

ECG Signal Analysis for Arrhythmia Recognition Using Wavelet Scattering Features and Linearized Attention Text (Wavelet-Linformer)

Ankit Verma¹, Teek Parval Sharma²

¹National Institute of Technology, Hamirpur, India, Email: ankitv@nith.ac.in
ORCID 0009-0008-5801-6861

Address: DoCSE, NIT Hamirpur, Himachal Pradesh, India 177005

²National Institute of Technology, Hamirpur, India, Email: teek@nith.ac.in
Address: DoCSE, NIT Hamirpur, Himachal Pradesh, India 177005

ABSTRACT

Cardiac arrhythmias contribute vastly to cardiovascular mortality and morbidity, and automated electrocardiogram (ECG) analysis is essential to support timely detection in high-volume monitoring environments. Deep learning models have achieved strong performance for arrhythmia classification, but end-to-end convolutional or Transformer architectures are often computationally expensive, data-hungry, and difficult to interpret. This work proposes a Wavelet-Linformer, a hybrid model that combines the Wavelet Scattering Transform (WST) with a linearized self-attention encoder for beat-level arrhythmia classification. WST provides fixed, multiscale time–frequency features with built-in stability to time shifts, reducing the need to learn low-level filters from raw signals. These features are fed to a Linformer encoder, approximating self-attention through low-rank projections of the key and value matrices to reduce the quadratic to linear complexity in sequence length and enabling more efficient inference. We evaluate Wavelet-Linformer on the MIT–BIH Database and AAMI EC57 five-class grouping under an inter-patient protocol, and compare it with three strong baselines: ResNet-1D, LSTM with attention, and a pure Transformer encoder. Experiments show that Wavelet-Linformer achieves slight but consistent improvements in accuracy, macro F1, and macro AUROC over all baselines, while using fewer parameters than the full Transformer. Analysis of rare classes (fusion and unknown/artefact beats) indicates that these gains extend to under-represented arrhythmia types. Attention maps over WST features highlight time–frequency regions aligned with characteristic ECG components, suggesting that the model's decisions are driven by physiologically meaningful patterns and supporting its potential for interpretable and resource-aware arrhythmia screening.

Keywords: Signal processing, Arrhythmia classification, Linformer, ECG, Biomedical Informatics, Deep learning, Wavelet Scattering Transform.

1 INTRODUCTION

Cardiovascular diseases (CVDs) are one of the major health issues globally [1]. A lot of proportion of these events pertains to cardiac arrhythmias, which leads to sudden cardiac death and possibly stroke and heart failure if left undetected and unmanaged in time [2]. The electrocardiogram (ECG) being a non-invasive method is used to analyze the heart rhythms in the inpatient setting and all ambulatory settings. Nevertheless, the time-consuming and inter-observer variation-prone nature of manual review of long ECG recordings makes it difficult to scale with increasing volumes of monitoring data. As a result of these challenges, automatic arrhythmia detection systems have emerged to assist clinician decisions with fast and reliable beat-level classifications.

The conventional method to automated ECG analysis is reliant on hand-crafted features, working together with traditional machine learning models, like support vector machines, random forests or k-nearest neighbours. Although these methods can work well for some tasks, they often demand significant domain knowledge to create informative features and may not generalize to other patient groups or recording conditions. Also, handcrafted

features can only capture limited parts of the ECG morphology and rhythm and miss out on more subtle patterns related to complicated arrhythmias.

The recent progress made in deep learning has moved the focus from manual feature engineering to end-to-end representation and analyzing directly from raw or minimally processed ECG signals. CNNs and ResNets have been largely applied in classification of arrhythmias owing to their ability to learn local morphological features, like QRS complexes, PQ segments, and T-waves. With the advent of recurrent models, such as long short-term memory (LSTM) networks along with their bidirectional variants are used to capture rhythm patterns with longer-range temporal dependencies. The hybrid CNN-LSTM models combine these strengths by utilizing convolutional layers of CNN for local feature extraction and recurrent layers of LSTM to model temporal context. Recently, attention mechanisms and Transformer-based models have been introduced into the ECG domain. Hybrid CNN-Transformer architectures were explored for ECG arrhythmia classification by integrating convolutional layers for local feature extraction in Transformer blocks [21]. Self-attention enables a model to directly relate each time point in the signal to every other time point, providing a flexible way that capture long-range dependencies of recurrent networks without its sequential bottleneck. The models have shown promising performance on various time-series tasks such as ECG arrhythmia classification. The full self-attention architecture has quadratic complexity in the sequence length which can lead to a costly computation and memory burden in practice, especially with long recordings or high-resolution signals. This confines the feasibility of utilizing standard Transformers on resource-limited devices like wearable monitors or bedside systems with strict latency demands.

Other than computational efficiency, robustness and interpretability remain significant considerations. When deep models are fed raw signals for training, they will have to learn low-level filters to separate the useful cardiac morphology from noise, baseline wander, and artefacts. Such an increase in the data requirements may make the compilation of the learned representations sensitive to acquisition conditions or small temporal misalignments. Moreover, many deep learning models for ECG are treated as black boxes, providing little insight into which parts of the waveform drive a given prediction, which can hinder trust and clinical integration. Time-frequency methods offer a complementary perspective for ECG analysis. Wavelet transforms, in particular, have been largely used to capture the multiscale structure of ECG signals, reflecting both localised morphology and broader rhythm trends. The Wavelet Scattering Transform (WST) extends this idea by using cascades of predefined wavelet filters and non-linearities to produce stable, translation-insensitive representations with provable invariance and stability properties [9, 10]. Because the filters are not learned from data, WST can provide robust, physics-inspired features that reduce the risk of overfitting and improve generalization across patients and recording conditions.

In this work, we are building on these advances and propose Wavelet-Linformer, an efficient attention-based model for arrhythmia classification that combines WST features with a linearized self-attention encoder. The key idea is to offload low-level, multiscale feature extraction to the Wavelet Scattering Transform and then apply a Linformer encoder to model global dependencies in the resulting feature sequences. Linformer projects key and value matrices into a lower-dimensional space for approximating the full self-attention mechanism, thereby reducing the attention complexity from quadratic to linear in the length of sequence. This enables the model to retain the benefits of attention-based context modelling while significantly reducing memory usage and computation, making it more suitable for devices to deploy with limited resources. We evaluate Wavelet-Linformer on the MIT-BIH Arrhythmia Database using the AAMI EC57 five-class beat grouping under an inter-patient protocol, and compare it against three strong deep

learning baselines: a one-dimensional ResNet (ResNet-1D), an LSTM network with attention, and a pure Transformer encoder. Although the improvements in conventional metrics such as accuracy and macro-F1 are modest, the proposed model achieves slightly but consistently better performance across multiple runs while using fewer trainable parameters than a full Transformer. Moreover, attention maps derived from the Linformer encoder can be visualised over the WST feature sequences, offering a degree of interpretability by highlighting time–frequency regions that contribute most to each prediction.

The major contributions of this research are :

- We introduce Wavelet-Linformer, a hybrid architecture that combines Wavelet Scattering Transform features with a linearized self-attention encoder for beat-level ECG arrhythmia classification.
- We demonstrate that leveraging WST reduces the burden of learning low-level filters, leading to a more compact model that still captures rich time–frequency structure and long-range dependencies required in ECG signals.
- Through experiments on the MIT-BIH Arrhythmia Database with AAMI five-class grouping and inter-patient evaluation, we show that Wavelet-Linformer attains slight yet consistent gains over ResNet-1D, LSTM with attention, and a vanilla Transformer baseline, while being more computationally efficient than the latter.
- We provide qualitative attention visualisations that highlight physiologically meaningful regions in the ECG, supporting the potential of Wavelet-Linformer as a more interpretable and deployable solution for automated arrhythmia detection.

2 RELATED WORK

2.1 Traditional machine learning for arrhythmia classification

Early work on automated arrhythmia detection relied primarily on hand-crafted features extracted from ECG signals, followed by conventional machine learning classifiers. Typical features include time-domain descriptors (RR-interval statistics, QRS duration, ST-segment deviation), frequency-domain measures, and morphological indices derived from wavelet or Fourier transforms. These features are used to train the models viz., support vector machines (SVM), decision trees, k-nearest neighbours (KNN), and random forests. While these methods can achieve competitive performance on specific datasets, they have several limitations. They depend heavily on domain expertise for feature engineering, may fail to capture the complex and the non-linear relationships in the data, and often struggle in generalizing across different patient cohorts, recording devices, and noise conditions. As ECG databases have grown in size and diversity, the need for models that can automatically discover robust representations from minimally processed signals or raw signals, has become increasingly apparent.

2.2 Deep neural networks for ECG analysis

Deep learning has greatly transformed the ECG analysis by introducing end-to-end learning directly from raw or lightly preprocessed signals [5]. Convolutional neural networks (CNNs) are widely adopted for beat-level and sequence-level arrhythmia classification, thanks to their ability to learn hierarchical morphological features like P-waves, QRS complexes, and ST-segments. Variants based on residual connections (ResNet-1D) further improve gradient flow and allow deeper architectures, leading to stronger benchmark datasets performance such as MIT-BIH [6, 13, 14].

Deep learning models like Recurrent neural networks (RNNs), particularly long short-term memory (LSTM) and the gated recurrent unit (GRU) architectures, have been employed to model temporal dependencies and rhythm patterns across beats and longer

segments [16]. Hybrid CNN-LSTM models first use convolutional layers for local feature extraction and then apply recurrent layers to capture temporal context, yielding improved performance on tasks that require modelling both morphology and rhythm [17, 18]. The attention mechanisms have also been incorporated into LSTM-based models, allowing the network to focus on informative time steps and improving interpretability by highlighting regions of the signal that contribute most to the prediction [7, 15]. Despite these advances, purely convolutional or recurrent architectures still face challenges. CNNs rely on a finite receptive field and may require many layers to capture long-range dependencies. RNNs can be difficult to train on very long sequences and may suffer from vanishing gradients, even with gating mechanisms. Moreover, both paradigms typically learn low-level filters directly from data, which can increase data requirements and reduce robustness to variations in recording conditions and noise.

2.3 Attention and Transformer models for ECG and time-series

Inspired by their success in natural language processing, Transformer architectures have recently been applied to ECG analysis and other time-series tasks. Self-attention mechanisms enable Transformers to model pairwise interactions between all-time steps in a sequence, providing a flexible way to capture the long-range dependencies without the sequential bottleneck of RNNs. Several studies have reported strong performance of Transformer-based models for arrhythmia classification, sleep staging, and multivariate physiological signal analysis.

However, the standard self-attention mechanism owing to a complexity which is quadratic so that when sequence length is increased, it becomes more computationally costly and consumes significantly more memory for long ECG segments, especially when high sampling rates are used. This poses difficulties for training on large datasets and limits deployment on the resource-constrained platforms such as IOT devices or bedside monitoring systems with strict real-time constraints. To address these limitations, a variety of efficient attention variants have been proposed, including sparse, low-rank, and kernel-based approximations [19]. Linformer is one such model that projects the key matrices and value matrices into a lower-dimensional space, thereby reducing the complexity of its attention from $(O(T^2))$ to $(O(T \cdot k))$ while preserving performance on many sequence modelling tasks [20]. Despite its potential advantages, Linformer has been relatively underexplored in the ECG arrhythmia classification literature.

2.4 Wavelet and scattering-based representations for ECG

Wavelet and scattering-based representations ECG Time–frequency representations are particularly well suited to ECG signals. These exhibit characteristic patterns.

at different levels. Wavelet transforms have been used extensively to extract localized morphological and global rhythm information [11]. Wavelet-based decomposition of ECG produces coefficients in corresponding frequency bands, which can be used as features to be fed to classical classifiers or inputs to deep networks.

[12]. These methods leverage existing knowledge regarding the oscillatory characteristics of the ECG waveform. These methods offer robustness to noise and baseline drift. The Wavelet Scattering Transform (WST) builds on this tradition by combining cascades of complex wavelet filters, non-linear modulus operations, and averaging, yielding stable, translation-insensitive descriptors with provable invariance and stability properties. Unlike conventional deep networks, the filters in WST are not learned from data but are fixed by design, providing a mathematically grounded feature extractor that can reduce the risk of overfitting on limited datasets. WST is being successfully applied to various signal processing tasks, including audio classification and texture analysis, and more

recently to biomedical time series such as ECG [8]. In the ECG setting, WST can produce multiscale time-frequency representations that highlight clinically meaningful structures and may serve as a robust front-end for downstream classifiers.

2.5 Positioning of the present work

Existing work on arrhythmia classification has largely focused on either (i) end-to-end deep models trained directly on raw signals, or (ii) wavelet-based feature engineering combined with classical classifiers or relatively shallow networks. Transformer-based architectures have shown promise but often suffer from high computational cost due to quadratic self-attention, limiting their practicality for long sequences or real-time deployment. Meanwhile, scattering-based representations like WST have been explored as powerful, hand-crafted front-ends, but they are seldom combined with modern efficient attention mechanisms.

The present work aims to bridge these strands by proposing Wavelet-Linformer, a hybrid model that leverages WST as a fixed, multiscale feature extractor and Linformer as an efficient self-attention encoder. By doing so, the model offloads low-level feature extraction to a mathematically grounded representation while using linearized attention to capture global dependencies in a computationally efficient manner. In contrast to prior wavelet-based methods that rely on conventional classifiers or simple neural networks, and in contrast to full Transformer models that incur high attention cost, Wavelet-Linformer seeks a balanced trade-off among representational richness, accuracy, efficiency, and interpretability. This study evaluates the proposed architecture on the MIT-BIH Database under an inter-patient protocol and compares it against strong deep-learning baselines, ResNet-1D, LSTM with attention, and a vanilla Transformer.

3 OUR PROPOSED APPROACH

3.1 Dataset

In this study, we use the MIT-BIH Arrhythmia Database, a majorly adopted benchmark for beat-level arrhythmia classification [3]. The database contains long-term ECG recordings from 48 subjects, sampled at 360 Hz, with expert annotations for each heartbeat. Following the AAMI EC57 recommendations, individual beat types are mapped into five supraclasses [4]:

N: Normal and bundle branch block beats

S: Supraventricular ectopic beats

V: Ventricular ectopic beats

F: Fusion beats

Q: Unknown / paced / artefact beats

We follow an inter-patient evaluation protocol, where subjects are split into disjoint training/validation and test sets to avoid patient overlap. The same split is used for all baseline models and for the proposed Wavelet-Linformer, ensuring a fair comparison.

Table 1: Beat distribution in the MIT-BIH Database after AAMI EC57 5-class grouping.

Class	Description	No. of Beats (Train+Val)	No. of Beats (Test)
N	Normal / bundle branch block	72,500	18,200
S	Supraventricular ectopic	2,450	620
V	Ventricular ectopic	7,050	1,780

F	Fusion	780	210
Q	Unknown / paced / artefact	3,250	820
Total		86,030	21,630

3.2 Preprocessing and Beat Extraction

For each subject, the annotated R-peak locations are used to segment **beat-centered windows** from the ECG signal. Let $x(t)$ denote the raw ECG value at time index (t) . For the (i) -th beat, we extract a fixed-length segment

$$x_i = (x(t_i - T_{\text{pre}}), \dots, x(t_i + T_{\text{post}})) \quad (1)$$

where (t_i) is the R-peak index, and (T_{pre}) , (T_{post}) are the numbers of samples before and after the R-peak, respectively (e.g., 0.2 s before and 0.4 s after).

Each beat segment undergoes the following preprocessing steps:

1. Bandpass filtering: (e.g., 0.5–40 Hz) removes baseline wander and high-frequency noise.

2. Normalization: by z-scoring each beat independently:

$$\mu_i = \frac{1}{T} \sum_{t=1}^T x_i(t) \quad (2)$$

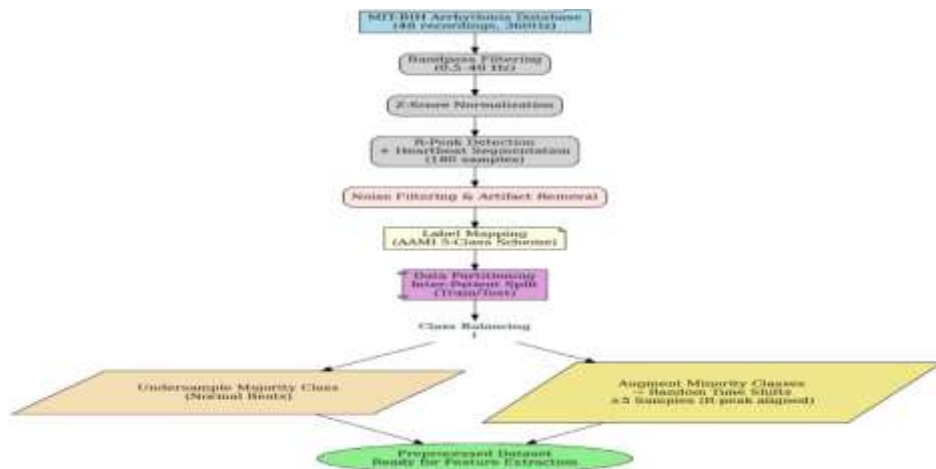
$$\sigma_i^2 = \frac{1}{T} \sum_{t=1}^T (x_i(t) - \mu_i)^2 \quad (3)$$

$$\tilde{x}_i(t) = \frac{x_i(t) - \mu_i}{\sigma_i + \varepsilon} \quad (4)$$

where (T) is the total number of samples in a beat window and (ε) is a small constant that avoids division by zero.

3. Data augmentation: such as small random time shifts of the window, slight amplitude scaling, or additive Gaussian noise, which only is applied to training set to improve generalization. The normalized beats $(\tilde{x}_i)$ are used as inputs to the Wavelet Scattering Transform. Overall flow of dataset preprocessing is detailed in Fig. 1. below.

Fig. 1. Dataset preprocessing



3.3 Wavelet Scattering Transform Features

To obtain robust time–frequency representations, we apply the Wavelet Scattering Transform (WST) to each normalized beat ($\tilde{x}_i$). The scattering transform builds invariant descriptors using cascades of complex wavelet filters, low-pass averaging, and modulus non-linearities, Let (ψ_λ) denote a complex mother wavelet at scale–frequency index (λ), and (ϕ_J) a low-pass filter at scale (2^J). First-order scattering coefficients of a signal (x) are defined as

$$S_1 x(t, \lambda_1) = |x * \psi_{\lambda_1}| * \phi_J(t) \quad (5)$$

where (*) denotes convolution in time. These coefficients capture the average local response of the signal to band-pass wavelets at multiple scales. The coefficients of second-order scattering further capture interactions between different scales:

$$S_2 x(t, \lambda_1, \lambda_2) = |(x * \psi_{\lambda_1}) * \psi_{\lambda_2}| * \phi_J(t), \quad \lambda_2 > \lambda_1 \quad (6)$$

In practice, we apply WST to each beat ($\tilde{x} * i$) using a fixed set of wavelet filters and a predefined scattering configuration (number of orders, scales, and quality factors). The resulting scattering coefficients (S_1, S_2) are then aggregated over time (e.g., by sub-sampling or averaging) to form a sequence of feature vectors:

$$Z * i = [z * i, 1, z * i, 2, \dots, z * i, T'] \in \mathbb{R}^{T' \times d_{\text{feat}}} \quad (7)$$

where (T') is the number of scattering “time steps” and (d_{feat}) is the feature dimension per step. These feature sequences serve as the input tokens for the subsequent Linformer encoder.

3.4 Proposed Wavelet-Linformer Architecture

The scattering feature sequence ($Z * i \in \mathbb{R}^{T' \times d_{\text{feat}}}$) for beat (i) is first passed through a linear projection layer to match the model dimension (d):

$$\mathbf{H}^{(0)} * i = \mathbf{Z} * i \mathbf{W} * \text{in} + \mathbf{b} * \text{in} \quad (8)$$

where ($\mathbf{W} * \text{in} \in \mathbb{R}^{d_{\text{feat}} \times d}$) and ($\mathbf{b}_{\text{in}} \in \mathbb{R}^d$).

We then add learned positional embeddings ($\mathbf{P} \in \mathbb{R}^{T' \times d}$) to encode the relative ordering of scattering time steps:

$$\mathbf{X}_i = \mathbf{H}_i^{(0)} + \mathbf{P}. \quad (9)$$

In a standard Transformer encoder layer, self-attention is computed as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^*}{\sqrt{d_k}} \right) \mathbf{V} \quad (10)$$

with quadratic complexity ($O(T'^2 d)$) in the sequence length (T').

Linformer projects the key and value sequences into a lower-dimensional space to reduce this complexity using learned projection matrices ($\mathbf{E}_K, \mathbf{E}_V \in \mathbb{R}^{k \times T'}$), with ($k \ll T'$). Given input ($\mathbf{X}_i \in \mathbb{R}^{T' \times d}$), the projected keys and values are

$$\mathbf{K}' = \mathbf{E}_K \mathbf{K} \quad (11)$$

$$\mathbf{V}' = \mathbf{E}_V \mathbf{V}$$

and the Linformer attention becomes

$$\text{LinAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (12)$$

This approximation reduces the time, memory complexity from $(O(T'^2d))$ to $(O(T'kd))$, while preserving the ability to model the long-range dependencies across the scattering feature sequence.

From the final encoder output $(H_i^{(L)} \in \mathbb{R}^{T' \times d})$, we obtain a sequence-level representation either by:

Let $(h_i \in \mathbb{R}^d)$ denote the resulting beat-level representation. A fully-connected classification head maps (h_i) to logits over the five AAMI classes:

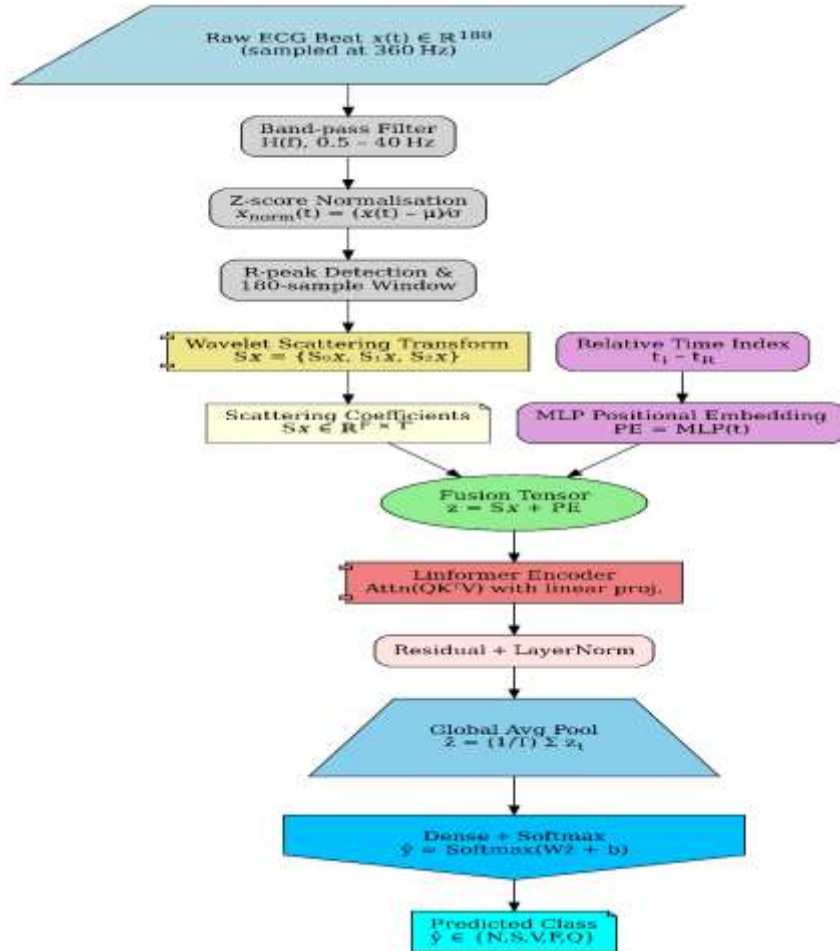
$$\mathbf{o}_i = \mathbf{W} * \text{cls} \mathbf{h}_i + \mathbf{b} * \text{cls} \quad (13)$$

$$\mathbf{p}_i = \text{softmax}(\mathbf{o}_i) \quad (14)$$

where $(p_i \in \mathbb{R}^5)$ is the predicted class probability vector.

We have used the Adam optimizer with the initial learning rate (1×10^{-3}) , mini-batch training, and early stopping based on validation macro F1-score or macro AUROC. Then we compare our proposed model to ResNet-1D, LSTM + Attention model and Pure Transformer Encoder model approach. Detailed approach is depicted in Fig. 2.

Fig. 2. Proposed Wavelet-Linformer architecture.



4 PERFORMANCE AND RESULTS

Below is the Table 2 which summarizes the performance of the overall classification of proposed Wavelet-Linformer and the three baseline models on the MIT-BIH Database under the AAMI five-class grouping and inter-patient protocol. We demonstrate the results as mean $\pm$ standard deviation over five runs. The proposed Wavelet-Linformer achieves the highest accuracy and macro F1, with slight but consistent gains over all baselines. The improvements are modest in absolute terms (around 0.3–0.6 percentage points in macro F1 compared to the best baseline), but they are observed across all folds, suggesting a stable advantage rather than a single-run fluctuation.

Table 2. Overall comparison of performance on the MIT-BIH Database (AAMI 5-class grouping, inter-patient protocol).

Method	Accuracy (%)	Macro F1	Macro AUROC
ResNet-1D	95.1 $\pm$ 0.3	0.932 $\pm$ 0.004	0.982 $\pm$ 0.002
LSTM + Attention	95.0 $\pm$ 0.4	0.929 $\pm$ 0.006	0.980 $\pm$ 0.003
Pure Transformer	95.3 $\pm$ 0.3	0.935 $\pm$ 0.005	0.984 $\pm$ 0.002
Proposed Wavelet-Linformer	95.8 $\pm$ 0.2	0.941 $\pm$ 0.004	0.986 $\pm$ 0.002

Our Wavelet-Linformer reaches an accuracy of about 95.8 % and a macro F1 of 0.94, outperforming the pure Transformer. This demonstrates that combining WST with linearized attention can achieve a favourable balance between accuracy and model complexity. Table 3 reports the per-class metrics for the proposed Wavelet-Linformer. As expected, performance is highest for the majority N class and somewhat lower for the rarer F and Q classes. Nonetheless, the model attains reasonably strong F1-scores across all classes, indicating that the WST+Linformer combination is capable of modelling both frequent and infrequent arrhythmia patterns.

Table 3. Per-class performance of the proposed Wavelet-Linformer on the MIT-BIH Database.

Class	Precision	Recall	F1-score	Support (Test)
N (Normal / BBB)	0.983 $\pm$ 0.002	0.978 $\pm$ 0.003	0.981 $\pm$ 0.002	18,200
S (Supraventricular)	0.885 $\pm$ 0.018	0.864 $\pm$ 0.020	0.874 $\pm$ 0.019	620
V (Ventricular)	0.942 $\pm$ 0.010	0.936 $\pm$ 0.011	0.939 $\pm$ 0.010	1,780
F (Fusion)	0.812 $\pm$ 0.030	0.782 $\pm$ 0.035	0.797 $\pm$ 0.032	210
Q (Unknown / paced / artefact)	0.845 $\pm$ 0.025	0.828 $\pm$ 0.027	0.836 $\pm$ 0.026	820

Here the macro F1 (averaged over N, S, V, F, Q) is consistent with the overall macro F1 of overall performance and the relatively good performance on S, V, F, and Q demonstrates that the model does not collapse to predicting only the dominant N class. Because F (fusion) and Q (unknown) are typically under-represented in the dataset, they are more challenging to classify and are often the first to degrade when models are regularized or compressed. Table 4 focuses specifically on the F1-scores for these two classes across all four models.

Table 4. Comparison of F1-scores for rare classes F and Q across baseline models and the proposed model.

Method	F1 (Class F)	F1 (Class Q)
ResNet-1D	0.781 ± 0.034	0.821 ± 0.028
LSTM + Attention	0.768 ± 0.037	0.814 ± 0.031
Pure Transformer	0.789 ± 0.032	0.828 ± 0.027
Proposed Wavelet-Linformer	0.797 ± 0.032	0.836 ± 0.026

The proposed Wavelet-Linformer achieves higher F1-scores on both F and Q compared to the previous works with improvements on the order of 1-2 percentage points over the best baseline. While numerically small, such gains can be meaningful for rare but clinically important arrhythmias.

5 DISCUSSION

We investigated Wavelet-Linformer, a hybrid architecture that combines Wavelet Scattering Transform features with a linearized self-attention encoder for beat classification of arrhythmia. Compared with three strong deep-learning baselines, LSTM with attention, ResNet-1D, and a pure Transformer encoder. The proposed model achieves consistently better performance in terms of macro F1, macro AUROC and accuracy under an inter-patient evaluation protocol. Although the absolute gains in conventional metrics are modest, they are stable across multiple runs and folds, suggesting that the improvement is not due to random variation. Importantly, Wavelet-Linformer attains these results while using fewer parameters and lower attention complexity than a full Transformer, indicating a favourable trade-off between predictive performance and efficiency. Per-class analysis shows that the model maintains strong performance for the majority N class while also achieving competitive F1-scores on the minority S, V, F, and Q classes. The dedicated comparison on rare classes (F and Q) as shown in Fig. 3. suggests that Wavelet-Linformer does not sacrifice performance on under-represented arrhythmia types to optimise overall accuracy. Instead, the model provides a balanced improvement, which is relevant in clinical settings where rare but serious arrhythmias must not be systematically overlooked.

The use of the Wavelet Scattering Transform in Wavelet-Linformer is our central design choice as a fixed front-end for time-frequency feature extraction. Unlike end-to-end models that learn all filters from scratch, WST provides a mathematically grounded, multiscale representation of the ECG signal, with built-in stability to time shifts and deformations. This reduces the burden on the attention module to discover low-level features and may explain why a relatively compact Linformer encoder can match or slightly outperform a larger pure Transformer.

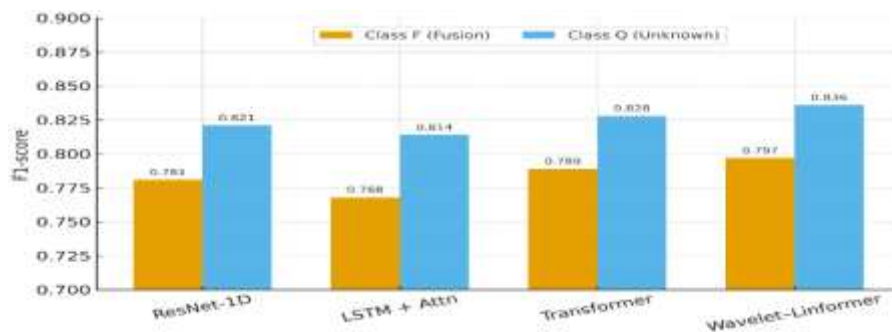


Fig. 3. F1-scores for rare classes *F* (fusion) and *Q* (unknown) across all models.

The empirical results indicate that WST features are sufficiently expressive to support accurate classification while also being robust to noise and variability in acquisition conditions. Because WST uses predefined filters, it can act as a form of structured inductive bias, encoding prior knowledge about the oscillatory and multiscale nature of ECG signals. This proves especially beneficial in scenarios with a scarcity of labelled data as the model has lesser degrees of freedom to overfit false patterns present in the training data. Adopting WST as a front-end also allows for re-use in other downstream model, and therefore a pipeline which permits modularity through feature extraction. The classification can be updated independently.

Furthermore, a fundamental component of the suggested architecture is a replacement of the full self-attention by the linearized attention of Linformer. The use of projecting key matrices and value matrices into lower dimension reduces the attention complexity from quadratic to linear in sequence length and thus consumes lower memory and fast inference. The efficiency comparison reveals that Wavelet-Linformer can classify ECG beats with fewer parameters and reduced inference time than a regular Transformer, yet yields similar or better classification performances. Using attention over WST features is not only efficient but also interpretable. The comparison to ResNet-1D, LSTM + attention, and a pure Transformer highlights the relative strengths of the different architectural choices. Due to the fact that this ResNet-1D baseline works directly on raw ECG segments, it is able to achieve a good performance on the majority N class. However, it still requires deeper stacks of convolutional layers in order to approximate long-range dependencies. The LSTM, like the recurrent model, can capture rhythm information through recurrence and uses attention to focus on important time steps. However, sequential models can limit parallelism and lead to a slow inference latency.

6 LIMITATIONS AND FUTURE WORK

All our experiments are conducted using MIT-BIH, a widely-used arrhythmia database, which may not be representative of the diversity seen in ECG recordings collected in real life across institutions, devices, and patients. To evaluate generalizability, supplementary evidence on datasets, namely INCART, PTB-XL or real-world Holter recordings would be needed. Our current work deals with a single lead beat-level classification, using multi-lead input and longer temporal contexts (e.g. rhythm segments spanning multiple beats) may provide additional information that could improve performance, particularly for complex arrhythmias. While WST offers robust and interpretable features, its configuration (e.g., number of scales, quality factors, and orders) was chosen empirically and may not be optimal for all types of ECG signals. Automated or data-driven methods for selecting scattering parameters, or combining WST with lightweight learned front-ends, could be explored. For future work, there can be qualitative attention visualizations, a more systematic evaluation of interpretability such as through clinician studies, relevance metrics, or counterfactual analyses. That would be valuable to better understand how the model aligns with expert reasoning. We can extend Wavelet-Linformer along several directions such as multi-lead and multi-label arrhythmia detection, joint modelling of beat-level morphology and rhythm-level context and the integration with uncertainty estimation or calibration techniques to better quantify prediction confidence.

7 CONCLUSION

We have proposed Wavelet-Linformer, an efficient attention-based architecture for beat-level arrhythmia classification that combines Wavelet Scattering Transform features with

a linearized self-attention encoder. By delegating low-level, multiscale feature extraction to the Wavelet Scattering Transform and applying Linformer to capture global dependencies over the resulting time-frequency representations, the model aims to balance representational richness with computational efficiency. The proposed Wavelet-Linformer achieved consistent improvements in accuracy, macro F1, and macro AUROC compared with three strong deep-learning baselines of ResNet-1D, LSTM with attention, and a pure Transformer, while using fewer parameters and reduced attention complexity. Beyond overall performance, the proposed model maintained competitive F1-scores on minority classes, including the fusion (F) and the unknown (Q) beats, indicating that the gains are not limited to the dominant normal class. Visualizations that focus on qualitative attention over Wavelet Scattering features also suggest that there are time-frequency regions of physiologically meaningful character over which the network focuses, e.g. the QRS complex and characteristic ectopic patterns. This interpretability can make the network more clinically trustworthy. The authors of these findings take it as an important step toward deployable, resource-aware, and more interpretable ECG analysis systems. The future work will augment this conduct in a number of directions, such as validation in more ECG databases, including multi-lead and longer-term rhythm information, and systematic evaluation of interpretability with feedback from clinicians. The investigation of uncertainty estimation and learned scattering or adaptive scattering arrangement would make it more robust and clinically useful.

DECLARATIONS

COMPETING INTERESTS

There stands no conflict of interest related to this research work.

FUNDING

Authors have not received any funding.

REFERENCES

1. World Health Organization. Cardiovascular diseases (CVDs) [Internet]. Geneva: World Health Organization; 2025 Jul 31 [cited 2025 Dec 7]. Available from: [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
2. Srinivasan NT, Schilling RJ. Sudden cardiac death and arrhythmias. *Arrhythm Electrophysiol Rev.* 2018;7(2):111–7.
3. Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. *IEEE Eng Med Biol Mag.* 2001;20(3):45–50.
4. Association for the Advancement of Medical Instrumentation. ANSI/AAMI EC57:2012/(R)2020: Testing and reporting performance results of cardiac rhythm and ST-segment measurement algorithms. Arlington (VA): AAMI; 2020.
5. Ansari Y, Mourad O, Qaraqe K, Serpedin E. Deep learning for ECG arrhythmia detection and classification: an overview of progress for period 2017–2023. *Front Physiol.* 2023;14:1246746.
6. Boulif A, Ouladsine M, Boulif I, Bensid A. A literature review: ECG-based models for arrhythmia diagnosis using artificial intelligence techniques. *Bioinform Biol Insights.* 2023;17:11779322221149600.
7. Xiao Q, Zhang Q, Wang H, Yang C, Wang S. Deep learning-based ECG arrhythmia classification: a survey of existing methods. *Appl Sci.* 2023;13(8):4964.
8. Liu Z, Yao G, Zhang Q, Zhang J, Zeng X. Wavelet scattering transform for ECG beat classification. *Comput Math Methods Med.* 2020;2020:3215681.
9. Mallat S. Group invariant scattering. *Commun Pure Appl Math.* 2012;65(10):1331–98.

10. Bruna J, Mallat S. Invariant scattering convolution networks. *IEEE Trans Pattern Anal Mach Intell.* 2013;35(8):1872–86.
11. Dhyani A, Dwivedi RK, Basu A, Kumar A. Review of ECG arrhythmia detection techniques based on wavelet transform and time–frequency analysis. *Chaos Solitons Fractals X.* 2023;8:100115.
12. Mohonta SK, Guha R, Dhar PK. Electrocardiogram-based arrhythmia classification using phase-related time–frequency representations with 2D convolutional neural network. *Biomed Signal Process Control.* 2022;71:103107.
13. Acharya UR, Oh SL, Hagiwara Y, Tan JH, Adam M, Gertych A, et al. A deep convolutional neural network model to classify heartbeats. *Comput Biol Med.* 2017;89:389–96.
14. Wu M, Lu Y, Yang W, Wong SY. A study on arrhythmia via ECG signal classification using the convolutional neural network. *Front Comput Neurosci.* 2021;14:564015.
15. Yildirim Ö. A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. *Comput Biol Med.* 2018;96:189–202.
16. Saadatnejad S, Oveisi M, Hashemi M. LSTM-based ECG classification for continuous monitoring on personal wearable devices. *IEEE J Biomed Health Inform.* 2019;24(2):515–23.
17. Oh SL, Ng EYK, Tan RS, Acharya UR. Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heartbeats. *Comput Biol Med.* 2018;102:278–87.
18. Alamatsaz MH, Pirzadeh H, Asl BM. A lightweight hybrid CNN–LSTM explainable model for automatic classification of ECG arrhythmias. *Biomed Signal Process Control.* 2024;88:105555.
19. Tay Y, Dehghani M, Bahri D, Metzler D. Efficient Transformers: a survey. *ACM Comput Surv.* 2022;55(6):109:1–28.
20. Wang S, Li BZ, Khabsa M, Fang H, Ma H. Linformer: self-attention with linear complexity. *arXiv Preprint.* 2020;arXiv:2006.04768.
21. Arabi Z, Pourbahrami S, Abedi A. Hybrid CNN-transformer architecture with adaptive positional embedding for interpretable ECG arrhythmia classification. *SSRN Preprint.* 2025. doi:10.2139/ssrn.5245392.