EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Intelligent ECG Classification Based on Improved Swin Transformer Model

Bin Xu[1], Zongzhen Yue[1], Shoucheng Ji[1,*], Ning Sun[2], Jianyong Zheng[3]

[1]Faculty of Intelligence Technology, Shanghai Institute of Technology, Shanghai 201418, China
[2]College of Automobile and Traffic Engineering, Nanjing Forestry University, Nanjing 210037, China
[3]Institute of Artificial Intelligence, Shanghai University, Shanghai 200444, China

Abstract

INTRODUCTION: Clinical 12-lead Electrocardiogram (ECG) image classification faces key limitations, including insufficient capture of fine-grained waveform details, compromised integration of local-global rhythmic contexts, and suboptimal modeling of multi-lead spatial relationships.
OBJECTIVES: This study aimed to propose Swin-LGF-FPN, an intelligent image classification model based on Swin Transformer architecture, to address these challenges and improve the accuracy of ECG image classification for early cardiovascular disease screening.
METHODS: The enhanced framework integrated multi-scale Feature Pyramid Network (FPN) modules with an improved Swin Transformer backbone to effectively fuse local and global features. Axial Temporal Attention was incorporated to strengthen temporal feature extraction across ECG waveforms. Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations were used to demonstrate feature saliency. The model was validated on two publicly available datasets: the ECG Images dataset of Cardiac Patients and PTB-XL, with performance compared against baseline models including ResNet-34 and Vision Transformer (ViT).
RESULTS: The results indicated that Swin-LGF-FPN significantly outperformed baseline models in key metrics, including overall accuracy and F1-score. Grad-CAM visualizations showed significantly enhanced feature saliency in critical regions, as evidenced by heatmaps superimposed on original images.
CONCLUSION: The Swin-LGF-FPN model effectively classifies ECG images, showing robust performance and promising translational potential for early cardiovascular disease screening.

Keywords: ECG; Swin Transformer; FPN; Multi-scale Feature Fusion

Received on 05 May 2025, accepted on 10 December 2025, published on 02 February 2026

doi: 10.4108/eetpht.11.11675

1. Introduction

Cardiovascular diseases (CVDs), the leading global cause of mortality, claim approximately 17.9 million lives annually according to the World Health Organization. Early prevention and appropriate intervention for CVDs are critical to improve prognosis in high-risk populations [1]. Current CVD detection spans imaging modalities (ECG, echocardiography, cardiac MRI, CT) and laboratory biomarkers [2]. As a non-invasive cardiac signal recording technique, ECG is essential for early screening, disease assessment, and treatment monitoring, given its painlessness, low cost, operational simplicity, and clinical availability. These advantages underpin its status as a first-line diagnostic tool endorsed by clinical guidelines [3,4].

*Corresponding author. Email: jisc@sina.cn

Conventional ECG analysis depends on physician-based manual interpretation, entailing significant time and resource expenditures. With advances in artificial intelligence (AI), machine learning (ML) increasingly enables automated ECG abnormality detection [5,6]. Existing ML-based ECG analysis techniques primarily process raw temporal data from single-lead or multi-lead signals [7], categorized into two paradigms:

(1) *Feature-engineering-based methods,* which require expert-guided manual annotation of key characteristic points (e.g., R-peaks, QRS onset/offset) to build classification models [8];

(2) *End-to-end deep learning approaches* that directly learn abstract representations from raw signals or transformed images, thus minimizing human intervention [9].

Feature-engineering-based methods primarily employ composite features from specific waveforms, intervals, segments, or peaks [10,11]. Liu [12] developed an arrhythmia classification framework combining an enhanced dual-branch SE-ResNet with expert-defined features. Kraft [13] implemented a 1D U-Net architecture using convolutional blocks for QRS complex detection in normal sinus rhythm and premature ventricular contractions. Wu [14] achieved high-precision R-peak detection by applying a squared window variance transform to enhance QRS complexes and suppress noise, coupled with adaptive thresholding. Katamreddi [15] employed dual-tree complex wavelet transforms to extract morphological features and isolate R-peaks. Abdel-Rahman [16] adapted Faster R-CNN to estimate QRS duration from sparsely annotated ECG images. Despite strong clinical interpretability, such methods may compromise diagnostic integrity due to dissociation of local features from global rhythm patterns.

Compared with the substantial dependence of feature engineering on expert priors, end-to-end deep learning approaches directly learn intrinsic representations from raw ECG data. These methods are broadly categorized into two paradigms: raw signal-based and image-based techniques. Mantravadi [17] developed a lightweight multi-scale fusion network (CLINet) that achieved 99.94% accuracy on the MIT-BIH arrhythmia database. Fan [18] proposed the KEMT-MCAN framework to extract complex temporal features using a multi-level cross-attention network. Ahmad [19] converted raw ECG signals into three representations: Gramian angular fields, recurrence plots, and Markov transition fields, proposing dual multimodal fusion frameworks (MIF/MFF). Weimann [20] improved atrial fibrillation classification performance by 6.57% through CNN pre-training on large-scale raw ECG datasets with subsequent fine-tuning. Zhou [21] integrated hybrid-scale features with lead-encoder attention (LEA) mechanisms to fuse morphological-temporal information.

Although raw ECG signal-based classification achieves high precision, it faces two fundamental constraints. First, the scarcity of high-quality labeled datasets arises from stringent patient privacy regulations and ethical barriers, limiting publicly available resources. Existing databases typically have restricted sample sizes. For instance, the MIT-BIH Arrhythmia Database includes just 47 patient records [22], while the American Heart Association (AHA) database contains only 154 subjects [13]. Second, printed or digital ECG images are ubiquitous in medical institutions owing to universal applicability and archival convenience [23], offering a practical data source overcoming limitations in ECG anomaly classification.

Recent research has prioritized deep learning and computer vision techniques for direct diagnostic information extraction from ECG images. Jothiaruna [24] proposed a MobileNet-FPN architecture combining multi-scale feature maps with single-shot detectors (SSD) for anomaly localization, introducing weighted sigmoid focal loss to mitigate class imbalance and enhance pathological region detection. Demolder [25] developed a fully automated, deep learning-based ECG digitization method that achieves high-fidelity signal conversion from smartphone-captured images (PM-ECG-ID database) through grid correction and signal reconstruction. Sadad [26] designed an IoT-enabled cardiac monitoring system using lightweight CNNs with attention modules for four-class cardiac state classification. Hao [27] introduced an automated myocardial infarction screening framework for 12-lead ECG images, utilizing text-based lead segmentation with multi-branch feature extraction and deep fusion classification. Cao [28] proposed a weakly supervised fine-grained model identifying abnormalities in unprocessed ECG images using only image-level annotations. Fatema [29] applied artifact-removal preprocessing to enhance ECG image quality, constructing an InResNet-106 architecture integrating InceptionV3 and ResNet50. Khalid [30] introduced ECGConVT, fusing CNNs with Vision Transformers for myocardial infarction and arrhythmia classification via multilayer perceptron fusion. Ma [31] developed Mamba-RAYOLO, incorporating multi-branch feature extraction, dynamic attention mechanisms, and spatial fusion for real-time ECG image classification.

Although existing ECG image classification methods enhance feature extraction capabilities, they typically suffer from poor interpretability and frequently neglect spatial correlations across multi-lead configurations. Crucially, classification performance depends on precise fusion of local waveform details with global rhythm patterns, while confronting challenges such as lead spatial misalignment, limited image resolution, information loss, and interference from clinical annotations [28]. These factors substantially increase feature extraction and classification complexity.

To address these limitations, we developed Swin-LGF-FPN, an enhanced ECG image classification model based on Swin Transformer. The proposed model offers a novel approach for AI-assisted ECG interpretation, demonstrating potential for integration into clinical workflows. The principal contributions were:

(1) To enhance the model's capacity for perceiving details in key ECG waveforms and for integrating multi-scale features, the model incorporated a deep integration of the hierarchical windowed attention mechanism from the Swin Transformer and leveraged the strengths of the Feature Pyramid Network (FPN) in multi-scale feature extraction.

(2) The proposed ECG-LGF module enhanced local-global feature integration via an Axis-aware Temporal Attention

(ATA) mechanism, which sharpened the model's focus on diagnostically salient temporal features in ECG waveforms and augmented its overall feature extraction power.
(3) The model demonstrated superior performance across multiple evaluation metrics on two independent public ECG datasets, quantitatively confirming its efficacy in ECG image classification. Furthermore, Grad-CAM visualizations revealed that the model's decisions were consistently driven by pathologically critical features.

2. Proposed Methodology

The overall workflow of this study is schematically illustrated in Figure 1. Figure 2 depicts the architecture of the proposed Swin-LGF-FPN model. This model was designed to improve discriminatory accuracy for pathological categories through the extraction and fusion of features across multiple spatial scales and hierarchical levels in ECG images. An optimized Swin Transformer backbone was employed to extract multi-level features, which were subsequently enhanced by an ECG-LGF module and integrated into a multi-scale feature pyramid via an FPN neck. The final classification output was generated by a classifier. To address class imbalance in the dataset, Focal Loss was utilized during training, while translation augmentation was applied to mitigate overfitting.

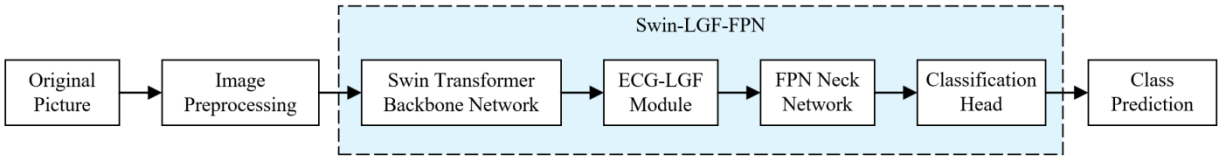


Figure 1. The overall workflow of this study

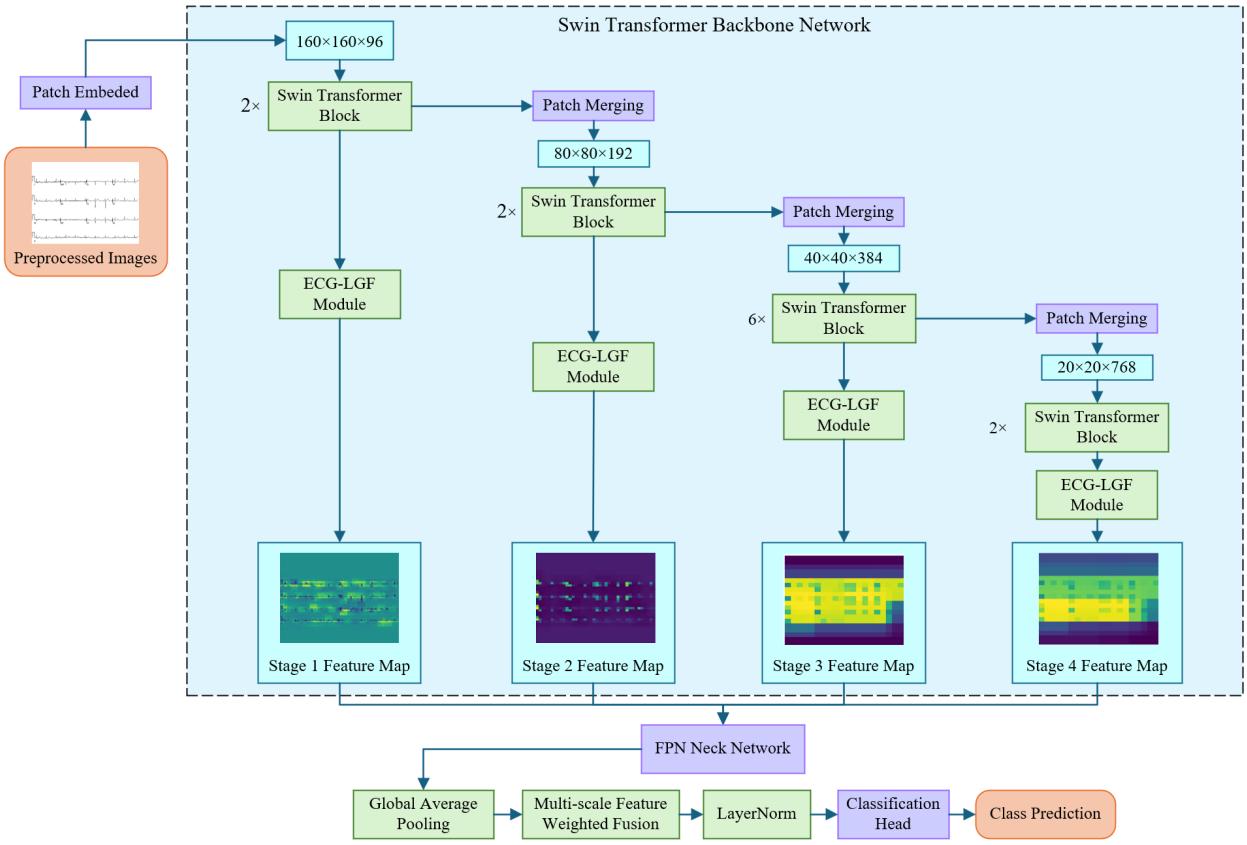


Figure 2. Structure of the proposed Swin-LGF-FPN model

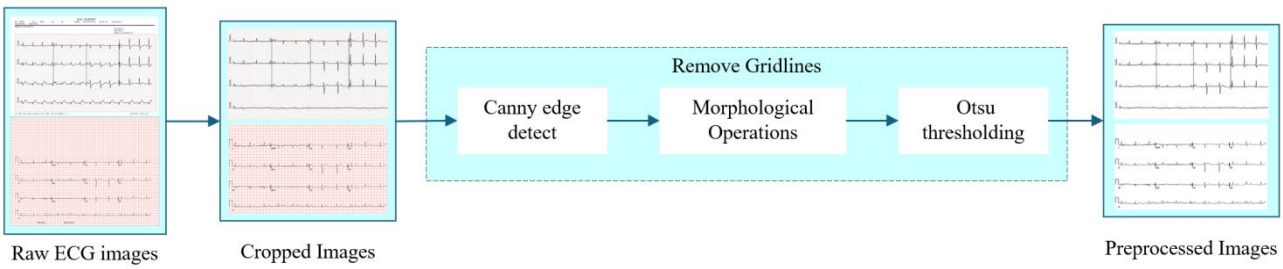


Figure 3. Workflow of Image Preprocessing

2.1 Image Preprocessing

Before model input, raw ECG images were processed through a preprocessing workflow (Figure 3) to remove redundant components, including background grid lines and header/footer annotations. This process mitigated noise interference in feature extraction, prevented feature acquisition bias, and reduced prediction distortion, consequently improving the accuracy of ECG classification [29]. Following vertical-edge cropping, canny edge detection was applied to precisely identify ECG waveform boundaries. This step comprised computing gradient magnitude and orientation via the Sobel operator, with subsequent subpixel edge localization achieved through non-maximum suppression and double-threshold hysteresis. The gradient magnitude G and direction θ are calculated as:

$$G = \sqrt{G_X^2 + G_Y^2} \tag{1}$$

$$\theta = arctan\left(\frac{G_y}{G_x}\right) \tag{2}$$

where G_X^2 and G_Y^2 are the convolution results of the Sobel operator.

A composite morphological operation combining opening and closing was implemented to eliminate minor noise artifacts and fill small cavities in binary edge images, thereby enhancing edge smoothness and continuity. This process utilizes closing to fill internal cavities and opening to remove isolated noise points. The combined operation is mathematically defined in Equation (3).

$$I_{enhanced} = \left[(I_{canny} \oplus B_3) \ominus B_3\right] \ominus B_3 \oplus B_3 \tag{3}$$

where $I_{enhanced}$ is the image after closing-opening operations, I_{canny} is the canny edge detection result, $\oplus$ and $\ominus$ represent dilation and erosion operators, respectively, and B_3 denotes a 3×3 structuring element. Morphological closing executes dilation before erosion, whereas opening performs erosion before dilation.

Otsu's adaptive thresholding method eliminated background interference while separating ECG waveforms from the background. The optimal threshold T^* is automatically determined through inter-class variance maximization, as formulated in Equation (4).

$$\begin{cases} T^* = arg\ max_T[\omega_0(T)\omega_1(T)(\mu_0(T) - \mu_1(T))^2] \\ I_{final}(x,y) = \begin{cases} 1 & I_{enhanced}(x,y) > T^* \\ 0 & otherwise \end{cases} \end{cases} \tag{4}$$

where $\omega_0(T)$ is the proportion of pixels in the background area, $\omega_1(T)$ is the proportion of pixels in the waveform area, $\mu_0(T)$ is the average gray level of the background area, $\mu_1(T)$ is the average gray level of the waveform area, and $I_{final}(x,y)$ is the value of the binarized image at position (x, y).

Computational efficiency and ECG waveform edge feature preservation were balanced by resizing pre-processed ECG images to 640 pixels via image scaling. This standardization optimized input quality for subsequent model training.

2.2 Swin Transformer Backbone Network

Functioning as the backbone network, the optimized Swin Transformer balanced computational efficiency and global modeling capability using Window Multi-Head Self-Attention (*W-MSA*) and Shifted Window Multi-Head Self-Attention (*SW-MSA*). Hierarchical feature maps were generated through a four-stage downsampling, with an enhanced Local-Global Feature Fusion Module (ECG-LGF Module) integrated at each stage terminus to augment ECG waveform feature extraction.

Following Patch Embedding, preprocessed ECG images were processed through the four-stage Swin Transformer backbone. Three critical operations enabled hierarchical feature extraction:

(1) *Intra-stage computation*: Multiple consecutive Swin Transformer blocks per stage facilitated cross-window global context modeling with reduced computational complexity. Figure 4 illustrates the core architecture, where *W-MSA* and *SW-MSA* [32] underwent alternate computation governed by Equations (5)-(8), where F_{i-2} denotes the input features to the Swin Transformer block, and F_{i+2} represents the output features from the block.

$$F_{i-1} = W - MSA(LN(F_{i-2})) + F_{i-2} \quad (5)$$

$$F_i = MLP(LN(F_{i-1})) + F_{i-1} \quad (6)$$

$$F_{i+1} = SW - MSA(LN(F_i)) + F_i \quad (7)$$

$$F_{i+2} = MLP(LN(F_{i+1})) + F_{i+1} \quad (8)$$

(2) *Inter-stage downsampling*: A Patch Merging layer was employed to connect consecutive stages, halving spatial resolution while doubling channel depth to generate downsampled features for the next stage.

(3) *Stage-terminal enhancement*: An ECG-LGF Module was integrated following each stage, taking the Swin Transformer block's output features as input to fuse local features with global rhythm patterns.

2.3 ECG-LGF Module

To enhance the model's synergistic perception of both local morphological details in ECG waveforms and global rhythm patterns, an ECG-LGF Module incorporating Axial Temporal Attention (ATA) for ECG signals was proposed, with its detailed architecture depicted in Figure 5. This module consisted of two parallel pathways defined by Equation (9) and Equation (10):

(a) Local pathway: Depthwise separable convolution ($DWConv_{3\times3}$) extracted local features, succeeded by 1×1 convolution for channel adjustment, GroupNorm, and GELU activation.

$$Local(x) = GELU(GroupNorm(Conv_{1\times1}(DWConv_{3\times3}(x)))) \quad (9)$$

(b) Global pathway: Global max pooling captured image-level features, upsampled to original spatial dimensions through 1×1 convolution and GELU activation.

$$Global(x) = Upsample(GELU(Conv_{1\times1}(Global\_Max\_Pool(x)))) \quad (10)$$

Channel-wise concatenation of dual-path output features was performed according to Equation (11):

$$f = Concat(Local(x), Global(x)) \quad (11)$$

The fused features underwent channel compression and normalization via a 1×1 convolution and GroupNorm, as mathematically defined in Equation (12):

$$f_1 = (GroupNorm(Conv_{1\times1}(f))) \quad (12)$$

The ATA mechanism augmented temporal features in key waveform regions along the ECG image's horizontal axis using adaptive weighting. Spatial weight mappings were acquired by two successive 1×1 convolutional layers separated by GELU activation. The axial temporal attention weight map is generated through a Sigmoid function (σ), which modulates the feature tensor f_1 by element-wise multiplication ($\odot$):

$$Axial\_Attn = f_1 \odot \sigma(Conv_{1\times1}(GELU(Conv_{1\times1}(f_1)))) \quad (13)$$

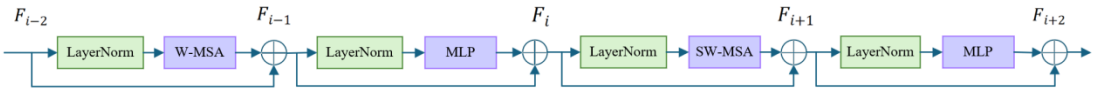


Figure 4. Structure of Swin Transformer Block.

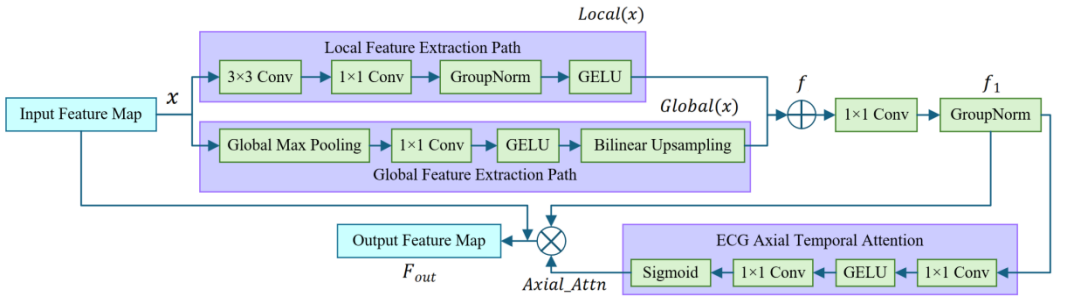


Figure 5. Structure of ECG-LGF Module.

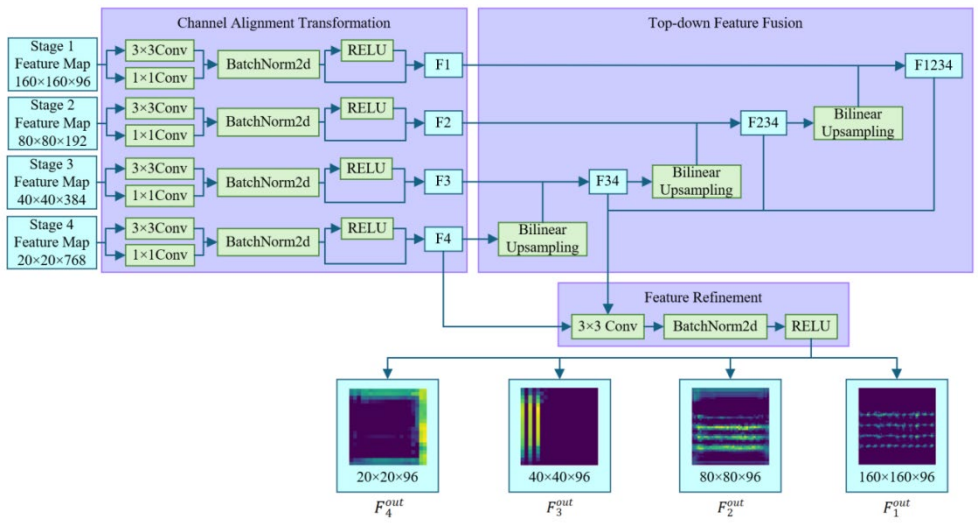


Figure 6. Structure of FPN Neck Network.

The ATA-weighted features were combined with the module's original input via residual connection, yielding enhanced output features as formulated in Equation (14):

$$F_{out} = Axial\_Attn + x \tag{14}$$

Integrating local morphological details, global rhythm patterns, and the ATA mechanism, the ECG-LGF Module produced optimized hierarchical feature maps for the backbone network. These maps are transferred directly to the FPN neck, maintaining original information while improving discriminative feature extraction of key waveforms. The enhanced feature maps delivered spatially and semantically expressive representations to support subsequent ECG image classification.

2.4 FPN Neck Network

A Feature Pyramid Network (FPN) integrated multi-level features optimized by the ECG-LGF Module from the backbone, transmitting deep-layer semantic information to shallow high-resolution features. This constructed a feature pyramid with complementary spatial resolution and semantic content, as illustrated in Figure 6.

(1) *Channel alignment transformation*: A dual-path residual architecture was employed. The primary path utilized 3×3 convolution for simultaneous spatial feature extraction and channel dimension transformation, while the residual path achieved low-complexity channel alignment via 1×1 convolution. Both paths fused outputs through element-wise addition, mitigating gradient vanishing while enhancing information integrity through original feature preservation. For each hierarchy level *i*, the channel alignment transformation followed Equation (15):

$$F_i = ReLU(BN(Conv_{3\times3}(F_i^{in}))) + BN(Conv_{1\times1}(F_i^{in})) \tag{15}$$

Where F_i denotes the transformed output, F_i^{in} denotes the input feature map, and BN denotes Batch Normalization.

(2) *Top-down feature fusion*: Utilizing aligned multi-scale features, this module is initiated from the highly abstract top-level features (F_4). It upsampled features to match adjacent lower-level spatial dimensions via bilinear interpolation (Equation (16)), then fused the upsampled results with current-level aligned features (F_3) through element-wise addition. The fused features (F_{34}) iteratively propagated downward for subsequent layer-wise integration.

$$U(F_i) = \sum_{h'}\sum_{\omega'} F[c,h',\omega'] \cdot \beta(h-\alpha_h h') \cdot \beta(\omega-\alpha_\omega \omega') \tag{16}$$

Where c is the number of feature channels, h' is the time axis coordinate of the feature map, and ω is the lead axis

coordinate of the feature map, α_h is the time axis scaling ratio, and α_ω is the lead axis scaling ratio.

The top-down fusion module produced three hierarchical outputs (Equations (17)-(19)). High-level features encapsulated global rhythm characteristics, while low-level features preserved local waveform morphology. Their element-wise summation intrinsically combined rhythm and morphological information across scales.

$$F_{34} = U(F_4) + F_3 \quad (17)$$
$$F_{234} = U(F_{34}) + F_2 \quad (18)$$
$$F_{1234} = U(F_{234}) + F_1 \quad (19)$$

(3) *Feature refinement module*: This module suppressed discontinuous artifacts from direct cross-scale fusion by optimizing the spatial smoothness of fused features. The refined output produced the feature pyramid's final feature map F_i^{out} as defined in Equation (20)-(23).

$$F_4^{out} = ReLU(Conv_{3\times3}(BN(F_4))) \quad (20)$$
$$F_3^{out} = ReLU(Conv_{3\times3}(BN(U(F_4) + F_3))) \quad (21)$$
$$F_2^{out} = ReLU(Conv_{3\times3}(BN(U(F_{34}) + F_2))) \quad (22)$$
$$F_1^{out} = ReLU(Conv_{3\times3}(BN(U(F_{234}) + F_1))) \quad (23)$$

These refined feature maps preserved critical morphological details and encoded global rhythm semantics, constituting a hierarchical pathological representation. Cross-scale fusion amplified pathological signatures across resolutions, enhancing adaptability to lesion size variations while delivering multi-semantic features for classification.

Multi-scale feature maps extracted from the FPN were first processed by global average pooling and flattening. Their contributions were then balanced through a weighted fusion mechanism employing learnable weights, followed by layer normalization. The resulting normalized features were finally fed into a two-layer multilayer perceptron (MLP) for classification:

$$y_{out} = W_2 \cdot Dropout(\Gamma_{LN}(GELU(W_1 x + b_1))) + b_2 \quad (24)$$

Where x is the weighted fused feature vector, W is the weight matrix, and b represents the bias vector, Γ_{LN} denotes the Layer Normalization operation, Dropout stands for the operation of randomly dropping neurons during training, and y_{out} represents the final classification prediction score.

2.5 Training process optimization

During model training, Focal Loss, an AdamW optimizer coupled with a linear Warm-up Cosine Annealing scheduler, and translation augmentation were employed to mitigate class imbalance and overfitting risks while enhancing model accuracy.

Focal Loss modified the standard Cross-Entropy loss [33] by down-weighting the loss contributions from easy-to-classify samples and focusing training on hard examples, which helped mitigate class imbalance in training data as defined in Equation (25).

$$Focal\_Loss(p_t) = -\alpha_t(1 - p_t)^\gamma \, log(p_t) \quad (25)$$

where p_t is the prediction probability of the model for the correct class, γ is the adjustable focusing parameter, and α_t is the class balance weight.

Throughout the training process, AdamW utilized decoupled weight decay to enhance regularization in parameter updates, countering overfitting [34]. Equation (26) mathematically defines this training-specific mechanism:

$$\theta_t = \theta_{t-1} - \eta\left(\frac{\widehat{m_t}}{\sqrt{\widehat{v_t}} + \epsilon} + \lambda\theta_{t-1}\right) \quad (26)$$

where the weight decay coefficient λ=0.05, the initial learning rate η=3e-5, θ_t denotes the model parameter at step t, ϵ is a numerical stability constant, and $\widehat{m_t}$, $\widehat{v_t}$ represent bias-corrected first-moment and second-moment estimates, respectively.

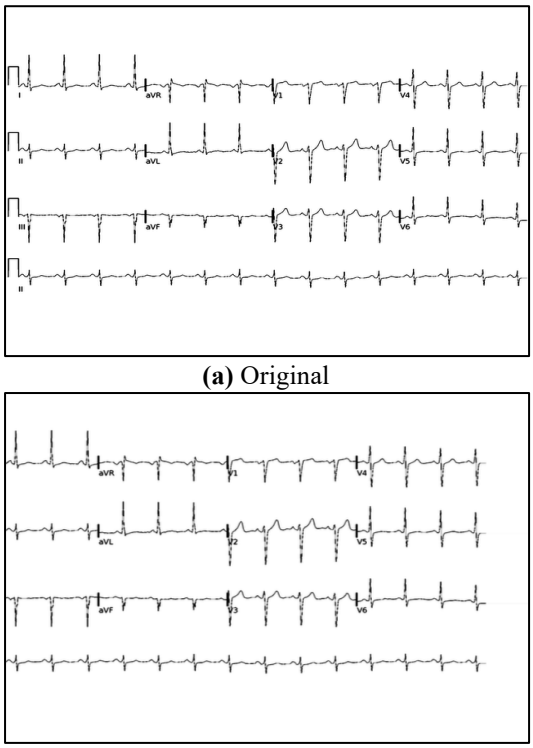
The learning rate was dynamically adjusted during training via a Cosine Annealing scheduler with linear warmup (Equation (27)):

$$\eta_t = \begin{cases} \eta_{min} + \frac{1}{2}(\eta - \eta_{min})\frac{t}{T_{warm}} & t < T_{warm} \\ \eta_{min} + \frac{1}{2}(\eta - \eta_{min})\left[1 + \cos\left(\frac{T_{warm}}{T_{max}}\right)\right] & else \end{cases} \quad (27)$$

where the minimum learning rate η_{min} is set to 0.001 times the initial learning rate, the warmup period T_{warm} is 30 epochs, and the cosine decay period T_{max} accounts for 70% of the total epochs. Empirical results demonstrate that AdamW with this scheduler reduced overfitting while maintaining stable convergence.

This study employed translation augmentation to enhance the ECG image dataset. This approach boosted model generalization, reduced overfitting risks, and preserved critical waveform attributes (morphology, amplitude, duration). During training, images were randomly translated along temporal and voltage axes by

≤10% of image dimensions, simulating natural variations such as temporal phase shifts and electrode-induced baseline drift. Original ECG images are shown in Figure 7a, while the augmented results are displayed in Figure 7b.



(a) Original

(b) After translation augmentation

Figure 7. Original and translated enhanced images

3. Experiments and Results

3.1 Datasets and Evaluation Metrics

To validate the ECG classification performance and feature fusion efficacy of the Swin-LGF-FPN model, two independent public datasets were utilized: the ECG Images Dataset of Cardiac Patients and the PTB-XL dataset. Their key characteristics were detailed in Table 1.

The first dataset, the ECG Images dataset of Cardiac Patients, was a public repository curated by the Ch. Pervaiz Elahi Institute of Cardiology, Multan, Pakistan [35]. It comprised 928 patient records spanning four diagnostic categories, featuring 12-lead ECG images acquired via tele-health diagnostic tools. All images exceeded 800 KB in file size, with representative samples illustrated in Figure 8.

The second dataset, PTB-XL, was a large-scale public ECG database that contained 21,837 clinical 12-lead recordings from 18,885 patients, each spanning 10 seconds [36]. Its multi-label coexistence and broad age/gender distribution reflected real-world complexity, although the original data were in digital signal format. To ensure compatibility with image-based models, signals were converted into a standardized image format (matching the resolution and lead layout of the ECG Images Dataset of Cardiac Patients) using ECG-Image-Kit [37] (Figure 9). PTB-XL employed super-class annotation (5 labels) for fair benchmarking.

Figure 8. Sample images from the ECG Images dataset of Cardiac Patients

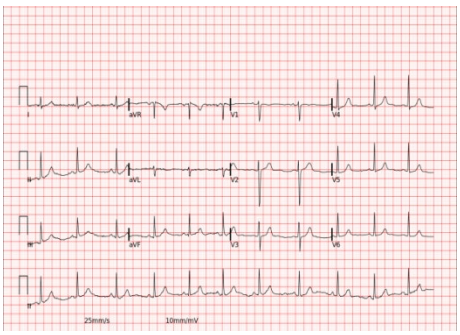


Figure 9. Sample images converted from the PTB-XL dataset

Table 1. Summary statistics of the two datasets

Dataset	Number of Classes	Class Name	Total Samples

ECG Images Dataset of Cardiac Patients	4	Abnormal Heartbeats, Myocardial Infarction, History of Myocardial Infarction, Normal Heartbeats	928
PTB-XL	5	Normal, Myocardial Infarction, ST-T Changing, Conduction Disturbance, Hypertrophy	21837

Both datasets underwent stratified random sampling without replacement: 20% per class was allocated to validation sets, while the remaining 80% was used for training. Fixed random seeds ensured reproducibility of this split.

Comprehensive evaluation of Swin-LGF-FPN on the ECG Images Dataset of Cardiac Patients and PTB-XL datasets utilizes standard classification metrics: Overall Accuracy, Specificity, Recall, Precision, and F1 Score. Mathematical definitions are given by Equation (28)-(32):

$$Overall\ accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (28)$$

$$Specificity = \frac{TN}{TN + FP} \quad (29)$$

$$Recall = \frac{TP}{TP + FN} \quad (30)$$

$$Precision = \frac{TP}{TP + FP} \quad (31)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (32)$$

where *TP*, *TN*, *FP*, and *FN* denote True Positives, True Negatives, False Positives, and False Negatives, respectively.

As complementary to standard metrics, Receiver Operating Characteristic (ROC) curves assess model classification performance and generalization capability. These curves graphically represent the trade-off between True Positive Rate (*TPR*) and False Positive Rate (*FPR*) at varying classification thresholds. Through continuous adjustment of classification thresholds τ, each threshold-specific point ($TPR(\tau)$, $FPR(\tau)$) is computed via Equations (33)-(34). Sequentially connecting these points generates curves that visually validate the model's discriminative capacity for distinguishing positive and negative samples [38].

$$TPR(\tau) = \frac{\sum_{i\epsilon\ positive\ samples} I\,(s_i \geq \tau)}{TP + FN} \quad (33)$$

$$FPR(\tau) = \frac{\sum_{j\epsilon\ negative\ samples} I\,(s_j \geq \tau)}{TN + FP} \quad (34)$$

The core evaluation metric for ROC curves is the Area Under the Curve (AUC), representing the area beneath the ROC curve, defined by Equation (35).

$$AUC = \int_0^1 TPR(FPR)\, dFPR \quad (35)$$

3.2 Experimental Setup

Experimental hardware comprised an NVIDIA GeForce RTX 4070 Ti GPU (16GB VRAM) paired with an Intel Core i5-14600KF processor. Software environments utilized Python 3.12 and the PyTorch framework. During training, each epoch updated model parameters based on the training set, with performance evaluated on the validation set. A unified batch size of 16 was employed across all datasets to optimize the trade-off between computational efficiency and model performance. It is known that smaller batches increase parameter update frequency, which accelerates convergence but extends training duration, while larger batches demand greater GPU memory resources. Iterative testing confirmed that a batch size of 16 maintained an optimal balance between training efficiency and model effectiveness.

To prevent overfitting and enhance generalization, an early stopping criterion was implemented to terminate training if the validation loss failed to decrease for 15 consecutive epochs. The embedding dimension was set to 96 to fully characterize ECG signal details and higher-order features. Preliminary experiments had established 96 as the optimal value for preserving critical features while controlling model complexity. To accelerate computation, enhance training stability, and reduce memory footprint, the proposed model adopted mixed-precision training (FP16) with L2-norm gradient clipping at a maximum threshold of 0.5, thereby preventing gradient explosion.

3.3 Performance comparison of ECG image classification

To validate the superior ECG image classification performance of Swin-LGF-FPN and assess feature fusion efficacy, we benchmarked against established state-of-the-art image classification models: ResNet-34, Vision Transformer (ViT), MobileViT, and ConvNeXt. As evidenced in Table 2 and Table 3, the proposed model

consistently significantly surpassed all baselines across every metric on both datasets.

The proposed model achieved an overall accuracy of 0.9945 and an F1-score of 0.9945 in Table 2, with specificity reaching 0.9983. In Table 3, notwithstanding class imbalance in PTB-XL, the model maintained robust generalization capability and high performance levels. To visually demonstrate cross-model performance disparities, Figures 10 and 11 compare overall accuracy and F1-scores across models on both datasets.

Table 2. Comparison with baseline methods on the ECG Images dataset of the Cardiac Patients dataset

Method	Overall accuracy	Specificity	Recall	Precision	F1-Score
Resnet-34	0.9672	0.9889	0.9607	0.9679	0.9669
Vision-Transformer	0.9727	0.9907	0.9652	0.9726	0.9722
MobileViT	0.9781	0.9925	0.9781	0.9783	0.9781
Swin Transformer	0.9891	0.9966	0.9891	0.9897	0.9890
ConvNeXt	0.9891	0.9964	0.9891	0.9893	0.9891
Our Model	**0.9945**	**0.9983**	**0.9945**	**0.9947**	**0.9945**

Table 3. Comparison with baseline methods on the PTB-XL dataset

Method	Overall accuracy	Specificity	Recall	Precision	F1-Score
Resnet-34	0.7698	0.9250	0.7698	0.7717	0.7604
Vision Transformer	0.7713	0.9257	0.7713	0.7690	0.7630
MobileViT	0.7727	0.9286	0.7727	0.7700	0.7676
Swin Transformer	0.7779	0.9334	0.7779	0.7804	0.7764
ConvNeXt	0.7765	0.9313	0.7765	0.7752	0.7733
Our Model	**0.7894**	**0.9336**	**0.7894**	**0.7840**	**0.7842**

Table 4. Ablation study of different components on the PTB-XL dataset

Method	Overall accuracy	Specificity	Recall	Precision	F1-Score
Backbone only	0.7779	0.9334	0.7779	0.7804	0.7764
without ECG-LGF Module	0.7843	0.9326	0.7843	0.7788	0.7797
without FPN Neck Network	0.7819	0.9267	0.7819	0.7761	0.7694
without Feature Refinement	0.7860	0.9314	0.7860	0.7829	0.7804
Proposed Method	**0.7894**	**0.9336**	**0.7894**	**0.7840**	**0.7842**

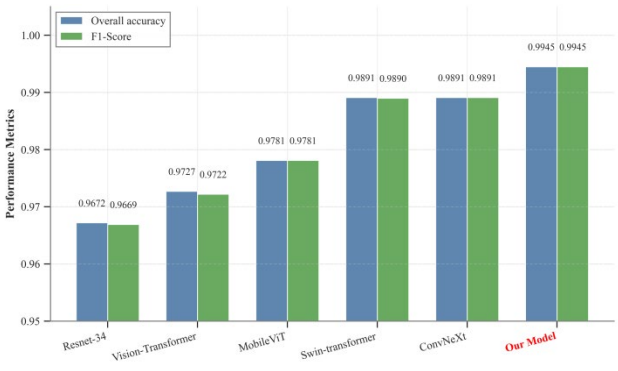


Figure 10. Model performance comparison (Accuracy & F1) on the ECG Images dataset of Cardiac Patients

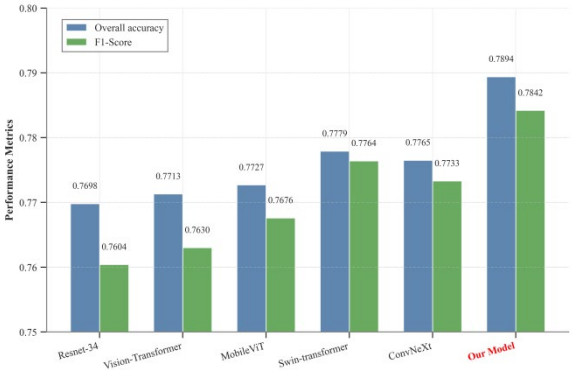


Figure 11. Model performance comparison (Accuracy & F1) on PTB-XL dataset

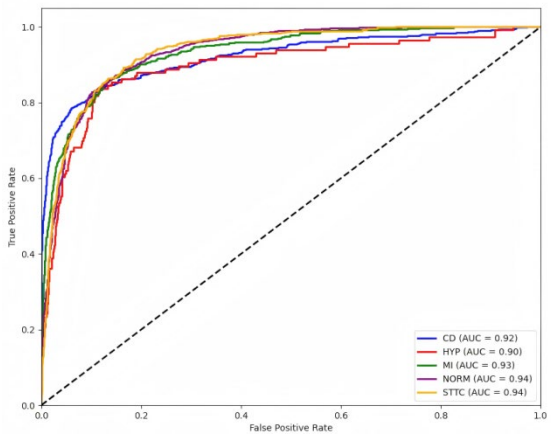


Figure 12. ROC curve of the proposed method on PTB-XL

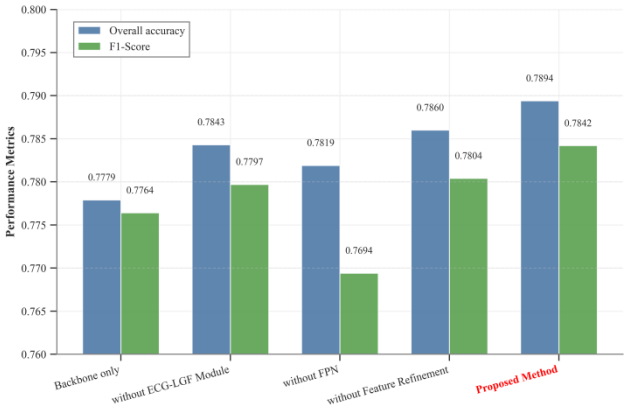


Figure 13. Ablation model performance comparison (Accuracy & F1) on the PTB dataset

To validate generalization on complex ECG data, the ROC curves on PTB-XL (Figure 12) showed that AUC values were stably distributed between 0.90 and 0.94. The consistent curve morphology indicated a balanced discriminative capacity across pathologies, confirming sustained diagnostic robustness in heterogeneous ECG data. The performance variations between the PTB-XL dataset and the ECG Images Dataset of Cardiac Patients objectively reflected real-world clinical challenges, such as diagnostic heterogeneity, signal noise, and acquisition-device discrepancies.

Ablation studies validated the necessity of each component by sequentially removing the ECG-LGF Module, FPN Neck Network, and Feature Refinement Module, using the Swin Transformer backbone-only as the baseline reference (Table 4). Considering the limited data volume of the ECG Images Dataset of Cardiac Patients (928 images), which was insufficient for thoroughly validating model components in complex diagnostic environments, the PTB-XL dataset was selected as the primary ablation platform due to its advantages in clinical complexity and annotation scale. Figure 13 visually compares performance disparities (overall accuracy and F1-score) among the ablation variants on PTB-XL.

The results demonstrated that the full model (Proposed Method) achieved optimal metrics. Removing any component degraded performance, confirming the architectural integrity and effectiveness. Although all ablation variants surpassed the backbone-only baseline in overall accuracy, the variant without the FPN Neck Network exhibited a 0.7% F1-score reduction versus the backbone and a 1.48% decline relative to the full model, representing the most significant performance drop. This

evidenced the critical role of FPN in multi-scale feature integration.

3.4 Grad-CAM-Enhanced Decision Interpretability

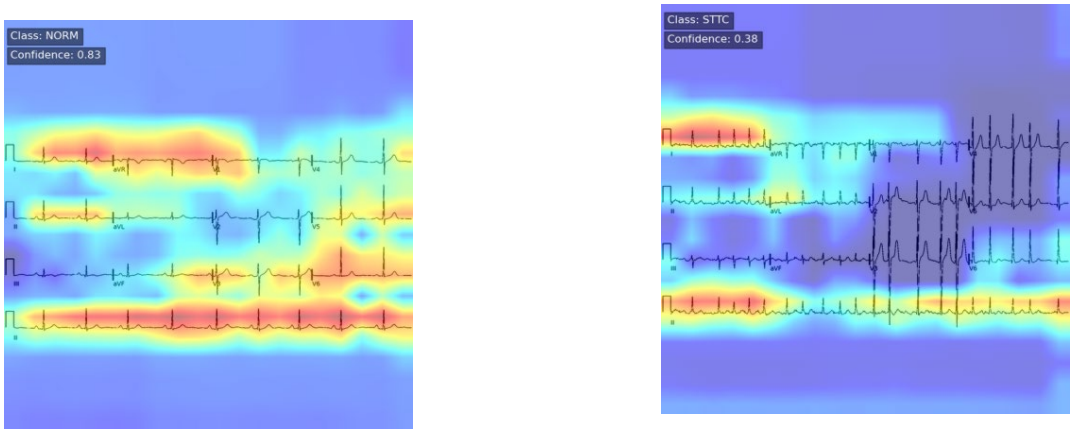
To enhance model interpretability and clinical credibility, this study employed Grad-CAM, a gradient-based visualization technique [39]. The computation followed Equation (36) and Equation (37):

$$\alpha_k^c = \frac{1}{z}\sum_i \sum_j \frac{\partial y_{out}}{\partial A_{ij}^k} \tag{36}$$

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A_{ij}^k) \tag{37}$$

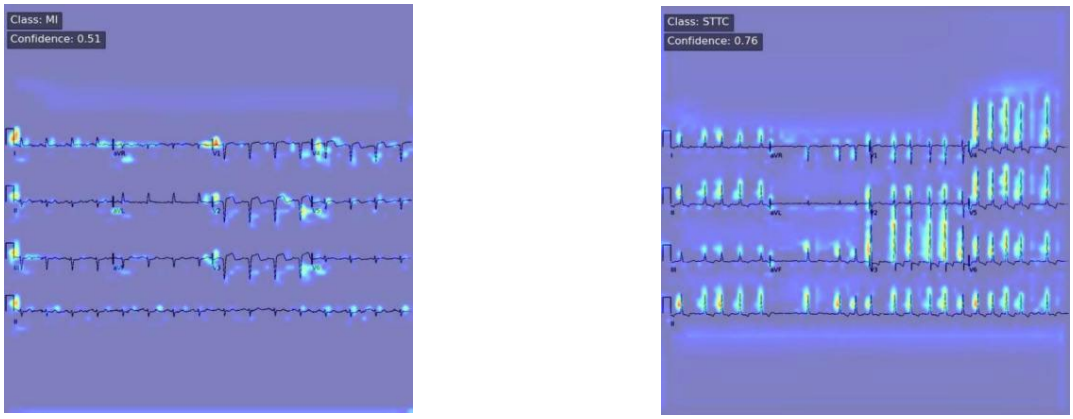
where y_{out} is the prediction score of the target class c, A_{ij}^k is the k-th feature map, α_k^c is the importance weight of feature map k for class c, and $L_{Grad-CAM}^c$ is the class activation map with original resolution.

Grad-CAM localized model decision-critical regions by computing target-class gradients relative to final convolutional features, generating weight coefficients through global average pooling, linearly combining weighted feature maps with *ReLU* activation, and up-sampling to input resolution. Figure 14 visualizes Swin Transformer's attention overlays, contrasting with our proposed model's results in Figure 15. Red/yellow regions indicated high model attention during decisions, while blue/purple areas denoted low attention. Although Swin Transformer captured overall waveform trends, its global attention mechanism induced gradient smoothing effects that dispersed attention and blurred critical node localization, weakening pathological focus. Conversely, our model demonstrated enhanced focus specificity.



(**a**) Normal (**b**) ST-T Changing

Figure 14. Examples of attention overlay visualization generated by Swin Transformer.



(**a**) Myocardial Infarction (**b**) ST-T Changing

Figure 15. Examples of attention overlay visualization generated by Swin-LGF-FPN.

4. Conclusions

(1) This work proposed the Swin-LGF-FPN, a Swin Transformer-based architecture for ECG image classification. The backbone network extracted multi-scale features from preprocessed ECG images, which were enhanced by the ECG-LGF Module to refine waveform-specific feature representation. A Feature Pyramid Networks (FPN) served as the neck network, fusing multi-scale features to preserve local morphology and capture axial-temporal dependencies through global context modeling, thus enabling intelligent ECG classification.

During training, Focal Loss was employed to mitigate class imbalance, while the AdamW optimizer with a Warmup Cosine Annealing scheduler was used to ensure stable convergence and prevent overfitting. Translation augmentation of ECG images was applied to enhance model generalization.

Evaluations on public ECG datasets (the ECG Images Dataset of Cardiac Patients and PTB-XL) showed that the proposed model achieved efficient classification with superior performance across all key metrics compared to baseline methods. ROC curves on PTB-XL confirmed robust generalization under complex data distributions, while ablation studies validated the essential contributions of each architectural component. Grad-CAM visualizations demonstrated that, in contrast to the dispersed attention patterns of baseline models, our model exhibited a concentrated focus on pathological regions within ECG waveforms, enhancing interpretability.

The findings suggest that this model has the potential to transform ECG diagnostics through data-driven intelligence, providing clinical decision support that could enable population screening, rapid triage, and timely referral. By shifting from experience-based to data-driven interpretation paradigms, it may enhance early detection and intervention for cardiovascular diseases. Given the current performance limitations of image-based ECG classification on complex multi-label datasets such as PTB-XL, we emphasize that it should be designated as a clinical adjunct tool rather than a diagnostic replacement.

Author Contribution

Conceptualization, Bin Xu and Shoucheng Ji; Methodology, Zongzhen Yue; Validation, Bin Xu, Zongzhen Yue; Writing—Original Draft Preparation, Zongzhen Yue; Writing—Review and Editing, Bin Xu and Shoucheng Ji; Visualization, Ning Sun; Supervision, Jianyong Zheng. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

Funding

This paper was sponsored by the National Natural Science Foundation of China (No.51675345); Bio-medical science and technology support project of Shanghai Science and Technology Commission (20S31905500); Public welfare projects of the State Administration for Market Regulation (2020MK030, 2024MK030); Pudong District Science and Technology Development Fund-Industry-Academia-Research Special Project (Future Vehicles, PKX2024-W06).

Ethics Statement

This study did not involve an ethical problem, and all experiments provided written informed consent.

Conflict of interest statement

The authors declare that they have no conflict of interest.

References

[1] Viliani D, Cecconi A, López-Melgar B, Muñiz ÁM, Martínez-Vives P, Cuenca S, et al. Machine-learning computer-assisted ECG analysis to predict myocardial fibrosis in patients with hypertrophic cardiomyopathy. Journal of Electrocardiology 2025; Volume 90, 153892. https://doi.org/10.1016/j.jelectrocard.2025.153892.

[2] Swathy M, Saruladha K. A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques. ICT Express 2022; Volume 8, Issue 1, Pages 109-116. https://doi.org/10.1016/j.icte.2021.08.021.

[3] Li X, Zhang D, Li X, Gao X, Liang Y, Tse G, et al. Exploring artificial intelligence methods for cardiac syncope diagnosis combined with electrocardiogram parameters and clinical characteristics. Journal of Electrocardiology 2025; Volume 91, 154018. https://doi.org/10.1016/j.jelectrocard.2025.154018.

[4] Palermi S, Vecchiato M, Ng F, Attia Z, Cho Y, Anselmino M, et al. Artificial intelligence and the electrocardiogram: A modern renaissance. European Journal of Internal Medicine 2025; https://doi.org/10.1016/j.ejim.2025. 04.036.

[5] Seoni S, Molinari F, Acharya U.R, Lih OS, Barua PD, García S, et al. Application of spatial uncertainty predictor in CNN-BiLSTM model using coronary artery disease ECG signals. Information Sciences 2024; Volume 665, 120383. https://doi.org/10.1016/ j.ins.2024.120383.

[6] Al-Zaiti S, Martin-Gill C, Zègre-Hemsey JK., Bouzid Z, Faramand Z, Alrawashdeh MO., et al. Machine learning for ECG diagnosis and risk stratification of occlusive myocardial infarction. Nature Medicine 2023; 29, 1804–1813. https://doi.org/10.1038/ s41591-023-02396-3.

[7] Xiao Q, Lee K, Mokhtar SA, Ismail I, Pauzi ALM, Zhang Q, et al. Deep Learning-Based ECG Arrhythmia Classification: A Systematic Review. Applied Sciences 2023; 13, no.8:4964. https://doi.org/10.3390/ app13084964.

[8] Hannun AY., Rajpurkar P, Haghpanahi M, Tison GH., Bourn C, Turakhia MP., et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine 2019; 25 (1) 65-69. https://doi.org/10.1038/s41591-018-0268-3.

[9] Islam MS, Kalmady SV, Hindle A, Sandhu R, Sun W, Sepehrvand N, et al. Diagnostic and Prognostic Electrocardiogram-Based Models for Rapid Clinical Applications. Canadian Journal of Cardiology 2024; Volume

40, Issue 10, Pages 1788-1803. https://doi.org/10.1016/j.cjca.2024.07.003.
[10] Al-Zaiti S, Besomi L, Bouzid Z, Faramand Z, Frisch S, Martin-Gill C, et al. Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. Nature Communications 2020; 11, 3966. https://doi.org/10.1038/s41467-020-17804-2.
[11] Kolliyil JJ, Brindise MC. Automated detection of arrhythmias using a novel interpretable feature set extracted from 12-lead electrocardiogram. Computers in Biology and Medicine 2025; Volume 189, 109957. https://doi.org/10.1016/j.compbiomed.2025.109957.
[12] Liu J, Liu Y, Jin Y, Li Z, Qin C, Chen X, et al. A novel diagnosis method combined dual-channel SE-ResNet with expert features for inter-patient heartbeat classification. Medical Engineering & Physics 2024; Volume 130, 104209. https://doi.org/10.1016/j.medengphy.2024. 104209.
[13] Kraft D, Bieber G, Jokisch P, Rumm P. End-to-End Premature Ventricular Contraction Detection Using Deep Neural Networks. Sensors 2023; 23, 8573. https://doi.org/10.3390/s23208573.
[14] Wu L, Xie X, Wang Y, ECG Enhancement and R-Peak Detection Based on Window Variability. Healthcare 9 2021; no. 2: 227. https://doi.org/10.3390/healthcare 9020227.
[15] Katamreddi H.K.P, Battula TK. A hybrid approach for machine learning based beat classification of ECG using different digital differentiators and DTCWT. Computers in Biology and Medicine 2025; Volume 194, 110426. https://doi.org/10.1016/j.compbiomed.2025.110426.
[16] Abdel-Rahman S, Antiperovitch P, Tang A, Daoud MI., Parsa V, Lacefield JC. Faster R-CNN approach for estimating global QRS duration in electrocardiograms with a limited quantity of annotated data. Computers in Biology and Medicine 2025; Volume 192, Part A, 110200. https://doi.org/10.1016/j.compbiomed. 2025.110200.
[17] Mantravadi A, Saini S, Teja R. SC, Mittal S, Shah S, Devi R. S, et al. CLINet: A novel deep learning network for ECG signal classification. Journal of Electrocardiology 2024; Volume 83, Pages 41-48. https://doi.org/10.1016/j.jelectrocard. 2024.01.004.
[18] Fan L, Chen B, Zeng X, Zhou J, Zhang X. Knowledge-enhanced meta-transfer learning for few-shot ECG signal classification. Expert Systems with Applications 2025; Volume 263, 125764. https://doi.org/10.1016/j.eswa.2024. 125764.
[19] Ahmad Z, Tabassum A, Guan L, Khan NM. ECG heartbeat classification using multimodal fusion. IEEE Access, 2021, 9: 100615-100626. https://doi.org/10.1109/ACCESS.2021.3097614.
[20] Weimann K, Conrad TO.F.. Transfer learning for ECG classification. Sci Rep 2021; 11, 5251. https://doi.org/10.1038/s41598-021-84374-8.
[21] Zhou F, Fang D. Classification of multi-lead ECG based on multiple scales and hierarchical feature convolutional neural networks. Sci Rep 2025; 15, 16418. https://doi.org/10.1038/s41598-025-94127-6.
[22] Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng in Med and Biol 20(3): 45-50.
[23] Liu Z, Cao Q, Jin Q, Lin J, Lv G, Chen K. Accurate detection of arrhythmias on raw electrocardiogram images: An aggregation attention multi-label model for diagnostic assistance. Medical Engineering & Physics 2023; Volume 114, 103964, https://doi.org/10.1016/j.medengphy.2023. 103964.
[24] Jothiaruna N, Anny Leema A, SSDMNV2-FPN: A cardiac disorder classification from 12 lead ECG images using deep neural network. Microprocessors and Microsystems 2022; Volume 93,104627. https://doi.org/10.1016/j.micpro. 2022.104627.
[25] Demolder A, Kresnakova V, Hojcka M, Boza V, Iring A, Rafajdus A, et al. High precision ECG digitization using artificial intelligence[J], Journal of Electrocardiology, Volume 90, 2025, 153900. https://doi.org/10.1016/j.jelectrocard.2025.153900.
[26] Sadad T, Safran M, Khan I, Alfarhood S, Khan R, Ashraf I. Efficient Classification of ECG Images Using a Lightweight CNN with Attention Module and IoT. Sensors 2023; 23, no. 18: 7697. https://doi.org/10.3390/ s23187697.
[27] Hao P, Gao X, Li Z, Zhang J, Wu F, Bai C. Multi-branch fusion network for Myocardial infarction screening from 12-lead ECG images. Computer Methods and Programs in Biomedicine 2020; Volume 184, 105286. https://doi.org/10.1016/j.cmpb.2019.105286.
[28] Cao Q, Du N, Yu L, Zuo M, Lin J, Liu N, et al. Practical fine-grained learning based anomaly classification for ECG image. Artificial Intelligence in Medicine 2021; Volume 119, 102130. https://doi.org/10.1016/j.artmed. 2021.102130.
[29] Fatema K, Montaha S, Rony M. AH, Azam S, Hasan M. Z, Jonkman M. A Robust Framework Combining Image Processing and Deep Learning Hybrid Model to Classify Cardiovascular Diseases Using a Limited Number of Paper-Based Complex ECG Images. Biomedicines 2022; 10, 2835. https://doi.org/10.3390/biomedicines10112835.
[30] Khalid M, Pluempitiwiriyawej C, Abdulkadhem AA. , Afzal I, Truong T, ECGConVT: A Hybrid CNN and Vision Transformer Model for Enhanced 12-Lead ECG Images Classification. IEEE Access 2024; vol. 12, pp. 193043-193056. https://doi.org/10.1109/ACCESS.2024.3516495.
[31] Ma L, Zhang F. A Novel Real-Time Detection and Classification Method for ECG Signal Images Based on Deep Learning Sensors 2024; 24, no. 16: 5087. https://doi.org/10.3390/s24165087.
[32] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) 2021; pp. 9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986.
[33] Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal Loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017; pp. 2980-2988.
[34] AbuKaraki A, Alrawashdeh T, Abusaleh S, Alksasbeh MZ, Alqudah B, Alemerien K, et al. Pulmonary Edema and Pleural Effusion Detection Using EfficientNet-V1-B4 Architecture and AdamW Optimizer from Chest X-Rays Images. Computers, Materials and Continua 2024; Volume 80, Issue 1, Pages 1055-1073. https://doi.org/ 10.32604/cmc.2024.051420.
[35] Khan AH, Hussain M. ECG Images dataset of Cardiac Patients[dataset], Mendeley Data, V2, 2021. https://doi.org/10.17632/gwbz3fsgp8.2.
[36] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI., Samek W, et al. Ptb-xl, a large publicly available electrocardiography dataset. Sci Data 2020; 7:154. https://doi.org/ 10.1038/s41597-020-0495-6.
[37] Shivashankara KK, Deepanshi, Shervedani AM, Reyna MA., Clifford GD., Sameni R. ECG-image-kit: a synthetic image generation toolbox to facilitate deep learning-based electrocardiogram digitization. In Physiological Measurement 2024; IOP Publishing. https://doi.org/ 10.1088/1361 - 6579/ad4954.
[38] Bianco AM., Boente G. Addressing robust estimation in covariate-specific ROC curves, Econometrics and Statistics 2023; https://doi.org/10.1016/j.ecosta.2023. 04.001.

[39] Selvaraju RR., Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis 2020; 128, 336-359. https://doi.org/10.1007/s11263- 019-01228-7.