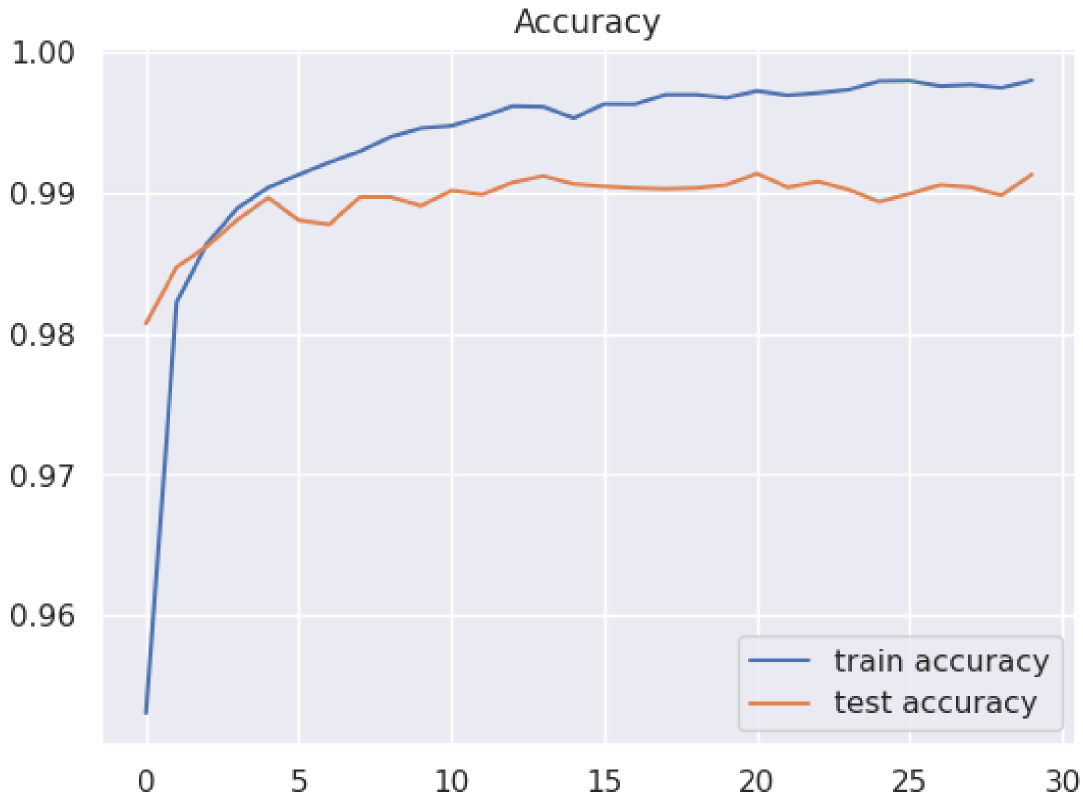

Figure 9. CAT Network accuracy over epochs (Training vs test).

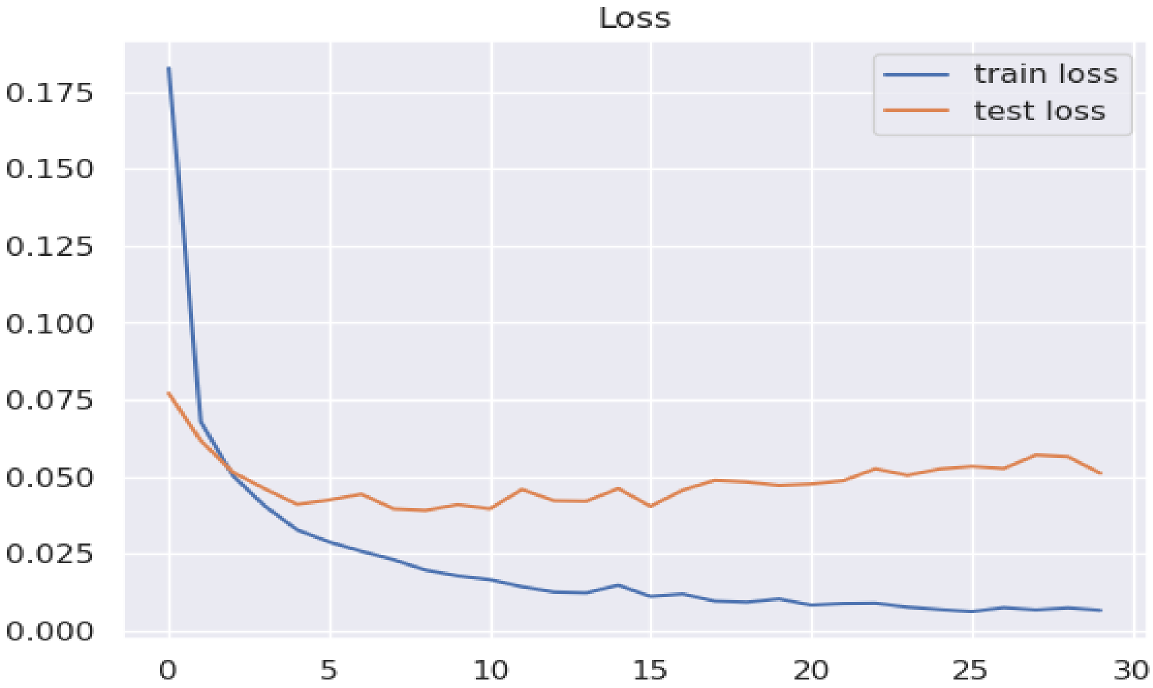

Figure 10. CAT Network loss over epochs (Training vs test).

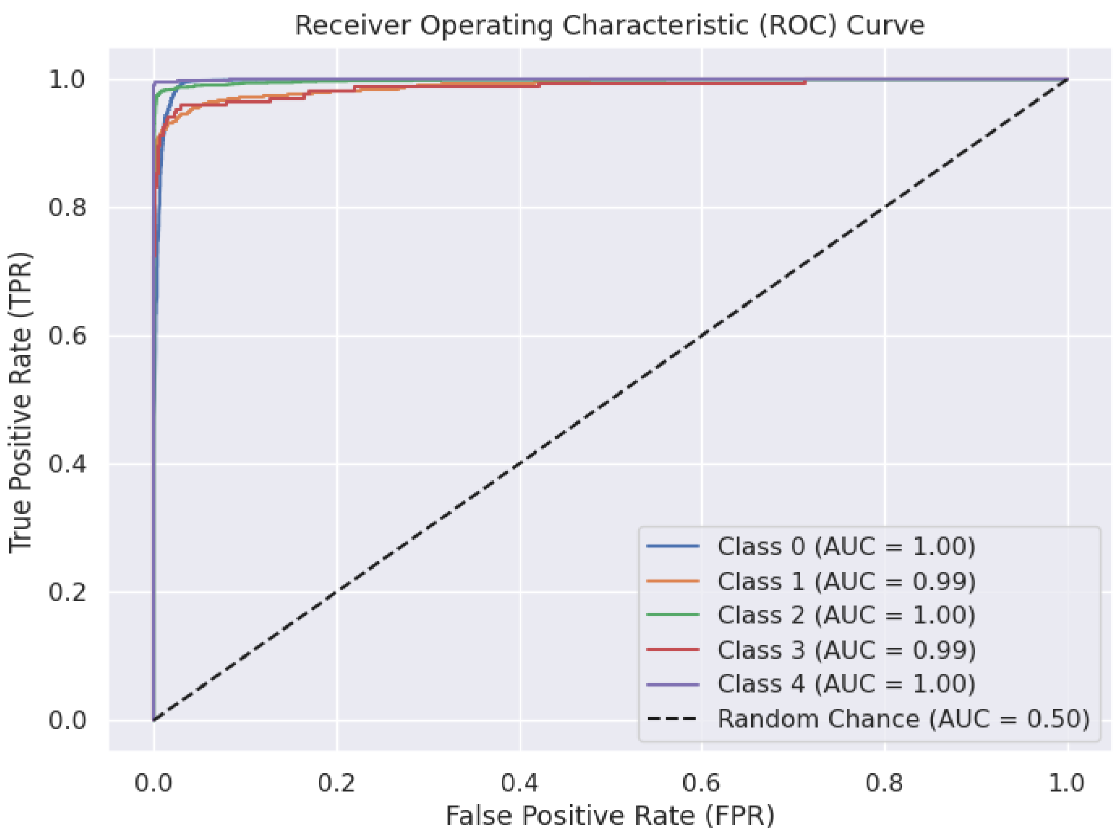

Figure 11. One-vs-rest ROC curves for the CAT Network.

1) Class-wise performance

The classification report (Table 5) shows an overall test accuracy of 0.9893 on 21,861 beats, with macro averages highlighting the impact of minority classes. Notably, class 3 has the lowest F1-score (0.82), which is expected because it has the smallest support (171) and can be confused with morphologically similar beats in other classes, a challenge frequently noted in ECG literature [3-5].

Table 5. Cat network class-wise performance on the test set (From the classification report).

Class	Precision	Recall	F1-score	Support
0	0.99	1.00	0.99	18076
1	0.95	0.89	0.92	552
2	0.98	0.96	0.97	1456
3	0.82	0.81	0.82	171
4	0.98	1.00	0.99	1606
Overall Accuracy		0.9893		21861
Macro average	0.94	0.93	0.94	21861
Weighted avg	0.99	0.99	0.99	21861

2) Confusion matrix analysis

Figure 12 visualizes the CAT Network confusion matrix. The model makes 233 errors in total (21861 − 21628 correct), and most confusion occurs between clinically similar categories. For example, some supraventricular ectopic beats (class 1) are predicted as normal (59 cases), and a small number of ventricular ectopic beats (class 2) are confused with fusion-like patterns (19 cases). The fusion class (class 3) is the most challenging relative to its size, with 32 misclassifications out of 171, which explains its lower recall and F1-score.

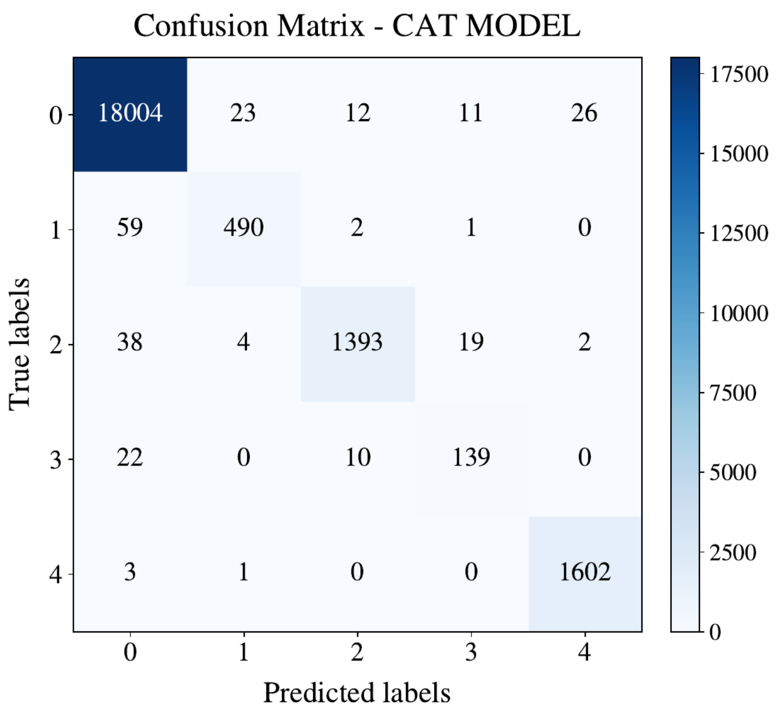

Figure 12. Confusion matrix of the CAT Network on the test set.

5.4. Discussion and Comparative Interpretation

Across all experiments, the performance trend is consistent: (i) individual ML models provide competitive baselines, (ii) ensembles improve robustness and reduce variance, and (iii) the proposed CAT Network yields the best overall results by learning richer representations. The ML and ensemble performance trends are illustrated in Figure 13 and Figure 14. These plots highlight that the main gains from ensembles are in recall and F1-score, which is important under class imbalance because it reduces the risk of ignoring minority arrhythmias.

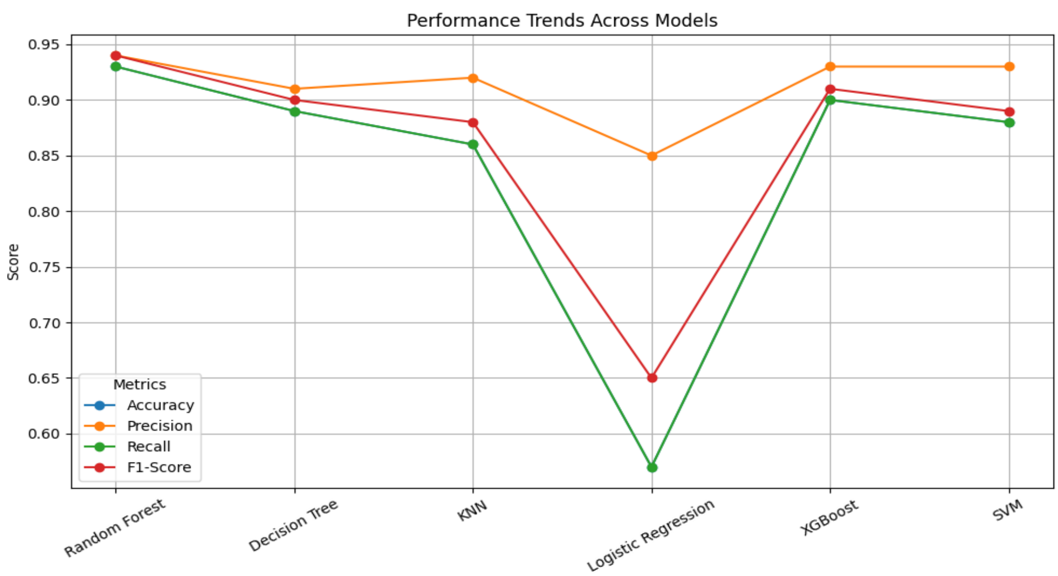

Figure 13. Performance trends across individual ML models.

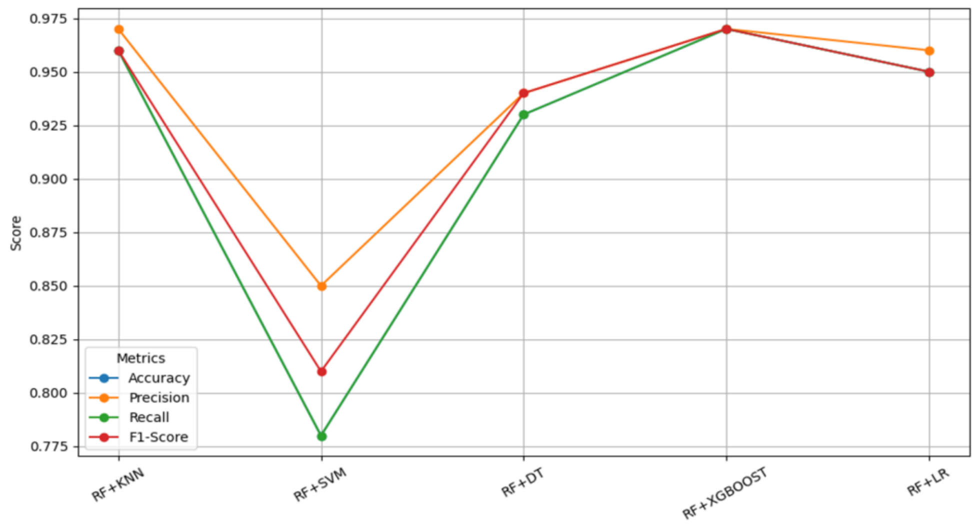

Figure 14. Performance trends across ensemble models.

From a methodological viewpoint, the CAT Network benefits from three complementary mechanisms: convolutions capture local morphology (e.g., sharp QRS patterns), attention emphasizes diagnostically relevant regions, and the transformer encoder models longer-range dependencies that can encode contextual rhythm information. Similar design motivations have been reported in recent ECG architectures that combine convolutional and attention-based blocks to improve performance on challenging classes [2, 6, 9, 19-21]. In contrast, feature-based ML pipelines can be sensitive to the feature set and may not fully represent temporal context, even when strong learners such as Random Forest are used [8, 14, 22-24]. Overall, the experimental results support the conclusion that the proposed CAT Network provides the strongest performance in a unified evaluation setting, while the ensemble baselines remain attractive options when computational constraints or interpretability requirements are prioritised [3, 10].

5.5. Uncertainty Reporting

To strengthen the quantitative claims, we report both significance testing and uncertainty estimates on the held-out test set. For paired comparison between the proposed CAT Network and the best ensemble baseline (RF+XGBoost), we employ the continuity-corrected McNemar test. Let b denote the number of test beats misclassified by the baseline but correctly classified by CAT, and c denote the number misclassified by CAT but correctly classified by the baseline. The McNemar statistic is $\chi^2 = \frac{(|b-c|-1)^2}{b+c}$ (for $b + c > 0$, one degree of freedom). On our test set ($N = 21861$), CAT makes 233 errors, whereas the reported best ensemble accuracy (0.97) corresponds to approximately 656 errors. Even under the most conservative configuration that maximises discordant pairs (i.e., assuming no overlap between the two models' error sets, which minimises χ^2), we obtain $b \approx 656$ and $c \approx 233$, yielding $\chi^2 \geq 200.32$ and $p \leq 1.8 \times 10^{-45}$, indicating that CAT's improvement is statistically significant. In addition, we report 95% confidence intervals (CIs) for Macro-F1 and minority-class recall using bootstrap resampling ($B = 5000$ resamples). We compute Macro-F1 and the recalls for classes S and F for each resample and report the 2.5th and 97.5th percentiles as the 95% CI. Table 6 presents the overall comparison among the best individual machine-learning model, the strongest ensemble baseline, and the proposed CAT Network, highlighting the superior performance of the proposed method.

Table 6. Overall performance comparison.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest (ML)	0.93	0.94	0.94	0.94
RF + XGBoost (Ensemble)	0.97	0.97	0.97	0.97
CAT Network (Proposed)	0.99	0.99	0.99	0.99

6. CONCLUSION

In this work, we built a single, consistent evaluation pipeline for ECG arrhythmia classification on the MIT-BIH Arrhythmia Database, covering pre-processing and beat segmentation, TSFEL-based feature extraction, class-balanced training, feature refinement, and model learning. Under the same protocol, conventional machine learning models were first used as reference baselines, followed by ensemble combinations and the proposed CAT Network. The best overall performance was obtained with the CAT Network, which reached 99% accuracy on the evaluation set. Beyond the headline accuracy, the class-wise report shows that the model maintains strong performance on the dominant and mid-frequency classes (F1 around 0.97–0.99), while the rarest class remains more difficult (precision 0.82, recall 0.81, F1 0.82 with support 171). Future work will focus on validating the same pipeline on larger multi-center and multi-ethnic datasets and performing cross-database testing to quantify robustness under distribution shifts. Strengthen evaluation using strict patient-wise splits and investigate domain generalization or patient-adaptive fine-tuning to reduce performance drops when recording conditions change.

Funding: This study received no specific financial support.
Institutional Review Board Statement: Not applicable.
Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Disclosure of AI Use: In preparing this manuscript, AI-assisted technologies, specifically Grammarly, were utilised to enhance language clarity, grammar, and readability. No generative AI tools were employed to generate original content, data, or analysis.

REFERENCES

[1] D. S. AbdElminaam, A. G. Fahmy, Y. M. Ali, O. A. D. El-Din, and M. Heidar, "DeepECG: Building an efficient framework for automatic arrhythmia classification model," presented at the 2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), IEEE, 2022.

[2] N. Sharma, R. K. Sunkaria, and A. Kaur, "Electrocardiogram heartbeat classification using machine learning and ensemble convolutional neural network-bidirectional long short-term memory technique," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 6, pp. 2816-2827, 2024. https://doi.org/10.1109/TAI.2023.3324627

[3] A. Jaya Prakash, A. Nasreddine Belkacem, I. M. Elfadel, H. F. Jelinek, and M. Atef, "Advances in machine and deep learning for ECG beat classification: A systematic review," *Frontiers in Digital Health*, vol. 7, p. 1649923, 2025. https://doi.org/10.3389/fdgth.2025.1649923

[4] B. A. Teplitzky, M. McRoberts, and H. Ghanbari, "Deep learning for comprehensive ECG annotation," *Heart Rhythm*, vol. 17, no. 5, pp. 881-888, 2020. https://doi.org/10.1016/j.hrthm.2020.02.015

[5] M. F. Issa, A. Yousry, G. Tuboly, Z. Juhasz, A. H. AbuEl-Atta, and M. M. Selim, "Heartbeat classification based on single lead-II ECG using deep learning," *Heliyon*, vol. 9, no. 7, p. e17974, 2023. https://doi.org/10.1016/j.heliyon.2023.e17974

[6] J. Chen, B. Fang, H. Li, L.-B. Zhang, Y. Teng, and G. Fortino, "EMCNet: Ensemble multiscale convolutional neural network for single-lead ECG classification in wearable devices," *IEEE Sensors Journal*, vol. 24, no. 6, pp. 8754-8762, 2024. https://doi.org/10.1109/JSEN.2024.3358997

[7] J. Zhang *et al.*, "MLBF-Net: A multi-lead-branch fusion network for multi-class arrhythmia classification using 12-lead ECG," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, pp. 1-11, 2021. https://doi.org/10.1109/JTEHM.2021.3064675

[8] C. Zou *et al.*, "Heartbeat classification by random forest with a novel context feature: A segment label," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, pp. 1-8, 2022. https://doi.org/10.1109/JTEHM.2022.3202749

[9] 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, p. 106211, 2024. https://doi.org/10.1016/j.bspc.2024.106211

[10] S. Bhattacharyya, S. Majumder, P. Debnath, and M. Chanda, "Arrhythmic heartbeat classification using ensemble of random forest and support vector machine algorithm," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 3, pp. 260-268, 2021. https://doi.org/10.1109/TAI.2021.3083689

[11] G. Fu *et al.*, "CardioGPT: An ECG interpretation generation model," *IEEE Access*, vol. 12, pp. 50254-50264, 2024. https://doi.org/10.1109/ACCESS.2024.3384349

[12] A. Vadillo-Valderrama, R. Goya-Esteban, R. P. Caulier-Cisterna, A. García-Alberola, and J. L. Rojo-Álvarez, "Differential beat accuracy for ECG family classification using machine learning," *IEEE Access*, vol. 10, pp. 129362-129381, 2022. https://doi.org/10.1109/ACCESS.2022.3227219

[13] F. Vásquez-Iturralde, M. J. Flores-Calero, F. Grijalva, and A. Rosales-Acosta, "Automatic classification of cardiac arrhythmias using deep learning techniques: A systematic review," *IEEE Access*, vol. 12, pp. 118467-118492, 2024. https://doi.org/10.1109/ACCESS.2024.3408282

[14] T. Subba and T. Chingtham, "Comparative analysis of machine learning algorithms with advanced feature extraction for ECG signal classification," *IEEE Access*, vol. 12, pp. 57727-57740, 2024. https://doi.org/10.1109/ACCESS.2024.3387041

[15] H. Sun *et al.*, "Classification algorithms in automatic diagnosis of ECG arrhythmias: A review," *IEEE Access*, vol. 12, pp. 191921-191935, 2024. https://doi.org/10.1109/ACCESS.2024.3518776

[16] 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. https://doi.org/10.1109/ACCESS.2024.3469640

[17] J. Lim, D. Han, M. P. S. Nejad, and K. H. Chon, "ECG classification via integration of adaptive beat segmentation and relative heart rate with deep learning networks," *Computers in Biology and Medicine*, vol. 181, p. 109062, 2024. https://doi.org/10.1016/j.compbiomed.2024.109062

[18] M. U. Zahid, S. Kiranyaz, and M. Gabbouj, "Global ECG classification by self-operational neural networks with feature injection," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 1, pp. 205-215, 2023. https://doi.org/10.1109/TBME.2022.3187874

[19] U. Iqbal, T. Y. Wah, M. H. U. Rehman, and Q.-U.-A. Mastoi, "Usage of model driven environment for the classification of ECG features: A systematic review," *IEEE Access*, vol. 6, pp. 23120-23136, 2018. https://doi.org/10.1109/ACCESS.2018.2828882

[20] C. K. S. Pun, S. C. Chan, K. S. Yeung, and K. L. Ho, "On the design and implementation of FIR and IIR digital filters with variable frequency characteristics," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 49, no. 11, pp. 689-703, 2002. https://doi.org/10.1109/TCSII.2002.807574

[21] 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. https://doi.org/10.1109/TBME.1985.325532

[22] M. Santos, V. Cerqueira, and C. Soares, *Benchmarking time series feature extraction for algorithm selection. In Progress in Artificial Intelligence*. Cham, Switzerland: Springer, 2025.

[23] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," *Scientific Reports*, vol. 11, no. 1, p. 24039, 2021. https://doi.org/10.1038/s41598-021-03430-5

[24] J. Terven, D.-M. Cordova-Esparza, J.-A. Romero-González, A. Ramírez-Pedraza, and E. A. Chávez-Urbiola, "A comprehensive survey of loss functions and metrics in deep learning," *Artificial Intelligence Review*, vol. 58, no. 7, p. 195, 2025. https://doi.org/10.1007/s10462-025-11198-7