WonHEE: A Multimodal Transformer for ECG Time Series

FULTON SCIENCE ACADEMY — Private School —

Jinseop Song **Daniel Kang**
Jooheon Ryu **Jiwon Kim**

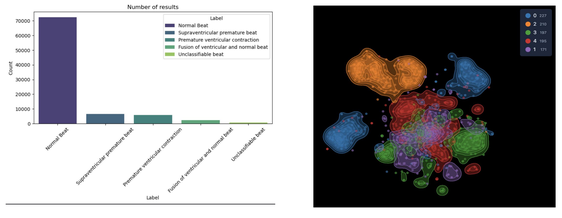
Overview

The electrocardiogram (ECG), an electrical representation of cardiac activity, is widely employed in the diagnosis of heart attacks and related cardiovascular conditions. While prior studies using machine learning for ECG analysis have largely focused on single tasks like heartbeat classification, there is a growing need for more unified approaches for handling dynamic situations. Recent progress in time-series research has introduced frameworks that evaluate models across multiple tasks, including prediction, imputation, and labeling. In this work, we present ***WonHEE***, a new novel multi-transformer block architecture capable of simultaneously performing all three tasks on ECG datasets with high precision. This framework advances ECG time-series analysis by integrating multi-task learning within a single, highly effective model. Our code can be found in: https://github.com/axion66/wonhee

Background

For many decades, machine learning analysis of ECG data has been dominated by single-task models for the most part for heartbeat classification [6]. While working well for discrete tasks, this methodology leaves much to be desired for decades-long clinical use wherein data are potentially dynamic, incomplete, and noisy. Even as multi-task learning has become an influential tool for general time-series modeling, its use for ECG analysis has not yet proceeded far. One clear gap for the field: the unavailability of truly multimodal models able to synthesize disparate data sources beyond the ECG itself. By far the most common algorithms for identifying many cardiac abnormalities are multiclass convolutional or recurrent neural network ones that return a probability for every disease. This method has the tendency of diluting diagnostic confidence and returning questionable interpretations wherein multiple arrhythmias are simultaneously present. In addition, there exists at present no widely accepted AI model capable of identifying rare but lethal heart conditions from standard ECG records, leaving patients without access to advanced cardiac imaging or monitoring-specific knowledge with a diagnostic gap. In their present form, conventional clinical procedures wherein cardiac diagnosis comes from Holter monitoring, echocardiography, and stress tests all necessitate personnel with specific expertise for their proper reading. These deficiencies underscore the serious need for inclusive and ubiquitous AI-based algorithms able to discriminate with accuracy many heart disease types from standard ECG signals without any dependence on expert-specific knowledge.

Dataset



Dataset Analysis: ***Left***: Distribution of ECG data categories from the MIT-BIH Arrhythmia Database before preprocessing. ***Right***: It displays a t-SNE plot of the dataset after being balanced. The labels are as follows: 0 (Normal Beat), 1 (Supraventricular Premature Beat), 2 (Premature Ventricular Contraction), 3 (Fusion of Ventricular and Normal Beat), 4 (Unclassifiable Beat), and 5 (Paced Beat).

Dataset Preprocess

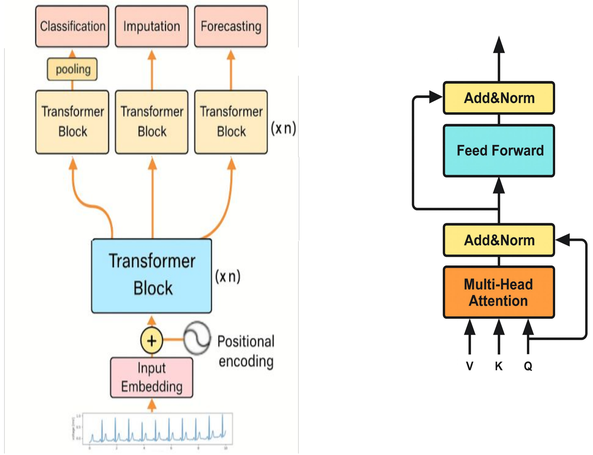
Our dataset comes from the MIT-BIH Arrhythmia Database (PhysioNet) [7]. The raw ECG recordings are sampled at 360 Hz. For quick training and with the aim of minimizing computational complexity without losing performance, we re-sampled the data at 128 Hz utilizing a polyphase filter. The rationale for this choice comes from an earlier observation [5] that performance of deep learning models remains strong with ECG sampling rates higher than 100 Hz, which suffices for extracting the salient features of the QRS complex. At **a sampling rate of 128 Hz**, we get 5-second long sequences of the ECG signal with 640 data points per sample (128 Hz×5 seconds=640).

First, we split the dataset into training and test sets with an **80:20 ratio**. The original dataset has an extreme class imbalance; e.g., Normal beats constitute 89.2% of all beats, whereas the most common class of arrhythmia has only 4.5%. To counter this, we use a weighted random sampler **[torch.utils.data.WeightedRandomSampler]** for the training set. This method assigns weights inversely proportional to class frequency so that the minority classes are represented correctly at each training batch. This prevents the model from getting biased toward the dominant classes and promotes better generalization at the less-common arrhythmia types.

t-SNE

To test the performance of the feature extraction and preprocessing workflow, we employed t-distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction of the high-dimensional ECG features and projection onto a two-dimensional space. This algorithm is especially suitable for data with many features and moderate to large sample sizes for data visualization and preserving local structure and highlighting inter-class separations. As Figure 2 indicates, the t-SNE plot presents clear clusters for varying arrhythmia classes. The clear presence of groupings indicates strong qualitative support for the raw ECG signals having been properly transformed with the preprocessing steps for representation as a feature space amenable to classification and hence for the development of a high-performing classification model.

Architecture Overview



Overview of the WonHEE pipeline. ***Left***: Full depiction of the WonHEE architecture. For simplicity, the [CLS] token used after the pooling layer in the classification head is omitted. The Transformer Blocks shown in blue and yellow correspond to the same underlying architecture. ***Right***: Detailed view of a Transformer Block within the model, adapted from the encoder block of the standard Transformer architecture [4].

The proposed model leverages a Transformer-based architecture tailored for ECG analysis, incorporating advanced components to enhance performance and efficiency. Multi-Head Attention (MHA) allows the model to attend to different parts of the input sequence simultaneously, capturing various temporal relationships. Formally, MHA can be expressed as: $\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$

and the scaled dot-product attention is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Here, Q, K, and V are the query, key, and value matrices, and d_k is the dimension of the key vectors.

The Feed-Forward Network (FFN) in each block uses the SwiGLU activation function, which has been shown to improve expressivity. SwiGLU is defined as: $\text{SwiGLU}(x) = \text{Swish}(xW_1 + b_1) \odot (xW_2 + b_2)$

where $\text{Swish}(x) = x \cdot \sigma(\beta x)$, σ is the sigmoid function, β is a learnable parameter, and $\odot$ denotes element-wise multiplication.

For efficient training, each block applies Root Mean Square Normalization (RMSNorm) instead of LayerNorm shown in original "Attention is All you Need" paper, which normalizes activations based on their root mean square:

$$\text{RMSNorm}(x) = \frac{x}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2 + \epsilon}}\gamma$$

where N is the number of elements, ϵ is a small constant for numerical stability, and γ is a learnable scaling parameter.

Model Architecture and Components

The overall architecture consists of an encoder with multiple Transformer blocks that process ECG sequences to capture complex temporal dependencies. Task-specific Transformer heads handle classification, forecasting, and imputation. Missing values are embedded with a learnable vector, and positional information is incorporated using sinusoidal embeddings. The implementation explored two versions: a standard one with a Feed-Forward Network (FFN) and another with a Soft Mixture-of-Experts (Soft-MoE) [3] layer, though no meaningful performance difference was found between them. [cls] token was added at the beginning of the sequence during the classification step, then later used for classification. It was first introduced in BERT model in 2018.

Hyperparameters and Training Details

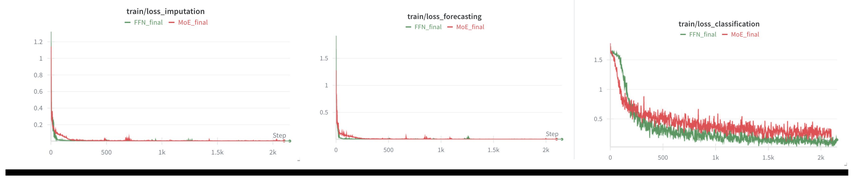
The model was implemented in PyTorch with a hidden dimension of 128, 4 attention heads, 4 encoder layers, and a sequence length of 640 (where 128 sequences = 1 second of ECG data). The Soft-MoE used 5 experts. For each task-specific transformer block, each consists of two Transformer Blocks. Training was done using the Adam optimizer with weight decay, a learning rate of 3e-4 for 3 epochs, and a batch size of 128. Initially, *PC-Grad* [2] was used for multi-task learning, but a simple sum of individual task losses proved more effective for most case by ablation study done by Kurin et al. [1].

Multi-Task Learning Approach

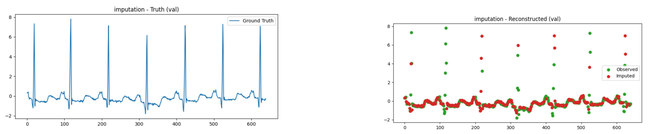
A multi-task learning approach was employed, where the model was trained on those three modes: imputation, forecasting, or classification within each batch. For imputation and forecasting, the model's objective was to reconstruct masked or future portions of the 640-point time series. The number of points to be forecasted or imputed ranged from 200 to 540. Imputation points were randomly selected, while forecasting involved predicting a continuous block from the beginning. Cross-Entropy Loss was used for classification, and Mean Squared Error (MSE) loss was used for the other two tasks. Input data was standardized across entire training dataset, and validation was performed after the first 50 steps of each epoch.

Results

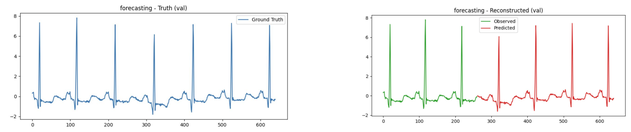
Train Losses. ***Left***: Imputation, ***Center***: Forecasting, ***Right***: Classification
[red indicates MoE version, while Green indicates Non-MoE model]



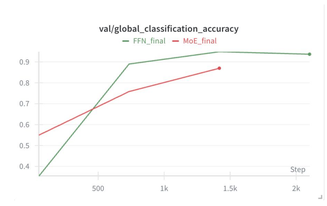
ECG Reconstruction (Imputation) on a Test Sample. *Left*: Ground Truth, ***Right***: Green dots indicate observed datapoints and red dots indicate interpolated values. Non-MoE model was used.



ECG Forecasting on a Test Sample. *Left*: Ground Truth, ***Right***: Green dots indicate observed datapoints and red dots indicate predicted values Non-MoE Model was used.



ECG Classification on Entire Test Samples. The Non-MoE (FFN_final) model achieves ***94.8%*** accuracy. Green line represents Non-MoE model and Red line represents MoE model.



Conclusion

We proposed ***WonHEE***, a novel multi-transformer block architecture that can perform labeling, imputation, and forecasting on ECG datasets simultaneously. ***WonHEE*** represents a significant step beyond traditional single-task models, opening a new gate toward a multimodal approach for ECG analysis. Moreover, the ability to forecast and impute ECG data can be beneficial for handling noisy or incomplete recordings from wearable devices that's beyond mere prediction of ECG types. We truly hope our research will contribute to the future development of ECG diagnostics!

* All training and evaluation were conducted on a single NVIDIA RTX 5090.

Sources

[1] Kurin et al., In Defense of the Unitary Scalarization for Deep Multi-Task Learning, NeurIPS 2022.
[2] Yu et al., Gradient Surgery for Multi-Task Learning, NeurIPS 2020.
[3] Puigcerver et al., From Sparse to Soft Mixtures of Experts, ICLR 2024.
[4] Vaswani et al., Attention Is All You Need, NeurIPS 2017.
[5] Kwon, Jeong, Kim et al., Electrocardiogram Sampling Frequency Range Acceptable for Heart Rate Variability Analysis, Healthcare Informatics Research, 2018.
[6] Aziz, Ahmed & Alouini, ECG-based Machine-Learning Algorithms for Heartbeat Classification, Scientific Reports, 2021.
[7] Moody, G.B. & Mark, R.G. The impact of the MIT-BIH Arrhythmia Database, IEEE Eng in Med and Biol, 2001.