

HIERARCHICAL TRANSFORMER FOR ELECTROCARDIOGRAM DIAGNOSIS

Xiaoya Tang¹ Jake Berquist^{1,2,4} Benjamin A. Steinberg⁵ Tolga Tasdizen^{1,3}

¹ Scientific Computing and Imaging Institute, University of Utah, SLC, UT, USA

² Department of Biomedical Engineering, University of Utah, SLC, UT, USA

³ Electrical and Computer Engineering, University of Utah, SLC, UT, USA

⁴ Nora Eccles Harrison Cardiovascular Research and Training Institute, University of Utah, SLC, UT, USA

⁵ University of Colorado Anschutz Medical Campus, Denver, CO, USA

ABSTRACT

Transformers, originally prominent in NLP and computer vision, are now being adapted for ECG signal analysis. This paper introduces a novel hierarchical transformer architecture that segments the model into multiple stages by assessing the spatial size of the embeddings, thus eliminating the need for additional downsampling strategies or complex attention designs. A classification token aggregates information across feature scales, facilitating interactions between different stages of the transformer. By utilizing depth-wise convolutions in a six-layer convolutional encoder, our approach preserves the relationships between different ECG leads. Moreover, an attention gate mechanism learns associations among the leads prior to classification. This model adapts flexibly to various embedding networks and input sizes while enhancing the interpretability of transformers in ECG signal analysis.

Index Terms— Hierarchical Transformer, Multi-scale, ECG Classification, Depth-wise Convolution, Attention

1. INTRODUCTION

Since Dosovitskiy et al. [1] adapted the Transformer model from natural language processing (NLP) to the computer vision (CV) domain, the purely Transformer-based Encoder architecture has demonstrated remarkable potential across various vision benchmarks. When pre-trained on extensive datasets such as ImageNet and JFT-300M, Transformers have outperformed state-of-the-art Convolutional Neural Networks (CNNs), including ResNets, in smaller image recognition tasks. The Transformer Encoder employs a patch tokenization process that converts the input into a sequence of equal-length feature embeddings, termed tokens. The tokenization is followed by a series of Multi-Head Self-Attention (MSA) layers, which project these tokens and learn the interrelationships among them. This architecture excels at capturing global dependencies within the input. Despite its high model capability, Transformer-based models often lack inductive biases inherent in CNNs—such as translation equivariance, locality, and hierarchical representations—which are vital for

signal processing tasks. To address this deficiency without the necessity of large-scale data, researchers in NLP and CV domains have explored various approaches to integrate necessary inductive biases into Transformers, including convolutional incorporations, novel windowed or local attentions, hierarchical structuring, and auxiliary self-supervised tasks.

Cardiovascular diseases remain a significant health threat worldwide [2, 3]. Early diagnosis of cardiac disorders is crucial, enabling timely interventions that can significantly improve patient outcomes [2]. Traditional manual arrhythmia detection by clinicians is labor-intensive and prone to errors. Advances in computer-aided diagnosis have aimed to enhance the accuracy of ECG interpretation and reduce associated costs [4]. Recently, deep learning has been applied to develop more effective computer-aided diagnosis systems. These models, capable of analyzing comprehensive data sets, detect complex patterns essential for diagnosing heart conditions, with minimal or no reliance on predefined features or manual intervention. Previous methodologies applied to ECG tasks often involved elaborate convolutional and recursive structures [5]. Given the sequential nature of ECG signals, the application of Transformers has been promising due to their superior capacity to learn dependencies across sequences [2]. Recent implementations of Transformer Encoders in cardiac abnormality classification [2], arrhythmia detection [3], phonocardiography (PCG)-based valvular heart diseases (VHD) detection [6], and constrained loss models [7] have showcased their advantages in simplicity and training efficiency over CNNs and recurrent neural networks (RNNs). A few recent studies have experimented with hierarchical transformers to better leverage the capabilities of Transformers on ECG data while addressing their inductive bias limitations. For instance, Li et al. [8] utilized a shifted-window-based Transformer for heartbeat classification. Deng et al. [9] constructed a CNN encoder and decoder, utilizing a transformer to bridge the gap between them for left ventricle segmentation tasks. Wahid et al. [10] integrated ResNet, ViT, and channel attention mechanisms to introduce inductive biases for myocardial infarction detection. Similarly, Dong et al. [4]

employed depthwise separable convolutions based on an Inception module and a deformable vision transformer for the classification of arrhythmias. However, the application of transformers to ECG signals is still in its early stages and needs further exploration.

Building on these insights, this paper proposes a novel hierarchical transformer model to advance the field of ECG classification. Our contributions are as follows:

- We apply a simple yet effective six-layer convolutional encoder to extract features from raw ECG data. The depthwise convolution mechanism is utilized to maintain the information across multiple leads, enabling further exploration at the model’s end.
- Through the convolutional encoder, we obtain hierarchical representations that bring necessary locality and multi-scale information to the transformer. We propose a unique method that employs a CLS token to aggregate information post-attention and transmit it to subsequent transformer stages without additional pooling or downsampling strategies. This approach uses vanilla multi-head attention without the need for specialized window-based or local attention mechanisms, avoiding the need for meticulously handcrafted designs.
- We integrate a lightweight attention-gated module comprising three linear layers to learn associations between different ECG leads. This module enhances performance when combined with the depth-wise encoder.

The paper demonstrates the efficacy of our model in harnessing multi-scale and lead-specific information through comprehensive experiments. The innovations in multi-scale processing and lead dependencies signify an important departure from previous transformer-based methods in ECG analysis. Furthermore, the depiction of attention maps elucidates how the hierarchical transformer design enhances interpretability compared to conventional architectures, effectively identifying critical patterns in ECG signals. This work provides deeper insights into ECG analysis, bridging the gap between advanced computational techniques and clinical utility.

2. METHODOLOGY

2.1. Depthwise Convolutional Feature Encoder

The framework of our model is illustrated in fig. 1. We employ a six-layer simple encoder to extract useful features from the input ECG signal. Each layer features varying kernel sizes and strides to facilitate progressive downsampling, adopted from [2]. The downsampling rates are adjustable based on needs and can be configured according to user preferences. This simple encoder creates a multi-scale feature representation for the ECG input, which has been proven beneficial in conjunction with transformers for vision tasks [11]. To preserve the critical yet unknown associations between different

ECG leads, we employ depth-wise convolutions in all layers of our encoder. Depthwise convolutions [12] employ a distinct filter for each input channel, capturing spatial relationships without cross-channel interactions. In the context of multi-lead ECG signals, these convolutions are applied individually to each lead. Subsequently, the resulting feature maps from each lead can be transformed separately onto a new space. [4]. Although depth-wise convolutions and pyramid structures have been previously noted [4], they haven’t fully mined the inter-lead information nor optimized the use of hierarchical features. Our experiments indicate that depth-wise convolutions enhance our model’s performance by not mixing valuable hidden information between leads.

2.2. Three-stage transformer

According to previous researches, the effective receptive field of ViT shifts from local to global as it progresses through the layers. To leverage this characteristic and bring multi-scale inductive biases, we here propose a novel yet straightforward hierarchical transformer. This design utilizes the standard transformer encoder along with Multi-Head Self-Attention (MSA) mechanisms. We structure the transformer into three stages, each containing a stack of MSA layers, with the division of layers tailored to specific needs. Our approach involves feeding hierarchical feature embeddings (called contextual tokens in fig. 1) into three stages, derived from different layers of our convolutional encoder using three distinct downsampling rates from the input ECG segment. Each stage begins by integrating embeddings with a learnable CLS token, commonly used in classification tasks. After each stage, the CLS token is extracted and concatenated with a new sequence of embeddings at a larger downsampling rate, then passed into the next transformer stage. This progressive feeding of downsampled features compels the model to transition its focus from detailed to more abstract, global patterns. Utilizing the CLS token allows us to efficiently aggregate and transfer multi-scale information to the final classification layer.

2.3. Attention-Gated Module

Given an output from three-stage transformer x with dimensions $x \in \mathbb{R}^{B \times C \times S}$, where B represents the batch size, C the number of channels, and S the sequence length. The information for each lead remains distinct and uncombined. Thus we utilize an attention-gated module to model dependencies between leads, inspired by [13]. This module comprises three linear layers designed to uncover latent dependencies between channels, which correspond to associations between ECG leads in this context. The attention score a is computed through an element-wise multiplication of the query and key vectors, resulting in $a \in \mathbb{R}^{B \times C \times S}$, as shown in eq. (1). $W_q \in \mathbb{R}^{S \times S}$, $b_q \in \mathbb{R}^S$, $W_k \in \mathbb{R}^{S \times S}$, and $b_k \in \mathbb{R}^S$

represent the weights and biases of the linear layer for learning query and key, with σ denotes the Sigmoid function.

$$\begin{aligned} q &= \tanh(W_q x + b_q) \\ k &= \sigma(W_k x + b_k) \\ a &= q \odot k \end{aligned} \quad (1)$$

A linear project is applied to the attention scores resulting in $a' \in \mathbb{R}^{B \times C \times N}$, where N is the number of classes. The raw attentions are then normalized by a softmax and multiplied with the output from the three-stage transformer, yielding $v \in \mathbb{R}^{B \times N \times S}$, shown in eq. (2). These operations are analogous to the MSA mechanism. Finally, a separate classifier for each class is applied across the sequences, where v_i denotes the segment of v corresponding to the i -th class.

$$\begin{aligned} a' &= \text{Projection}(a) \\ a'' &= \text{softmax}(\text{transpose}(a', (0, 2, 1))) \\ v &= a'' @ x \\ \text{logits}_i &= W_i v_i + b_i \quad \text{for each } i \in \{1, \dots, N\} \end{aligned} \quad (2)$$

3. RESULTS AND ANALYSIS

3.1. Data and Evaluation Metrics

We utilize the public training data from the 2020 PhysioNet/CinC Challenge [14] and KCL data from our group. The public dataset comprises 43,101 recordings, and we adopt the 10-fold split used by the winner model 'Prna' [2]. This setup involves a multi-label classification task related to 24 diagnoses. Following the preprocessing steps of 'Prna', we resample all recordings to 500Hz, apply an FIR band-pass filter, and perform normalization. We also randomly crop multiple fixed-length ECG segments of $T = 15$ seconds from the input, adding padding when necessary for segments shorter than 15s. We also leveraged the wide features that they used. For evaluation metrics, we report macro F_β , G_β , geometric mean(GM) combining precision and recall and the challenge score defined by the challenge organizers [14], detailed in eq. (3). The score S generalizes standard accuracy by fully crediting correct diagnoses and penalizing incorrect ones based on the similarity between arrhythmia types. Here a_{ij} represents an entry in the confusion matrix corresponding to the number of recordings classified as class c_i but actually belonging to class c_j , with different weights w_{ij} assigned based on the similarity of classes c_i, c_j :

$$\begin{aligned} F_\beta &= (1 + \beta^2) \cdot \frac{TP}{(1 + \beta^2) \cdot TP + FP + \beta^2 FN} \\ G_\beta &= \frac{TP}{TP + FP + \beta FN} \\ GM &= \sqrt{F_\beta \cdot G_\beta}, \beta = 2 \\ S &= \sum_{ij} w_{ij} a_{ij} \end{aligned} \quad (3)$$

For the KCL potassium classification, all recordings maintain a uniform sampling rate of 500Hz. After applying normalization, we randomly crop these to fixed segments of $T = 5s$. The dataset includes 54,419 recordings for training and 6,245 for testing. We report the macro-averaged area under the receiver operating characteristic curve (AUC) on test data.

3.2. Experiments

Our model achieved outstanding results in the 2020 PhysioNet/CinC Challenge dataset, surpassing other commonly used architectures and even exceeding the performance of previous challenge winners, Prna [2] and Res-SENet [15], across all evaluated metrics. Results are shown in table 1. Notably, our model can compete with semi-supervised methods, such as those described by [16], which employed supervised contrastive learning. An important observation from these results is the enhanced performance of the standard ViT model, Prna, upon integration of a CLS_token, which validates the CLS_token's significance in classification tasks. Additionally, our model demonstrates that both with and without the Attention_gated module, it maintains competitive performance, demonstrating the efficiency of our hierarchical transformer.

To further validate the efficiency and generalizability of our approach, we conducted additional tests on the KCL binary classification task, comparing our model against other prominent architectures. Our model showcased the highest AUC, outperforming models such as SpatialTemporalNet and ViT (Prna), shown in table 2. These results confirm the robustness and adaptability of our model, effectively identifying complex patterns essential for precise ECG classification.

3.3. Interpretability

The multi-head self-attention(MSA) allows each head to learn distinct attention patterns across the time sequence. These patterns can be analogized to distinct attentions across different ECG leads, facilitated by our depthwise encoder. During the evaluation phase of the KCL classification, we randomly selected an abnormal sample, with the attention map at the final stage shown in fig. 2 and fig. 3. We qualitatively assessed the attention map by examining which areas of the ECG signals garnered the highest attention in this unhealthy case. Notably, our model exhibited heightened attention to clinically significant features such as the QRS complex, S-T segment, and T-wave, which are recognized as clinical indicators of changes in serum potassium levels. The proposed approach also underscores how the model's attention shifts across different stages. While we do not leverage lead attention here, the attn_gated module after MSA allowed us to discern dependencies among multiple leads. This capability further provides valuable insights into how the model relies on different leads, enhancing our understanding of deep learning models for ECG diagnosis.

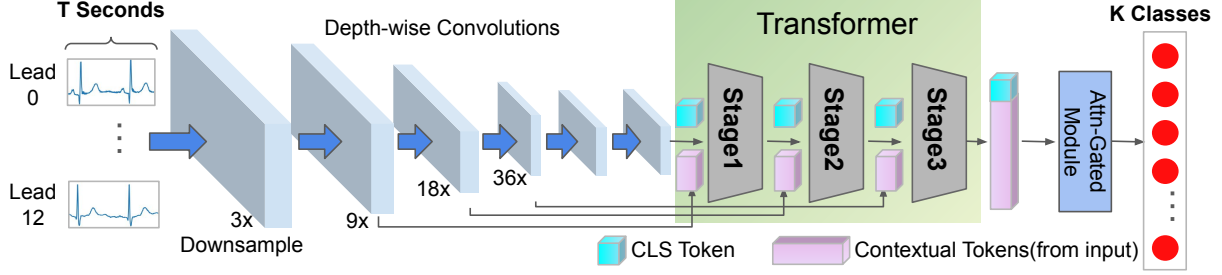


Fig. 1: Framework: Six-layer encoder, three-stage transformer, and attention-gated module for classification(left to right).

Table 1: Performance comparison of various models on multi-label classifications. All models were evaluated using 10-fold validation. Results for the first six models are sourced from [8]. Following [2], for computational efficiency, results for all other models are averaged over three of the ten folds.

Model	Fbeta_measure	Gbeta_measure	Geometric_mean	Challenge_metric	Params.
LSTM	0.4323 ± 0.0024	0.2742 ± 0.0052	0.3443 ± —	0.4372 ± 0.0073	-
CNN	0.4519 ± 0.0070	0.2862 ± 0.0083	0.3596 ± —	0.4542 ± 0.0076	-
ResNet	0.5088 ± 0.0021	0.3278 ± 0.0088	0.4084 ± —	0.5158 ± 0.0041	-
ViT	0.3263 ± 0.0054	0.1970 ± 0.0037	0.2535 ± —	0.3197 ± 0.0078	-
Swin Transformer	0.4812 ± 0.0042	0.3045 ± 0.0020	0.3828 ± —	0.4811 ± 0.0068	-
BaT [8]	0.5011 ± 0.0034	0.3125 ± 0.0036	0.3957 ± —	0.4958 ± 0.0041	-
Res-SENet [15]	0.5607 ± 0.0073	0.3264 ± 0.0096	0.4278 ± 0.0090	0.5939 ± 0.0018	8.84M
SpatialTemporalNet	0.4296 ± 0.0121	0.2403 ± 0.0072	0.3212 ± 0.0050	0.4322 ± 0.0424	4.52M
Prna [2]	0.4975 ± 0.0257	0.2679 ± 0.0187	0.3650 ± 0.0219	0.5463 ± 0.0176	13.64M
Prna + CLS_Token	0.5211 ± 0.0051	0.2926 ± 0.0072	0.3905 ± 0.0068	0.5732 ± 0.0121	13.64M
Ours-No Attn_gated	0.5672 ± 0.0034	0.3296 ± 0.0110	0.4309 ± 0.0076	0.6174 ± 0.0065	16.62M
Ours	0.5778 ± 0.0044	0.3407 ± 0.0074	0.4436 ± 0.0032	0.5980 ± 0.0051	16.78M

Table 2: Performance comparison of typical models for KCL.

Model	Test AUC	Params.
Prna(ViT)	0.8126 ± 0.0088	13.63M
Swin Transformer-Tiny	0.7954 ± 0.0009	47.47M
SpatialTemporalNet	0.8218 ± 0.0029	4.41M
Res-SENet	0.8203 ± 0.0007	8.82M
Ours	0.8232 ± 0.0027	10.94M

Fig. 2: Attentions in an abnormal case (high potassium) for leads 1 to 4, illustrating final stage attentions.

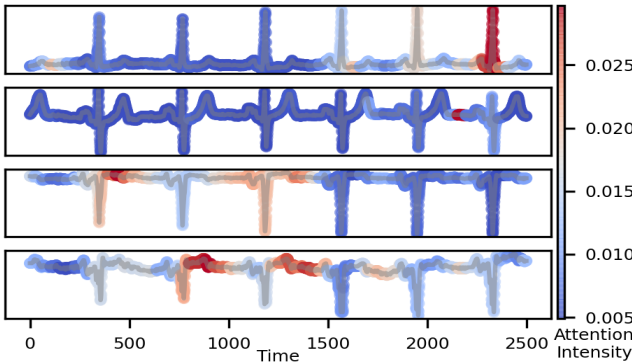
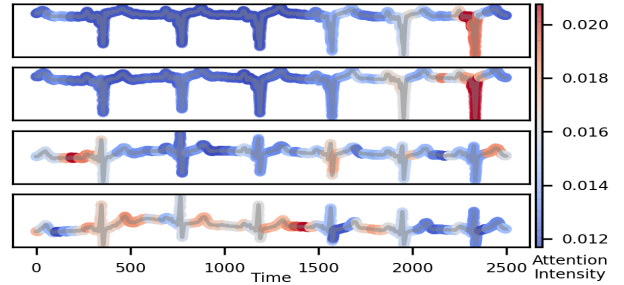


Fig. 3: Attentions in an abnormal case (high potassium) for leads 5 to 8, illustrating final stage attentions.



4. CONCLUSION

We proposed a hierarchical transformer for ECG diagnosis that includes a depthwise encoder, a three-stage transformer, and an attention-gated module. The experimental results demonstrate our model’s efficiency in handling varied and challenging ECG diagnostic tasks. The attention maps illustrate that our model focuses on clinically significant features relevant to the diagnostic task. Additionally, our model is capable of learning dependencies between multiple leads, enhancing interpretability compared to previous models.

5. COMPLIANCE WITH ETHICAL STANDARDS

This study was conducted under IRB 00140935 12-Lead ECG Machine Learning Analysis to Identify Unique Non-invasive Clinical Indicators of Disease at the University of Utah. All studies and data acquisition were subject to and complied with the University of Utah institutional review board review and requirements.

6. ACKNOWLEDGMENTS

This work was supported by Grant NIH R21HL172288. We thank Dr. Man Minh Ho for his insightful suggestions.

7. REFERENCES

- [1] Alexey Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [2] Annamalai Natarajan, Yale Chang, Sara Mariani, Asif Rahman, Gregory Boverman, Shruti Vij, and Jonathan Rubin, "A wide and deep transformer neural network for 12-lead ecg classification," in *2020 Computing in Cardiology*. IEEE, 2020, pp. 1–4.
- [3] Rui Hu, Jie Chen, and Li Zhou, "A transformer-based deep neural network for arrhythmia detection using continuous ecg signals," *Computers in Biology and Medicine*, vol. 144, pp. 105325, 2022.
- [4] Yanfang Dong, Miao Zhang, Lishen Qiu, Lirong Wang, and Yong Yu, "An arrhythmia classification model based on vision transformer with deformable attention," *Micromachines*, vol. 14, no. 6, pp. 1155, 2023.
- [5] Genshen Yan, Shen Liang, Yanchun Zhang, and Fan Liu, "Fusing transformer model with temporal features for ecg heartbeat classification," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 898–905.
- [6] Sonain Jamil and Arunabha M Roy, "An efficient and robust phonocardiography (pcg)-based valvular heart diseases (vhd) detection framework using vision transformer (vit)," *Computers in Biology and Medicine*, vol. 158, pp. 106734, 2023.
- [7] Chao Che, Peiliang Zhang, Min Zhu, Yue Qu, and Bo Jin, "Constrained transformer network for ecg signal processing and arrhythmia classification," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, pp. 184, 2021.
- [8] Xiaoyu Li, Chen Li, Yuhua Wei, Yuyao Sun, Jishang Wei, Xiang Li, and Buyue Qian, "Bat: Beat-aligned transformer for electrocardiogram classification," in *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 320–329.
- [9] Kaizhong Deng, Yanda Meng, Dongxu Gao, Joshua Bridge, Yaochun Shen, Gregory Lip, Yitian Zhao, and Yalin Zheng, "Transbridge: A lightweight transformer for left ventricle segmentation in echocardiography," in *Simplifying Medical Ultrasound: Second International Workshop, ASMUS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 2*. Springer, 2021, pp. 63–72.
- [10] Junaid Abdul Wahid, Xu Mingliang, Muhammad Ayoub, Shabir Hussain, Lifeng Li, and Lei Shi, "A hybrid resnet-vit approach to bridge the global and local features for myocardial infarction detection," *Scientific Reports*, vol. 14, no. 1, pp. 4359, 2024.
- [11] Xiaoya Tang, Bodong Zhang, Beatrice S Knudsen, and Tolga Tasdizen, "Duoformer: Leveraging hierarchical visual representations by local and global attention," *arXiv preprint arXiv:2407.13920*, 2024.
- [12] Andrew G Howard, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [13] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood, "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16144–16155.
- [14] Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, et al., "Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020," *Physiological measurement*, vol. 41, no. 12, pp. 124003, 2020.
- [15] Zhibin Zhao, Hui Fang, Samuel D Relton, Ruqiang Yan, Yuhong Liu, Zhijing Li, Jing Qin, and David C Wong, "Adaptive lead weighted resnet trained with different duration signals for classifying 12-lead ecgs," in *2020 Computing in Cardiology*. IEEE, 2020, pp. 1–4.
- [16] Duc Le, Sang Truong, Patel Brijesh, Donald A Adjero, and Ngan Le, "scl-st: Supervised contrastive learning with semantic transformations for multiple lead ecg arrhythmia classification," *IEEE journal of biomedical and health informatics*, vol. 27, no. 6, pp. 2818–2828, 2023.