Neurocomputing 680 (2026) 133287

Contents lists available at ScienceDirect

Neurocomputing

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

CARDxnosis: Anatomico-clinical knowledge-driven explainable network with multi-domain information fusion for ECG diagnosis and report generation

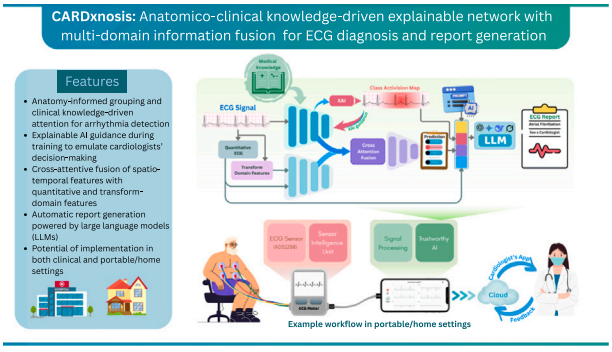
Tushar Talukder Showrav[a], Soyabul Islam Lincoln[b], Md. Kamrul Hasan[a,*]

[a] Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology (BUET), Dhaka, 1205, Bangladesh
[b] Department of Electronics & Communication Engineering, Khulna University of Engineering and Technology, Khulna, 9203, Bangladesh

HIGHLIGHTS

- Rhythm encoder with anatomy-informed lead grouping and clinical attention.
- Feature encoder integrates quantitative time- and frequency-domain features.
- Cross-attention fuses rhythm and feature encoders for higher accuracy.
- Explainable AI with annotated regions enhances interpretability and trust.
- LLM-driven ECG report generation connects AI outputs to clinical practice.

GRAPHICAL ABSTRACT



ARTICLE INFO

Communicated by R. Yang

Keywords:
Medical knowledge
Lead grouping attention
Variational mode decomposition
ECG quantitative features
Explainable AI guidance
ECG report generation

ABSTRACT

Electrocardiography (ECG) is indispensable for diagnosing cardiovascular diseases; however, accurate interpretation remains challenging due to waveform complexity and limited annotated datasets. We propose CARDxnosis, an anatomico-clinical knowledge-driven deep learning framework, that integrates anatomical lead grouping with attention mechanisms to capture clinically meaningful relationships among ECG leads. The architecture employs two encoders: one dedicated to learning spatio-temporal patterns from raw signals, and another that extracts quantitative waveform descriptors and frequency-domain representations generated through Variational Mode Decomposition (VMD). A cross-attention fusion module integrates these complementary encodings into a unified representation. Moreover, explainable AI guidance is incorporated into the training process, wherein model learning is augmented through the use of region-level annotations of abnormal ECG segments. This guidance, validated on a subset of strategically annotated data, demonstrates that embedding interpretability within the optimization framework enhances both predictive accuracy and clinical trust, although its broader application remains constrained by substantial annotation requirements. Additionally, the framework incorporates an automated ECG report-generation scheme powered by large language models (LLMs) that translate predictions into structured human-readable clinical interpretations. Validated across four benchmark datasets, CARDxnosis achieves multilabel accuracies of 78.73% (CPSC), 70.69% (Chapman), 35.78% (Georgia), 53.09% (PTBXL), and multiclass accuracies of 82.24% (CPSC), 95.17% (Chapman), and 67.69% (Georgia). Visualizations using t-SNE and Grad-CAM further confirm that the model learns robust and clinically interpretable features. Despite the challenges of full clinical integration, this work provides a reliable and explainable foundation for automated ECG analysis.

* Corresponding author.
Email addresses: 1706087@eee.buet.ac.bd (T.T. Showrav), islamsoyabul98@gmail.com (S.I. Lincoln), khasan@eee.buet.ac.bd (Md.K. Hasan).

https://doi.org/10.1016/j.neucom.2026.133287
Received 18 November 2025; Received in revised form 9 February 2026; Accepted 6 March 2026
Available online 7 March 2026
0925-2312/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

1. Introduction

Cardiovascular diseases are among the most common fatal illnesses. Studies estimate that in 2019, the number of deaths caused by cardiovascular diseases reached as high as 18.6 million [1]. Electrocardiograms (ECGs), as a tool for recording the electrophysiological activity of the heart, have become a crucial basis for diagnosing cardiovascular diseases. An inexpensive and non-invasive 12-lead ECG can provide a wealth of information to help detect abnormalities associated with the heart's electrical activity [2]. Early detection of abnormal heart rhythms can significantly reduce the risk of sudden death caused by cardiovascular diseases, making rapid and accurate ECG diagnosis essential in clinical practice. However, ECG annotation is time-consuming, labor-intensive, and is performed by a specialized cardiologist. In recent years, with the rapid development of smart medicine, ECG-based arrhythmia detection systems have become crucial tools for physicians, enhancing the efficiency and precision of diagnosis [3–5]. Despite these achievements, a key remaining challenge is improving the accuracy of automatic heart abnormality detection.

Over the past decade, hidden Markov models, machine learning techniques, and artificial neural networks have been widely applied for ECG signal classification to detect arrhythmias, involving preprocessing, feature extraction, and classification [6,7]. For machine learning approaches, proper feature extraction is crucial for obtaining decent results. Feature extraction includes intralead features (P wave, QRS complex, ST segment, T wave, P-R interval) extracted from individual leads, interlead features capturing relationships across leads, frequency component-based features from Intrinsic Mode Functions (IMFs), and patients' morphological features. Frequency-domain analysis often relies on the discrete wavelet transform (DWT) [8] and continuous wavelet transform (CWT) [9], which can be improved through complex DWT, dual-tree complex wavelet transform, or empirical mode decomposition (EMD) [10]. However, EMD has limitations, particularly its susceptibility to mode mixing, which can reduce its ability to distinguish signals with similar frequencies [11]. In contrast, Variational Mode Decomposition (VMD) decomposes signals into non-overlapping modes, achieving a cleaner and more distinct separation of components. This makes VMD especially effective for analyzing complex, non-stationary signals such as the ECG, facilitating more precise feature extraction and improving the accuracy of arrhythmia detection.

In recent years, deep learning has made remarkable strides in ECG classification, demonstrating significant potential for precise diagnosis and timely intervention of arrhythmias. Researchers have developed a variety of approaches to enhance both the accuracy and efficiency of ECG analysis. Wei et al. [12] employed knowledge distillation through a teacher-student framework, enabling targeted learning of subtle features by capturing differences between teacher and student outputs. Huang et al. [13] introduced a mutual learning strategy that aligns multi-resolution features at each layer to improve robustness and performance. In another study, Liu et al. introduced ECG-STAR (Spatio-Temporal Attention Residual), a deep learning model for multi-label classification of ECG abnormalities that combines linear layers, LSTM networks, spatial-temporal convolutions, and channel attention mechanisms to capture complex ECG patterns [14]. Cheng et al. [15] proposed a dual-stream architecture using 9-second ECG segments, where SHapley Additive exPlanations (SHAP) identified key features from the first model for the second, effectively combining feature extraction and classification. To address data imbalance, Msigwa et al. [16] integrated DBSCAN clustering with GANs to generate high-quality synthetic ECG samples from clean data filtered through signal quality evaluation. Mahel et al. [17] presented a hybrid deep learning framework combining multi-scale 1D convolutional layers, Efficient Channel Attention (ECA), MobileViT blocks, and LSTMs to capture both local and global dependencies without extensive preprocessing. Complementing these efforts, Liu et al. [18] proposed a multimodal architecture that fuses spatio-temporal and frequency-domain features, overcoming the limitations of single-modality approaches. In another work, Wang et al. proposed the Wave Masked Autoencoder (WMAE), a self-supervised model that incorporates ECG wave-specific prior knowledge to learn robust representations from unlabeled data and improve diagnostic performance [19]. Collectively, these advancements illustrate the growing sophistication of deep learning in extracting complex patterns from ECG signals, paving the way for more accurate, efficient, and clinically meaningful arrhythmia detection.

Most existing methods treat 12-lead ECGs as conventional 2D arrays, overlooking the complementary and interdependent roles of individual leads in capturing diverse aspects of cardiac electrical activity [20]. To address this limitation, several strategies have been proposed to better leverage lead-specific information. Xiaoyun et al. [21] introduced a multibranch network that organizes leads in varying configurations, from all leads combined to each lead separately. Shunxiang et al. [22] proposed an anatomically inspired grouping based on signal origin, including inferior, lateral, anterior, and atrial leads. Liu et al. [23] explored lead fusion using the Barlow Twins method during pretraining along with multi-branch concatenation, enhancing encoder performance. These approaches highlight the growing recognition of lead-specific relationships and demonstrate the importance of structuring networks to respect the physiological contributions of each lead for improved ECG classification.

Beyond the anatomical orientation of leads, incorporating clinically informed knowledge and interpretability is crucial for reliable diagnosis. Traditional deep neural networks often function as "black boxes," making it difficult for clinicians to understand or trust model decisions. To address this, Hui et al. [24] integrated beat-level and record-level ECG information through a dual-view CNN-Transformer network, using an external attention block to emphasize key spatio-temporal signal regions, while Jiang et al. [25] employed a dual-branch, domain-specific RR interval attention network to account for both intra-lead similarities and inter-lead variances. On the other hand, explainable AI (XAI) techniques have become widely adopted for interpreting and evaluating the reliability of network predictions. For instance, Grad-CAM has been used to generate heatmaps that explain 1D-CNN decisions for ECGs [26], while Tao et al. [6] extended this approach to multi-lead ECGs using 2D-ECGNet, which captures both inter- and intra-lead relationships through DKR blocks. Despite their usefulness in assessing prediction reliability, a key limitation of existing XAI methods is that they primarily provide post hoc explanations and do not actively contribute to improving the network's inherent reliability or performance. Recent studies also combined such explainable approaches with large language models: Wu et al. [27] used Grad-CAM heatmaps as input to GPT-4, producing clinically aligned textual explanations that enhance interpretability and trust. Liu et al. [28] further improved model transparency and interoperability using a Transformer-XL encoder with hierarchical attention to consolidate key features while reducing reliance on convolution.

Interpreting ECGs and generating comprehensive medical reports are inherently complex tasks in cardiology, requiring specialized expertise and substantial time [29]. Early automated systems approached this challenge using rule-based algorithms and classical AI methods, which relied on expert-defined thresholds to identify waveform features and detect arrhythmias. Recent advances in generative AI, however, have transformed automated medical report generation, attracting growing research interest. For example, Bleich et al. [30] developed an encoder-decoder framework that embeds 1D ECG signals using a modified 34-layer ResNet and produces clinically coherent automated reports through Transformer or LSTM decoders. Similarly, Ashfaq et al. [31] integrated Byte Pair Encoding (BPE) into a GPT-2 architecture, customizing the tokenizer to include ECG-specific terminology, thereby enabling the generation of highly accurate reports directly from raw ECG waveforms.

Despite notable advancements, existing literature still faces challenges in fully leveraging ECG data for highly reliable abnormality

detection. Although some methods acknowledge correlations between leads, they often lack attention guidance mechanisms tailored to specific cardiac conditions, limiting the network's ability to dynamically focus on the most relevant aspects of the ECG signal. Many studies also fail to integrate morphological features with deep learning architectures, missing the opportunity to combine domain-specific knowledge with the robust feature extraction capabilities of neural networks for improved classification performance. While frequency-based components, statistical features, and morphological characteristics have been utilized independently, they are rarely incorporated into a unified framework that assigns distinct importance to each feature type, potentially reducing the model's ability to capture the full complexity of ECG signals. Moreover, research on multi-lead ECG interpretation remains limited, and current automated report generation methods often suffer from inadequate input prompting, failing to provide structured feature data, including key waveform measurements and clinical indicators, systematically for generating comprehensive and clinically meaningful reports. To address these limitations, we propose CARDxnosis, an innovative architecture for ECG diagnosis. The key contributions of this work are as follows:

i. We present an anatomico-clinical knowledge-driven explainable network, CARDxnosis, that harnesses multi-domain information fusion, forming the foundation for reliable ECG diagnosis and automated report generation through LLMs.
ii. Building upon this foundation, we design an anatomically knowledge-inspired rhythm encoder that clusters ECG leads according to their anatomical origins and applies clinically informed attention mechanisms, thereby enhancing the model's capacity to capture and emphasize clinically relevant spatial relationships within cardiac signals.
iii. To complement rhythm encoding, we introduce a feature encoder that seamlessly integrates quantitatively derived characteristics from both raw ECG waveforms and frequency-domain representations, thereby capturing a more comprehensive and holistic cardiac profile.
iv. These representations are further enriched through a novel cross-attention fusion mechanism, which effectively merges signal-level and feature-level information. This integration significantly enhances the model's discriminative capacity, leading to improved classification performance across both multiclass and multilabel diagnostic tasks.
v. To ensure clinical trustworthiness, we embed explainable AI guidance directly into the training process. This mechanism enables the network to focus on clinically relevant regions of the ECG, with ground-truth annotations for such regions generated through an innovative automated annotation pipeline.
vi. Finally, the model's diagnostic capability is seamlessly translated into practice through automated ECG report generation powered by an LLM, thereby bridging predictive performance with clinical interpretability and supporting real-world deployment in healthcare.

By integrating these innovative components into a unified framework, we present a robust and adaptable solution that has the potential to be utilized in both clinical and remote monitoring environments. We evaluated our model on four widely used benchmark ECG datasets, demonstrating compelling performance compared to recent state-of-the-art methods across both multiclass and multilabel classification tasks. The incorporation of automated ECG report generation enhances clinical interpretability, bridging the gap between algorithmic output and practical use. The remainder of this paper is organized as follows: Section 2 details the datasets, preprocessing, network architecture, and loss functions. Section 3 describes the experimental setup, evaluation metrics, and presents comparative results. Section 4 offers an ablation study, reliability analysis, and discussion of anomaly cases, along with identified limitations. Finally, Section 5 concludes with a summary of our contributions and future directions.

2. Materials and methods

2.1. Materials

2.1.1. Datasets

For the evaluation of the study, four publicly available datasets were utilized: the China Physiological Signal Challenge 2018 (CPSC) [32], PTB-XL [33], the Chapman dataset from Chapman University, Shaoxing People's Hospital [34], and the Georgia database [35]. The PTB-XL dataset comprises 21,837 clinical ECG recordings, each 10 s long, annotated by up to two ECG experts. Each recording includes multiple ECG statements, adhering to SCP-ECG criteria, with a total of 71 distinct statements for multilabel classification. The CPSC 2018 dataset includes 6877 multilead, multilabel ECG recordings, ranging from 6 to 60 s, sampled at 500 Hz. The Chapman dataset contains 10,248 ECG recordings. The fourth source is the Georgia database which is comprised of 10,344 samples in total. This database represents a unique demographic characteristic of the Southeastern United States. To ensure fair comparison across these datasets, the following preprocessing conditions were applied:

i. Sampling frequency standardized to 500 Hz.
ii. Recordings limited to 5000 samples, taking only the first 5000 samples if longer.
iii. Multilabel annotations restricted to no more than three labels per recording.
iv. Each class required to have at least 200 samples to ensure sufficient representation.

After meeting all these conditions, the ECG records for CPSC, PTB-XL, Chapman, and Georgia datasets are 6861, 16184, 9194, and 8163, respectively. For the multiclass classification task, a primary requirement is that each sample must contain only a single label. The datasets were processed to satisfy this criterion. Additionally, a threshold was set such that only classes with more than 100 samples were retained. Classes that did not meet this requirement were excluded from the multiclass setting. During this process, it was observed that the PTB-XL dataset contained only two classes meeting the threshold, with a significant class imbalance between them. Therefore, the PTB-XL dataset was excluded from the multiclass classification task. After filtering, the remaining sample counts were 6387 for CPSC, 6011 for Chapman, and 4087 for the Georgia dataset. The distribution of heart conditions within the dataset is shown in Table 1.

2.1.2. Preprocessing

The dataset pre-processing consists of two steps: denoising and normalization. For data denoising, we have only removed the baseline drifts from the signal. The process of baseline removal can be represented by the equation below:

$$X_{\text{filtered}}(n) = X(n) - \text{MedFilt}_{m_2}\left(\text{MedFilt}_{m_1}\left(X(n)\right)\right) \tag{1}$$

The original ECG signal $X(n)$ is filtered to obtain $X_{\text{filtered}}(n)$ by applying two median filters: MedFilt_{m_1} with a window size corresponding to 0.2 seconds, and MedFilt_{m_2} with a window size of 0.6 seconds. The window sizes are computed as

$$m_i = \lfloor f_s \cdot t_i \rfloor + \left(\lfloor f_s \cdot t_i \rfloor \bmod 2\right) - 1, \tag{2}$$

where the sampling frequency is $f_s = 500$ Hz and $t_i \in \{0.2, 0.6\}$ seconds. The signal is then normalized by subtracting its mean and dividing by its standard deviation, expressed as $X_{\text{norm}}(n) = \frac{X(n)-\mu}{\sigma}$, where $X(n)$ represents the original ECG signal at sample index n, μ is the mean of the signal, σ is the standard deviation, and $X_{\text{norm}}(n)$ is the normalized signal.

Table 1
Comparison of heart conditions based on lead attention regions and their presence across the Chapman, PTB, CPSC, and Georgia datasets in both multilabel and multiclass settings.

Heart Conditions	Lead Attention (A)				Dataset (Multilabel)				Dataset (Multiclass)		
	Ant.	Lat.	Inf.	Atr. + Sep.	Chap	PTB	CPSC	Georgia	Chap	CPSC	Georgia
Sinus Rhythm	0	0	1	0	1765	14,060	918	1732	1453	918	1732
Sinus Bradycardia	0	0	1	0	3697	491	✗	1336	2554	✗	526
Sinus Tachycardia	0	0	1	0	1384	468	✗	938	864	✗	288
Sinus Arrhythmia	0	0	1	0	✗	578	✗	352	✗	✗	146
Atrial Flutter	0	0	1	1	362	✗	✗	✗	192	✗	✗
Atrial Fibrillation	0	0	1	1	1435	761	1218	355	583	973	114
SVT	0	0	1	1	485	✗	✗	✗	365	✗	✗
1D AV Block	0	0	1	0	✗	202	721	480	✗	685	✗
PAC	0	0	1	0	✗	✗	612	422	✗	529	105
PVC	1	1	1	1	✗	393	697	201	✗	606	✗
RBBB	0	1	0	1	322	✗	1852	369	✗	1530	✗
Incomplete RBBB	0	1	0	1	✗	652	✗	220	✗	✗	✗
LBBB	0	1	0	1	✗	379	235	✗	✗	178	✗
LVH	0	1	1	0	997	987	✗	814	✗	✗	262
MI Side Wall	0	1	0	0	✗	1786	✗	✗	✗	✗	✗
Myocradial Ischemia	0	1	1	0	✗	680	✗	✗	✗	✗	✗
Pacing Rhythm	0	1	1	1	✗	288	✗	✗	✗	✗	✗
ST drop	1	1	1	1	✗	201	868	✗	✗	783	✗
ST Tilt up	1	1	1	1	✗	✗	220	✗	✗	185	✗
ST Changes	1	1	1	1	✗	497	✗	✗	✗	✗	✗
ST-T Changes	1	1	1	1	711	✗	✗	1301	✗	✗	315
ST Extension	1	1	1	1	✗	✗	✗	361	✗	✗	✗
Abnormal QRS	1	1	1	1	✗	1042	✗	✗	✗	✗	✗
T change	1	1	1	1	1312	1557	✗	1208	✗	✗	315
T opposite	1	1	1	0	✗	✗	✗	385	✗	✗	✗
Low Volt QRS	1	1	1	1	✗	✗	✗	284	✗	✗	✗
Axis Left	0	1	1	0	254	2078	✗	568	✗	✗	140
Intraventricular Block	0	1	0	1	✗	265	✗	✗	✗	✗	✗
Anterior Fascicular Block	0	1	1	0	✗	436	✗	✗	✗	✗	✗
QT Interval	1	1	1	1	✗	✗	✗	610	✗	✗	✗
Lateral Ischemia	0	1	1	0	✗	✗	✗	411	✗	✗	✗
Atrail Enlargement	1	1	1	1	✗	✗	✗	517	✗	✗	138

2.2. Proposed method

In this paper, we propose CARDxnosis, an innovative deep learning framework, for ECG abnormality classification, guided by medical domain knowledge in its architectural design. A simplified block diagram of the overall system is presented in Fig. 1. As illustrated, the architecture comprises two parallel encoder branches, namely the rhythm encoder and the feature encoder. The rhythm encoder processes the raw ECG signal to extract temporal and spatial features, while the feature encoder extracts quantitative features derived from waveform morphology and frequency-domain decompositions. These dual encoder features are subsequently integrated through a joint fusion module that employs a cross-attention mechanism to effectively capture complementary information from both modalities. The fused representation is then used to predict the underlying ECG abnormalities. In addition, an innovative

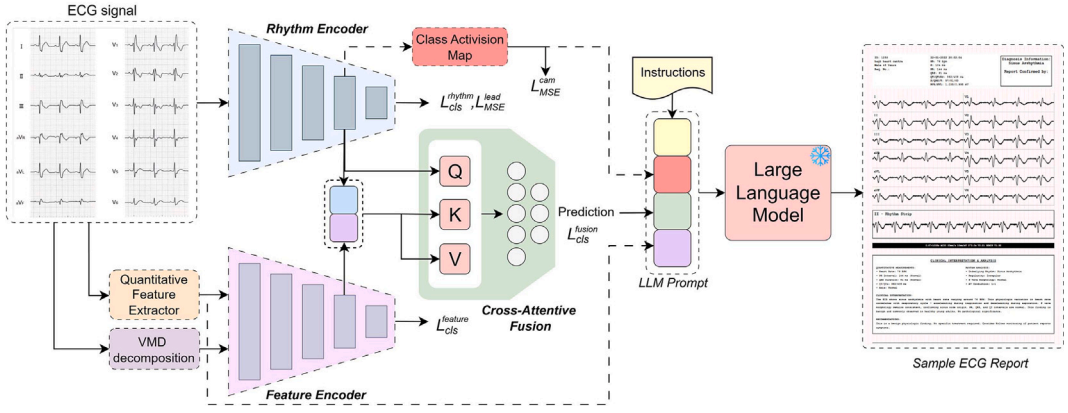


Fig. 1. Overview of the full architecture of the proposed CARDxnosis. The model consists of two parallel encoders: the Rhythm Encoder and the Feature Encoder. The raw twelve-lead ECG signal is directly processed by the Rhythm Encoder. In parallel, a set of quantitative features extracted from both the morphological and frequency domains of the ECG signal is passed to the Feature Encoder. The intermediate feature representations from both encoders are fused using a cross-attention mechanism, which is used to produce the final prediction. Furthermore, the final prediction, along with the extracted quantitative features, the class activation map, and a structured instruction, is used to generate an ECG report using a large language model.

train-time XAI guidance strategy is introduced to refine the rhythm encoder's focus on clinically relevant regions. In the final stage, an LLM is utilized to generate an automated diagnostic report. The subsequent sections provide a detailed explanation of each component of the proposed framework, along with the rationale and motivation behind the corresponding design choices.

2.2.1. Lead-specific medical knowledge induction

Traditional deep learning models have demonstrated impressive performance across various domains, including medical diagnostics; however, their black-box nature raises concerns regarding interpretability and trustworthiness, especially in clinical applications. One effective strategy to enhance trust in such models is the integration of domain-specific medical knowledge into their architecture. In this research, we integrated anatomical knowledge related to ECG lead configurations directly into the network design. Specifically, for 12-lead ECG-based abnormality classification, conventional approaches often use a single branch based encoder that processes all leads collectively. Some studies have proposed employing 12 independent branches in the encoder (one for each lead), but this approach significantly increases the model's complexity and parameter count. A more recent approach leverages clinical insight to group leads, enabling the encoder to utilize multiple branches, each dedicated to processing a medically coherent subset of leads. For instance, MVMSnet [22] demonstrated improved performance by grouping leads according to their anatomical relevance.

Building upon this idea, we propose a lead-specific attention mechanism that enables the network to focus on the most diagnostically relevant lead groups for each cardiac abnormality. This design aligns with the diagnostic practices of cardiologists, who prioritize certain leads depending on the suspected condition. It is important to note that not all ECG leads contribute equally to identifying all abnormalities; specific patterns are more pronounced in particular leads corresponding to different anatomical regions of the heart. Table 2 summarizes the grouping of ECG leads based on their anatomical projections. Leads are categorized into four regions: anterior, lateral, inferior, and right atrial/septal. For each heart condition, we generate attention weights over these lead groups to guide the network's focus toward the most informative regions, thereby enhancing the model's explainability. The attention weights are derived from established cardiology books [36,37]. Importantly, a single disease may manifest in multiple lead groups, and in cases where multiple abnormalities are present in one ECG signal, their respective attention weights are combined using a logical OR operation. The resulting lead-specific attention distribution is presented in Table 1, which also details the dataset prevalence of each condition.

Table 2
Overview of standard ECG lead groupings based on their anatomical projection on the heart surface.

Leads	Location
V3, V4	Anterior surface
I, aVL, V5, V6	Lateral surface
II, III, aVF	Inferior surface
V1, V2, aVR	Right atrium + Septal

2.2.2. Rhythm encoder

Both temporal and spatial features play a crucial role in effectively analyzing time-series data such as ECG signals. Temporal features capture the progression of electrical activity over time, while spatial features reflect variations across different leads, which correspond to different regions of the heart. In this work, we carefully design the encoder to extract both types of features to improve the classification of ECG abnormalities.

To capture these characteristics, four identical parallel encoder branches are developed, each dedicated to one group of ECG leads. The architecture of the encoder is shown in Fig. 2. Each encoder branch is composed of eight Multiresolution Residual (MultiRes) blocks. These blocks are designed to capture both short and long-term temporal features using multiple parallel convolutions with a fixed kernel size of 15. Following the design principles in Ibtehaz et al. [38], and Showrav et al. [39], the convolutional sequence is arranged so that, despite using only 15 sized kernels, the resulting feature combination effectively captures the equivalent receptive fields of 15, 29, and 43. This configuration allows the network to extract multi-resolution temporal features efficiently without increasing the number of parameters. As can be seen in Fig. 2, within each block, the input is first projected to a higher-dimensional feature space and then processed through sequential convolutions with progressively increasing channel widths. Specifically, the sequential convolution block divides the total number of block

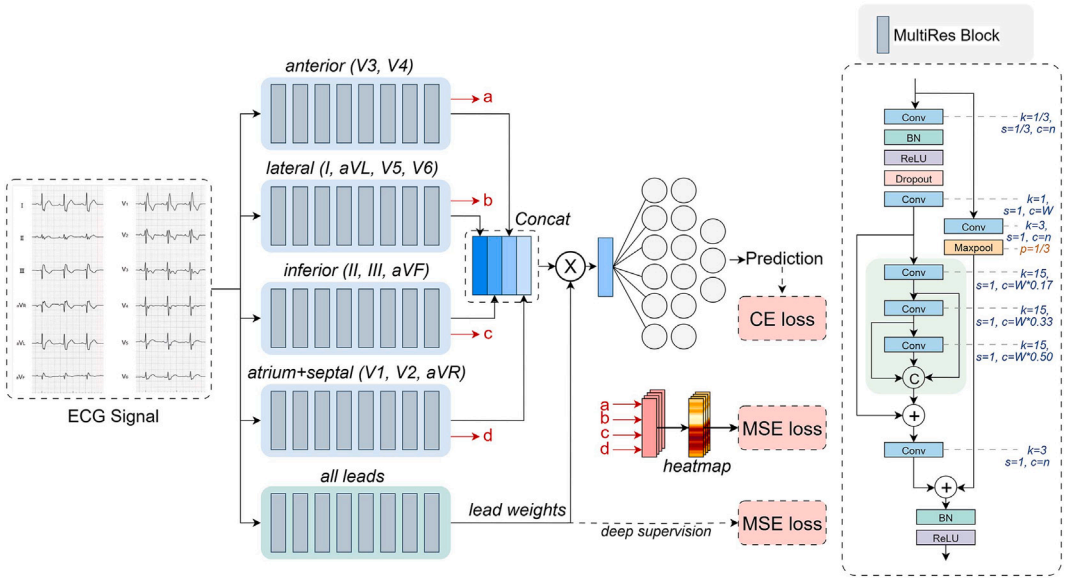


Fig. 2. Detailed architecture of the Rhythm Encoder, which directly processes the raw ECG signal. The design is driven by medical knowledge, incorporating lead group attention based on the anatomical placement of the ECG leads. Besides, the network has train-time XAI guidance capability to focus on clinically important regions.

output channels into three proportions of approximately 17%, 33%, and 50% for the first, second, and third convolutional paths, respectively, following the channel allocation strategy for efficient multi-scale feature extraction shown in Ibtehaz et al. [38]. The outputs of these convolutions are concatenated to form a rich representation across scales. Internal and external skip connections are used to maintain consistent dimensions, enable efficient feature reuse, and stabilize gradient flow. Odd-numbered blocks 1, 3, 5, and 7 apply downsampling by a factor of three, while even-numbered blocks 2, 4, 6, and 8 preserve the original resolution. Starting from 5000 samples per lead and per signal, the temporal dimension is gradually reduced along the encoder to extract global features while retaining important local information. To improve spatial resolution along the lead dimension, the network uses separate branches for different lead groups, capturing relationships within each group.

Following the encoding process, outputs from all four branches are concatenated and modulated using lead-specific attention weights guided by clinical importance. This allows the network to focus on the most relevant spatial regions (i.e., lead groups) based on the disease being analyzed. This operation can be expressed as

$$y_{\text{out}} = \overbrace{\begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \end{bmatrix}}^{G} \odot \overbrace{\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}}^{\alpha} = \sum_{j=1}^{4} g_j a_j \tag{3}$$

where G represents the encoded feature vectors from the four branches (spatial groups), α denotes the clinical importance-based attention weights for each lead group, and $\odot$ is the dot product. The weighted output y_{out} is then passed through a global average pooling layer, followed by three fully connected layers to generate the prediction.

2.2.3. XAI guidance during training

XAI plays a crucial role in evaluating the reliability and trustworthiness of predictive deep learning models, particularly in healthcare applications. By addressing the "black box" nature of these models, XAI provides valuable insights into the decision-making process. It allows researchers to verify whether the model's predictions are based on clinically relevant and meaningful regions of the input data. While XAI enhances interpretability and supports confidence in a model's decisions, it does not by itself improve the model's ability to focus on diagnostically significant regions or make it inherently more reliable.

To overcome this limitation, Showrav et al. [40] introduced EXGnet, a novel architecture that integrates XAI guidance, referred to as enforced XAI, directly into the model training process, in contrast to the prevailing literature. In EXGnet, class activation maps (CAMs) are generated within the network and refined using automatically constructed CAM ground truths derived based on heart rate variability (HRV). This approach aligns the model's attention with physiologically meaningful patterns. In the present work, we extend this idea by enriching the CAM ground truths with additional diagnostic cues that capture clinically relevant abnormalities. These include the absence of P waves, wide QRS complexes, PR interval variations, and RSR′ patterns, in addition to HRV. The methodology for constructing these enriched CAM ground truths is illustrated in Fig. 3. Since large-scale validation would require manually annotated CAMs produced by expert cardiologists, which is highly resource-intensive, we instead focused on demonstrating the correctness of our proposed approach. Specifically, we selected four abnormality classes from the CPSC dataset: Atrial Fibrillation, First Degree AV Block, Ventricular Ectopics or PVC, and Right Bundle Branch Block.

To automate the ground truth generation, we employed the Neurokit2 [41] library to detect the P, Q, R, S, and T wave positions. Because these waves are most clearly visible in lead II of the ECG, their locations were first identified in lead II and then propagated to generate CAMs for the corresponding lead groups. Identifying ectopic beat locations in PVC-labeled ECG signals posed a unique challenge, since no direct method exists in the Neuorokit2 library. To address this, we extracted all beats from the signal and applied K-means clustering with $k = 2$ to separate normal and ectopic beats, thereby determining their positions. Fig. 3 shows representative ground truth CAMs overlaid on ECG signals. In these CAMs, diagnostically important regions are labeled as 1, while all other regions are labeled as 0. Each CAM contains 61 values per lead, with each value representing a distinct segment of the ECG. As the rhythm encoder is organized into four lead groups, the ground truth CAMs are generated separately for each branch, resulting in four CAMs for a single twelve-lead ECG, each with 61 values.

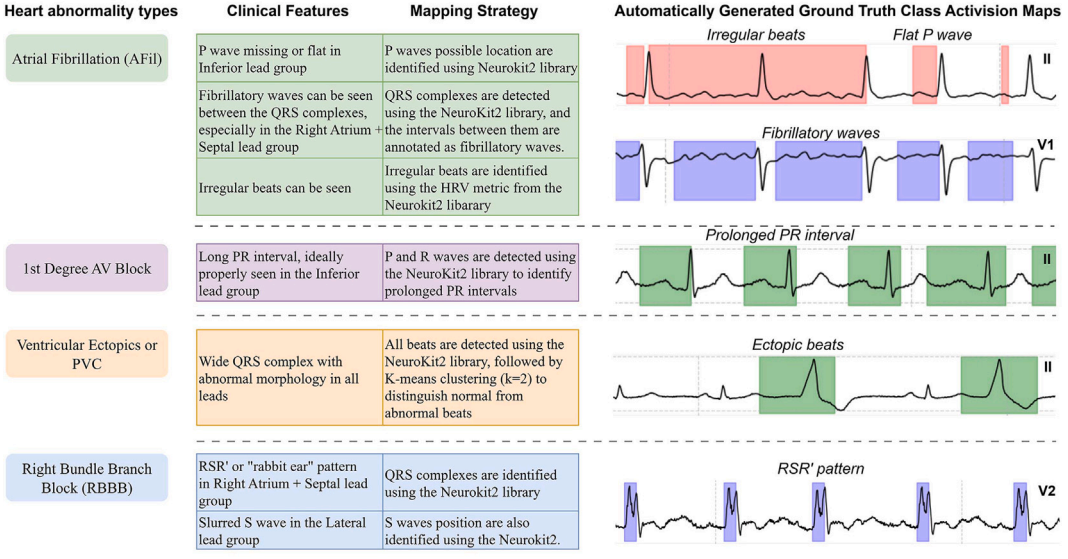


Fig. 3. Ground truth class activation map (P) generation strategy for the four ECG abnormalities: Atrial Fibrillation, First-degree AV Block, Ventricular Ectopics or PVC, and Right Bundle Branch Block.

Algorithm 1 Network integrated multi-branch CAM generation.

Input: Model M, Input Signal X, Target Layers $L_{targets}$
Output: List of CAMs for each branch H_{list}

procedure GENERATE MULTI-BRANCH CAM($M, X, L_{targets}$)
  $[A_1, \ldots, A_4], Y_{scores} \leftarrow \text{ForwardPass}(M, X, L_{targets})$
  $H_{list} \leftarrow []$
  for $i \in \{1, 2, 3, 4\}$ **do**
    $A_{branch} \leftarrow A_i$
    $Y_{branch} \leftarrow Y_{scores}[:, :, i-1]$ ▷ Select scores for the i-th branch
    $y^c \leftarrow \max(Y_{branch})$ ▷ Get the score of the predicted class
    $G \leftarrow \frac{\partial y^c}{\partial A_{branch}}$ ▷ Compute gradients via backpropagation
    $P \leftarrow A_{branch} \odot G$ ▷ Element-wise product
    $H_{raw} \leftarrow \text{mean}(P, \text{axis=channel})$ ▷ Average across channels
    $H \leftarrow \text{Normalize}(H_{raw})$ ▷ Scale heatmap values to [0, 1]
    Append H to H_{list}
  end for
  return H_{list}
end procedure

In our network, CAMs are estimated using a GradCAM-inspired technique. It is important to note that this approach requires access to gradients during training, which is typically restricted in the PyTorch environment. However, we devised an effective strategy to achieve this in the TensorFlow–Keras framework. The detailed CAM generation method is presented in Algorithm 1. Our strategy is inspired both by Showrav et al. [40] and by the original GradCAM [42] implementation. By supervising the model with ground-truth CAMs, we encourage it to concentrate on diagnostically significant regions, closely emulating the interpretive reasoning of expert cardiologists. This unique guided learning paradigm not only improves predictive accuracy but also enhances the model's inherent interpretability and clinical trustworthiness.

2.2.4. Feature encoder

Several existing studies have demonstrated the effectiveness of using quantitative ECG features such as P-wave count, QRS duration, heart rate, and other morphological and interval-based metrics [43]. Many cardiac abnormalities can be identified directly from these features, although more complex cases may still require analysis of the raw signal to capture temporal features. In addition, frequency-domain features are also valuable in ECG-based diagnosis [18]. Extracting quantitative features by decomposing the ECG into the frequency domain can further enhance the model's robustness and performance.

Motivated by these factors, we incorporate quantitative features derived from both the raw ECG signal and its frequency-domain representation alongside the primary signal encoder. For raw ECG-based features, we utilize the NeuroKit2 [41] library to extract a comprehensive set of physiological descriptors. In this work, we use 15 ECG waveform-based features and 17 HRV features. These 32 features are concatenated and passed through the first branch of the quantitative feature encoder. A complete list of features is provided in the Appendix A.

To extract features from the frequency domain, we apply VMD, which decomposes the ECG signal $f(t)$ into a finite number of band-limited intrinsic mode functions (IMFs), each associated with a center frequency ω_k. The goal is to minimize the total bandwidth of these IMFs while ensuring their sum reconstructs the original signal. The optimization problem is defined as

$$\min_{\{u_k\},\{\omega_k\}} \left\{ \sum_{k=1}^{K} \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \cdot e^{-j\omega_k t} \right] \right\|_2^2 \right\} \quad \text{subject to} \quad \sum_{k=1}^{K} u_k(t) = f(t) \tag{4}$$

Here, $*$ denotes convolution, $\delta(t)$ is the Dirac delta function, $||'||_2$ denotes l_2 norm, and $\frac{j}{\pi t}$ represents the Hilbert transform kernel. The exponential term $e^{-j\omega_k t}$ shifts each IMF to baseband. The constraint ensures that the summation of all IMFs reconstructs the original ECG signal, $f(t)$.

In VMD, the lower-order IMFs (e.g., IMF1) typically capture low-frequency components, while higher-order IMFs (e.g., IMF3 and IMF5) capture higher-frequency features, including those related to the P-wave, T-wave, and QRS complex. Based on this, IMF1, IMF3, and IMF5 are selected in this work. From each selected IMF of all leads, 13 statistical features (e.g., spectral moments, leakage measures) are extracted and passed to three parallel encoder branches. A complete list of these features is provided in the Appendix A.

The encoder branches for both raw ECG-based and frequency domain feature-based inputs follow an identical architecture. Each encoder branch consists of three sequential residual blocks, as illustrated in Fig. 4, followed by average pooling. The outputs from the final residual block are then passed to the feature refinement module. In this module, a linear attention mechanism is applied. It is implemented with two dense layers using ReLU and sigmoid activations along with residual multiplication. This highlights the most informative features and suppresses the less relevant ones. The refined features are then processed by a Transformer layer with four attention heads, which captures interactions among features and enhances the representation of complex ECG patterns.

To integrate information across domains, a cross-attention mechanism is applied. Here, the output S of the Transformer block representing raw ECG quantitative features, as in Fig. 4, is used as the *Query*. The concatenated outputs from the frequency-domain branches serve as the *Key* and *Value*. The output of the cross-attention is then passed through two dense layers for prediction. The ReLU-activated output of the first dense layer is also fed into a joint cross-attentive fusion network to generate the final prediction.

2.2.5. Joint cross-attentive fusion

We extracted both temporal and spatial features from raw ECG signals and quantitative features from the raw ECG signals as well as from the IMFs obtained via VMD. A strategic fusion of these feature sets can significantly enhance the network's ability to learn from both domains. To gain this advantage, we introduce a cross-attention mechanism between the two encoders.

As illustrated in Fig. 1, the intermediate features from the rhythm encoder (m) and the feature encoder (n) are passed to a Cross-Attentive Fusion module. Specifically, the output m from the rhythm encoder is used as the Query (Q), while the concatenation of m and n is used to form the Key (K) and Value (V) matrices. This configuration is motivated by the observation that raw ECG signals inherently carry fine-grained temporal and morphological patterns that are essential for interpreting cardiac dynamics. Using signal-derived features as the Query ensures that the attention mechanism is guided by the underlying physiological waveform, helping the model focus on diagnostically important segments such as P-waves, QRS complexes, and T-wave abnormalities. On the other hand, the concatenated features used as K and V include both raw signal and high-level quantitative representations, such as heart rate, number of P-waves, or statistical moments, which provide a richer and more informative context. This allows the attention mechanism to selectively retrieve and aggregate relevant contextual information in response to temporal cues from the signal. Consequently, this cross-attentive fusion enables the network to effectively integrate low-level signal features with high-level representations, resulting in a more comprehensive and context-aware representation that enhances the network's performance.

2.2.6. Automated report generation

Automated clinical report generation has gained significant attention in recent years, particularly with the advent and widespread adoption of LLMs. In the context of ECG analysis, generating automated diagnostic

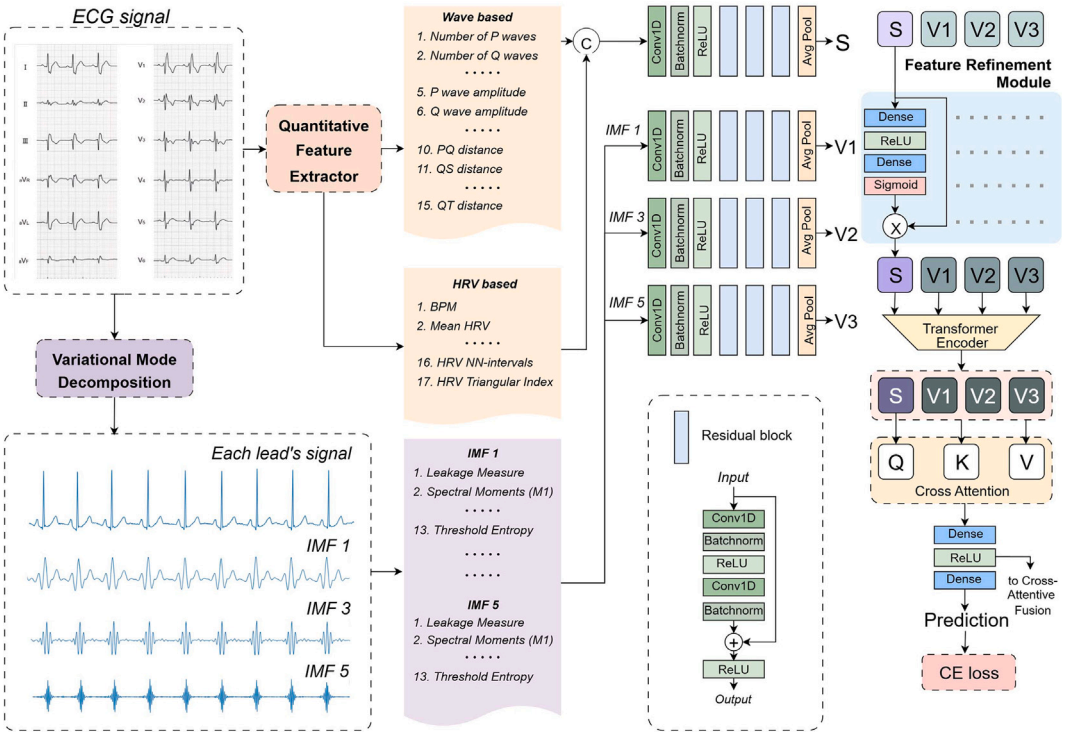


Fig. 4. Detailed architecture of the Feature Encoder. First, quantitative ECG features are extracted from both the signal domain and the frequency domain using variational mode decomposition. These features are then passed to the Feature Encoder for further processing.

reports is of critical importance; however, this area remains relatively underexplored in the existing literature within this ECG domain. In this research, we propose an automated ECG report generation framework that incorporates the capabilities of state-of-the-art LLMs. Fig. 1 illustrates a simplified block diagram of the proposed pipeline.

Through empirical evaluation, we observed that integrating powerful pretrained LLMs such as LLaMA or Gemini outperforms custom-trained small language models (SLMs) in both linguistic fluency and medical coherence. Hence, we employ LLMs via API integration for the final report generation. Nonetheless, our experiments indicate that relying solely on LLMs to generate diagnostic reports directly from raw ECG signals often leads to inconsistencies, raising reliability concerns. To address this limitation, we propose a structured prompting strategy that distinguishes our framework from existing approaches. Specifically, we construct prompts derived from our reliable and explainable diagnostic network, where CAMs and quantitative ECG features are integrated alongside the predicted diagnostic class.

The prompt is designed to follow a sequential structure. First, it provides the network's predicted diagnostic classes. To further enhance interpretability, CAMs generated using the Grad-CAM technique are incorporated to highlight regions of interest and communicate the reliability of predictions. This is followed by a set of quantitative ECG descriptors, including heart rate, heart rate variability, QRS duration, and other morphological features. Finally, a structured instruction is provided to guide the LLM in producing a coherent, clinically sound diagnostic report. Fig. 5 presents a representative prompt used in the proposed framework.

Prompt

You are a cardiologist analyzing an ECG recording. Generate a very concise medical report based on the following automated analysis.

Classification Results: {top_class_string }

Primary diagnosis: {class_name}

Model Attention Analysis (Grad-CAM): {cam_interpretation}

Quantitative Features: {Heart Rate, HRV SDNN, HRV RMSSD}

Instructions:
Generate a professional, and brief medical report.
1. State the primary diagnosis.
2. Briefly describe the findings that support the diagnosis, informed by the stamps from the medical attention analysis (Grad-CAM).
3. Recommend follow-up actions if necessary.

Fig. 5. A sample prompt for LLM-based report generation, integrating classification labels, primary diagnoses, Grad-CAM attention insights, and extracted ECG quantitative features.

2.3. Loss functions

To train the network for ECG abnormality classification, we used different loss functions depending on the nature of the task. For multiclass classification, we adopted the Categorical Cross-Entropy (CE) loss, while for multilabel classification, we employed the Binary Cross-Entropy (BCE) loss. These choices reflect the differing output structures: a single class per sample for multiclass tasks, and multiple simultaneous labels for multilabel settings.

The Categorical Cross-Entropy loss used for multiclass classification is defined as

$$L_{CE} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \tag{5}$$

where N is the number of samples, C is the number of classes, $y_{i,c}$ is the ground truth label, and $\hat{y}_{i,c}$ is the predicted probability for class c of sample i.

For multilabel classification, we used the Binary Cross-Entropy loss:

$$L_{BCE} = -\sum_{i=1}^{N}\sum_{c=1}^{C}\left[y_{i,c}\log(\hat{y}_{i,c}) + (1-y_{i,c})\log(1-\hat{y}_{i,c})\right] \tag{6}$$

which evaluates the predictions for each label independently, allowing multiple labels to be active for a single sample.

In addition to the classification loss, we incorporated the Mean Squared Error (MSE) loss to supervise the lead attention values. This auxiliary objective encourages the network to align its learned attention maps with the ground truth lead-specific relevance. The MSE loss is defined as

$$L_{MSE}^{lead} = \frac{1}{M}\sum_{i=1}^{M}\left(A_i - B_i\right)^2 \tag{7}$$

where A and B represent the ground truth and predicted attention values, respectively, M is the batch size. The ground-truth attention values, denoted by A, are provided in Table 1, specifically in columns 2 through 5, which are collectively labeled as "Lead Attention." To provide XAI guidance during training, we employ a MSE loss between the estimated and ground truth CAM values. This loss can be expressed as

$$L_{\text{MSE}}^{cam} = \frac{1}{D}\sum_{i=1}^{D}\frac{1}{C}\sum_{j=1}^{C}\left(P_{i,j} - Q_{i,j}\right)^2 \tag{8}$$

where P and Q denote the ground truth and predicted CAM values, C is the number of spatial locations in each CAM, and D is the batch size. The ground truth CAM values P are obtained using the methodology explained in Section 2.2.3 and Fig. 3. This auxiliary supervision encourages the network to focus its attention on diagnostically relevant regions.

The total loss function combines the classification and attention supervision losses from various components of the network:

$$L_{total} = \alpha\cdot L_{cls}^{rhythm} + \beta\cdot L_{cls}^{feature} + \gamma\cdot L_{cls}^{fusion} + \delta\cdot L_{MSE}^{lead} + \phi\cdot L_{MSE}^{cam} \tag{9}$$

where L_{cls}^{rhythm} is the classification loss from the rhythm encoder, $L_{cls}^{feature}$ is from the feature encoder, and L_{cls}^{fusion} is from the joint cross-attentive fusion network. Each L_{cls} term corresponds to either CE or BCE, depending on whether the classification setting is multiclass or multilabel. In our experiments, we found the optimal weights for the loss components to be: $\alpha = 1$, $\beta = 1$, $\gamma = 2$, $\delta = 1$, and $\phi = 1$ when XAI guidance is applicable; otherwise, $\phi = 0$.

3. Results

3.1. Implementation details

All experiments in this research were conducted using an NVIDIA RTX 4090 GPU. The dataset was split into three parts: training, validation, and test in an overall split ratio of 70:10:20, respectively. No label-wise stratification was applied, as in some multi-label datasets, certain label combinations occur only once in the entire dataset. The models were optimized using the Adam optimizer with an initial learning rate of 0.0001 and a batch size of 8. Each network was trained for a maximum of 200 epochs, as empirical observations of the training and validation loss curves indicated convergence within this range. To further refine learning in the later stages of training, a Cosine Annealing learning rate scheduler was employed starting from epoch 100, with a reduced initial learning rate of 0.00005 applied over the subsequent 50 epochs.

We investigated several ECG-specific augmentation strategies to address class imbalance, including amplitude scaling, time warping, random temporal shifting, shape-based deformations, and oversampling methods such as SMOTE. These approaches did not meaningfully improve performance, likely due to the morphology-sensitive nature of ECG signals. Augmentations can distort clinically important features, such as interval durations or voltage criteria, producing samples that are physiologically inconsistent and add noise rather than useful variability. GAN-based methods were also not pursued, as multi-label ECG generation is computationally expensive and often fails to preserve dependencies among co-occurring arrhythmias. Consistent with prior observations [44], such methods were excluded to maintain clinical realism and true class prevalence.

3.2. Evaluation metrics

To comprehensively evaluate the performance of the proposed method and compare it with existing approaches in the literature, we adopted four widely used evaluation metrics: Accuracy, Macro-F_1, Macro-F_2, and Area Under the ROC Curve (AUC). Each metric offers unique insights into the classification performance, particularly in imbalanced and multi-class settings.

Accuracy quantifies the overall correctness of the model by computing the proportion of correctly classified instances across all classes:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. It is important to note that all accuracy results reported for the multilabel setting are computed using subset accuracy, also referred to as the exact match ratio (Tables 3–6).

Macro-F_1 is the unweighted average of F_1-scores computed independently for each class. It treats all classes equally, regardless of their frequency, making it suitable for imbalanced datasets:

$$\text{Macro-F}_1 = \frac{1}{C}\sum_{i=1}^{C}\frac{2\cdot\text{Precision}_i\cdot\text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{11}$$

where C is the total number of classes.

Macro-F_2 extends Macro-F_1 by placing greater emphasis on recall. It is useful in applications where missing a positive instance is more critical than a false positive:

$$\text{Macro-F}_2 = \frac{1}{C}\sum_{i=1}^{C}\frac{5\cdot\text{Precision}_i\cdot\text{Recall}_i}{4\cdot\text{Precision}_i + \text{Recall}_i} \tag{12}$$

AUC measures the model's ability to distinguish between classes. It is threshold independent and particularly informative in multilabel or imbalanced settings. A higher AUC indicates better separability between positive and negative classes.

By using these metrics, we ensure a balanced and rigorous evaluation of the proposed method, highlighting both its overall accuracy and its ability to handle class imbalance and multiclass discrimination.

3.3. Comparative results

We compare the performance of our proposed model against six recent state-of-the-art methods: MVMSNet [22], DMFSNet [45], MobileNetViT [17], MRM [13], LFGNet [15], and BarlowTwin [23]. Evaluation is conducted on four publicly available ECG datasets under both multilabel and multiclass classification settings. We report four standard metrics: accuracy, macro-averaged F_1, macro-averaged F_2, and AUC, all expressed as percentages. These metrics provide a balanced view of both overall correctness and the model's ability to handle label imbalance and the clinical priority of positive predictions. We utilize four datasets for multilabel evaluation: CPSC, Chapman, Georgia, and PTB-XL. For multiclass evaluation, we exclude PTB-XL due to limited label diversity and focus on CPSC, Chapman, and Georgia, which provide adequate class granularity. Table 3 through Table 6 present the performance of all models across the multilabel datasets. While Table 7

Table 3
Quantitative comparison on the CPSC dataset under the multilabel setting.

Models	Accuracy	macro-F_1	macro-F_2	AUC
DMFSNet [45]	68.61	72.06	70.48	95.19
MVMSnet [22]	67.81	71.05	69.46	93.38
MRM [13]	66.49	68.93	66.27	95.44
LFGnet [15]	75.07	76.59	74.89	95.84
MobileNetViT [17]	53.68	51.19	48.90	88.45
BarlowTwin [23]	69.30	75.18	72.23	94.43
Ours	78.73	77.10	76.31	95.30

Table 4
Quantitative comparison on the Chapman dataset under the multilabel setting.

Models	Accuracy	macro-F_1	macro-F_2	AUC
DMFSNet [45]	66.94	68.05	66.24	94.71
MVMSnet [22]	67.11	68.92	67.38	95.36
MRM [13]	61.93	64.40	62.49	95.14
LFGnet [15]	67.14	68.25	67.29	94.56
MobileNetViT [17]	54.38	48.78	47.50	91.36
BarlowTwin [23]	67.46	71.14	68.88	95.83
Ours	70.69	72.91	70.44	96.18

Table 5
Quantitative comparison on the Georgia dataset under the multilabel setting.

Models	Accuracy	macro-F_1	macro-F_2	AUC
DMFSNet [45]	33.56	39.61	37.48	87.18
MVMSnet [22]	29.33	35.03	33.03	84.59
MRM [13]	31.06	35.56	35.56	85.94
LFGnet [15]	32.79	36.33	32.64	86.88
MobileNetViT [17]	24.75	26.27	26.44	78.59
BarlowTwin [23]	30.27	42.75	40.20	85.80
Ours	35.78	43.59	41.84	91.83

Table 6
Quantitative comparison on the PTB-XL dataset under the multilabel setting.

Models	Accuracy	macro-F_1	macro-F_2	AUC
DMFSNet [45]	51.76	48.63	45.41	88.28
MVMSnet [22]	52.75	51.88	50.18	91.74
MRM [13]	51.82	42.74	39.54	91.46
LFGnet [15]	53.83	49.26	46.72	91.72
MobileNetViT [17]	50.39	42.07	39.82	87.78
BarlowTwin [23]	54.02	53.06	51.39	92.49
Ours	53.09	53.25	51.48	92.11

Table 7
Quantitative comparison on the CPSC dataset under the multiclass setting.

Models	Accuracy	macro-F_1	macro-F_2	AUC
DMFSNet [45]	77.54	72.36	72.06	95.47
MVMSnet [22]	78.09	73.13	72.41	95.73
MRM [13]	70.74	63.81	63.94	93.36
LFGnet [15]	77.62	72.53	72.37	95.79
MobileNetViT [17]	61.89	57.31	56.82	88.03
BarlowTwin [23]	75.43	68.77	68.51	94.05
Ours	82.24	78.55	78.67	96.26

through Table 9 present the performance of all models across the multiclass datasets. Additionally, since ground truth CAMs are not available for all datasets and their manual annotation would require substantial time and resources, we do not include this approach in our comparative results. However, the effectiveness of this strategy is demonstrated in a separate section (refer to Section 3.4).

Table 8
Quantitative comparison on the Chapman dataset under the multiclass setting.

Models	Accuracy	macro-F_1	macro-F_2	AUC
DMFSNet [45]	95.01	87.48	87.12	99.12
MVMSnet [22]	91.60	85.52	85.36	98.78
MRM [13]	86.78	69.02	70.31	95.19
LFGnet [15]	91.11	78.49	78.94	97.21
MobileNetViT [17]	82.71	64.81	65.46	94.24
BarlowTwin [23]	93.18	88.14	88.26	98.93
Ours	96.26	89.82	90.36	99.29

Table 9
Quantitative comparison on the Georgia dataset under the multiclass setting.

Models	Accuracy	macro-F_1	macro-F_2	AUC
DMFSNet [45]	65.85	45.88	46.21	88.06
MVMSnet [22]	63.27	41.89	42.34	85.28
MRM [13]	57.25	32.79	34.21	79.82
LFGnet [15]	63.14	42.02	41.42	86.03
MobileNetViT [17]	56.39	30.40	30.34	78.39
BarlowTwin [23]	64.37	38.87	39.21	84.26
Ours	67.69	52.05	51.67	88.02

3.3.1. *Multilabel classification*

On the CPSC dataset, LFGNet demonstrated strong performance among the compared methods (Table 3), achieving an accuracy of 75.07%, a macro-F_1 score of 76.59%, a macro-F_2 score of 74.89%, and an AUC of 95.84%. In contrast, MobileNetViT exhibited the weakest performance across all evaluation metrics. Our proposed model outperforms all existing methods, attaining the highest accuracy of 78.73%, a macro-F_1 score of 77.10%, a macro-F_2 score of 76.31%, and a competitive AUC of 95.30%. In comparison with LFGNet, which ranked second in performance, our model shows a significant improvement in accuracy by 3.66% (p <0.001) and in macro-F_1 by 0.51% (p <0.005). The F_2 score is a metric used to evaluate the performance of a classification model, particularly when false negatives are considered more costly than false positives. In this regard, our model demonstrates superior performance, improving upon LFGNet's macro-F_2 score of 74.89% by a margin of 1.42% (p <0.001), which underscores the model's effectiveness in clinically sensitive scenarios.

On the Chapman dataset, our model consistently outperforms all compared methods, achieving an accuracy of 70.69%, a macro-F_1 score of 72.91%, and a macro-F_2 score of 70.44%. The most competitive baseline, BarlowTwin, attains an accuracy of 67.46%, a macro-F_1 score of 71.14% and a macro-F_2 score of 68.88%, significantly trailing behind our model by 3.23% (p <0.001), 1.77% (p <0.001) and 1.56% (p <0.001), respectively. Although LFGNet and MVMSnet also deliver respectable results, their relatively lower macro-F_2 scores suggest limitations in effectively capturing false negatives. The F_2 score, which places greater emphasis on recall by penalizing false negatives more heavily, reflects the clinical importance of sensitivity in detecting cardiac abnormalities. In terms of AUC, our model achieves the highest score of 96.18%, surpassing all other methods (Table 4).

The Georgia dataset poses a substantial challenge due to its high sparsity and limited class co-occurrence, which often hampers the performance of conventional multilabel classification methods. Despite these difficulties, our model consistently outperforms all competing methods across all evaluation metrics (Table 5). Specifically, our approach achieves the highest AUC of 91.83%, surpassing the second-best model, DMFSNet, which reaches 87.18%. This AUC improvement of 4.65% is statistically significant (p <0.001), underlining the robustness of our model in capturing meaningful clinical patterns under sparse label conditions. Beyond AUC, our model also attains the highest macro-F_1 score of 43.59% and macro-F_2 score of 41.84%, demonstrating superior balance between precision and recall, especially for rare but clinically

critical conditions. In contrast, although BarlowTwin achieves a macro-F_1 of 42.75%, its macro-F_2 drops to 40.20%, and its AUC remains substantially lower at 85.80%. Moreover, our model achieves the best classification accuracy of 35.78%, highlighting its comparatively better performance not only in probabilistic estimation (AUC) but also in discrete prediction performance. The combination of a relatively high AUC alongside a lower accuracy is expected and reflects the extreme class sparsity in the dataset, as summarized in Table 1. Overall, these results demonstrate the all-around superiority of our method on this challenging dataset.

The PTB-XL dataset presents a relatively balanced challenge in multilabel ECG classification, with many methods achieving competitive performance across key metrics (Table 6). Notably, BarlowTwin secures the top AUC of 92.49%, closely followed by our model at 92.11%. Although the absolute margin is narrow, our method achieves the best overall macro-F_1 (53.25%) and macro-F_2 (51.48%) scores, slightly edging out BarlowTwin in sensitivity to rare but important classes. In terms of classification accuracy, our model (53.09%) performs comparably to BarlowTwin (54.02%), with only a marginal difference of 0.93%. However, this slight trade-off in accuracy comes with improved balance between precision and recall, particularly in macro-F_2, where our model maintains higher sensitivity. Other baselines such as MVMSNet and LFGnet, perform well in AUC and accuracy but fall short in F-measures, indicating weaker handling of class imbalance. Meanwhile, MobileNetViT and MRM underperform across most metrics, showing limited capacity to generalize effectively in multilabel settings.

Finally, our model demonstrates consistently strong performance across diverse and challenging ECG datasets, surpassing existing methods in key metrics such as macro-F_1, macro-F_2, and AUC. Its ability to maintain high sensitivity, particularly in clinically relevant settings where recall is critical, highlights its potential practical utility in real-world medical applications. These results affirm the robustness and generalizability of our approach for multilabel cardiac abnormality detection.

3.3.2. Multiclass classification

To further assess the generalization ability of our model, we evaluate its performance on multiclass classification tasks using the CPSC, Chapman, and Georgia datasets. Each dataset presents distinct characteristics in terms of class distribution, complexity, and signal variability, enabling a comprehensive comparison of method robustness across diverse clinical settings. The CPSC dataset includes a balanced distribution of common cardiac arrhythmias, making it a suitable benchmark for multiclass ECG classification. Our proposed model consistently outperforms all baseline methods (Table 7), achieving the highest accuracy of 82.24%, a macro-F_1 score of 78.55%, a macro-F_2 score of 78.67%, and an AUC of 96.26%. Compared to MVMSnet, the second-best model in terms of accuracy, macro-F_1 and F_2, our method shows notable improvements of 4.15% in accuracy, 5.42% in macro-F_1 and 6.26% in macro-F_2, both statistically significant (p <0.001). These gains are indicative of better class separation and improved sensitivity to subtle signal variations. Models like MobileNetViT and MRM perform poorly across all metrics, likely due to insufficient feature extraction or underfitting, demonstrating limited capability for fine-grained classification in complex ECG signals.

The Chapman dataset contains high-quality ECG recordings and well-separated arrhythmia classes, enabling most models to perform competitively. In this context, our model achieves the best performance across all metrics (Table 8), including an accuracy of 96.26%, a macro-F_1 score of 89.82%, a macro-F_2 score of 90.36%, and an AUC of 99.29%. Among the baselines, DMFSNet ranks second in both accuracy (95.01%) and AUC (99.12%), indicating strong overall classification capability. Meanwhile, BarlowTwin performs second-best in terms of macro-F_1 (88.14%) and macro-F_2 (88.26%), reflecting its proficiency in balancing precision and recall. Compared to BarlowTwin, our model achieves relative improvements of 1.68% in macro-F_1 and 2.10% in macro-F_2, while also slightly surpassing DMFSNet in accuracy and AUC. These improvements suggest our model's superior ability to capture both dominant and subtle diagnostic patterns, particularly in minimizing false negatives. Other models, such as LFGNet and MVMSNet, demonstrate moderate performance, with adequate accuracy but larger drops in macro-F_2.

The Georgia dataset presents a particularly challenging multiclass classification scenario due to lower signal quality and a higher prevalence of rare diagnostic categories. Despite these challenges, our model achieves the best performance across all key metrics except AUC (Table 9), obtaining the highest accuracy (67.69%), macro-F_1 score (52.05%), and macro-F_2 score (51.67%). In comparison, DMFSNet ranks second in these metrics with 65.85% accuracy, 45.88% macro-F_1, and 46.21% macro-F_2. Our model achieves relative improvements of 1.84% in accuracy (p <0.001), 6.17% in macro-F_1 (p <0.001), and 5.46% in macro-F_2 (p <0.001), highlighting its stronger performance in both precision and recall. However, DMFSNet records the highest AUC (88.06%), slightly surpassing our model's AUC of 88.02%. This significantly narrow margin (p >0.05) suggests comparable overall class ranking ability between the two models, though our model offers superior discrete classification performance and greater sensitivity to rare conditions. Other baselines, such as BarlowTwin and LFGNet, display limited capability in handling class imbalance, with substantially lower macro-F_1 and F_2 scores. MobileNetViT again exhibits the weakest results across all metrics, emphasizing its limited suitability in this complex diagnostic setting.

Across all three multiclass datasets, our model demonstrates superior performance in every evaluation criterion, highlighting its ability to effectively capture class-specific features and maintain sensitivity in both balanced and imbalanced diagnostic conditions.

3.4. Assessment of the impact of XAI guidance

To rigorously evaluate the effectiveness of XAI guidance during training, we conducted experiments with our proposed network both with and without this strategy. The experiment was conducted on a subset of the CPSC dataset, as shown in Fig. 3 and described in Section 2.2.3. Fig. 6 presents a comprehensive comparison, including classification metrics, confusion matrices, and representative CAMs superimposed on the ECG signals.

The quantitative results demonstrate a clear advantage of XAI guidance. The guided model achieved an overall accuracy improvement of 2.11% (p < 0.001) compared to the unguided version, alongside consistent gains in macro-F1, macro-F2, and AUC scores. These improvements confirm that XAI guidance positively influences the network's discriminative ability. The confusion matrices provide further insights into class-specific benefits. The largest improvement is observed in the Ventricular Ectopics or PVC class, where the number of correctly identified samples increased from 112 to 125 when XAI guidance was applied. Similarly, notable gains are observed for First-Degree AV Block (1D-AVB) and Right Bundle Branch Block (RBBB). Although overall performance for the Atrial Fibrillation (AFib) class remained stable, the guided model exhibited improved precision, as fewer other arrhythmias were misclassified as AFib. This indicates that the inclusion of XAI not only enhances accuracy but also reduces misleading predictions, thereby improving diagnostic reliability.

Fig. 6 also highlights two representative case studies from the modified CPSC test set, further illustrating the qualitative impact of XAI guidance. Case (a) represents a PVC signal. PVC signals are primarily characterized by the presence of ectopic beats, which are visually distinct from the other ECG beats. The unguided model failed to correctly classify this signal, and its CAM revealed minimal focus on the ectopic beats, the most critical diagnostic feature. In contrast, the XAI-guided model correctly identified the sample as PVC, with its CAM sharply highlighting the ectopic beats. This indicates that XAI guidance effectively directs the model's attention toward clinically relevant features, thereby improving both prediction accuracy and interpretability.

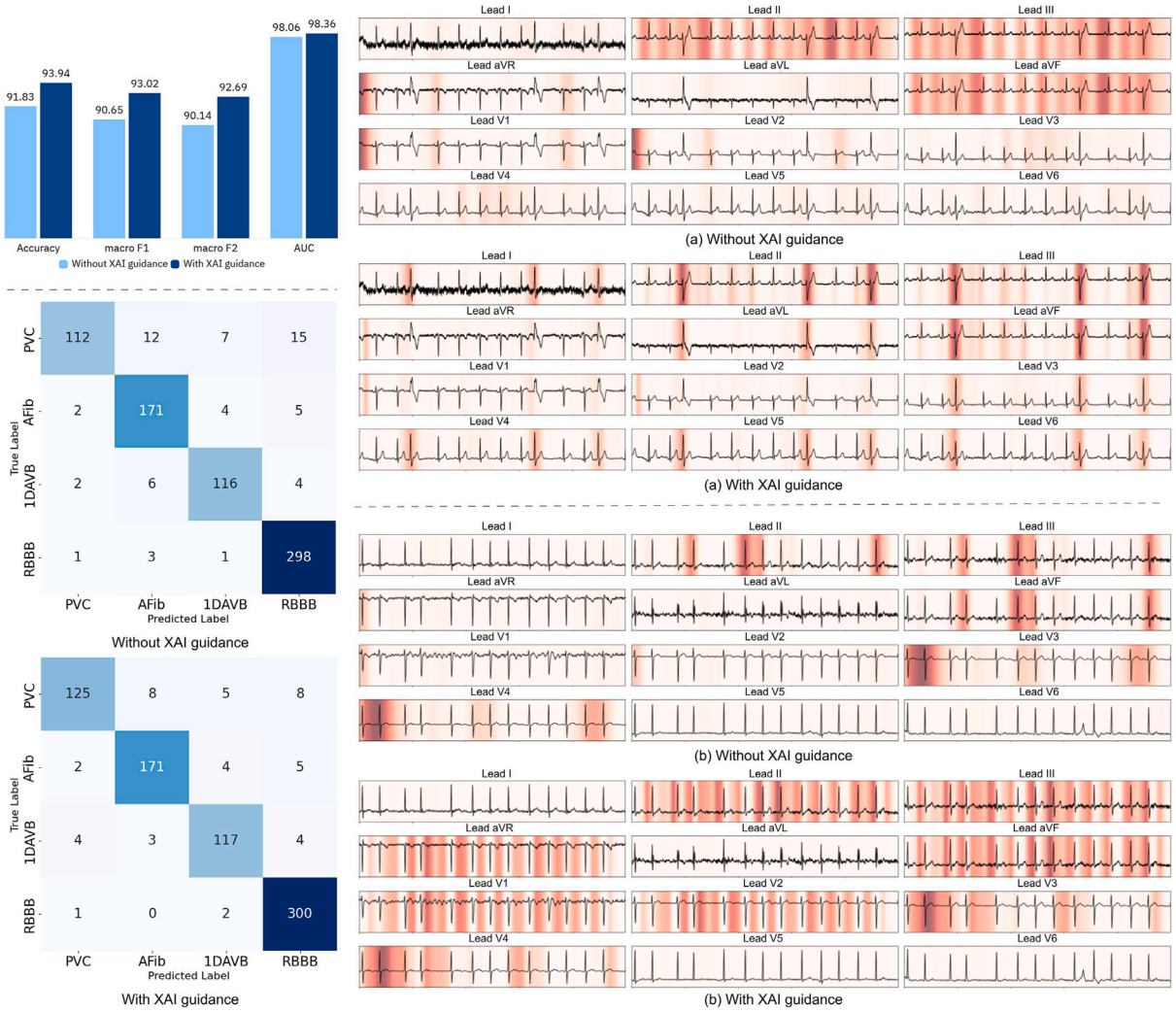


Fig. 6. Quantitative and qualitative visualization of the impact of XAI guidance. On the right, cases (a) and (b) are drawn from the modified CPSC test set. In case (a), originally labeled as PVC (Ventricular Ectopics), the XAI-guided model correctly identified the signal, whereas the non–XAI-guided model failed. In case (b), both models produced correct predictions; however, their attention patterns reveal that the XAI-guided model relied on more clinically meaningful regions, making its decision more reliable. The overlaid CAM on the signal highlights model attention, where darker regions indicate stronger focus and lighter regions indicate weaker focus.

Another interesting example is shown in case (b), which is originally an Atrial Fibrillation (AFib) signal. AFib is typically identified by the absence or flattening of P waves in the inferior leads (II, III, aVF), the presence of fibrillatory waves in leads V1-V2, and irregular beat-to-beat variability. Interestingly, both guided and unguided models correctly classified this signal. However, the CAM analysis reveals an important distinction: the unguided model predominantly focused on rhythm irregularities, while neglecting other diagnostic markers. In contrast, the XAI-guided model simultaneously attended to the P wave regions in inferior leads, fibrillatory wave activity in V1-V2, and irregular heart rhythms. This holistic focus more closely resembles the diagnostic reasoning of expert cardiologists.

These results confirm that the proposed innovative XAI guidance not only improves raw performance metrics but also enforces clinically meaningful attention patterns. By encouraging the model to attend to diagnostically critical regions, XAI guidance enhances the reliability, trustworthiness, and clinical interpretability of predictions. In practice, this is crucial, as clinical decision support systems must not only provide accurate outputs but also justify them in a manner consistent with expert reasoning. Thus, the incorporation of XAI guidance significantly increases the validity and acceptance of model predictions in real-world clinical applications.

3.5. Case-based analysis of generated reports

The ECG datasets used in this study focus on signal-level annotations such as rhythm labels and beat-level information, but they do not include corresponding textual clinical reports. As a result, ground-truth ECG reports are not available for direct comparison, and standard quantitative evaluation methods for text generation cannot be applied. This limitation is common in publicly available ECG datasets, as clinical reports are often not released due to privacy concerns and the high cost of expert annotation. To address this challenge, we collaborated with an expert cardiologist to support the evaluation of the generated reports. The cardiologist created reference clinical reports for a representative subset of ECG recordings, which were then used as expert-aligned ground truth for qualitative assessment. Importantly, the diagnostic content of the generated reports is primarily determined by the predictions of the core diagnostic network, while the large language model is used only to translate these predictions into structured clinical text.

Fig. 7 presents three representative cases comparing the reports generated by CARDxnosis with those written by the cardiologist. The automated and expert reports are shown side by side, with yellow-highlighted sections indicating similar or matching clinical statements.

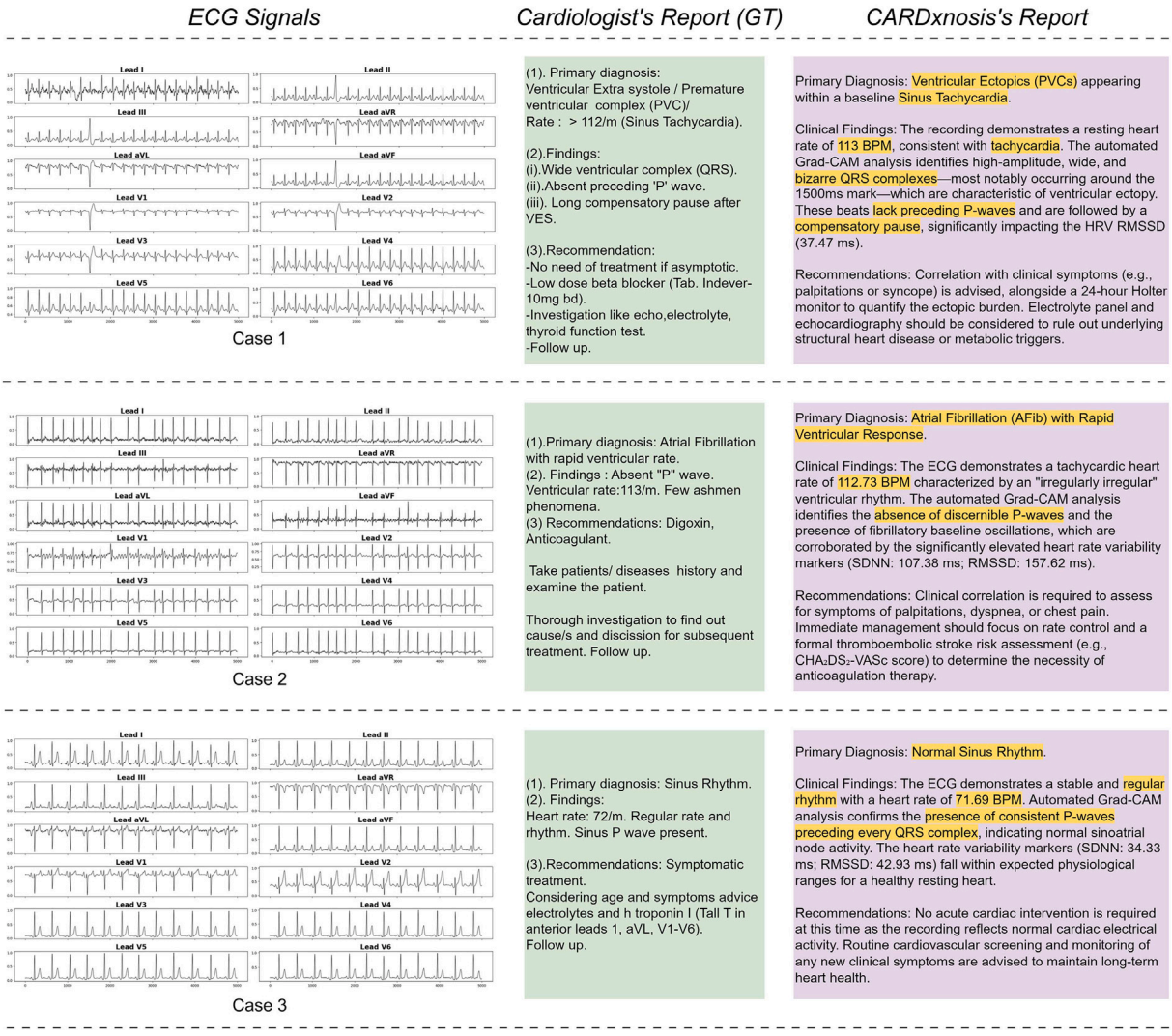


Fig. 7. Comparative analysis of ECG reports generated by CARDxnosis and an expert cardiologist (Ground Truth) across three clinical cases. Yellow highlights in the CARDxnosis reports indicate specific clinical findings, diagnostic terms, and heart rate measurements that closely align with the expert's assessment, demonstrating the model's accuracy in identifying critical diagnostic markers.

These highlights make it easier to visually assess agreement between the two reports. In the first case, both reports describe ventricular ectopic activity with sinus tachycardia, including premature ventricular beats and compensatory pauses. In the second case, both reports identify atrial fibrillation with rapid ventricular response, including consistent observations regarding absent P-waves and irregular R–R intervals. As can be seen in the third case, both reports describe a normal sinus rhythm with regular heart rate and normal QRS morphology. Across all cases, the highlighted similarities demonstrate that the generated reports capture the key diagnostic findings identified by the expert cardiologist.

4. Discussion

4.1. Ablation studies

A comprehensive ablation study was conducted to evaluate the contribution of each major design choice in the proposed method. Table 10 presents the results for the Georgia and Chapman datasets across four evaluation metrics in a multiclass setting: accuracy, macro-F_1, macro-F_2, and AUC. The configurations are incrementally built as follows: (a) a single-branch rhythm encoder, (b) a multi-branch rhythm encoder in which ECG leads are grouped based on their corresponding

Table 10
Ablation study demonstrating the effect of the design choices in the proposed method.

Setting	Georgia				Chapman			
	Accuracy	macro F1	macro F2	AUC	Accuracy	macro F1	macro F2	AUC
(a) Single branch rhythm encoder	63.76	44.28	43.58	86.84	93.60	85.94	86.21	98.66
(b) Lead grouping based rhythm encoder	65.11	46.70	46.09	87.25	94.51	87.92	88.13	98.89
(c) b + lead attention	65.84	48.22	47.68	87.91	95.34	88.49	88.75	99.06
(d) Feature encoder	46.07	34.73	33.16	77.42	84.46	67.12	66.83	96.45
(e) c + d + cross attention fusion	67.69	52.05	51.67	88.02	96.26	89.82	90.36	99.29

anatomical regions, (c) a multi-branch rhythm encoder with lead group attention to adaptively focus on the most informative anatomical lead groups, (d) a standalone feature encoder, and (e) the combination of (c) and (d) with cross-attention fusion between the rhythm and feature encoders.

Starting from the single-branch rhythm encoder (Setting a), replacing it with the anatomically grouped multi-branch design (Setting b) yields consistent improvements across both datasets, indicating that processing leads within their anatomical context enhances rhythm representation learning. Incorporating the attention of the lead group (Setting c) further improves performance, particularly in the macro-F_1 and macro-F_2 scores, enabling the model to emphasize the most diagnostically relevant anatomical lead groups. The standalone feature encoder (Setting d), which extracts complementary morphological and statistical descriptors from ECG signals, achieves lower performance when used in isolation compared to rhythm-based encoders, showing that rhythm features are more discriminative in this task. However, integrating the feature encoder with the attention-enhanced multi-branch rhythm encoder through cross-attention fusion (Setting e) produces the best results across all metrics and datasets. In the Georgia dataset, the full model achieves 67.69% accuracy, 52.05% macro-F_1, and 88.02% AUC, while in the Chapman dataset it reaches 96.26% accuracy, 89.82% macro-F_1, and 99.29% AUC, surpassing all intermediate configurations. These results confirm that each design choice contributes positively to the final model, with the largest gains achieved when anatomically grouped rhythm representations and complementary feature representations are jointly optimized via cross-attention fusion.

To identify the optimal weights for the loss calculation, we evaluated different sets of values for α, β, γ, δ, and ϕ across all datasets. For the Georgia Dataset, we present the results for a subset of representative

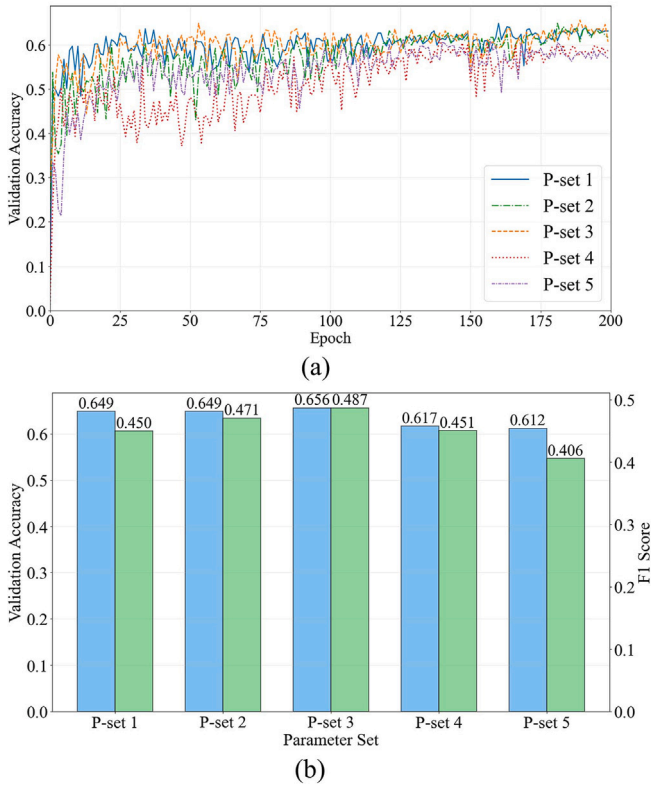


Fig. 8. Performance comparison across different hyperparameter weight configurations, denoted as P-set: $(\alpha, \beta, \gamma, \delta)$; (a) validation accuracy curves, (b) maximum validation accuracy and F1-score. The evaluated parameter sets are defined as follows: P-set 1: $(1, 1, 1, 1)$; P-set 2: $(1, 1, 2, 0.5)$; P-set 3: $(1, 1, 2, 1)$; P-set 4: $(1, 1, 0.5, 1)$; P-set 5: $(1, 2, 2, 1)$.

weight combinations in a multiclass classification task over 200 epochs in Fig. 8. In this experiment, XAI guidance was not incorporated, and thus ϕ was set to 0. The graph in Fig. 8(a) shows that combination set-3 achieves the best balance between rapid convergence and sustained high performance. Unlike set-1, which exhibits a slow initial rise, or set-4, which consistently lags in accuracy, the configuration $\alpha = 1$, $\beta = 1$, $\gamma = 2$, and $\delta = 1$ quickly reaches a high validation score within the first few epochs and maintains this performance throughout training. Fig. 8(b) further demonstrates that the chosen configuration achieves the highest accuracy and F1 score among the best results of all configurations, thereby justifying its selection. Based on these observations, this loss-weight configuration, $\alpha = 1$, $\beta = 1$, $\gamma = 2$, and $\delta = 1$ was used in the experiments reported in this study. This setting provided a stable balance between the primary diagnostic objective and the auxiliary optimization terms that guide intermediate representation learning. Overall, with the recommended parameter settings, the framework exhibits stable convergence across the multiple loss terms. Nevertheless, minor adjustments to these coefficients may be beneficial when applying the framework to different computational environments or datasets, as small variations can help accommodate changes in data distribution or training dynamics.

4.2. Generalizability of XAI-guided supervision

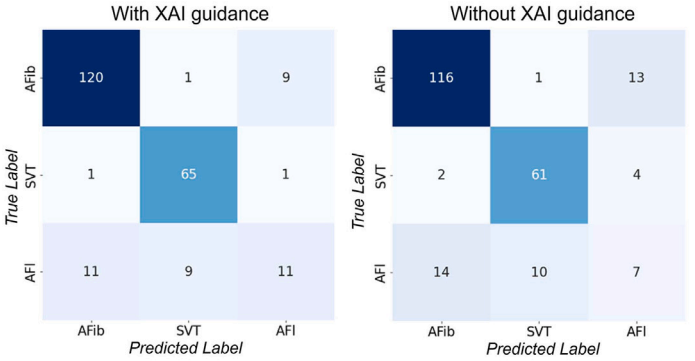
In this work, the primary evaluation was conducted on the CPSC dataset to establish the effectiveness of the proposed XAI-guided supervision framework. To examine the extent to which these benefits generalize, we performed additional experiments across different datasets and class configurations. These analyses are presented here to provide evidence on whether the observed improvements extend beyond the dataset used for the main evaluation. The XAI guidance mechanism itself is implemented as a supervised learning strategy that leverages ground truth attention information to highlight clinically relevant ECG regions. While such guidance would ideally be provided by expert cardiologists, large-scale expert-annotated CAMs are not currently available in public ECG datasets. To mitigate this limitation, we employed an automated CAM generation approach based on established ECG signal characteristics. Although these automated CAMs may introduce minor inaccuracies, they capture diagnostically meaningful regions and serve as effective guidance during training.

The first supplementary experiment was designed to assess dataset-level generalizability using the Chapman dataset, which includes three abnormality classes: Atrial Fibrillation, Supraventricular Tachycardia, and Atrial Flutter. Despite notable differences in data distribution and labeling protocols between the CPSC and Chapman datasets, the incorporation of XAI-guided supervision consistently improved classification performance and reduced inter-class confusion (see Fig. 9a). These results suggest that directing the model's attention toward clinically relevant ECG regions enhances robustness when applied to external datasets, thereby supporting the broader applicability of the proposed framework. A second experiment further examined the influence of XAI guidance in a multi-dataset, multi-class setting. In this configuration, the four supervised classes from the CPSC dataset were combined with two additional classes, Supraventricular Tachycardia from Chapman and Left Ventricular Hypertrophy from Georgia, which did not receive spatial supervision. For these unsupervised classes, CAM masks were set uniformly to all ones, effectively removing region-specific guidance. This design allowed us to consider whether supervision restricted to a subset of classes could still influence overall feature learning, an effect that is evident in the outcomes shown in Fig. 9(b).

The results demonstrate that models trained with XAI guidance achieve improved performance across all classes, including those without explicit spatial supervision. Notably, reductions in false positives and false negatives are observed even for Supraventricular Tachycardia and Left Ventricular Hypertrophy, where no attention masks were provided. This outcome indicates that XAI-guided supervision promotes the

Metrics	With XAI guidance	Without XAI guidance
Accuracy	85.96	80.70
F1	75.15	67.26
F2	74.90	67.44
AUC	87.04	85.47

(a)



Metrics	With XAI guidance	Without XAI guidance
Accuracy	93.11	90.85
F1	93.01	90.41
F2	92.72	90.09
AUC	98.99	98.42

(b)

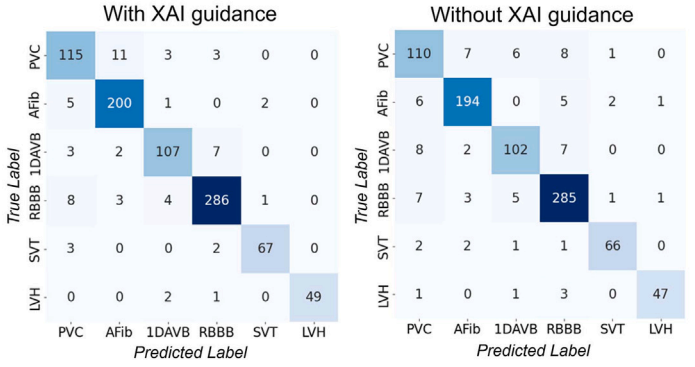


Fig. 9. Effect of XAI-guided supervision on model generalizability. (a) Chapman dataset results show improved classification and reduced inter-class confusion with XAI guidance. (b) Multi-dataset evaluation with four supervised CPSC classes and two unsupervised classes (SVT from Chapman, LVH from Georgia, using all-one CAM masks) demonstrates that XAI guidance on a subset of classes enhances overall learning, directly and indirectly benefiting both supervised and unsupervised classes.

learning of more meaningful and shared representations, which extend beyond the supervised classes. While the largest gains occur for the CPSC classes with explicit guidance, the consistent improvements observed for unsupervised classes highlight a broader influence on feature learning. Taken together, these findings suggest that the proposed framework is capable of generalizing across datasets and abnormality types, reinforcing its potential for wider application when clinically meaningful attention guidance, whether expert-defined or reliably automated, is available.

4.3. Interpretability analysis

To gain insight into how different models learn and separate cardiac rhythm classes, we perform a qualitative interpretability analysis using t-SNE visualizations of the feature embeddings extracted from the immediate final layers of each model. Fig. 10 presents the results for the three best-performing models, namely DMFSNet, LFGNet, and MVMSNet, along with the proposed method, on the CPSC test set.

Among the visualizations, our proposed method demonstrates the most distinct and compact clustering of rhythm classes, reflecting its strong discriminative capacity. Notably, classes such as “Atrial Fibrillation”, “Ventricular Ectopics”, “Right bundle branch block”, and “Left bundle branch block” form well-separated clusters, suggesting that the learned representations effectively capture class-specific patterns and minimize intra-class variance. In contrast, MVMSNet shows relatively good separation, but certain classes (e.g., “Atrial Fibrillation” and “Atrial premature beats”) still exhibit partial overlap, indicating a lack of a clear boundary in feature space. LFGNet reveals denser cluster formations for several classes, yet suffers from considerable overlapping between “Atrial premature beats”, “ST tilt up”, and “ST down”, which could lead to misclassification. Similarly, DMFSNet produces moderately clustered representations, but the boundaries are fuzzier, and several classes are entangled, indicating suboptimal feature separability. Overall, the t-SNE plots corroborate the quantitative results, highlighting the superior feature extraction ability of our model. The improved cluster separation validates that our architecture not only performs well numerically but also learns interpretable and discriminative features conducive to clinical reliability.

4.4. Computational complexity analysis

Understanding the computational complexity of a model is crucial for evaluating its suitability for real-world applications, particularly in scenarios with limited hardware resources. As shown in Table 11, our proposed method exhibits a well-balanced efficiency profile compared to state-of-the-art approaches. With 3.143 GFLOPs (Giga Floating Point Operations), 1.571 GMACs (Giga Multiply-Accumulate Operations), and a GPU latency of 31.41 ms, the model demonstrates superior efficiency relative to computationally intensive architectures such as DMFSNet (14.658 GFLOPs) and MobileNetViT (34.479 GFLOPs), while exceeding the most lightweight baselines, BarlowTwin (0.761 GFLOPs) and LFGNet (0.762 GFLOPs). Despite this, our model contains a substantial number of trainable parameters (21.997 million), second only to MRM (34.27 million), indicating a more sophisticated architecture. Its memory usage of 83.91 MB is moderate, positioned between memory-heavy models such as MRM (130.73 MB) and lightweight designs like MobileNetViT (2.10 MB). Overall, the comparison in Table 11 highlights that the proposed approach achieves significant performance improvements over existing methods while maintaining competitive computational efficiency, latency, and memory footprint. This combination makes the model particularly well-suited for practical ECG classification tasks, where both high performance and resource efficiency are essential.

4.5. Potential anomalies in datasets

During the data processing stage, several anomalies were identified across the datasets. One of the most prominent issues was the presence of leads with no recorded values in certain ECG recordings. In particular,

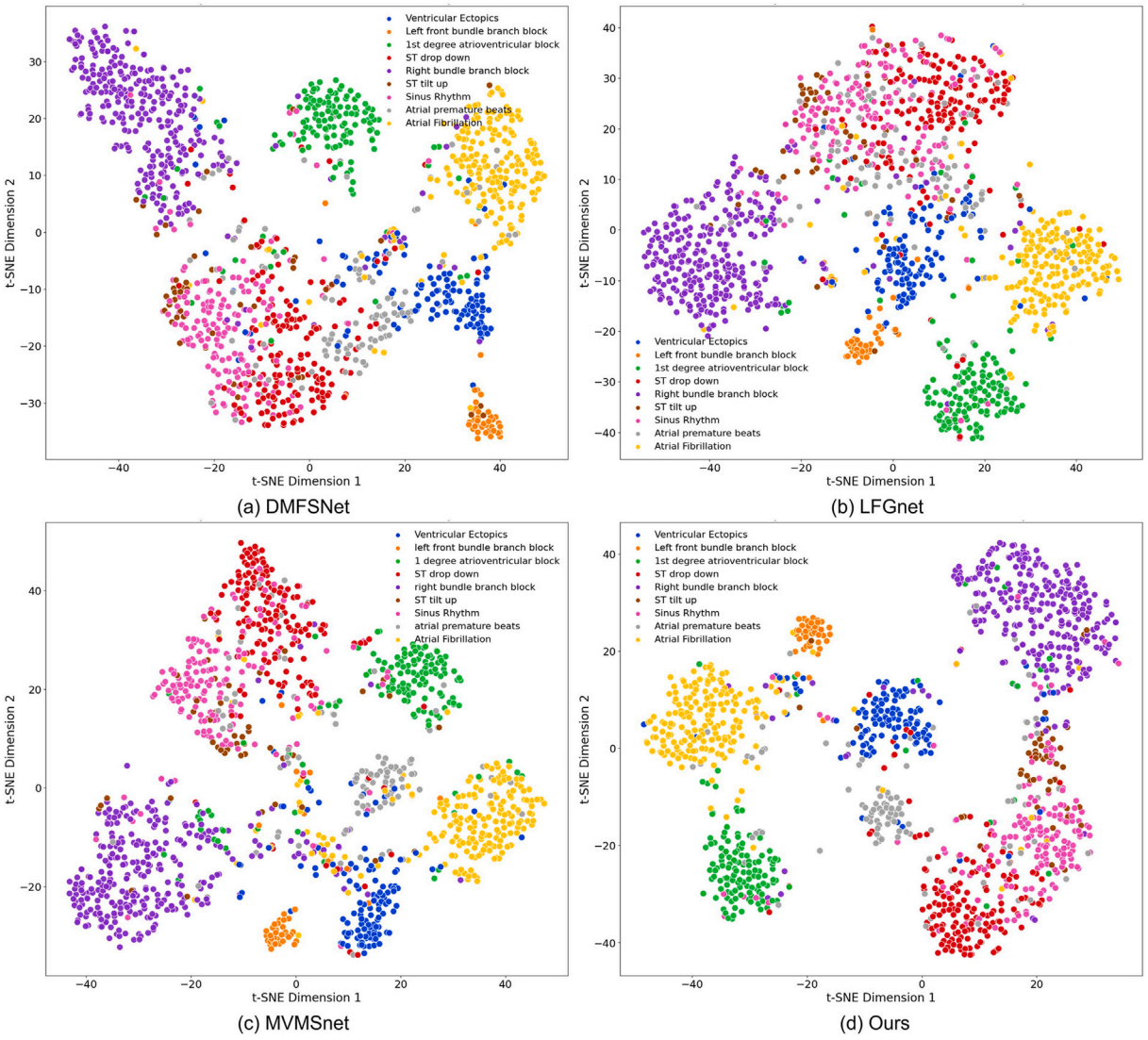


Fig. 10. t-SNE visualization of the feature representations from the final layer of the three best-performing models on the CPSC test set in a multiclass setting: (a) DMFSNet, (b) LFGNet, (c) MVMSNet, along with (d) the proposed method.

Table 11
Computational complexity comparison of the proposed method with state-of-the-art-methods.

Models	GFLOPs	GMACs	Parameters (Million)	Estimated GPU Latency (ms)	Total Memory (MB)
DMFSNet [45]	14.658	7.329	7.079	146.59	27.01
MVMSnet [46]	2.892	1.446	2.245	28.93	8.57
MRM [13]	5.531	2.765	34.27	55.32	130.73
LFGnet [15]	0.762	0.381	1.025	191.28	5.56
MobileNetViT [17]	34.479	17.239	0.666	344.79	2.10
BarlowTwin [23]	0.761	0.381	0.629	7.62	2.40
Ours	3.143	1.571	21.997	31.41	83.91

some leads in the Chapman, PTB-XL, CPSC, and Georgia datasets were found to be entirely blank, and these records were excluded during preprocessing. For example, Fig. 11(a), taken from the CPSC dataset, illustrates a case where lead V6 contains no data except a sharp spike at the end. A considerable number of ECG signals were found to have missing data in multiple chest leads. In addition, a noticeable subset of signals contained no discernible ECG waveform at all, as shown in Fig. 11(d), which is labeled as Atrial Fibrillation. Such cases pose a major challenge. While most signals with missing leads were automatically detected and removed from evaluation, ambiguous cases like those in Fig. 11(a) and (d) were retained, since they were not composed entirely of zero values.

Beyond missing data, another major source of concern was the presence of mislabeled records across all datasets. These misannotations significantly hindered model training, as they introduce label noise that may confuse the network during learning and prediction. Examples of such discrepancies are illustrated in Fig. 11. In case (a), the signal is labeled as Sinus Rhythm; however, clear irregular rhythms are present,

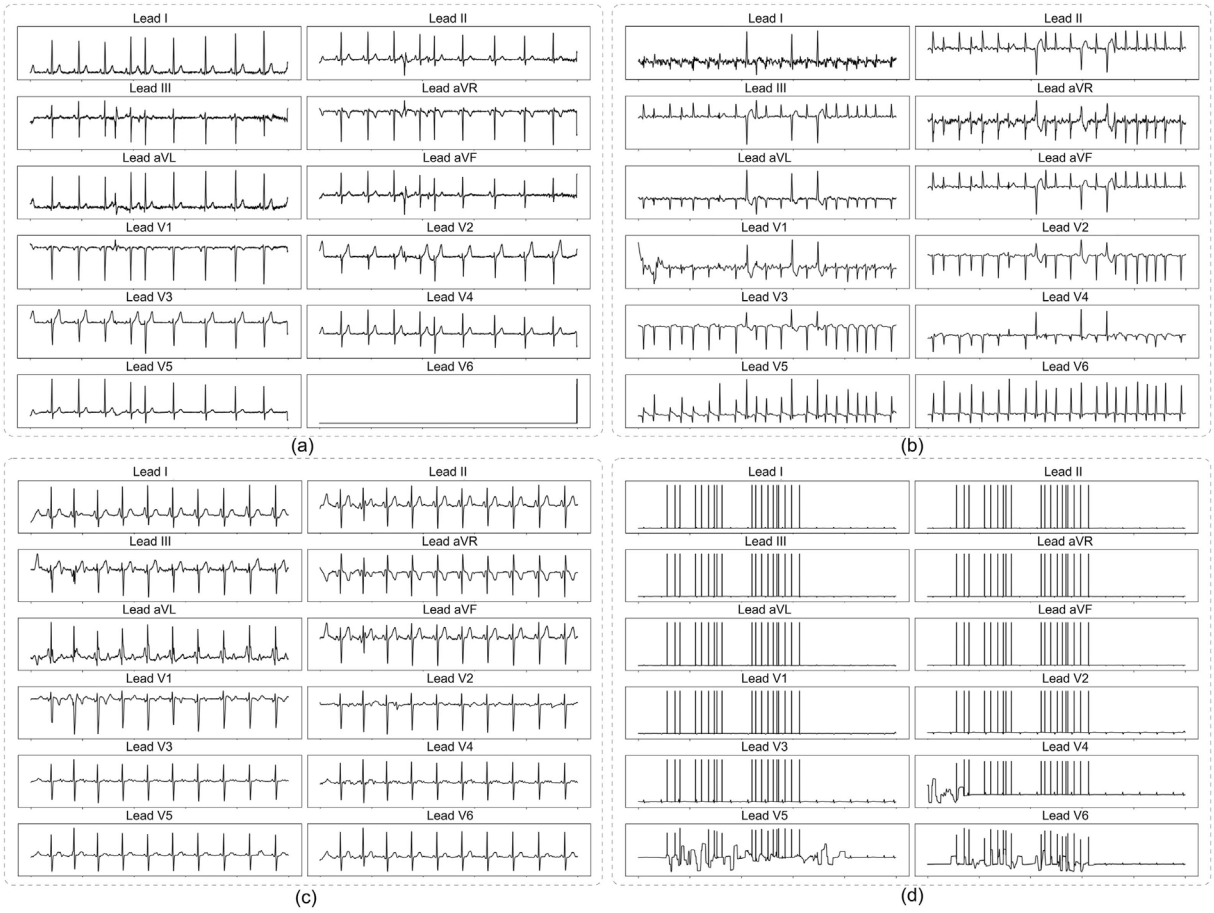


Fig. 11. Examples of anomalies identified in the ECG datasets (cases are taken from the CPSC dataset). (a) An ECG record labeled as Sinus Rhythm that displays irregular rhythms and a missing V6 lead. (b) A recording annotated as Atrial Fibrillation, which more closely resembles Ventricular Ectopics due to the presence of ectopic beats. (c) An instance labeled as Ventricular Ectopics despite the absence of any visible ectopic beats. (d) A signal labeled as Atrial Fibrillation that lacks any discernible ECG waveform.

suggesting it should not be considered a normal sinus rhythm signal. In case (b), the record is annotated as Atrial Fibrillation, but the presence of three ectopic beats indicates that it is more consistent with Ventricular Ectopics or PVC. Even more interesting is case (c), which is labeled as Ventricular Ectopics despite the absence of any ectopic beats upon careful inspection. These examples highlight the prevalence of annotation errors within the datasets, which occur at a non-negligible frequency. Although such anomalies undoubtedly affect the performance and generalization ability of deep learning models trained on this data, we retained the original labels to remain consistent with the existing literature.

4.6. Limitations and future works

Our proposed network demonstrates reliable and promising results compared to state-of-the-art approaches, yet several limitations remain that highlight directions for future work. Multilabel classification accuracy is not yet optimal, and improving this will require large-scale, accurately annotated datasets, which are difficult to obtain in the medical domain due to privacy concerns and labeling complexity. Additionally, the XAI-guided framework is inherently supervised and depends on clinically meaningful CAM annotations, but generating such annotations requires expert cardiologist involvement and is challenging to scale. While automated CAM generation provided a practical solution in this study, large-scale validation will ultimately require expert-annotated CAMs to ensure reliability. It is important to note that, all XAI-guidance experiments in this study were conducted in a multi-class setting. Extending the XAI-guidance framework to multilabel classification is particularly demanding because overlapping pathologies and shared ECG regions complicate annotation and supervision. Addressing these issues will necessitate scalable annotation strategies and closer integration of expert knowledge. Future work will therefore focus on developing such pipelines and exploring partially or fully unsupervised XAI guidance mechanisms, which could broaden applicability and enable validation across diverse ECG pathologies and larger datasets.

We also aim to integrate language or textual features, such as clinical annotations or physician notes, alongside signal-based features in the future. Combining these complementary modalities has the potential to improve diagnostic accuracy and interpretability. Additionally, while our current approach focuses on signal-level feature extraction, incorporating beat-level features may provide more granular insights into ECG patterns. A hybrid approach incorporating both beat-level and signal-level representations could further strengthen the model's ability to detect complex arrhythmias. Furthermore, if the system is to be deployed on edge devices such as smartphones for privacy preservation, the automated report generation must be performed locally on-device. This introduces additional challenges and research opportunities to develop efficient, lightweight natural language processing models suitable for such platforms. Furthermore, we identified significant anomalies across all datasets. A key direction for future work is the development of methods to systematically identify and exclude these records from the evaluation. Addressing these directions will be central to advancing the robustness and clinical applicability of our framework.

5. Conclusion

In this work, we propose CARDxnosis, an innovative and interpretable deep learning framework for automated ECG analysis, addressing challenges posed by waveform complexity and limited annotated data. By combining spatial and temporal features from raw signals with domain-specific quantitative attributes extracted from both time and frequency domains, the dual-branch encoder captures a comprehensive representation of cardiac signals. A joint cross attention fusion network effectively integrates these features to improve classification performance. A novel XAI guidance strategy is also introduced, which assists the network in concentrating on clinically relevant regions of the ECG. The proposed model was evaluated in both multiclass and multilabel settings, demonstrating strong adaptability across diverse clinical scenarios. Extensive experiments on benchmark datasets show consistent outperformance over existing state-of-the-art methods. To support prediction reliability and transparency, XAI techniques such as t-SNE and GradCAM were applied. Furthermore, the integration of a large language model for automated ECG report generation enhances clinical readability. Although significant progress is still needed before automated ECG systems can be fully integrated into clinical workflows, our work advances the field by moving several important steps closer to that goal.

CRediT authorship contribution statement

Tushar Talukder Showrav: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Soyabul Islam Lincoln:** Writing – original draft, Validation, Software, Investigation, Data curation. **Md. Kamrul Hasan:** Supervision, Conceptualization, Methodology, Validation, Funding acquisition, Project administration, Writing – review & editing.

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

The authors would like to express their sincere gratitude to Teton Private Ltd. for providing partial financial support for the execution of this project, as well as to Bangladesh University of Engineering and Technology (BUET) for awarding a basic research grant. They also gratefully acknowledge Prof. Dr. S. T. M. Abu Azam, cardiologist, for his valuable assistance in reviewing and interpreting the ECG data, which significantly enhanced the quality of this study.

Appendix A. ECG quantitative features list

A.1. Wave-based ECG features

The quantitative features extracted from the ECG waveform include:

- Wave counts: P, Q, R, S, and T waves
- Wave amplitudes: P, Q, R, S, and T waves
- Intervals between key points: P–Q, Q–S, S–T, P–S, P–T, and Q–T.

A.2. Heart rate variability (HRV) features

The HRV-based quantitative features include:

- Basic time-domain measures: Beats per minute (BPM), mean of normal-to-normal intervals (Mean NN), standard deviation of successive differences (SDSD), standard deviation of NN intervals (SDNN), and root mean square of successive differences (RMSSD)
- Variation measures: Coefficient of variation of successive differences (CVSD), coefficient of variation of NN intervals (CVNN)
- Robust statistics: Median NN intervals (Median NN), median absolute deviation of NN intervals (MAD NN), median coefficient of variation of NN intervals (MCV NN)
- Additional measures: Interquartile range of NN intervals (IQR NN), standard deviation of RMSSD (SDRMSSD), 20th and 80th percentiles of NN intervals (Prc20 NN, Prc80 NN), minimum and maximum NN intervals (Min NN, Max NN), and heart rate turbulence index (HTI).

A.3. Features from IMFs of ECG leads

For each IMF of each ECG lead, the extracted quantitative features include:

- Signal measures: Leakage measure
- Spectral moments: M1, M2, M3
- Auxiliary counts: AC1, AC2, AC3
- Statistical features: Skewness, kurtosis
- Entropy measures: Shannon entropy, normalized entropy, logarithmic entropy, threshold entropy.

Data availability

Data will be made available on request.

References

[1] H. Khraishah, B. Alahmad, R.L. Ostergard Jr, A. AlAshqar, M. Albaghdadi, N. Vellanki, M.M. Chowdhury, S.G. Al-Kindi, A. Zanobetti, A. Gasparrini, et al., Climate change and cardiovascular disease: implications for global health, Nat. Rev. Cardiol. 19 (12) (2022) 798–812.
[2] V. Gliner, I. Levy, K. Tsutsui, M.R. Acha, J. Schliamser, A. Schuster, Y. Yaniv, Clinically meaningful interpretability of an AI model for ECG classification, NPJ Digit. Med. 8 (1) (2025) 109.
[3] Y.D. Daydulo, B.L. Thamineni, A.A. Dawud, Cardiac arrhythmia detection using deep learning approach and time frequency representation of ECG signals, BMC Med. Inform. Decis. Mak. 23 (1) (2023) 232.
[4] G. Wang, M. Chen, Z. Ding, J. Li, H. Yang, P. Zhang, Inter-patient ECG arrhythmia heartbeat classification based on unsupervised domain adaptation, Neurocomputing 454 (2021) 339–349.
[5] T.J. Poterucha, L. Jing, R.P. Ricart, M. Adjei-Mosi, J. Finer, D. Hartzel, C. Kelsey, A. Long, D. Rocha, J.A. Ruhl, et al., Detecting structural heart disease from electrocardiograms using AI, Nature (2025) 1–10.
[6] R. Tao, L. Wang, Y. Xiong, Y.-R. Zeng, Im-ECG: an interpretable framework for arrhythmia detection using multi-lead ECG, Expert Syst. Appl. 237 (2024) 121497.
[7] Z. Wang, C. Ma, S. Zhang, M. Zhao, Y. Liu, L. Zhao, T. Zhu, J. Li, C. Liu, Multi-teacher self-distillation based on adaptive weighting and activation pattern for enhancing lightweight arrhythmia recognition, Inf. Fusion. 122 (2025) 103178.
[8] A.K. Singh, S. Krishnan, ECG signal feature extraction trends in methods and applications, Biomed. Eng. Online 22 (1) (2023) 22.
[9] F.E. Alsaadi, N.S. Alharbi, Dynamic-aware self-attention with embedded CWT for imbalanced ECG classification in cardiac disease detection, Neurocomputing (2025) 131945.
[10] A. Mert, ECG feature extraction based on the bandwidth properties of variational mode decomposition, Physiol. Meas. 37 (4) (2016) 530.
[11] K. Dragomiretskiy, D. Zosso, Variational mode decomposition, IEEE Trans. Signal Process. 62 (3) (2013) 531–544.
[12] X. Wei, Z. Li, Y. Tian, M. Wang, J. Liu, Y. Jin, W. Ding, C. Liu, Differentiated knowledge distillation: patient-specific single-sample personalization for electrocardiogram diagnostic models, Eng. Appl. Artif. Intell. 136 (2024) 108880.
[13] W. Huang, N. Wang, P. Feng, H. Wang, Z. Wang, B. Zhou, A multi-resolution mutual learning network for multi-label ECG classification, in: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2024, pp. 3303–3306.
[14] C.-L. Liu, B. Xiao, C.-F. Tsai, ECG-star: spatio-temporal attention residual networks for multi-label ECG abnormality classification, Inf. Sci. (2025) 122273.
[15] Y. Cheng, D. Li, D. Wang, Y. Chen, L. Wang, Multi-label arrhythmia classification using 12-lead ECG based on lead feature guide network, Eng. Appl. Artif. Intell. 129 (2024) 107599.
[16] G. Msigwa, E. Ntambala, J. Yun, Enhancing atrial fibrillation classification from single-lead electrocardiogram signals using attention-based networks and generative adversarial networks with density-based clustering, Eng. Appl. Artif. Intell. 133 (2024) 108607.
[17] A.S.B. Mahel, M.S. Al-Gaashani, R.I. Alkanhel, D.S.M. Hassan, M.S.A. Muthanna, A. Muthanna, A. Aziz, A multi-scale deep learning framework combining mobilevit-eca and LSTM for accurate ECG analysis, IEEE Access 13 (2025).
[18] C.-L. Liu, B. Xiao, C.-H. Hsieh, Multimodal fusion of spatial-temporal and frequency representations for enhanced ECG classification, Inf. Fusion. (2025) 102999.
[19] L. Wang, X. Wang, R. Tao, Wave masked autoencoder: an electrocardiogram signal diagnosis model based on wave making strategy, Inf. Sci. 690 (2025) 121516.

[20] X. Liu, H. Wang, Z. Li, L. Qin, Deep learning in ECG diagnosis: a review, Knowl. Based Syst. 227 (2021) 107187.
[21] X. Xie, H. Liu, D. Chen, M. Shu, Y. Wang, Multilabel 12-lead ECG classification based on leadwise grouping multibranch network, IEEE Trans. Instrum. Meas. 71 (2022) 1–11.
[22] S. Yang, C. Lian, Z. Zeng, B. Xu, J. Zang, Z. Zhang, A multi-view multi-scale neural network for multi-label ECG classification, IEEE Trans. Emerg. Top. Comput. Intell. 7 (3) (2023) 648–660.
[23] W. Liu, S. Pan, Z. Li, S. Chang, Q. Huang, N. Jiang, Lead-fusion barlow twins: a fused self-supervised learning method for multi-lead electrocardiograms, Inf. Fusion 114 (2025) 102698.
[24] H. Li, J. Han, H. Zhang, X. Zhang, Y. Si, Y. Zhang, Y. Liu, H. Yang, Clinical knowledge-based ECG abnormalities detection using dual-view cnn-transformer and external attention mechanism, Comput. Biol. Med. 178 (2024) 108751.
[25] R. Jiang, B. Fu, R. Li, R. Li, D.Z. Chen, Y. Liu, G. Xie, K. Li, A dual-branch convolutional neural network with domain-informed attention for arrhythmia classification of 12-lead electrocardiograms, Eng. Appl. Artif. Intell. 139 (2025) 109480.
[26] S. Vijayarangan, B. Murugesan, R. Vignesh, S.P. Preejith, J. Joseph, M. Sivaprakasam, Interpreting deep neural networks for single-lead ECG arrhythmia classification, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2020, pp. 300–303.
[27] S. Wu, J. Zhou, Y. Dong, F. Chen, Enhancing explainability of deep learning-based ECG diagnosis using large language models, in: Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence, 2024, pp. 61–65.
[28] X. Liu, Y. Xu, H. Xu, L. He, S. Long, Y. Huang, Y. Wang, Y. Lu, Y. Huang, J. Wu, et al., Advancing interpretable cardiac disease diagnosis via a transformer-convolutional hybrid network on electrocardiograms, Eng. Appl. Artif. Intell. 152 (2025) 110675.
[29] L.S. Johnson, P. Zadrozniak, G. Jasina, A. Grotek-Cuprjak, J.G. Andrade, E. Svennberg, S.Z. Diederichsen, W.F. McIntyre, S. Stavrakis, J. Benezet-Mazuecos, et al., Artificial intelligence for direct-to-physician reporting of ambulatory electrocardiography, Nat. Med. 31 (3) (2025) 925–931.
[30] A. Bleich, A. Linnemann, B.H. Diem, T.O.F. Conrad, Automated medical report generation for ecg data: Bridging medical text and signal processing with deep learning, arXiv preprint arXiv:2412.04067, 2024.
[31] F. Ashfaq, N.Z. Jhanjhi, N.A. Khan, D. Javaid, M. Masud, M. Shorfuzzaman, Enhancing ECG report generation with domain-specific tokenization for improved medical NLP accuracy, IEEE Access 13 (2025).
[32] F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu, Y. Liu, C. Ma, S. Wei, Z. He, et al., An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection, J. Med. Imaging Health Inform. 8 (7) (2018) 1368–1373.
[33] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F.I. Lunze, W. Samek, T. Schaeffter, PTB-XL, a large publicly available electrocardiography dataset, Sci. Data 7 (1) (2020) 1–15.
[34] J. Zheng, H. Guo, H. Chu, A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0. 0), PhysioNet (2022).
[35] M.A. Reyna, E.A.P. Alday, A. Gu, C. Liu, S. Seyedi, A.B. Rad, A. Elola, Q. Li, A. Sharma, G.D. Clifford, Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020, in: 2020 Computing in Cardiology, IEEE, 2020, pp. 1–4.
[36] A.B.M. Abdullah, ECG in Medical Practice, Jaypee Brothers Medical Publishers, 2021.
[37] G.S. Wagner, D.G. Strauss, Marriott's Practical Electrocardiography, Lippincott Williams & Wilkins, 2014.
[38] N. Ibtehaz, M.S. Rahman, Multiresunet: rethinking the U-Net architecture for multimodal biomedical image segmentation, Neural Netw. 121 (2020) 74–87.
[39] T.T. Showrav, M.K. Hasan, Hi-gmisnet: generalized medical image segmentation using DWT based multilayer fusion and dual mode attention into high resolution PGAN, Phys. Med. Biol. 69 (11) (2024) 115019.
[40] T.T. Showrav, S.I. Lincoln, M.K. Hasan, An explainable artificial intelligence framework for trustworthy arrhythmia detection from heart rhythm data, Healthc. Anal. 2026, 100455.
[41] D. Makowski, T. Pham, Z.J. Lau, J.C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, S.H.A. Chen, NeuroKit2: A python toolbox for neurophysiological signal processing, Behav. Res. Methods 53 (4) (2021) 1689–1696, https://doi.org/10.3758/s13428-020-01516-y
[42] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[43] A. Ameen, I.E. Fattoh, T. Abd El-Hafeez, K. Ahmed, Advances in ECG and pcg-based cardiovascular disease classification: a review of deep learning and machine learning methods, J. Big Data 11 (1) (2024) 159.
[44] A. Raghu, D. Shanmugam, E. Pomerantsev, J. Guttag, C.M. Stultz, Data augmentation for electrocardiograms, in: Conference on Health, Inference, and Learning, PMLR, 2022, pp. 282–310.
[45] R. Wang, J. Fan, Y. Li, Deep multi-scale fusion neural network for multi-class arrhythmia detection, IEEE J. Biomed. Health Inform. 24 (9) (2020) 2461–2472, https://doi.org/10.1109/JBHI.2020.2981526
[46] S. Yang, C. Lian, Z. Zeng, B. Xu, J. Zang, Z. Zhang, A multi-view multi-scale neural network for multi-label ECG classification, IEEE Trans. Emerg. Top. Comput. Intell. 7 (3) (2023) 648–660, https://doi.org/10.1109/TETCI.2023.3235374

Author biography



Tushar Talukder Showrav is an electrical engineer and researcher focused on advancing healthcare through technology. He earned his B.Sc. in Electrical and Electronic Engineering from the Bangladesh University of Engineering and Technology (BUET). Currently, he is a full-time Research Assistant at the DSP Research Lab at BUET, where he works on AI-driven solutions for medical challenges. His research spans artificial intelligence in healthcare, medical image analysis, and signal processing. He is passionate about designing data-driven tools and intelligent systems that support clinical decision-making, improve diagnostic accuracy, and contribute to the development of innovative medical technologies.



Soyabul Islam Lincoln received his B.Sc. degree in Electronics and Communication Engineering from Khulna University of Engineering and Technology (KUET), Bangladesh, in 2023. His academic interests include machine learning and deep learning, with emphasis on medical imaging, signal processing, natural language processing, and generative AI. He is currently engaged in research on data-driven intelligent systems, with a strong interest in developing robust, interpretable, and real-world applicable artificial intelligence models.



Md. Kamrul Hasan received the B.Sc. and M.Sc. degrees in Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh, in 1989 and 1991, respectively. He earned the M.Eng. and Ph.D. degrees in Information and Computer Sciences from Chiba University, Japan, in 1995 and 1997. In 1989, he joined BUET as a Lecturer in the Department of Electrical and Electronic Engineering, where he is currently a Professor. He has held several international research positions, including Postdoctoral Fellow at Chiba University, Japan; Research Associate at Imperial College London, UK; Invited Research Fellow at the University of Tokyo, Japan; and Professor of International Scholars at Kyung Hee University, Korea.
Dr. Hasan has authored or coauthored more than 150 scientific publications. His research interests include digital signal processing, artificial intelligence for digital health applications, quantitative ultrasound, elastography, medical imaging, and intelligent battery management systems. He currently serves as an Associate Editor for IEEE Access.