Supplementary Information

Unmasking Hidden Waves: A Proof-of-Concept Transformer-Based Electrocardiogram Signal Decomposition for Complex Arrhythmias

Michał Szafarczyk, B.Eng., Peter M. van Dam, PhD

Table of Contents

- Supplementary Methods S1: Detailed ESD-former Architecture ..... Page 2
- Supplementary Methods S2: Comprehensive Synthetic Generation Pipeline ...Page 3
- Supplementary Methods S3: Mathematical Formulation of ShapeAwareLoss ..Page 4

Supplementary Methods S1: Detailed ESD-former Architecture

Encoder

The encoder is responsible for extracting local morphological features from the 1D ECG signal. It consists of two downsampling levels utilizing 1D convolutional layers with a relatively large kernel size of 7 and padding of 3. This specific kernel size was chosen to provide a wider receptive field early in the network, which is crucial for capturing prolonged, low-frequency components like the T wave. Each convolutional layer is followed by Batch Normalization and a ReLU activation function. Between the spatial levels, downsampling is achieved using strided convolutions (kernel size of 3, stride of 2), progressively reducing the temporal sequence length by a total factor of 4 while increasing the feature depth to `d_model` = 128. High-resolution feature maps from each encoder level are preserved for skip connections.

Transformer Bottleneck

While convolutional layers excel at local morphology, understanding arrhythmias requires global context. To achieve this, the deepest feature map (now at one-quarter of the original sequence length) is passed into a Transformer Encoder bottleneck. By operating on a temporally reduced sequence, the model avoids the steep computational cost of attention mechanisms while maintaining global receptive capability. A learnable positional encoding is added to the sequence to retain temporal ordering. The bottleneck consists of 4 Transformer encoder layers, each utilizing 8 attention heads and a feedforward dimension of 512, with a dropout rate of 0.1.

Decoder and Multi-Head Output

The decoder rebuilds the temporal resolution and translates the high-level contextual features back into morphological waveform predictions. Upsampling is performed via 1D transposed convolutions (stride of 2, kernel size of 4). To recover the fine-grained localization boundaries lost during the encoding phase, the upsampled features are concatenated with the corresponding high-resolution maps from the encoder via skip connections. The final layer branches into three independent regression heads. Each head utilizes a 1x1 convolution to map the final 32-channel feature map to a single continuous output channel.

Supplementary Methods S2: Comprehensive Synthetic Generation Pipeline

Template Extraction and Intra-Patient Consistency

To prevent the model from learning artificial boundaries, we strictly enforced intra-patient consistency. For each synthetic sequence, a set of isolated physiological templates was randomly sampled from a single patient's normal sinus rhythm record. A dynamic trimming

algorithm removed isoelectric buffers from the wave margins (amplitudes below 5-10% of the maximum peak).

Modeling Atrioventricular Blocks and Tachycardias

To simulate 1AVB, we prolonged the PR interval while keeping the 1:1 ratio intact. For 3AVB, the atrial rate and ventricular rate were generated as entirely independent, asynchronous physiological oscillators, allowing P waves to naturally drift across the cardiac cycle. Simulating reentrant tachycardias required advanced adaptations. To mirror natural repolarization changes, extracted T wave templates were temporally compressed by 20–45%. Deep ST-segment depressions (-0.15 to -1.0 mV) were mathematically fused with the T waves. For AVNRT, retrograde P waves were generated and partially superimposed onto the terminal portion of the QRS complex. For AVRT, the P waves were inverted, scaled, and injected deep into the initial downslope of the ST-segment. To prevent overfitting, real clinical noise (scaled up to 25% of the P wave amplitude) was superimposed.

Supplementary Methods S3: Mathematical Formulation of ShapeAwareLoss

The structural component of the objective function is governed by the Cosine Distance (L_{Cos}), rendering it invariant to absolute amplitude scaling:

$$L_{Cos} = 1 - \frac{\sum_{i=1}^{N} y_i \hat{y}_i}{\sqrt{\sum_{i=1}^{N} y_i^2} \sqrt{\sum_{i=1}^{N} \hat{y}_i^2}} \tag{1}$$

The magnitude component is driven by the standard Mean Squared Error (L_{MSE}):

$$L_{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2}$$

where y is the ground truth signal, $\hat{y}$ is the model's prediction, and N is the sequence length. All models were trained using a batch size of 32, with a maximum learning rate set to 0.001 modulated by a OneCycleLR scheduler.