Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SpecMFM-Net: A Spectrogram-Enhanced Multi-Scale Feature Modulation Network with Medical-Prior-Guided Transfer Learning for Atrial Fibrillation Detection

Yujia Sun
Northeastern University
Qi Yan
Northeastern University
Haotian Tang
Northeastern University
Yunjia Wu
Northeastern University
Rongrong Fu
Northeastern University
Hanyu Cui
Northeastern University
Xidong Wu
Northeastern University
Liping Xie
xieliping@bmie.neu.edu.cn
Northeastern University

Article

Keywords:

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11471660/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SpecMFM-Net: A Spectrogram-Enhanced Multi-Scale Feature Modulation Network with Medical-Prior-Guided Transfer Learning for Atrial Fibrillation Detection

Yujia Sun[†], Qi Yan[†], Haotian Tang, Yunjia Wu, Rongrong Fu, Hanyu Cui, Xidong Wu, Liping Xie[*]

[1]College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, 110169, China.

*Corresponding author(s). E-mail(s): xieliping@bmie.neu.edu.cn;
[†]These authors contributed equally to this work.

Abstract

Atrial fibrillation (AF) is a prevalent cardiac arrhythmia whose timely detection is critical for improving patient outcomes. Deep learning has achieved remarkable success in data-rich image, video, and text tasks and has also been widely applied to ECG-based AF detection. However, ECG analysis is often constrained by limited high-quality annotations because rhythm-level labeling requires expert interpretation and substantial time cost. Models trained under such annotation-limited conditions are prone to overfitting, resulting in restricted detection performance and weak generalization in practical scenarios. To address this issue, this paper proposes a spectrogram-enhanced multi-scale feature modulation network (SpecMFM-Net) with medical-prior-guided transfer learning for AF detection. Raw ECG signals are first converted into super-resolution spectrograms using the fractional adaptive superlet transform (FASLT), which adaptively balances time-frequency resolution and better captures rhythm evolution and transient AF-related patterns. SpecMFM-Net is then pretrained on a source dataset to acquire domain-relevant medical priors and then fine-tuned on the target data. It further integrates multi-scale feature modulation and BiLSTM temporal modeling to enhance discriminative temporal-spectral representations for AF classification. The SpecMFM-Net achieved an average macro F1 score of 90.04% in 5-fold cross-validation on CinC2017 dataset and 91.54% in cross-dataset testing on AFDB dataset, demonstrating promising precision and generalization for AF detection under limited high-quality annotations.

Keywords: Atrial fibrillation, Deep learning, Transfer Learning, ECG spectrogram, Multi-scale feature modulation

1 Introduction

Atrial fibrillation (AF) is the most common cardiac arrhythmia in clinical practice and has attracted substantial attention because of its strong association with cardioembolic stroke and heart failure[1]. As a global epidemic, the number of patients continues to rise with accelerating population aging. It is an urgent public health priority to implement effective early screening strategies to mitigate complications and improve prognostic outcomes[2]. Electrocardiography (ECG) is a noninvasive technique for recording cardiac electrical activity. It contains rich rhythm and morphological information and is widely used for arrhythmia screening. However, conventional ECG diagnosis relies heavily on manual beat-by-beat interpretation by clinicians, which is time-consuming and susceptible to inter-physician variability and fatigue, thereby increasing the risk of misdiagnosis and missed diagnosis[3]. With the rapidly growing volume of clinical monitoring data, the development of efficient and robust automated methods for AF detection has become an urgent need in this field.

In recent years, automated AF recognition has progressively shifted from handcrafted feature engineering to deep representation learning. CNN-based models are effective in extracting local morphological patterns from ECG signals[4], recurrent models such as LSTM and BiLSTM can describe rhythm evolution over time[5], and Transformer-based models have been explored to strengthen long-range dependency modeling[6]. However, ECG signals are inherently nonstationary, and purely time-domain representations may not sufficiently capture both transient waveform changes and frequency-dependent rhythm variations. Therefore, many studies convert ECG signals into time-frequency representations before classification[7, 8]. Common choices include the short-time Fourier transform (STFT) and the continuous wavelet transform (CWT). Nevertheless, STFT is constrained by a fixed time-frequency resolution, whereas CWT still involves a predefined scale-resolution trade-off. These limitations motivate the use of a more adaptive spectrogram representation that can better preserve AF-related rhythm irregularities and transient morphological patterns.

In addition, the real-world clinical application of AF recognition is still broadly constrained by the limited availability of high-quality annotations. Rhythm-level ECG annotation requires professional expertise and is costly and time-consuming, making it difficult to construct large-scale, reliably labeled target-domain datasets [9]. Models trained with limited high-quality labels are prone to overfitting and often show restricted generalization in practical clinical scenarios. To alleviate performance degradation under annotation-limited conditions, transfer learning has been introduced into ECG tasks to improve learning efficiency in the target domain. Wei et al.[9] attempted to use an ImageNet-pretrained EfficientNet-B0 to extract features from ECG spectrograms. Mavaddati et al.[10] also used an ImageNet-pretrained ResNet-34 for CVD classification. However, despite the performance improvements brought by transfer

learning, the inherent disparity in feature distributions between natural images and medical data frequently constrains the advantages of cross-domain pretraining [11].

Motivated by the above considerations, this paper proposes a spectrogram-enhanced multi-scale feature modulation network (SpecMFM-Net) with medical-prior-guided transfer learning. Instead of relying only on limited high-quality annotations in the target domain, the proposed framework first pretrains the model on a related source dataset to learn domain-relevant medical priors, and then transfers the learned representations to the target AF detection task for fine-tuning. By providing a more informative initialization, this strategy reduces the dependence on large-scale target-domain annotations, alleviates overfitting caused by limited labels, and improves generalization to unseen ECG data. Meanwhile, SpecMFM-Net further enhances AF feature extraction from the representation level. Specifically, a super-resolution spectrogram construction strategy based on the fractional adaptive superlet transform (FASLT) is adopted to improve the separability of multi-scale rhythm structures and transient morphological variations. The network then incorporates a multi-scale feature modulation (MSFM) module to emphasize discriminative temporal-spectral features, along with a bidirectional long short-term memory (BiLSTM) module to model rhythm dynamics. Experimental results show that SpecMFM-Net achieved an average accuracy of 95.82% and an average F1 score of 90.04% on the CinC2017 three-class classification task. In addition, it achieved an F1 score of 91.54% in cross-dataset testing on AFDB, suggesting favorable generalization potential for AF detection under limited high-quality annotations. The main contributions are summarized as follows:

1. A medical-prior-guided transfer learning framework is proposed for AF detection, which learns medical-specific priors from a related source dataset and transfers them to the target task, reducing dependence on large-scale high-quality annotations and alleviating overfitting under annotation-limited conditions.

2. A spectrogram-enhanced multi-scale feature modulation network, SpecMFM-Net, is introduced to capture discriminative AF-related rhythm and morphological patterns by integrating FASLT-based super-resolution spectrograms, MSFM-based feature enhancement, and BiLSTM temporal modeling.

3. Extensive experiments on CinC2017 and AFDB demonstrate that the proposed method achieves strong classification performance and cross-dataset generalization, indicating its potential for AF detection with limited high-quality annotations.

2 Method

The overall workflow of this study is illustrated in Figure 1. The raw ECG signals are first segmented and denoised, and then transformed into standardized spectrograms using FASLT[12]. During training, Synthetic Minority Over-sampling Technique (SMOTE)[13] is applied to mitigate class imbalance, and the proposed SpecMFM-Net is trained under a medical-prior-guided transfer learning setting to perform classification.

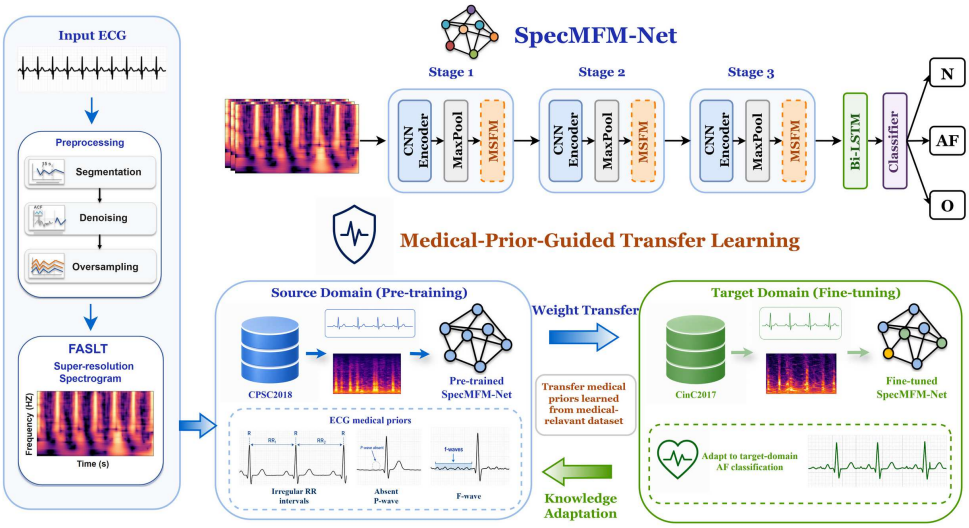

Fig. 1 Flowchart of the SpecMFM-Net model based on medical-prior-guided transfer learning.

2.1 Dataset

To comprehensively evaluate the effectiveness of the proposed method and to verify the performance gains brought by the transfer learning strategy, this study uses three public ECG datasets: PhysioNet/Computing in Cardiology Challenge 2017 Dataset (CinC2017)[14], China Physiological Signal Challenge 2018 (CPSC2018)[15], and Atrial Fibrillation Database (AFDB)[16].

The CinC2017 dataset contains 8,528 single-lead ECG recordings sampled at 300 Hz and processed with a 0.5–40 Hz band-pass filter. The recording length ranges from 9 to 61 seconds. Each record is annotated by experts into four categories: Normal rhythm, AF, Other rhythm, and Noisy.

The CPSC2018 dataset includes 6,877 12-lead ECG recordings collected from 11 hospitals, with a sampling rate of 500 Hz. To construct a subset suitable for the three-class task, we simplify the labels into three categories, Normal, AF, and Other, resulting in 918 Normal records, 1,098 AF records, and 4,861 Other records. The first lead signal is used in subsequent experiments.

The AFDB dataset contains long-term ECG recordings from 25 subjects, approximately 10 hours per record, sampled at 250 Hz. In this study, the first channel is used, and the recordings are segmented into AF and non-AF clips according to the annotations for analysis.

2.2 Data Preprocessing

To unify the input dimensionality, all ECG recordings were standardized into 15 s segments. Wavelet-thresholding denoising method were used reduce noise[17]. To address class imbalance in multi-class arrhythmia recognition, SMOTE proposed by Chawla et al.[13] was applied as the default data augmentation strategy only to the training

set within each cross-validation fold, where minority-class samples were generated by interpolating between nearest neighbors from the same class.

Fractional adaptive superlet transform (FASLT)[12] provides a more flexible multi-resolution representation, enabling improved localization of both slowly varying rhythms and transient morphological components. Therefore, we adopt FASLT to map the signals into a two-dimensional temporal-spectral space and represent rhythm irregularities together with dynamic changes in energy distribution. Specifically, for a given signal $x(t)$, a set of Morlet wavelets is constructed at a center frequency f, sharing the same center frequency but using different numbers of cycles. Let $\psi_{f,c}(t)$ be the complex Morlet wavelet with center frequency f and bandwidth parameter c. The time-scale is defined as $\sigma_t = \frac{c}{2\pi f}$. The wavelet response $W_{f,c}(t)$ is obtained by the convolution $x(t) * \psi_{f,c}(t)$. Under a superlet combination of order o, the response is defined as the geometric mean of the responses obtained with different cycle numbers and can be expressed as follow:

$$\mathrm{FASLT}_o(f,t) = \left(\prod_{j=1}^{o} \left| W_{f,c_j}(t) \right| \right)^{\frac{1}{o}}, \quad c_j = j \cdot c_1 \tag{1}$$

where c_1 is the base numbers of cycles, and o is the combination order. To accommodate resolution requirements across different frequency bands, FASLT lets the order vary linearly with frequency. A corresponding $o(f)$ is used at each f to generate the spectrogram as the network input.

$$o(f) = o_{\min} + (o_{\max} - o_{\min}) \cdot \frac{f - f_{\min}}{f_{\max} - f_{\min}} \tag{2}$$

Considering the physiological characteristics of ECG, we set $f_{\min} = 0.5$ Hz, $f_{\max} = 30$ Hz, the base number of cycles $c_1 = 3$, $o_{\min} = 1$ and $o_{\max} = 15$. Under this configuration, low-frequency bands use low-order superlets close to CWT to preserve time resolution for slowly varying components such as heart-rate rhythms, whereas high-frequency bands use high-order superlets to enhance frequency discriminability for transient components such as the QRS complex, thereby yielding a more focused and clearer time–frequency representation. Finally, the FASLT magnitude spectrum is normalized and uniformly resized to 224×224 as the input to the model.

2.3 Medical-prior-guided Transfer Learning

Transfer learning improves target-task performance by reusing knowledge learned from a related source task, instead of training a model entirely from scratch[18]. This strategy is especially useful when labeled data are limited. In ECG arrhythmia classification, the dominant approach is to transform ECG signals into two-dimensional (2D) representations and use ImageNet pretrained models, such as ResNet, AlexNet, and VGG, as feature extractors or initialization networks[19]. These pretrained ImageNet models capture low-level and generic visual patterns, such as edges, contours, and

local textures, which can facilitate the interpretation of two-dimensional ECG representations by providing transferable feature extraction capabilities, thereby improving classification performance and reducing reliance on extensive newly collected and manually labeled ECG data. However, given the substantial semantic distribution gap exists between generic natural image datasets and ECG spectrogram tasks, directly transferring ImageNet-pretrained weights is inherently suboptimal. Although such models can capture generic low-level structural patterns, they offer limited medically relevant prior knowledge for arrhythmia analysis, thereby restricting performance gains and potentially introducing representation bias. In contrast, pretraining on a related ECG dataset may be more effective, even if the source dataset is also limited in size. It can provide domain-specific representations and medically relevant priors that are better aligned with the target task. As a result, it may outperform models pretrained on large-scale natural image datasets such as ImageNet. Therefore, we adopt a medical-prior-guided transfer learning strategy tailored to the target scenario. Rather than depending on pretraining from natural image datasets, we first pretrain the model on a related ECG dataset so that it can learn medically relevant and task-aligned priors from ECG-specific patterns. Specifically, ECG signals from the CPSC2018 dataset are converted into FASLT-based spectrograms with clear clinical characteristics and used for pretraining. The pretrained model is then transferred to the CinC2017 dataset for fine-tuning, enabling an evaluation of its effectiveness and adaptability under domain-relevant pretraining.

2.4 Model

This study proposes a spectrogram-enhanced multi-scale feature modulation network, termed SpecMFM-Net, to accurately identify AF-related temporal and spectral pathological features from high-resolution ECG spectrograms. As shown in Figure 2, the model uses a convolutional backbone to extract hierarchical local representations from the spectrogram, providing the basis for subsequent feature modulation and sequence modeling. A multi-scale feature modulation module (MSFM) is then introduced to modulate temporal-spectral features, highlighting discriminative responses while suppressing redundancy and residual artifacts. Finally, the enhanced features are serialized and fed into a BiLSTM to model bidirectional temporal dependencies, capturing contextual information of rhythm evolution and compensating for the limited ability of pure convolution to characterize long-range dependencies. The classification head outputs class probabilities to complete the classification.

2.4.1 Convolutional Feature Extraction

The convolutional backbone is used to learn hierarchical local representations from the FASLT processed spectrograms. The backbone consists of three convolutional stages with progressively increasing depth from 64 to 256. Each stage comprises convolution operations followed by batch normalization and nonlinear activation, and ends with pooling for scale reduction, while Dropout is employed for standard regularization. This design implements a progressive trade-off, where the spatial resolution is

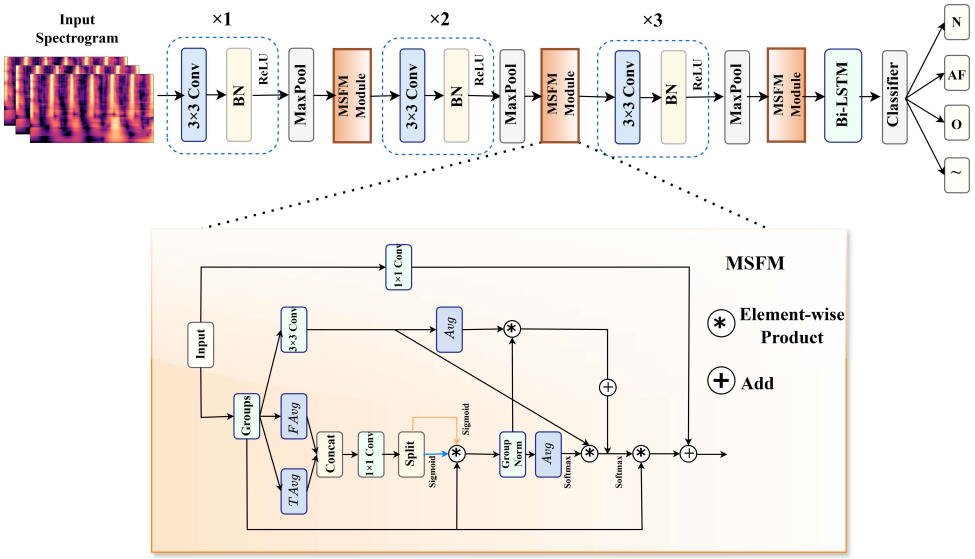

Fig. 2 Structure of the SpecMFM-Net.

downsampled and the number of feature channels is expanded, effectively abstracting representations toward key diagnosing temporal-spectral patterns.

2.4.2 MSFM Module

Standard convolutions are limited by fixed receptive fields, making it challenging to simultaneously capture localized morphological perturbations and long-range or cross-band distributed patterns. Moreover, high-dimensional channel-wise computation introduces redundancy. To mitigate these issues and further enhance feature discriminability, we propose the Multi-Scale Feature Modulation(MSFM) module, which dynamically modulates temporal-spectral features driven from convolutional backbone in multi-scales. The MSFM module employs a dual-path architecture to achieve selective feature modulation within temporal-spectral representations. Its core design balances an efficient gating mechanism for discriminative feature enhancement with a residual pathway dedicated to preserving informational fidelity, particularly for subtle diagnostic components.

The modulation branch follows a grouped parallel modeling framework, where the input channels are evenly partitioned into G groups to reduce computational overhead and enable efficient subspace-wise feature learning. In our implementation, we set $G = 8$ across all MSFM modules as a trade-off between computational efficiency and representation capacity. Each group is processed independently within a low-dimensional subspace, allowing the model to capture diverse local patterns while maintaining computational efficiency. This branch consists of two parallel subbranches applied within each group: an Axial Context Branch that aggregates feature responses along the temporal and spectral axes via long-range pooling to capture global structural priors and long-range dependencies, and a Local Detail Branch that

extracts fine-grained patterns through more localized operations. The outputs of these sub-branches are fused through cross-spatial interaction to capture complementary semantic patterns, ultimately generating a recalibration mask M that emphasizes salient regions in the time–frequency domain.

Critically, to prevent the over-modulation of high-energy components (e.g., QRS complexes) and ensure the preservation of low-energy yet diagnostically critical oscillations (e.g., atrial fibrillatory f-waves), MSFM incorporates an Ungated Residual Information Pathway. Unlike the gating branches, this pathway maintains global contextual information via a learnable 1×1 convolutional projection $\mathcal{F}_{\text{proj}}(\cdot)$, which performs adaptive channel alignment while preserving the original feature resolution. This provides a stable, low-loss channel for transmitting original features, mitigating potential information bottlenecks caused by deep gating operations. The final output Y is formulated as the sum of this residual stream and the gated, enhanced features, enabling an additive–multiplicative fusion:

$$Y = \mathcal{F}_{\text{proj}}(X) + \text{Reshape}\left(X_{\text{grouped}} \odot M\right) \tag{3}$$

Where $\odot$ denotes element-wise multiplication and Reshape$(\cdot)$ restores the grouped tensor to its original dimensionality. In this formulation, the multiplicative term selectively emphasizes discriminative temporal-spectra patterns, while the additive linear term safeguards low-energy diagnostic signals and stabilizes gradient flow. This dual-path design enables robust feature modulation with enhanced representational fidelity for complex physiological signals.

2.4.3 Bi-directional Temporal Modeling

To characterize the contextual information of rhythm patterns as they evolve over time, SpecMFM-Net introduces a BiLSTM on top of the MSFM-modulated features to model bidirectional temporal dependencies. Specifically, the two-dimensional temporal–spectral features are flattened over the spatial grid and serialized to form a feature sequence that can be processed by recurrent networks. By incorporating a bidirectional architecture, the model can leverage both past and future context, thereby more effectively capturing long-range temporal dependencies and improving its ability to recognize fine-grained temporal–spectral abnormal patterns associated with rhythm evolution.

3 Experiment

3.1 Experimental Settings

The model was trained using the Adam optimizer with cross-entropy loss. The input size was set to $224\times224\times1$, with a batch size of 48. The learning rate was initialized at 1×10^{-3} for pre-training and 1×10^{-4} for fine-tuning. To mitigate overfitting, Dropout (0.2) and weight decay (1×10^{-4}) were applied. A ReduceLROnPlateau scheduler (patience = 5) and early stopping were adopted, and the best model was selected based on validation performance.

All experiments were conducted on a NVIDIA RTX PRO 6000 GPU. The experimental pipeline was implemented in Python 3.12 and PyTorch 2.8.0.

3.2 Evaluation Metrics

To comprehensively assess model performance, this study uses accuracy, precision, recall, and F1 score as quantitative metrics. For the multi-class task, precision, recall, and F1 score are computed on a per-class basis using a one-vs-all strategy, and macro-averaging is reported as the overall summary metric to mitigate the impact of class imbalance on evaluation.

4 Results and Discussion

4.1 Model Evaluation

To evaluate the classification performance and robustness of SpecMFM-Net in complex ECG scenarios, we performed a 5-fold cross-validation on the target dataset CinC2017. To mimic a medical-prior-guided transfer setting, the model was first pretrained on the source-domain dataset CPSC2018 using a supervised classification objective, and subsequently fine-tuned on CinC2017 under the same training protocol. Given the discrepancy between the label spaces of the source and target domains, only the backbone parameters were transferred, while the classification head was randomly initialized and jointly optimized during fine-tuning. Following prior studies[20, 21] and the official CinC2017 evaluation protocol, we report the macro-averaged F1-score over three clinically relevant classes (Normal, AF, and Other).

The aggregated 5-fold cross-validation results are summarized in Table 1. The proposed model achieved an average accuracy of 95.82% and an average F1 score of 90.04%, demonstrating strong overall performance of the proposed framework. Across different folds, the model maintained generally consistent results, indicating good robustness and generalization ability under different data partitions. Overall, these results suggest that the proposed method can effectively balance sensitivity and precision in AF detection.

Table 1 Classification performance of SpecMFM-Net on the CinC2017 dataset.

Fold	F1 (%)	Acc (%)	Recall (%)	Precision (%)
Fold 1	90.05	95.82	87.15	93.68
Fold 2	91.41	96.27	89.73	93.31
Fold 3	90.33	95.78	88.87	91.96
Fold 4	88.22	95.20	84.38	93.56
Fold 5	90.18	96.04	85.73	96.58
Avg	**90.04**	**95.82**	**87.17**	**93.82**

4.2 Evaluation of Fractional Adaptive Superlet Transform

To evaluate the effectiveness of the fractional adaptive superlet transform (FASLT) for ECG representation, we compared it with two widely used time-frequency analysis methods, including short-time Fourier transform (STFT) and continuous wavelet transform (CWT). To ensure a fair comparison, the same classification backbone of SpecMFM-Net and the same training setting without transfer learning were used for all three input representations. The corresponding classification results are summarized in Table 2. STFT and CWT showed weaker overall performance, suggesting that conventional time-frequency methods remain limited by the trade-off between temporal and spectral resolution when handling nonstationary ECG signals. As a result, they are less effective at simultaneously preserving transient waveform details and frequency-structure information. In contrast, FASLT achieved the best overall performance. This result indicates that the superior time-frequency focusing capability of FASLT can generate more discriminative spectrogram representations, thereby improving the detection of subtle pathological patterns and reducing the risk of missed diagnoses.

Table 2 Classification performance of SpecMFM-Net using different ECG time–frequency representations.

Method	F1 (%)	Acc (%)	Recall (%)	Precision (%)
STFT	73.63	85.59	70.25	78.99
CWT	75.54	86.76	73.88	77.60
FASLT	**85.63**	**91.11**	**85.46**	**86.02**

4.3 Evaluation of Medical-Prior-Guided Transfer Learning

To examine the contribution of medical-prior-guided transfer learning under limited high-quality annotations, we evaluated SpecMFM-Net under two initialization settings. The first one was training from scratch without domain knowledge, denoted as Scratch. The second one was fine-tuning from a model pretrained on CPSC2018, denoted as Pretrained. For a fair comparison, both settings used the same input preprocessing, network architecture, optimization strategy, and evaluation protocol on the CinC2017 test set. As shown in Table 3, the model initialized with transferred medical-domain knowledge consistently outperforms its scratch counterpart across all evaluation metrics. This finding suggests that ECG-specific prior knowledge learned from CPSC2018 provides a more suitable inductive bias for downstream AF recognition and places the model closer to the physiological signal distribution before fine-tuning. Consequently, SpecMFM-Net can capture discriminative time-frequency patterns and rhythm-related characteristics more effectively under limited labeled data.

Table 3 Comparison of SpecMFM-Net with and without medical-prior-guided transfer learning on the CinC2017 dataset.

Method	F1 (%)	Acc (%)	Recall (%)	Precision (%)
Scratch	85.63	91.11	85.46	86.02
Pretrained	**90.04**	**95.82**	**87.17**	**93.82**

Table 4 Comparison of medical-prior-guided transfer learning and ImageNet pretraining for AF classification on the CinC2017 dataset.

Method	F1 (%)	Acc (%)	Recall (%)	Precision (%)
SpecMFM-Net (images)	87.28	91.79	87.81	86.88
SpecMFM-Net (medical-prior)	**90.04**	**95.82**	**87.17**	**93.82**
EfficientNet-B0 (images)	85.15	91.11	84.32	86.27
EfficientNet-B0[22] (medical-prior)	**87.65**	**91.84**	**86.99**	**88.35**
MobileNetV2 (images)	78.94	87.20	77.88	80.07
MobileNetV2[23] (medical-prior)	**87.83**	**88.03**	**87.67**	**88.07**
Vision Transformer (images)	82.32	89.93	80.69	84.18
Vision Transformer[24] (medical-prior)	**87.81**	**88.12**	**87.87**	**87.76**

Building on the above results, we further compared medical-domain-guided transfer learning with conventional pretraining methods on ImageNet. Specifically, the models were initialized either from ImageNet or from ECG spectrograms generated from CPSC2018, and then fine-tuned under the same protocol on CinC2017. As shown in Table 4, initialization from CPSC2018 consistently yielded higher F1-scores across all architectures compared to ImageNet initialization. Given that the F1-score is particularly sensitive to the classification performance of the minority class in imbalanced medical datasets, this consistent improvement indicates that medical-prior-guided transfer learning enhances the model's ability to identify true diseased cases. Although other metrics exhibited slight fluctuations in certain settings, the robust gains in overall classification performance demonstrate that medical-prior-guided transfer learning provides more domain-specific and clinically relevant prior knowledge for AF classification than generic visual pretraining. Even with a smaller source dataset, pretraining on ECG spectrograms enables models to learn representations that better capture the morphological and spectral characteristics of arrhythmic signals, thereby improving generalization to the target task.

Table 5 Ablation experiment on the progressive integration of model components.

Configuration	F1 (%)	Acc (%)	Recall (%)	Precision (%)
3-Stage CNN (Baseline)	77.97	87.02	77.69	78.28
+ MSFM Module	80.01	88.92	76.93	84.05
+ BiLSTM (SpecMFM-Net)	**85.63**	**91.11**	**85.46**	**86.02**

Table 6 Performance comparison with peer works on the CinC2017 dataset.

Model	F1 (%)	Acc (%)	Recall (%)	Precision (%)
CAM-SC-CL	85.90	94.08	83.30	89.19
MFEG	87.45	94.96	83.28	93.49
2DCNN-GRU	83.76	93.36	80.54	88.20
CWT-SqueezeNet	87.99	88.28	88.03	87.96
StarNet	80.48	92.48	76.12	87.84
ConvNeXt	84.64	93.68	81.50	88.87
Vision-Transformer	87.81	88.12	87.87	87.76
MobileNet-V2	87.83	88.03	87.67	88.07
Ours	**90.04**	**95.82**	**87.17**	**93.82**

4.4 Model Ablation Experiment

To evaluate the contribution of each component of the SpecMFM-Net, we conducted a progressive ablation study by incrementally integrating the convolutional backbone, the MSFM module, and the BiLSTM, as summarized in Table 5. The 3-stage convolutional backbone establishes a robust baseline by hierarchically abstracting local temporal-spectral patterns through progressive channel expansion and spatial downsampling. Integrating the MSFM module yields a notable improvement in precision and overall accuracy. Its dual-path modulation selectively enhances discriminative time–frequency regions while the ungated residual pathway preserves low-energy diagnostic features, effectively suppressing false positives. Finally, appending the BiLSTM branch achieves the highest classification performance. By capturing bidirectional long-range temporal dependencies, it compensates for the limited receptive field of convolutional layers, enabling the model to contextualize rhythm evolution and fine-grained temporal abnormalities. Collectively, these results confirm that each component plays a distinct and complementary role, jointly driving the superior performance of SpecMFM-Net.

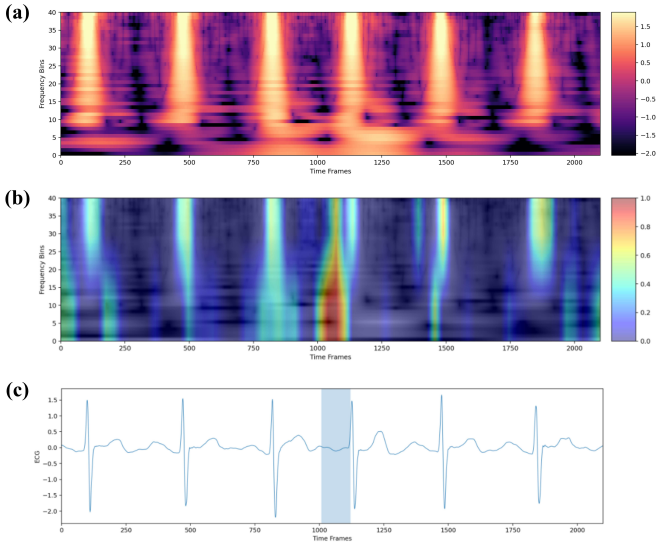

Fig. 3 Grad-CAM visualization of SpecMFM-Net attention. (a) FASLT-derived super-resolution spectrogram; (b) Attention heatmap; (c) Raw ECG.

4.5 Comparison with Existing Works

To ensure fair comparison, we re-implemented 4 state-of-the-art AF detection methods (CAM-SC-CL[4], MFEG[25], 2DCNN-GRU[26], CWT-SqueezeNet[8]) and 4 representative vision backbones (StarNet[27], ConvNeXt[28], ViT, MobileNet-V2), all trained under identical protocols with our medical-prior-guided transfer learning setting.

As shown in Table 6, our method achieves the best overall classification performance on the CinC2017 dataset, with notable improvements in F1-score and accuracy. This demonstrates the effectiveness of our joint time-frequency representation and residual multi-scale attention design.

For cross-dataset evaluation, we directly tested the CinC2017-trained models on AFDB without fine-tuning. As reported in Table 7, our method maintains superior performance, outperforming all peer works. These results confirm strong generalization capability and clinical applicability of SpecMFM-Net under domain shift.

4.6 Grad-CAM Visualization

Furthermore, we utilized Gradient-weighted Class Activation Mapping (Grad-CAM) technology to visually highlight the key focus areas of the proposed model in the FASLT spectrogram. The brighter areas in the spectrogram indicate higher importance assigned by the model to these specific time-frequency regions. As shown in Figure 3, the heatmap reveals that the model predominantly attends to the low-frequency region (0-20 Hz) during specific temporal segments. Notably, this high-attention frequency band corresponds precisely to the characteristic spectral range of atrial activity[29]. When mapped back to the temporal domain, these critical attention regions align precisely with the absence of P-waves preceding QRS complexes, which is a crucial

Table 7 Cross-dataset generalization performance on the AFDB dataset.

Model	F1 (%)	Acc (%)	Recall (%)	Precision (%)
CAM-SC-CL	90.70	90.74	91.65	90.82
MFEG	90.66	90.70	91.69	90.87
2DCNN-GRU	88.29	88.31	89.49	88.82
CWT-SqueezeNet	88.56	90.08	88.69	88.44
StarNet	82.56	82.56	84.44	84.73
ConvNeXt	88.85	88.87	90.00	89.28
Vision-Transformer	87.21	87.52	87.32	87.12
MobileNet-V2	86.84	87.07	86.65	87.15
Ours	**91.54**	**92.19**	**90.73**	**92.70**

feature of atrial fibrillation[25]. This result demonstrates that our model has learned to identify the same discriminative frequency-temporal patterns that clinical experts use for AF diagnosis, thereby validating the model's interpretability and clinical potential in automated AF detection.

5 Conclusion

In this study, we proposed SpecMFM-Net, an AF detection framework that integrates FASLT-based spectrogram representation, medical-prior-guided transfer learning, and multi-scale temporal modeling for ECG classification under limited high-quality annotations. By transferring ECG-specific medical priors from a related source dataset to the target AF task, the framework reduces dependence on large-scale target-domain annotations and improves generalization. Experiments on CinC2017 demonstrate strong classification performance, and cross-dataset testing on AFDB further verifies its generalization to unseen ECG data. Future work will focus on multi-source external validation, noise robustness, lightweight optimization, and wearable-device implementation to enhance clinical applicability.

Data availability. The datasets analysed in this study are publicly available. The PhysioNet/Computing in Cardiology Challenge 2017 dataset (CinC2017) is available from PhysioNet at https://physionet.org/content/challenge-2017/1.0.0/. The China Physiological Signal Challenge 2018 dataset (CPSC2018) is publicly available through the PhysioNet/Computing in Cardiology Challenge 2020 repository at https://physionet.org/content/challenge-2020/1.0.2/training/cpsc_2018/. The MIT-BIH Atrial Fibrillation Database (AFDB) is available from PhysioNet at https://physionet.org/content/afdb/1.0.0/.

Author contributions. Yujia Sun: Conceptualization, Methodology, Software, Writing—Original Draft. Qi Yan: Methodology, Software, Investigation, Formal Analysis, Writing—Original Draft. Haotian Tang: Software, Investigation, Formal Analysis, Writing—Review & Editing. Yunjia Wu: Data Curation, Validation. Rongrong Fu:

Validation, Formal Analysis. Hanyu Cui: Visualization, Writing—Review & Editing. Xidong Wu: Visualization. Liping Xie: Supervision, Project Administration, Writing—Review & Editing. All authors reviewed and approved the final manuscript. Yujia Sun and Qi Yan contributed equally to this work.

Competing interests. The authors declare no competing interests.

Funding. This work was supported by the National Training Program for College Students' Innovation and Entrepreneurship (20th batch at Northeastern University, Project No. 261551).

References

[1] Xie, L., Wang, L., Mo, D., Zhang, Z., Liang, M.: Intelligent algorithms powered smart devices for atrial fibrillation discrimination. Biomedical Signal Processing and Control **103**, 107480 (2025) https://doi.org/10.1016/j.bspc.2024.107480

[2] Tang, H., Yan, M., Wu, X., Lai, H., Zhang, Y., Cui, H., Xie, L.: Lightweight element-wise product enhanced neural network for efficient arrhythmia detection on embedded devices. Computer Methods and Programs in Biomedicine, 109101 (2025)

[3] Raghunath, S., Pfeifer, J.M., Ulloa-Cerna, A.E., Nemani, A., Carbonati, T., Jing, L., VanMaanen, D.P., Hartzel, D.N., Ruhl, J.A., Lagerman, B.F., Rocha, D.B., Stoudt, N.J., Schneider, G., Johnson, K.W., Zimmerman, N., Leader, J.B., Kirchner, H.L., Griessenauer, C.J., Hafez, A., Good, C.W., Fornwalt, B.K., Haggerty, C.M.: Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ecg and help identify those at risk of atrial fibrillation-related stroke. Circulation **143**(13), 1287–1298 (2021) https://doi.org/10.1161/CIRCULATIONAHA.120.047829

[4] Zhu, Q., Zhang, L., Lu, F., Fang, L., Pan, Q.: Class activation map-based slicing-concatenation and contrastive learning: A novel strategy for record-level atrial fibrillation detection. Expert Systems with Applications **262**, 125619 (2025) https://doi.org/10.1016/j.eswa.2024.125619

[5] Zhang, H., Gu, H., Gao, J., Lu, P., Chen, G., Wang, Z.: An effective atrial fibrillation detection from short single-lead electrocardiogram recordings using mcnn-blstm network. Algorithms **15**(12), 454 (2022) https://doi.org/10.3390/a15120454

[6] Liu, X., Xu, Y., Xu, H., He, L., Long, S., Huang, Y., Wang, Y., Lu, Y., Huang, Y., Wu, J., Gao, H., Liu, X.: Advancing interpretable cardiac disease diagnosis via a transformer-convolutional hybrid network on electrocardiograms. Engineering Applications of Artificial Intelligence **152**, 110675 (2025) https://doi.org/10.1016/j.engappai.2025.110675

[7] Hong, S., Zhang, W., Sun, C., Zhou, Y., Li, H.: Practical lessons on 12-lead ecg classification: Meta-analysis of methods from physionet/computing in cardiology challenge 2020. Frontiers in Physiology **12** (2022) https://doi.org/10.3389/fphys.2021.811661

[8] Malleswari, P.N., Odugu, V.k., Rao, T.S., Aswini, T.: Deep learning-assisted arrhythmia classification using 2-d ecg spectrograms. EURASIP Journal on Advances in Signal Processing **2024**(1), 104 (2024)

[9] Wei, T.-R., Lu, S., Yan, Y.: Automated atrial fibrillation detection with ecg. Bioengineering **9**(10), 523 (2022) https://doi.org/10.3390/bioengineering9100523

[10] Mavaddati, S.: Ecg arrhythmias classification based on deep learning methods and transfer learning technique. Biomedical Signal Processing and Control **101**, 107236 (2025)

[11] Raghu, M., Kleinberg, J., Zhang, C., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging, Vancouver, BC, Canada, pp. 3347–3357 (2019)

[12] Barzan, H., Moca, V.V., Ichim, A.-M., Muresan, R.C.: Fractional superlets. In: 2020 28th European Signal Processing Conference (EUSIPCO), pp. 2220–2224. IEEE, Amsterdam, Netherlands (2020)

[13] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002) https://doi.org/10.1613/jair.953

[14] Clifford, G., Liu, C., Moody, B., Lehman, L.-w., Silva, I., Li, Q., Johnson, A., Mark, R.: Af classification from a short single lead ecg recording: the physionet computing in cardiology challenge 2017. In: Computing in Cardiology (CinC), Rennes, France, pp. 1–4 (2017)

[15] Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., Liu, Y., Ma, C., Wei, S., He, Z., Li, J., Ng, E.: An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. Journal of Medical Imaging and Health Informatics **8**, 1368–1373 (2018) https://doi.org/10.1166/jmihi.2018.2442

[16] Moody, G.B., Mark, R.G.: A new method for detecting atrial fibrillation using r-r intervals. Computers in Cardiology **10**, 227–230 (1983)

[17] Yu, K., Feng, L., Chen, Y., Wu, M., Zhang, Y., Zhu, P., Chen, W., Wu, Q., Hao, J.: Accurate wavelet thresholding method for ecg signals. Computers in Biology and Medicine **169**, 107835 (2024) https://doi.org/10.1016/j.compbiomed.2023.107835

[18] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proceedings of the IEEE **109**(1), 43–76 (2020)

[19] Ismail, S.N.M.S., Razak, S.F.A., Ab Aziz, N.A.: Ecg-based transfer learning for cardiovascular disease: A scoping review. International Journal of Cognitive Computing in Engineering **6**, 280–297 (2025)

[20] Yildirim, Ö.: A novel wavelet sequence based on deep bidirectional lstm network model for ecg signal classification. Computers in Biology and Medicine **96**, 189–202 (2018) https://doi.org/10.1016/j.compbiomed.2018.03.016

[21] Gao, Y., Wang, H., Liu, Z.: An end-to-end atrial fibrillation detection by a novel residual-based temporal attention convolutional neural network with exponential nonlinearity loss. Knowledge-Based Systems **212**, 106589 (2021) https://doi.org/10.1016/j.knosys.2020.106589

[22] Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research 97, pp. 6105–6114 (2019)

[23] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)

[24] Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM computing surveys (CSUR) **54**(10s), 1–41 (2022)

[25] Wu, X., Yan, M., Wang, R., Xie, L.: Multiscale feature enhanced gating network for atrial fibrillation detection. Computer Methods and Programs in Biomedicine **261**, 108606 (2025) https://doi.org/10.1016/j.cmpb.2025.108606

[26] Mahel, A.S.B., Cao, S., Zhang, K., Chelloug, S.A., Alnashwan, R., Muthanna, M.S.A.: Advanced integration of 2dcnn-gru model for accurate identification of shockable life-threatening cardiac arrhythmias: a deep learning approach. Frontiers in Physiology **15** (2024) https://doi.org/10.3389/fphys.2024.1429161

[27] Ma, X., Dai, X., Bai, Y., Wang, Y., Fu, Y.: Rewrite the stars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5694–5703 (2024)

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

[29] Traykov, V.B., Pap, R., Sághy, L.: Frequency domain mapping of atrial fibrillation: Methodology, experimental data and clinical implications. Current Cardiology Reviews **8**(3), 231–238 (2012) https://doi.org/10.2174/157340312803217229

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.csv
- Table2.csv
- Table3.csv
- Table4.csv
- Table5.csv
- Table6.csv
- Table7.csv