A CNN-based Local-Global Self-Attention via Averaged Window Embeddings for Hierarchical ECG Analysis

Anonymous Author(s)
Affiliation
Address
email

Abstract

Electrocardiograms (ECGs) are multivariate time series where clinically relevant patterns span both local waveform morphology and long-range rhythm structure. We introduce LGA-ECG, a hierarchical Transformer architecture that integrates convolutional inductive biases into self-attention. Queries are extracted from overlapping local windows to retain morphological fidelity, while keys and values are globally derived to enable full temporal context. This design eliminates the need for explicit positional encodings by leveraging convolutional locality. On the CODE-TEST benchmark, LGA-ECG achieves a macro F1-score of 0.885, recall of 0.872, and precision of 0.907, outperforming CNN and Transformer baselines. Ablation studies confirm the effectiveness of combining local queries with global key-value pairs.[1]

1 Introduction

Time series data are central to modern healthcare. From electronic health records and continuous monitoring devices to population-level trends, temporal information drives medical decision-making across scales. Among these signals, electrocardiograms (ECGs) stand out as one of the most structured and widely studied physiological time series. As one-dimensional signals capturing the heart's electrical activity over time, ECGs encode both fine-grained morphological patterns (e.g., P, QRS, and T waves) and long-range temporal dynamics (e.g., rhythm and rate). This dual structure makes ECGs a powerful yet challenging benchmark for time series modeling, requiring architectures that can operate across multiple temporal resolutions.

Cardiovascular diseases (CVDs) remain the leading cause of death worldwide, accounting for 17.9 million deaths in 2019 (32% of all global deaths), according to the World Health Organization (WHO) [19]. Electrocardiograms (ECGs), being non-invasive and widely available, are essential in diagnosing and monitoring heart conditions, and their role has expanded with the growth of digital health technologies [10]. In this context, artificial intelligence has become a powerful tool for automating ECG analysis, supporting clinical decision-making, reducing telemedicine backlogs, and enabling tasks such as arrhythmia classification [14, 15, 3], atrial fibrillation detection [18], age estimation [9], and wave segmentation [5].

Deep learning, particularly convolutional neural networks (CNNs), has driven these advances by autonomously extracting morphological and temporal ECG features [14, 15]. CNNs leverage inductive biases such as spatial locality and translation equivariance, enabling hierarchical modeling from localized waveform details to global rhythm patterns. Hybrid models combining CNNs and recurrent

[1]To preserve anonymity during review, we will release the code and pretrained models upon acceptance.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

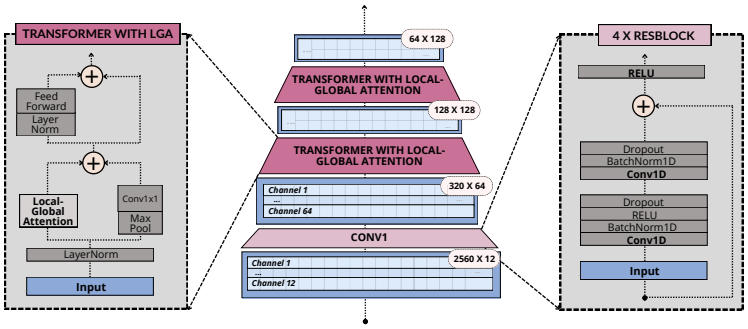

Figure 1: Overall architecture of the proposed LGA-ECG network.

layers have also been proposed to jointly capture morphology and temporal dependencies [18]. More recently, transformer architectures have been introduced to ECG analysis, motivated by their success in sequence modeling. CNN-transformer hybrids [7, 4] combine convolutional front-ends for local feature extraction with global self-attention layers, while beat-aligned transformers (BaT) [8] leverage local attention on segmented beats, and hierarchical CNN-transformer pipelines further improve multi-scale modeling [2].

These efforts reflect broader trends in time series modeling. As researchers seek general-purpose architectures for diverse health signals, there is increasing emphasis on combining inductive biases for local structure with mechanisms capable of modeling long-range dependencies. General advances in local attention mechanisms, such as Swin Transformers with shifting windows [11], CoAtNet's convolution-attention integration [1], and ELSA's enhanced local sensitivity [20], demonstrate the importance of combining convolutional inductive biases with global self-attention. These approaches underscore that effective ECG analysis requires models capable of capturing both localized waveform morphology and long-range temporal dependencies.

Motivated by these insights, we propose a novel hierarchical transformer architecture for ECG signals that embeds convolutional biases directly into the attention mechanism. Queries are derived from overlapping convolutional projections, preserving locality, while global key-value pairs enable broader temporal modeling. This design allows the model to jointly capture fine-grained morphological variations and global rhythm context, aiming to advance the state of automated ECG interpretation. Our proposed Local-Global Attention ECG model (LGA-ECG) treats ECGs not just as clinical artifacts, but as representative of a broader class of complex, multiscale time series in healthcare. By integrating local convolutional inductive biases with global self-attention, LGA-ECG captures both fine-grained waveform morphology and long-range temporal patterns. Experiments show that this hybrid architecture outperforms state-of-the-art baselines across multiple ECG classification tasks.

2 Methods

The proposed LGA-ECG model integrates a convolutional front-end with a hierarchical transformer backbone (Figure 1). The design explicitly targets two complementary aspects of ECG interpretation: (i) fine-grained waveform morphology (P, QRS, T waves and intervals) and (ii) broader rhythm patterns across multiple beats. Convolutional layers provide strong locality biases, while transformer blocks equipped with the proposed Local–Global (LG) self-attention mechanism capture dependencies across multiple temporal scales.

2.1 Convolutional Front-End

The convolutional encoder consists of four sequential residual blocks (ResBlocks), each composed of two 1D convolutions (kernel sizes 7 and 3), BatchNorm, ReLU activations, and Dropout. These layers project raw ECG signals into a high-dimensional feature space while preserving spatial locality and morphology. All ResBlocks maintain 64 output channels and progressively reduce the temporal resolution, forming a feature sequence suitable for transformer-based processing.

2.2 Local–Global Self-Attention

Let the input sequence be $\mathbf{X} \in \mathbb{R}^{B\times N\times D}$, where B is the batch size, N the sequence length, and D the embedding dimension. Queries are derived from overlapping temporal windows: for each window of length l, a 1D convolution followed by averaging yields a single query vector

$$\mathbf{Q}^{(i)} = \frac{1}{l}\sum_{t=1}^{l} \text{Conv1D}_Q\big(\mathbf{X}^{(i)}\big)_{:,t}, \tag{1}$$

capturing localized waveform morphology. Stacking across M windows produces $\mathbf{Q} \in \mathbb{R}^{B\times M\times D}$. In contrast, keys and values are obtained from convolutional projections over the entire sequence, $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{B\times N\times D}$, enabling global context modeling while retaining a locality bias.

Attention is then computed via standard multi-head scaled dot-product attention:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}, \tag{2}$$

where $d = D/H$ is the head dimension. Residual connections preserve the morphology captured in queries, and each block reduces the temporal length by half through strided pooling, yielding progressively shorter but semantically richer sequences. This mechanism overcomes the quadratic cost of global attention while embedding clinically relevant local structure. A full mathematical derivation of this mechanism, including normalization, residual pathways, and hierarchical reduction, is provided in Appendix A.

Each Local-Global attention block reduces the temporal dimension by half, forming a hierarchical sequence representation. This structure mirrors clinical reasoning: earlier layers focus on wave morphology, while deeper layers model inter-beat patterns and rhythm. Moreover, this design reduces computational complexity from $\mathcal{O}(N^2)$ in global attention to $\mathcal{O}(MN)$ with $M = N/2^L$, enabling scalability to long ECG sequences.

2.3 Transformer Block and Overall Architecture

Each Transformer Block integrates the LG self-attention layer with a two-layer feed-forward network (MLP), both wrapped by normalization and residual connections. A parallel pooling–projection branch ensures dimensional compatibility when the temporal resolution is reduced. Stacking multiple blocks produces a hierarchical representation: early blocks emphasize waveform morphology, intermediate blocks capture intra-beat intervals, and deeper blocks encode rhythm-level dependencies. The complete architecture consists of: (i) a convolutional front-end of residual blocks that embed raw ECG signals into feature sequences, and (ii) a cascade of Transformer Blocks equipped with LG attention, progressively abstracting features toward clinically meaningful rhythm analysis.

Notably, LGA-ECG does not require explicit positional encodings. The convolutional projections inherently encode spatial and temporal position via their receptive fields. Ablation studies (Appendix B) show that neither absolute nor relative positional encodings improve overall performance, reinforcing the effectiveness of convolutional inductive biases.

Following prior work, we partitioned the dataset by patient ID to prevent leakage. Specifically, 90% of CODE-15 was used for training, 5% for validation, and 5% as a development set for ablations and hyperparameter tuning. Final evaluation was conducted on CODE-TEST, which includes 827 ECGs with consensus labels from expert cardiologists. A complete description of the datasets, experimental protocol, and implementation details is provided in Appendix C.

3 Results

We evaluate the proposed LGA-ECG model against state-of-the-art (SOTA) approaches for ECG abnormality classification, including ResNet-1 [15], ResNet-2 [6], BAT [8], ECG-DETR [7], and HiT [2]. Table 1 summarizes the average performance across accuracy, precision, recall, and F1-score.

LGA-ECG achieves the highest accuracy (0.994) and F1-score (0.885), outperforming all baselines. While BAT exhibits slightly higher precision (0.918 vs. 0.907), our model significantly improves recall (0.872 vs. 0.799), surpassing the 0.8 threshold for the first time. This increase in recall, crucial

Table 1: Average performance of LGA-ECG when compared to other SOTA methods.

Metrics	ResNet-1	ResNet-2	ECG-Transform	BAT	ECG-DETR	HiT	LGA-ECG
Accuracy	0.991	0.989	0.981	0.991	0.984	0.991	**0.994**
Precision	0.875	0.908	0.711	**0.918**	0.776	0.909	0.907
Recall	0.778	0.743	0.687	0.799	0.661	0.798	**0.872**
F1-Score	0.814	0.811	0.677	0.848	0.699	0.841	**0.885**

Table 2: Per class F1-score of LGA-ECG and baseline methods in the test set.

Abnormality	ResNet-1	ResNet-2	ECG-Transform	BAT	ECG-DETR	HiT	LGA-ECG
1st AVB	0.661	0.719	0.489	0.689	0.631	0.682	**0.8**
RBBB	**0.924**	0.890	0.909	0.922	0.747	0.886	0.923
LBBB	0.927	0.843	0.886	0.945	0.826	0.909	**0.983**
SB	0.767	0.821	0.535	**0.836**	0.588	0.824	0.778
AF	0.703	0.758	0.478	0.818	0.563	0.833	**0.880**
ST	0.897	0.833	0.763	0.870	0.838	0.914	**0.946**
Avg. F1	0.814	0.811	0.677	0.848	0.699	0.841	**0.885**

in medical diagnosis where false negatives are costly, is obtained without a drastic drop in precision, demonstrating the robustness of LGA-ECG for practical deployment.

Per-class results (Table 2) show that LGA-ECG sets new benchmarks in four abnormalities: ST (0.946), LBBB (0.983), AF (0.880), and 1st AVB (0.800). Performance is nearly equivalent to the best baseline for RBBB (0.923 vs. 0.924), while the only underperformance occurs in SB, likely due to difficulties in capturing longer RR intervals. Despite this, the model still maintains competitive results.

Finally, we benchmarked model performance against human annotators using the CODE-TEST dataset. Figure 2 shows that LGA-ECG consistently outperforms 4th-year cardiology residents, 3rd-year emergency residents, and 5th-year medical students across all key metrics, using consensus labels from specialist cardiologists as ground truth. These findings demonstrate that LGA-ECG not only surpasses existing machine learning baselines but also exceeds the diagnostic performance of medical professionals at varying levels of expertise.

4 Conclusion and Future Work

This study introduced LGA-ECG, a novel deep learning model for ECG classification that integrates local convolutional inductive biases with global self-attention mechanisms. Our approach effectively captures both fine-grained morphological features and broader temporal dependencies, leading to improvements over state-of-the-art methods. LGA-ECG achieved the highest F1-score among all evaluated models, demonstrating the benefits of local-global attention in medical signal analysis.

A promising and important future direction is extending LGA-ECG with self-supervised learning techniques to pretrain the model on large unlabeled ECG datasets before fine-tuning it for classification. This approach could enhance generalization and robustness, particularly for rare abnormalities with limited labeled data.

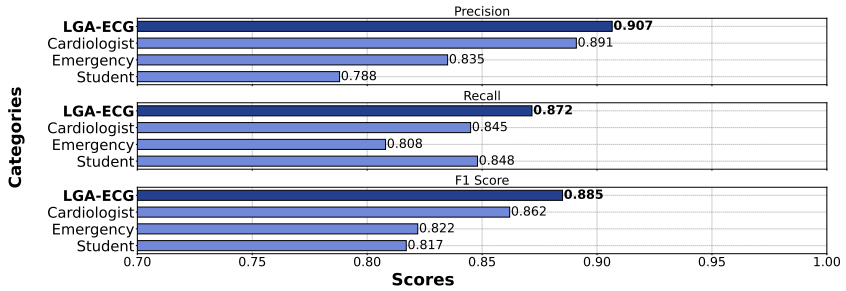

Figure 2: Comparison of the average metrics between LGA-ECG and human performance.

References

[1] Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. In *Advances in Neura l Information Processing Systems*, volume 34, pages 3965–3977, 2021.

[2] Pedro Robles Dutenhefner, Turi Andrade Vasconcelos Rezende, Gisele Lobo Pappa, Gabriela Miana de Matos Paixão, Antônio Luiz Pinho Ribeiro, and Wagner Meira Jr. Um transformer hierárquico para classificação e diagnóstico de eletrocardiograma. *Journal of Health Informatics*, 16(Especial), 2024.

[3] Zahra Ebrahimi, Mohammad Loni, Masoud Daneshtalab, and Arash Gharehbaghi. A review on deep learning methods for ecg arrhythmia classification. *Expert Systems with Applications: X*, 7:100033, 2020.

[4] Hany El-Ghaish and Emadeldeen Eldele. Ecgtransform: Empowering adaptive ecg arrhythmia classification framework with bidirectional transformer. *Biomedical Signal Processing and Control*, 89:105714, 2024.

[5] Nobuaki Fujita, Akira Sato, and Masatoshi Kawarasaki. Performance study of wavelet-based ecg analysis for st-segment detection. In *2015 38th International Conference on Telecommunications and Signal Processing (TSP)*, pages 430–434. IEEE, 2015.

[6] Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69, January 1 2019.

[7] 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*, 144:105325, 2022.

[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)*, pages 320–329. IEEE, 2021.

[9] Emilly M Lima, Antônio H Ribeiro, Gabriela MM Paixão, Manoel Horta Ribeiro, Marcelo M Pinto-Filho, Paulo R Gomes, Derick M Oliveira, Ester C Sabino, Bruce B Duncan, Luana Giatti, et al. Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nature communications*, 12(1):5117, 2021.

[10] Xinwen Liu, Huan Wang, Zongjin Li, and Lang Qin. Deep learning in ecg diagnosis: A review. *Knowledge-Based Systems*, 227:107187, 2021.

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[12] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[13] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[14] P Rajpurkar, AY Hannun, M Haghpanahi, C Bourn, and AY Ng. Cardiologist-level arrhythmia detection with convolutional neural networks. arxiv 2017. *arXiv preprint arXiv:1707.01836*, 2017.

[15] Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1–9, 2020.

[16] Antonio Luiz P Ribeiro, Gabriela MM Paixao, Paulo R Gomes, Manoel Horta Ribeiro, Antonio H Ribeiro, Jessica A Canazart, Derick M Oliveira, Milton P Ferreira, Emilly M Lima, Jermana Lopes de Moraes, et al. Tele-electrocardiography and bigdata: the code (clinical outcomes in digital electrocardiography) study. *Journal of electrocardiology*, 57:S75–S78, 2019.

[17] Diogo Tuler, Pedro Robles Dutenhefner, Jose Geraldo Fernandes, Turi Rezende, Gabriel Lemos, Gisele L Pappa, Gabriela Paixao, Antônio Ribeiro, and Wagner Meira Jr. Leveraging cardiologists prior-knowledge and a mixture of experts model for hierarchically predicting ecg disorders. *Computing in Cardiology*, 2024.

[18] J Wang and W Li. Atrial fibrillation detection and ecg classification based on cnn-bilstm. arxiv 2020. *arXiv preprint arXiv:2011.06187*, 2020.

[19] World Health Organization. Cardiovascular diseases, 2024. Accessed: 2024-08-16.

[20] Jingkai Zhou, Pichao Wang, Fan Wang, Qiong Liu, Hao Li, and Rong Jin. ELSA: Enhanced local self-attention for vision transformer. *arXiv preprint arXiv:2112.12786*, 2021.

A Expanded Methodology

ECG analysis requires capturing information across multiple temporal scales: wave morphology (P, QRS, T), intra-heartbeat intervals (PR, QT), and inter-beat distances essential for rhythm analysis. We propose a novel self-attention mechanism tailored for ECG signals, which effectively balances fine-grained morphological details with global heartbeat patterns.

The proposed model first uses convolutional layers to project the ECG into an embedding space. Its core comprises layers of a novel windowed self-attention and feed-forward blocks with residual connections. Unlike traditional global self-attention, our method extracts queries (Q) from small overlapping windows to preserve local detail, while keys (K) and values (V) are computed globally, capturing long-range dependencies. Additionally, each self-attention block progressively reduces the sequence length, similar to convolutional pooling, allowing hierarchical abstraction from local waveform characteristics toward global rhythm and beat-to-beat features.

A.1 Local-Global Self-Attention

We now formalize the local–global attention mechanism, assuming the input $\mathbf{X}$ has already been projected into an embedding space by the convolutional encoder described in Section 2.

Step 1: Normalization. First, we apply a standard layer normalization along the embedding dimension to stabilize and normalize the input:

$$\tilde{\mathbf{X}} = \text{LayerNorm}(\mathbf{X}), \quad \tilde{\mathbf{X}} \in \mathbb{R}^{B\times N\times D}. \tag{3}$$

Step 2: Local Windowed Query Generation. To effectively capture precise wave-level morphological details from ECG signals, we introduce a local window-based query generation strategy. Starting from the normalized input tensor $\tilde{\mathbf{X}} \in \mathbb{R}^{B\times N\times D}$, we extract a series of overlapping windows along the temporal dimension to form localized queries ($\mathbf{Q}$).

Formally, given a window length l and stride s, we extract M overlapping windows from the sequence, where:

$$M = \left\lfloor \frac{N-l}{s} \right\rfloor + 1. \tag{4}$$

For each window indexed by $i \in \{0, 1, \ldots, M-1\}$, we select a contiguous subset of the input sequence:

$$\tilde{\mathbf{X}}^{(i)} = \tilde{\mathbf{X}}\left[:,\ (i\cdot s) : (i\cdot s + l),\ :\right], \quad \tilde{\mathbf{X}}^{(i)} \in \mathbb{R}^{B\times l\times D}. \tag{5}$$

Next, each extracted window $\tilde{\mathbf{X}}^{(i)}$ undergoes a convolutional projection along the temporal dimension. Specifically, we apply a 1D convolution with kernel size k_q, stride 1, padding p_q, and D output channels, obtaining:

$$\mathbf{Q}^{(i)}_{\text{conv}} = \text{Conv1D}_Q\left(\tilde{\mathbf{X}}^{(i)}\right), \quad \mathbf{Q}^{(i)}_{\text{conv}} \in \mathbb{R}^{B\times D\times l}. \tag{6}$$

The output $\mathbf{Q}^{(i)}_{\text{conv}}$ represents an enhanced embedding of the original local window, where each temporal position within the window has been projected into a new feature space through convolution.

To summarize this detailed local information into a single representative query vector per window, we then average these embeddings along the temporal dimension of length l. For each window i, the averaged query vector is calculated as:

$$\mathbf{Q}^{(i)} = \frac{1}{l}\sum_{t=1}^{l} \mathbf{Q}^{(i)}_{\text{conv}}[:,:,t], \quad \mathbf{Q}^{(i)} \in \mathbb{R}^{B\times D}. \tag{7}$$

Finally, stacking all the averaged queries across the M extracted windows results in the complete query tensor for attention:

$$\mathbf{Q} = \left[\mathbf{Q}^{(0)}, \mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(M-1)}\right], \quad \mathbf{Q} \in \mathbb{R}^{B\times M\times D}. \tag{8}$$

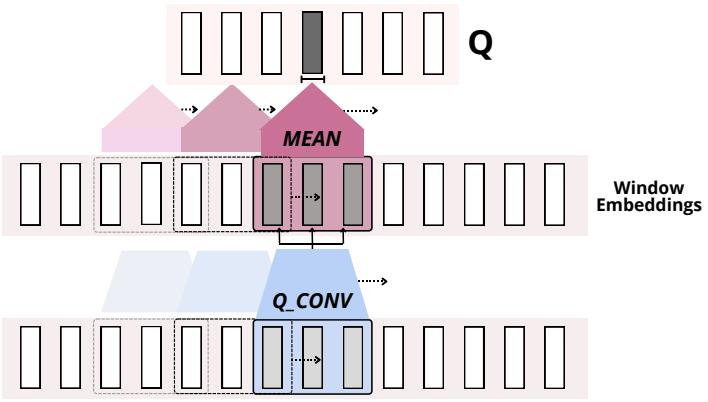

Figure 3: Mean query extraction process for each ECG window.

To enhance stability and facilitate residual connections in deeper layers, we retain a copy of the query tensor as a residual term. This preserves local morphological details captured by convolution, ensuring stable gradients and improved convergence.

This process can be implemented in a simple and effective manner using a combination of a 1D convolutional layer that preserves the input shape, followed by an average pooling layer. The kernel size of the pooling operation determines the temporal compression factor. This approach is illustrated in Figure 3.

Step 3: Global Key and Value Generation. In contrast to the localized queries, keys ($\mathbf{K}$) and values ($\mathbf{V}$) are computed from the entire normalized sequence, enabling each local query to attend globally. We define these global projections using convolutional layers to retain a locality inductive bias while still allowing global context modeling:

$$\mathbf{K}_{\text{conv}} = \text{Conv1D}_K(\tilde{\mathbf{X}}), \quad \mathbf{V}_{\text{conv}} = \text{Conv1D}_V(\tilde{\mathbf{X}}), \tag{9}$$

both producing tensors of shape:

$$\mathbf{K}_{\text{conv}}, \mathbf{V}_{\text{conv}} \in \mathbb{R}^{B\times D\times N}. \tag{10}$$

We permute them back to match the original embedding format:

$$\mathbf{K} = \mathbf{K}_{\text{conv}}^{\top} \in \mathbb{R}^{B\times N\times D}, \quad \mathbf{V} = \mathbf{V}_{\text{conv}}^{\top} \in \mathbb{R}^{B\times N\times D}. \tag{11}$$

Step 4: Multi-Head Local-Global (LG) Attention Computation. We now apply a multi-head attention mechanism. For H attention heads, we split the embedding dimension D into H sub-dimensions of size $D_h = D/H$:

$$\mathbf{Q}_h \in \mathbb{R}^{B\times M\times D_h}, \quad \mathbf{K}_h, \mathbf{V}_h \in \mathbb{R}^{B\times N\times D_h}, \quad h = 1, \ldots, H. \tag{12}$$

For each head h, the scaled dot-product attention scores are computed as:

$$\mathbf{A}_h = \text{softmax}\left(\frac{\mathbf{Q}_h\mathbf{K}_h^{\top}}{\sqrt{D_h}}\right) \in \mathbb{R}^{B\times M\times N}. \tag{13}$$

Subsequently, we calculate the features as a weighted sum of values:

$$\mathbf{O}_h = \mathbf{A}_h\mathbf{V}_h \in \mathbb{R}^{B\times M\times D_h}. \tag{14}$$

Concatenating across all heads, we get the combined multi-head attention output:

$$\mathbf{O} = \text{concat}(\mathbf{O}_1, \ldots, \mathbf{O}_H) \in \mathbb{R}^{B\times M\times D}. \tag{15}$$

Step 5: Residual Connection and Sequence Reduction. Finally, we reintroduce the residual query information by adding back the previously stored queries $\mathbf{Q}_{\text{res}}$, maintaining strong local fidelity:

$$\mathbf{Y} = \mathbf{O} + \mathbf{Q}_{\text{res}}, \quad \mathbf{Y} \in \mathbb{R}^{B\times M\times D}. \tag{16}$$

The sequence length is effectively reduced from N to M by selecting a stride $s = 2$, ensuring $M = N/2$. This hierarchical summarization progressively condenses ECG features, capturing local and global information.

Our LG self-attention combines standard self-attention, convolution, and hierarchical transformers while overcoming their limitations. Unlike traditional self-attention, which lacks locality and scales quadratically, or convolutions, which struggle with long-range dependencies, our method extracts locally-informed queries via overlapping convolutional projections while maintaining global attention through sequence-wide keys and values. Additionally, convolutional projections inherently encode positional information, removing the need for explicit positional encodings. The local-global attention is illustrated in Figure 4

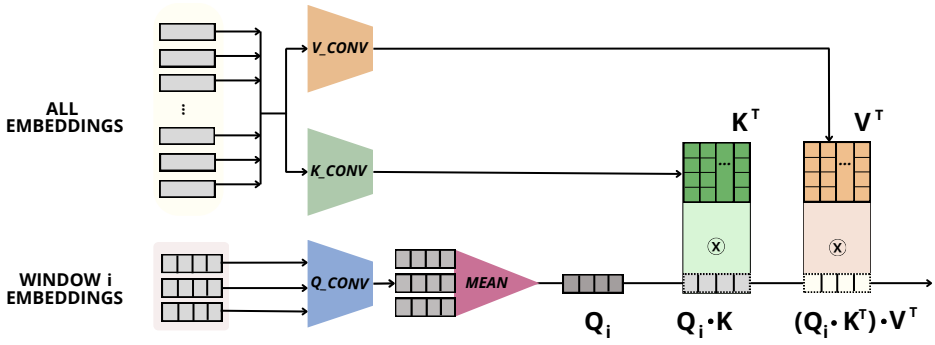

Figure 4: Local-global self-attention operation for one ECG embedding window.

A.2 Transformer Block with Local–Global Self-Attention

Each Transformer Block integrates the LGA layer within a residual architecture. Its formal computation is detailed below.

Given an input tensor $\mathbf{X} \in \mathbb{R}^{B\times N\times D}$, where B is the batch size, N is the sequence length, and D is the embedding dimension, the Transformer Block initially applies layer normalization along the embedding dimension:

$$\tilde{\mathbf{X}} = \text{LayerNorm}(\mathbf{X}), \quad \tilde{\mathbf{X}} \in \mathbb{R}^{B\times N\times D}. \tag{17}$$

Subsequently, the normalized sequence is processed by the local-global self-attention layer. Due to the windowed attention design, the spatial dimension N is effectively reduced approximately by half, from N to $M = N/2$, resulting in an output tensor $\mathbf{Y}_{\text{attn}}$:

$$\mathbf{Y}_{\text{attn}} = \text{LocalGlobalAttention}(\tilde{\mathbf{X}}), \quad \mathbf{Y}_{\text{attn}} \in \mathbb{R}^{B\times M\times D}. \tag{18}$$

To maintain a consistent residual connection despite the reduction in sequence length, we apply a pooling operation followed by a 1×1 convolution to the normalized input $\tilde{\mathbf{X}}$, ensuring dimensional compatibility:

$$\mathbf{X}_{\text{res}} = \text{Conv1D}\left(\text{MaxPool1D}\left(\tilde{\mathbf{X}}\right)\right), \quad \mathbf{X}_{\text{res}} \in \mathbb{R}^{B\times M\times D}. \tag{19}$$

Here, the max pooling operation reduces the temporal dimension by half, from N to M, while the 1×1 convolution adjusts embedding dimensions and reinforces the residual pathway. The resulting residual tensor $\mathbf{X}_{\text{res}}$ is added to the self-attention output, stabilizing training and enhancing gradient flow:

$$\mathbf{Z} = \mathbf{Y}_{\text{attn}} + \mathbf{X}_{\text{res}}, \quad \mathbf{Z} \in \mathbb{R}^{B\times M\times D}. \tag{20}$$

Next, we apply a second-layer normalization followed by a feed-forward neural network, often called the Multi-Layer Perceptron (MLP). This MLP consists of two linear layers with an intermediate non-linearity (ReLU). The dimensionality of the intermediate MLP layer, denoted as D_{MLP}, dynamically increases at each transformer block stage i, defined explicitly as $D_{\text{MLP}} = D_{\text{base}} \times 2 \times i$. Specifically, the MLP initially projects each embedding vector from the input dimension D to this expanded dimension D_{MLP}:

$$\mathbf{Z}_{\text{MLP}}^{(i)} = \text{ReLU}\left(\mathbf{Z}^{(i)}\mathbf{W}_1^{(i)} + \mathbf{b}_1^{(i)}\right), \quad \mathbf{Z}_{\text{MLP}}^{(i)} \in \mathbb{R}^{B\times M\times (D_{\text{base}}\times 2\times i)}, \tag{21}$$

and subsequently project it back to the original embedding dimension D:

$$\mathbf{Z}_{\text{out}}^{(i)} = \mathbf{Z}_{\text{MLP}}^{(i)}\mathbf{W}_2^{(i)} + \mathbf{b}_2^{(i)}, \quad \mathbf{Z}_{\text{out}}^{(i)} \in \mathbb{R}^{B\times M\times D}. \tag{22}$$

This incremental expansion of the MLP dimensionality at successive transformer stages allows the model to progressively capture more complex and abstract features. A second residual connection then integrates the MLP output back into the main pathway, resulting in the final output tensor of each transformer block:

$$\mathbf{X}_{\text{final}}^{(i)} = \mathbf{Z}^{(i)} + \mathbf{Z}_{\text{out}}^{(i)}, \quad \mathbf{X}_{\text{final}}^{(i)} \in \mathbb{R}^{B\times M\times D}. \tag{23}$$

This staged expansion of the MLP dimension allows deeper layers to encode increasingly complex and abstract features, naturally aligning with the progressive shift from fine-grained morphological details to broader, long-range inter-beat relationships.

Each Transformer Block hierarchically condenses and enriches representations, aligning with clinical ECG analysis. Early layers capture fine-grained wave morphology, intermediate layers focus on intra-heartbeat intervals, and deeper layers model long-range dependencies across heartbeats, effectively identifying rhythm abnormalities. This structured progression inherently encodes clinically relevant inductive biases.

B Ablations

To assess the effectiveness of our proposed local-global attention mechanism, we perform a series of ablation studies to isolate its contributions and better understand its impact on ECG feature extraction.

B.1 Alternative Attention Mechanisms

First, we compare the proposed LGA against alternative attention strategies. Our goal is to evaluate how different query, key, and value configurations influence the model's ability to capture fine-grained ECG morphology and global contextual dependencies.

ViT-like: We begin by examining a standard ViT-like approach, which applies global self-attention across the entire sequence using linear projections for queries, keys, and values. While this method captures the global context effectively, it lacks local inductive biases.

Swin-like: Next, we compare our method with a local attention mechanism inspired by Swin Transformer [11], where self-attention is restricted to non-overlapping windows. This approach captures local features while progressively integrating global context through stacked local attention and inter-block pooling.

Global Q, K, V: We also analyze a global attention variant, which follows the standard attention mechanism but replaces linear projections with convolutional and average pooling layers. In this configuration, queries are computed in the same manner as keys and values, ensuring that all positions attend to each other globally. Although this setup preserves global context awareness, it may fail to efficiently encode localized waveform structures.

Local Q, K, V: Finally, we examine a fully localized variant, where the query Q is the mean of the embeddings within a window, while the keys K and values V correspond only to the embeddings of that window, without global context. We extract overlapping windows, ensuring that each window is

Table 3: Per class F1-score comparison between different attention mechanisms.

Abnormality	ViT-like	Swin-like	Global Q, K, V	Local Q, K, V	LGA-ECG
1st AVB	0.653	0.682	0.809	0.782	**0.800**
RBBB	0.862	0.886	0.925	**0.955**	0.923
LBBB	0.875	0.909	0.909	0.982	**0.983**
SB	0.768	**0.824**	0.733	0.750	0.778
AF	0.792	0.833	0.833	0.782	**0.880**
ST	0.887	0.914	0.870	0.885	**0.946**
Avg. F1	0.806	0.841	0.847	0.856	**0.885**

condensed into a single embedding after the attention operation. This progressively reduces the data by half at each stage, establishing a hierarchical processing framework.

The results in Table 3 show that LGA-ECG achieves the highest F1-score (0.885), outperforming all alternative attention mechanisms. By integrating local convolutional inductive biases with global context, LGA-ECG surpasses both fully global (ViT-like, global QKV) and fully local (Swin-like, local QKV) approaches, demonstrating superior feature extraction for ECG classification.

B.2 Positional Encoding

We further evaluate whether convolutional biases introduced by the adapted projections sufficiently capture positional information, which is crucial in ECG analysis due to the diagnostic relevance of intervals between waves and heartbeats. Specifically, we investigate three positional encoding strategies:

Absolute sinusoidal positional encoding: Predefined sinusoidal functions of varying frequencies are computed based on absolute positions and directly summed to the embeddings after the convolutional projection, explicitly embedding absolute positional information into each token.

Absolute learnable positional encoding: A trainable embedding vector for each absolute position is learned during training and summed to the embeddings immediately after convolutional projection, enabling the model to adaptively capture position-specific patterns.

Relative positional encoding: A learnable relative position matrix, matching the attention matrix dimensions, is added directly to the attention scores before the softmax operation. This matrix encodes pairwise relative distances between token positions, allowing the model to flexibly emphasize or suppress interactions based on relative position.

Table 4: Per class F1-score comparison between positional encoding strategies.

Abnormality	Sinusoidal APE	Learnable APE	RPE	Without PE
1st AVB	0.681	0.526	0.667	**0.800**
RBBB	0.857	0.844	0.928	**0.923**
LBBB	0.966	0.947	0.909	**0.983**
SB	0.743	0.643	**0.800**	0.778
AF	0.769	0.667	0.621	**0.880**
ST	0.873	0.899	0.853	**0.946**
Avg. F1	0.815	0.754	0.796	**0.885**

The results in Table 4 indicate that LGA-ECG achieves the highest performance without explicit positional encoding, suggesting that the convolutional projections effectively encode spatial dependencies inherent in ECG signals. While relative positional encoding improves certain classes, neither absolute nor relative positional encodings consistently enhances performance, reinforcing the effectiveness of the learned convolutional inductive biases in capturing diagnostic temporal structures. Notably, relative positional encoding (RPE) improved SB detection, likely aiding R-R interval analysis for bradycardia and rhythm abnormalities. A similar trend in the Swin-like attention, which also uses RPE, emphasizes its role in enhancing rhythm irregularity detection.

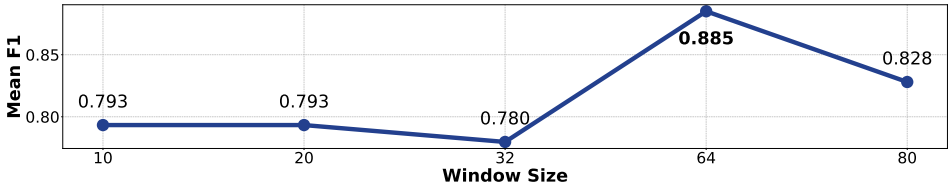

Figure 5: F1-score comparison across different window sizes.

B.3 Window Size Analysis

We investigate the impact of varying the window size on the proposed LGA-ECG architecture. This hyperparameter controls both the kernel size of convolutional projections and the temporal length of local segments used to compute the local queries. By testing different window sizes, we aim to evaluate the sensitivity of the model's performance to the temporal scale at which local morphological features are captured.

As shown in Figure 5, the best performance was achieved with a window size of 64. This setting provides a trade-off between capturing fine-grained waveform details and maintaining sufficient temporal context for effective local-global feature integration.

C Experiment setup

C.1 Datasets

Our model was trained and evaluated using CODE-15, a publicly available 15% subset of the CODE (Clinical Outcomes in Digital Electrocardiography) dataset [16]. CODE contains over 2 million ECGs from Minas Gerais, Brazil, annotated by cardiologists for six cardiac abnormalities: first-degree atrioventricular block (1st AVB), right bundle branch block (RBBB), left bundle branch block (LBBB), sinus bradycardia (SB), atrial fibrillation (AF), and sinus tachycardia (ST). These conditions indicate an increased risk for cardiovascular events, including stroke, heart failure, and sudden death, and require targeted clinical interventions. CODE-15 comprises 345,779 exams from 233,770 patients and has been widely adopted in ECG research, serving as a benchmark dataset for developing and evaluating deep learning models [15] [17].

We evaluated our model using the publicly available CODE-TEST dataset, also collected by the Telehealth Network of Minas Gerais (TNMG). CODE-TEST comprises 827 ECGs labeled by consensus among two or three cardiologists, covering the same six cardiac abnormalities. The high-quality, expert-consensus labels provide a robust benchmark for performance assessment.

For developing and validating the LGA-ECG model, the dataset is divided into four subsets by patient IDs: 90% of CODE-15 is used as the training set to train the model, while 5% of CODE-15 serves as the validation set for early stopping. An additional 5% of CODE-15 is designated as the development set, which is utilized for hyperparameter tuning and ablation studies. Finally, the entire CODE-TEST dataset is used as the test set to evaluate the final model performance against baseline methods.

C.2 Implementation details and Benchmarks

For comparison, we assessed LGA-ECG against a suite of baseline models spanning diverse architectural families, including traditional CNN and transformer-based architectures. This selection ensured a rigorous and comprehensive evaluation across distinct modeling paradigms. The baselines were implemented using their original authors' codebases, with training settings configured according to their recommendations. All models were trained on the same Training Set and evaluated on the Test Set to ensure consistent comparisons. We employed standard classification metrics to evaluate the models: accuracy, F1-score, precision, and recall. These metrics were computed for each cardiac condition individually to provide a detailed understanding of model performance across different diseases, as well as averaged (macro).

The training process utilized the AdamW optimizer [12] and employed a cosine annealing learning rate schedule [13]. The initial learning rate was set to 0.0001 and was decreased cosine-wise to

0.00001 throughout the training. Additionally, early stopping was implemented, which terminates training if the validation error does not decrease for seven consecutive epochs. The training was conducted in parallel using 4 NVIDIA V100 GPUs.