AI Engineering
https://www.sciltp.com/journals/aieng

Article

Multi-Source Feature Fusion with Self-Supervised Contrastive Learning for AF Detection under Label Scarcity

Zhengyang Miao †, Hexin Wan †, Yuying Xie, Qi Yan, Dongchen Wu, Haotian Tang and Liping Xie *

College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110169, China
* Correspondence: xielp@bmie.neu.edu.cn
† These authors contributed equally to this work.

How To Cite: Miao, Z.; Wan, H.; Xie, Y.; et al. Multi-Source Feature Fusion with Self-Supervised Contrastive Learning for AF Detection under Label Scarcity. *AI Engineering* **2026**, *2*(1), 8. https://doi.org/10.53941/aieng.2026.100008

Received: 26 January 2026
Revised: 15 June 2026
Accepted: 23 June 2026
Published: 30 June 2026

Abstract: Accurate atrial fibrillation (AF) screening from short, single-lead wearable ECG remains challenging due to noise contamination, limited computation budgets, and scarce annotations. We present a multi-source feature fusion framework that combines rhythm, morphology, and implicit representations to improve AF detection under label scarcity. The pipeline performs denoising, robust R-wave localization, and heartbeat segmentation to construct rhythm sequences and beat-level waveform inputs for downstream modeling. Complementary representations are learned from rhythm variability statistics, morphology-aware encoders, and self-supervised contrastive pre-training on external data with subsequent adaptation to the target domain. The resulting embeddings are integrated using attention-guided fusion for classification. Experiments on multiple public datasets demonstrate consistent performance in low-label settings, indicating the potential of the proposed approach for wearable AF screening. Further evaluation of computational efficiency and on-device performance is still required.

Keywords: atrial fibrillation detection; unlabeled signals; wearable devices; multi-feature fusion; self-supervised learning; multi-head self-attention mechanism

1. Introduction

Cardiovascular diseases remain a major global health burden, and atrial fibrillation (AF) is among the most prevalent arrhythmias [1,2]. AF can be intermittent and asymptomatic, which makes timely diagnosis difficult [3]. Wearable ECG devices enable continuous monitoring and provide an opportunity for earlier screening, yet the recorded single-lead signals are often short and contaminated by baseline drift and motion artifacts [4]. These characteristics increase the difficulty of robust AF detection under the computation constraints of edge devices.

Electrocardiogram (ECG) signals record the electrical activity of the human heart, capturing the processes of cardiac contraction and relaxation, and indirectly revealing abnormalities in cardiac rhythm and function [5,6]. ECG serves as the gold standard for diagnosing arrhythmias such as atrial fibrillation. ECG records electrical signals generated by ion movement during cardiac activity, originating from the sinoatrial node in the conduction system. The sinoatrial node transmits electrical signals downward, triggering excitation and contraction of myocardial cells, followed by ventricular muscle fiber contraction that pumps blood to the lungs and throughout the body. In ECG recordings, a complete cardiac cycle consists of multiple waves and segments, each representing different cardiac events. During a single heartbeat cycle, six characteristic waveforms serve as diagnostic features: P-wave, PR interval, QRS complex, ST segment, T-wave, and QT interval [5,7]. Changes in the morphology, amplitude, and duration of these characteristic waveforms can reflect cardiac physiological and pathological states, providing crucial diagnostic information for clinicians [8].

In single-lead ECG signals acquired by wearable devices, atrial fibrillation (AF) exhibits a series of typical characteristics. The most prominent feature is the irregularity of RR intervals, characterized by highly disordered ventricular activation intervals lacking the periodicity of normal sinus rhythm [9]. Compared to traditional multi-lead

clinical ECG signals, wearable devices face limitations in acquisition conditions, making P-waves difficult to clearly identify. In AF, P-waves disappear or are replaced by low-amplitude, rapidly oscillating f-waves, making this waveform feature even more difficult to observe explicitly [5]. Additionally, in wearable signals, AF rhythm manifests as dramatic heart rate fluctuations with high RR interval variability, often quantified through statistical indicators such as SDNN, RMSSD, and pNN50. In the frequency domain, f-wave oscillations cause enhanced energy distribution between 5–10 Hz, with power spectral density exhibiting low-amplitude, widely dispersed characteristics. Although wearable signal quality falls short of clinical recordings, the significant abnormalities in rhythm and local waveform patterns of AF still provide an effective recognition basis for our proposed algorithm [10,11].

Prior AF detectors often rely on handcrafted rhythm or morphology features with conventional classifiers [12–15]. While interpretable, such pipelines can be sensitive to signal quality and dataset shifts, and they typically require sufficient labeled data to maintain performance. In recent years, AF detection algorithms have evolved along three paths: traditional machine learning, deep learning, and self-supervised/contrastive learning [10]. Traditional machine learning methods rely on signal processing and statistical modeling, including feature engineering, feature selection, and classifier construction. Tateno and Glass proposed an "AF index" based on RR interval irregularity, achieving over 96% accuracy on PhysioNet [12]. Mohebbi and Ghassemian utilized wavelet transforms to extract f-wave and P-wave features, combined with SVM to achieve 91.5% accuracy and 90.3% recall on MIT-BIH AFDB [13]. These methods offer strong interpretability but depend on expert-designed features, exhibit poor generalization, and require large amounts of labeled data.

Deep learning enables automatic feature learning. Zhang et al. employed 1D-CNN on the CinC_2017 dataset using ECG segments as input, achieving an F1-score of 0.89 [16]. Hannun et al. achieved physician-level screening capability using a 34-layer residual network on Apple Watch data. While these methods demonstrate strong expressive power, their complex models and high computational requirements make deployment on wearable devices challenging [5]. Recent deep and self-supervised methods alleviate label dependence by learning representations from large-scale unlabeled ECG data, but transferring these representations to short, noisy wearable recordings and combining heterogeneous evidence sources remains nontrivial [10,11]. Self-supervised and contrastive learning reduce label dependency. Zheng et al.'s ECG-CL achieved superior performance to fully supervised methods across multiple datasets through contrastive pre-training on unlabeled data followed by minimal fine-tuning [11]. TS-TCC combines temporal and frequency domain augmentation to learn time-series invariance, demonstrating robustness in UCR and AFDB datasets. These methods learn features from unlabeled data, making them suitable for wearable devices. To balance performance and edge deployment, lightweight methods have been proposed. Chen et al.'s low-complexity SVM model achieved an F1-score of 0.84 on CinC_2017 [14]. Xia et al. employed EMD to extract IMF features, combined with LDA to maintain 93.4% accuracy under high noise conditions, providing references for practical deployment [15].

Despite significant progress, the "few/unlabeled + wearable device" scenario still lacks lightweight, robust, and transferable solutions. This work addresses AF detection under limited annotations by integrating three complementary cues. We model rhythm irregularity through RR-interval statistics, capture waveform structure via a multi-scale convolutional residual encoder, and learn implicit embeddings through self-supervised contrastive pre-training on SHDB-AF, followed by supervised fine-tuning on a small labeled subset of CPSC_2025. The resulting embeddings are fused with attention and gated aggregation to form a compact representation for classification. This study integrates multimodal features, self-supervised pre-training, and attention mechanisms to propose a novel AF detection framework that learns discriminative features from unlabeled data and adapts to target scenarios through minimal fine-tuning. The proposed design is evaluated across CPSC_2025, CinC_2017, and SHDB-AF to examine its performance under limited annotations and different data distributions.

2. Materials and Methods

2.1. Overview

To achieve effective AF recognition from wearable ECG signals with high accuracy and robustness, we design an attention-based multi-source feature fusion network for AF detection. The overall framework comprises key modules including R-wave detection, multi-source feature modeling, multi-source feature fusion network, loss optimization, and contrastive learning.

The technical pipeline, illustrated in Figure 1, consists of three main stages: (1) *Data preprocessing*: including wavelet denoising of raw signals, R-wave localization, heartbeat segmentation, and RR interval sequence construction; (2) *Feature extraction*: utilizing morphological and rhythm modeling branches to extract QRS morphological features and RR interval rhythm features, respectively; (3) *Feature fusion and classification*: completing multi-source feature integration through an attention-guided feature fusion module, while jointly optimizing features and weight

configuration using combined loss functions to achieve AF signal detection.

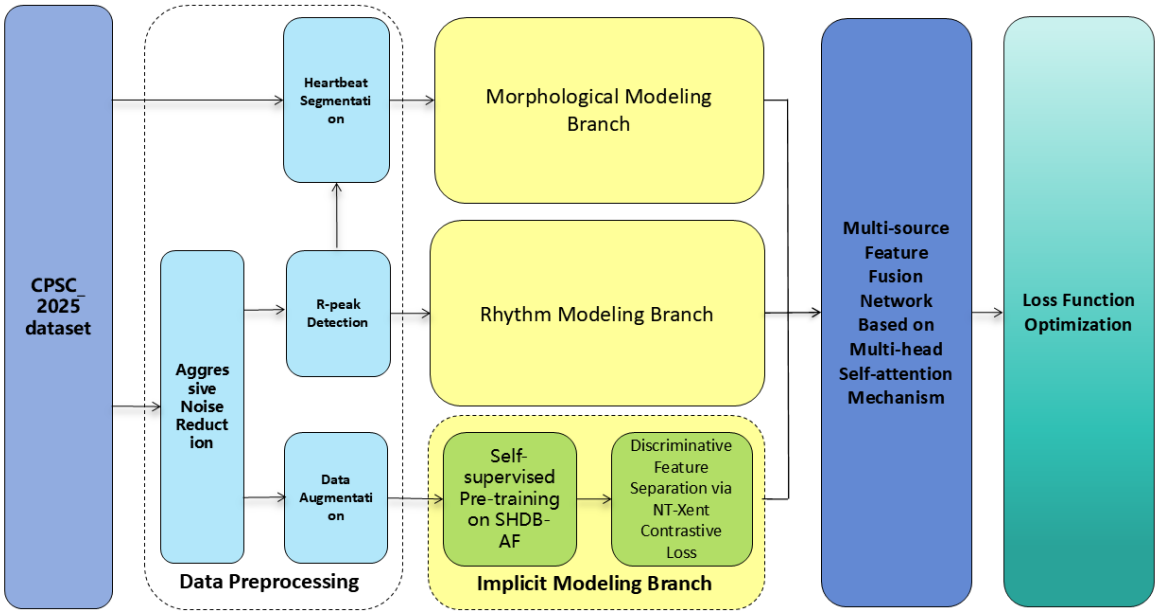

Figure 1. Overall technical pipeline of the proposed AF detection framework.

2.2. Data Preprocessing

2.2.1. Dataset Description

This study employs three datasets: the China Physiological Signal Challenge dataset (CPSC_2025), the Saitama Heart Database for Atrial Fibrillation (SHDB-AF) [17], and the 2017 PhysioNet/CinC Challenge dataset (CinC_2017) [18]. Table 1 summarizes the main characteristics and experimental usage of the three datasets.

Table 1. Summary of the three ECG datasets used in this study.

Dataset	Signal Type	Sampling Rate	Data Format/Duration	Usage in This Study
CPSC_2025	Single-lead ECG	400 Hz	MAT format; each recording lasts 10 s; 20,000 training samples including 1000 labeled samples, and 10,000 unlabeled test samples	Used as the target-domain dataset, with 200 labeled samples for supervised fine-tuning and the remaining 800 samples for held-out testing.
CinC_2017	Single-lead ECG from wearable or portable devices	300 Hz	MATLAB V4 WFDB-compatible format; recordings range from 9 s to over 60 s; 8528 training records and 3658 test records	Used exclusively for cross-dataset evaluation and not used during pre-training or fine-tuning.
SHDB-AF	Holter ECG with two leads: modified CC5 and NASA leads	125 Hz	WFDB format; ECG1 corresponds to modified CC5 and ECG2 corresponds to NASA lead	Used for self-supervised pre-training of the implicit feature encoder.

The CPSC_2025 dataset is publicly available and contains single-lead ECG signals sampled at 400 Hz, with each signal lasting 10 s, stored in .mat format. The dataset provides training data (partially labeled) and unlabeled test data, including 20,000 training samples (1000 labeled) and 10,000 test samples.

The SHDB-AF dataset stores all ECG recordings in WFDB format. Signals were recorded using Fukuda Holter monitors, digitized at 125 Hz, with two leads: modified CC5 and NASA leads. Each ECG file (.dat extension) contains two channels: “ECG1” representing the modified CC5 lead and “ECG2” representing the NASA lead.

The CinC_2017 dataset primarily originates from wearable and portable ECG acquisition devices, containing 8528 single-lead ECG records with durations ranging from 9 s to slightly over 60 s. The test set contains 3658 similar-length ECG records. ECG signals are sampled at 300 Hz and bandpass filtered by AliveCor devices. All data are provided in MATLAB V4 WFDB-compatible format (each including a .mat file containing the ECG and a .hea file containing waveform information).

In the experimental protocol, SHDB-AF was used for self-supervised pre-training of the implicit encoder. From the 1000 labeled CPSC_2025 records, 200 samples were selected for supervised fine-tuning, while the remaining 800 samples were held out for final testing. CinC_2017 was used exclusively for cross-dataset evaluation and did not contribute to pre-training, fine-tuning, or parameter selection.

2.2.2. Standardization

To unify the formats and parameters of CPSC_2025, SHDB-AF, and CinC_2017 datasets, we standardize them to CPSC_2025 specifications (single-lead, 400 Hz, 10 s, .mat format). SHDB-AF is extracted from WFDB format using the ECG1 lead and converted to .mat format, while CinC_2017 is used directly. Both datasets are upsampled from 125 Hz and 300 Hz to 400 Hz, respectively. Signals are uniformly truncated to 10 s, with longer segments cropped and shorter segments zero-padded. All signals undergo 0.5–40 Hz bandpass filtering and Z-score normalization. Metadata formats are uniformly transcribed to CPSC_2025 format. Finally, signal quality is verified through QRS detection to ensure data suitability for unified modeling and analysis.

2.2.3. Data Denoising

To enhance model accuracy and robustness in multi-task feature extraction, we design a hierarchical denoising scheme tailored to different signal processing requirements, serving the rhythm modeling branch and heartbeat morphological modeling branch separately.

For the rhythm modeling branch, to ensure R-wave localization stability and high signal-to-noise ratio, we employ strong denoising processing. Specifically, we use wavelet transform denoising based on autocorrelation threshold adjustment, selecting the db7 wavelet basis function for four-level decomposition of the original signal. The optimal soft threshold is determined through normalized zero-crossing non-periodic peak (NZOPP) and autocorrelation functions. After compressing high-frequency noise, the signal is reconstructed to enhance QRS energy concentration and improve R-wave candidate quality.

For the heartbeat morphological modeling branch, to preserve structural details such as QRS waves and P-waves while avoiding damage to morphological details, we employ only mild wavelet denoising, aiming to eliminate baseline drift and minimal high-frequency noise interference while maximally preserving the spatial structure of the original waveform.

2.2.4. R-wave Localization

Precise R-wave localization is a crucial step for calculating RR interval sequence irregularity coefficients. We employ a multi-scale QRS wave identification method combining wavelet packet energy superposition with second-order derivative difference joint criteria, demonstrating better robustness on poor-quality signals compared to the Pan-Tompkins algorithm. First, three-level wavelet packet reconstruction is performed on the strongly denoised signal to generate multi-scale energy response maps, and R-wave candidate points are identified through sliding windows in local energy peaks. Subsequently, slope mutation points are detected and combined with signal polarity judgment to further correct candidate positions. Finally, redundant filtering and missed detection compensation are completed based on empirical RR interval constraints (refractory period set between 250 ms and 2 s), forming a physiologically reasonable R-wave sequence (Figure 2).

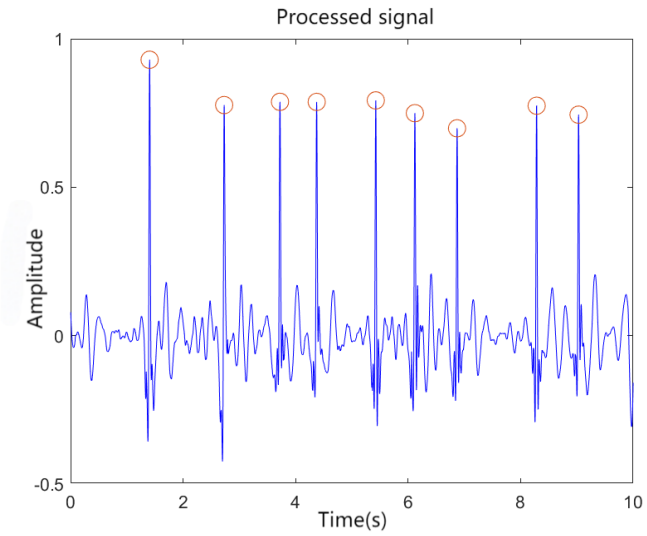

Figure 2. R-wave localization and sequence generation.

2.2.5. Heartbeat Segmentation

On the original signal, we perform only mild wavelet denoising to eliminate baseline offset and high-frequency noise. When constructing heartbeat segments, based on the R-wave sequence, for each R-wave, we extract 100 sampling points forward and 200 sampling points backward, forming a standard heartbeat input of length 300 (Figure 3). Considering that some signals may have dense R-waves due to premature beats, artifacts, or T-wave misidentification, leading to overlapping heartbeat segments and abnormal morphology, we introduce a refractory window mechanism during segmentation, excluding overlapping heartbeat segments with intervals less than a set threshold (e.g., 200 ms) before and after, ensuring independence and structural consistency of segments in the training set. The final retained high-quality heartbeat segments are used for the morphological modeling branch.

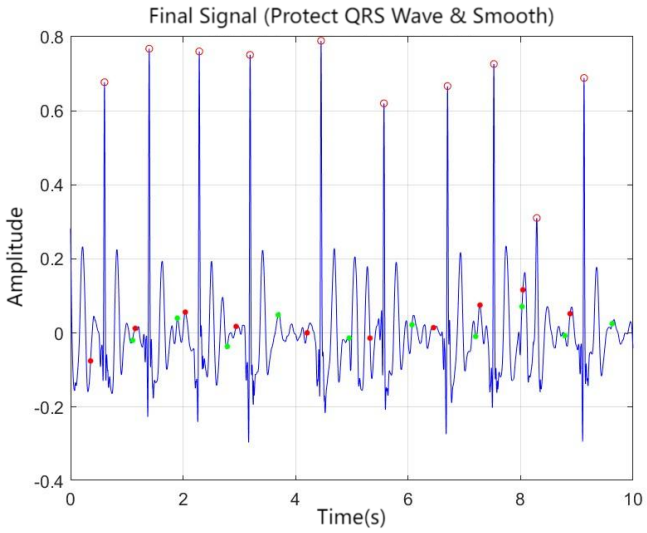

Figure 3. Heartbeat segmentation process.

2.3. Feature Modeling Branches

In ECG signal detection, P-wave absence, irregular RR intervals, and abnormal QRS wave morphology are typical characteristics of AF. In this project, we adopt multi-feature combination for AF signal detection. We extract features from three perspectives: rhythm features, morphological features, and potential implicit features, and prepare for subsequent model fusion by unifying output dimensions.

2.3.1. Rhythm Modeling Branch

Rhythm features are core criteria for arrhythmias, especially AF, reflecting the variability, stability, and sudden rhythm changes of heartbeat time series. Based on previous R-wave localization results, we select RR interval

sequences of 10 consecutive heartbeats for each signal to construct a rhythm analysis framework and employ statistical feature extraction methods for variability quantification. Input RR interval sequences are first normalized to unify dimensions and enhance feature stability, then we calculate standard deviation (SDNN), root mean square of successive differences (RMSSD), and the proportion of RR interval differences greater than 50 ms (pNN50), characterizing overall variability, short-term fluctuations, and sudden rhythm changes, respectively.

Additionally, considering that AF rhythms often exhibit alternating long-short patterns and periodic absence of abnormal RR interval patterns, we introduce the coefficient of variation (CV) as a supplementary indicator, further emphasizing significantly deviant rhythm features through the ratio of standard deviation to mean, enhancing the model's responsiveness to irregular signals. The coefficient of variation formula is:

$$\mathrm{CV} = \frac{\sigma}{\mu} \tag{1}$$

where σ represents the standard deviation and μ represents the mean of RR intervals.

The rhythm features are concatenated and mapped to a unified feature vector of dimension (1×128) through fully connected layers for input to subsequent models, ensuring rhythm features have consistent dimensions and expressive capabilities with other modalities in the feature space. In the classification process, rhythm representations based on statistical features help effectively distinguish AF from non-AF heartbeats in high-dimensional space, improving the model's discriminative robustness under complex rhythm backgrounds.

2.3.2. Morphological Modeling Branch

Morphological features are another key dimension in AF recognition, primarily reflected in the shape, amplitude, and local variation patterns of ECG waveforms. To address this characteristic, we design a morphological feature extraction branch based on multi-scale convolution and residual attention mechanisms, aiming to extract highly expressive local morphological feature representations from individual heartbeat segments. The model input is a single-lead ECG signal segment of length 300, covering a complete heartbeat cycle.

In the initial stage, the signal input is fed into a multi-scale convolution module, where three parallel one-dimensional convolution channels are constructed with kernel sizes of 3, 5, and 7, respectively, to capture ECG waveform changes at different scales. This design enhances the model's ability to identify rhythm abnormality-related morphologies such as P-wave absence, QRS wave distortion, and f-wave oscillations. Features extracted from each channel are then fused along the channel dimension to generate a set of local feature maps with multi-scale expressive capabilities.

Subsequently, multi-scale features are fed into three groups of cascaded residual structures, each residual block consisting of two layers of 1D convolution with Squeeze-and-Excitation (SE) attention mechanisms inserted in the middle or end. This mechanism strengthens the model's perception of key morphological structures through channel attention, particularly in enhancing responses to R-wave rising edges, T-wave width changes, and subtle f-wave fluctuations. Meanwhile, skip connection structures preserve low-level feature information, helping the model maintain detail stability while extracting deep features.

Finally, the residual module output is compressed through global average pooling to obtain a morphological feature vector of dimension (1×128) (consistent with other dimensions) for representing the global morphological features of the current heartbeat segment. This vector serves as input to the subsequent discriminative model, helping to accurately characterize the differences between AF and non-AF ECG waveforms in feature space, improving the overall model's recognition performance and discriminative robustness. The branch framework is illustrated in Figure 4.

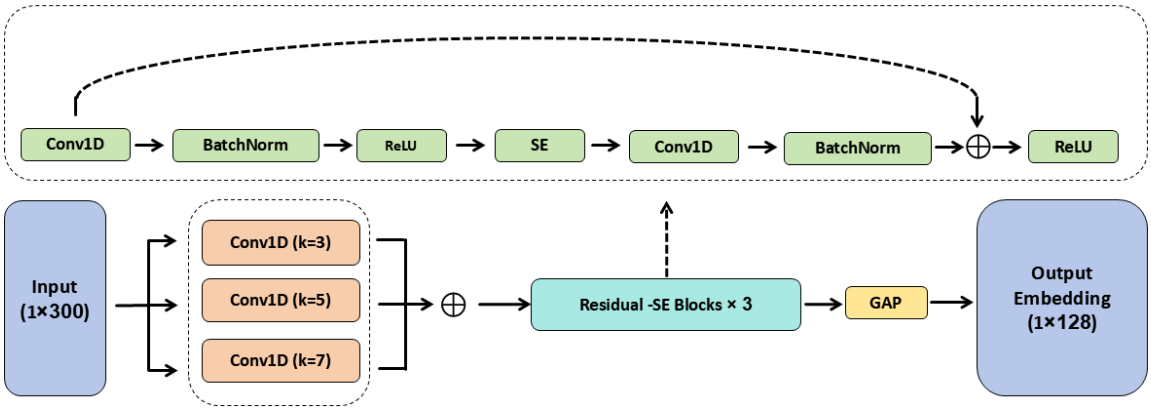

Figure 4. Architecture of the morphological modeling branch.

2.3.3. Implicit Feature Modeling Branch

Although rhythm modeling and morphological modeling can capture temporal variations and local waveform structures of arrhythmias, respectively, AF signals still contain a large amount of implicit feature information that is difficult to explicitly quantify, such as subtle but continuous perturbations, small-amplitude non-periodic fluctuations, cross-scale structural abnormalities, and individual-specific interference backgrounds. Such potential features are often difficult to model in traditional statistical models or explicit convolutional encoding structures, limiting the model's recognition capability for boundary samples and complex backgrounds.

To address this, we introduce an implicit feature modeling branch based on the TS-TCC self-supervised contrastive learning framework. The implicit encoder was first pre-trained on SHDB-AF without using class labels. It was subsequently fine-tuned using 200 labeled samples selected from the 1000 labeled CPSC_2025 records. The remaining 800 labeled CPSC_2025 records were held out and used only for final testing. CinC_2017 was not used during pre-training or fine-tuning and was used exclusively for cross-dataset evaluation.

Structurally, this branch adopts a dual-channel contrastive encoding architecture: one channel inputs the original ECG segment, while the other inputs its time-frequency joint perturbation-enhanced view, extracting temporal embedding features through encoders, respectively. To strengthen the model's structural cognition of key micro-perturbation features, during training, we minimize distances between homologous samples (enhanced views of the same signal) in the embedding space while maximizing differences between heterologous samples, thereby explicitly guiding the model to learn cross-view invariant features and class-discriminative structures.

Finally, this branch outputs a feature vector of dimension (1×128) to downstream modules, complementing the morphological and rhythm modeling branches to effectively enhance the model's overall discriminative capability for complex AF signals.

2.4. Attention-Based Multi-Source Feature Fusion Network

This module integrates embedding vectors output from three branches, including rhythm features (Rhythm Embedding), morphological features (Morphology Embedding), and implicit features (Implicit Embedding), each of dimension 128. First, a multi-path gating mechanism assigns learnable selective weights to each original modal pathway, controlling the flow intensity of the three input features separately. The gated features then undergo pairwise interaction concatenation, followed by linear mapping regression to a unified 128 dimensions to reduce interference between modalities. Subsequently, multi-head self-attention mechanism (6-heads) is employed to strengthen structural associations between features. Finally, residual connections preserve original interaction information, generating a fused output vector (Final Fusion Embedding, 128 dimensions) as input to the downstream classifier. This module enhances collaborative expressive capabilities between different modal features, providing unified and robust representation support for AF detection. The process is illustrated in Figure 5.

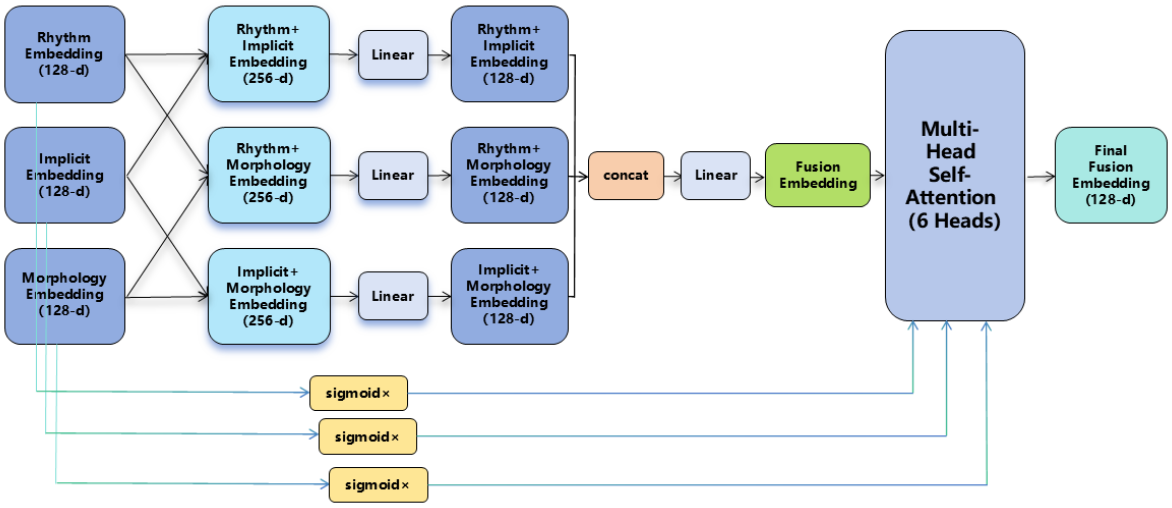

Figure 5. Architecture of the attention-based multi-source feature fusion network.

2.4.1. Multi-Path Gating Mechanism

The input to this module consists of embedding representations generated by three subnetworks, all 128-dimensional vectors, including: (1) Rhythm Embedding, rhythm statistical features based on RR intervals; (2) Morphology Embedding, QRS waveform structural features extracted by multi-scale convolutional ResNet models; (3) Implicit Embedding, deep embedding representations from self-supervised TS-TCC encoders (all modules have been unified to 128 dimensions in this network's input).

We observe that different modal inputs (Rhythm, Morphology, Implicit) exhibit significant differences in expressive capabilities. Equal-weight fusion may mask certain key features. Therefore, this module introduces a multi-path gating mechanism (Multi-Path Gating) to assign learnable selective weights to each original modal pathway before feature interaction and attention processing.

Specifically, for the i-th input pathway, the gating coefficient is calculated as:

$$g_i = \sigma(\mathbf{W}_g \cdot \mathbf{e}_i + \mathbf{b}_g) \tag{2}$$

where $\mathbf{e}_i$ represents the i-th input feature, $\mathbf{W}_g$ and $\mathbf{b}_g$ are learnable parameters, and σ denotes the Sigmoid activation function, compressing gating coefficients to the interval $(0, 1)$.

Each original input is multiplied by its corresponding gating weight, producing weighted features $\mathbf{e}'_i = g_i \cdot \mathbf{e}_i$. These gated features are then fed into the subsequent interaction and attention modules, thereby controlling the influence of each modality on the overall representation. This mechanism enables selective enhancement of informative modalities before feature fusion and attention computation.

2.4.2. Input and Feature Interaction

To achieve information fusion between modalities, we adopt a pairwise interaction strategy on the gated features. The three gated embedding features are concatenated pairwise (Rhythm-Morphology, Rhythm-Implicit, Morphology-Implicit) to form three interaction feature vectors of length 256. Then, each interaction vector is mapped back to 128 dimensions through independent linear transformation layers, thereby avoiding information interference caused by modal redundancy. The formula is as follows:

$$\mathbf{h}_{ij} = \mathbf{W}_{ij} \cdot [\mathbf{e}'_i; \mathbf{e}'_j] + \mathbf{b}_{ij} \tag{3}$$

where $\mathbf{e}'_i$ and $\mathbf{e}'_j$ represent two different gated embedding vectors, $[\cdot;\cdot]$ denotes concatenation, and $\mathbf{W}_{ij}$ and $\mathbf{b}_{ij}$ are learnable parameters.

Finally, the three transformed interaction vectors are concatenated and uniformly mapped to a fusion vector (128 dimensions), providing a structurally rich fusion feature foundation for subsequent attention mechanisms.

2.4.3. Multi-Head Self-Attention Mechanism

To further enhance global associations between features, we introduce a multi-head self-attention mechanism. This module takes the fusion vector (obtained from pairwise interaction of gated features) as input, constructs query

(Q), key (K), and value (V) matrices, and computes attention based on the standard self-attention formula:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where d_k is the dimension scaling factor for the attention head.

This module employs 6 independent attention heads (6-heads), modeling multi-subspace relationships of input features in parallel, effectively enhancing fine-grained modeling capabilities of representations. Outputs from each head are concatenated and linearly transformed to obtain structurally enhanced representation vectors (still 128 dimensions) for transmission to the next stage.

2.4.4. Residual Connection

To alleviate potential gradient vanishing problems in deep networks and maintain the integrity of original inter-modal interaction features, residual connections (Residual Add) are introduced after the attention module output. Specifically, the structurally enhanced representation is element-wise added to the interaction fusion features:

$$\mathbf{h}_{\text{final}} = \mathbf{h}_{\text{att}} + \mathbf{h}_{\text{interact}} \tag{5}$$

This module not only improves network training stability but also preserves key information from the initial interaction process, enabling the model to maintain robustness and expressive diversity during feature integration. The final output vector (128 dimensions) serves as a unified fusion representation transmitted to downstream modules.

2.5. Loss Function Optimization

The proposed framework adopts different loss functions at different training stages. During unsupervised pre-training, NT-Xent loss is used to learn transferable ECG representations from unlabeled signals. During supervised fine-tuning, cross-entropy loss is used for AF classification, and center loss is added to improve feature compactness within each class. This staged design allows the model to benefit from unlabeled ECG data while maintaining a clear supervised objective for the final AF detection task.

2.5.1. Cross-Entropy Loss

Cross-entropy loss is used as the primary supervised classification objective. Since the task is formulated as AF versus non-AF detection, the loss encourages the predicted class probability to match the ground-truth label for each ECG segment. It is defined as

$$L_{\text{CE}} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c), \tag{6}$$

where C is the number of classes, y_c is the ground-truth label indicator, and $\hat{y}_c$ is the predicted probability for class c. In this study, cross-entropy loss provides the main decision boundary for the downstream AF classification task.

2.5.2. Center Loss

Center loss is introduced during supervised fine-tuning to reduce the intra-class variation of learned features. This is useful for AF detection because ECG segments from the same class may still show large waveform and rhythm differences across subjects and datasets. The center loss is defined as

$$L_{\text{center}} = \frac{1}{2}\sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{c}_{y_i}\|_2^2, \tag{7}$$

where $\mathbf{x}_i$ denotes the feature representation of sample i, and $\mathbf{c}_{y_i}$ denotes the feature center of its corresponding class. The supervised fine-tuning objective is therefore written as

$$L_{\text{fine}} = L_{\text{CE}} + \lambda L_{\text{center}}, \tag{8}$$

where λ controls the contribution of center loss. By combining cross-entropy loss and center loss, the model learns both discriminative class boundaries and more compact class-specific feature distributions.

2.5.3. NT-Xent Loss

NT-Xent loss is used only in the unsupervised pre-training stage of the implicit representation branch. Given two augmented views of the same ECG segment as a positive pair, the loss pulls their embeddings closer while pushing embeddings from different samples apart. The loss for a positive pair (i, j) is

$$L_{\text{NT-Xent}} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{k \neq i} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}, \tag{9}$$

where $\mathbf{z}_i$ and $\mathbf{z}_j$ are embedding vectors, $\text{sim}(\cdot)$ denotes cosine similarity, and τ is the temperature parameter. The pre-trained encoder is then transferred to the downstream fusion framework. NT-Xent loss is not jointly optimized with cross-entropy loss or center loss during supervised fine-tuning.

3. Results

3.1. Model Evaluation

Following self-supervised pre-training on SHDB-AF, 200 samples selected from the 1000 labeled CPSC_2025 records were used for supervised fine-tuning. The remaining 800 samples were held out and used only for final testing. On the 800 held-out CPSC_2025 samples, the model achieved an accuracy of 95.6%, a precision of 99.3%, a recall of 91.8%, and an F1-score of 95.4%. The high precision indicates that relatively few non-AF recordings were incorrectly classified as AF, while the recall shows that most AF recordings were successfully detected.

In contrast, on the CinC_2017 test set, the model achieved 94.64% accuracy, 89.70% precision, 81.76% recall, and a corresponding F1-score of 85.51%. The decline in recall suggests that ECG signal patterns are more diverse in this dataset, and there remains room for improvement in balancing detection rate and accuracy.

On the SHDB-AF test set, the model achieved 95.74% accuracy, with precision and recall of 90.74% and 85.57%, respectively, corresponding to an F1-score of 88.13%. These results indicate that while overall accuracy is comparable to that on the CPSC_2025 dataset, due to the atypical and complex nature of AF waveforms, the model still needs to strengthen control over false positive rates while further improving detection efficiency.

Overall, the model demonstrates relatively balanced performance on the CPSC dataset, while performance metrics decline on CinC_2017 and SHDB-AF datasets due to differences in signal characteristics. Future research could employ more robust feature extraction methods or combine data augmentation strategies to further enhance the model's generalization capability across different ECG signal datasets (Table 2).

Table 2. Detailed evaluation metrics on different datasets.

Dataset	Accuracy	Precision	Recall	F1-Score
CPSC (800 samples)	0.956	0.993	0.918	0.954
CinC_2017	0.946	0.897	0.817	0.855
SHDB-AF	0.957	0.907	0.857	0.881

3.2. Ablation Study

When using only cross-entropy loss, the model relies on labeled signal classification to update parameters, enabling rapid learning of class boundaries but struggling to ensure intra-class sample aggregation and inter-class sample separation.

After adding center loss, the model maintains classification accuracy while enhancing intra-class feature consistency by constraining the distance between sample features and class centers, reducing misclassification risk.

When NT-Xent pre-training is introduced before supervised fine-tuning, the encoder learns more transferable ECG representations from unlabeled signals. This improves inter-class separability without jointly optimizing contrastive loss during the supervised fine-tuning stage.

The final staged strategy combines NT-Xent-based unsupervised pre-training with cross-entropy and center loss during supervised fine-tuning. In this setting, cross-entropy provides the main classification objective, center loss enhances intra-class compactness, and NT-Xent pre-training improves the transferability and separability of learned representations. The ablation results on the CinC_2017 dataset are shown in Table 3.

Table 3. Ablation study results on CinC_2017 dataset.

Training Strategy	Acc. (%)	Intra-Class Dist.	Inter-Class Dist.	Boundary Acc. (%)
CE only	85.2	0.62	0.45	78.3
CE + Center	88.7	0.41	0.48	83.5
NT-Xent pretrain + CE	89.5	0.50	0.67	87.2
NT-Xent pretrain + CE + Center	92.3	0.44	0.75	91.0

3.3. Model Comparison

On the same test set, three unsupervised AF detection models each have different emphases across various evaluation metrics, while our model demonstrates more balanced comprehensive performance.

First, the Sample Entropy + K-Means model calculates sample entropy from R-R interval sequences as features, achieving high recall (approximately 0.90) under low noise conditions, but its precision typically drops to the 0.85–0.88 range, indicating that while most AF segments are detected, the number of false positives for sinus rhythm samples remains high.

Second, the Wavelet Packet Decomposition + DBSCAN model utilizes multi-scale energy features combined with density clustering algorithms, maintaining precision exceeding 0.92 in high-noise environments, but its recall is often only 0.80–0.83, reflecting that while pursuing strict elimination of negative class samples, some AF samples are incorrectly classified as sinus rhythm.

Third, the Phase Space Reconstruction Hierarchical Clustering model performs hierarchical clustering based on nonlinear temporal features such as phase space attractor reconstruction and fractal dimensions, providing high overall accuracy (reaching above 0.94), but still faces trade-offs in discriminative precision and detection capability for boundary samples: its precision is approximately 0.89 and recall is approximately 0.86.

Our model achieved an accuracy of 0.946, a precision of 0.897, a recall of 0.817, and an F1-score of 0.855 on the CinC_2017 evaluation set, consistent with the results reported in Table 2. Compared with the traditional unsupervised baselines, the proposed method achieved competitive accuracy and precision, although its recall and F1-score remained lower than those of some comparison methods. These results indicate that further improvement is needed to reduce missed AF detections under cross-dataset distribution shifts. The complete comparison is presented in Table 4.

Table 4. Model comparison results on CinC_2017 dataset.

Model	Recall	Precision	Accuracy	F1-Score
Sample Entropy + K-Means	0.90	0.86	0.92	0.88
Wavelet Packet Decomposition + DBSCAN	0.82	0.93	0.93	0.87
Phase Space Reconstruction Hierarchical Clustering	0.86	0.89	0.94	0.87
Our Model	0.817	0.897	0.946	0.855

4. Discussion

Across datasets, the proposed framework achieves the strongest overall performance on the target-domain CPSC_2025 setting, while recall decreases on CinC_2017 and SHDB-AF. This behavior is consistent with distribution shifts introduced by heterogeneous acquisition devices, sampling rates, and noise characteristics. These results suggest that combining rhythm, morphology, and self-supervised implicit representations improves robustness, but cross-dataset generalization remains a key challenge for short, noisy wearable ECG.

The ablation results indicate that objectives promoting intra-class compactness and inter-class separability contribute to improved discrimination, particularly when annotations are limited. In addition, the attention-guided fusion mechanism provides a principled way to integrate heterogeneous cues and reduce the risk that a single modality dominates the decision under adverse signal conditions.

This study has several limitations. First, the pipeline relies on the stability of R-wave localization and beat segmentation, which may degrade under severe artifacts. Second, the current evaluation focuses on fixed-length recordings and may not fully represent long-term monitoring scenarios. Future work will further study domain adaptation and robustness-enhancing augmentation strategies, and will evaluate the approach under device-level constraints.

5. Conclusions

5.1. Summary

This study presents a multi-source feature fusion framework for AF detection in wearable devices. The framework integrates rhythm features derived from RR interval statistics, morphological features extracted from QRS waveforms through multi-scale convolution, and implicit features learned from TS-TCC self-supervised embeddings. These three feature modalities are combined through multi-head attention and gating mechanisms to achieve cross-modal information synergy.

To address different task requirements, we employ hierarchical wavelet denoising strategies. Strong denoising is applied to highlight rhythm features for the rhythm modeling branch, while mild denoising preserves morphological details for the morphological modeling branch. This approach ensures optimal feature extraction for each modeling branch. The staged training strategy uses NT-Xent loss for self-supervised pre-training, followed by cross-entropy and center loss during supervised fine-tuning.

Experimental validation was conducted on three major datasets including CPSC_2025, CinC_2017, and SHDB-AF. The model achieved accuracies of 95.60%, 94.64%, and 95.74% respectively, with corresponding F1-scores of 95.40%, 85.50%, and 88.10%. These results demonstrate competitive accuracy and precision compared with the evaluated traditional unsupervised baselines, while also indicating room for improvement in recall and F1-score under cross-dataset evaluation. Ablation experiments show that the staged pre-training and fine-tuning strategy significantly improves accuracy to 92.3%, increases average inter-class distance to 0.75, and enhances boundary sample recognition capability.

The proposed approach reduces the dependency on large-scale labeled data that limits traditional algorithms, providing a universal AF detection solution suitable for resource-constrained scenarios such as wearable devices. The technical framework exhibits transferability and can be extended to recognition of various arrhythmia types. The feature modeling approach aligns with clinical diagnostic logic, offering strong interpretability and promising potential for medical applications.

5.2. Future Work

Several research directions will be explored to further improve the framework. For datasets with atypical AF waveforms such as SHDB-AF, generative adversarial networks or temporal data augmentation techniques will be introduced to simulate individual differences and noise interference in real-world scenarios, thereby enhancing model recognition capability for boundary samples. Future work will investigate knowledge distillation and parameter quantization to reduce model size. On-device feasibility will then be evaluated in terms of computational complexity, memory consumption, and inference latency on representative wearable hardware.

Multi-modal data fusion represents another important direction. By combining heart rate variability, accelerometer data, or blood pressure signals, cross-modal fusion models can be constructed to enhance the robustness and clinical credibility of AF detection. The introduction of time-varying and time-invariant features will enable personalization of wearable device AF detection algorithms while improving robustness and mining periodic characteristics of ECG signals. Finally, personalized model training will utilize transfer learning frameworks to fine-tune model parameters based on patient historical data, constructing personalized AF early warning systems that reduce misjudgments caused by individual differences.

Author Contributions

H.W.: conceptualization, methodology, software; Z.M.: data curation, writing—original draft preparation; Y.X.: visualization, investigation; Q.Y.: supervision; D.W.: software, validation; H.T.: writing—reviewing and editing; L.X.: supervision, project administration, funding acquisition, and writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the National Undergraduate Innovation and Entrepreneurship Training Program (Project No. 260235).

Institutional Review Board Statement

Not applicable. This study was based on secondary analysis of de-identified ECG datasets and did not involve new data collection from human participants.

Informed Consent Statement

Not applicable. This study used de-identified secondary ECG datasets and did not involve identifiable personal information or direct recruitment of human participants.

Data Availability Statement

The datasets analyzed in this study are publicly available or available through their corresponding official challenge/database platforms.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

1. Cardiovascular Diseases Fact Sheet. Available online: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed on 20 December 2024).
2. National Center for Cardiovascular Diseases. Report on Cardiovascular Health and Diseases in China 2023: An updated summary. *Biomed. Environ. Sci.* **2024**, *37*, 949–992.
3. Healey, J.S.; Connolly, S.J.; Gold, M.R.; et al. Subclinical atrial fibrillation and the risk of stroke. *N. Engl. J. Med.* **2012**, *366*, 120–129.
4. Shukla, P.K.; Roy, V.; Shukla, P.K.; et al. An advanced EEG motion artifacts eradication algorithm. *Comput. J.* **2023**, *66*, 429–440.
5. Zhang, D.; Wei, G.; Geng, S.; et al. Artificial intelligence algorithm for atrial fibrillation recognition from paper-based electrocardiograms. *J. Pract. Electrocardiol.* **2023**, *32*, 1–7. https://doi.org/10.13308/j.issn.2095-9354.2023.01.001.
6. Acharya, U.R.; Fujita, H.; Lih, O.S.; et al. A survey on ECG analysis. *Biomed. Signal Process. Control* **2018**, *43*, 216–235.
7. Liu, C.; Zhang, X.; Wang, H.; et al. Foundation Model of ECG Diagnosis: Diagnostics and explanations of any form and rhythm on ECG. *Cell Rep. Med.* **2024**, *5*, 100789.
8. Dhingra, L.S.; Aminorroaya, A.; Sangha, V.; et al. Ensemble Deep Learning Algorithm for Structural Heart Disease Screening Using Electrocardiographic Images PRESENT SHD. *J. Am. Coll. Cardiol.* **2025**, *85*, 1302–1313.
9. Duan, J.; Wang, Z.; Ji, Y.; et al. Accurate detection of atrial fibrillation events with R–R intervals from ECG signals. *PLoS ONE* **2022**, *17*, e0271596.
10. Sîngeap, M.S.; Corneanu, L.E.; Prodaniuc, A.; et al. Diagnostic accuracy of wearable ECG devices for atrial fibrillation and ST-segment changes: A systematic review. *Diagnostics* **2025**, *15*, 3162.
11. Lv, J.; Zhang, R.; Zhou, T. Nonstationary Feature-Based Atrial Fibrillation Detection from Single-Lead ECG. Available online: https://www.docin.com/p-4732079905.html (accessed on 15 June 2025).
12. Tateno, K.; Glass, L. Automatic detection of atrial fibrillation using RR intervals. *Med. Biol. Eng. Comput.* **2001**, *39*, 664–671.
13. Mohebbi, M.; Ghassemian, H. A support vector machine approach for atrial fibrillation classification from a short single-lead ECG recording. In Proceedings of the Computing in Cardiology Conference, Rennes, France, 24–27 September 2017; pp. 1–4.
14. Lown, M.; Brown, M.; Brown, C.; et al. Machine learning detection of atrial fibrillation using a wearable heart rate monitor and SVM classifier. *PLoS ONE* **2020**, *15*, e0244341.
15. Izci, E.; Ozdemir, M.A.; Sadighzadeh, R.; et al. Arrhythmia Detection on ECG Signals by Using Empirical Mode Decomposition. In Proceedings of the IEEE Conference on Technologies Applied to Electronics Teaching, Magusa, Cyprus, 28–30 November 2018.
16. Zhang, Y. Energy-Efficient Convolutional Neural Network for Arrhythmia Detection on Portable Devices. Available online: https://wenku.csdn.net/doc/56g0t0j0w7 (accessed on 15 June 2025).
17. Tsutsui, K.; Brimer, S.B.; Ben-Moshe, N.; et al. SHDB-AF: A Japanese Holter ECG database of atrial fibrillation (Version 1.0.1). *Sci. Data* **2025**, *12*, 454. https://doi.org/10.1038/s41597-025-04777-4.
18. Clifford, G.D.; Liu, C.; Moody, B.; et al. AF classification from a short single-lead ECG recording: The PhysioNet/Computing in Cardiology Challenge 2017. In Proceedings of the Computing in Cardiology Conference, Rennes, France, 24–27 September 2017; pp. 1–4.