IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 November 2025, accepted 21 December 2025, date of publication 12 January 2026, date of current version 15 January 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3651029

RESEARCH ARTICLE

Generalized ECG Heartbeat Classification Using Time-Series Transformers

TIMO DE WAELE, JARON FONTAINE, ELI DE POORTER, AND ADNAN SHAHID, (Senior Member, IEEE)
IDLab, Department of Information Technology, Ghent University, 9052 Ghent, Belgium

Corresponding author: Timo De Waele (timo.dewaele@ugent.be)

This work was supported in part by the Fund for Scientific Research Flanders, Belgium, FWO-Vlaanderen, FWO-SB, under Grant 1S33924N; and in part by the Erasmus+ Program of European Commission through the Capacity Building for Digital Health Monitoring and Care Systems in Asia (DigiHealth-Asia) under Agreement 619193-EPP-1-2020-BE-EPPKA2-CBHE-JP.

ABSTRACT This research investigates Time-Series Transformer architectures for Electrocardiogram (ECG) heartbeat classification, particularly focusing on their generalization capabilities towards new patients and varying signal sampling rates, a critical challenge in real-world clinical applications. This study conducts a systematic comparison between Transformers and Convolutional Neural Network (CNN) models using the St. Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database (INCART). Key aspects explored include the impact of different input modalities (raw ECG, Continuous Wavelet Transform (CWT) scalograms, and their combination), various Positional Encoding (PE) schemes for Transformers, and the effect of integrating expert-derived RR interval features through different feature fusion techniques. Transformers, especially with concatenation-based PE schemes and CWT or combined ECG+CWT inputs, consistently outperformed CNNs in classification accuracy and generalization when expert features were not used. They demonstrated more than 30% better generalization to unseen patients, and 20% better generalization to unseen patients and sampling rates. Ultimately, this study emphasizes that robust ECG classifiers depend heavily on deliberate architectural choices, positional encoding schemes, input representations, and the integration of expert features to handle inter-patient and sampling rate variations. Importantly, it also demonstrates that Time-Series Transformers can achieve strong results even with relatively modest model sizes and datasets.

INDEX TERMS Deep learning, transformer, time-series, sampling rate, ECG, classification.

I. INTRODUCTION

Cardiovascular Diseasess (CVDs) represent a spectrum of disorders affecting the heart and blood vessels, including coronary heart disease, cerebrovascular disease, rheumatic heart disease, and other conditions. They stand as the foremost global health challenge and the leading cause of mortality worldwide. In 2019 alone, an estimated 17.9 million lives were lost to CVDs, accounting for 32% of all global deaths, with a disproportionate impact on low- and middle-income countries where preventative measures and access to advanced healthcare are often limited [1], [2]. Beyond mortality, CVDs impose a significant burden of disability and economic cost, underscoring the urgent need for effective diagnostic and management strategies [3].

Central to the diagnosis and monitoring of a wide array of cardiovascular conditions is the Electrocardiogram (ECG). This non-invasive and relatively inexpensive tool provides a graphical representation of the heart's electrical activity, offering crucial insights for detecting arrhythmias, heart disease, and electrolyte imbalances, among other pathologies [4], [5]. However, the manual interpretation of ECG recordings is a time-consuming process that requires significant expertise, and its large volume in clinical practice can lead to a substantial workload for cardiologists and trained medical personnel. This reliance on manual analysis can introduce challenges such as delays in diagnosis, inter-observer variability in interpretation, and limited accessibility

The associate editor coordinating the review of this manuscript and approving it for publication was Sajid Ali.

to expert review, particularly in resource-constrained environments or remote geographical areas [6].

Automated analysis of the ECG signal is therefore crucial for advancing cardiovascular healthcare, providing scalable, consistent, and accessible methods for arrhythmia detection and broader cardiac assessment. By leveraging machine learning algorithms automated ECG classification systems can assist clinicians by providing rapid initial screenings, highlighting potential abnormalities, and improving diagnostic efficiency. This not only has the potential to alleviate the burden on healthcare professionals but also to facilitate earlier detection of critical heart conditions, ultimately enabling more timely interventions and improving patient outcomes, especially in underserved populations. While deep learning has shown considerable promise in this domain, a critical challenge is the development of models that exhibit strong generalization capabilities. This is especially challenging due to the inherent variability in real-world ECG datasets, which includes significant inter-patient differences and, crucially, inconsistencies in sampling rates across various clinical datasets and ECG capturing devices [7], [8]. This paper explores the potential of Time-Series Transformer architectures to address these challenges, focusing on their ability to generalize effectively to unseen patients and adapt to new, previously unseen sampling rates.

To achieve such robust generalization, this paper performs a comprehensive investigation into Time-Series Transformers, comparing their performance systematically against the Convolutional Neural Networks (CNN) approach commonly employed in ECG classification research. We rigorously evaluate how these two architecture families handle data heterogeneity, particularly focusing on generalization towards unseen patients and robustness to variations in sampling rates. This comparative analysis extends to an exploration of different input types (raw ECG, Continuous Wavelet Transform (CWT) scalograms, and their combination), various Positional Encodings (PEs) schemes for the Transformer architecture, and the impact of integrating expert clinical knowledge, specifically RR interval features, through different feature fusion techniques. A significant aspect of this work is to demonstrate that while Transformers are often associated with training on enormous datasets and requiring billions of parameters, they can effectively challenge established architectures like CNNs even with considerably smaller models (a couple of million parameters) and more modestly sized datasets (less than 100,000 training samples). The insights gained are crucial for developing deployable diagnostic tools and may guide the architecture of future large-scale, pre-trained foundation models for ECG [9], [10].

The main contributions of this paper are:

- A systematic comparison of Time-Series Transformers and CNN architectures for ECG heartbeat classification, evaluating their performance and generalization capabilities across both unseen patients and unseen sampling rates on two gold-standard databases (INCART and MIT-BIH), including raw ECG, CWT, and combined inputs.
- An empirical evaluation of four distinct Positional Encoding schemes for Transformer models (standard additive sinusoidal, learned, timestamp concatenation, and linspace concatenation).
- A comparative analysis of incorporating expert RR interval features, detailing the effects of early versus late fusion within Transformers compared to multi-headed integration in CNNs.
- A demonstration that Time-Series Transformers can achieve competitive, and often superior (e.g., more than 30% more accurate than CNNs without expert features), performance in ECG classification with relatively modest model sizes and datasets, challenging the notion that they are only suited for scenarios with massive datasets and computational resources.
- The public release of the source code and trained models developed in this paper (https://github.com/timodw/heartbeat_classification_using_transformers) to ensure reproducibility and facilitate future research building upon these findings.

The remainder of this paper is structured as follows. Section II reviews relevant literature concerning CNNs and Transformers for ECG analysis, addressing sampling rate invariance, positional encoding, and feature fusion. Section III details the INCART dataset, preprocessing steps, implemented model architectures, experimental configurations, and used evaluation metrics. Section IV presents and interprets the findings from our comparative experiments, particularly focusing on the impact of RR interval features and generalization capabilities. Building on this, Section V summarizes the key insights derived from the study. Finally, Section VI concludes the paper and suggests directions for future research.

II. RELATED WORK

The pursuit of robust and generalizable deep learning models for automated ECG analysis is critical for advancing cardiovascular healthcare. While CNNs have become a cornerstone in this domain, demonstrating success in arrhythmia detection and heartbeat classification [11], achieving consistent generalization towards unseen patients and across datasets with varying characteristics, such as different sampling rates, remains a significant hurdle.

CNNs excel at learning hierarchical features from raw 1D ECG segments [12] or 2D time-frequency representations like CWT scalograms [13]. The utility of such transformed inputs is further evidenced by studies improving deep CNN performance using wavelet transforms [14] or by classifying ECG scalograms with architectures that combine CNN feature extractors with classifiers like SVMs [15]. Common architectures range from custom designs to established networks like ResNet [16], and are often combined

with Recurrent Neural Networks (RNNs) like Bidirectional Long Short-Term Memory Network (Bi-LSTM) to better capture temporal dependencies [17]. While effective for local morphological patterns, the fixed receptive field of standard convolutions limits the modeling of long-range dependencies crucial for rhythm analysis. Similarly, RNNs can struggle with gradient stability and capturing truly global context in long ECG recordings. These limitations have motivated the adoption of Transformer-based architectures.

In contrast, Transformers leverage self-attention mechanisms to capture global dependencies across entire sequences, a characteristic increasingly explored for physiological time series, including Electroencephalogram (EEG) and ECG [18], [19], [20]. A common and effective strategy when applying Transformers to ECG data involves utilizing CNNs as initial feature extractors or to generate patch embeddings, which are then fed into the Transformer blocks. This hybrid CNN-Transformer architecture is exemplified in models such as Cat-Net [21], and has been extended to use multi-scale convolutional inputs [22], combine a Transformer's output with scalar clinical features [23], or create complex parallel models. These more advanced designs include dual-branch systems that fuse features from a CNN and a Vision Transformer (ViT) via cross-attention [24] or process time-domain and frequency-domain representations in separate network branches before fusion [25]. In addition to fusion-based strategies, other work has focused on modifying the core Transformer architecture itself, for example by proposing constrained Transformer networks better tailored for ECG signal characteristics [26]. While these sophisticated architectures show promise, their evaluations do not always employ a patient-independent data split and typically rely on a fixed, standardized sampling rate, leaving key generalization challenges unaddressed.

Despite this progress, several aspects critical for building truly deployable models remain underexplored. First, robustness to sampling rate variability is a major practical hurdle, as clinical data is often captured at different frequencies (e.g., 257 Hz, 360 Hz, 500 Hz) [7], [8]. While some methods exist for rate invariance [27], [28], a systematic comparison of how modern architectures like CNNs and Transformers generalize to unseen sampling rates is largely absent. Second, while an effective PEs scheme is essential for Transformers [29], the comparative efficacy of different PEs approaches—beyond standard sinusoidal or learned embeddings [30], [31], [32] to include concatenation-based methods [33], [34] on their generalization capability for ECG analysis has not been thoroughly investigated. Third, while incorporating expert domain knowledge via RR-interval fusion is a known strategy to boost performance [13], [35], [36], and early work has explored fusing Transformers with temporal features [37], its specific influence on the sampling rate robustness of different architectures is not well understood. This paper is designed to address these three specific gaps through a series of systematic experiments comparing architectures, PEs schemes, and feature fusion strategies, with a primary focus on generalization to unseen patients and new, interpolated sampling rates.

To contextualize our contributions and highlight the comprehensive nature of our evaluation, we systematically review how prior works address several critical aspects of model assessment. Table 1 provides a comparative summary, indicating whether selected studies explicitly evaluate their models for:

- Generalization towards unseen patients
- Generalization towards unseen sampling rates
- Evaluation of different Positional Encoding schemes
- Direct comparison against CNNs
- Use of feature fusion
- Classify individual heartbeats

Ultimately, addressing these multifaceted aspects is crucial for advancing towards truly robust and generalizable ECG analysis tools. The insights from existing literature, coupled with the identified gaps that our work aims to fill, motivate the experimental design of this paper. By systematically evaluating Transformers against CNNs across these dimensions, particularly focusing on generalization to unseen patients and sampling rates, and demonstrating their efficacy even with compact model sizes and datasets, this research aims to provide critical insights for developing reliable clinical decision support systems.

III. METHODOLOGY

This section details the used dataset, preprocessing pipeline, evaluated input representations, model architectures, and experimental configurations employed in this study.

A. DATASET

The dataset used in this research is the St. Petersburg INCART 12-lead Arrhythmia Database, available on PhysioNet.[1] This database contains 75 annotated recordings from 32 patients, each sampled at a frequency (f_s) of 257 Hz. For this study, we exclusively utilized the Lead II data from each recording, as this lead was consistently present across all records in the database, ensuring a uniform input signal for our analysis. For our heartbeat classification task, we focused on three primary classes present in the INCART dataset: Normal (N), Supraventricular Ectopic Beat (SVEB), and Ventricular Ectopic Beat (VEB). Following the AAMI EC57 standard, heartbeats can be classified into five superclasses: Normal, SVEB, VEB, Fusion (F), and Unknown (Q). However, beats annotated as Fusion or Unknown were excluded from our analysis due to their extremely limited sample size in the INCART dataset. The rarity of F and Q samples would make reliable model training difficult and, more critically, prevent meaningful evaluation of generalization performance for these categories, particularly given our strict patient-independent evaluation protocol where these rare beats only appear in a handful of patients.

[1] https://physionet.org/content/incartdb/1.0.0/

TABLE 1. Comparison of evaluation strategies in the ECG classification literature presented in this section. (✓ indicates the aspect was explicitly evaluated; – indicates not explicitly evaluated or not applicable).

Study	Unseen Patients	Unseen Sampling Rates	Different PEs	Comparison with CNN	Feature Fusion	Individual Heartbeats	Used Input Modalities
Abid & Cheikhrouhou [38]	–	–	–	✓	–	–	12-Lead ECG
Alexakis *et al.* [28]	–	–	–	–	–	–	Time-interval features
Che *et al.* [26]	✓	–	–	✓	–	–	12-Lead ECG
Cheng *et al.* [17]	✓	–	–	✓	–	–	1-Lead ECG
De Waele *et al.* [39]	✓	–	–	–	✓	✓	CWT, RR features, scalar features
Garcia *et al.* [36]	✓	–	–	–	–	✓	Temporal Vectorcardiogram (TVCG)
Hannun *et al.* [12]	✓	–	–	–	–	–	1-Lead ECG
Hu *et al.* [20]	–	–	–	✓	–	✓	1-Lead ECG
Islam *et al.* [21]	–	–	–	✓	–	✓	1-Lead ECG
Liu *et al.* [25]	✓	–	–	✓	✓	–	12-Lead ECG, Wavelet Coefficients
Natarajan *et al.* [23]	✓	–	–	–	✓	–	1-Lead ECG, scalar features
Ozaltin *et al.* [15]	–	–	–	–	–	✓	CWT, Short-Time Fourier Transform (STFT)
Wang *et al.* [13]	✓	–	–	–	✓	✓	CWT, RR features
Wei & Li [24]	–	–	–	✓	✓	–	12-Lead ECG
Yan *et al.* [37]	✓	–	–	✓	✓	✓	1-Lead ECG, RR features
Ye *et al.* [35]	✓	–	–	–	✓	✓	Wavelet coefficients, Independent Component Analysis (ICA), RR features
Zhang *et al.* [22]	–	–	✓	✓	–	–	1-Lead ECG
Zhao *et al.* [14]	✓	–	–	–	–	–	1-Lead ECG
This Work	✓	✓	✓	✓	✓	✓	1-Lead ECG, CWT, RR features

To assess the models' ability to generalize to new individuals, a strict patient-independent split was implemented. We specifically opted for a 50/50 training-validation split to create a challenging evaluation scenario with a large number of unseen patients in the validation set. This split was achieved using the `StratifiedGroupKFold` function from the scikit-learn library [40] with two splits, which ensures that no patient appears in both sets while maintaining a comparable class distribution across them. The resulting split assigned patients to the training and evaluation sets as follows:

- **Training Set Patients:** 1, 2, 4, 5, 7, 8, 10, 12, 18, 19, 20, 23, 25, 27, 28, 30, 32 (17 patients)
- **Evaluation Set Patients:** 3, 6, 9, 14, 22, 26, 11, 13, 15, 16, 17, 21, 24, 29, 31 (15 patients)

This patient-independent split resulted in the class distributions detailed in Table 2.

TABLE 2. Class distribution of heartbeat samples in the training and evaluation sets after patient-independent splitting.

Set	Class	Number of Samples
Training	N	62,354
	SVEB	605
	VEB	11,088
Evaluation	N	62,620
	SVEB	1,051
	VEB	8,462

To mitigate the effects of the inherent class imbalance shown in Table 2, a balanced batch sampling strategy was employed during training (described in Section III-H), and performance was assessed using metrics robust to imbalance (Section III-I).

B. PREPROCESSING AND SEGMENTATION

Prior to heartbeat segmentation, the raw continuous ECG signals underwent several preprocessing steps. First, baseline wander was removed using two consecutive passes of a median filter. The first pass used a window size equivalent to $0.2 \times f_s$ samples, and the second pass used a window size of $0.6 \times f_s$ samples, where f_s is the sampling rate of the signal (original or resampled). Second, after baseline correction, the signals were normalized relative to the average R-peak amplitude by dividing the signal by the mean amplitude calculated across all annotated R-peaks within that specific recording.

Individual heartbeats were subsequently segmented using the R-peak annotations provided within the INCART dataset. Each segment was defined from the R-peak of the preceding beat (R_{i-1}) to the R-peak of the subsequent beat (R_{i+1}). To ensure segments could accommodate resampling up to 500 Hz within a fixed input size, any heartbeat segment where the duration between R_{i-1} and R_{i+1} exceeded 2 seconds was discarded. All remaining segments were then padded with a value of -1 to achieve a uniform length of 1000 samples. This padding value was chosen to be distinct from the expected normalized signal range (0 to 1). An example of a resulting segmented and padded heartbeat window is illustrated in Figure 1.

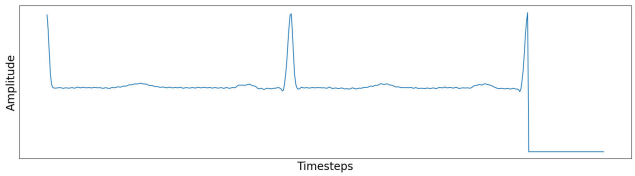

FIGURE 1. An example of a segmented and padded heartbeat window.

C. INPUT MODALITIES

To evaluate the impact of the used input modality, we utilized three different representations for the segmented heartbeats:

1) **Raw ECG:** The preprocessed, segmented, and padded 1D time series signal, resulting in an input vector of shape (1000, 1).
2) **Continuous Wavelet Transform (CWT):** The time-frequency representation was obtained using

the CWT with the Mexican Hat ('mexh') wavelet function. The CWT is a type of time-frequency domain transformation that uses the same base idea as the STFT but makes it more adjustable through the inclusion of a selectable wavelet function and the use of scale and translation parameters. Given a signal $x(t)$, the CWT is defined as shown in Equation (1):

$$C_a(b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \cdot \phi\left(\frac{t-b}{a}\right) dt, \qquad (1)$$

where a and b represent scale and translation parameters, respectively and $\phi(t)$ is the wavelet function used. The wavelet function $\phi(t)$ that we used is the Mexican Hat wavelet and is represented in Equation (2):

$$\phi(t) = \frac{2}{\sqrt{3}\sqrt[4]{\pi}} \exp\left(-\frac{t^2}{2}\right)\left(1-t^2\right). \qquad (2)$$

The scale parameter (a) can also easily be transformed to its corresponding frequency by Equation (3):

$$f = \frac{f_c \cdot f_s}{a}, \qquad (3)$$

where f_c indicates the center frequency of the used wavelet and f_s is the sampling frequency of the original signal. By iterating over the different time steps while varying the scale parameter, and thus the frequency, one can use this transformation to obtain a 2D scalogram in the time-frequency domain of the original signal. This paper uses the same wavelet as used in previous work, as it closely matches the shape of ECG waveform and is already widely used in ECG signal analysis. For this paper's specific implementation, the CWT was computed using `pywt.cwt` [41] to produce a (1000, 100) time-frequency grid for each heartbeat segment. An example visualization is provided in Figure 2.

3) **RR Features:** We employed four RR interval features: previous RR, next RR, local RR (mean of the last 10 preceding RR intervals), and the ratio of the next to the previous RR. A key difference in this work is that all interval durations were calculated in milliseconds (ms) based on the annotated R-peak timings, rather than in number of samples. This conversion ensures the features are inherently independent of the signal's sampling rate, which is crucial for the invariance analysis performed in this paper. Figure 3 visualizes how each RR feature relates to the ECG signal.

Following segmentation and feature/representation generation, specific normalization procedures were applied to each input modality:

- **Raw ECG:** Each segmented and padded heartbeat window underwent an additional min-max scaling step applied only to the non-padded values. The minimum and maximum values were determined from the non-padded portion of that specific window, and these values were used to scale this portion to the range [0, 1]. The padded values (-1) remained unchanged by this scaling.
- **CWT:** The computed CWT magnitudes for each heartbeat window were normalized independently, following the same min-max scaling strategy applied to the Raw ECG. Specifically, the minimum and maximum magnitude values within each CWT window were used to scale the values of that window to the range [0, 1], ensuring that this scaling was applied only to the relevant CWT coefficients, analogous to the non-padded values in the ECG data.
- **RR Features:** Similarly, the four RR interval features (now in ms) were normalized using per-patient min-max scaling to the range [0, 1].

This strategy applies a local, per-beat normalization (min-max scaling) independently to both the primary ECG waveform and its CWT representation, while the auxiliary RR features are normalized on a broader, per-patient basis.

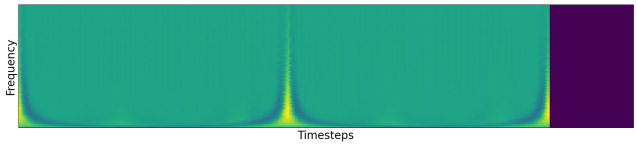

FIGURE 2. An example of the CWT representation of a segmented heartbeat.

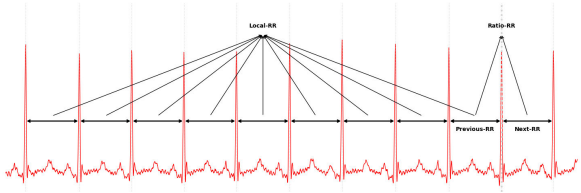

FIGURE 3. A visualization of how the RR features relate to the ECG signal.

D. DATASET RESAMPLING

A central goal of this paper is to evaluate the sampling rate invariance capabilities of CNNs and Transformers, specifically focusing on their ability to generalize to unseen frequencies interpolated between those encountered during training. To facilitate this analysis, the preprocessed and segmented heartbeat data were resampled to various target frequencies using the Fourier method implemented in `scipy.signal.resample` [42]. Consistent with the preprocessing pipeline, the CWT representation was computed after resampling the raw ECG signal to each target frequency.

The experimental setup was designed to assess interpolation performance. Models were trained using data resampled to a set of nine distinct frequencies:
(100, 150, 200, 250, 300, 350, 400, 450, 500) Hz. Following training, the models' ability to handle intermediate sampling rates was evaluated using data resampled to a different set of eight interleaved, previously unseen frequencies: (125, 175, 225, 275, 325, 375, 425, 475) Hz. Crucially, the patient-independent data splitting described in Section III-A

was strictly maintained across all resampled datasets. This ensures that the evaluation reflects generalization not only to unseen intermediate sampling frequencies but also to unseen patients.

E. MODEL ARCHITECTURES

Two main deep learning architectures were implemented and compared: a Convolutional Neural Network (CNN) based on prior work [13], [39], and a Transformer-based encoder model. Both architectures were designed to handle different combinations of input modalities (Raw ECG, CWT, and RR Features).

1) CONVOLUTIONAL NEURAL NETWORK (CNN)

The CNN architecture, visualized in Figure 4, utilizes separate processing heads for each input modality, a method that facilitates the late fusion of scalar features with features extracted from time-series data. This architectural choice draws inspiration from the multi-head design used in [39] and builds upon the CWT-based architecture from [13].

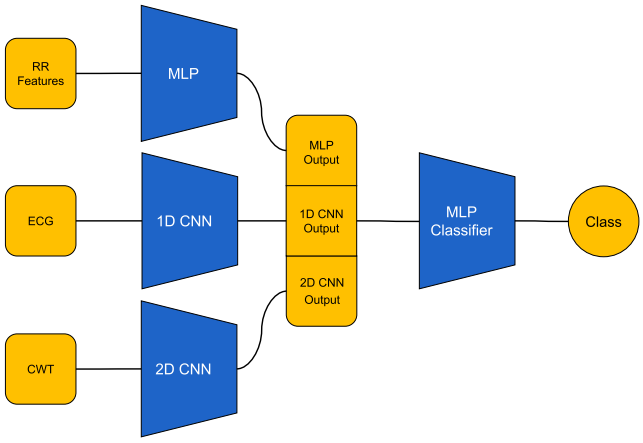

FIGURE 4. A visualization of the implemented CNN architectures showing its 3 different heads.

- **ECG Head (1D CNN):** Processes the raw ECG signal $(B, 1, 1000)$. It consists of three sequential blocks, each containing a `Conv1d` layer (output channels: 16, 32, 64; kernel sizes: 7, 3, 3 respectively), `BatchNorm1d`, `LeakyReLU` activation (negative slope = 0.2), and `MaxPool1d` (kernel sizes: 5, 3). The final block uses `AdaptiveMaxPool1d(1)` instead of standard max pooling, followed by a `Flatten` layer to produce a 64-dimensional feature vector.
- **CWT Head (2D CNN):** Processes the CWT representation $(B, 1, 1000, 100)$. Its structure is the same as the ECG Head but it employs 2D operations instead of 1D (`Conv2d`, `BatchNorm2d`, `MaxPool2d`, `AdaptiveMaxPool2d((1, 1))`). It also outputs a 64-dimensional feature vector after the use of a `Flatten` layer.
- **RR Feature Head (MLP):** Processes the 4 RR interval features $(B, 4)$. It comprises three linear layers with `LeakyReLU` activations (negative slope = 0.2): `Linear(4 → 128)`, `Linear(128 → 128)`, and `Linear(128 → 64)`, outputting a 64-dimensional vector.

Depending on the input modalities used in a specific experiment (e.g., Raw ECG only, CWT+RR, Raw ECG+CWT+RR), the outputs from the corresponding active heads are concatenated. This concatenated feature vector (size $64 \times N_{\text{modalities}}$, where $N_{\text{modalities}}$ is the number of active input types) is fed into a final classifier Multilayer Perceptron (MLP) consisting of four linear layers (`Linear(input → 256)`, `Linear(256 → 64)`, `Linear(64 → 32)`, `Linear(32 → 3)`) with intermediate `LeakyReLU` activations (negative slope = 0.2). All convolutional and linear layer weights were initialized using Kaiming Normal initialization [43], and biases were initialized to zero. The approximate number of trainable parameters for the full CNN model (all heads active) is 125 000.

2) TRANSFORMER

An encoder-only Transformer architecture, inspired by the BERT family [30] and visualized in Figure 5, was implemented.

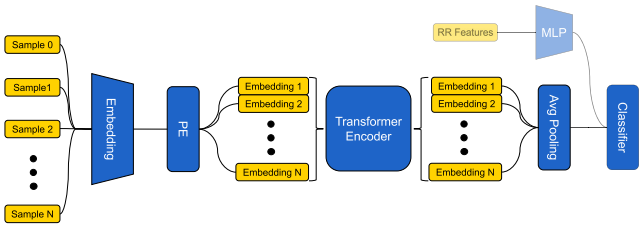

FIGURE 5. A visualization of the implemented Transformer architecture.

- **Input Embedding:** Input samples (concatenated ECG, CWT, and potentially RR features for early fusion) are projected into an embedding space of dimension $d_{embedding} = 256$ using a single linear layer (`Linear(input → 256)`). The number of input features to this layer depends on the selected input modalities and PEs scheme.
- **Positional Encoding (PE):** Various PEs schemes (detailed in Section III-F) are applied to inject positional information into the embeddings.
- **Transformer Encoder:** The core of the model consists of $N_{layers} = 2$ Transformer encoder layers. Each layer utilizes multi-head self-attention ($N_{heads} = d_{embedding}/64 = 4$) and a position-wise Feedforward Neural Network (FNN) (FNN dimension $d_{FNN} = d_{embedding} \times 4 = 1024$). We employed pre-layer normalization (`norm_first=True`) and a dropout rate of 0.3 within the encoder layers [29]. The standard `ReLU` activation was used within the FNN.
- **Output Pooling:** After the final encoder layer and layer normalization, the output sequence $(B, 1000, d_{embedding})$ is processed by swapping dimensions, applying average pooling (`AvgPool1d`) across the time dimension, and

flattening to produce a single vector of size $d_{embedding} = 256$ per heartbeat.

- **RR Feature Head (Late Fusion):** For the late fusion configuration, a separate MLP processes the 4 RR features (`Linear(4 → 128) → ReLU → Linear(128 → 128) → ReLU → Linear(128 → 64) → ReLU`), yielding a 64-dimensional vector.
- **Classifier Head:** The pooled output from the Transformer encoder (potentially concatenated with the RR feature vector from the late fusion head) is passed through a final classifier MLP (`Linear(input → 512) → ReLU → Linear(512 → 3)`).

The approximate number of trainable parameters for the Transformer model (configuration dependent on fusion/PE) is 1 800 000.

F. POSITIONAL ENCODING SCHEMES

Since the self-attention mechanism in Transformers is permutation-invariant, explicit positional information must be provided. We evaluated four distinct PEs schemes for the Transformer model:

1) **Standard Addition:** The original fixed sinusoidal PEs proposed by Vaswani et al. [29]. These position-dependent vectors of dimension $d_{embedding}$ are added element-wise to the input embeddings. Visualization: Figure 6.

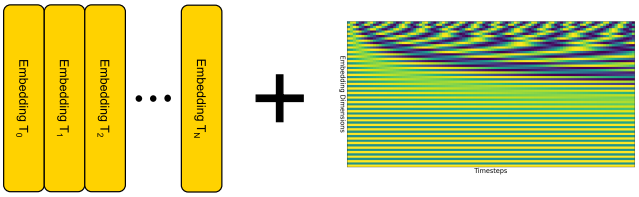

FIGURE 6. A visualization of the standard addition positional encoding scheme.

2) **Zero (Learned):** A learnable embedding layer implemented as `torch.nn.parameter.Parameter` with shape $(1, N_{input}, d_{embedding})$, initialized randomly using `torch.randn`. This learned positional vector is added element-wise to the input embeddings, similar to the PEs used in models like BERT and GPT, as shown in Figure 7.

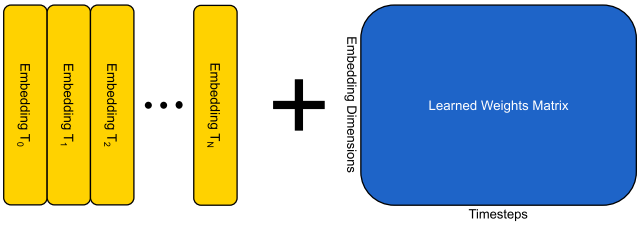

FIGURE 7. A visualization of the zero (learned) positional encoding scheme.

3) **Timestamp Concatenation:** The timestamp corresponding to each of the 1000 input samples is used as positional information. This approach was chosen to explicitly provide the model with information about the temporal distance between samples, and thus indirectly, information about the signal's sampling rate. The goal was to aid the model in learning a more sampling rate invariant representation. This scalar timestamp value is concatenated as an additional feature dimension after the main linear embedding layer. To maintain the target embedding dimension $d_{embedding} = 256$, the linear embedding layer projects the input features to $d_{embedding} - 1 = 255$ dimensions. A visualization of this scheme is provided in Figure 8.

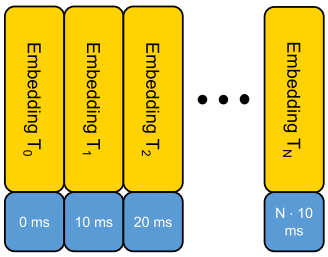

FIGURE 8. A visualization of the timestamp concatenation positional encoding scheme, an example sampling rate of 100 Hz is used here.

4) **Linspace Concatenation:** Similar to timestamp concatenation, but uses a fixed, linearly spaced vector (e.g., `torch.linspace(0, 2, 1000)`) as the positional feature. This scalar value per position is concatenated after the linear embedding layer (which outputs $d_{embedding} - 1$ dimensions). Visualization: Figure 9.

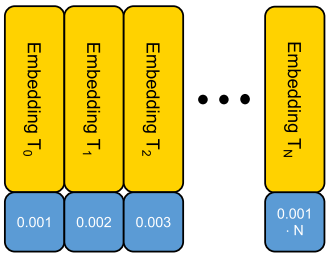

FIGURE 9. A visualization of the linspace concatenation positional encoding scheme.

G. RR INTERVAL FEATURE FUSION

We compared different strategies for incorporating the four RR interval features into the CNN and Transformer models:

- **CNN Method:** A late fusion strategy was employed via a dedicated MLP head for processing the RR features, concatenating its 64-dimensional output with ECG/CWT features before the final classifier (Section III-E). This approach was chosen over early fusion (adding RR features as input channels) as standard CNN filters are less suited to jointly process time-varying signals (ECG/CWT) and time-invariant scalar features (RR intervals) directly in their input channels.

- **Transformer - Early Fusion:** The four RR features are treated as additional input channels alongside the ECG and/or CWT data. For each heartbeat, the 4 scalar RR values are repeated across all 1000 time steps and concatenated to the feature dimension of the input sequence before it is passed to the initial embedding layer.
- **Transformer - Late Fusion:** The RR features are processed independently by a dedicated MLP head to generate a 64-dimensional embedding. This embedding is then concatenated with the 256-dimensional pooled output vector from the main Transformer Encoder before the final classifier MLP.

H. EXPERIMENTAL SETUP

All models were trained and evaluated using the patient-independent splits defined in Section III-A.

- **Sampling Rate Invariant Training Strategy:** For the sampling rate invariance analysis (Section III-D), models were trained on a mixed-frequency dataset. For each training patient, heartbeats from 5 distinct sampling rates were randomly chosen from the training frequency set, effectively quintupling the training samples per patient. The evaluation dataset was constructed analogously using the corresponding evaluation frequency set.
- **Optimization and Regularization:** Models were trained using the Adam optimizer [44] with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 1 \times 10^{-9}$. We employed the Noam learning rate schedule [29] with 800 warmup steps and a base learning rate factor adjusted to yield a peak learning rate around 1×10^{-4}, using the Transformer's embedding dimension ($d_{embedding} = 256$) as the model size parameter in the schedule calculation. L2 regularization (weight decay) with a factor of 1×10^{-2} was applied directly within the optimizer to combat overfitting.
- **Training Procedure:** Due to the large effective dataset size after resampling, which was too large to fit into the available system memory, a chunked training approach was adopted. The training process consisted of 20 chunks. In each chunk, 5000 samples were randomly selected from the full training dataset (either the original dataset or the dataset containing mixed sampling rates), and the model was trained on this chunk for 10 epochs using a batch size of 128. This resulted in 200 effective epochs over approximately 100,000 unique samples. To address the significant class imbalance (Table 2), we employed a balanced batch sampling strategy, ensuring that each minibatch contained an equal number of samples from each of the three classes (N, SVEB, VEB) [45]. The standard Cross-Entropy Loss function was used. For the CNN models, which were found to be more susceptible to overfitting than the Transformer architecture, early stopping was implemented with a patience of 5 chunks (50 effective epochs), monitoring the validation loss on the evaluation set (which was also processed in chunks).
- **Hardware:** All experiments were conducted on cloud computing instances equipped with one NVIDIA V100 GPU, 8 CPU cores, and 64 GB of RAM.[2] Models were implemented using PyTorch [46].

I. EVALUATION STRATEGY

Model performance was primarily evaluated using balanced accuracy. This metric, calculated as the arithmetic mean of the recall (sensitivity) for each class, provides a robust measure of performance across imbalanced classes by giving equal weight to each class regardless of its frequency.

To ensure the reliability and robustness of our findings, each experiment (defined by a specific model architecture, PEs scheme, and input modalities or feature fusion method) was repeated five times using different random seeds for weight initialization and data shuffling. The final results for each experiment are reported as the mean $\pm$ standard deviation across these five runs. The code and trained model weights are made publicly available to ensure reproducibility (See I).

IV. RESULTS AND DISCUSSION

This section presents and discusses the results from our experiments. The primary objectives were twofold: firstly, to evaluate the classification accuracy and generalization capabilities of CNNs versus Transformer architectures, considering various input modalities (raw ECG, CWT, and the combined ECG+CWT) and different Positional Encoding (PEs) schemes for the Transformer; and secondly, to assess the influence of incorporating expert RR interval features on these performance metrics using different fusion strategies. Full detailed results for all performed experiments can be found in the appendix. All experiments were conducted on the INCART dataset, as detailed in Section III-A.

A. PERFORMANCE WITHOUT RR INTERVAL FEATURES

The results for the initial experiment, evaluating model performance on the original dataset without the inclusion of RR interval features, are presented in the heatmap in Figure 10. This figure displays the balanced accuracy for each model configuration across the different input modalities.

Overall, the CNN architecture was generally outperformed by the Transformer model, particularly when the latter was paired with effective Positional Encoding (PEs) schemes. For instance, while the CNN's highest balanced accuracy reached $67.8\% \pm 5.7$ with the CWT input, most Transformer configurations surpassed this value. The choice of PEs scheme was indeed a significant determinant for the Transformer's performance. Concatenation-based PEs (timestamp and linspace) consistently demonstrated strong performance, achieving balanced accuracies exceeding 90% when provided

[2]https://idlab.ugent.be/resources/gpulab

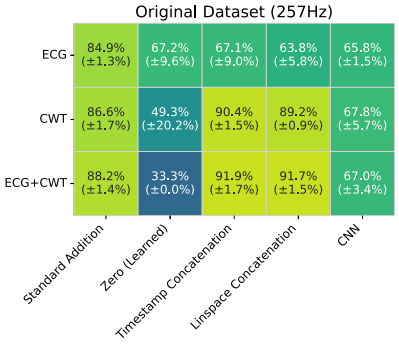

FIGURE 10. **Balanced accuracy results for the evaluation set (containing patients not used during training) on the original INCART dataset without RR features. The first four columns show the Transformer model with different positional encodings.**

with CWT input data. The standard additive sinusoidal PEs also proved competitive across various inputs. This general superiority of the Transformer architecture, when appropriately configured, can likely be attributed to fundamental architectural differences. The Transformer's self-attention mechanism allows for a global receptive field, enabling the capture of long-range dependencies across the entire heartbeat segment. In contrast, the CNN's convolutional operations inherently possess a more localized receptive field, potentially limiting their capacity to model global contextual information crucial for discriminating complex and highly personalized heartbeat patterns. Furthermore, the explicit temporal information introduced by PEs schemes in Transformers, providing a form of awareness of the distance between samples, is absent in the standard CNN architecture, which might hinder its ability to learn features that depend on this inter-sample distance.

In stark contrast to effective PEs schemes, the Zero (learned) PEs scheme markedly underperformed. Its performance with raw ECG input was 67.2%±9.6, barely matching the CNN's performance. However, its efficacy dropped substantially with other inputs, yielding only 49.3% ± 20.2 with CWT data and a significantly lower 33.0% ± 0.0 when using combined ECG and CWT data, making it the only PEs configuration that consistently failed to outperform the CNN. This consistent underperformance of the learned PEs scheme, characterized by low mean accuracies and high variance, may stem from several factors. Its fully learned nature, for instance, could make it prone to overfitting on the training dataset, thus impairing its ability to generalize to the unseen patients in the validation set. Additionally, there appear to be challenges with its optimization for this task. The hypothesis that these optimization issues are due to poor gradient flow, possibly from unfortunate weight initializations, is supported by observations of stagnant training loss curves in several runs.

Regarding input modalities across the more effective PEs schemes, the combination of raw ECG and CWT data typically yielded the highest overall classification performance in this set of experiments without RR features. The peak performance of 91.9% ± 1.7 was achieved by the Transformer with the timestamp concatenation PEs using this combined input. Generally, models utilizing CWT data as the only input modality also outperformed those using only the raw ECG signal. The utility of the CWT was evident, as transforming the raw time-series data into a time-frequency representation generally enhanced performance across most model configurations, particularly for Transformers. This improvement may arise because CWT can expose features in the frequency domain that are more invariant to inter-patient morphological differences than raw signal amplitudes alone. The combination of raw ECG and CWT inputs often yielded further, albeit sometimes marginal, improvements for Transformer models, likely by providing complementary information. For the CNN, however, combining raw ECG with CWT offered no significant advantage over using CWT alone, suggesting a potential saturation in the information it could effectively leverage from these inputs without further context.

A particularly noteworthy observation emerged specifically with the raw ECG signal input. While the CNN achieved 65.8% ± 1.5 and Transformer models with most PEs showed similarly modest results on this input type, the Transformer using the standard additive sinusoidal PEs achieved a significantly higher balanced accuracy of 84.9%± 1.3. This highlights a specific synergy between this PEs scheme and the raw ECG data, suggesting that the structured, periodic nature of sinusoidal PEs might align well with the inherent patterns in the 1D raw ECG signal, allowing it to outperform some configurations that used more complex inputs.

Next, we evaluated the models' generalization capabilities when trained and tested on datasets incorporating a sweep of varying sampling rates, with the results presented in Figure 11. This analysis reveals the models' capabilities to generalize towards sampling rates that were not part of its training dataset.

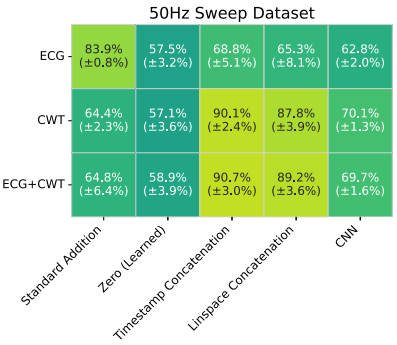

FIGURE 11. **Balanced accuracy results for the evaluation set (containing patients not used during training) on the resampled dataset using a sweep with 50Hz spacing without RR features. The first four columns show the Transformer model with different positional encodings.**

In experiments evaluating generalization to varying sampling rates, Transformer models employing concatenation-based PEs with CWT or combined ECG+CWT inputs

significantly outperformed CNNs. Specifically, the timestamp concatenation PEs achieved balanced accuracies of 90.7%±3.0 (combined input) and 90.1%±2.4 (CWT alone), while the linspace concatenation PEs registered comparable scores of 89.2% ± 3.6 (combined) and 87.8% ± 3.9 (CWT). In contrast, the CNN's generalization performance was more modest, peaking at 70.1% ± 1.3 with CWT input and showing lower scores with ECG (62.8%±2.0) and combined ECG+CWT (69.7% ± 1.6) inputs. Interestingly, for the CNN, this performance on the sampling rate sweep dataset with CWT or combined inputs actually surpassed its results on the original fixed-rate dataset. This counter-intuitive improvement suggests that training the CNN with data resampled to multiple frequencies may serve as an effective form of data augmentation when using these time-frequency representations . Furthermore, this multi-rate training strategy also notably reduced the performance variance across runs for the CNN, pointing to enhanced training stability when exposed to diverse sampling rates. However, the choice of PEs was critical for Transformer performance. Other PEs schemes exhibited distinct generalization profiles: the Standard Addition PEs uniquely excelled with raw ECG input (83.9% ± 0.8) but its effectiveness diminished considerably with CWT (64.4%±2.3) or combined ECG+CWT (64.8%± 6.4) inputs. The learned PEs, consistent with its performance on the fixed-rate dataset, generalized poorly across all input modalities (e.g., 57.5%±3.2 for ECG, 57.1%±3.6 for CWT, and 58.9%±3.9 for ECG+CWT), indicating its unsuitability, in its current form, for this architecture in ECG classification.

The superior performance of Transformers when paired with concatenation-based PEs in handling sampling rate variations can be primarily attributed to the explicit temporal structure these encodings provide. By directly incorporating information about the spacing between samples (via timestamp or normalized linspace values), these PEs equip the Transformer's global self-attention mechanism to better adapt to the temporal scaling effects introduced by varying sampling rates. This contrasts with CNNs, which typically have localized receptive fields and lack such explicit sequence timing cues, making them less inherently adaptable to these distortions.

B. PERFORMANCE WITH RR INTERVAL FEATURES

Next, we evaluated the influence of adding RR interval features on model performance, first on the original dataset and subsequently on the resampled dataset. For these experiments, we focused on the most promising configurations identified in the previous experiments. Specifically, all models were evaluated using the combined raw ECG and CWT input, which generally resulted in the highest performance. Furthermore, the learned PEs scheme for the Transformer architecture was excluded from these evaluations due to its significantly lower performance and high variance.

The balanced accuracy results for the original dataset with RR interval features are presented in the heatmap in Figure 12. This figure compares the performance of the CNN architecture (with its multi-headed late fusion strategy for RR features) against the Transformer architecture using both early and late fusion strategies. We also included the results without RR features from the previous section for comparison.

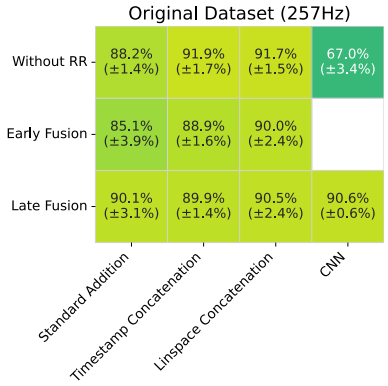

FIGURE 12. Balanced accuracy results for the evaluation set (containing patients not used during training) on the original INCART dataset with the inclusion of RR features. The first four columns show the Transformer model with different positional encodings.

The most notable impact of including the RR interval features was observed for the CNN architecture. Without RR features, the CNN achieved a balanced accuracy of 67.0% ± 3.4%. With the integration of RR features, the CNN's performance increased noticeably to 90.6% ± 0.6%. This substantial improvement resulted in the CNN matching the performance of the Transformer model, highlighting the significant added value of these features for the CNN architecture. This highlights the significant value of these temporally explicit, rate-invariant features for dramatically improving CNN performance on the original dataset. Given that standard CNN architectures lack the explicit temporal information provided by PEs schemes in Transformers these RR features likely fill a critical informational gap by supplying essential temporal context directly, thereby significantly enhancing the CNN's discriminative capabilities.

In contrast, for the Transformer model, the addition of RR interval features did not result in a clear improvement and adding these features, in many cases, actually resulted in a small decrease in performance on the original dataset. When employing the early fusion strategy, all PEs schemes showed a decline in accuracy: Standard Addition dropped to 85.1% ± 3.9% (from 88.2% ± 1.4% without RR), Timestamp Concatenation decreased to 88.9% ± 1.6% (from 91.9% ± 1.7% without RR), and Linspace Concatenation fell to 90.0% ± 2.4% (from 91.7% ± 1.5% without RR).

Late fusion of RR features was generally more effective than early fusion for Transformers on this dataset. However, even with this strategy, performance was still often lower than that of the Transformer without RR features. Specifically, Timestamp Concatenation with late fusion achieved 89.9% ± 1.4%, and Linspace Concatenation PEs reached 90.5% ± 2.4%, both lower than their respective performances without RR features. A small improvement was observed for the

Standard Addition, which achieved 90.1% ± 3.1%. Critically, the peak performance of 91.9% ± 1.7% from the previous experiment without including extra features (achieved by the Timestamp Concatenation PEs) was not matched by any model configuration when RR features were added to the original dataset. This suggests that with rich combined inputs (ECG+CWT), the Transformer might already capture sufficient contextual information, and the simple addition of these specific RR features may not provide substantial new information or could even introduce noise if not optimally integrated when the sampling rate is fixed.

Next, we assessed how the inclusion of RR interval features impacts the models' generalization capabilities when trained and tested on datasets incorporating a sweep of varying sampling rates. The results for this analysis are presented in Figure 13.

FIGURE 13. Balanced accuracy results for the evaluation set (containing patients not used during training) on the resampled dataset using a sweep with 50Hz spacing and the inclusion RR features. The first four columns show the Transformer model with different positional encodings.

The integration of RR features significantly enhanced generalization performance for most configurations when tested on the sweep dataset, compared to their counterparts without RR features. The CNN, which exhibited modest generalization without RR features (69.7% ± 1.6%), saw its performance climb dramatically to 92.6% ± 0.8% with the inclusion of RR features. This score not only represents a substantial improvement but also positions the CNN as one of the best generalizing model configurations when RR features are incorporated, highlighting their role in improving generalization capabilities.

For the Transformer architecture, late fusion again proved generally superior to early fusion for generalization. With early fusion on the sweep dataset, the Standard Addition achieved 66.5% ± 6.0%, Timestamp Concatenation reached 90.4% ± 2.7%, and Linspace Concatenation scored a lower 78.7% ± 9.2%. While the Timestamp Concatenation with early fusion maintained a high level of performance, closely matching its score without RR features (90.7% ± 3.0%), the other early fusion configurations did not show as strong an advantage or even underperformed compared to their "without RR" counterparts when generalizing to new sampling rates.

Late fusion yielded more consistent and often higher generalization performance for the Transformer models on the sweep dataset. The Standard Addition with late fusion achieved a notable 89.1% ± 0.5%, a significant improvement over its performance without RR features (64.8%±6.4%) and its early fusion counterpart. The Timestamp Concatenation with late fusion reached the highest balanced accuracy among all configurations on the sweep dataset, 92.7% ± 0.8%, improving upon its already excellent generalization without RR features (90.7% ± 3.0%). Linspace Concatenation with late fusion also performed well, scoring 90.2% ± 1.3%, an improvement over its performance without RR features (89.2% ± 3.6%) and substantially better than its early fusion result. This indicates that the benefit of RR features for Transformers is notably context-dependent. While these features offered limited additional value, and potentially introduced noise, when Transformers processed rich combined inputs on the fixed-rate original dataset, their explicit and inherently rate-invariant temporal context becomes a crucial anchor when generalizing towards varying sampling rates. In such scenarios of input instability, especially with late fusion, RR intervals provide stable physiological cues that significantly enhance the Transformer's generalization and robustness.

The method of integrating RR interval features proved to be a significant factor where these features offered benefits for generalization. For the Transformer architecture, where both early and late fusion strategies were explicitly compared, late fusion consistently demonstrated superior accuracy over early fusion, particularly in enhancing generalization performance. This suggests that allowing the Transformer to first learn complex representations from the primary ECG and CWT data before integrating the RR feature embeddings is a more effective strategy than combining them at the input level, especially for sampling rate robustness.

C. GENERALIZATION ACROSS DATASETS

To evaluate generalization, we conducted experiments on the MIT-BIH Arrhythmia Database [7], a gold-standard benchmark containing 48 half-hour, two-channel ambulatory ECG records. Using the standard AAMI-recommended inter-patient split (DS1 for training, DS2 for evaluating), we focused on the N, SVEB, and VEB classes for consistency with the INCART evaluation [47]. Due to computational constraints, we evaluated only the top-performing INCART configurations: ECG+CWT inputs for both the CNN and the Transformer architectures, timestamp concatenation positional encoding and late fusing the RR interval features for the Transformer model. The full results are shown in tables 8 to 11 in the Appendix.

This cross-dataset evaluation reveals significant performance differences between the two architectures across all experimental conditions, as summarized in Table 3.

The cross-dataset results consistently show the CNN's higher robustness over the Transformer on the smaller,

TABLE 3. Balanced accuracies on the MIT-BIH dataset for the top-performing configurations (ECG+CWT inputs, Timestamp Concatenation Positional Encoding, Late Fusion for RR Interval Features).

Experiment	Model	Balanced Accuracy
Original SR (without RR)	Transformer	$65.59\% \pm 4.71\%$
	CNN	$82.81\% \pm 0.74\%$
Original SR (with RR)	Transformer	$77.13\% \pm 7.63\%$
	CNN	$92.70\% \pm 0.58\%$
Sweep SR (without RR)	Transformer	$70.71\% \pm 6.35\%$
	CNN	$82.36\% \pm 1.78\%$
Sweep SR (with RR)	Transformer	$70.15\% \pm 8.26\%$
	CNN	$91.89\% \pm 0.62\%$

noisier MIT-BIH dataset. In the first scenario (original sampling rate without RR Features), the CNN achieved 82.81% ± 0.74%, substantially outperforming the Transformer's 65.59% ± 4.71%. This performance difference is a significant change from the INCART results, where the Transformer architecture excelled, the high variance (±4.71%) suggests the Transformer suffered from training instability on this smaller dataset. This instability likely stems from overfitting, as evidenced by the training and evaluation loss curves; the training curves appeared smooth and similar to those from the INCART dataset, but the validation curves were significantly noisier and more variable, failing to follow a clear, steady learning trajectory.

Integrating RR Features through late fusion significantly improved the performance on both architectures. The CNN's performance increased to 92.70% ± 0.58%, nearly matching its INCART performance and demonstrating excellent cross-dataset robustness. The Transformer also saw a substantial gain, reaching 77.13% ± 7.63%, but still lagged behind the CNN with a lower mean value and higher standard deviation. This confirms again that RR features provide valuable, rate-invariant temporal context that aids both architectures, but they cannot fully compensate for the Transformer's tendency to overfit on the smaller, noisier, dataset.

When trained across multiple rates using the 50Hz sampling rate sweep dataset, the CNN maintained near-identical performance (82.36% ± 1.78%), confirming its robust sampling rate invariance. The Transformer surprisingly improved to 70.71% ± 6.35%, suggesting multi-rate training acts as a form of regularization, though the gain was modest compared to its INCART 50Hz sampling rate sweep performance.

When combining the 50Hz sampling rate sweep dataset with RR interval features the CNN achieved 91.89%±0.62%, maintaining its consistency on the MIT-BIH dataset. Conversely, the Transformer only reached a balanced accuracy of 70.15%±8.26%, showing no improvement when adding RR interval features. Its high standard deviation (±8.26%) again suggests training instability taking place.

Across all four scenarios, the CNN demonstrated exceptional robustness, consistently maintaining high performance (above 82% without RR, above 91% with RR). The Transformer, however, showed highly dataset-dependent performance, struggling on MIT-BIH compared to INCART across all conditions. Crucially, RR features consistently and substantially improved CNN performance across datasets, confirming their value as robust, dataset-agnostic temporal features. For the Transformer, RR benefits were context-dependent, failing to help or even slightly degrading performance in the most complex multi-rate training on this smaller, noisier dataset.

We hypothesize that the Transformer's poor generalization stems from its larger model size (±1.8M parameters) interacting with MIT-BIH's smaller size (±110K beats) and higher noise levels. This combination likely causes the Transformer to overfit to noise patterns, a problem that the CNN's architectural inductive biases inherently mitigate. These findings imply that while Transformers excel with sufficient high-quality data, CNNs with RR feature fusion offer more reliable and consistent generalization across diverse clinical settings with limited or variable-quality training data, making them more suitable for real-world clinical deployment.

V. INSIGHTS

This study aimed to answer several key research questions regarding Time-Series Transformers for ECG heartbeat classification. The main insights derived from the experimental results are summarized below:

1) **How do Time-Series Transformers compare to benchmark CNNs in terms of classification performance and, critically, generalization to both unseen patients and new sampling rates?** When expert RR interval features were not used, Time-Series Transformers, particularly those with concatenation-based Positional Encoding (PE) schemes (Timestamp or Linspace) and CWT or combined ECG+CWT inputs, generally outperformed benchmark CNNs in both classification performance on the original dataset and, critically, in generalization to unseen patients and new sampling rates. For instance, Transformers achieved balanced accuracies over 90% (e.g., 91.9% ± 1.7 on original, 90.7% ± 3.0 on sweep) while the CNN remained below 70% in these scenarios. With the inclusion of RR features (via late fusion or equivalent for CNNs), both architectures achieved comparable top-tier performance (around 90.6% for CNN and 90.1% − 90.5% for best Transformers on original; 92.6% for CNN and 92.7% for Transformer on sweep), significantly boosting the CNN's capabilities.
2) **What is the impact of different input representations on the generalization capabilities of these architectures?** For generalization capabilities, CWT scalograms and the combination of raw ECG + CWT were generally more effective than raw ECG alone, especially for Transformer models using concatenation-based PEs. Raw ECG input showed strong generalization (83.9% ± 0.8) only with the Transformer using standard additive sinusoidal PE. For CNNs, CWT input also provided better generalization

TABLE 4. Detailed results for the experiments using the INCART dataset, without RR interval features, and on the original sampling rate.

			Global				Normal			SVEB			VEB		
Model type	Input type	PE scheme	Accuracy	Precision	Recall	F1 score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Transformer	ECG	Standard Addition	89.9% ± 1.1	63.8% ± 2.5	84.9% ± 1.3	68.8% ± 2.4	98.1% ± 0.4	91.6% ± 1.6	94.8% ± 0.7	27.6% ± 7.9	85.7% ± 4.5	41.0% ± 8.3	65.5% ± 6.1	77.5% ± 2.8	70.7% ± 3.2
		Zero (Learned)	61.2% ± 18.9	47.3% ± 8.3	67.2% ± 9.6	43.7% ± 12.6	94.4% ± 2.4	59.8% ± 22.0	71.0% ± 18.2	7.3% ± 2.8	71.5% ± 10.6	13.0% ± 4.4	40.3% ± 21.6	70.4% ± 5.4	47.1% ± 16.8
		Timestamp Concatenation	69.3% ± 14.5	51.8% ± 15.4	67.1% ± 9.0	50.4% ± 16.3	94.6% ± 1.7	69.5% ± 16.3	79.2% ± 10.7	15.4% ± 19.6	63.6% ± 9.8	21.2% ± 21.9	45.4% ± 27.0	68.2% ± 6.0	50.8% ± 18.6
		Linspace Concatenation	56.2% ± 20.4	45.3% ± 5.5	63.8% ± 5.8	39.7% ± 13.1	95.0% ± 1.0	53.9% ± 24.5	64.5% ± 26.7	4.9% ± 0.8	65.5% ± 5.2	9.1% ± 1.3	36.0% ± 15.8	71.9% ± 8.7	45.4% ± 13.8
	CWT	Standard Addition	85.6% ± 3.1	58.8% ± 3.2	86.6% ± 1.7	63.7% ± 3.5	98.9% ± 0.3	85.1% ± 3.6	91.5% ± 2.2	16.3% ± 3.5	85.7% ± 2.8	27.2% ± 4.9	61.2% ± 7.6	89.0% ± 1.1	72.3% ± 5.4
		Zero (Learned)	48.3% ± 35.3	26.0% ± 21.1	49.3% ± 20.2	28.2% ± 21.9	55.8% ± 45.7	48.0% ± 41.3	50.8% ± 42.1	7.5% ± 10.0	50.1% ± 42.2	12.0% ± 15.3	14.8% ± 14.9	49.9% ± 41.9	21.6% ± 20.3
		Timestamp Concatenation	92.4% ± 2.9	69.8% ± 3.5	90.4% ± 1.5	74.5% ± 5.0	99.4% ± 0.2	93.0% ± 3.5	96.1% ± 2.0	31.6% ± 14.2	90.4% ± 2.4	44.8% ± 15.9	78.4% ± 5.8	87.9% ± 4.5	82.7% ± 4.1
		Linspace Concatenation	87.9% ± 3.5	62.8% ± 4.7	89.2% ± 0.9	67.5% ± 4.7	99.6% ± 0.2	87.7% ± 4.3	93.2% ± 2.4	21.0% ± 9.3	90.5% ± 1.4	33.1% ± 11.1	67.7% ± 13.6	89.4% ± 2.9	76.1% ± 8.4
	ECG + CWT	Standard Addition	86.9% ± 4.0	62.1% ± 4.9	88.2% ± 1.4	65.9% ± 4.7	98.6% ± 0.2	87.0% ± 4.8	92.4% ± 2.8	17.3% ± 6.2	91.8% ± 2.0	28.6% ± 8.5	70.3% ± 11.5	85.9% ± 2.6	76.7% ± 5.5
		Zero (Learned)	37.7% ± 40.3	12.6% ± 13.4	33.3% ± 0.0	14.2% ± 13.9	34.7% ± 42.5	40.0% ± 49.0	37.2% ± 45.5	0.6% ± 0.7	40.0% ± 49.0	1.2% ± 1.4	2.4% ± 4.7	20.0% ± 40.0	4.2% ± 8.4
		Timestamp Concatenation	93.4% ± 2.2	71.0% ± 5.0	91.9% ± 1.7	77.0% ± 5.4	99.6% ± 0.3	93.7% ± 2.6	96.6% ± 1.3	37.7% ± 12.4	90.7% ± 1.7	51.9% ± 13.0	75.8% ± 4.7	91.3% ± 4.6	82.6% ± 2.5
		Linspace Concatenation	92.6% ± 1.2	67.6% ± 2.4	91.7% ± 1.6	73.1% ± 2.2	99.7% ± 0.1	92.7% ± 1.2	96.0% ± 0.6	24.6% ± 3.6	90.4% ± 1.1	38.5% ± 4.3	78.6% ± 7.0	92.0% ± 3.5	84.7% ± 5.2
CNN	ECG	-	58.2% ± 4.6	46.3% ± 1.3	65.8% ± 1.5	42.9% ± 1.9	96.9% ± 0.6	55.3% ± 5.5	70.3% ± 4.3	3.6% ± 0.5	62.9% ± 8.7	6.8% ± 1.0	38.3% ± 3.6	79.3% ± 4.2	51.5% ± 3.4
	CWT	-	61.7% ± 4.6	50.5% ± 2.7	67.8% ± 5.7	47.6% ± 3.4	96.8% ± 1.2	58.8% ± 5.2	73.0% ± 4.1	3.2% ± 0.7	61.3% ± 13.0	6.1% ± 1.3	51.6% ± 6.5	83.3% ± 3.6	63.6% ± 5.6
	ECG + CWT	-	60.7% ± 6.1	50.3% ± 1.3	67.0% ± 3.4	47.0% ± 1.9	96.9% ± 0.8	57.7% ± 7.7	72.0% ± 5.5	3.0% ± 0.3	59.6% ± 14.0	5.7% ± 0.6	51.0% ± 3.6	83.6% ± 4.2	63.2% ± 2.7

TABLE 5. Detailed results for the experiments using the INCART dataset, without RR interval features, and on the resampled dataset using a sweep with 50Hz spacing.

			Global				Normal			SVEB			VEB		
Model type	Input type	PE scheme	Accuracy	Precision	Recall	F1 score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Transformer	ECG	Standard Addition	87.5% ± 1.8	59.8% ± 2.2	83.9% ± 0.8	64.1% ± 2.4	98.3% ± 0.1	88.9% ± 2.1	93.3% ± 1.2	18.6% ± 4.7	86.8% ± 1.8	30.4% ± 6.2	62.6% ± 6.0	76.0% ± 1.4	68.5% ± 3.3
		Zero (Learned)	37.7% ± 15.0	39.5% ± 4.6	57.5% ± 3.2	28.6% ± 9.2	92.4% ± 0.9	33.7% ± 17.4	47.3% ± 16.3	3.9% ± 0.3	72.9% ± 6.5	7.4% ± 0.6	22.3% ± 13.0	65.8% ± 4.7	31.1% ± 11.1
		Timestamp Concatenation	71.0% ± 9.9	46.7% ± 4.4	68.8% ± 5.1	47.9% ± 6.8	95.6% ± 1.5	71.7% ± 11.2	81.5% ± 7.8	11.2% ± 5.6	69.3% ± 6.3	18.8% ± 8.2	33.4% ± 8.4	65.3% ± 6.3	43.5% ± 7.0
		Linspace Concatenation	61.5% ± 24.6	47.2% ± 9.7	65.3% ± 8.1	42.8% ± 16.2	94.5% ± 3.0	60.7% ± 28.2	69.4% ± 27.7	6.6% ± 2.9	67.8% ± 6.5	11.9% ± 4.6	40.5% ± 23.8	67.4% ± 9.7	47.2% ± 19.1
	CWT	Standard Addition	58.3% ± 6.0	45.4% ± 3.1	64.4% ± 2.3	42.4% ± 4.6	94.9% ± 0.9	55.9% ± 6.9	70.1% ± 5.6	3.7% ± 0.4	58.6% ± 2.8	7.0% ± 0.7	37.6% ± 9.1	78.7% ± 4.6	50.2% ± 8.5
		Zero (Learned)	27.4% ± 5.4	37.5% ± 0.1	57.1% ± 3.6	22.3% ± 3.1	92.5% ± 0.3	20.2% ± 6.5	32.7% ± 9.1	3.8% ± 0.5	70.3% ± 15.1	7.1% ± 1.0	16.3% ± 0.3	80.8% ± 2.8	27.1% ± 0.3
		Timestamp Concatenation	91.1% ± 2.5	66.8% ± 3.8	90.1% ± 2.5	71.9% ± 4.6	99.1% ± 0.5	91.3% ± 2.6	95.0% ± 1.5	25.2% ± 9.6	88.8% ± 3.0	38.3% ± 11.5	76.1% ± 8.8	90.2% ± 4.7	82.4% ± 6.4
		Linspace Concatenation	89.2% ± 4.7	63.3% ± 7.9	87.8% ± 3.9	69.5% ± 8.0	99.1% ± 0.5	89.6% ± 4.7	94.0% ± 2.8	29.8% ± 10.1	88.1% ± 2.1	43.7% ± 11.6	61.0% ± 14.9	85.7% ± 5.7	70.7% ± 11.3
	ECG + CWT	Standard Addition	51.9% ± 18.2	45.3% ± 5.9	64.8% ± 6.4	39.0% ± 12.1	95.8% ± 2.1	47.9% ± 20.8	61.3% ± 20.0	4.3% ± 1.0	63.4% ± 3.7	8.0% ± 1.7	35.8% ± 16.0	83.0% ± 3.8	47.6% ± 16.4
		Zero (Learned)	30.8% ± 14.8	37.7% ± 2.6	58.9% ± 3.9	24.2% ± 8.7	91.5% ± 2.6	24.3% ± 17.1	36.0% ± 18.7	5.6% ± 2.6	73.1% ± 8.7	10.2% ± 4.2	15.9% ± 2.6	79.4% ± 1.9	26.4% ± 3.4
		Timestamp Concatenation	90.4% ± 3.4	67.8% ± 4.9	90.7% ± 3.0	73.3% ± 6.4	99.5% ± 0.3	90.1% ± 3.3	94.5% ± 1.9	35.3% ± 20.2	88.9% ± 1.4	47.2% ± 19.9	68.6% ± 12.1	93.1% ± 6.4	78.2% ± 9.0
		Linspace Concatenation	90.1% ± 4.7	65.1% ± 7.0	89.2% ± 3.6	70.8% ± 7.2	99.0% ± 0.3	90.6% ± 4.8	94.5% ± 2.8	28.9% ± 9.1	90.9% ± 2.4	43.1% ± 10.2	67.4% ± 15.4	86.1% ± 4.2	74.9% ± 11.5
CNN	ECG	-	57.4% ± 3.5	44.6% ± 1.3	62.8% ± 2.0	41.2% ± 1.7	96.1% ± 0.7	55.2% ± 4.2	70.0% ± 3.4	3.4% ± 0.2	57.2% ± 7.1	6.5% ± 0.3	34.3% ± 3.3	76.1% ± 3.1	47.2% ± 3.5
	CWT	-	65.8% ± 3.9	52.9% ± 2.1	70.2% ± 1.3	50.8% ± 1.6	97.7% ± 0.4	63.3% ± 4.5	76.7% ± 3.3	3.6% ± 0.5	60.0% ± 5.1	6.7% ± 0.9	57.4% ± 6.6	87.2% ± 2.2	69.0% ± 4.6
	ECG + CWT	-	64.5% ± 4.7	50.5% ± 2.2	69.7% ± 1.6	48.7% ± 3.1	97.2% ± 0.4	62.0% ± 5.1	75.6% ± 4.1	3.7% ± 0.3	60.7% ± 3.3	7.0% ± 0.6	50.5% ± 6.2	86.3% ± 2.8	63.5% ± 5.5

(70.1% ± 1.3) than raw ECG (62.8% ± 2.0) when RR features were absent. The inclusion of RR features further enhanced generalization for all input types, but the combined ECG+CWT with RR features generally yielded the most robust models for both architectures.

3) **For Transformer models, which Positional Encoding schemes are most effective for ECG heartbeat classification, especially concerning their effect on sampling rate invariance?** Concatenation-based PEs schemes (Timestamp Concatenation and Linspace Concatenation) were the most effective for Transformer models in ECG heartbeat classification, consistently yielding high accuracy and strong generalization to varying sampling rates, especially with CWT or combined inputs (e.g., Timestamp PE: 91.9% ± 1.7 on original, 90.7% ± 3.0 on sweep, without RR). Standard additive sinusoidal PEs performed exceptionally well with raw ECG input (84.9% ± 1.3 on original, 83.9% ± 0.8 on sweep) but less so with CWT-based inputs concerning sampling rate invariance. The learned PEs scheme significantly underperformed across all conditions.
4) **How does the integration of expert-derived RR interval features affect performance and generalization, comparing different feature fusion strategies?** The integration of expert-derived RR interval features had a profound positive impact, especially on the CNNs, increasing its performance and generalization significantly (e.g., from 69.7% to 92.6%). For Transformers, RR features were most beneficial for enhancing generalization to new sampling rates, particularly when using late fusion (e.g., Timestamp PE generalization improved from 90.7% to 92.7%). Early fusion of RR features in Transformers was generally less effective and sometimes detrimental, while late fusion proved superior. The results highlight that these rate-invariant features are crucial for achieving peak robustness to sampling rate variations in both architectures.
5) **How do these architectures generalize across different ECG databases?** Cross-dataset evaluation on the MIT-BIH database reveals that CNNs demonstrate superior robustness, maintaining consistent performance across both INCART and MIT-BIH datasets. In contrast, Transformers show dataset-dependent performance, excelling on INCART but severely degrading on the smaller, noisier MIT-BIH dataset. This suggests that while Transformers can achieve state-of-the-art results with sufficient high-quality data, CNNs with RR feature fusion offer more reliable generalization for clinical deployment scenarios with diverse recording conditions or limited training data.

VI. CONCLUSION

This study systematically compared Time-Series Transformers and CNN for ECG heartbeat classification, emphasizing generalization to new patients and varying sampling rates. Key findings demonstrate a crucial trade-off: while appropriately configured Transformers (e.g., with concatenation-based PE and CWT inputs) achieved state-of-the-art performance on the large INCART dataset, the simpler CNN architecture proved significantly more robust and consistent when evaluated on the smaller, noisier MIT-BIH dataset. Critically, the integration of rate-invariant RR interval features through late fusion was essential, not only boosting the Transformer's rate invariance but, more importantly, increasing the CNN's performance to consistently high levels (over 91% balanced accuracy) across both diverse datasets. This reliability confirms that models with strong inductive biases (like CNNs) are highly

TABLE 6. Detailed results for the experiments using the INCART dataset, with RR interval features, and on the original sampling rate.

			Global				Normal			SVEB			VEB		
Model type	PE scheme	Fusion	Accuracy	Precision	Recall	F1 score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Transformer	Standard Addition	Early	82.5% ± 6.0	57.1% ± 4.6	85.1% ± 3.9	61.1% ± 5.7	98.4% ± 0.9	81.7% ± 6.7	89.2% ± 4.4	14.6% ± 5.0	85.9% ± 4.0	24.6% ± 7.4	58.3% ± 10.1	87.7% ± 2.7	69.6% ± 7.7
		Late	91.8% ± 2.7	66.8% ± 4.3	90.2% ± 3.1	72.5% ± 4.3	99.5% ± 0.1	92.1% ± 2.9	95.6% ± 1.6	27.3% ± 5.7	88.2% ± 4.4	41.4% ± 6.8	73.6% ± 10.9	90.2% ± 4.6	80.5%% ± 7.5
	Timestamp Concatenation	Early	91.4% ± 2.4	66.3% ± 5.0	88.9% ± 1.7	71.0% ± 3.7	98.9% ± 0.7	92.4% ± 3.2	95.5% ± 1.7	25.1% ± 4.6	90.0% ± 3.5	39.0% ± 5.7	74.9% ± 10.7	84.3% ± 6.3	78.6% ± 4.3
		Late	91.2% ± 1.5	66.9% ± 2.7	90.0% ± 1.4	71.9% ± 3.0	99.4% ± 0.2	91.7% ± 1.6	95.4% ± 0.9	27.5% ± 9.4	91.0% ± 1.8	41.3% ± 11.2	73.9% ± 12.7	87.1% ± 3.9	79.1% ± 5.8
	Linspace Concatenation	Early	93.3% ± 3.7	72.0% ± 8.5	90.0% ± 2.4	76.8% ± 7.6	99.4% ± 0.3	94.3% ± 4.3	96.7% ± 2.2	39.6% ± 13.9	89.0% ± 3.4	53.2% ± 13.3	76.9% ± 15.5	86.8% ± 6.5	80.5% ± 9.2
		Late	91.3% ± 3.0	66.5% ± 5.5	90.5% ± 2.4	72.0% ± 5.8	99.7% ± 0.1	91.5% ± 2.9	95.4% ± 1.6	27.2% ± 9.7	90.2% ± 2.6	40.7% ± 10.8	72.8% ± 11.8	89.9% ± 5.7	80.0% ± 8.2
CNN	-	-	92.3% ± 1.4	67.9% ± 1.7	90.7% ± 0.6	72.1% ± 2.2	99.5% ± 0.1	92.7% ± 1.5	96.0% ± 0.8	21.2% ± 3.9	89.0% ± 1.2	34.1% ± 5.2	82.9% ± 3.5	90.3% ± 1.5	86.4% ± 2.4

TABLE 7. Detailed results for the experiments using the INCART dataset, with RR interval features, and on the resampled dataset using a sweep with 50Hz spacing.

			Global				Normal			SVEB			VEB		
Model type	PE scheme	Fusion	Accuracy	Precision	Recall	F1 score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Transformer	Standard Addition	Early	64.2% ± 7.0	48.3% ± 2.0	66.5% ± 6.0	46.6% ± 2.8	96.3% ± 1.4	62.3% ± 8.2	75.4% ± 6.2	3.8% ± 1.4	55.9% ± 20.4	7.1% ± 2.6	44.9% ± 5.9	81.2% ± 3.7	57.4% ± 4.1
		Late	87.4% ± 1.0	61.9% ± 2.7	89.1% ± 0.5	65.8% ± 1.4	99.6% ± 0.1	87.1% ± 1.2	92.9% ± 0.7	15.0% ± 1.8	90.4% ± 1.5	25.6% ± 2.6	71.1% ± 9.7	89.8% ± 1.5	78.9% ± 5.7
	Timestamp Concatenation	Early	90.9% ± 4.0	67.9% ± 6.0	90.4% ± 2.7	73.0% ± 7.6	99.4% ± 0.2	91.1% ± 4.3	95.1% ± 2.4	31.0% ± 18.1	90.9% ± 2.2	43.5% ± 19.4	73.4% ± 6.4	89.3% ± 4.1	80.3% ± 3.8
		Late	95.2% ± 1.0	73.8% ± 3.5	92.7% ± 0.8	79.6% ± 3.6	99.6% ± 0.1	95.6% ± 0.9	97.6% ± 0.4	40.3% ± 10.4	90.5% ± 3.4	54.7% ± 9.2	81.5% ± 3.3	91.8% ± 2.7	86.3% ± 2.6
	Linspace Concatenation	Early	78.3% ± 12.9	56.4% ± 8.7	78.8% ± 9.2	58.6% ± 12.5	97.2% ± 1.8	78.6% ± 13.9	86.3% ± 9.6	17.8% ± 15.2	81.8% ± 8.9	26.9% ± 19.4	54.2% ± 12.4	75.9% ± 10.7	62.7% ± 11.5
		Late	92.2% ± 1.1	66.2% ± 2.2	90.2% ± 1.3	71.8% ± 2.2	99.6% ± 0.2	92.7% ± 1.3	96.0% ± 0.7	25.5% ± 3.1	90.1% ± 1.6	39.6% ± 3.6	73.3% ± 4.2	87.8% ± 4.1	79.9% ± 3.5
CNN	-	-	94.6% ± 0.6	71.6% ± 1.5	92.6% ± 0.8	77.6% ± 1.6	99.6% ± 0.1	94.9% ± 0.6	97.2% ± 0.3	33.9% ± 4.7	90.6% ± 1.1	49.2% ± 4.9	81.4% ± 4.6	92.3% ± 1.3	86.4% ± 2.9

TABLE 8. Detailed results for the experiments using the MIT-BIH dataset, without RR interval features, and on the original sampling rate.

	Global				Normal			SVEB			VEB		
Model type	Accuracy	Precision	Recall	F1 score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Transformer	70.82% ± 7.49%	48.92% ± 4.32%	65.59% ± 4.71%	50.18% ± 6.06%	96.58% ± 1.31%	70.40% ± 8.89%	81.10% ± 5.78%	7.84% ± 2.19%	31.50% ± 14.61%	12.18% ± 3.52%	42.34% ± 12.18%	94.86% ± 3.97%	57.27% ± 11.95%
CNN	75.08% ± 0.94%	59.12% ± 1.79%	82.81% ± 0.74%	61.40% ± 1.18%	98.71% ± 0.14%	73.07% ± 1.02%	83.97% ± 0.67%	13.44% ± 0.92%	80.22% ± 1.88%	23.01% ± 1.34%	65.22% ± 5.96%	95.15% ± 0.91%	77.22% ± 3.99%

TABLE 9. Detailed results for the experiments using the MIT-BIH dataset, with RR interval features, and on the original sampling rate.

	Global				Normal			SVEB			VEB		
Model type	Accuracy	Precision	Recall	F1 score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Transformer	83.84% ± 2.20%	62.46% ± 4.81%	77.13% ± 7.63%	65.04% ± 3.99%	98.11% ± 0.71%	84.49% ± 3.51%	90.75% ± 2.06%	15.21% ± 2.93%	57.39% ± 21.61%	23.79% ± 5.36%	74.05% ± 12.07%	89.52% ± 8.10%	80.59% ± 8.99%
CNN	91.25% ± 1.34%	73.03% ± 1.55%	92.70% ± 0.58%	78.17% ± 1.81%	99.55% ± 0.06%	90.82% ± 1.49%	94.98% ± 0.83%	33.27% ± 3.69%	91.16% ± 1.15%	48.61% ± 3.88%	86.28% ± 1.66%	96.11% ± 0.71%	90.92% ± 1.04%

TABLE 10. Detailed results for the experiments using the MIT-BIH dataset, without RR interval features, and on the resampled dataset using a sweep with 50Hz spacing.

	Global				Normal			SVEB			VEB		
Model type	Accuracy	Precision	Recall	F1 score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Transformer	61.46% ± 7.15%	49.82% ± 3.55%	70.71% ± 6.35%	48.41% ± 4.76%	97.98% ± 0.75%	59.30% ± 8.40%	73.53% ± 6.92%	7.53% ± 2.22%	63.08% ± 13.92%	13.41% ± 3.80%	43.95% ± 10.06%	89.76% ± 9.37%	58.30% ± 9.62%
CNN	75.32% ± 1.74%	58.15% ± 1.52%	82.36% ± 1.78%	60.56% ± 1.33%	98.69% ± 0.28%	73.75% ± 1.95%	84.41% ± 1.28%	12.51% ± 1.26%	78.63% ± 4.73%	21.58% ± 2.00%	63.24% ± 5.12%	94.68% ± 0.94%	75.69% ± 3.70%

TABLE 11. Detailed results for the experiments using the MIT-BIH dataset, with RR interval features, and on the resampled dataset using a sweep with 50Hz spacing.

	Global				Normal			SVEB			VEB		
Model type	Accuracy	Precision	Recall	F1 score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Transformer	80.25% ± 8.57%	54.16% ± 6.51%	70.15% ± 8.26%	56.92% ± 7.60%	97.93% ± 0.71%	81.30% ± 10.29%	88.51% ± 6.48%	13.99% ± 7.37%	42.63% ± 20.86%	19.87% ± 9.43%	50.57% ± 13.82%	86.52% ± 12.43%	62.37% ± 12.31%
CNN	89.39% ± 1.26%	67.98% ± 1.87%	91.89% ± 0.62%	74.01% ± 1.73%	99.60% ± 0.07%	88.88% ± 1.43%	93.93% ± 0.79%	29.18% ± 2.56%	91.69% ± 1.28%	44.20% ± 2.97%	75.17% ± 4.46%	95.10% ± 0.53%	83.89% ± 2.73%

advantageous for clinical deployment where training data is often limited and noisy. This research highlights that robust ECG classifiers depend on careful design of architectures, positional encoding, input modalities, and expert feature integration, and underscores that architectural robustness is critically important for real-world generalization.

Future work should prioritize investigating more advanced PEs schemes, including strategies to improve learned PEs through refined optimization or initialization techniques. Enhancements to input processing are also crucial, such as exploring alternative embedding and patching strategies. For instance, replacing the current linear embedding with a CNNs-based encoder for patches could better capture local ECG morphology and potentially improve computational efficiency for both training and inference. The robustness and generalizability of these models should be further validated by evaluating performance across multiple diverse datasets, especially focusing on conditions where data is limited and noisy, and exploring training methodologies that leverage combined datasets to mitigate architectural overfitting issues. Finally, a significant avenue involves introducing self-supervised pretraining for Transformers to facilitate the development of larger, more complex architectures. Such advancements could also inform future efforts towards large-scale model development, such as ECG Foundation Models [9], [10], [48], [49], which will depend on effectively handling the inherent heterogeneity found in real-world clinical data.

APPENDIX

DETAILED RESULTS

See Tables 4–11.

REFERENCES

[1] World Health Organization. (2021). *Cardiovascular Diseases (CVDS)*. Accessed: May 28, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] G. A. Roth et al., "Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the gbd 2019 study," *J. Amer. College Cardiol.*, vol. 76, no. 25, pp. 2982–3021, 1990.

[3] G. A. Mensah, G. A. Roth, and V. Fuster, "The global burden of cardiovascular diseases and risk factors: 2020 and beyond," *J. American College Cardiol.*, vol. 74, no. 20, pp. 2529–2532, 2019.

[4] P. Kligfield, L. S. Gettes, J. J. Bailey, R. Childers, B. J. Deal, E. W. Hancock, G. Van Herpen, J. A. Kors, P. Macfarlane, D. M. Mirvis, O. Pahlm, P. Rautaharju, and G. S. Wagner, "Recommendations for the standardization and interpretation of the electrocardiogram: Part I: The electrocardiogram and its technology: A scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society endorsed by the International Society for Computerized Electrocardiology," *Circulation*, vol. 115, no. 10, pp. 1306–1324, 2007.

[5] J. M. Prutkin and J. A. Drezner, "Training and experience matter: Improving athlete ECG screening, interpretation, and reproducibility," *Circulat., Cardiovascular Quality Outcomes*, vol. 10, no. 8, Aug. 2017, Art. no. 003881.

[6] D. Massel, "Observer variability in ECG interpretation for thrombolysis eligibility: Experience and context matter," *J. Thrombosis Thrombolysis*, vol. 15, no. 3, pp. 131–140, Jun. 2003.

[7] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, Mar. 2001.

[8] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, no. 1, pp. 1–15, May 2020.

[9] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutiérrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Med.*, vol. 29, no. 8, pp. 1930–1940, 2023.

[10] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, Apr. 2023.

[11] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mech. Syst. Signal Process.*, vol. 151, Apr. 2021, Art. no. 107398.

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

[13] T. Wang, C. Lu, Y. Sun, M. Yang, C. Liu, and C. Ou, "Automatic ECG classification using continuous wavelet transform and convolutional neural network," *Entropy*, vol. 23, no. 1, p. 119, Jan. 2021.

[14] Y. Zhao, J. Cheng, P. Zhan, and X. Peng, "ECG classification using deep CNN improved by wavelet transform," *Comput., Mater. Continua*, vol. 64, no. 3, pp. 1615–1628, 2020.

[15] O. Ozaltin and O. Yeniay, "A novel proposed CNN–SVM architecture for ECG scalograms classification," *Soft Comput.*, vol. 27, no. 8, pp. 4639–4658, 2023.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[17] J. Cheng, Q. Zou, and Y. Zhao, "ECG signal classification based on deep CNN and BiLSTM," *BMC Med. Informat. Decis. Making*, vol. 21, no. 1, pp. 1–12, Dec. 2021.

[18] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun, "Transformers in time series: A survey," 2022, *arXiv:2202.07125*.

[19] J. Sun, J. Xie, and H. Zhou, "EEG classification with transformer-based models," in *Proc. IEEE 3rd Global Conf. Life Sci. Technol. (LifeTech)*, Mar. 2021, pp. 92–93.

[20] R. Hu, J. Chen, and L. Zhou, "A transformer-based deep neural network for arrhythmia detection using continuous ECG signals," *Comput. Biol. Med.*, vol. 144, May 2022, Art. no. 105325.

[21] M. R. Islam, M. Qaraqe, K. Qaraqe, and E. Serpedin, "CAT-Net: Convolution, attention, and transformer based network for single-lead ECG arrhythmia classification," *Biomed. Signal Process. Control*, vol. 93, Jul. 2024, Art. no. 106211.

[22] X. Zhang, M. Lin, Y. Hong, H. Xiao, C. Chen, and H. Chen, "MSFT: A multi-scale feature-based transformer model for arrhythmia classification," *Biomed. Signal Process. Control*, vol. 100, Feb. 2025, Art. no. 106968.

[23] A. Natarajan, Y. Chang, S. Mariani, A. Rahman, G. Boverman, S. G. Vij, and J. Rubin, "A wide and deep transformer neural network for 12-lead ECG classification," in *Proc. Comput. Cardiol.*, Sep. 2020, pp. 1–4.

[24] L. Wei and Y. Li, "A multi-scale CNN-transformer parallel network for 12-lead ECG signal classification," *Signal, Image Video Process.*, vol. 19, no. 8, pp. 1–11, Aug. 2025.

[25] C.-L. Liu, B. Xiao, and C.-H. Hsieh, "Multimodal fusion of spatial–temporal and frequency representations for enhanced ECG classification," *Inf. Fusion*, vol. 118, Jun. 2025, Art. no. 102999.

[26] C. Che, P. Zhang, M. Zhu, Y. Qu, and B. Jin, "Constrained transformer network for ECG signal processing and arrhythmia classification," *BMC Med. Informat. Decis. Making*, vol. 21, no. 1, p. 184, Dec. 2021.

[27] A. Baevski and M. Auli, "Adaptive input representations for neural language modeling," 2018, *arXiv:1809.10853*.

[28] C. Alexakis, H. O. Nyongesa, R. Saatchi, N. Harris, C. C. Davies, C. Emery, R. H. Ireland, and S. Heller, "Feature extraction and classification of electrocardiogram (ECG) signals related to hypoglycaemia," in *Proc. Comput. Cardiol.*, Sep. 2003, pp. 537–540.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2025, pp. 5998–6008.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, vol. 1, 2019, pp. 4171–4186.

[31] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," Tech. Rep., 2018.

[32] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2021, pp. 2114–2124.

[33] H. Ryu, S. Yu, and K. Y. Lee, "TI-former: A time-interval prediction transformer for timestamped sequences," in *Proc. IEEE/ACIS 21st Int. Conf. Softw. Eng. Res., Manage. Appl. (SERA)*, May 2023, pp. 319–325.

[34] D. Zhang, J. Gao, and X. Li, "Multivariate time series classification with crucial timestamps guidance," *Expert Syst. Appl.*, vol. 255, Dec. 2024, Art. no. 124591.

[35] C. Ye, B. V. K. Vijaya Kumar, and M. T. Coimbra, "Heartbeat classification using morphological and dynamic features of ECG signals," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 10, pp. 2930–2941, Oct. 2012.

[36] G. Garcia, G. Moreira, D. Menotti, and E. Luz, "Inter-patient ECG heartbeat classification with temporal VCG optimized by PSO," *Sci. Rep.*, vol. 7, no. 1, p. 10543, Sep. 2017.

[37] G. Yan, S. Liang, Y. Zhang, and F. Liu, "Fusing transformer model with temporal features for ECG heartbeat classification," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 898–905.

[38] A. Abid and O. Cheikhrouhou, "A data-augmented vision transformer model for robust multi-label ECG arrhythmia classification," *Int. J. Inf. Technol.*, vol. 17, no. 4, pp. 2287–2293, May 2025.

[39] T. De Waele, D. Peralta, E. De Poorter, and A. Shahid, "Improved deep learning based ECG classification through automated feature selection and weighted loss function," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2024, pp. 1–8.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. J. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2012.

[41] G. Lee, R. Gommers, F. Waselewski, K. Wohlfahrt, and A. O'Leary, "PyWavelets: A Python package for wavelet analysis," *J. Open Source Softw.*, vol. 4, no. 36, p. 1237, Apr. 2019.

[42] P. Virtanen et al., "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[45] R. Shimizu, K. Asako, H. Ojima, S. Morinaga, M. Hamada, and T. Kuroda, "Balanced mini-batch training for imbalanced image data classification with neural network," in *Proc. 1st Int. Conf. Artif. Intell. Industries (AI4I)*, Sep. 2018, pp. 27–30.

[46] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*.

[47] P. D. Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 7, pp. 1196–1206, Jul. 2004.

[48] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.

[49] T. Mehari and N. Strodthoff, "Towards quantitative precision for ECG analysis: Leveraging state space models, self-supervision and patient metadata," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 11, pp. 5326–5334, Nov. 2023.

TIMO DE WAELE received the M.Sc. degree in computer science from Ghent University, Belgium, in 2020. After his graduation, he started working as a Ph.D. Student with the iWINE-UGent/IDLab/imec Research Group. In 2021, he became a Ph.D. Fellow of the FWO-V (Research Foundation-Flanders). His scientific work mainly focuses on developing deep learning based algorithms using time-series data from bio signals, such as electrocardiograms and movement data from accelerometers and gyroscopes.

JARON FONTAINE received the M.Sc. degree (Hons.) in information engineering technology from Ghent University, Ghent, Belgium, in 2017, and the Ph.D. degree in information engineering technology from the IDLab, Department of Information Technology (INTEC), Ghent University. His main research interests include machine learning (ML) techniques for wireless network applications (e.g., indoor localization systems using ultra-wideband, wireless technology recognition, and healthcare activity monitoring). Specifically, he focuses on optimizations of ML to run on edge and embedded devices and foundation models for wireless networks to allow task agnostic AI in new environments and networks. He has already published and co-authored numerous papers on these topics in journals and presented his work at conferences. He was a recipient of the FWO-SB Grant and funded by the Scientific Research Flanders (Belgium, FWO-Vlaanderen).

ELI DE POORTER received the master's degree in computer science engineering from Ghent University, Belgium, in 2006, and the Ph.D. degree from the Department of Information Technology, Ghent University, in 2011. He became a Professor with IDLab, in 2015. Since 2017, he has been affiliated with the IMEC Research Institute. He is a Professor with the IDLab Research Group from Ghent University and imec (https://idlab.technology). His team performs research on wireless communication technologies, such as (indoor) localization solutions, wireless IoT solutions, and machine learning for wireless systems. He performs both fundamental and applied research. For his fundamental research, he is currently the coordinator of several research projects (H2020, SBO, FWO, and GOA) and has over 200 publications in international journals or in the proceedings of international conferences. For his applied research, he collaborates with (Flemish) industry partners to transfer research results to industrial applications as well as to solve challenging industrial research problems.

ADNAN SHAHID (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in computer engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2006 and 2010, respectively, and the Ph.D. degree in information and communication engineering from Sejong University, Seoul, South Korea, in 2015. From March 2015 to August 2015, he worked as a Postdoctoral Researcher at Yonsei University, Seoul. From September 2015 to June 2016, he worked as an Assistant Professor with Taif University, Taif, Saudi Arabia. He is currently working as a Senior Researcher with IDLab, which is a core research group of imec with research activities embedded with Ghent University and the University of Antwerp, Antwerp, Belgium. He is the Lead of ''artificial intelligence (AI)/machine learning (ML) for wireless,'' which is a research group at iWINE-UGent/IDLab/imec. He is leading the Erasmus+ DigiHealth-Asia Project, from 2021 to 2024, and European Space Agency (ESA) MRC100 Project, from 2022 to 2023. He is also involved in the imec ICON 5GECO Project, from 2022 to 2023, and the MSCA Staff Exchanges EVOLVE Project, from 2022 to 2025. He has been the Technical Manager of the ESA CODYSUN Project. He has also been involved in several projects: DARPA Spectrum Collaboration Challenge (SC2); European H2020 Research Projects, such as eWINE and WiSHFUL; and national projects, such as SAMURAI, IDEAL-IOT, Cognitive Wireless Networking Management, Smartwaterways, and UWB-AAA. His research interests include AI/ML for wireless communications and networks, resource optimization, interference management, self-organizing networks, the Internet of Things, localization, and 5G/6G networks. He is a Voting Member and the Secretary of IEEE P1900.8—Standard for Training, Testing, and Evaluating Machine-Learned Spectrum Awareness Models. He is also serving as an Associate Editor for various journals, such as IEEE Access and *Journal of Network and Computer Applications* (JNCA), Elsevier.