Computers in Biology and Medicine 203 (2026) 111462

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A hybrid hierarchical transformer model for ECG classification and age prediction

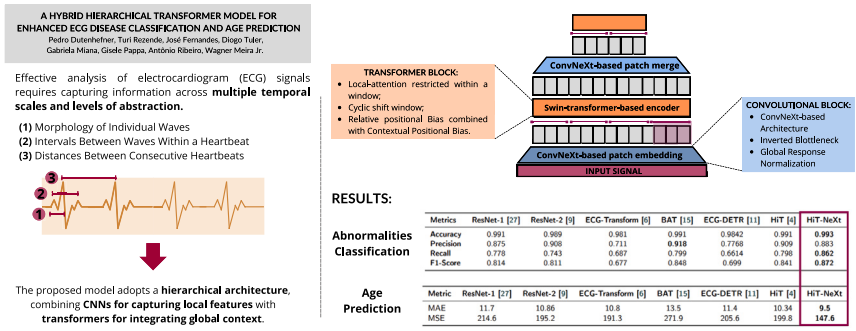
Pedro Dutenhefner[a,1,*], Turi Rezende[a,1], José Geraldo Fernandes[a], Diogo Tuler[a], Gabriela M.M. Paixão[b], Gisele Pappa[a], Antônio Ribeiro[b], Wagner Meira Jr.[a]

[a] Department of Computer Science, Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627, Belo Horizonte, 31270-901, Minas Gerais, Brazil
[b] School of Medicine, Universidade Federal de Minas Gerais, Av. Prof. Alfredo Balena, 190, Belo Horizonte, 30130-100, Minas Gerais, Brazil

HIGHLIGHTS

- HiT-NeXt is a hybrid hierarchical model combining CNNs and Transformers for ECG analysis.
- The model integrates local convolutional processing with contextualized local attention.
- It introduces a novel combination of Relative Position Bias and Contextual Positional Encoding (CoPE).
- HiT-NeXt surpasses state-of-the-art baselines in ECG abnormality classification and age prediction tasks.

GRAPHICAL ABSTRACT



Metrics	ResNet-1 [27]	ResNet-2 [9]	ECG-Transform [6]	BAT [15]	ECG-DETR [11]	HiT [4]	HiT-NeXt
Accuracy	0.991	0.989	0.981	0.991	0.9842	0.991	**0.993**
Precision	0.875	0.908	0.711	**0.918**	0.7768	0.909	0.883
Recall	0.778	0.743	0.687	0.799	0.6614	0.798	**0.862**
F1-Score	0.814	0.811	0.677	0.848	0.699	0.841	**0.872**

Metric	ResNet-1 [27]	ResNet-2 [9]	ECG-Transform [6]	BAT [15]	ECG-DETR [11]	HiT [4]	HiT-NeXt
MAE	11.7	10.86	10.8	13.5	11.4	10.34	**9.5**
MSE	214.6	195.2	191.3	271.9	205.6	199.8	**147.6**

ARTICLE INFO

Keywords:
ECG classification
Age prediction
Transformer model
Hierarchical model
Hybrid model

ABSTRACT

Electrocardiograms (ECGs) play a crucial role in cardiovascular healthcare, requiring effective analytical models. ECG analysis is inherently hierarchical, involving multiple temporal scales from individual waveforms to intervals within heartbeats, and finally to the distances between heartbeats. Convolutional Neural Networks (CNNs) have demonstrated strong performance in ECG classification tasks due to their inductive bias toward local connectivity and translation invariance. In other domains, Transformers have emerged as powerful models for capturing long-range dependencies. This paper introduces HiT-NeXt, a hybrid hierarchical model designed to capture both local morphological patterns and global temporal dependencies by combining CNNs with transformer blocks featuring restricted attention windows. The model incorporates ConvNeXt-based convolutional layers to extract local features and perform patch merging, enabling hierarchical representation learning. Transformer blocks are constrained with local attention windows and leverage relative contextual positional encoding to incorporate positional information effectively into embeddings, enhancing robustness to translations in ECG signal patterns. Experimental results demonstrate that HiT-NeXt outperforms state-of-the-art methods on tasks including ECG abnormality classification and cardiological age prediction, achieving superior performance compared to both existing models and cardiologist evaluations.[2]

* Corresponding author.
Email addresses: pedroroblesduten@ufmg.br (P. Dutenhefner), turi@ufmg.br (T. Rezende), josegeraldof@ufmg.br (J. Geraldo Fernandes), diogochaves@dcc.ufmg.br (D. Tuler), gabrielamiana@ufmg.br (G.M.M. Paixão), glpappa@dcc.ufmg.br (G. Pappa), alpr@ufmg.br (A. Ribeiro), meira@dcc.ufmg.br (W. Meira Jr.).

[1] These authors contributed equally to this work.
[2] PyTorch Implementation: https://github.com/pedroroblesduten/HiT-NeXt-ECG.

https://doi.org/10.1016/j.compbiomed.2026.111462
Received 28 March 2025; Received in revised form 30 November 2025; Accepted 7 January 2026
Available online 21 January 2026

1. Introduction

Cardiovascular diseases (CVDs) are the leading global cause of death, accounting for 17.9 million deaths in 2019, representing 32% of worldwide deaths, according to the World Health Organization (WHO) [37]. Electrocardiograms (ECGs), as simple and non-invasive exams, play a crucial role in diagnosing and monitoring cardiovascular conditions. They have gained greater relevance in digital health with the widespread use of digital ECGs [19]. In recent years, artificial intelligence (AI) in electrocardiography has emerged as a valuable tool for the automatic classification of ECG abnormalities, sex and age prediction [17], wave segmentation [17] and prediction of cardiac events [6]. AI models applied to ECGs often draw from innovations in other machine learning fields, including computer vision, text processing, and speech recognition, demonstrating the adaptability of techniques across domains.

Initially, machine learning applications in ECG analysis focused on extracting meaningful signal features to improve classification performance. Common techniques included analyzing the morphology of the QRS complex and RR intervals [13,18]. In this context, signal processing-based approaches, such as Fourier Transform, Discrete Wavelet Transform (DWT), and adaptive filtering, were applied to enhance signal quality, detect fiducial points, and extract critical features such as the ST segment and T wave. For example, DWT has been widely used to detect QRS complexes and other fiducial points due to its ability to analyze signals in both time and frequency domains [8,10,24]. Subsequently, these handcrafted features were used to train traditional classifiers, such as k-Nearest Neighbors (k-NN) and Support Vector Machines (SVM), enabling effective and computationally efficient solutions for the detection and classification of arrhythmias [25,35].

With the advent of deep learning, there has been a paradigm shift in ECG analysis methodologies. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been increasingly utilized given their ability to learn features automatically from raw data. CNNs leverage their inherent inductive biases, such as spatial locality and translation equivariance, to efficiently extract localized features from ECG signals [27,28]. RNNs, on the other hand, are adept at modeling temporal dependencies, effectively capturing the dynamic nature of heart rhythms [14,31]. These deep learning approaches have significantly advanced ECG analysis, achieving performance levels that, in some cases, exceed those of physicians [28]. Motivated by these advances, this work proposes a novel deep learning model for ECG analysis.

In recent years, transformer models have driven a paradigm shift in AI, significantly advancing domains such as computer vision [3,20] and natural language processing [26]. Despite their success, the application of transformer architectures to time-series data, particularly ECG, has not yielded comparable breakthroughs [1]. This discrepancy can be attributed to several inherent challenges. The global self-attention mechanism obscures crucial local and temporal patterns critical for accurate ECG analysis. Additionally, in Vision Transformers (ViT) approaches, the usual linear patch embedding process used can lead to a loss of fine-grained temporal information in ECG data, further diminishing model performance [5]. Furthermore, transformers typically rely on absolute positional encodings to incorporate sequential information; however, these traditional encodings break invariance properties that could be useful for ECG analysis [2]. Consequently, the effectiveness of transformers in this domain remains limited, highlighting the need for further research to adapt and enhance these models for time-series applications and unlock the potential benefits of transformers in ECG analysis.

According to cardiology specialists, a model developed for ECG analysis must be capable of extracting and learning the following features: the detailed local morphology of individual waves, the intervals between waves within a single heartbeat, general information regarding a heartbeat, and the distances between consecutive heartbeats. This underscores the necessity of feature extraction at multiple temporal and abstraction scales. The model must capture local information and effectively integrate it with information extracted from the global context. This naturally suggests combining models that excel at extracting local features, such as CNNs, with models capable of learning features in a global context, such as transformers.

Previous efforts were made to combine convolutional and transformer models, following one of three approaches:

Incorporating convolutional bias in transformers: This line of work is based on restricting the attention mechanism to enforce local feature extraction at multiple hierarchical levels, as proposed in [20].

Incorporating transformer bias in CNNs: This approach integrates transformer-like properties into convolutional models by employing large kernel sizes and inverted bottlenecks, which enhance the receptive field and aim to approximate global context [36].

Concatenating transformers with CNNs: This approach uses convolutional blocks in the first layers to extract local features, followed by transformer blocks to learn how to contextualize the extracted features globally [12].

The first two approaches face limitations inherent to their respective model designs: the Swin Transformer struggles to replicate the efficient feature extraction achieved by CNNs during resolution reduction [21], while ConvNeXt's large receptive field, despite enhancing global context, falls short of capturing the rich contextual representations enabled by attention mechanisms. The third approach, however, often leads to performance saturation, as existing implementations of convolutional and Transformer blocks struggle to balance efficiency and performance simultaneously [15].

In this work, we argue that the three approaches are not mutually exclusive and can be cohesively combined to harness the strengths of both Transformers and CNNs. We propose HiT-NeXt, a hybrid hierarchical model for ECG analysis that integrates interleaved convolutional and transformer blocks with restricted local attention windows. The local attention mechanism, combined with hierarchical and multi-level learning, allows HiT-NeXt to extract information from ECG signals across various levels of detail and temporal scales, effectively balancing global feature extraction through convolutions and contextual local processing via attention.

The HiT-NeXt model is built upon four main architectural ideas. First, convolutional layers with ConvNeXt-inspired improvements, such as inverted bottlenecks and Global Response Normalization (GRN) [36], are used to extract and aggregate local ECG patterns efficiently. Second, transformer blocks with restricted attention windows learn contextual representations by focusing on local temporal neighborhoods, mitigating the drawbacks of global self-attention for ECG signals. Third, these convolutional and transformer components are arranged in a multi-stage hierarchical architecture, enabling feature learning across multiple temporal and abstraction scales, from wave-level morphology to long-range rhythm patterns. Finally, a contextual relative positional encoding mechanism, which combines distance-based relative position bias with Contextual Position Encoding (CoPE) [9], enriches the attention mechanism with clinically meaningful temporal structure while preserving robustness to temporal translations.

HiT-NeXt surpasses state-of-the-art baselines in tasks related to the classification of ECG abnormalities and cardiological age prediction, demonstrating superior performance across these challenging domains. We also provide an extensive discussion of the model development process and the key insights obtained from intermediate results, which together help clarify how each architectural choice contributes to the final performance.

Motivation, contributions, and applications

The present work is motivated by two main limitations of current deep learning approaches for ECG analysis: (i) most architectures emphasize either local waveform morphology or global temporal context, but rarely model multiple temporal scales within a single unified design;

and (ii) transformer-based ECG models often rely on global self-attention and absolute positional encodings, which may obscure fine-grained ECG structures and disrupt translation-invariant properties that are desirable in this domain.

Building on these observations, this paper makes the following main contributions:

- We propose **HiT-NeXt**, a hybrid hierarchical architecture that interleaves ConvNeXt-inspired convolutional blocks and transformer blocks with restricted local attention windows, explicitly designed to capture ECG information ranging from local wave morphology to beat-level intervals and long-range rhythm patterns.
- We introduce a **contextual relative positional encoding** mechanism that combines distance-based relative position bias with Contextual Position Encoding (CoPE), preserving translation robustness while enriching attention scores with clinically meaningful temporal structure between embeddings.
- We conduct a comprehensive evaluation of HiT-NeXt on large-scale ECG datasets for both **abnormality classification** and **cardiological age prediction**, demonstrating performance improvements over strong CNN and transformer baselines and, in some metrics, over cardiology residents.
- We provide a **model development roadmap** with ablation experiments analyzing the impact of hierarchical design, convolutional patch merging, GRN, and the proposed positional encoding on overall performance, thereby offering practical guidelines for future ECG model design.

From a clinical and practical standpoint, the proposed model suggests several potential applications: (i) automated detection of common ECG abnormalities as a decision-support component in telemedicine platforms, emergency departments, and large-scale screening programs; (ii) estimation of ECG-based age as an auxiliary marker of cardiovascular risk that complements traditional clinical variables; and (iii) deployment of a single architecture capable of addressing classification and age prediction tasks, facilitating integration of deep learning models into real-world ECG workflows and longitudinal population studies. These applications, however, should be regarded as prospective: any real-world deployment would require a more robust experimental and validation setup, including prospective and external evaluations, assessment across diverse populations and recording conditions, and careful consideration of ethical, regulatory, and safety aspects before clinical use.

2. Related work

The application of artificial intelligence (AI) in ECG analysis has seen remarkable progress in recent years. A wide range of methodologies, from traditional machine learning techniques [35] to cutting-edge deep learning architectures [19], have been developed to enhance the accuracy and reliability of automated ECG interpretation. In this section, we review notable related work, focusing on papers that explore the use of convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers in advancing ECG analysis, aiming to build the technical trajectory that underpins our proposed work.

The authors in [27] introduced a deep CNN architecture trained end-to-end on a large-scale dataset, achieving cardiologist-level performance in arrhythmia detection. The model's success was attributed to a large volume of labeled data and careful design choices in the convolutional layers, prioritizing large kernel sizes to effectively capture long-range dependencies in the ECG signals and extract clinically relevant features.

Building on this, Ribeiro et al. [28] proposed a new ResNet architecture to analyze 12-lead ECGs for the automatic diagnosis of various cardiac abnormalities. In addition, this architecture was also employed to predict patient age from ECG signals, demonstrating that predicted age significantly higher than chronological age is correlated with increased mortality [17].

Regarding arrhythmia classification, Wang and Li [34] proposed a hybrid approach combining CNNs with bidirectional long short-term memory (BiLSTM) layers. This architecture leveraged CNNs for spatial feature extraction and BiLSTM networks to capture temporal dependencies, providing a robust solution for the accurate detection of atrial fibrillation.

To incorporate transformer-based architectures into ECG analysis, recent works have explored the synergy of convolutional layers for initial feature extraction and transformers for modeling global relationships. For instance, the authors in [12] proposed the ECG DETR model, which combines a CNN backbone with a transformer to reformulate arrhythmia classification as an object detection task, enabling simultaneous heartbeat localization and classification. This approach eliminates the need for explicit segmentation and leverages the self-attention mechanism to capture inter-heartbeat dependencies effectively. Similarly, a hybrid CNN-transformer model was developed in [7], where the CNN layers extract spatial features, and a transformer encoder learns temporal relationships across ECG segments.

Building upon the success of [20], the authors in [16] introduced a novel approach designed to leverage the inherent periodic structure of ECG signals using a beat-aligned framework (BaT). BaT employs local attention mechanisms and processes data progressively. This involves breaking down the input ECG into segments or "beats," applying self-attention locally within each beat window, and progressively merging these representations to capture both local and global features. However, the model has the limitation of relying on complex data preprocessing steps. Accurate beat alignment and segmentation are required for the model to function, potentially introducing preprocessing-related biases or errors, as variations in ECG morphology, noise, or irregular beats can disrupt accurate segmentation and alignment.

Drawing from the advancements of previous ECG models and harnessing recent breakthroughs in deep learning, this work proposes a novel approach to electrocardiogram analysis. The proposed model adopts a hierarchical transformer architecture while incorporating state-of-the-art concepts from CNNs and a contextual positional encoding mechanism.

3. ECG background, target abnormalities, and artifacts

An electrocardiogram (ECG) is a non-invasive exam that records the heart's electrical activity over time using electrodes placed on the body surface. In a standard 12-lead ECG, the electrical potentials are measured from different spatial orientations, providing complementary views of atrial and ventricular depolarization and repolarization. Each heartbeat is typically represented by a P wave (atrial depolarization), a QRS complex (ventricular depolarization), and a T wave (ventricular repolarization), along with clinically relevant intervals such as the PR interval (from the onset of the P wave to the onset of the QRS complex), the QT interval (from the onset of the QRS complex to the end of the T wave), and the RR interval (time between consecutive R peaks). Changes in the morphology of these waves, the duration of intervals, and the regularity of the rhythm are key indicators of a wide range of cardiac conditions.

In this study, we analyze standard 12-lead digital ECG exams labeled with six common diagnostic categories: first-degree atrioventricular block (1dAVb), right bundle branch block (RBBB), left bundle branch block (LBBB), sinus bradycardia (SB), atrial fibrillation (AF), and sinus tachycardia (ST). These labels are used as binary targets for the abnormality classification task. A detailed description of the dataset characteristics, label prevalence, and the specific protocol used for training and evaluation is provided later in Section 5.

3.1. Target ECG abnormalities

The six diagnostic categories considered in this work cover conduction disturbances, rhythm alterations, and rate abnormalities commonly reported in clinical practice.

First-degree atrioventricular block (1dAVb). First-degree atrioventricular block is a delay in atrioventricular conduction, classically defined by a prolonged PR interval (typically > 200 ms in adults) with preserved 1:1 conduction between P waves and QRS complexes. Waveform morphology is usually normal, and the abnormality is predominantly expressed as a systematic shift in intra-beat timing.

Right bundle branch block (RBBB). Right bundle branch block reflects delayed activation of the right ventricle due to impaired conduction in the right bundle. The ECG shows a widened QRS complex with characteristic rsR' or rSR' patterns in right precordial leads and broad S waves in lateral leads. It is therefore primarily a morphological disturbance of the QRS complex with relatively preserved rhythm.

Left bundle branch block (LBBB). Left bundle branch block arises from delayed conduction through the left bundle branch, leading to late activation of the left ventricle. It is characterized by QRS prolongation, broad or notched R waves in lateral leads (I, aVL, V5–V6), and deep S waves in right precordial leads, often accompanied by secondary repolarization changes. LBBB reflects a marked alteration of ventricular depolarization and is frequently associated with structural heart disease.

Sinus bradycardia (SB). Sinus bradycardia is a sinus rhythm with a reduced heart rate, generally below 60 beats per minute in adults, with otherwise normal P wave, PR interval, and QRS morphology. Its main manifestation is longer RR intervals and slower overall rhythm, which can be physiological or related to sinus node dysfunction or increased vagal tone.

Atrial fibrillation (AF). Atrial fibrillation is a supraventricular arrhythmia with disorganised atrial activity and irregular ventricular response. On the ECG, organised P waves are absent and replaced by fine or coarse fibrillatory activity, while RR intervals display an "irregularly irregular" pattern with nearly normal QRS morphology. AF thus combines loss of atrial wave organisation with long-range irregularity in inter-beat timing and is associated with increased cardiovascular risk.

Sinus tachycardia (ST). Sinus tachycardia is a sinus rhythm with an elevated heart rate, typically above 100 beats per minute in adults. The ECG shows normal P waves preceding each QRS complex and normal PR and QRS durations, but with shortened RR intervals. As in sinus bradycardia, the abnormality is mainly a global rate and timing change rather than a major alteration of waveform morphology.

3.2. ECG artifacts and challenges for automated analysis

In routine clinical practice, ECG recordings are often affected by non-cardiac artifacts that distort the underlying electrical activity. Common sources include baseline wander due to respiration and patient motion, powerline interference from the electrical grid, muscle (electromyographic) noise, and transient disturbances related to poor electrode contact or lead disconnection. These effects may shift the baseline, introduce oscillatory high-frequency components, or generate abrupt deflections superimposed on the true cardiac signal.

Such artifacts can interfere with both morphological and rhythm-based analysis. Distortions of the QRS complex or T wave can hinder the detection of bundle branch blocks and repolarization changes, while irregular baseline fluctuations or spurious peaks can mimic or obscure rhythm irregularities, complicating the identification of atrial fibrillation, sinus bradycardia, or sinus tachycardia. Robust ECG analysis, therefore, requires models and pre-processing strategies that tolerate these perturbations and preserve clinically relevant information across varying levels of signal quality.

4. Methods

Effective automated analysis of ECGs depends on learning rich feature representations that organise this information across multiple temporal scales. Instead of treating each beat or segment in isolation, a model must jointly encode (i) local descriptors of signal shape and sharp transitions, (ii) short-range temporal relations that govern intra-beat and beat-to-beat timing, and (iii) longer-range patterns that reflect sustained changes in rate or rhythm over several seconds. From a representation-learning perspective, this calls for a hierarchical feature extractor that can progressively aggregate short-window information into increasingly abstract, context-aware embeddings, preserving both fine-grained morphology and global temporal structure.

4.1. HiT-NeXt architecture

The HiT-NeXt model is designed to capture these hierarchical patterns in ECG data by employing a hybrid architecture that interleaves convolutional and transformer blocks across multiple stages. The model operates on multiple temporal and abstraction scales, progressively learning local and global patterns from the data.

Fig. 1 provides an overview of the HiT-NeXt model architecture. At each stage, convolutional blocks, referred to as patch merging blocks, are responsible for extracting and aggregating local features while reducing the dimensionality of the signal. These blocks capture fine-grained morphological information from small waveform segments, such as the shape of the QRS complex or P waves. After local feature extraction, transformer blocks model contextual representations by learning the relationships between these segments. To ensure the model remains focused on localized patterns relevant to ECG analysis, the attention mechanism is restricted to a fixed window size, ensuring that embeddings are contextualized based only on nearby segments.

This process of alternating convolutional and transformer layers continues across multiple stages. Each successive stage reduces the temporal resolution while expanding the receptive field, progressively capturing higher-order patterns such as the relationships between full heartbeats. This hierarchical approach ensures that local waveform features and global heart rate patterns are effectively modelled for precise ECG analysis.

The model processes input data in the shape `(B, seq_len, num_leads)` and progressively transforms it while preserving relevant local and global features, where `B` is the batch size, `seq_len` represents the number of temporal steps in the ECG signal, and `num_leads` corresponds to the number of ECG leads.

At each convolutional block, referred to as a patch merging block, the spatial dimension is reduced by a factor of 4, and the number of channels is doubled. Transformer blocks do not alter the dimensionality. After four stages, global average pooling is applied, followed by an MLP that outputs logits in the shape `(B, num_classes)`.

4.1.1. Patch merging

The patch merging process employs convolution layers to aggregate local information and reduce the dimensionality of the data. Each patch merging block consists of two residual and non-linear sub-blocks. Both sub-blocks follow the insights presented in [21,36], which present the new state-of-the-art for image classification, outperforming transformer-based models. This is achieved primarily through the use of an inverted bottleneck architecture together with the use of large kernel sizes to capture long-range dependencies effectively. This design contrasts with the traditional bottleneck structure used in networks like ResNet, where the data flow follows a compression-expansion process: a dimensionality-reducing convolution is applied first, followed by a larger convolution to expand the feature space. In the inverted bottleneck approach, however, the process is reversed. The tensor dimensions are initially expanded to enrich the representation, followed by a convolution with a large kernel size to capture broader spatial patterns, and finally compressed back using a 1x1 convolution. This structure enables richer representation learning, as the intermediate representations operate in a higher-dimensional space, allowing the model to better capture complex patterns and relationships.

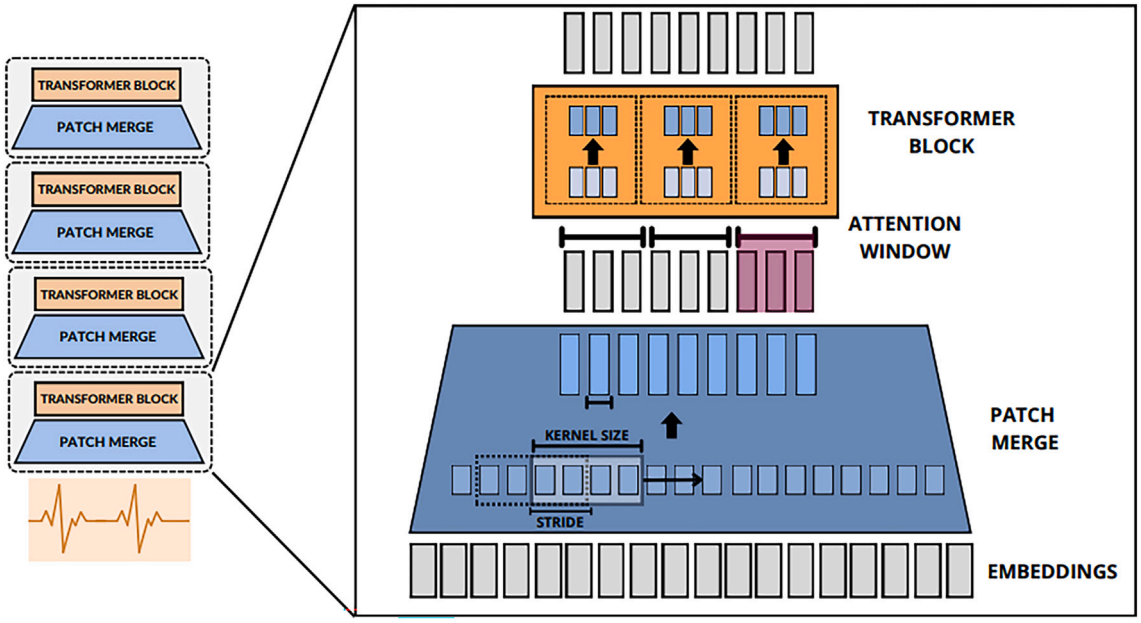


Fig. 1. High-level architecture of the HiT-NeXt model.

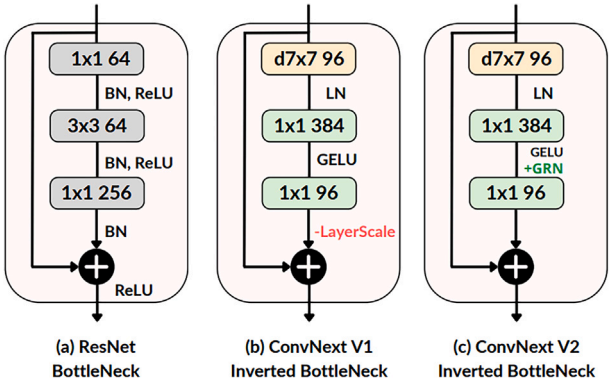


Fig. 2. Comparison between the classical ResNet bottleneck and the inverted bottleneck, first and second versions of ConvNeXt. Note that the values in the images are arbitrary for an intuitive example; in our work, we made small modifications to kernel size values, and the embedding dimensions change at each stage of the hierarchical model. However, the main idea of expansion-compression remains.

To provide a clear view of how temporal segments are progressively reduced and transformed before entering the hierarchical blocks, the patch merging module is summarized in Fig. 3. The first sub-block has a main branch consisting of convolutional layers interspersed with Layer Normalization, GELU activation as the non-linearity, and dropout for regularization. The process begins with a convolutional layer using a kernel size of 10, a stride of 4, and padding of 4, effectively reducing the number of embeddings by a factor of four. This layer takes input embeddings with dimension *in_c* and outputs embeddings with dimension *out_c*, where $out\_c = 2 \times in\_c$, followed by Layer Normalization to stabilize training. Subsequently, a 1×1 convolutional layer is applied to expand the feature dimensions by a factor of four, aligning with the inverted bottleneck principle. This expansion is followed by a GELU activation for enhanced expressiveness and dropout for improved generalization. This process is illustrated in Fig. 2.

Then, as presented in [36], we incorporate a Global Response Normalization (GRN) layer into our model. GRN is a normalization technique designed to enhance feature diversity in neural networks, consisting of three main steps: (1) *global feature aggregation*, where the L2 norm is used to calculate the magnitude of each channel, synthesizing global statistics; (2) *divisive normalization*, which adjusts these magnitudes relative to the channels, promoting competition and preventing the collapse of redundant features; and (3) *feature response calibration* by weighting, where the input features are adjusted based on the normalized magnitudes, ensuring that the diversity and relevance of the features are maintained.

To explicitly define how the Global Response Normalization (GRN) layer regulates channel-wise activations, the three aforementioned steps are applied to the input tensor as follows:

1. **Global feature aggregation**:

$$G(X)_i = \|X_i\|_2 \tag{1}$$

2. **Divisive normalization**:

$$N(G(X)_i) = \frac{\|X_i\|_2}{\sum_{j=1}^{\dim} \|X_j\|_2} \tag{2}$$

3. **Feature response calibration**:

$$X_i = \gamma \times X_i \times N(G(X)_i) + \beta + X_i \tag{3}$$

where X is the input tensor with shape $(B, \text{seq\_len}, \text{dim})$, where B is the batch size, seq_len represents the temporal dimension of the ECG signal, and dim is the feature dimension after transformation.

By normalizing feature magnitudes and adjusting them in a competitive manner, GRN helps mitigate feature collapse and reduces redundancy by balancing feature activations, ensuring that no single feature dominates the representation.

Finally, we use a final convolutional layer responsible for returning the embeddings to the dimension *out_c*, finalizing the main branch with the compression stage of the inverted bottleneck (expansion-compression).

The first sub-block also has a residual branch formed by a MaxPool layer, which reduces the number of embeddings by 4, and a 1x1 convolution, solely responsible for mapping the dimensions from *in_c* to *out_c*,

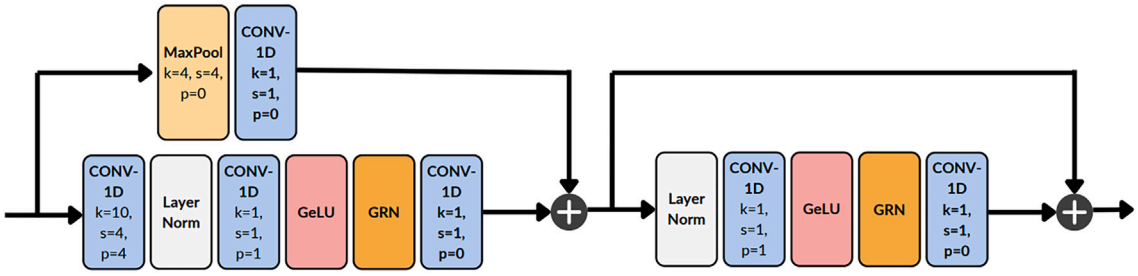


Fig. 3. Proposed convolutional patch merging architecture, outlining the sequence of operations used to reduce temporal resolution while enriching local feature representations.

taking the proposed architecture in [28] as inspiration. The result of the main branch is added to the result of the residual branch, as in a common residual network process.

The second sub-block is similar to the first but with important changes. In this sub-block, we still maintain the inverted bottleneck architecture, but the main branch does reduce the number of embeddings. However, we keep the same architecture with 1×1 convolutions, which are responsible for establishing the expansion and compression process of inverted bottlenecks. The residual branch in the second sub-block does not perform any operation, so the residue is the sum of the data before any changes made by the second sub-block, promoting a gradient corridor due to the partial derivative of the summation.

4.1.2. Transformer blocks

The transformer blocks in our model follow the standard architecture of a transformer encoder, but with important modifications. First, the attention mechanism is restricted to a fixed window based on [20]. The attention mechanism can be summarized as a process of contextual transformation, where the vector representation of the data is adjusted according to the context: the values of one embedding are transformed by the values of the other embeddings within the same context. This process is driven by attention scores, obtained through the dot product between embeddings, which enables a dynamic transformation conditioned on the context.

By restricting the attention mechanism to a fixed window, we force the model to transform the data based only on local information, resulting in more detailed and specific embeddings. Additionally, by interleaving transformer blocks with patch merging blocks, we progressively increase the level of abstraction and temporal scale. In the final stages, even with the restricted attention window, the model operates in a global context at the last stage. This ensures that learning is hierarchical and that each stage captures relationships at different temporal scales.

In the early stages, we work with a context smaller than a heartbeat, focusing on learning morphological details and small distances between waves, such as the PR segment distance. In the second stage, each window – after the information aggregation by the patch merging block – covers a context close to a full heartbeat. In subsequent stages, the model considers a context larger than one heartbeat, allowing the learning of more comprehensive information, such as complete heartbeat characteristics and the distances between them.

Following the structure proposed in [20], we implemented a cyclic shift window process, ensuring that the model is robust to translations. The translation robustness is crucial for electrocardiograms due to the heartbeat cycle and the possibility of beats being shifted between exams, a fact that should not change the diagnosis. This also eliminates the need for manual preprocessing steps such as dividing beats, calculating peaks, or explicitly detecting P, QRS, and T waves. As a result, our model can learn directly from raw ECG data, avoiding complex and potentially error-prone heuristic methods.

4.1.3. Relative positional encoding

The self-attention mechanism in transformers is inherently permutation invariant, lacking intrinsic awareness of the sequential positions of embeddings. Traditionally, positional information is injected into the model using *absolute positional encoding* (APE), where a position-specific vector—either learned or derived from a fixed function like a sinusoidal waveform—is added to each embedding at the input layer. However, this method introduces a dependency on absolute positions, disrupting robustness to translations.

Hence, we adopt *relative positional encoding* (RPE) to preserve these invariance properties. We compute positional relations between pairs of embeddings and incorporate them directly into the attention mechanism at each transformer block. Specifically, we introduce an additional term E in the attention formula, which encodes relative positional relationships. This approach allows the model to consider spatial dependencies dynamically during attention computation. To clarify how relative positions influence the attention weights, Equation (4) explicitly adds a learnable term that injects pairwise positional relationships directly into the attention score computation.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}} + E\right)V \tag{4}$$

where:

- $Q, K \in \mathbb{R}^{B\times H\times L\times d_k}$ and $V \in \mathbb{R}^{B\times H\times L\times d_v}$ are the query, key, and value matrices, where L represents the sequence length, H is the number of attention heads and B denotes the batch size. d_k and d_v denote the dimensions of the key and value embeddings, respectively.
- $E \in \mathbb{R}^{B\times H\times L\times L}$ is the relative positional encoding matrix.

The relative positional encoding term E introduces explicit positional dependencies, enabling the model to capture relationships between different positions dynamically rather than relying on absolute positions.

In our model, we employ a combination of two different forms of RPE to enrich the positional information: a learnable relative position bias (RPB) and the Contextual Position Encoding (CoPE) mechanism [9].

4.1.4. Learnable relative position bias (RPB)

RPB encodes the relative positions between pairs of embeddings within a window of size M [20].

To efficiently represent these positional biases, we initialize a learnable vector $\hat{B} \in \mathbb{R}^{2M-1}$, where each element $\hat{B}_i$ corresponds to a specific relative position within the window. This vector captures the bias for each possible relative distance between embeddings.

During the attention computation, we construct the bias matrix B by mapping the relative positional difference $d = i - j$ between any two positions i and j to the appropriate element in $\hat{B}$. Specifically, the bias B_{ij} is obtained as:

$$B_{ij} = \hat{B}_{d+M-1}, \tag{5}$$

where the shift $M-1$ adjusts the index to ensure it falls within the valid range of $\hat{B}$, which spans from 0 to $2M-1$.

As an example, consider a window size of $M = 3$. The possible relative positions are $-2, -1, 0, 1, 2$, resulting in a learnable vector $\hat{B} \in \mathbb{R}^5$. For any pair of positions i and j, the relative position difference is $d = i - j$. The corresponding bias is then:

$$B_{ij} = \hat{B}_{d+2}, \tag{6}$$

since $M - 1 = 2$. This effectively shifts the index range from $[-2, 2]$ to $[0, 4]$, aligning with the indices of $\hat{B}$.

By learning these biases, the model can better understand how the proximity of certain features affects the overall representation. However, it is important to note that this method focuses solely on the relative distances and does not account for contextual information or features that may exist between the two embeddings.

4.1.5. Contextual position encoding (CoPE)

To incorporate richer contextual information, we employ the *Contextual Position Encoding* (CoPE) mechanism [9]. Originally proposed for textual data, CoPE allows the model to capture counts of semantic entities or features within the distance between pairs of embeddings, effectively making positional measurements context-dependent rather than solely based on absolute distance.

In the context of ECG data, this mechanism enables the model to be aware of significant features occurring between two points, such as additional waves or intervals. CoPE achieves this by dynamically determining which embeddings should contribute to the positional measurement between any two positions.

Specifically, for each query-key pair (q_i, k_j), where q_i and k_j are the embedding vectors at positions i and j respectively, CoPE introduces a gating function defined as:

$$g_{ij} = \sigma(q_i^\top k_j), \tag{7}$$

where σ is the sigmoid function, and $j < i$. The gating value g_{ij} ranges between 0 and 1, representing the degree of similarity between the query at position i and the key at position j. A value of g_{ij} close to 1 indicates a high similarity, suggesting that k_j shares important features with q_i and should be included in the positional computation. Conversely, a value close to 0 implies low similarity, indicating that k_j is less relevant to q_i for positional measurement purposes and can be largely ignored. This mechanism allows the model to focus on embeddings that are contextually similar when calculating positional relationships, thereby enhancing the attention mechanism's sensitivity to meaningful patterns in the data.

The contextual position value p_{ij} between positions i and j is then calculated by summing the gating values over the positions between j and i:

$$p_{ij} = \sum_{k=j}^{i} g_{ik}. \tag{8}$$

This summation effectively accumulates the context-dependent contributions of the embeddings between j and i. If all gating values g_{ik} are equal to 1, p_{ij} simplifies to $i - j + 1$, recovering the standard relative positions based on token count.

Unlike traditional positional encodings based on fixed positions, CoPE's position values p_{ij} are continuous due to the sigmoid gating function. To handle this, an interpolation process is used to compute the positional embeddings $e[p_{ij}]$ based on the nearest integer embeddings, ensuring smooth transitions for non-integer positions, where e is a learnable vector of size M (window size).

The key idea behind CoPE is to measure positions not just by absolute or relative distances but by the content between embeddings, making the positional encoding sensitive to the actual data. In the context of ECG signals, this means the model can, for example, measure the distance between two points based on the number of significant cardiac events (like peaks or waves) between them, rather than just the number of samples.

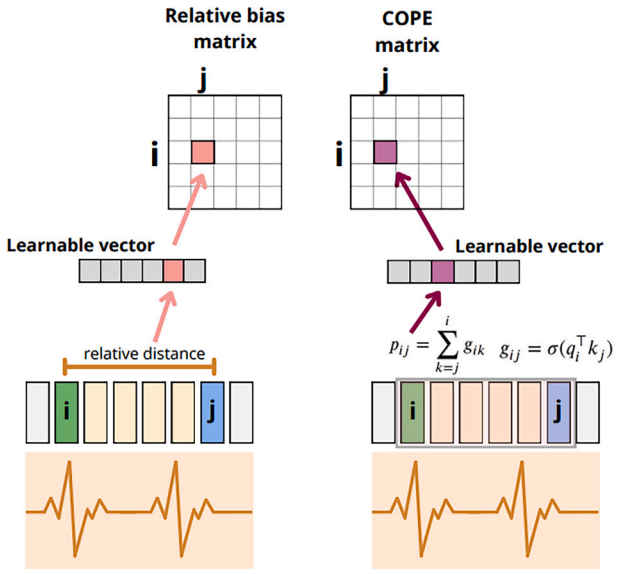


Fig. 4. High-level illustration of the proposed relative positional encoding, where pairwise distance-based bias and contextual position encoding are computed and combined to enrich the attention scores with translation-aware temporal structure.

4.1.6. Combining relative position bias and CoPE

To exploit the complementary advantages of both approaches, a learnable weight vector $\boldsymbol{\alpha} \in \mathbb{R}^2$ is introduced, allowing a linear combination of the two positional embeddings. To formalize how the two positional components are integrated into a single representation, the combined embedding is defined as follows.

$$\mathbf{E}_{\text{combined}} = \alpha_1 \mathbf{E}_{\text{CoPE}} + \alpha_2 \mathbf{E}_{\text{RPE}}. \tag{9}$$

However, directly learning $\boldsymbol{\alpha}$ without constraints can lead to one embedding type dominating the other. To mitigate this, $\boldsymbol{\alpha}$ is normalized (e.g., L2 normalization) before being applied. Concretely, if $\boldsymbol{\alpha} = [\alpha_1, \alpha_2]$, we compute:

$$\boldsymbol{\alpha}_{\text{norm}} = \frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|_2}. \tag{10}$$

Then, the combined positional embedding becomes:

$$\mathbf{E}_{\text{combined}} = \alpha_{\text{norm},1} \mathbf{E}_{\text{CoPE}} + \alpha_{\text{norm},2} \mathbf{E}_{\text{RPE}}. \tag{11}$$

To provide an intuitive view of how relative distances and contextual interactions are jointly encoded, Fig. 4 illustrates the mechanisms through which the model computes and integrates both components into the attention process.

By using a cyclic shift window in the hierarchical transformer model, as in [20], and only the two relative positional encodings, without absolute positional encodings, we ensure the model is still robust to translations. This is crucial in electrocardiograms due to the heartbeat cycle and the possibility of beats being shifted between exams, a fact that should not alter the diagnosis. Furthermore, despite the various technical details involved, once implemented, the HiT-NeXt model operates in a fully end-to-end manner, eliminating the need for manual preprocessing and feature engineering. Unlike traditional approaches that rely on wave segmentation, beat detection, and other heuristics, HiT-NeXt eliminates the need for manual preprocessing and feature engineering. The relative positional encoding is summarized in Fig. 4.

4.2. Computational complexity and inference efficiency

Let L denote the ECG sequence length. A vanilla Transformer with global self-attention has per-layer complexity $\mathcal{O}(L^2)$ with respect to L,

Table 1
Inference latency, throughput, and peak memory usage of HiT-NeXt for 12-lead ECGs of length 2560, measured on a single NVIDIA GeForce RTX 3090 Ti.

Batch size	Latency (ms)	Throughput (samples/s)	Peak memory (MB)
1	15.99	62.5	288.46
32	31.26	1023.6	381.76
128	104.45	1225.4	697.04

which is undesirable for long 1D signals such as 12-lead ECGs with $L = 2560$. In contrast, HiT-NeXt adopts a hierarchical architecture with local self-attention on fixed-size windows and patch merging. At stage s, self-attention is computed on windows of size w (independent of L), with complexity $\mathcal{O}(L_s w^2)$, where L_s is the current sequence length; patch merging reduces L_s across stages. Summing over stages yields an overall complexity that scales linearly with the input length, $\mathcal{O}(L)$, analogous to Swin-type hierarchical Transformers but instantiated for 1D ECG signals.

In the configuration used in our experiments, HiT-NeXt has 69,552,761 trainable parameters, corresponding to approximately 265.3 MB of weights in float32 precision. We empirically profiled inference on 12-lead ECGs of length 2560 using this configuration. The measurements were obtained on a single NVIDIA GeForce RTX 3090 Ti GPU. Table 1 reports latency, throughput, and peak memory for different batch sizes. The results are consistent with the linear-time analysis: the model maintains low latency for small batches and scales to more than 10^3 ECGs per second for larger batches while keeping the memory footprint below 1 GB.

5. Experiment setup

5.1. Datasets

For model development and training, we utilized the publicly available 15% sample of the CODE (Clinical Outcomes in Digital Electrocardiography) dataset [29], named CODE-15. The CODE dataset is a retrospective cohort comprising over 2 million digital electrocardiograms (ECGs) paired with mortality and hospital admission records from the state of Minas Gerais, Brazil. A team of cardiologists from the Telehealth Network of Minas Gerais analyzed the ECGs to detect six common cardiac abnormalities typically identified in ECG examinations: 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 abnormalities are important in clinical practice, as they are associated with an increased risk of cardiovascular events. They may require specific interventions and regular follow-up.

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 [28,32].

To evaluate the performance of our model, we used the publicly accessible CODE-TEST dataset, which was also collected by the Telehealth Network of Minas Gerais (TNMG). This dataset contains 827 ECG exams labeled through a rigorous consensus process involving two or three cardiology experts. The ECG diagnoses include the six previously mentioned 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). The high-quality, consensus-based labels in CODE-TEST provide a reliable benchmark for assessing model performance.

For the development and validation of the HiT-NeXt model, we adopted four non-overlapping subsets of data:

- **Training Set**: Comprising 90% of the CODE-15 dataset, this set was used to train the model.
- **Validation Set**: Consisting of 5% of the CODE-15 dataset, this set was utilized for early stopping during training to prevent overfitting.

Table 2
Distribution of the six ECG abnormalities, age, and sex in the datasets used.

Category	Variable	Train/Val/Dev (n = 345,779)	Test (n = 827)
Abnormality	1st AVB	5716 (1.7%)	28 (3.4%)
	RBBB	9672 (2.8%)	34 (4.1%)
	LBBB	6026 (1.7%)	30 (3.6%)
	SB	5605 (1.6%)	16 (1.9%)
	AF	7033 (2.0%)	13 (1.6%)
	ST	7584 (2.2%)	36 (4.4%)
Age Range	16–25	32,820 (9.5%)	43 (5.2%)
	26–40	66,729 (19.3%)	122 (14.8%)
	41–60	100,072 (28.9%)	340 (41.1%)
	61–80	112,181 (32.4%)	278 (33.6%)
	≥81	33,957 (9.8%)	44 (5.3%)
Sex	Male	206,576 (59.7%)	321 (38.8%)
	Female	139,203 (40.3%)	506 (61.2%)

- **Development Set**: Also comprising 5% of the CODE-15 dataset, this set was employed for hyperparameter tuning, model architecture decisions, and ablation studies.
- **Test Set**: The complete CODE-TEST dataset was used as the test set to evaluate the final performance of the model against baseline methods.

By partitioning the data in this manner, we ensured a rigorous and systematic approach to model development, validation, and evaluation, minimizing the risk of data leakage and providing robust performance metrics for the HiT-NeXt model. The division was performed randomly based on patient identification IDs to ensure that multiple exams from the same patient did not appear in different subsets.

Table 2 shows the class distribution of the 6 abnormalities presented in CODE-15, together with an analysis of patient age and sex. Note that some patients may exhibit more than one abnormality simultaneously, whereas others do not present any of the evaluated conditions.

5.2. Benchmarks and evaluation metrics

For comparison, we assessed HiT-NeXt against a suite of baseline models spanning diverse architectural families, including traditional convolutional neural networks (CNNs) 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, validated on the Validation 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, as well as in aggregate, to provide a detailed understanding of the model performance across different diseases.

5.3. Implementation details

The training process used the AdamW optimizer [22] and employed a cosine annealing learning rate schedule [23]. 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 terminated training if the validation error did not decrease for seven consecutive epochs. The training was conducted in parallel using 4 NVIDIA V100 GPUs.

For classification tasks, the model was trained using the Binary Cross-Entropy loss with logits (*BCEWithLogitsLoss*), having `num_classes` output neurons—one for each class. This setup allows the model to perform both multi-label and multi-class classification simultaneously. If

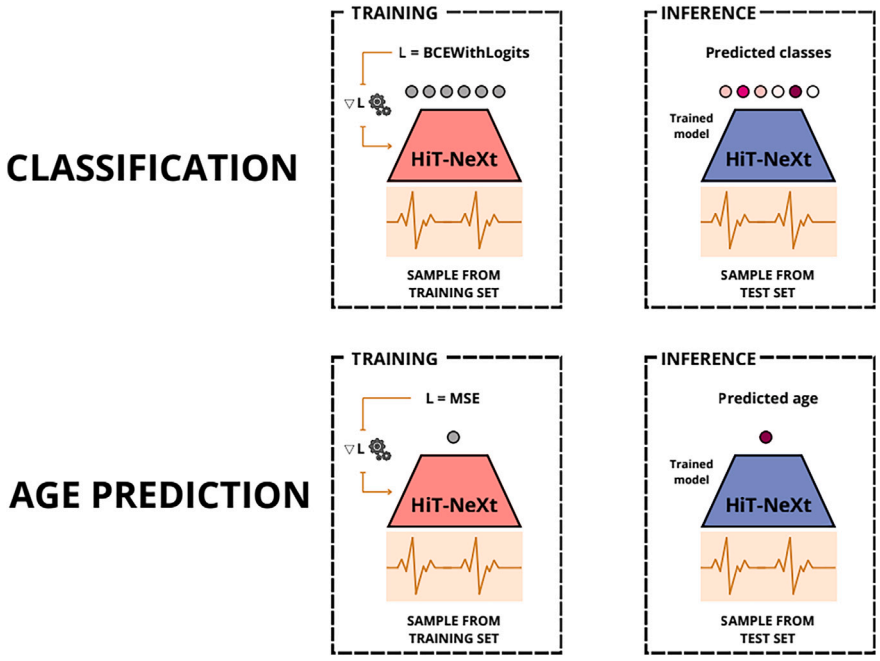


Fig. 5. Flowchart of the training and inference pipelines adopted for HiT-NeXt. The top panels correspond to the multi-label classification task, trained with BCEWithLogitsLoss and producing predicted cardiac abnormalities at inference time. The bottom panels correspond to the age regression task, trained with MSE and yielding an estimated patient age from ECG samples in the test set.

none of the classes are active, all output logits will be below the classification threshold after applying the sigmoid function, indicating the absence of all evaluated conditions.

For the age prediction task, the model was trained using Mean Squared Error (MSE) loss between the predicted and the true chronological age.

The flowchart in Fig. 5 summarizes the complete training and inference pipelines used for HiT-NeXt in both tasks. The top row depicts the multi-label classification setup, where the model is optimized with BCEWithLogitsLoss and outputs class probabilities during inference, while the bottom row illustrates the regression setup, in which the model is trained with MSE to estimate patient age from ECG signals.

6. Results

The method was tested in two scenarios: identification and classification of ECG abnormalities and age prediction, as detailed in the next sections.

6.1. Identification of ECG abnormalities

We first evaluated the performance of HiT-NeXt on the ECG abnormalities classification task, focusing on the six previously defined categories: 1st AVB, RBBB, LBBB, SB, AF, and ST. The next section introduces the evaluation metrics, followed by the results obtained for this task by HiT-NeXt and six other state-of-the-art baselines.

6.1.1. Evaluation metrics

We evaluate ECG abnormality classification using standard metrics for medical multi-label classification: accuracy, precision, recall, and F1-score. For each abnormality $k \in \{1, \ldots, K\}$ (here $K = 6$), we consider a binary decision problem and define the number of true positives (TP_k), false positives (FP_k), true negatives (TN_k), and false negatives (FN_k). The per-class metrics are given by

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k}, \qquad \text{Recall}_k = \frac{TP_k}{TP_k + FN_k}, \tag{12}$$

and the F1-score, which combines precision and recall into a single harmonic mean, is defined as

$$\text{F1}_k = \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}. \tag{13}$$

Overall accuracy is computed as

$$\text{Accuracy} = \frac{\sum_{k=1}^{K}(TP_k + TN_k)}{\sum_{k=1}^{K}(TP_k + TN_k + FP_k + FN_k)}. \tag{14}$$

In addition to reporting the metrics for each abnormality separately, we also compute macro-averaged scores by averaging Precision_k, Recall_k, and F1_k over k. This is particularly important in our setting, as the prevalence of each abnormality is imbalanced (Table 2), and macro-averaged metrics reduce the dominance of frequent classes. These metrics are widely used in ECG abnormality detection and facilitate comparison with prior work in automatic ECG classification [4,5,28,32].

6.1.2. Model performance

Table 3 presents the results of accuracy, precision, recall and F1-score obtained by the proposed model, HiT-NeXt, and 6 other networks used as baselines. HiT-NeXt, achieved the highest F1-score among all evaluated models. Despite presenting a slightly lower precision when compared to the baselines, HiT-NeXt achieved a significantly higher recall than all other methods, reaching 0.862 against 0.799 from BAT, the second-highest value. This result indicates that the model could identify a greater proportion of true cases among ECGs that present a given cardiac abnormality. Notably, this improvement in recall did not come at the cost of excessively low precision, as evidenced by the overall superior F1-score. In medical applications, recall is a particularly critical metric. It reflects the ability to detect positive cases among individuals with the condition, which is fundamental for patient care. A low recall (high false negative rate) can lead to tragic outcomes, as undiagnosed conditions may result in delayed treatment and severe consequences.

Table 4 presents the F1-scores for each class. HiT-NeXt achieved the best results for the majority of classes. In the few cases where it did

Table 3
Comparison of the performance of HiT-NeXt when compared to other SOTA methods.

Metrics	ResNet-1 [28]	ResNet-2 [11]	ECG-Transform [7]	BAT [16]	ECG-DETR [12]	HiT [5]	HiT-NeXt
Accuracy	0.991	0.989	0.981	0.991	0.9842	0.991	**0.993**
Precision	0.875	0.908	0.711	**0.918**	0.7768	0.909	0.883
Recall	0.778	0.743	0.687	0.799	0.6614	0.798	**0.862**
F1-Score	0.814	0.811	0.677	0.848	0.699	0.841	**0.872**

Table 4
Per-class f1-score of HiT-NeXt and baseline methods in the test set.

Abnormality	ResNet-1 [28]	ResNet-2 [11]	ECG-Transform [7]	BAT [16]	ECG-DETR [12]	HiT [5]	HiT-NeXt
1st AVB	0.661	0.719	0.489	0.689	0.631	0.682	**0.75**
RBBB	0.924	0.890	0.909	0.922	0.747	0.886	**0.941**
LBBB	0.927	0.843	0.886	0.945	0.826	0.909	**0.966**
SB	0.767	0.821	0.535	**0.836**	0.588	0.824	0.75
AF	0.703	0.758	0.478	0.818	0.563	0.833	**0.88**
ST	0.897	0.833	0.763	0.870	0.838	0.914	**0.944**
Avg. F1	0.814	0.811	0.677	0.848	0.699	0.841	**0.872**

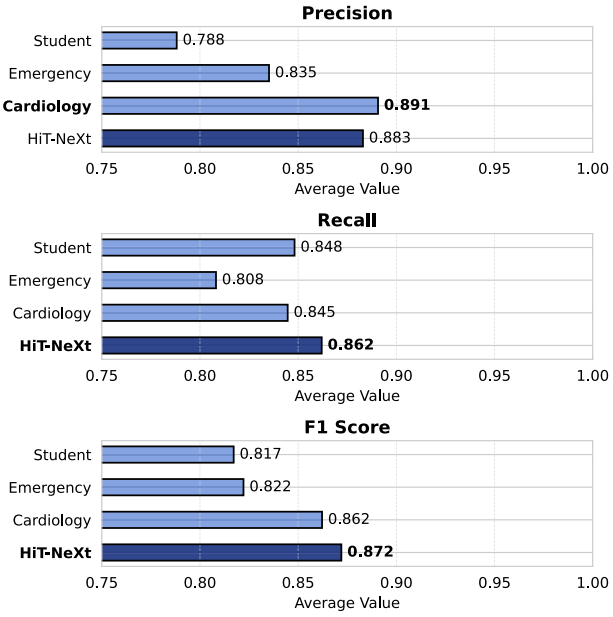


Fig. 6. Comparison of the average Precision, Recall, and F1-score between the proposed model and the evaluations from cardiology residents, emergency residents, and medical students.

not rank first, its performance remained very similar and competitive with the top-performing model, demonstrating its robustness and consistency. Notice that HiT-NeXt stands out in the 1st AVB class, where it achieved a significantly higher F1-score when compared to the baseline models. This improvement can be explained by HiT-NeXt's hierarchical local attention mechanism, which effectively captures local morphological patterns, combined with relative positional encoding, enabling the model to precisely learn temporal distances such as the PR interval length, a crucial diagnostic criterion for 1st AVB.

The CODE-TEST dataset contains labels provided by cardiologists at different levels of training: (i) 4th-year cardiology residents (cardio.), (ii) 3rd-year emergency residents (emerg.), and (iii) 5th-year medical students (stud.). The ground-truth reference used for metric calculation was established by consensus among three senior cardiologists.

Using this reference, we compared the classifications made by professionals at different training levels with those produced by the proposed HiT-NeXt model. The results, shown in Fig. 6, indicate that the model outperformed all evaluator groups in terms of recall and F1-score. In precision, the model performed worse than the cardiology residents but achieved higher precision than both medical students and emergency residents.

6.2. Age prediction

Age is one of the strongest predictors of cardiovascular disease, but chronological age does not necessarily reflect the *biological* or "cardiovascular" age of an individual. Deep neural networks trained to predict age directly from the raw 12-lead ECG can estimate an electrocardiographic age (ECG-age), and the difference between ECG-age and chronological age (Δage) has emerged as a meaningful marker of cardiovascular health [4,17].

Lima et al. [17] demonstrated this using over 1.5 million ECGs from the CODE cohort: individuals whose ECG-age was at least 8 years older than their chronological age had a substantially higher risk of all-cause mortality (hazard ratio ≈ 1.8), while those with ECG-age younger than their chronological age had a lower risk. Importantly, this association persisted after adjustment for traditional risk factors (e.g., hypertension, diabetes, smoking, dyslipidemia) and even when restricting the analysis to ECGs classified as normal by cardiologists. These findings suggest that ECG-age captures subtle electrophysiological signatures and accumulated cardiovascular damage that may not be evident through standard clinical interpretation, offering an interpretable biomarker of cardiovascular ageing.

In this study, we additionally evaluate HiT-NeXt on age prediction from ECGs.

6.2.1. Evaluation metrics

We assess age prediction using mean absolute error (MAE) and mean squared error (MSE). Given true ages y_i, predictions $\hat{y}_i$, and N test samples, we define

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|, \qquad \mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2. \tag{15}$$

MAE reflects the typical error in years and is easily interpretable, whereas MSE penalizes large deviations more heavily. Reporting both metrics provides a more complete assessment of the accuracy and robustness of ECG-age estimates.

6.2.2. Model performance

Table 5 presents the mean absolute error (MAE) and mean squared error (MSE) achieved by HiT-NeXt and all baselines. Our model attains the lowest error, indicating that HiT-NeXt estimates ECG-age with greater precision. This improved accuracy is clinically relevant, as a

Table 5
MAE and MSE of models for age prediction.

Metric	ResNet-1 [28]	ResNet-2 [11]	ECG-Transform [7]	BAT [16]	ECG-DETR [12]	HiT [5]	HiT-NeXt
MAE	11.7	10.86	10.8	13.5	11.4	10.34	**9.5**
MSE	214.6	195.2	191.3	271.9	205.6	199.8	**147.6**

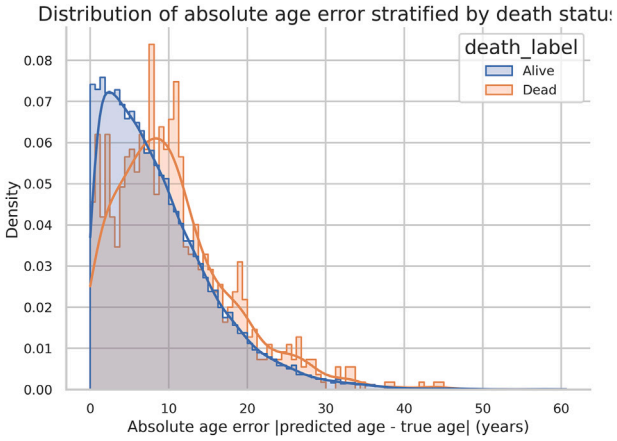


Fig. 7. Distribution of absolute age prediction error |Δage| stratified by vital status at follow-up.

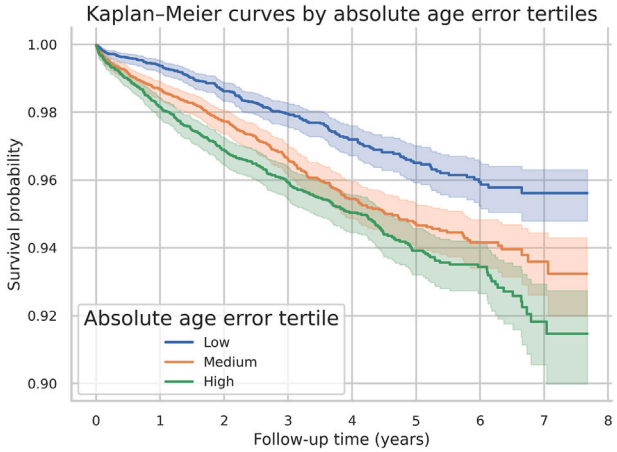


Fig. 8. Kaplan–Meier survival curves by tertiles of absolute age prediction error |Δage|.

more reliable ECG-age leads to a more informative Δage, strengthening its potential use as an auxiliary biomarker for cardiovascular risk assessment.

6.2.3. Age prediction error and mortality

Beyond evaluating the accuracy of ECG-based age prediction, we investigated whether the magnitude of the prediction error itself carries prognostic information. This type of analysis is aligned with previous studies that have related discrepancies between ECG-derived age and chronological age to all-cause mortality in adults and to the presence of relevant comorbidities in pediatric populations [4,17]. For each exam, we defined the age prediction error as

$$\Delta\text{age} = \widehat{\text{age}} - \text{age},$$

and focused on its absolute value,

$$|\Delta\text{age}| = \left|\widehat{\text{age}} - \text{age}\right|,$$

which quantifies how atypical the ECG-age is relative to chronological age, regardless of whether the ECG appears "older" or "younger". This analysis was restricted to the first exam of each patient in the validation subset of CODE-15, for which follow-up time and vital status were available, because the external CODE-TEST set does not contain mortality information; these exams were then linked to the corresponding age predictions produced by HiT-NeXt.

Fig. 7 shows the distribution of |Δage| stratified by vital status at follow-up. In both groups, the distribution is right-skewed, with most patients presenting relatively small absolute errors and a progressively thinner tail toward larger discrepancies. However, patients who died during follow-up exhibit a clear shift toward larger values of |Δage|: their density is slightly lower for small errors and relatively higher in the intermediate and upper ranges of the distribution. This pattern indicates that large deviations between ECG-age and chronological age are more common among individuals who subsequently died, suggesting that an ECG whose age signature is markedly atypical for the patient's chronological age is associated with a worse prognosis.

To further quantify this association over time, we performed a survival analysis using Kaplan–Meier curves. The Kaplan–Meier estimator is a non-parametric method that estimates the survival function $S(t) = \mathbb{P}(T > t)$ while explicitly accounting for right-censoring: at each observed death time, the curve is updated by multiplying the current survival estimate by the proportion of patients who survive at that instant among those still at risk [30]. In practice, this yields a stepwise trajectory that describes how the probability of being alive evolves over follow-up. In our setting, patients were divided into tertiles of |Δage| (low, medium and high absolute error), and we estimated a Kaplan–Meier survival curve for each group using time from the baseline ECG to death or censoring as the time-to-event variable.

As shown in Fig. 8, the resulting curves display a graded relationship: the group with the lowest absolute age error shows the highest survival probabilities throughout follow-up, the intermediate group lies in between, and the group with the largest errors consistently presents the lowest survival probabilities. This separation emerges early and is maintained over the entire observation window, indicating that increasing |Δage| is followed by a monotonic increase in cumulative mortality.

6.3. Limitations

While HiT-NeXt achieved strong results in both abnormality classification and age prediction, some aspects remain to be explored. In this work, the model was trained on CODE-15 and evaluated on CODE-TEST, both from the same Brazilian tele-ECG system. Although these datasets are large and widely adopted in the literature, future work should include external validation on additional cohorts to more thoroughly assess generalization across populations, acquisition protocols, and clinical settings.

7. Model development roadmap

In this section, we present an ablation study structured as a roadmap to explore how each component of the proposed HiT-NeXt model contributes to its final performance. We chose the roadmap format for its clarity in presenting intermediate results and the insights gained throughout the model's development. This approach not only provides a comprehensive understanding of the model's evolution but also serves as a foundation for future work in developing models for cardiology.

For all the experiments detailed in this section, we report metrics on the development set. It is important to note that the development set consistently yielded poorer results compared to the test set. This corroborates the approach reported in [28]. In this regard, the reported results represent the mean values obtained from 10,000 iterations of bootstrap resampling.

7.1. Vision transformer (ViT) based model

Arbitrarily, the model development began based on the Vision Transformer (ViT). Similar to a classic ViT, we divided the ECG signal into patches of 12 channels, corresponding to the 12 leads of the ECGs, and applied a linear projection to each patch to obtain the initial embeddings. These embeddings were then passed through consecutive transformer blocks with a global attention mechanism. Additionally, to incorporate positional information, sinusoidal absolute positional encoding was used. Patches of 16 signal measurements were used, resulting in 160 patches for a 2560-measurement signal.

For this initial model, we achieved an average F1-score of 0.535. However, we observed significantly unsatisfactory performance in identifying the cardiac conditions of first-degree atrioventricular block (1dAVb) and atrial fibrillation (AF), with 0.098 and 0.337 f1-score, respectively. These conditions are characterized by subtle details in the ECG trace, such as the prolongation of the PR interval in the case of 1dAVb, and the absence of well-defined P waves, replaced by irregular and wavy “f” waves in the case of AF.

These results suggest that the Vision Transformer (ViT)-based model, with its global attention mechanism, is not effective in capturing detailed information in small time windows and short intervals.

7.2. Hierarchical structure with restricted attention

To address the ViT limitation, we implemented a hierarchical structure inspired by the approach proposed by [20]. In this model, at each transformer block, we employed a patch merging process, initially implemented as a pooling layer, responsible for aggregating spatial and temporal information while simultaneously reducing the dimensionality of the data. This process was essential for enabling the model to operate at multiple temporal scales.

Moreover, we implemented the mechanism of attention restricted to local windows, forcing the model to focus on smaller regions of the signal at each stage. Additionally, we implemented the cyclic shift process to ensure that the model is robust to small translations in the temporal signals.

These structural changes brought significant improvements in the overall performance of the model. With the hierarchical model, we achieved a mean f1-score of 0.653. We also noticed a significant improvement in f1 for the detection of first-degree atrioventricular block (1dAVb) and atrial fibrillation (AF) abnormalities, which increased from 0.098 to 0.485 and from 0.337 to 0.641, respectively.

7.2.1. Introduction of convolutional block in patch projection

After the latest improvements, we still believed that the process of mapping the original ECG signal to a sequence of embeddings was crucial for an effective model, as it captures important information that will be transformed by the model.

To test this hypothesis, we enhanced the robustness of the patch embedding process by adopting a convolutional residual network architecture, based on the blocks proposed in [28]. The residual block was then used to project the input signal into embeddings, following a non-linear, residual-based mechanism.

With this modification, the model’s performance improved, achieving a mean F1 score of 0.656. This enhancement underscores the importance of robust embedding techniques in extracting meaningful features from the ECG signals, which in turn optimizes the model’s overall predictive capability.

7.2.2. Integration of convolutional blocks for patch merging

Building on the improvements obtained with the enhanced patch embedding process, we applied a similar approach to the patch merging stage. Specifically, we replaced the traditional pooling operations with a single-layer convolutional network designed to aggregate local information, thereby decreasing the spatial dimensions of the data.

Additionally, we rigorously tested a more robust architecture using residual and non-linear convolutional blocks, the same architecture employed in the patch embedding process. This approach allowed for more efficient aggregation of the embeddings while preserving key characteristics of the original ECG signal. As a result, the model was able to maintain the richness of the information throughout the merging process, further enhancing its ability to capture meaningful patterns from the data.

The patch merging stage using a single-layer convolutional network achieved a mean F1 score of 0.678. In comparison, the patch merging approach using the complete residual convolutional block architecture yielded a higher mean F1 score of 0.695, demonstrating the importance of employing robust blocks for information aggregation.

7.2.3. ConvNeXt-based architecture

Building on the promising results obtained thus far, additional fundamental modifications were introduced to further enhance the model’s performance. While retaining the residual structure in the convolutional blocks, the internal operations of each block were redesigned to incorporate inverted bottlenecks, large kernel sizes, layer normalization, and global response normalization, drawing inspiration from the design principles outlined in [21,36].

At this point, we also replaced the absolute positional encoding with relative positional encoding, relying solely on the relative positional distances.

These enhancements improved the model’s F1-score from 0.695 to 0.704 and showed that the inclusion of state-of-the-art architectural designs in the merging patches process is essential for attaining superior performance.

7.3. Combined contextual positional encoding

As mentioned before, initially, the method proposed in [20] was adopted, where a learnable matrix B is used, and the relative distances between embeddings serve as indices to query values from B.

Finally, the Contextual Positional Encoding (CoPE) mechanism was implemented to incorporate contextual information between embeddings into the calculation of positional encoding. The final solution combined these two approaches linearly, leveraging the strengths of both methods. This hybrid strategy further improved the average F1 score, raising it from 0.704 to 0.710.

Table 6 shows the average F1-score obtained by each type of positional encoding considered. We tested relative positional encodings by considering only distance-based relative encoding, as well as context-based encoding (COPE), in addition to the proposed combined method. For each of these three types, we evaluated both the presence and absence of absolute positional encoding.

First, note that the models without absolute positional encoding achieved better overall results. This can be explained by their resilience to translation, as they do not rely on absolute positions. The best overall result was achieved by the combined method without absolute positional encoding.

Table 6
Comparison of RPE Types with and without APE.

RPE Type	With APE [33]	Without APE
Only RPB [20]	0.698	0.704
Only CoPE [9]	0.706	0.706
Combined	0.703	**0.710**

8. Conclusions

This paper introduces HiT-NeXt, a method for analyzing ECGs that combines convolutions and transformer blocks together with state-of-the-art advances in the deep learning field to address the local and global patterns of ECG signals.

The method was tested in two tasks: classifying 6 different ECG abnormalities and patient age prediction. In both cases, HiT-NeXt was better than all the baselines in terms of f-measure and MSE. For the ECG abnormalities classification, the results were better than those obtained by medical students, showing the accuracy and robustness of the approach.

A valuable avenue for future research is to use the HiT-NeXt architecture in a self-supervised setup. By pretraining the model on extensive unlabeled ECG datasets before fine-tuning it for classification, this strategy could improve generalization and robustness, especially in detecting rare abnormalities where labeled data is scarce. Furthermore, investigating domain adaptation techniques could enhance the model's adaptability to diverse populations and varying recording conditions, ultimately broadening its clinical utility.

CRediT authorship contribution statement

Pedro Dutenhefner: Writing – original draft, Software, Methodology, Investigation, Formal analysis. **Turi Rezende:** Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization. **José Geraldo Fernandes:** Validation, Investigation. **Diogo Tuler:** Validation, Investigation. **Gabriela M.M. Paixão:** Writing – review and editing, Supervision, Funding acquisition. **Gisele Pappa:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Antônio Ribeiro:** Supervision, Funding acquisition. **Wagner Meira Jr.:** Writing – review & editing, Supervision, Funding acquisition.

Ethics statement

To address ethical concerns surrounding data privacy and patient confidentiality, this study exclusively utilized publicly available datasets that were fully anonymized and managed in accordance with ethical standards. No personally identifiable information was accessed or used, aligning with data protection regulations and reinforcing the commitment to responsible AI research in healthcare.

Funding

This work was partially funded by CNPq, CAPES, FAPEMIG, and CIIA-Saúde.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the Telehealth Center of Minas Gerais for providing access to the data and for their collaboration in the context of this work. This work was partially funded by CNPq, CAPES, FAPEMIG, and CIIA-Saúde.

References

[1] S. Ahmed, I.E. Nielsen, A. Tripathi, S. Siddiqui, G. Rasool, R.P. Ramachandran, Transformers in time-series analysis: A tutorial. arxiv 2022, arXiv preprint arXiv:2205.01138, 2022.

[2] B. Chen, Y. Li, N. Zeng, Centralized wavelet multiresolution for exact translation invariant processing of ECG signals, IEEE Access 7 (2019) 42322–42330.

[3] A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, 2020.

[4] P.R. Dutenhefner, G. Lemos, T. Rezende, J.G. Fernandes, D. Tuler, G.L. Pappa, G.M. Paixao, A.L.P. Ribeiro, W. Meira Jr, Ecg-Resnext: age prediction in pediatric electrocardiograms and its correlations with comorbidities, in: Encontro Nacional De Inteligência Artificial E Computacional (ENIAC), SBC, 2024a, pp. 49–60.

[5] P.R. Dutenhefner, T.A.V. Rezende, G.L. Pappa, G.M. de Matos Paixão, A.L.P. Ribeiro, W. Meira Jr, Um transformer hierárquico para classificação e diagnóstico de eletrocardiograma, J. Health Inform. 16 (2024b).

[6] Z. Ebrahimi, M. Loni, M. Daneshtalab, A. Gharehbaghi, A review on deep learning methods for ECG arrhythmia classification, Expert Syst. Appl.: X 7 (2020) 100033.

[7] H. El-Ghaish, E. Eldele, Ecgtransform: empowering adaptive ECG arrhythmia classification framework with bidirectional transformer, Biomed. Signal Process. Control 89 (2024) 105714.

[8] N. Fujita, A. Sato, M. Kawarasaki, Performance study of wavelet-based ECG analysis for st-segment detection, in: 2015 38th International Conference on Telecommunications and Signal Processing (TSP), IEEE, 2015, pp. 430–434.

[9] O. Golovneva, T. Wang, J. Weston, S. Sukhbaatar, Contextual position encoding: Learning to count what's important, arXiv preprint arXiv:2405.18719, 2024.

[10] V. Gupta, Wavelet transform and vector machines as emerging tools for computational medicine, J. Ambient Intell. Humanized Comput. 14 (2023) 4595–4605.

[11] A. Hannun y, P. Rajpurkar, M. Haghpanahi, G.H. Tison, C. Bourn, M.P. Turakhia, A. Ng y, Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network, Nat. Med. 25 (2019) 65–69.

[12] R. Hu, J. Chen, L. Zhou, A transformer-based deep neural network for arrhythmia detection using continuous ECG signals, Comput. Biol. Med. 144 (2022) 105325.

[13] A.K. Ka, Ecg beats classification using waveform similarity and rr interval, arXiv preprint arXiv:1101.1836, 2011.

[14] S. Kuila, N. Dhanda, S. Joardar, ECG signal classification and arrhythmia detection using elm-rnn, Multimed. Tools Appl. 81 (2022) 25233–25249.

[15] S. Li, C. Wu, N. Xiong, Hybrid architecture based on CNN and Transformer for strip steel surface defect classification, Electronics 11 (2022) 1200, https://doi.org/10.3390/electronics11081200

[16] X. Li, C. Li, Y. Wei, Y. Sun, J. Wei, X. Li, B. Qian, Bat: beat-aligned transformer for electrocardiogram classification, in: 2021 IEEE International Conference on Data Mining (ICDM), IEEE, 2021, pp. 320–329.

[17] E.M. Lima, A.H. Ribeiro, G.M. Paixão, M.H. Ribeiro, M.M. Pinto-Filho, P.R. Gomes, D.M. Oliveira, E.C. Sabino, B.B. Duncan, L. Giatti, et al., Deep neural network-estimated electrocardiographic age as a mortality predictor, Nat. Commun. 12 (2021) 5117.

[18] C.-C. Lin, C.-M. Yang, Heartbeat classification using normalized RR intervals and morphological features, Math. Probl. Eng. 2014 (2014) 712474.

[19] X. Liu, H. Wang, Z. Li, L. Qin, Deep learning in ECG diagnosis: a review, Knowl.-Based Syst. 227 (2021a) 107187.

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021b, pp. 10012–10022.

[21] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.

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

[23] I. Loshchilov, F. Hutter, Sgdr: Stochastic gradient descent with warm restarts, arXiv preprint arXiv:1608.03983, 2016.

[24] J.P. Martínez, R. Almeida, S. Olmos, A.P. Rocha, P. Laguna, A wavelet-based ECG delineator: evaluation on standard databases, IEEE Trans. Biomed. Eng. 51 (2004) 570–581.

[25] R.J. Martis, U.R. Acharya, L.C. Min, ECG beat classification using PCA, LDA, ICA and discrete wavelet transform, Biomed. Signal Process. Control 8 (2013) 437–448.

[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[27] P. Rajpurkar, A.Y. Hannun, M. Haghpanahi, C. Bourn, A.Y. Ng, Cardiologist-level arrhythmia detection with convolutional neural networks. arxiv 2017, arXiv preprint arXiv:1707.01836, 2011.

[28] A.H. Ribeiro, M.H. Ribeiro, G.M. Paixão, D.M. Oliveira, P.R. Gomes, J.A. Canazart, M.P.S. Ferreira, C.R. Andersson, P.W. Macfarlane, W. Meira Jr, et al., Automatic diagnosis of the 12-lead ECG using a deep neural network, Nat. Commun. 11 (2020) 1760.

[29] A.L.P. Ribeiro, G.M. Paixao, P.R. Gomes, M.H. Ribeiro, A.H. Ribeiro, J.A. Canazart, D.M. Oliveira, M.P. Ferreira, E.M. Lima, J.L. de Moraes, et al., Tele-electrocardiography and bigdata: the code (clinical outcomes in digital electrocardiography) study, J. Electrocardiol. 57 (2019) S75–S78.

[30] J.T. Rich, J.G. Neely, R.C. Paniello, C.C. Voelker, B. Nussenbaum, E.W. Wang, A practical guide to understanding Kaplan-Meier curves, Otolaryngol.-Head Neck Surg. 143 (2010) 331–336.

[31] G. Ruffini, D. Ibanez, M. Castellano, S. Dunne, A. Soria-Frisch, Eeg-driven RNN classification for prognosis of neurodegeneration in at-risk patients, in: Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part I 25, Springer, 2016, pp. 306–313.

[32] D. Tuler, P.R. Dutenhefner, J.G. Fernandes, T. Rezende, G. Lemos, G.L. Pappa, G. Paixao, A. Ribeiro, W. Meira Jr, Leveraging cardiologists prior-knowledge and a mixture of experts model for hierarchically predicting ECG disorders, in: Computing in Cardiology (CinC), vol. 51, 2024, pp. 1–4, https://doi.org/10.22489/CinC.2024.4773.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems 30 (NIPS 2017), Curran Associates, Inc., 2017, pp. 5998–6008.
[34] J. Wang, W. Li, Atrial fibrillation detection and ecg classification based on cnn-bilstm. arxiv 2020, arXiv preprint arXiv:2011.06187, 2020.
[35] M. Wasimuddin, K. Elleithy, A.-S. Abuzneid, M. Faezipour, O. Abuzaghleh, Stages-based ECG signal analysis from traditional signal processing to machine learning approaches: a survey, IEEE Access 8 (2020) 177782–177803.
[36] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I.S. Kweon, S. Xie, Convnext v2: co-designing and scaling convnets with masked autoencoders, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 16133–16142.
[37] World Health Organization, Cardiovascular diseases (CVDs), 31 July 2025, https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (Accessed 14 January 2026).