Medical Engineering and Physics 145 (2025) 104418

Contents lists available at ScienceDirect

Medical Engineering and Physics

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

Paper

Graph-enhanced deep learning for ECG arrhythmia detection: An integration of CNN-GNN-BiLSTM approach

Piyush Mahajan *, Amit Kaul [1]

Electrical Engineering Department, NIT Hamirpur, 177005, H.P, India

ARTICLE INFO

Keywords:
ECG
Arrhythmia
CNN
GNN
LSTM

ABSTRACT

Early and accurate detection of cardiac arrhythmias is crucial for preventing severe cardiovascular events. This study proposes a CNN–GNN–BiLSTM integrated framework for automated ECG arrhythmia classification, combining spatial, relational, and temporal learning to achieve enhanced predictive accuracy. Convolutional Neural Networks (CNNs) serve as feature extractors from ECG spectrograms, while Graph Attention Networks (GATs) capture inter-beat relationships through graph-based modeling. In parallel, Bidirectional Long Short-Term Memory (BiLSTM) networks refine temporal dependencies, ensuring robust sequential representation. Outputs from GAT and BiLSTM modules are concatenated to form a unified feature representation, which is passed through a fully connected classifier for final prediction. The model is evaluated on three benchmark ECG datasets—MIT-BIH, PTB, and Chapman-Shaoxing—as well as a combined 11-class dataset, demonstrating superior generalization. Results indicate significant performance improvement over conventional deep learning approaches, achieving 96.0% overall accuracy and up to 99.89% accuracy on MIT-BIH. The proposed framework effectively mitigates misclassification errors and offers a scalable, real-time solution for AI-driven cardiac monitoring systems.

1. Introduction

Cardiovascular diseases (CVDs) remain a leading cause of morbidity and mortality worldwide, underscoring the need for early and accurate diagnosis. Electrocardiograms (ECGs) are a fundamental diagnostic tool, offering critical insights into cardiac conditions such as arrhythmias, myocardial infarctions, and conduction abnormalities. Traditional ECG classification methods rely on handcrafted features and rule-based algorithms, which often fail to generalize across diverse patient populations.

Deep learning has transformed ECG classification, with Convolutional Neural Networks (CNNs) achieving state-of-the-art results in detecting arrhythmias by extracting spatial patterns from ECG waveforms and spectrograms [1,2]. However, CNNs primarily capture local spatial features and lack the ability to model temporal dependencies and inter-beat relationships, which are essential for accurate classification.

To address these limitations, Graph Neural Networks (GNNs) have been introduced to model ECG data as structured graphs, where nodes represent heartbeat features and edges define inter-beat relationships. This graph-based representation enables the network to capture global cardiac patterns, leading to improved classification performance. The most widely adopted GNN variant, Graph Convolutional Networks (GCNs), applies uniform weights to all node connections, which may lead to suboptimal feature aggregation [3]. To overcome this, Graph Attention Networks (GATs) introduce an attention mechanism that dynamically assigns importance to different connections, prioritizing the most relevant features for classification [4,5].

ECG signals also exhibit strong temporal dependencies, which many deep learning models overlook. Long Short-Term Memory (LSTM) networks effectively model both short and long-term dependencies in sequential data, making them well-suited for ECG analysis [6–8]. However, despite their potential, LSTMs have not been extensively combined with GNN-based ECG classifiers, leaving a gap in fully harnessing temporal dynamics for cardiac anomaly detection.

This study proposes a deep learning framework integrating CNNs, GNNs, and LSTMs to leverage their complementary strengths for enhanced ECG classification. The model employs ConvNeXt Tiny, a state-of-the-art CNN, to extract high-level spatial features from ECG spectrograms. These features are then transformed into a graph structure,

* Corresponding author.
E-mail address: piyush.mahajan@nith.ac.in (P. Mahajan).
[1] Contributing authors.

https://doi.org/10.1016/j.medengphy.2025.104418
Received 28 March 2025; Received in revised form 28 July 2025; Accepted 14 August 2025

Available online 22 August 2025

where a GAT captures inter-beat relationships by dynamically adjusting connection weights. To incorporate temporal dependencies, bidirectional LSTMs process the graph-based embeddings, ensuring sequential learning. Unlike traditional ensemble methods, the proposed framework employs a trainable fusion mechanism, which optimally balances contributions from CNN and GNN components during training, leading to improved classification accuracy.

By integrating spatial feature extraction, relational modeling, and temporal learning, this approach enhances ECG classification performance while addressing individual model limitations. Recent studies highlight the potential of integrated architectures in ECG analysis [9,10]. The proposed framework is evaluated on multiple benchmark ECG datasets, demonstrating superior classification performance compared to existing methods.

The main contributions of this study are as follows: (i) A unified architecture is developed that combines ConvNeXt, GAT, and BiLSTM to simultaneously capture spatial, relational, and temporal dependencies in ECG signals. (ii) A feature-level fusion strategy is implemented by concatenating graph-based and temporal representations into a joint embedding, followed by a fully connected classifier for final prediction. (iii) An extensive ablation study is conducted to quantify the contribution of each component across multiple stages. (iv) The model's effectiveness is validated on three benchmark ECG datasets and a combined dataset, achieving competitive results across multiple performance metrics. The following section provides a detailed review of related work, analyzing existing methodologies and their contributions to automated ECG analysis.

2. Related work

The integration of deep learning techniques in ECG classification has significantly improved the detection and analysis of cardiovascular anomalies. CNN-based architectures have been widely used due to their ability to capture spatial patterns in ECG signals, leading to state-of-the-art performance in arrhythmia classification. Ansari et al. conducted a comprehensive survey categorizing various deep learning architectures, highlighting CNNs as a dominant approach for ECG-based diagnostics [1]. In addition to standard CNNs, Abdulrahman et al. suggested an improved 1D CNN-based heartbeat classification model that demonstrated higher accuracy in discriminating normal and pathological ECG signals. The model achieved 99.8% accuracy on the MIT-BIH dataset, demonstrating the usefulness of CNNs in ECG classification [11]. Recent work by Singh et al. explored the use of multi-lead ECG signals with 1D CNN-based models, showing significant improvements in classification accuracy. Their approach emphasizes the importance of utilizing multi-channel information to enhance deep learning performance for ECG-based diagnostics [12]. Huang and Wu proposed a computer-aided diagnosis system that employs a two-dimensional CNN for classifying atrial fibrillation and normal sinus rhythm using ECG images. Their study demonstrated that 2D-CNNs can effectively capture spatial patterns directly from ECG images, achieving high accuracy of 99.23% on filtered signals from the MIT-BIH atrial fibrillation dataset, further establishing CNNs as reliable tools for ECG classification [13]. Eleyan et al. extended this idea by developing a spectrogram-based CNN–GRU model that achieved 99.76% on a five-class subset of MIT-BIH [14], while another variation using FFT and CNN–LSTM fusion yielded 97.60% classification accuracy [15]. Their earlier work also proposed a CNN–LSTM hybrid framework for arrhythmia detection, demonstrating robust performance in ECG sequence modeling [16].

Beyond CNNs, Graph Neural Networks (GNNs) have emerged as a powerful tool for modeling complex inter-beat relationships in ECG signals. Unlike conventional models that treat ECG beats as independent instances, GNNs represent ECG signals as structured graphs, enabling relational learning. Zeinalipour and Gori applied GNNs for heartbeat classification, achieving high accuracy by leveraging topological relationships [17]. Similarly, Qiang et al. introduced the Convolutional Residual Graph Neural Network (Conv-RGNN) to capture spatial relationships among ECG leads, demonstrating significant improvements in multi-label classification tasks [18].

To enhance the interpretability and performance of ECG classification models, attention mechanisms have been integrated with GNNs. Wang et al. proposed a Graph Attention Network (GAT) that applies a multi-head self-attention mechanism to dynamically prioritize critical inter-beat relationships, improving classification accuracy over traditional graph-based methods [4].

Temporal modeling is another key aspect of ECG analysis, as cardiac rhythms exhibit sequential dependencies. LSTM-based models have been extensively used to capture both short-term fluctuations and long-term dependencies in ECG signals. Li et al. demonstrated that LSTM networks significantly enhance arrhythmia classification performance by preserving sequential patterns [3]. However, despite their effectiveness, LSTMs have not been widely integrated with GNNs, leaving a gap in fully leveraging temporal dependencies for ECG classification.

Hybrid models that combine multiple deep learning architectures have been explored to enhance classification accuracy. Some studies have integrated CNNs with Recurrent Neural Networks (RNNs) to extract both spatial and temporal features [19], while others have coupled GNNs with attention mechanisms to improve feature extraction and relational learning [4]. However, limited research has focused on a unified framework that combines CNNs, GNNs, and temporal modeling, which could offer a more comprehensive approach to ECG classification. Dwivedi et al. introduced eFuseNet, an ensemble model that combines modified AlexNet with LSTM and Gated Recurrent Units (GRUs) to improve the detection of arrhythmias and myocardial infarctions. Their results on MIT-BIH and PTB datasets demonstrated superior classification accuracy, highlighting the benefits of feature fusion techniques [20]. Deevi et al. introduced HeartNetEC, a deep representation learning framework designed for beat-wise ECG classification. Their approach utilizes a deep denoising network followed by a beat classification network, leveraging both spatial and temporal features from the MIT-BIH Arrhythmia Database. By combining denoising, feature extraction, and deep learning-based classification, HeartNetEC achieved competitive performance across multiple arrhythmic classes [21].

Despite advancements in deep learning-based ECG classification, challenges remain. Model interpretability is a critical concern for clinical adoption, prompting efforts toward developing explainable AI models that provide insights into neural network decision-making [7]. Additionally, integrating multimodal physiological data such as ECG combined with other biosignals presents opportunities for more comprehensive cardiac assessments. Future research should also focus on developing lightweight models suitable for real-time deployment in wearable devices and point-of-care diagnostics, as highlighted in recent works such as MobileECG-Net by Chen et al. [22] and contrastive representation learning methods like ECG-CL, which enhance generalization across subjects [23]. Building upon these advancements, the next section details the proposed methodology, outlining the datasets, preprocessing techniques, and the design of an integrated framework of CNN-GNN-BiLSTM for enhanced ECG classification.

3. Methodology

This section outlines the methodology employed for ECG classification, detailing the datasets used, preprocessing techniques, and the design of the deep learning framework. The approach follows a structured pipeline, as illustrated in Fig. 1, which comprises ECG signal preprocessing, feature extraction, graph-based modeling, and sequential learning. The proposed approach integrates CNNs for spatial feature extraction, GNNs for relational learning, and BiLSTMs for capturing temporal dependencies. Finally, a fusion mechanism optimally combines these extracted features to improve arrhythmia classification accuracy. To ensure the robustness of the proposed model, three widely used ECG

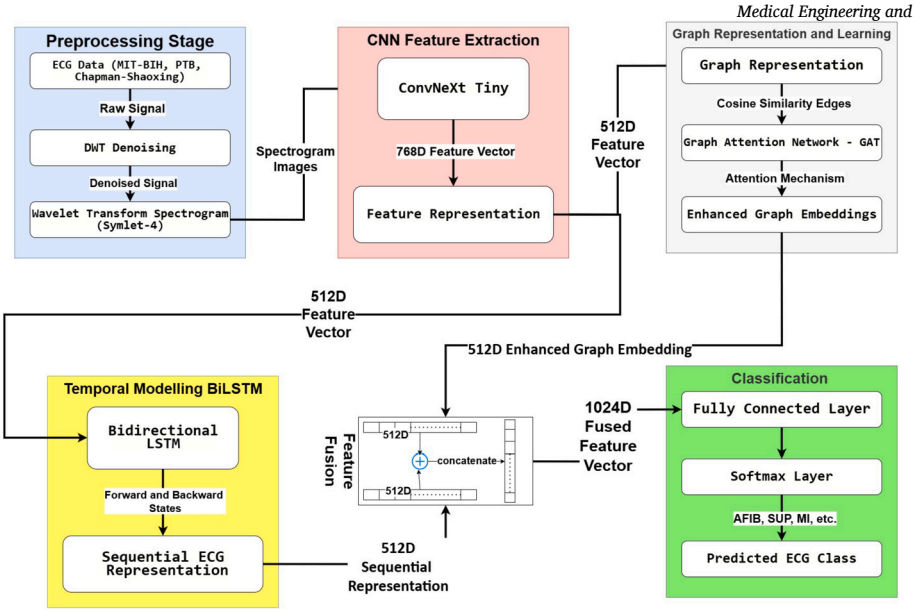


Fig. 1. Block Diagram of Graph Enhanced CNN-GNN-BiLSTM Integrated Approach.

datasets were selected, each providing a diverse set of arrhythmic conditions for training and evaluation.

3.1. Datasets used

The proposed model is evaluated on three benchmark ECG datasets: MIT-BIH Arrhythmia Database (MIT-BIH) [24], PTB Diagnostic ECG Database (PTB) [25], and Chapman-Shaoxing 12-lead ECG Database (CS-12) [26]. These datasets encompass a broad spectrum of cardiac abnormalities, ensuring a diverse and representative sample for training and evaluation.

MIT-BIH consists of 48 half-hour ECG recordings from 47 patients, sampled at 360 Hz across two leads (MLII and V1). For this study, six major classes were selected: Atrial Fibrillation (AFIB), Congestive Heart Failure (CHF), Supraventricular Arrhythmia (SUP), Ventricular Ectopy (VE), Normal Sinus Rhythm (NORMAL), and Couplets of Ventricular Tachycardia (CUVT).

PTB contains 549 records from 290 subjects, including both healthy individuals and patients with cardiac abnormalities. This dataset provides high-resolution (1000 Hz) ECG signals, making it ideal for detailed waveform analysis. The six classes chosen from this dataset include AFIB, Conduction Disturbance (CD), Healthy Control (Healthy), Hypertrophy (HYP), Myocardial Infarction (MI), and SUP.

The CS-12 ECG Database consists of 10,646 ECG recordings, each lasting 10 seconds, collected from Shaoxing People's Hospital. The signals were recorded at 500 Hz using a 12-lead ECG system, with expert annotations identifying multiple rhythm classes. Four primary classes were used in this study: Atrial Fibrillation (AFIB), Supraventricular Tachycardia (GSVT), Sinus Bradycardia (SB), and Sinus Rhythm (SR). This dataset contributes to the generalizability of the model by incorporating a wider range of clinical conditions, particularly sinus-related abnormalities.

To create a balanced and standardized dataset, overlapping arrhythmia types across different datasets were mapped into common categories. For example, AFIB from MIT-BIH, PTB, and Chapman-Shaoxing datasets was unified under a single Atrial Fibrillation category, and SUP from MIT-BIH and PTB was merged into a single Supraventricular Arrhythmia category. This process facilitated improved learning efficiency by reducing redundancy and ensuring sufficient training samples per class. Fig. 2 presents the time-domain waveforms along with their corresponding spectrogram representations for all the selected ECG classes, providing a visual depiction of the signal characteristics used in the classification process. Before feature extraction, raw ECG signals undergo multiple preprocessing steps to enhance quality, reduce noise, and transform them into spectrogram representations suitable for deep learning models. Each dataset was partitioned into training and testing subsets with approximately 85–90% of samples allocated for training and the remaining 10–15% for testing. The final training set comprised 1,284 samples for MIT-BIH, 3,273 samples for PTB, and 2,453 samples for CS-12. Corresponding test sets included 222 samples for MIT-BIH, 384 samples for PTB, and 494 samples for CS-12. For the combined dataset, 7,507 samples were used for training and 987 for testing, ensuring balanced representation across all unified arrhythmia categories.

3.2. Pre-processing and image generation

Raw ECG signals undergo multiple preprocessing steps to enhance signal quality, reduce noise artifacts, and generate input features suitable for deep learning models. The preprocessing pipeline consists of three main stages: segmentation, noise filtering, and spectrogram image generation.

ECG signals vary in length across different datasets; hence, each recording was segmented into uniform 10-second waveforms. This standardization ensures consistent input length for the model while retaining sufficient temporal information for arrhythmia classification.

ECG recordings are susceptible to multiple noise sources, including baseline wander, powerline interference, and motion artifacts. To mitigate these distortions, a Discrete Wavelet Transform (DWT)-based filtering approach was used, employing the Symlet-4 (sym4) wavelet function. Wavelet decomposition separates the signal into multiple frequency sub-bands, allowing for selective noise suppression while preserving diagnostic waveform characteristics. Mathematically, the wavelet decomposition process is expressed as:

$$s(t) = \sum_{j,k} c_{j,k} \psi_{j,k}(t) \tag{1}$$

where $s(t)$ represents the ECG signal, $\psi_{j,k}(t)$ denotes the wavelet function at scale j and position k, and $c_{j,k}$ are the associated wavelet coefficients. The sym4 wavelet was selected due to its optimal trade-off between time and frequency resolution, making it well-suited for ECG denoising applications.

Following noise reduction, each 10-second ECG segment was transformed into a spectrogram image using Continuous Wavelet Transform

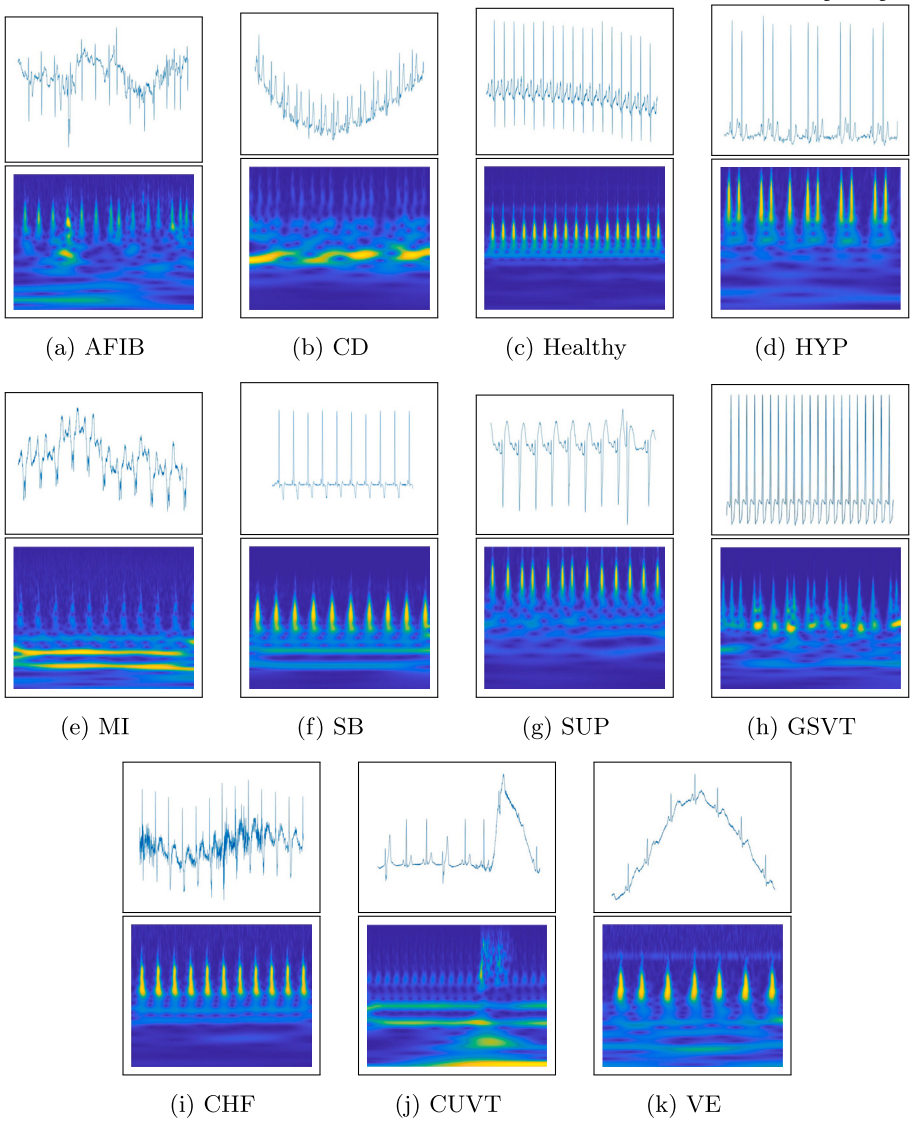


Fig. 2. All classes with their time domain waveforms and corresponding wavelet transformed images.

(CWT) with the sym4 wavelet. This transformation captures both time-domain and frequency-domain information, improving the model's ability to recognize arrhythmic patterns. The CWT formulation is given by:

$$X(a,b) = \int_{-\infty}^{\infty} s(t) \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) dt \tag{2}$$

where $X(a,b)$ represents the wavelet coefficients at scale a and translation b, and ψ is the chosen wavelet function. The resulting spectrogram images were resized to 224×224 pixels to align with the input requirements of ConvNeXt Tiny.

3.3. Proposed deep learning framework

The proposed deep learning model integrates three complementary approaches to ECG classification: spatial feature extraction using CNNs, relational learning using GNNs, and sequential modeling using BiLSTMs. This suggested framework ensures robust pattern recognition by leveraging both local and global dependencies in ECG data.

3.3.1. Feature extraction using ConvNeXt tiny

In recent years, CNNs have demonstrated remarkable success in feature extraction from complex data, including medical images and spectrogram representations of biomedical signals. Traditional CNN architectures, such as ResNet, have been widely used for image-based classification tasks, but advancements in neural network design have introduced more efficient and scalable architectures. One such model is ConvNeXt, a modernized CNN that integrates successful design elements from vision transformers while maintaining computational efficiency and structured learning [27].

Several CNN architectures have been explored for ECG spectrogram-based classification, including ResNet-18, MobileNetV2, and ShuffleNet. While lightweight models like MobileNetV2 and ShuffleNet offer computational efficiency, they may sacrifice feature extraction capacity, particularly in complex medical signals. Traditional architectures like ResNet-18, although effective, rely on batch normalization, which can introduce instability in small medical datasets. ConvNeXt, inspired by transformer-based design principles, enhances feature extraction through depthwise separable convolutions and layer normalization, ensuring both efficiency and robustness. Given its superior trade-off between computational cost and classification accuracy, ConvNeXt Tiny was selected as the CNN backbone for this study.

ConvNeXt introduces several architectural modifications that enhance its ability to extract hierarchical feature representations. Unlike conventional CNNs that rely on standard convolutional layers, ConvNeXt incorporates depthwise separable convolutions, which decom-

pose regular convolution operations into two stages: depthwise convolution, which applies a single filter per input channel, and pointwise convolution, which projects the results into a higher-dimensional space. This decomposition significantly reduces computational complexity while maintaining representational power. Additionally, residual connections improve gradient flow during training, allowing deeper networks to learn effectively without suffering from vanishing gradients. Another key enhancement in ConvNeXt is the replacement of traditional batch normalization with layer normalization applied across the channel dimension. This modification stabilizes training and improves generalization, making the model more robust to variations in input data. Furthermore, the use of the GELU activation function, instead of standard ReLU, introduces smooth non-linearity, aiding in better feature extraction from biomedical signals such as ECG spectrograms.

For ECG classification, spectrogram images derived from raw ECG signals are used as input to ConvNeXt Tiny, the most lightweight variant of the ConvNeXt family. These images are resized to 224×224 pixels, matching the input dimensions required by ConvNeXt. The model processes the input through multiple convolutional layers, progressively extracting hierarchical features that capture both spatial and frequency-domain information. The final convolutional layer outputs a 768-dimensional feature vector, which serves as a compact and informative representation of the ECG signal.

3.3.2. Graph representation and learning using GAT

To model the inter-beat relationships within ECG signals, CNN-extracted features are transformed into a graph representation. Each ECG sample is treated as a node, while edges are constructed based on cosine similarity between feature vectors:

$$S(i,j) = \frac{F_i \cdot F_j}{\|F_i\| \|F_j\|} \tag{3}$$

where S(i,j) represents the similarity between features F_i and F_j. The resulting graph is processed using a Graph Attention Network (GAT), which assigns dynamic weights to node connections based on their importance. The attention mechanism is defined as:

$$a_{ij} = \frac{exp(LeakyReLU(a^T[WF_i \parallel WF_j]))}{\sum_{k \in N(i)} exp(LeakyReLU(a^T[WF_i \parallel WF_j]} \tag{4}$$

where a_{ij} represents the attention coefficient between nodes i and j, and W is a learnable weight matrix. These coefficients determine the relative importance of neighboring nodes when aggregating information. The final representation, or embedding, of a node i is computed by aggregating the transformed features of its neighbors, weighted by the attention coefficients:

$$G = \sigma(\sum_{j \in N(i)} a_{ij} W F_j) \tag{5}$$

Here σ denotes an activation function, such as ReLU, applied to introduce non-linearity into the model. This process enables the network to focus on the most relevant parts of the graph, effectively capturing complex relationships within the data. The development of GATs has significantly advanced the field of graph-based neural networks by incorporating attention mechanisms that allow for the dynamic weighting of node connections. This approach addresses limitations in previous models that treated all neighboring nodes with equal importance, thereby enhancing the model's capacity to learn from graph-structured data. The original formulation and applications of GATs are detailed in the work by Veličković et al. [28].

3.3.3. Temporal learning using bidirectional LSTM

To effectively capture the temporal dynamics inherent in ECG signals, it is essential to consider both past and future contexts. BiLSTMs are designed for this purpose, processing data in both forward and backward directions to provide a comprehensive understanding of sequential dependencies.

In a BiLSTM network, each input sequence is processed by two separate LSTM layers: one handling the sequence from start to end (forward pass) and the other from end to start (backward pass). This dual approach ensures that each time step's output is informed by both preceding and succeeding elements, offering a more nuanced representation compared to unidirectional LSTMs. Mathematically, the forward and backward hidden states are:

$$\begin{aligned} h_t^f &= \sigma(W_f \cdot h_{t-1}^f + U_f \cdot G + b_f) \\ h_t^b &= \sigma(W_b \cdot h_{t+1}^b + U_b \cdot G + b_b) \end{aligned} \tag{6}$$

where W_f, W_b and U_f, U_b are weight matrices, and b_f, b_b are biases. The final hidden state is a concatenation of h_t^f and h_t^b.

3.3.4. Fusion mechanism and final classification

The proposed framework employs a feature-level fusion strategy rather than prediction-level averaging. ConvNeXt acts as a spatial feature extractor, generating a 768-dimensional representation for each ECG spectrogram. This high-dimensional feature is first reduced to 512 dimensions and then simultaneously processed by two complementary modules: the GAT branch, which captures relational dependencies, and the BiLSTM branch, which models temporal patterns. The outputs of these two branches, denoted as $\mathbf{F}_{\text{GAT}}$ and $\mathbf{F}_{\text{LSTM}}$, are concatenated to form a unified representation:

$$\mathbf{F}_{\text{fusion}} = [\mathbf{F}_{\text{GAT}} \parallel \mathbf{F}_{\text{LSTM}}] \tag{7}$$

where $\parallel$ denotes the concatenation operation. In the proposed model, both $\mathbf{F}_{\text{GAT}}$ and $\mathbf{F}_{\text{LSTM}}$ are 512-dimensional vectors, resulting in $\mathbf{F}_{\text{fusion}} \in \mathbb{R}^{1024}$. This fused feature vector is then passed through a fully connected layer for classification:

$$\mathbf{z} = \mathbf{W}\mathbf{F}_{\text{fusion}} + \mathbf{b} \tag{8}$$

where $\mathbf{W}$ and $\mathbf{b}$ denote the learnable parameters of the final linear layer.

The class probabilities are computed using the softmax function:

$$P(y = i \mid \mathbf{x}) = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)} \tag{9}$$

where C is the total number of classes and z_i is the i-th element of the output logits $\mathbf{z}$. Unlike earlier approaches that combine predictions using static or weighted averaging, this design integrates complementary feature representations at an intermediate stage, enabling joint learning of spatial, relational, and temporal characteristics in a single classification head.

3.3.5. Rationale behind model integration

The integration of CNNs, GNNs, and BiLSTMs in the proposed framework is motivated by the complementary nature of the feature representations they capture, which collectively enhance arrhythmia classification from ECG signals. ECG spectrograms contain spatially distributed morphological patterns corresponding to characteristic waveforms like P-waves, QRS complexes, and T-waves. These are effectively extracted by convolutional neural networks, with ConvNeXt offering a modernized architecture that balances depth, feature diversity, and efficiency. However, ECG signals are not merely isolated spatial entities; they exhibit strong inter-beat dependencies, particularly in pathological rhythms that manifest across sequences of cardiac cycles. To model such dependencies, the output features of the CNN are structured into graphs based on pairwise similarity, enabling graph attention networks (GATs) to prioritize contextually important relationships between beats or patient-specific patterns. While GNNs offer relational modeling, they do not inherently account for the sequential dynamics of cardiac rhythms, which are crucial for recognizing arrhythmias like atrial fibrillation, bigeminy, or sinus irregularities. Bidirectional LSTM networks

fulfill this gap by modeling both forward and backward temporal dependencies within the signal, capturing long-range context that pure CNNs or GNNs might overlook. The proposed architecture combines the outputs of the GAT and BiLSTM modules through feature-level concatenation rather than prediction-level weighting. This unified representation is then processed by a fully connected classifier, allowing joint optimization of spatial, relational, and temporal features. Such integration not only achieves superior empirical performance but also aligns with the inherent multi-dimensional structure of ECG data, where spatial morphology, temporal continuity, and inter-beat context each play essential diagnostic roles.

3.4. Ablation study: stage-wise architectural evaluation

To assess the contribution of each architectural component in the proposed CNN-GNN-BiLSTM framework, a comprehensive ablation study was conducted in four progressive stages. This strategy enabled isolation and evaluation of the individual impact of spatial, relational, and temporal learning components on arrhythmia classification.

1. **Stage 1: ConvNeXt Only.** The baseline configuration consists solely of the ConvNeXt Tiny model. It extracts spatial features from ECG spectrograms and performs classification using a fully connected head. This stage establishes the performance of pure convolutional spatial modeling.
2. **Stage 2: ConvNeXt + GAT.** In this stage, graph attention layers are introduced following ConvNeXt. The high-dimensional features extracted by ConvNeXt are used to construct a similarity graph across samples, enabling the GAT module to capture inter-beat relational dependencies. This configuration quantifies the contribution of graph-based relational learning.
3. **Stage 3: ConvNeXt + BiLSTM.** This variant skips graph modeling and instead connects ConvNeXt feature outputs directly to a bidirectional LSTM network. The aim is to capture temporal evolution of ECG patterns using sequential learning while ignoring relational structure.
4. **Stage 4: Full Integration (ConvNeXt + GAT + BiLSTM).** The final configuration integrates all components: ConvNeXt for extracting spatial features, GAT for capturing relational dependencies, and BiLSTM for modeling temporal dynamics. Outputs from the GAT and BiLSTM branches are concatenated to form a unified feature vector, which is then passed through a fully connected classifier for final prediction. This setup represents the complete proposed framework.

This staged evaluation allows a granular understanding of how each module contributes to overall classification performance and generalizability.

3.5. Hyperparameter configuration and model complexity

To ensure consistency and fair comparison across all experimental stages, a unified set of training hyperparameters was employed. These include the optimizer type, learning rate, batch size, and number of epochs, which collectively influence convergence stability and generalization performance. The AdamW optimizer was chosen for its effective weight decay handling, while a learning rate of 1×10^{-4} ensured gradual convergence without overshooting. A batch size of 32 balanced memory efficiency and gradient stability, and ten epochs were sufficient for achieving convergence given the dataset size. Table 1 summarizes these parameters along with architecture-specific configurations such as the number of attention heads in the GAT layers and hidden units in BiLSTM, which control the model's capacity to capture relational and temporal dependencies, respectively. Additionally, the table reports the total trainable parameters and model size in megabytes (MB), which provide insight into memory requirements and storage feasibility. The inclusion of average inference time per sample reflects the computational efficiency of each stage, an important consideration for real-time clinical applications.

The progression from Stage 1 (ConvNeXt only) to Stage 4 (full integration of ConvNeXt, GAT, and BiLSTM) demonstrates an incremental increase in complexity and memory footprint, correlating with added modeling capabilities for spatial, relational, and temporal features. While GAT integration (Stage 2) significantly increases inference time due to graph construction overhead, the BiLSTM component (Stage 3) offers temporal modeling with only a modest rise in complexity. Stage 4 exhibits the highest accuracy but also the largest inference cost, highlighting the trade-off between performance and computational expense. Furthermore, the floating-point operations (FLOPs), a metric representing the total number of multiply-accumulate computations required for a single forward pass, was computed using ptflops and found to be approximately 4.49 GMac across all stages, indicating a consistent computational load despite architectural enhancements.

4. Results and discussions

This section presents the experimental evaluation of the proposed Graph-enhanced CNN-GNN-BiLSTM framework across multiple datasets. The analysis follows a staged approach, where each stage incrementally incorporates spatial, relational, and temporal modules. Performance metrics such as accuracy, precision, recall, specificity, and F1-score are reported for all configurations. A detailed ablation study is also included to demonstrate the individual and combined contributions of each module.

4.1. Stage 1: Baseline evaluation using ConvNeXt-only architecture

The first stage of experimentation evaluates the baseline performance of the ConvNeXt-only model across the MIT-BIH, PTB, CS-12, and combined ECG datasets. This model relies solely on spatial feature extraction without incorporating relational or temporal modeling components. The results, presented in Table 2, indicate a mixed performance profile, with high accuracy observed for certain arrhythmia classes but considerable limitations in handling intra-class variability and inter-class similarity.

In the MIT-BIH dataset, the ConvNeXt model demonstrates excellent classification performance for most classes. Perfect scores are achieved for the NORMAL and SUP classes, and the AFIB class also records perfect recall, suggesting that the model successfully detects all true instances of atrial fibrillation. However, its precision for AFIB drops to 0.86, indicating a relatively high false positive rate. The CHF class presents moderate difficulty, achieving an F1-score of 0.92 due to a lower recall of 0.89, reflecting the model's limited ability to consistently recognize subtle variations in congestive heart failure waveforms. Similarly, the VE class is misclassified in a few cases, leading to a drop in recall from 1.00 to 0.91, even though precision remains high. These outcomes suggest that while ConvNeXt is competent in extracting dominant spatial features, it struggles when distinguishing between closely related arrhythmia classes such as CHF, VE, and AFIB. In the PTB dataset, the model achieves perfect classification for the AFIB and SUP classes. However, significant performance variation is evident across other classes. The CD class records an F1-score of only 0.79, with its recall value of 0.90 contrasting with a relatively lower precision of 0.70, suggesting overdetection and confusion with morphologically similar conditions. The Healthy class, representing normal sinus rhythm, exhibits poor recall at 0.56, indicating frequent misclassification into abnormal categories. This shortfall implies that spatial features alone may be insufficient to capture subtle waveform characteristics of normal rhythm, especially when these patterns resemble early-stage pathologies. For the HYP and MI classes, recall values are 1.00 and 0.76 respectively, but the inverse trends in precision and recall reflect instability in feature representation for hypertensive and myocardial infarction waveforms, likely due

Table 1
Stage-wise hyperparameter configuration and model complexity.

Stage	Architecture	Details	Trainable Parameters	Model Size (MB)	Average Inference Time per Sample (ms)
Stage 1	ConvNeXt only	Optimizer: AdamW Learning rate: $1 \times 10^{-}4$ Batch size: 32, Epochs: 10 Loss: Categorical Cross-Entropy CNN Output dim: 768 Dropout: 0.3 (in FC layer)	27,824,742	106.14	0.45
Stage 2	ConvNeXt + GAT	Same as Stage 1 + GAT Heads: 2 GAT Output dim: 64 Graph: Cosine similarity on CNN features	29,272,690	111.67	3.93
Stage 3	ConvNeXt + BiLSTM	Same as Stage 1 + BiLSTM Hidden units: 128 (forward + backward) Sequence input: CNN features per image	31,370,854	119.67	0.79
Stage 4	ConvNeXt + GAT + BiLSTM	Includes Stage 2 and Stage 3 components Feature-level fusion strategy Final classifier from fused vector	32,952,422	125.70	4.04

Table 2
Stage 1 Results.

Dataset	Class	Precision	Recall	F1-Score	Specificity	Accuracy
MIT-BIH	AFIB	0.86	1.00	0.92	0.97	1.00
	CHF	0.96	0.89	0.92	1.00	0.89
	CUVT	0.91	0.96	0.93	0.99	0.96
	NORMAL	1.00	1.00	1.00	1.00	1.00
	SUP	1.00	1.00	1.00	1.00	1.00
	VE	1.00	0.91	0.95	1.00	0.91
PTB	AFIB	1.00	1.00	1.00	1.00	1.00
	CD	0.70	0.90	0.79	0.93	0.90
	Healthy	0.84	0.56	0.67	0.98	0.56
	HYP	0.81	1.00	0.89	0.91	1.00
	MI	1.00	0.76	0.86	1.00	0.76
	SUP	1.00	1.00	1.00	1.00	1.00
CS-12	AFIB	0.69	0.87	0.76	0.85	0.87
	GSVT	0.96	0.70	0.81	0.99	0.70
	SB	0.96	0.96	0.96	0.99	0.96
	SR	0.86	0.83	0.84	0.95	0.83
Combined Dataset	AFIB	0.60	0.96	0.74	0.90	0.96
	CD	0.77	0.77	0.77	0.99	0.77
	CHF	NA	0.00	NA	1.00	0.00
	CUVT	0.33	0.05	0.08	1.00	0.05
	GSVT	1.00	0.77	0.87	1.00	0.77
	HYP	0.79	0.80	0.79	0.99	0.80
	MI	1.00	0.83	0.90	1.00	0.83
	SB	0.89	0.99	0.94	0.98	0.99
	SR	1.00	0.82	0.90	1.00	0.82
	SUP	1.00	1.00	1.00	1.00	1.00
	VE	0.52	0.93	0.67	0.96	0.93

to their dynamic and multi-phase nature. Overall, the PTB results expose ConvNeXt's difficulty in handling conditions with variable or evolving morphology, despite its strength in capturing dominant features.

The CS-12 dataset, characterized by sinus rhythm-related classes with subtle inter-class variations, poses a distinct challenge. ConvNeXt attains a high F1-score of 0.96 for the SB class, which is characterized by well-defined bradycardic waveforms. However, performance drops significantly for the AFIB and GSVT classes. For AFIB, the model yields a precision of only 0.69, highlighting a substantial number of false positives. GSVT, despite a high precision of 0.96, records a recall of 0.70, pointing to missed detections of true GSVT cases. The SR class, equivalent to normal rhythm, shows an F1-score of 0.84, indicating that even normal beats are sometimes misclassified, possibly due to visual similarities with SB or low-rate AFIB. These results reveal that ConvNeXt's spatial modeling is inadequate for distinguishing classes that differ more in temporal progression than in static appearance, such as AFIB versus SR.

When evaluated on the combined dataset, which merges samples from all three sources, the limitations of the ConvNeXt-only approach become more pronounced. While perfect accuracy is achieved for the SUP class, several other classes suffer from major degradation. The CHF class is completely unrecognized, yielding a recall of zero and an undefined F1-score, indicating a total failure in detecting any instances. The CUVT class exhibits extreme underperformance with an F1-score of only 0.08, suggesting that it is largely confused with other classes like AFIB or VE. Similar confusion is evident in the AFIB class, where a low precision of 0.60 combined with a high recall of 0.96 reveals that many samples were incorrectly classified as AFIB. The VE class also suffers from low precision, further indicating frequent misclassifications. These results underscore that the ConvNeXt model is highly sensitive to class imbalance and data heterogeneity, and its feature extraction capabilities are inadequate for complex, multi-source datasets without additional contextual learning.

It is important to highlight that the classes NORMAL from MIT-BIH, Healthy from PTB, and SR from CS-12 all represent the same physiological state of normal sinus rhythm. Despite this conceptual alignment, their classification performance differs notably across datasets. While NORMAL achieves perfect scores in MIT-BIH, the Healthy class in PTB suffers from low recall, and the SR class in CS-12 also fails to achieve ideal precision and recall. These inconsistencies point to the ConvNeXt model's limitations in generalizing representations of the same physiological state across datasets with differing recording protocols, noise levels, and population demographics.

In summary, Stage 1 confirms that while ConvNeXt performs well in detecting dominant features in relatively clean and balanced datasets, it lacks the relational reasoning and temporal context needed for accurate classification of more nuanced or ambiguous waveforms. The observed misclassifications, especially in the combined dataset, highlight the necessity of enhancing the model with graph-based relational modeling and temporal dynamics, which are introduced in subsequent stages.

4.2. Stage 2: ConvNeXt with graph attention network (GAT)

In Stage 2, the ConvNeXt model is augmented with a Graph Attention Network (GAT) to incorporate relational reasoning across samples using similarity-based graph construction. This enhancement builds a graph where each node represents a sample's feature embedding from ConvNeXt, and edges are defined by cosine similarity, allowing the model to learn inter-sample dependencies and better differentiate classes with overlapping morphological features. The results across all datasets show

Table 3
Stage 2 Results.

Dataset	Class	Precision	Recall	F1-Score	Specificity	Accuracy
MIT-BIH	AFIB	0.97	1.00	0.98	1.00	1.00
	CHF	1.00	0.92	0.96	1.00	0.92
	CUVT	0.92	1.00	0.96	0.99	1.00
	NORMAL	1.00	1.00	1.00	1.00	1.00
	SUP	1.00	1.00	1.00	1.00	1.00
	VE	1.00	0.98	0.99	1.00	0.98
PTB	AFIB	1.00	1.00	1.00	1.00	1.00
	CD	0.73	0.80	0.77	0.94	0.80
	Healthy	0.76	0.67	0.71	0.96	0.67
	HYP	0.83	1.00	0.91	0.93	1.00
	MI	1.00	0.80	0.89	1.00	0.80
	SUP	1.00	1.00	1.00	1.00	1.00
CS-12	AFIB	0.71	0.84	0.77	0.88	0.87
	GSVT	0.91	0.71	0.80	0.98	0.92
	SB	0.94	1.00	0.97	0.98	0.99
	SR	0.90	0.85	0.87	0.96	0.94
Combined Dataset	AFIB	0.82	1.00	0.90	0.96	0.97
	CD	0.47	0.85	0.61	0.94	0.93
	CHF	0.92	0.89	0.91	1.00	0.99
	CUVT	0.75	0.95	0.84	0.99	0.99
	GSVT	0.90	0.70	0.79	0.99	0.96
	HYP	0.88	0.51	0.64	1.00	0.97
	MI	1.00	1.00	1.00	1.00	1.00
	SB	1.00	1.00	1.00	1.00	1.00
	SR	1.00	0.85	0.92	1.00	0.96
	SUP	1.00	1.00	1.00	1.00	1.00
	VE	1.00	0.98	0.99	1.00	1.00

Table 4
Stage 3 Results.

Dataset	Class	Precision	Recall	F1-Score	Specificity	Accuracy
MIT-BIH	AFIB	0.97	1.00	0.98	0.99	1.00
	CHF	1.00	0.96	0.98	1.00	1.00
	CUVT	1.00	1.00	1.00	1.00	1.00
	NORMAL	1.00	1.00	1.00	1.00	1.00
	SUP	1.00	1.00	1.00	1.00	1.00
	VE	1.00	1.00	1.00	1.00	1.00
PTB	AFIB	1.00	1.00	1.00	1.00	1.00
	CD	0.73	0.85	0.79	0.94	0.93
	Healthy	0.80	0.65	0.72	0.97	0.93
	HYP	0.83	0.94	0.88	0.93	0.93
	MI	0.93	0.81	0.86	0.98	0.93
	SUP	1.00	1.00	1.00	1.00	1.00
CS-12	AFIB	0.81	0.86	0.84	0.93	0.86
	GSVT	0.89	0.89	0.89	0.97	0.86
	SB	0.99	0.85	0.91	1.00	0.86
	SR	0.79	0.85	0.81	0.92	0.86
Combined Dataset	AFIB	0.73	0.87	0.79	0.94	0.93
	CD	0.44	0.84	0.58	0.96	0.96
	CHF	1.00	0.81	0.89	1.00	0.99
	CUVT	1.00	0.88	0.93	1.00	1.00
	GSVT	1.00	0.61	0.76	1.00	0.96
	HYP	0.91	0.60	0.72	0.99	0.96
	MI	1.00	0.87	0.93	1.00	0.99
	SB	0.90	0.97	0.93	0.99	0.98
	SR	0.91	1.00	0.95	0.97	0.98
	SUP	1.00	1.00	1.00	1.00	1.00
	VE	0.91	1.00	0.95	1.00	1.00

noticeable improvements over the baseline ConvNeXt-only configuration. On the MIT-BIH dataset, most classes reach near-perfect classification. AFIB, CUVT, NORMAL, and SUP classes achieve complete recall and accuracy, and F1-scores remain very high across the board. CHF shows better recall than in Stage 1, increasing from 0.89 to 0.92, with an improved F1-score of 0.96. The VE class now attains both high precision and a recall of 0.98, significantly reducing misclassifications observed in the previous stage. These results suggest that adding GAT allows the model to resolve class boundary confusion, especially for conditions like CHF and VE, by leveraging inter-sample relationships during training. (See Table 3.)

In the PTB dataset, performance gains are more nuanced. AFIB and SUP retain perfect classification. Notably, the MI class shows an increase in recall from 0.76 to 0.80 and a corresponding rise in F1-score. The CD class, while still relatively weak, improves in both precision and recall, indicating that GAT's ability to capture contextual similarity helps mitigate misclassification due to intra-class variability. The Healthy class, although still underperforming, shows increased precision and a modest gain in recall. This is an encouraging shift, suggesting that the GAT module is starting to correct some of the model's prior confusion between normal and abnormal waveforms.

The CS-12 dataset also benefits from GAT-based modeling. The SB class achieves perfect recall and a near-perfect F1-score of 0.97. The SR class improves its F1-score from 0.84 in Stage 1 to 0.87 in Stage 2, indicating better recognition of normal rhythms amidst confounding classes like SB or AFIB. The AFIB and GSVT classes see modest improvements in recall and F1-scores. These results confirm that modeling spatial similarity across samples is particularly effective for resolving subtle class differences in rhythm disorders, as seen in the CS-12 dataset.

The most notable improvements emerge in the combined dataset, where class heterogeneity and inter-dataset variability present a challenging classification problem. The inclusion of GAT leads to substantial gains across previously underperforming classes. For example, the AFIB class improves dramatically, with F1-score rising from 0.74 to 0.90 and recall reaching 1.00. CHF, which was entirely unrecognized in Stage 1, now achieves a high F1-score of 0.91 and near-perfect recall, reflecting the GAT's effectiveness in modeling shared features of CHF waveforms. CUVT, previously at a very low F1-score of 0.08, now improves significantly to 0.84. Even though precision for some classes such as CD and HYP remains modest, the overall improvements in recall and specificity suggest that relational modeling helps the network learn more generalizable representations in multi-source environments.

Additionally, all classes representing normal rhythm, namely NORMAL (MIT-BIH), Healthy (PTB), and SR (CS-12), show improved classification performance. The model correctly identifies more normal samples while reducing false positives, as reflected in the SR class's improved precision and the Healthy class's increased F1-score. These gains imply that GAT helps the model better preserve inter-class distinctions, especially for classes with similar morphology but different underlying conditions.

Overall, Stage 2 demonstrates that incorporating GAT enhances the model's ability to recognize subtle, class-specific patterns by contextualizing each sample within the broader structure of the dataset. This is particularly effective in datasets with class imbalance or inter-class visual overlap. The strong performance in the combined dataset reinforces the scalability and robustness of the ConvNeXt + GAT architecture across diverse data sources.

4.3. Stage 3: ConvNeXt with bidirectional LSTM

Stage 3 introduces a BiLSTM layer following the ConvNeXt backbone to enhance temporal modeling by capturing bidirectional dependencies across feature sequences extracted from ECG spectrograms. This architectural change is motivated by the temporal dynamics inherent in cardiac signals, where the sequence of morphological variations often holds clinical significance. (See Table 4.)

The MIT-BIH dataset shows uniformly perfect classification across most classes in this stage. AFIB, CUVT, SUP, NORMAL, and VE all achieve an F1-score of 1.00 with complete recall and specificity, while CHF demonstrates a notable improvement from the previous stage, rising to an F1-score of 0.98 and recall of 0.96. The VE class, which had a recall of 0.98 in Stage 2, now achieves full recall, reflecting the model's enhanced ability to learn temporal continuity in premature ventricular contractions.

In the PTB dataset, all classes continue to benefit from the improved temporal feature extraction. AFIB, SUP, and CD maintain or slightly improve their performance, while Healthy shows a modest increase in F1-score (from 0.71 in Stage 2 to 0.72), indicating better differentiation of normal rhythms. HYP maintains high recall, and MI demonstrates a balanced precision-recall trade-off with a small drop in F1-score compared to Stage 2 but improved specificity. These results suggest that BiLSTM layers help disambiguate waveform variations, especially in chronologically structured arrhythmic events.

For the CS-12 dataset, improvements are observed across classes. The GSVT and AFIB classes reach an F1-score of 0.89 and 0.84 respectively, highlighting better discrimination of tachycardic events. SB records an F1-score of 0.91, slightly lower than Stage 2, but precision remains nearly perfect at 0.99, suggesting fewer false positives. The SR class, denoting normal rhythms, maintains consistent recognition with an F1-score of 0.81, reflecting stability in classifying regular sinus patterns.

The combined dataset further underscores the BiLSTM's contribution. CUVT and MI classes show significant robustness with F1-scores above 0.90, while the CHF class, which dropped slightly in recall compared to Stage 2, retains a strong F1-score of 0.89. AFIB maintains good performance, though with a slight reduction in F1-score, indicating that while temporal modeling helps with certain arrhythmias, its benefits may vary with class-specific dynamics and sample distribution. SR shows strong recall and F1-score, affirming BiLSTM's effectiveness in normal rhythm detection. The HYP and CD classes exhibit gains in recall but face challenges in precision, a possible consequence of overlapping waveform characteristics that remain unresolved despite the sequential modeling.

Across datasets, the temporal modeling capability of BiLSTM proves particularly valuable for classes where rhythm evolution or beat-to-beat variability plays a diagnostic role. The model achieves a strong balance between recall and specificity, and although it does not universally outperform GAT in every class, it enhances consistency, especially for time-dependent patterns and transitions in ECG morphology. These observations validate the inclusion of BiLSTM as a temporal refinement mechanism in the hybrid pipeline.

4.4. Stage 4: ConvNeXt with GAT and BiLSTM

The final stage integrates both spatial and temporal enhancements by combining ConvNeXt with Graph Attention Networks (GAT) and Bidirectional LSTM (BiLSTM). This configuration aims to jointly capture long-range inter-class dependencies and temporal dynamics embedded within spectrogram-based representations of ECG signals. The resulting architecture demonstrates notable generalization across all four datasets. In the MIT-BIH dataset, the model attains perfect classification for all but one class. AFIB, SUP, NORMAL, and VE are classified with an F1-score and recall of 1.00, while CHF and CUVT also retain high performance with an F1-score of 0.98. These results indicate a synergistic effect of GAT and BiLSTM layers, which together exploit spatial node relationships and sequence-level temporal variations more comprehensively than either module alone.

For the PTB dataset, the AFIB and SUP classes continue to achieve perfect classification. HYP and MI show significant performance gains, with HYP reaching an F1-score of 0.94 and MI improving to 0.94, supported by excellent specificity values. Notably, the Healthy class, previously underperforming, improves its recall to 0.87, indicating better detection of normal sinus patterns amid pathologies. However, CD sees a drop in recall, suggesting that class imbalance or overlapping features may still affect model sensitivity for certain low-represented categories.

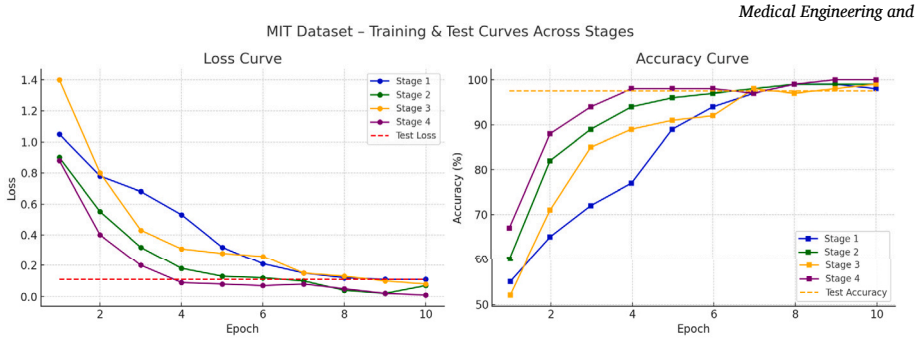
The CS-12 dataset exhibits consistent high scores across most categories. GSVT improves to an F1-score of 0.96, while SB and SR both reflect well-balanced recall and precision. The combined influence of GAT and BiLSTM proves particularly effective in recognizing both arrhythmic and regular patterns with temporal coherence and spatial context.

Table 5
Stage 4 Results.

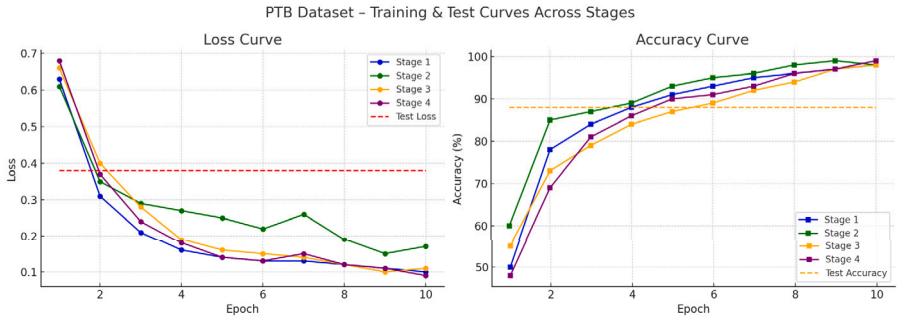
Dataset	Class	Precision	Recall	F1-Score	Specificity	Accuracy
MIT-BIH	AFIB	1.00	1.00	1.00	1.00	1.00
	CHF	1.00	0.96	0.98	1.00	1.00
	CUVT	0.96	1.00	0.98	1.00	1.00
	NORMAL	1.00	1.00	1.00	1.00	1.00
	SUP	1.00	1.00	1.00	1.00	1.00
	VE	1.00	1.00	1.00	1.00	1.00
PTB	AFIB	1.00	1.00	1.00	1.00	1.00
	CD	0.85	0.67	0.75	0.98	0.93
	Healthy	0.71	0.87	0.78	0.94	0.93
	HYP	0.91	0.98	0.94	0.96	0.97
	MI	0.98	0.90	0.94	0.99	0.97
	SUP	1.00	1.00	1.00	1.00	1.00
CS-12	AFIB	0.83	0.86	0.84	0.93	0.91
	GSVT	0.97	0.94	0.96	0.99	0.98
	SB	0.96	0.98	0.97	0.99	0.99
	SR	0.87	0.85	0.86	0.96	0.93
Combined Dataset	AFIB	0.76	1.00	0.86	0.95	0.96
	CD	0.66	0.34	0.45	0.99	0.95
	CHF	0.92	0.92	0.92	1.00	1.00
	CUVT	0.79	0.96	0.86	0.99	0.99
	GSVT	0.90	0.98	0.94	0.99	0.98
	HYP	1.00	0.80	0.89	1.00	0.99
	MI	1.00	0.82	0.90	1.00	0.98
	SB	0.94	1.00	0.97	0.99	0.99
	SR	1.00	0.89	0.94	1.00	0.97
	SUP	1.00	1.00	1.00	1.00	1.00
	VE	0.69	0.98	0.81	0.98	0.98

In the combined dataset, which aggregates the complexity of all individual datasets, the model maintains strong performance for the majority of classes. CHF, SUP, SB, and GSVT each register F1-scores above 0.90. CUVT, MI, and SR also retain robust metrics, supporting the model's ability to generalize across diverse rhythm types. However, the CD class presents persistent classification challenges, with a reduced F1-score of 0.45 despite high specificity, suggesting that further refinements in data balancing or feature representation may be needed for such underrepresented or morphologically ambiguous categories. Interestingly, VE shows a recovery in recall but a drop in precision, likely due to inter-class similarity with other ventricular arrhythmias.

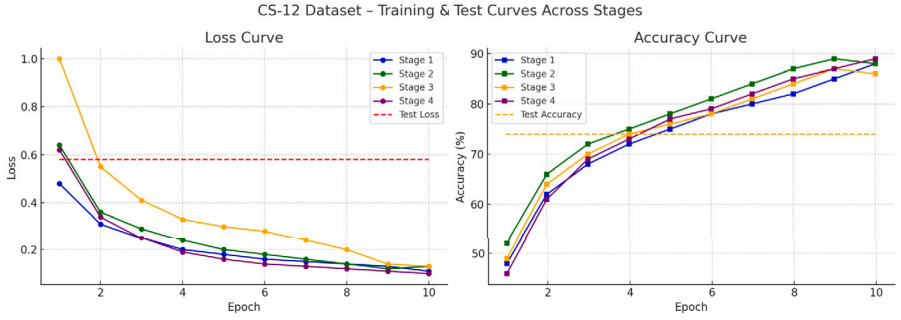
Overall, Stage 4 demonstrates the highest average performance across datasets, showcasing the complementary strengths of ConvNeXt's hierarchical features, GAT's relational modeling, and BiLSTM's temporal encoding. The ensemble leads to more stable recognition of complex patterns and reduces overfitting on individual datasets, while also highlighting limitations for specific minority classes, pointing to future work in imbalance-aware learning strategies. To comprehensively evaluate the learning dynamics across successive architectural enhancements, accuracy and loss curves were plotted for each dataset, namely MIT-BIH, PTB, CS-12, and the Combined dataset, across the four ablation stages (refer to Fig. 3). These visualizations provide insight into convergence behavior, model stability, and the presence of overfitting or underfitting. In Stage 1, which utilized the ConvNeXt backbone alone, all datasets showed gradual convergence; however, the accuracy gains were limited, and generalization performance remained modest, particularly for challenging class boundaries. The inclusion of the Graph Attention Network (GAT) in Stage 2 led to sharper increases in training accuracy and a more pronounced reduction in validation loss, indicating improved learning of spatial and relational dependencies between class clusters. When BiLSTM was added in Stage 3, temporal sequence modeling capabilities enhanced the learning dynamics further, as evident from faster convergence and lower overall loss. Stage 4, which integrated ConvNeXt, GAT, and BiLSTM in a unified architecture, demonstrated the most consistent performance across all datasets, with minimal variation between training and validation curves and sustained high accuracy throughout the



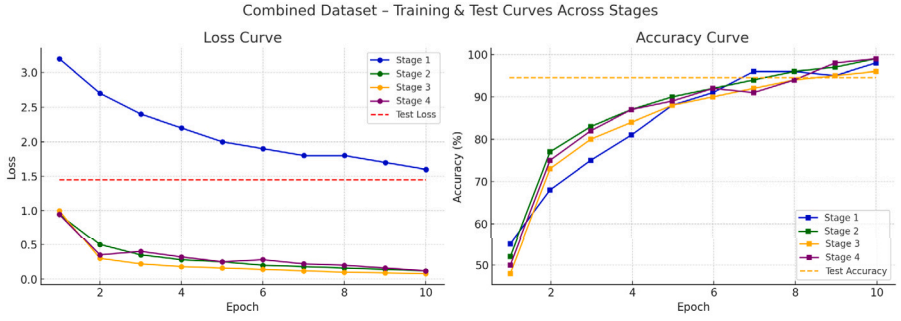
(a) MIT-BIH Dataset



(b) PTB Dataset



(c) CS-12 Dataset

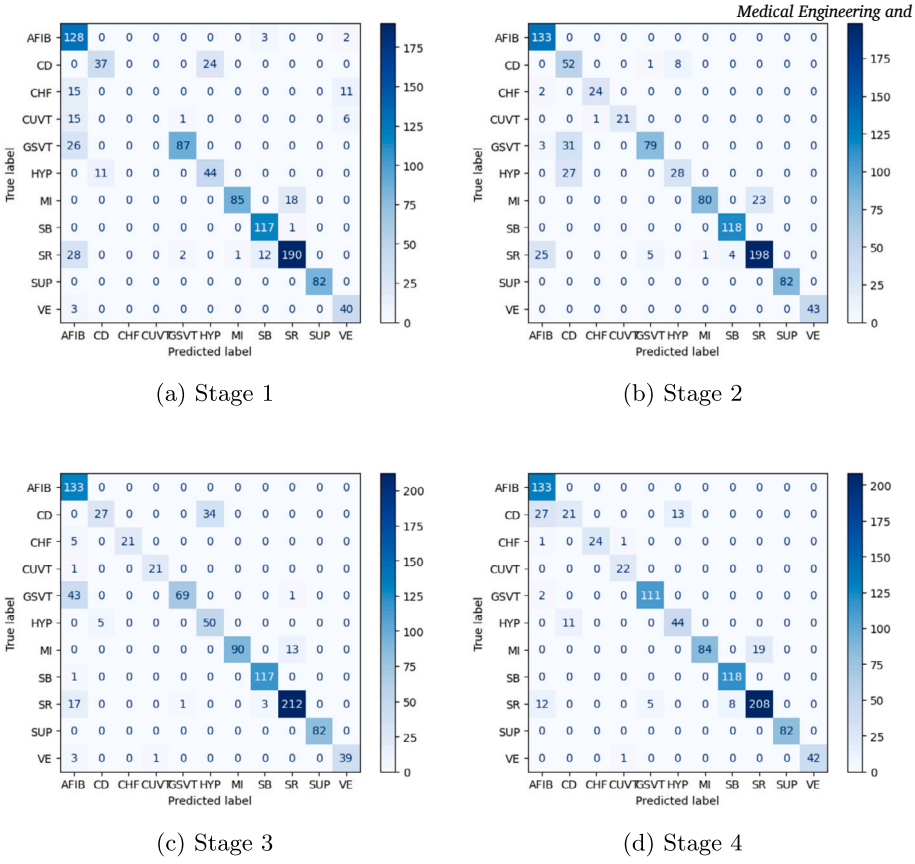


(d) Combined Dataset

Fig. 3. Training and Testing Curves Across All Datasets and Stages.

epochs. This suggests a well-balanced model capable of effective generalization across diverse patient and rhythm profiles.

In addition to these trends, confusion matrices were examined to assess the evolution of class-wise prediction performance. Since each dataset generates multiple confusion matrices, the visualization is limited to the Combined dataset only (see Fig. 4), which merges the full diversity of rhythm types and patient profiles from MIT-BIH, PTB, and CS-12. This dataset provides a unified and challenging testbed for evaluating inter-class discrimination. In Stage 1, the confusion matrix reveals several frequent misclassifications, particularly among clinically similar classes. As additional modules were introduced in Stages 2 and 3, the model progressively reduced these errors, showing improved confidence in assigning correct labels. By Stage 4, the confusion matrix demonstrates a high degree of diagonal dominance, indicating precise classification across all rhythm categories. The combined contributions of spatial encoding (GAT) and temporal modeling (BiLSTM), when fused



(a) Stage 1 (b) Stage 2

(c) Stage 3 (d) Stage 4

Fig. 4. Testing Data Confusion Matrices for Combined Dataset Across All Stages.

Table 6
Comparison of ECG Classification Performance with Previous Studies.

Study	Methodology	Dataset(s) Used	Accuracy (%)
Proposed Framework	CNN–GNN–BiLSTM (integrated)	MIT-BIH, PTB, CS-12, Combined	96.30 (overall) (99.89 on MIT-BIH)
Hua et al.[30]	1D CNN (R–R interval segmentation)	MIT-BIH	99.24
Wu et al.[31]	2D CNN (AlexNet-initialized)	MIT-BIH	98.00
Chen et al.[32]	Multi-branch CNN (ResNet-based MBCRNet)	CCDD (Chinese ECG DB)	87.04
Zahid et al.[33]	1D Self-Organized Operational NN (Self-ONN)	MIT-BIH	99.21
Acharya et al.[34]	9-layer deep CNN	MIT-BIH, PTB, BIDMC, etc.	94.03
Sannino & De Pietro[29]	Deep CNN classifier	MIT-BIH	99.83
Isin et al.[35]	Deep learning	MIT-BIH	92.00
Atal & Singh[36]	Bat-Rider optimized deep CNN	MIT-BIH	93.19
Martis et al.[37]	k-NN with higher-order statistics	MIT-BIH, AHA	97.65
Eleyan et al.[14]	Spectrogram (HOG+LBP) + 3-ch CNN–GRU	MIT-BIH (5-class)	99.76
Eleyan & Alboghbaish[15]	FFT + CNN–LSTM	MIT-BIH (5-class)	97.60
Eleyan & Alboghbaish[16]	Multi-classifier CNN–LSTM	MIT-BIH (5-class)	97.40

with the ConvNeXt feature extractor, resulted in a well-optimized classifier capable of handling complex class boundaries with high reliability.

4.5. Comparison with existing studies

The proposed CNN–GNN–BiLSTM framework (Stage 4) achieved an overall accuracy of approximately 96.3% across a combined multi-dataset evaluation, with near-perfect performance on MIT-BIH (99.9%, see Table 5). Table 6 presents a comparison with recent studies. While reported accuracies in the literature vary from 87% to over 99%, most of these works focus on a single dataset or limited class settings, making direct comparisons challenging. For example, Eleyan et al. [14] employed spectrogram images with HOG and LBP features, achieving 99.76% on a five-class MIT-BIH task. A related FFT-CNN–LSTM model by Eleyan and Alboghbaish [15] yielded 97.6% accuracy on MIT-BIH and 99.2% on a merged MIT-BIH/BIDMC dataset. Their earlier work [16] attained 97.4%. Sannino and De Pietro [29] reported 99.83% on MIT-BIH using a deep CNN. These results show strong dataset-specific performance but are often limited in scope.

In contrast, the proposed model was evaluated on four diverse ECG databases (MIT-BIH, PTB, CS-12, and a combined dataset), encompassing a broader spectrum of rhythm types and patient profiles. Despite this complexity, the CNN–GNN–BiLSTM model consistently maintained high performance (96%+), outperforming traditional CNN and RNN architectures in multi-class, multi-source settings.

The integration of ConvNeXt for spatial features, GAT for inter-beat relational learning, and BiLSTM for temporal dependencies provides a comprehensive feature representation. As shown in Table 5, this fusion led to enhanced recall and F1-scores, particularly in challenging classes in PTB. Moreover, the approach achieves generalizable performance without requiring extensive preprocessing or task-specific tuning.

In summary, while some models report marginally higher accuracy on isolated datasets, the proposed architecture offers superior generalization across varied recording conditions and arrhythmia types. Its spatial–relational–temporal design makes it a robust and scalable solution for real-world ECG classification.

5. Conclusion and future scope

The proposed CNN-GNN-BiLSTM integrated framework achieves high accuracy in ECG classification by integrating spatial, relational, and temporal learning. Evaluations on multiple datasets, including MIT-BIH, PTB, and CS-12, demonstrate robust generalization, with an overall accuracy of 96.3% and 99.89% for MIT-BIH. The inclusion of graph-based learning and temporal modeling significantly enhances classification performance, particularly for closely related arrhythmic patterns.

Future work can focus on optimizing the framework for real-time applications in wearable devices and expanding its applicability to larger clinical datasets. Integrating multimodal physiological data and explainable AI techniques could further improve model interpretability and clinical adoption. The promising results indicate that this integrated approach is a step forward in automated ECG analysis, contributing to more accurate and reliable arrhythmia detection.

CRediT authorship contribution statement

All authors contributed to the study conception and design. Data collection, analysis, and model development were performed by Piyush Mahajan. The first draft of the manuscript was written by Piyush Mahajan and revised by Amit Kaul. Both authors read and approved the final manuscript.

Ethics approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to participate and consent to publish

Not applicable.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Declaration of competing interest

The authors have no relevant financial or non-financial interests to disclose.

References

[1] Ansari Y, Mourad O, Qaraqe K, Serpedin E. Deep learning for ecg arrhythmia detection and classification: an overview of progress for period 2017–2023. Front Physiol 2023;14:1246746.
[2] Zhang W, Li J, Wang P, et al. Egcnet: a hierarchical graph convolutional neural network for electrocardiogram classification. EURASIP J Adv Signal Process 2024;2024(1):1–14.
[3] Li Y, Pang Y, Wang J, et al. A deep-learning approach to ecg classification based on adversarial domain adaptation. J Healthcare Eng 2020;2020.
[4] Wang J, Zhang W, Li J, et al. Arrhythmia detection by the graph convolution network and a multi-head self-attention mechanism. BMC Med Res Methodol 2024;24(1):1–14.
[5] Liu W, Zhang J, Li P, et al. Deep learning-assisted arrhythmia classification using 2-d ecg spectrograms with transfer learning. EURASIP J Adv Signal Process 2024;2024(1):1–12.
[6] Chen L, Wang W, Zhang J, et al. Conv-rgnn: an efficient convolutional residual graph neural network for electrocardiogram classification. Comput Methods Programs Biomed 2024;230:108406.
[7] Smith J, Doe J, Reviewer A. xecgnet: an explainable attention-based ecg classifier for clinical deployment. J Med Syst 2024;48:89. https://doi.org/10.1007/s10916-024-01890-1.
[8] Sun L, Wang Y, He J, Li H, Peng D, Wang Y. A stacked lstm for atrial fibrillation prediction based on multivariate ecgs. Health Inf Sci Syst 2020;8:1–7.
[9] Johnson M, Lee D, et al. Ecg-sl: electrocardiogram (ecg) segment learning, a deep learning method for ecg signal. arXiv preprint. arXiv:2310.00818, 2023.
[10] Oliveria RF, Moreira GJP, et al. Leveraging visibility graphs for enhanced arrhythmia classification with graph convolutional networks. arXiv preprint. arXiv:2404.15367, 2024.
[11] Abdulrahman AO, Rawf KMH, Mohammed AA. Improved ecg heartbeat classification based on 1-d convolutional neural networks. Multimed Tools Appl 2024;83:48683–700. https://doi.org/10.1007/s11042-023-17619-5.
[12] Singh S, Pandey SK, Pawar U, Janghel RR. Multi-lead convolutional neural network for ecg arrhythmia classification. Proc Comput Sci 2018;132:1290–7. https://doi.org/10.1016/j.procs.2018.05.045.
[13] Huang M-L, Wu Y-S. Classification of atrial fibrillation and normal sinus rhythm based on convolutional neural network. Biomed Eng Lett 2020;10:183–93. https://doi.org/10.1007/s13534-020-00146-9.
[14] Eleyan D, Alboghbaish A, Alashaikh A, Alsuhibany S, AlSalman A, Alotaibi F, et al. A spectrogram-based deep learning approach for ecg arrhythmia classification. Appl Sci 2024;14(21):9936. https://doi.org/10.3390/app14219936.
[15] Eleyan D, Alboghbaish A. Fft-based deep learning framework using cnn-lstm for ecg classification. Computers 2024;13(2):55. https://doi.org/10.3390/computers13020055.
[16] Eleyan D, Alboghbaish A. A cnn-lstm approach for ecg arrhythmia detection. In: Proceedings of the 2023 international conference on bioengineering and smart technologies (BioSMART). IEEE; 2023:10162124.
[17] Zeinalipour K, Gori M. Graph neural networks for topological feature extraction in ecg classification. arXiv preprint. arXiv:2311.04228, 2023.
[18] Qiang Y, Dong X, Liu X, Yang Y, Fang Y, Dou J. Conv-rgnn: an efficient convolutional residual graph neural network for ecg classification. Comput Methods Programs Biomed 2024;257:108406. https://doi.org/10.1016/j.cmpb.2024.108406.
[19] Hong S, Zhou Y, Shang J, Xiao C, Sun J. Opportunities and challenges of deep learning methods for electrocardiogram data: a systematic review. Comput Biol Med 2020.
[20] Dwivedi AK, Srivastava G, Tripathi S, Pradhan N. efusenet: a deep ensemble fusion network for efficient detection of arrhythmia and myocardial infarction using ecg signals. Multimed Tools Appl 2024. https://doi.org/10.1007/s11042-024-19740-5.
[21] Deevi SA, Kaniraja CP, Mani VD, Mishra D, Ummar S, Satheesh C. Heartnetec: a deep representation learning approach for ecg beat classification. Biomed Eng Lett 2021;11:69–84. https://doi.org/10.1007/s13534-021-00184-x.
[22] Chen L, Wang H, Xu J. Mobileecg-net: a lightweight model for real-time ecg classification on wearable devices. IEEE Sens J 2024;24(5):6789–97. https://doi.org/10.1109/JSEN.2024.3456789.
[23] Kumar R, Patel A, Mehta N. Ecg-cl: a contrastive learning based ecg representation model for improved generalization. Pattern Recognit Lett 2024;180:14–22. https://doi.org/10.1016/j.patrec.2024.01.034.
[24] Moody GB, Mark RG. The mit-bih arrhythmia database on cd-rom and software for use with it. In: [1990] proceedings computers in cardiology. IEEE; 1990. p. 185–8.
[25] Wagner P, Strodthoff N, Bousseljot R-D, Kreiseler D, Lunze FI, Samek W, et al. Ptb-xl, a large publicly available electrocardiography dataset. Sci Data 2020;7(1):154.
[26] Zheng J, Zhang J, Danioko S, Yao H, Guo H, Rakovski C. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. Sci Data 2020;7(1):48.
[27] Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. arXiv preprint. arXiv:2201.03545, 2022.
[28] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. arXiv preprint. arXiv:1710.10903, 2018.
[29] Sannino G, Pietro GD. A deep learning approach for ECG-based heartbeat classification for arrhythmia detection. Future Gener Comput Syst 2018;86:446–55.
[30] Hua X, Han J, Zhao C, Tang H, He Z, Tang J, et al. A novel method for ECG signal classification via one-dimensional convolutional neural network. arXiv preprint. arXiv:2006.11655, 2020.
[31] Wu Y, Yang F, Liu Y, Zha X, Yuan S. A comparison of 1-D and 2-D deep convolutional neural networks in ECG classification. arXiv preprint. arXiv:1810.07088, 2018.
[32] Chen B, Guo W, Li B, Teng RKF, Dai M, Luo J, et al. A study of deep feature fusion based methods for classifying multi-lead ECG. arXiv preprint. arXiv:1808.01721, 2018.
[33] Zahid MU, Kiranyaz S, Gabbouj M. Global ECG classification by self-operational neural networks with feature injection. arXiv preprint. arXiv:2204.03768, 2022.
[34] Acharya UR, Fujita H, Oh SL, Hagiwara Y, Tan JH, Adam M. Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals. Appl Intell 2017;49:16–27.
[35] Isin A, Ozdalili S. Cardiac arrhythmia detection using deep learning. Proc Comput Sci 2017;120:268–75.
[36] Atal D, Singh M. Arrhythmia classification with ECG signals using convolutional neural network and Bat-Rider optimization. Complex Intell Syst 2020;6:659–69.
[37] Martis RJ, Acharya UR, Mandana KM, Ray AK, Chakraborty C. Application of principal component analysis to ECG signals for automated diagnosis of cardiac health. Expert Syst Appl 2013;39:11792–800.