Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Advancing interpretable cardiac disease diagnosis via a transformer-convolutional hybrid network on electrocardiograms

Xiaoqiang Liu [a,1], Yinlong Xu [b,1], Hongxia Xu [c,*], Liang He [d], Siyu Long [d], Yisen Huang [a], Yubin Wang [a], Yingzhou Lu [e], Yingxuan Huang [a], Jian Wu [f], Honghao Gao [g,*], Xiaobo Liu [a,d,*]

[a] First Hospital of Quanzhou Affiliated to Fujian Medical University, Quanzhou, China
[b] College of Computer Science & Technology, Zhejiang University, Hangzhou, China
[c] Innovation Institute for Artificial Intelligence in Medicine, Zhejiang University, Hangzhou, China
[d] McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada
[e] School of Medicine, Stanford University, Stanford, USA
[f] School of Public Health, Zhejiang University, Hangzhou, China
[g] School of Computer Engineering and Science, Shanghai University, Shanghai, China

ARTICLE INFO

Keywords:
Deep learning neural network
Electrocardiogram
Classification
Automatic diagnosis
Transformer
Explainable Artificial Intelligence

ABSTRACT

Manual heart disease diagnosis with the electrocardiogram (ECG) is intractable due to the intertwined signal features and lengthy diagnosis procedure, especially for the 24-hour dynamic ECG signals. Consequently, even experienced cardiologists may face difficulty in producing all accurate ECG reports. In recent years, Artificial Intelligence (AI), particularly neural network-based automatic ECG diagnosis methods have exhibited promising performance, suggesting a potential alternative to the labor-intensive examination conducted by cardiologists. However, many existing approaches failed to adequately consider the temporal and channel dimensions when assembling features and ignored interpretability. And clinical theory underscores the necessity of prolonged signal observations for diagnosing certain ECG conditions such as tachycardia. Moreover, specific heart diseases manifest primarily through distinct ECG leads represented as channels. In response to these challenges, this paper introduces a novel neural network architecture for ECG classification (diagnosis). The proposed model incorporates Lead Fusing blocks, transformer-XL (meaning extra long) encoder-based Encoder modules, and hierarchical temporal attentions. Importantly, this classifier operates directly on raw ECG time-series signals rather than cardiac cycles. Signal integration begins with the Lead Fusing blocks, followed by the Encoder modules and hierarchical temporal attentions, enabling the extraction of long-dependent features. Furthermore, existing convolution-based methods have been argued to compromise interpretability, whereas the proposed neural network provides improved clarity in this regard. Experimental evaluations on a comprehensive public dataset confirm the superiority of the proposed classifier over state-of-the-art methods. Moreover, a visualization method was employed to generate a location map that demonstrates the areas of the signal emphasized by the model, thereby enhancing interpretability.

1. Introduction

Heart diseases, such as tachycardia (Ganz and Friedman, 1995), atrial fibrillation (Wijesurendra and Casadei, 2019) and atrial premature beats (Gladstone et al., 2015), pose significant life-threatening risks. Diagnosis of these conditions relies on electrocardiogram (ECG) signals (Sörnmo and Laguna, 2006) recorded by specialized monitoring devices. However, the interpretation of ECG signals for diagnoses, particularly conditions like atrial fibrillation, is a labor-intensive and intricate process. This complexity arises from the inherent noise present in ECG signals and the fact that features of heart diseases are often subtle and entangled within the ECG data. Analogous to the manual analysis of various time-series data, ECG signal interpretation necessitates a meticulous examination of diverse waveforms (Natarajan et al., 2020). This process lacks the intuitive clarity afforded by the analysis of two-dimensional images. Consequently, there is a pressing need for accurate ECG signal classification methods tailored to specific heart diseases.

In recent times, increased attention has been directed towards methods for automating heart disease diagnosis. Notably, automatic diagnosis (classification) methods (Chen et al., 2020b; Pourbabaee et al.,

[*] Corresponding authors.
E-mail addresses: Einstein@zju.edu.cn (H. Xu), gaohonghao@shu.edu.cn (H. Gao), xiaobo.liu@mail.mcgill.ca (X. Liu).
[1] Co-first authors.

https://doi.org/10.1016/j.engappai.2025.110675
Received 11 August 2024; Received in revised form 13 March 2025; Accepted 23 March 2025
Available online 11 April 2025

2018; Golany et al., 2020; Kachuee et al., 2018; Bian et al., 2022; Azar et al., 2012; Emary et al., 2014) have demonstrated significant performance advancements leveraging machine learning techniques. Initially, automatic ECG diagnosis methods involved a phase of feature extraction through statistical summarizations (Sayadi et al., 2009; Inan et al., 2006), followed by a classification stage employing classifiers such as the support vector machine (Faziludeen and Sabiq, 2013) or the multi-layer neural network (Inan et al., 2006). While manually engineered features may offer reasonable insights into signal analysis, the paradigm of automated feature extraction has shown to be more adept at ensuring accurate classification (Chen et al., 2020b; Rajpurkar et al., 2017) and with the development of the deep learning, the features extracted from numerous neural networks have demonstrated significant superiority over traditional machine learning methods. End-to-end neural networks (Acharya et al., 2017; Jin and Dong, 2017; Kiranyaz et al., 2015; Chen et al., 2020b; Golany et al., 2020) have been effectively applied to ECG signal classification, seamlessly integrating feature extraction and classification within a unified framework. Rajpurkar et al. (2017) introduced a straightforward neural network for ECG classification, achieving performances comparable to human cardiologists. Several other neural network approaches (Kachuee et al., 2018; Chen et al., 2020b; Golany et al., 2020; Chen et al., 2020a) have been proposed, each offering unique perspectives, such as data augmentation, multi-label regularization, and robust feature extraction.

Although prior works have achieved high precision in ECG signal classification, they often neglected the modeling of long-term dependencies in ECG signals and the interpretability of their results. The long-term depending is particularly critical when dealing with extended time-series data, such as the continuous 24-hour signals recorded by dynamic ECG monitoring. Neglecting long-dependent features can hinder the accurate diagnosis of heart diseases. For instance, tachycardia (Ganz and Friedman, 1995; Cierpka-Kmieć and Hering, 2020), defined as a state with an average heart rate exceeding 100 beats per minute, requires extended observation for diagnosis. Clinical practice also demands the scrutiny of several cardiac cycles to diagnose atrioventricular block. Furthermore, mainstream methods have several cardiac cycles to diagnose the atrioventricular block. Additionally, the known mainstream methods extracted features by convolutions that are essentially weighted moving average operations (Acharya et al., 2017; Chen et al., 2020b). Summary features were extracted through the average, but the exact location information is vanished. Meanwhile, the interpretability of the results is an important reference in the actual diagnosis process, a model with high performance but low weak interpretability will not be so acceptable in diagnosis field as in others.

To satisfy the demand of the high precision in long-term signals, this paper presents a novel neural network framework designed for ECG signal classification, incorporating both local feature extraction and long-dependence capture. After initial pre-processing steps, such as de-noising, the ECG signals are input into a series of Lead Fusing blocks. These blocks consist of multiple 1D convolutions with ReLU activation (Agarap, 2018) and incorporate the Squeeze-and-Excitation Attention (SE-Attention) mechanism (Hu et al., 2018). Their purpose is to extract time-aligned features and fuse the ECG signals in the channel dimension. Following this initial feature extraction stage, the processed data is then forwarded through two stacked Encoder modules. Each Encoder module employs a sliding window approach to partition the learned time-aligned features into smaller local and real-time segments in the length dimension. The Transformer-XL encoder, a variant of the Transformer architecture (Dai et al., 2019), is then applied to these segments to extract non-local features. Notably certain diseases can only be accurately diagnosed by closely observing specific cardiac cycles. To address this challenge, this paper introduces a hierarchical temporal attention module that plays a crucial role in discerning critical non-local features and acquiring a comprehensive understanding of global features. Subsequently, these global features are directed into a fully connected layer for classification, following a methodology similar to prior research (Chen et al., 2020b; Rajpurkar et al., 2017). Experimental results demonstrate our proposed neural classifier surpasses the performance of previous state-of-the-art methods on a public ECG dataset comprising a large collection of multi-lead ECG signals.

Our method also enhances the interpretability of our model by applying a visualization technique GradCAM (Selvaraju et al., 2017), a technique that generates a location map by calculating the gradient flow of the target concept to the final convolution layer, highlighting areas of the image that are important for the prediction concept. The results indicate that, compared to previous convolution-based methods, the proposed framework employs fewer convolutions while demonstrating excellent performance in focusing on critical classification features.

This work presents several significant contributions:

- This paper introduces a novel neural classifier for ECG signal classification (diagnosis), which leverages the Transformer-XL encoder to extract non-local features and model long-term dependencies effectively. The proposed model also incorporates a hierarchical temporal attention module to consolidate the features derived from the Encoder modules and predict global features.
- Experimental evaluations conducted on the public Tianchi ECG dataset confirm the superior performance of our proposed neural classification method in comparison to previous state-of-the-art approaches. The higher precision shows better application potential in real diagnosis scenario.
- Interpretability is improved through minimizing the use of convolutions in the processing of ECG data. The visualization underscores the contents of the model focuses and thus enhances the interpretability of the model. The location of the specific wave concerned by the model can intuitively explain the reason for the prediction of the model.

2. Related work

2.1. Deep learning for ECG disease diagnosis

ECG-based cardiac diagnosis has evolved from manual feature engineering to end-to-end deep learning. Early methodologies leveraged handcrafted features such as QRS complex duration and ST-segment elevation, combined with classifiers like SVM (Venkatesan et al., 2018) or random forests (Kropf et al., 2017). While offering interpretability, these approaches exhibited limited robustness to noise and struggled to detect subtle pathological patterns in long-term recordings.

The introduction of 1D CNNs (Dutta and Das, 2024; Li et al., 2017) marked a paradigm shift by enabling automated feature extraction directly from raw ECG signals. However, conventional CNNs remain constrained by their local receptive fields, hindering their ability to model long-range dependencies essential for episodic conditions like paroxysmal atrial fibrillation (Feyisa et al., 2022). Subsequent efforts to expand receptive fields through dilated convolutions (Zhang et al., 2024; Hejc et al., 2024) introduced architectural complexity without resolving fundamental limitations in global context modeling. To address temporal dependencies, Recurrent Neural Networks (RNNs) (Kuila et al., 2022; Wang et al., 2023) were explored, yet their susceptibility to gradient vanishing effects impeded effective learning from hour-long ECG sequences. Hybrid CNN-RNN architectures (Liang and Lu, 2023) improved short-term rhythm analysis but proved computationally intractable for real-time applications like Holter monitoring. These limitations persist despite recent advances in transformer-based models (Natarajan et al., 2020; Yan et al., 2019), which prioritize sequential relationships but often neglect spatial correlations inherent in multi-lead ECG systems.

The field has further diversified through innovations in residual networks (Kachuee et al., 2018), dynamic data augmentation via GANs (Golany et al., 2020), and hybrid rule-CNN frameworks (Jin and Dong, 2017). While comparative studies (Andreotti et al., 2017) consistently affirm the superiority of neural networks over traditional methods, critical gaps persist in two clinically pivotal dimensions: multi-lead synergistic analysis and long-range rhythm modeling.

2.2. *Transformer architectures in ECG analysis*

Initially, Transformers (Vaswani et al., 2017; Tay et al., 2020) exhibited their efficiency in natural language processing (Wu et al., 2024; Du et al., 2024), before finding application in computer vision (Parmar et al., 2018; Carion et al., 2020). In earlier research, strategies for processing time-series data were predominantly characterized by original and variant methods of recurrent neural networks (Sak et al., 2014; Zaytar and El Amrani, 2016; Cho et al., 2014). In the current landscape, transformer-based approaches (Vaswani et al., 2017; Dai et al., 2019) have made significant contributions by utilizing self-attention mechanisms (Wang et al., 2018) to capture data relationships.

Although transformers have achieved great performance, the original transformers introduced substantial computational complexities when dealing with lengthy sequences. Subsequent efforts aimed at refining these models with improved computational efficiency. The reformer (Kitaev et al., 2019) employed a hashing method to identify critical features before self-attention computations. Linformer (Wang et al., 2020) utilized low-rank approximation to reformulate attention weights, effectively replacing the $O(n^2)$ operations of the original transformer (Vaswani et al., 2017) with an $O(n)$ operation. Similarly, the Performer (Choromanski et al., 2020; Lan, 2023) introduced a concept using orthogonal random features to eliminate the need for storing and computing attention weights. Transformer-XL (Dai et al., 2019) drew inspiration from recurrent neural networks by establishing connections between adjacent local segments. Specifically, segment-based recurrent methods, such as the Transformer-XL, demonstrated remarkable flexibility in managing long sequences.

These advances in sequential data analysis led to the introduction of some transformer-based methods (Natarajan et al., 2020; Yan et al., 2019; Han et al., 2023) for ECG classification, resulting in commendable performance. However, it is worth noting that these methods did not explicitly model long-term dependencies, potentially omitting critical features in the process.

2.3. *Explainable Artificial Intelligence*

The primary goal of Explainable Artificial Intelligence (XAI) is to obtain human-interpretable models especially for applications in sensitive sectors such as military, banking, and healthcare applications (Ali et al., 2023). The specialists of these domains want to solve their problems more effectively while with trustworthy outputs not only in the results but the reason behind the outputs. Ribeiro et al. (2016) proposed LIME, a novel explanation technique that explains the predictions of any classifier. LIME understands prediction reasons by learning explainable models, and presents representative individual predictions and explanations in a non-redundant manner. (Selvaraju et al., 2017) introduced GradCAM, a technique using gradients to generate positioning maps to highlight important areas of the image and is suitable for various CNN models without retraining.

XAI is extremely important in medical and healthcare fields, and numerous work has been carried on these fields (Salahuddin et al., 2022). Schutte et al. (2021) introduced StyleGAN, a new method that can be used to understand the predictions of any black-box model on images by showing how the input image would be modified in order to produce different predictions. (Kim et al., 2021) proposed XProtoNet, a globally and locally interpretable diagnosis framework for chest radiography. XProtoNet predicts the area where a sign of the disease is likely to appear and compares the features in the predicted area with the prototypes thus enhances its interpretability.

These new interpretative approaches cannot only provide more informative and meaningful explanations, but can also help uncover new biomarkers, reveal patterns learned by models, and ultimately uncover potential biases (Fan et al., 2021). The research and application of explainability in the medical direction will play a crucial role in enhancing the credibility and application value of machine learning algorithms in clinical practice.

3. Method

3.1. *Problem definition*

ECG is a cost-effective, non-invasive tool for diagnosing cardiovascular diseases, in which multiple leads (channels) record the heart's electrical signals over time. Formally, a typical multi-lead ECG recording can be represented as $X \in \mathbb{R}^{c \times t}$, where c is the number of leads and t is the temporal length. Each channel $X_{i,:}$ captures a view of the cardiac activity sampled at a specific rate. In this setting, the complexity arises from several factors:

(i) long-term dependencies. The classification function f is explicitly modeled as a mapping from the full temporal sequence:

$$f : \mathbb{R}^{c \times t} \rightarrow \{0,1\}^K, \tag{1}$$

where c denotes the number of leads (channels), t represents the temporal length, and K is the number of possible conditions. To emphasize long-range dependencies, a "global" operator (e.g., attention mechanism, LSTM unit, etc.) is introduced, for instance:

$$f(X) = g\big(\Phi(X)\big), \tag{2}$$

where Φ captures extended temporal context (e.g., a Transformer, an LSTM, or stacked 1D convolutions), and g is a subsequent classification mapping. Consequently, the model leverages global information from the entire time series rather than relying solely on local windows.

(ii) interpretability requirements. Clinicians often require insight into which segments or leads drive the model's decision for a specific condition. To address this, an additional "explanation" function,

$$h : \mathbb{R}^{c \times t} \rightarrow \mathbb{R}^{c \times t}, \tag{3}$$

can be introduced. The output $h(X)$ may be viewed as an "attention map" or "importance score", indicating which time segments or leads receive greater focus. A common practice is to include an extra interpretability regularization or constraint in the loss function. Combining the above aspects enables the incorporation of noise, global dependencies, and interpretability into a single framework. Suppose a multi-lead ECG signal is observed:

$$X_{\text{obs}} = X + \eta, \quad \eta \sim \mathcal{D}_\eta. \tag{4}$$

A classification function f and an explanation function h are learned to satisfy the following properties: Robustness, ensuring the model maintains stable performance under noise corruption η; Global Dependency, where f integrates information across the entire temporal window (t time steps) and multiple leads (c channels); and Interpretability, whereby h highlights critical time segments or leads, thus facilitating clinical evaluation.

3.2. *Our proposed neural classifier*

Therefore, advanced Transformer and attention technologies were combined to develop a new classifier. Noted that our proposed architecture based on a length-adapter structure called Transformer-XL (Dai et al., 2019). The self-attention module in the original transformer (Vaswani et al., 2017) was limited to processing data with a predefined length, resulting in high computational complexity when dealing with lengthy input data. However, Transformer-XL addressed this limitation by incorporating segment-level recurrence into the attention mechanism and introducing relative position encoding to replace absolute position encoding. In Transformer-XL, the self-attention is defined as:

$$\begin{aligned} &\tilde{h}^{n-1}_{\gamma+1} = [\mathrm{sg}(h^{n-1}_{\gamma}) \odot h^{n-1}_{\gamma+1}], \\ &q^{n}_{\gamma+1}, k^{n}_{\gamma+1}, v^{n}_{\gamma+1} = h^{n-1}_{\gamma+1} W_q^T, \tilde{h}^{n-1}_{\gamma+1} W_k^T, \tilde{h}^{n-1}_{\gamma+1} W_v^T, \\ &h^{n}_{\gamma+1} = \text{Transformer-Layer}(q^{n}_{\gamma+1}, k^{n}_{\gamma+1}, v^{n}_{\gamma+1}). \end{aligned} \tag{5}$$

where h^n_γ denotes the γ-th segment within the hidden states of the nth layer, $\mathrm{sg}(\cdot)$ indicates the stop-gradient operation, and $[h \odot h]$ means the concatenation along the length dimension. The Transformer-Layer$(\cdot)$ stands for the self-attention computing in the original transformer. This design allows the previous segment to influence the current attention computation during forward propagation, but it does not affect backward propagation. This approach enables Transformer-XL to effectively learn long-term dependencies in sequential data.

A complete ECG signal usually contains several leads which are collected from different parts of the body (e.g. the Lead I records the voltage difference between the left and right arm electrodes and the Lead II records the voltage difference between the electrodes of the left leg and the right arm). Although the waves of the different leads look similar generally which reflect the general state of the heart health, the difference between them contains abundant information and some diseases can only be diagnosed by combining several leads. To enable the model to integrate and synthesize information from various sources is of critical importance.

A novel framework for heart disease classification is introduced here, comprising Lead Fusing blocks, Encoder modules with Transformer-XL encoders, and hierarchical temporal attention modules, as illustrated in Fig. 2. The input is a multi-lead ECG tensor $X \in \mathbb{R}^{c \times t}$, where c denotes the leads and t the temporal length. First, the Lead Fusing blocks fuse features across channels using local convolution and channel-wise attention, which adaptively weights each lead according to its diagnostic importance. The fused output is then fed into the Encoder modules, each employing a Transformer-XL backbone with a sliding window strategy: overlapping segments of the time-series are processed while carrying "memory" states forward, allowing long-range dependencies to be captured. Next, the resulting representations are passed to the hierarchical temporal attention modules, which operate in two stages: an initial temporal attention emphasizes the most relevant time segments within each window, and a subsequent attention layer consolidates these window-level features into a global representation. Finally, a fully connected layer produces the multi-label classification results for various cardiac abnormalities. By integrating channel-wise feature fusion, non-local temporal modeling, and multi-level attention, this framework highlights clinically significant waveforms and leads, while maintaining interpretability and robustness in heart disease diagnosis.

3.2.1. *Lead fusing block*

Following data pre-processing, the ECG signals $x \in \mathbb{R}^{l \times c}$ (l and c are the length and channel number) are forwarded through several Lead Fusing blocks for channel fusion. Prior approaches often processed ECG signals using 1D convolutions (Chen et al., 2020b; Rajpurkar et al., 2017), overlooking the unique relationships among different leads. While 1D convolutions can compute a weighted average over the signal leads, the use of fixed weights is impractical. Specifically, signals recorded by various leads capture the same cardiac activities from different perspectives. Each signal conveys specific information, which can be of vital importance in certain cases (Anderson et al., 1994). Moreover, the significance of these signals can vary under different circumstances.

To alleviate this concern, an innovative Lead Fusing block is introduced. In particular, following a residual learning strategy, SE-Attention (Hu et al., 2018) is incorporated to enhance the fusion of features from different channels, as depicted in Fig. 2. Formally, for a signal (or features) x_{c_i} from lead c_i, the output is computed as:

$$u_{c_i} = v_i(x) * x_{c_i}, \tag{6}$$

where $v_i(x)$ squeezes the time dimension of the ith channel into a single scalar, and "$*$" denotes broadcast multiplication. Fig. 2 shows that each Lead Fusing block has two pipelines in parallel: a 1D convolution pipeline and a Squeeze-and-Excitation (SE) pipeline. The 1D convolution pipeline captures local features by convolving each lead's signal, applying Batch Normalization, and using ReLU. The SE pipeline reweights each channel by collapsing the time dimension into one statistic, learning a scaling factor, and applying this factor to the feature map. These two outputs are then merged by element-wise addition and linked with a residual connection to preserve low-level details. Four Lead Fusing blocks are stacked in sequence, where the first block uses a kernel size of 50 and later blocks use 15, thus progressively consolidating multi-lead ECG information before higher-level temporal modeling.

3.2.2. *Encoder module*

The Encoder modules in our framework handle the features produced by the final Lead Fusing block. Unlike text, ECG signals lack clear breakpoints (e.g., full stops or paragraphs). Prior transformer-based approaches have either operated on entire ECGs at once (Natarajan et al., 2020), causing high computational cost, required manual R-peak annotations (Yan et al., 2019), or used hard segmentation (Mousavi et al., 2020).

To handle long ECG recordings, a sliding window of length $L = 32$ and step size $s = 24$ is employed. Consequently, each window overlaps with its predecessor by $o = 8$ samples. Formally, let the post-fusion ECG signal be:

$$X \in \mathbb{R}^{c \times T},$$

where c denotes the number of channels (leads), and T indicates the total temporal length. X is partitioned into N segments based on:

$$N = \left\lfloor \frac{T - L}{s \cdot o} \right\rfloor + 1. \tag{7}$$

Each segment is defined as:

$$S_i \in \mathbb{R}^{c \times L}, \quad i = 1, 2, \ldots, N,$$

where S_i represents the ith contiguous slice of the signal in the time dimension. Next, each segment S_i is processed by a two-layer Transformer-XL encoder (see Eq. (5)), as illustrated in Fig. 1. Concretely, the outputs of these encoders for all segments are denoted as:

$$\{ h_1, h_2, \ldots, h_N \},$$

which are subsequently concatenated along the temporal axis, yielding:

$$H = \mathrm{Concat}(h_1, h_2, \ldots, h_N).$$

Two such Encoder modules are stacked in sequence: the hidden states from the first module are concatenated and fed into the second, thus capturing more non-local information without incurring prohibitive computational overhead. Finally, the output H is passed into a hierarchical temporal attention module, which fuses the learned features into a global ECG representation for downstream classification.

3.2.3. *Hierarchical temporal attention*

In the Encoder module (Section 3.2.2), the emphasis is on capturing non-local features along the length dimension. However, we have not yet captured the global features. Since Transformer-XL encoders break down the self-attention computation into several steps, a global feature extractor is required to fuse the features from each step (for each segment). Inspired by spatial attention mechanisms, a hierarchical temporal attention operation is proposed to address these non-local features and achieve higher-level (global) semantics. The operation in the temporal attention is defined as follows:

$$\begin{aligned} &\mathrm{sigmoid}(f(x)) = A \\ &A * x = y \end{aligned} \tag{8}$$

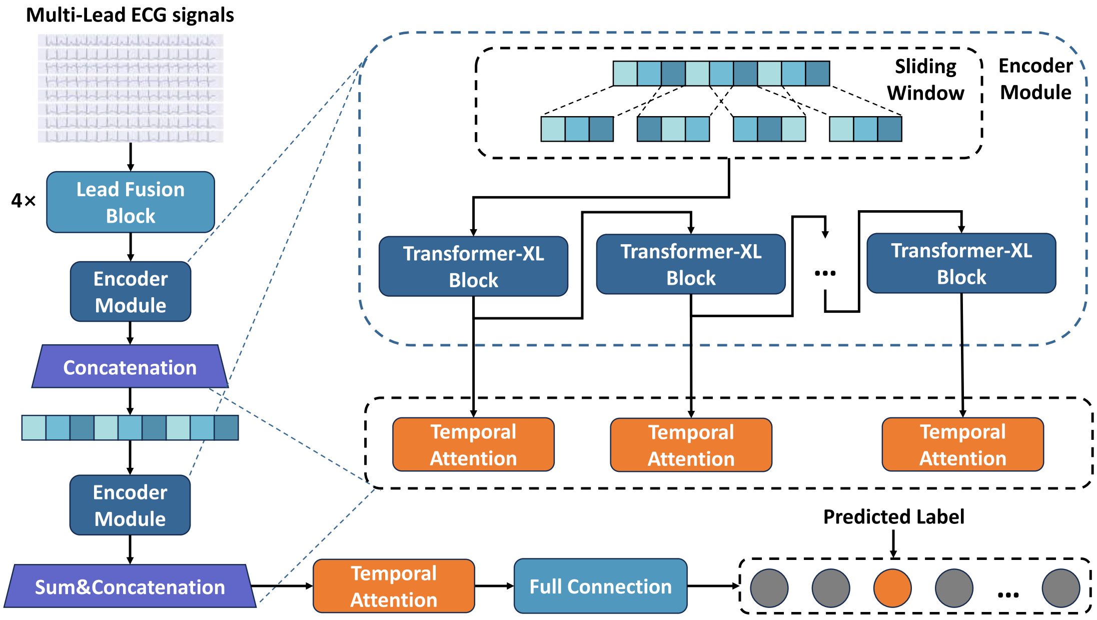

Fig. 1. The overview of our proposed neural classifier for heart disease diagnosis. An encoder module is built based on the transformer-XL encoder, and the parameters of the temporal attentions are shared in the same layers. The input ECG signal is recorded by multiple leads, and the signal from each lead is treated as one input channel in our method.

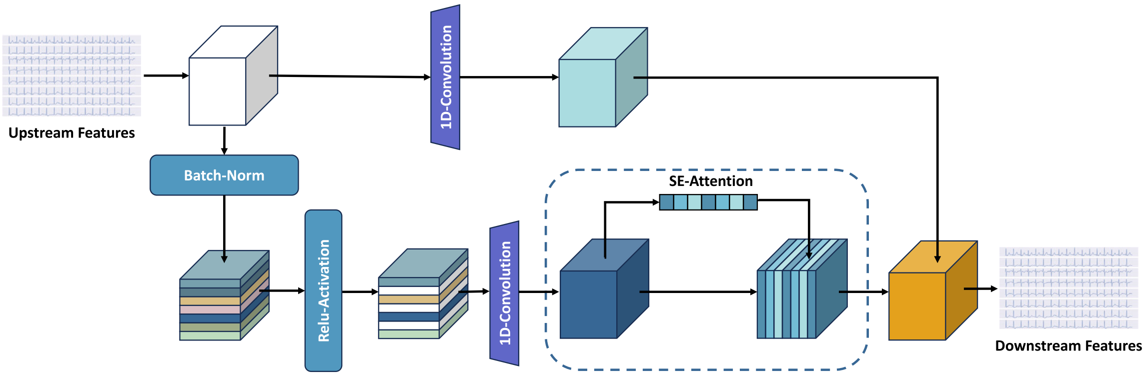

Fig. 2. The detailed structure of the proposed Lead Fusing block. The input data follows two pipeline: 1D-convolution and SE-Attention, and then two pipelines join together as the input data for the next layer of the model. Note that: SE-Attention represents Squeeze-and-Excitation Attention.

where $x \in \mathbb{R}^{l\times c}$ (l and c indicate the length and the number of channels, respectively), and the function f denotes a mapping function realized through neural operations, such as convolution with non-linear activation. The attention weight A in Eq. (8) shares the same dimensions as the input x and is applied through element-wise multiplication, denoted by the "*" operator.

In the first attention layer, the temporal attentions are parameter-shared. The length of the resulting features is reduced to a single value through averaging, and these values are concatenated before being fed into the second attention layer. This procedure can be formally defined as:

$$\begin{aligned} \mathbb{E}_l\{y^{(i)}_{(l\times c)}\} &= r^{(i)}_{(1\times c)} \\ r_{(n\times c)} &= [r^{(1)}, r^{(2)}, \dots, r^{(i)}, \dots, r^{(n)}] \end{aligned} \tag{9}$$

where $y^{(i)}$ denotes the ith produced features from the temporal attention, and the subscripts indicate the feature sizes. $\mathbb{E}$ signifies an average operation. The symbol r represents the reduced feature representation obtained after applying the hierarchical temporal attention mechanism. Therefore, the hierarchical temporal attention further consolidates the non-local features: the first attention layer processes the segment features, and the second combines these features to capture global information.

Finally, the extracted features are fed into a fully connected layer to make predictions regarding ECG diseases. Similar to the previous works on multi-label classification, the model is trained using the binary cross-entropy loss function, defined as:

$$\mathcal{L}(p, y) = -y \log p - (1 - y) \log(1 - p) \tag{10}$$

where y and p indicate the target label and predicted probability, respectively.

The binary classification is employed because an ECG may exhibit more than one type of pathology or none at all. The relationships among different diseases can be considered independent.After training using the hyperparameters specified in Section 4.2, performance was evaluated on the validation set, followed by testing on the test set. During the application phase, data preprocessing and inference follow the model's forward workflow. The predicted result is determined by selecting the target label with the highest probability within the specified range:

$$\widetilde{Y} = \arg\max p, \tag{11}$$

where $\widetilde{Y}$ represents the predicted outcome for the given case during inference.

4. Experiments

4.1. Data preparation

In this section, experiments are conducted on the Tianchi ECG dataset[2] to evaluate the effectiveness of the proposed method. From the total of 20,038 8-lead ECG records, 19,759 samples were selected, covering the most common 20 ECG categories, each representing a specific disease. These signals are recorded at a frequency of 500 Hertz, and each belongs to more than one category. Before being fed into the proposed networks, the ECG signals were pre-processed through a denoising procedure (using the Python package (Makowski et al., 2020)) and linear scaling normalization into the 0–1 range, thereby enhancing neural network performance by:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{12}$$

where $x_{\min}$ and $x_{\max}$ are the minimum and the maximum values of the ECG signals respectively, x' represents the normalized ECG signal scaled to a range between 0 and 1.

4.2. Experimental setup

The proposed networks were implemented using Python 3.8 and PyTorch 1.7 on RTX2080Ti GPUs. The batch size was set to 128, and the stochastic gradient descent (SGD) optimizer (Robbins and Monro, 1951) was utilized during the training phase. The dataset was divided into 70% for training, 15% for validation, and 15% for testing, ensuring a consistent category distribution across each subset. During training, the SGD optimizer with a momentum of 0.9 and an initial learning rate of 0.1 was applied. When the training loss remained unchanged for 5 consecutive epochs, the learning rate was decayed by a factor of 0.1. A batch size of 128 was used, and the model was trained for 50 epochs to ensure convergence. For comparison, the performances of previous works are obtained by running the reimplemented codes in their default experimental setup as specified in their original papers.

4.3. Classification performances

In comparison with state-of-the-art methods—including the 1D ResNet-34 (He et al., 2016) (the champion network in the Tianchi ECG Classification Competition[3]), ECG-Net (Mousavi et al., 2019), and Hannun's model (Rajpurkar et al., 2017) (considered comparable to cardiologists)—our method exhibits superior performance across Precision, Recall, and Micro F1 indicators. These metrics serve as the official evaluation criteria in the Tianchi ECG Classification Competition. As shown in Table 1, specifically, the proposed model achieved a Precision of 94.3%, which is 0.5% higher than that of 1D ResNet-34, 3.2% higher than ECG-Net, and 2.0% higher than Hannun's model. In terms of Recall, the proposed model reached 89.0%, [showing an improvement over 1D ResNet-34's 87.2% and ECG-Net's 85.1%.] Considering both Precision and Recall, our model achieved a Micro F1 score of 91.6%, surpassing Hannun's model by 1.4%.

Furthermore, the model's performance in detecting different cardiac abnormalities was evaluated by analyzing the ROC-AUC curves for each ECG category, as shown in Fig. 3. The proposed model exhibited higher ROC-AUC values in most categories. For instance, in detecting left ventricular high voltage, our model achieved an ROC-AUC of 92.9%, which is 2.6% higher than Hannun's model's (Rajpurkar et al., 2017) 90.3%. In detecting atrial premature beats and sinus arrhythmia, the model achieved ROC-AUCs of 91.4% and 92.2%, respectively, outperforming Hannun's model (Rajpurkar et al., 2017) by 8.2% and 3.0%.

Table 1
The classification performances of the proposed methods and the current state-of-the-art work. The best performances are marked **in bold**.

Methods	Precision	Recall	Micro F1
Ours	**0.943**	**0.890**	**0.916**
Champion Network	0.935	0.888	0.912
1D ResNet34	0.938	0.872	0.904
1D ResNet34+Mixup	0.930	0.887	0.908
ECG-Net	0.911	0.851	0.880
Hannun's	0.923	0.881	0.902

The application of ECG diagnosis through deep learning focuses mainly on the performance of the model, which refers to metrics like accuracy and sensitivity (Acharya et al., 2017; Jin and Dong, 2017; Kiranyaz et al., 2015; Pourbabaee et al., 2018). Consequently, models are often designed to be as complex and deep as possible to extract features. However, the features extracted in this way are abstract and difficult for humans to understand, making it challenging to explain the reasoning behind them. On the other hand, features extracted manually are more apparent to humans, but the portability and performance of such models are weak due to the small and simple feature space representation (Majeed and Alkhafaji, 2023; Qin et al., 2017). To address these two problems, constructing models that are both precise and interpretable is urgent to meet the desire of doctors for a diagnosis system with good performance to accelerate their work and with good interpretability to enhance the reliability of their diagnosis (Ali et al., 2023). The performance improvement of the proposed method is mainly attributed to the introduction of multi-lead adaptive fusion and the Transformer-XL encoder. While the traditional 1D ResNet-34 model (He et al., 2016) performs well in local feature extraction, it is limited in multi-lead signal integration and long-term dependency modeling. ECG-Net (Mousavi et al., 2019), although incorporating visual attention mechanisms, has shortcomings in multi-lead integration. Hannun's model (Rajpurkar et al., 2017), despite its deep network structure, has limited effectiveness in capturing long-term features when processing long signal sequences.

Given that traditional neural networks, especially well-performing convolutional networks, lack interpretability while machine learning methods are weak in performance, it is more challenging to improve performance than to enhance interpretability (Salahuddin et al., 2022). Designing an architecture based on neural networks and endowing it with interpretability is technically feasible. This work is produced under this circumstance. The proposed network, based on a transformer-convolution hybrid architecture, is adequate for the task of electrocardiogram classification while not as complicated as ResNet-34, allowing interpretability studies to be carried out (Natarajan et al., 2020; Yan et al., 2019).

The proposed Lead Fusing module can dynamically weight different lead signals, allowing the model to adaptively adjust weights according to different signal sources, thereby enhancing the integration of multi-lead signals. Meanwhile, the Transformer-XL encoder captures long-range dependency features in the temporal dimension (Gao et al., 2023), and the hierarchical temporal attention mechanism further enhances the model's ability in global feature extraction and interpretability. These structural innovations effectively improve the model's performance in various ECG classification tasks.

4.4. Ablation study

To deeply analyze the contributions of each module, ablation experiments were conducted, and the results are shown in Table 2. The results indicate that the proposed modules are beneficial for ECG signal diagnosis. Notably, after removing the SE-Attention module, the Precision dropped from 0.943 to 0.939, and the Micro F1 score decreased to 0.913. Adjusting the number of Encoder modules yielded different effects. When using only one Encoder module, the Precision was 0.940,

[2] https://tianchi.aliyun.com/competition/entrance/231754/information?lang=en-us

[3] https://tianchi.aliyun.com/competition/entrance/231754/rankingList/1

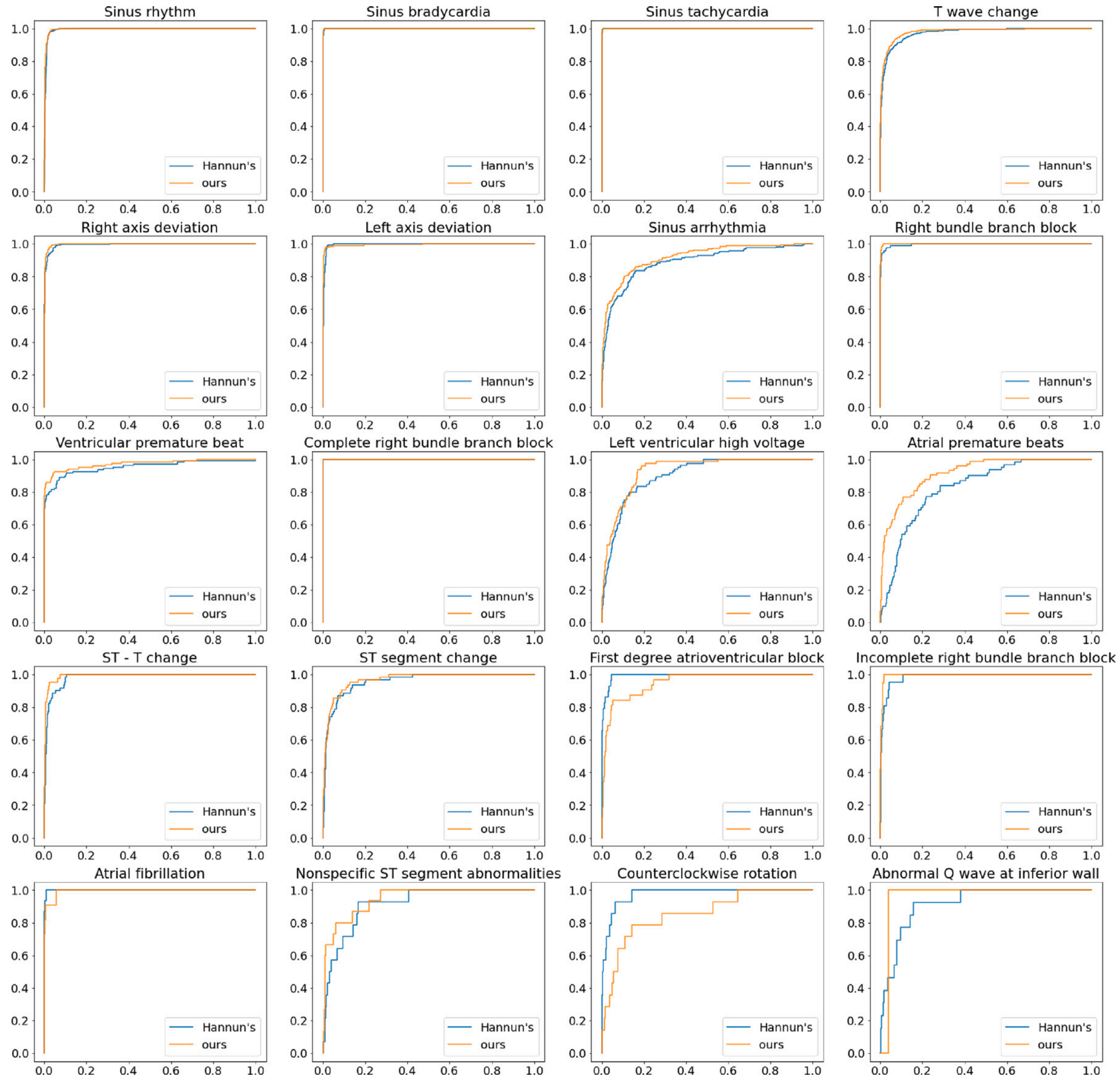

Fig. 3. Illustrating the ROC curves of our proposed method and the Hannun's (Rajpurkar et al., 2017). Notably, our proposed method outperform the Hannun's (Rajpurkar et al., 2017) in the most types of disease diagnosis. Each sub-figure represents a specific type of ECG abnormality, with ROC curves demonstrating the model's classification performance. Key differences in the performance are highlighted, particularly in categories like left ventricular high voltage, atrial premature beats, and sinus arrhythmia, where our model shows a clear advantage.

and the Recall slightly decreased to 0.888, indicating that a small number of Encoder modules are insufficient in capturing non-local features. Increasing to five Encoder modules, although the Precision increased to 0.947, the Recall dropped to 0.877, suggesting that too many modules lead to overfitting issues.

This approach assumes that there are relational features between lower and higher features extracted by a sequence model block. These relational features enhance the core features and weaken the less important ones, leading the final features to concentrate on the core areas. Meanwhile, due to the lightweight architecture, the visualization work can be presented clearly and precisely, demonstrating the interpretability of the model. In the past, convolution-based models have dominated the field of ECG classification (Farag et al., 2022). While moving and averaging operations can group and fuse local features to reveal smoother representations, models with a large number of stacked convolutional layers tend to blur the precise locations of these features (Huang et al., 2023). To address this problem, the use of convolutions was limited, and attention mechanisms were employed for feature extraction — such as SE-Attention in the Lead Fusing blocks, self-attention mechanisms in the Encoder modules, and hierarchical temporal attention — to enhance the model's interpretability.

The SE-Attention module effectively enhances the fusion of lead signals by adaptively adjusting weights among different leads, significantly enhancing signal-specific information. In contrast, excessive Encoder modules increase complexity and tend to overfit minor features. Therefore, a reasonable module configuration and appropriate attention mechanisms achieve a balance between model performance and generalization ability.

4.5. Visualizations and interpretability analysis

To evaluate whether the proposed classifier extracts essential features, the GradCAM method (Selvaraju et al., 2017) was applied to ECG signals, and key segments in the automatic diagnosis were visualized. Two cardiac abnormalities, complete right bundle branch block and ventricular premature beat, were selected for analysis.

As depicted in Fig. 4, the results show that our model can precisely focus on key waveforms such as P, Q, R, and S waves, with particularly

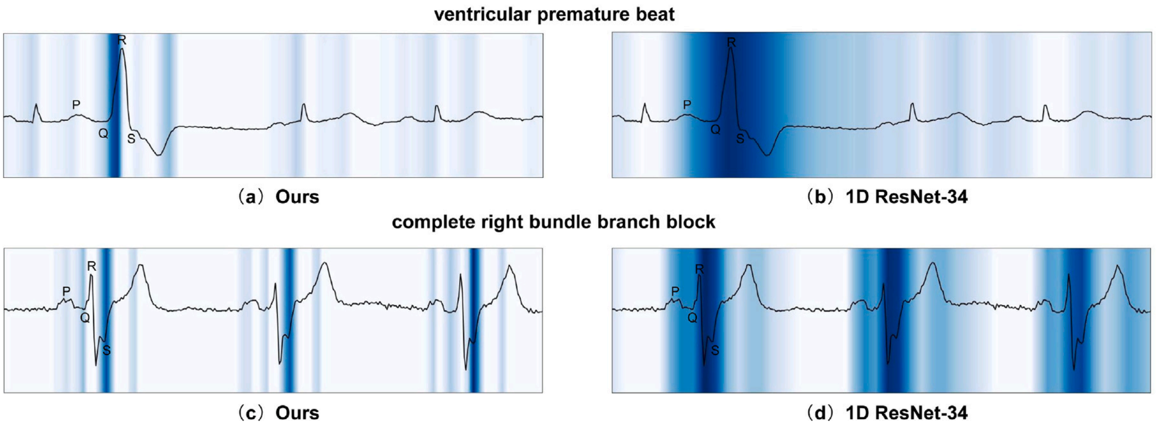

Fig. 4. Visualization comparisons between our approach with convolutional neural networks (using 1D ResNet as example).

Table 2
The ablation study of the proposed modules.

Methods	Precision	Recall	Micro F1
Ours (original)	0.943	0.890	0.916
(1) w/o SE-Attention	0.939	0.889	0.913
(2) w/o Temporal Attention	0.938	0.888	0.912
(3) w/ 1 * Encoder module	0.940	0.888	0.913
(4) w/ 3 * Encoder module	0.943	0.889	0.915
(5) w/ 4 * Encoder module	0.943	0.885	0.913
(6) w/ 5 * Encoder module	0.947	0.877	0.911

higher attention on the R wave compared to the comparative methods. For example, As in Fig. 4:a, in ventricular premature beat, the model accurately focuses on the abnormal QRS complex, which is crucial for correct diagnosis. In contrast, the attention of the ResNet-34 model is relatively dispersed, covering most of the cycle and lacking precise localization of specific waveform positions (see Fig. 4:b). This difference stems from the proposed model design using the Transformer-XL and hierarchical temporal attention mechanisms, reducing the number of convolutional layers, enabling the model to capture both global and local features in long time sequences, increasing attention to important cardiac events, and avoiding the feature location blurring issue in traditional convolutional networks.

Explainable Artificial Intelligence (XAI) has two main research directions: interpretability and explainability. The former focuses on interpreting results through visualization, statistical analysis, or other methods, while the latter aims to formalize the entire process, from feature extraction to model prediction, into mathematical formulations. Given the difficulty of simplifying and formalizing the ECG feature extraction process, this study concentrates on interpretability, that is, explaining the model's behavior by analyzing its predictions. For instance, when processing an input $x \in \mathbb{R}^{10}$, it would require 4 stacked convolutional layers with kernel sizes of 3 and a stride of 1 to fuse the first value $x^{(1)}$ and the last value $x^{(10)}$. In contrast, an attention module (e.g., self-attention) may achieve the same result in just one step. Therefore, the attention-based design facilitates the extraction of long-range dependencies in ECG signals with fewer steps, preserving the precise feature locations.

By adopting these attention mechanisms, the proposed model enhances interpretability in the following ways: First, the model can focus on specific parts of the input, maintaining the spatial and temporal resolution of important features. Second, the model can dynamically adjust the importance of different leads based on the input data, which is particularly critical when certain heart diseases primarily manifest in specific leads (Hammad et al., 2018). Third, visualizing the attention weights highlights which segments of the ECG signals contribute most to the classification decisions, aligning with clinical practice. By focusing on signals consistent with the features considered important by clinicians, our model's predictions become more understandable and trustworthy to medical professionals.

These results indicate that the model has a superior ability to precisely focus on the key parts of ECG signals with fewer convolutions, demonstrating the interpretability of the approach. However, there are still some limitations in this work. The reusability of the model architecture is not optimal. This construction appears to be specifically designed for the task at hand; the Lead Fusion Block and Transformer Sequence employed are not yet as universally applicable as linear or ResNet blocks, which represents a potential direction for future research (Zhao et al., 2022). Regarding interpretability, this work demonstrates what the model focuses on but not why the model makes its predictions. Although the proposed model achieves high performance on the current dataset, its generalization ability to unseen or out-of-distribution ECG signals has not been fully explored. Future work will focus on evaluating the model's performance across diverse datasets and implementing techniques such as data augmentation and domain adaptation to enhance its generalization ability.

5. Conclusions

In summary, this paper has introduced a neural classifier utilizing self-attention operations to model long-dependencies in time signals for ECG signal diagnosis. Our approach initially fuses features from various leads using the Lead Fusing block. Subsequently, Transformer-XL encoders with sliding windows are employed to capture non-local features and further learn global features. The final step involves a fully connected layer for ECG signal diagnosis (classification). Experiments conducted on the Tianchi ECG dataset have demonstrated that the proposed method outperforms state-of-the-art approaches. In addition, ablation studies provide insights into the contribution of each module and confirm their effectiveness. In the end, the GradCAM visualizations underscore the contents of the model focuses and thus enhance the interpretability of our approach while maintaining strong performance. However, the proposed method only visualizes the degree to which the model focuses on the features, and a thorough exploration of causality among these features remains insufficient.

In the future, research will focus on enhancing interpretability, improving robustness, and expanding clinical applications of AI-driven ECG diagnosis. Beyond GradCAM visualization, causality-based interpretation methods will be explored to clarify feature contributions, increasing clinical trust. To improve generalizability, efforts will incorporate multi-source ECG datasets, domain adaptation, and noise-resistant learning. The model's application will extend to early disease detection, wearable monitoring, and multimodal data fusion. Additionally, optimization techniques like pruning and quantization will enable real-time deployment on edge devices, ensuring accessibility. These advancements will drive AI-powered cardiovascular diagnostics towards greater accuracy, usability, and clinical impact.

CRediT authorship contribution statement

Xiaoqiang Liu: Writing – original draft, Conceptualization. **Yinlong Xu:** Writing – review & editing. **Hongxia Xu:** Writing – original draft. **Liang He:** Formal analysis, Data curation. **Siyu Long:** Methodology, Investigation. **Yisen Huang:** Project administration, Methodology. **Yubin Wang:** Validation, Supervision. **Yingzhou Lu:** Project administration, Methodology. **Yingxuan Huang:** Visualization, Funding acquisition. **Jian Wu:** Supervision. **Honghao Gao:** Supervision. **Xiaobo Liu:** Supervision.

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

This research was partially supported by National Natural Science Foundation of China under grants No. 82202984 and China Scholarship Council under Grant No. 201906070289.

Data availability

Data will be made available on request.

References

Acharya, U.R., Oh, S.L., Hagiwara, Y., Tan, J.H., Adam, M., Gertych, A., San Tan, R., 2017. A deep convolutional neural network model to classify heartbeats. Comput. Biol. Med. 89, 389–396.

Agarap, A.F., 2018. Deep learning using rectified linear units. arXiv preprint arXiv:1803.08375.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F., 2023. Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Inf. Fusion 99, 101805.

Anderson, S.T., Pahlm, O., et al., 1994. Panoramic display of the orderly sequenced 12-lead ECG. J. Electrocardiol.

Andreotti, F., Carr, O., Pimentel, M.A., Mahdi, A., De Vos, M., 2017. Comparing feature-based classifiers and convolutional neural networks to detect arrhythmia from short segments of ECG. In: 2017 Computing in Cardiology. CinC, IEEE, pp. 1–4.

Azar, A.T., Member, I., Hassanien, A.E., Kim, T.-h., 2012. Expert system based on neural-fuzzy rules for thyroid diseases diagnosis. In: International Conference on Bio-Science and Bio-Technology. Springer, pp. 94–105.

Bian, Y., Chen, J., Chen, X., Yang, X., Chen, D.Z., Wu, J., 2022. Identifying electrocardiogram abnormalities using a handcrafted-rule-enhanced neural network. IEEE/ACM Trans. Comput. Biol. Bioinform.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: European Conference on Computer Vision. ECCV, Springer, pp. 213–229.

Chen, H., Huang, C., Huang, Q., Zhang, Q., Wang, W., 2020a. ECGadv: Generating adversarial electrocardiogram to misguide arrhythmia classification system. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 3446–3453.

Chen, J., Yu, H., Feng, R., Chen, D.Z., et al., 2020b. Flow-Mixup: Classifying multi-labeled medical images with corrupted labels. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine. BIBM, pp. 534–541.

Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y., 2014. On the properties of neural machine translation: Encoder–decoder approaches. In: Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. pp. 103–111.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al., 2020. Rethinking attention with performers. arXiv preprint arXiv:2009.14794.

Cierpka-Kmieć, K., Hering, D., 2020. Tachycardia: The hidden cardiovascular risk factor in uncomplicated arterial hypertension. Cardiol. J. 27 (6), 857–867.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V., Salakhutdinov, R., 2019. Transformer-XL: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860.

Du, W., Luo, T., Qiu, Z., Huang, Z., Shen, Y., Cheng, R., Guo, Y., Fu, J., 2024. Stacking your transformers: A closer look at model growth for efficient llm pre-training. arXiv preprint arXiv:2405.15319.

Dutta, A., Das, M., 2024. ECG disease classification using 1D CNN. In: 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation, Vol. 2. IATMSI, IEEE, pp. 1–6.

Emary, E., Zawbaa, H.M., Hassanien, A.E., Schaefer, G., Azar, A.T., 2014. Retinal blood vessel segmentation using bee colony optimisation and pattern search. In: 2014 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 1001–1006.

Fan, F.L., Xiong, J., Li, M., Wang, G., 2021. On interpretability of artificial neural networks: A survey. IEEE Trans. Radiat. Plasma Med. Sci. 5 (6), 741–760.

Farag, M.M., et al., 2022. A self-contained STFT CNN for ECG classification and arrhythmia detection at the edge. IEEE Access 10, 94469–94486.

Faziludeen, S., Sabiq, P., 2013. ECG beat classification using wavelets and SVM. In: 2013 IEEE Conference on Information & Communication Technologies. IEEE, pp. 815–818.

Feyisa, D.W., Debelee, T.G., Ayano, Y.M., Kebede, S.R., Assore, T.F., 2022. Lightweight multireceptive field CNN for 12-lead ECG signal classification. Comput. Intell. Neurosci. 2022 (1), 8413294.

Ganz, L.I., Friedman, P.L., 1995. Supraventricular tachycardia. N. Engl. J. Med. 332 (3), 162–173.

Gao, Z., Cui, X., Zhuo, T., Cheng, Z., Liu, A.A., Wang, M., Chen, S., 2023. A multitemporal scale and spatial–temporal transformer network for temporal action localization. IEEE Trans. Human-Mach. Syst. 53 (3), 569–580.

Gladstone, D.J., Dorian, P., Spring, M., Panzov, V., Mamdani, M., Healey, J.S., Thorpe, K.E., or Operations Committee, E.S.C., Aviv, R., Boyle, K., et al., 2015. Atrial premature beats predict atrial fibrillation in cryptogenic stroke: results from the EMBRACE trial. Stroke 46 (4), 936–941.

Golany, T., Lavee, G., Yarden, S.T., Radinsky, K., 2020. Improving ECG classification using generative adversarial networks. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Hammad, M., Maher, A., Wang, K., Jiang, F., Amrani, M., 2018. Detection of abnormal heart conditions based on characteristics of ECG signals. Measurement 125, 634–644.

Han, H., Lian, C., Zeng, Z., Xu, B., Zang, J., Xue, C., 2023. Multimodal multi-instance learning for long-term ECG classification. Knowl.-Based Syst. 270, 110555.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 770–778.

Hejc, J., Redina, R., Kolarova, J., Starek, Z., 2024. Multi-channel delineation of intracardiac electrograms for arrhythmia substrate analysis using implicitly regularized convolutional neural network with wide receptive field. Biomed. Signal Process. Control. 94, 106274.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Huang, R., Lu, H., Xing, Y., Fan, W., 2023. Multi-scale convolutional feature approximation for defocus blur detection. In: 2023 26th International Conference on Computer Supported Cooperative Work in Design. CSCWD, pp. 1172–1177.

Inan, O.T., Giovangrandi, L., Kovacs, G.T., 2006. Robust neural-network-based classification of premature ventricular contractions using wavelet transform and timing interval features. IEEE Trans. Biomed. Eng. 2507–2515.

Jin, L., Dong, J., 2017. Classification of normal and abnormal ECG records using lead convolutional neural network and rule inference. Sci. China Inf. Sci. 60 (7).

Kachuee, M., Fazeli, S., Sarrafzadeh, M., 2018. ECG heartbeat classification: A deep transferable representation. In: 2018 IEEE International Conference on Healthcare Informatics. ICHI, IEEE, pp. 443–444.

Kim, E., Kim, S., Seo, M., Yoon, S., 2021. XProtoNet: diagnosis in chest radiography with global and local explanations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15719–15728.

Kiranyaz, S., Ince, T., Gabbouj, M., 2015. Real-time patient-specific ECG classification by 1-D convolutional neural networks. IEEE Trans. Biomed. Eng. 63 (3), 664–675.

Kitaev, N., Kaiser, L., Levskaya, A., 2019. Reformer: The efficient transformer. In: International Conference on Learning Representations. ICLR.

Kropf, M., Hayn, D., Schreier, G., 2017. ECG classification based on time and frequency domain features using random forests. In: 2017 Computing in Cardiology. CinC, IEEE, pp. 1–4.

Kuila, S., Dhanda, N., Joardar, S., 2022. ECG signal classification and arrhythmia detection using ELM-RNN. Multimedia Tools Appl. 81 (18), 25233–25249.

Lan, E., 2023. Performer: A novel PPG-to-ECG reconstruction transformer for a digital biomarker of cardiovascular disease detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1991–1999.

Li, D., Zhang, J., Zhang, Q., Wei, X., 2017. Classification of ECG signals based on 1D convolution neural network. In: 2017 IEEE 19th International Conference on E-Health Networking, Applications and Services (Healthcom). IEEE, pp. 1–6.

Liang, H., Lu, Y., 2023. A CNN-RNN unified framework for intrapartum cardiotocograph classification. Comput. Methods Programs Biomed. 229, 107300.

Majeed, R.R., Alkhafaji, S.K., 2023. ECG classification system based on multi-domain features approach coupled with least square support vector machine (LS-SVM). Comput. Methods Biomech. Biomed. Eng. 26 (5), 540–547.

Makowski, D., Pham, T., et al., 2020. NeuroKit2: A Python toolbox for neurophysiological signal processing. URL: https://github.com/neuropsychology/NeuroKit.

Mousavi, S., Afghah, F., Acharya, U.R., 2020. HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks. Comput. Biol. Med. 104057.

Mousavi, S., Afghah, F., Razi, A., Acharya, U.R., 2019. ECGNET: Learning where to attend for detection of atrial fibrillation with deep visual attention. In: 2019 IEEE EMBS International Conference on Biomedical & Health Informatics. BHI, IEEE, pp. 1–4.

Natarajan, A., Chang, Y., Mariani, S., Rahman, A., Boverman, G., Vij, S., Rubin, J., 2020. A wide and deep transformer neural network for 12-lead ECG classification. In: 2020 Computing in Cardiology. IEEE, pp. 1–4.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D., 2018. Image transformer. In: International Conference on Machine Learning. ICML, PMLR, pp. 4055–4064.

Pourbabaee, B., Roshtkhari, M.J., Khorasani, K., 2018. Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients. IEEE Trans. Syst. Man, Cybern.: Syst. 48 (12), 2095–2104.

Qin, Q., Li, J., Zhang, L., Yue, Y., Liu, C., 2017. Combining low-dimensional wavelet features and support vector machine for arrhythmia beat classification. Sci. Rep. 7 (1), 6067.

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

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144.

Robbins, H., Monro, S., 1951. A stochastic approximation method. Ann. Math. Stat. 400–407.

Sak, H., Senior, A., Beaufays, F., 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Fifteenth Annual Conference of the International Speech Communication Association.

Salahuddin, Z., Woodruff, H.C., Chatterjee, A., Lambin, P., 2022. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Comput. Biol. Med. 140, 105111.

Sayadi, O., Shamsollahi, M.B., Clifford, G.D., 2009. Robust detection of premature ventricular contractions using a wave-based Bayesian framework. IEEE Trans. Biomed. Eng. 353–362.

Schutte, K., Moindrot, O., Hérent, P., Schiratti, J.B., Jégou, S., 2021. Using stylegan for visual interpretability of deep learning models on medical images. arXiv preprint arXiv:2101.07563.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Sörnmo, L., Laguna, P., 2006. Electrocardiogram (ECG) signal processing. Wiley Encycl. Biomed. Eng.

Tay, Y., Dehghani, M., Bahri, D., Metzler, D., 2020. Efficient transformers: A survey. arXiv preprint arXiv:2009.06732.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. arXiv preprint arXiv:1706.03762.

Venkatesan, C., Karthigaikumar, P., Paul, A., Satheeskumaran, S., Kumar, R., 2018. ECG signal preprocessing and SVM classifier-based abnormality detection in remote healthcare applications. IEEE Access 6, 9767–9773.

Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803.

Wang, S., Li, B., Khabsa, M., Fang, H., Ma, H., 2020. Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768.

Wang, M., Rahardja, S., Fränti, P., Rahardja, S., 2023. Single-lead ECG recordings modeling for end-to-end recognition of atrial fibrillation with dual-path RNN. Biomed. Signal Process. Control. 79, 104067.

Wijesurendra, R.S., Casadei, B., 2019. Mechanisms of atrial fibrillation. Heart 105 (24), 1860–1867.

Wu, D., Nie, L., Mumtaz, R.A., Agarwal, K., 2024. A LLM-based hybrid-transformer diagnosis system in healthcare. IEEE J. Biomed. Heal. Inform.

Yan, G., Liang, S., Zhang, Y., Liu, F., 2019. Fusing transformer model with temporal features for ECG heartbeat classification. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine. BIBM, IEEE, pp. 898–905.

Zaytar, M.A., El Amrani, C., 2016. Sequence to sequence weather forecasting with long short-term memory recurrent neural networks. Int. J. Comput. Appl. 143 (11), 7–11.

Zhang, T., Lian, C., Xu, B., Su, Y., Zeng, Z., 2024. Cardiac signals classification via optional multimodal multiscale receptive fields CNN-enhanced transformer. Knowl.-Based Syst. 300, 112175.

Zhao, Z., Murphy, D., Gifford, H., Williams, S., Darlington, A., Relton, S.D., Fang, H., Wong, D.C., 2022. Analysis of an adaptive lead weighted ResNet for multiclass classification of 12-lead ECGs. Physiol. Meas. 43 (3), 034001.