Progress in
Medical Devices

Zentime

RESEARCH ARTICLE

Cardiac function state recognition model based on bimodal time–frequency representation

Mingzhi Zhang, Piding Li

School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China.

Corresponding author: Piding Li.

Address correspondence to: Piding Li, School of Health Science and Engineering, University of Shanghai for Science and Technology, No. 516 Jungong Road, Yangpu District, Shanghai 200093, China. E-mail: lpdbyusst@163.com.

Received November 28, 2025; Accepted February 27, 2026; Published June 24, 2026

DOI: 10.61189/784716ypyhmm

Abstract

Objective: This study uses dual-modality signals, including phonocardiogram (PCG) and electrocardiogram (ECG), together with machine learning methods to distinguish cardiac function states in subjects. **Methods:** We developed a model based on time–frequency representations. The model includes data preprocessing, a time–frequency conversion module, a feature extraction module, and a feature-fusion classifier module. The system uses complete ensemble empirical mode decomposition with adaptive noise to remove noise from the PCG and applies filters to reduce noise in the ECG. The system extracts Mel-frequency cepstral coefficients from the PCG and uses Fourier synchrosqueezed transform for the ECG. This study also improves VGG16 and ResNet18 as feature extractors by inserting a variant attention mechanism into the feature extraction networks. Finally, the system feeds the feature vector into a support vector machine for classification. **Results:** The dual-modality time–frequency method achieves 95.4% accuracy and 97.4% sensitivity for positive cases on public datasets, demonstrating strong performance in cardiac function classification. **Conclusion:** This research shows that the approach improves both diagnostic accuracy and sensitivity. The system provides valuable support for the preliminary screening of cardiac dysfunction.

Keywords: Multi-modal, Phonocardiogram signal, Electrocardiogram signal, Feature encoding, Heart disease screening

Highlights

- We use two types of cardiac physiological signals together. They complement each other and help improve the final classification accuracy.
- This study converts phonocardiograms and electrocardiograms into time–frequency images, which helps increase the positive detection rate and enables automatic learning of modality-specific features through a neural network.
- This study modifies the baseline model to achieve a more streamlined neural network architecture and incorporates an attention mechanism to better focus on information correlations.

1 INTRODUCTION

Cardiovascular disease remains one of the leading causes of death and illness worldwide, currently accounting for 16% of all deaths [1]. The prevalence of cardiovascular disease in China continues to rise, with an estimated 330 million people now affected [2]. Phonocardiogram (PCG) and electrocardiogram (ECG) serve as critical physiological parameters for evaluating cardiac function. PCG mainly reflect the mechanical and acoustic features of valve movement and blood flow, while ECG records the heart's electrophysiological activity [3, 4].

Current research has widely examined single-modality classification of PCG and ECG [5, 6]. Advances in sensing and signal-

View Online

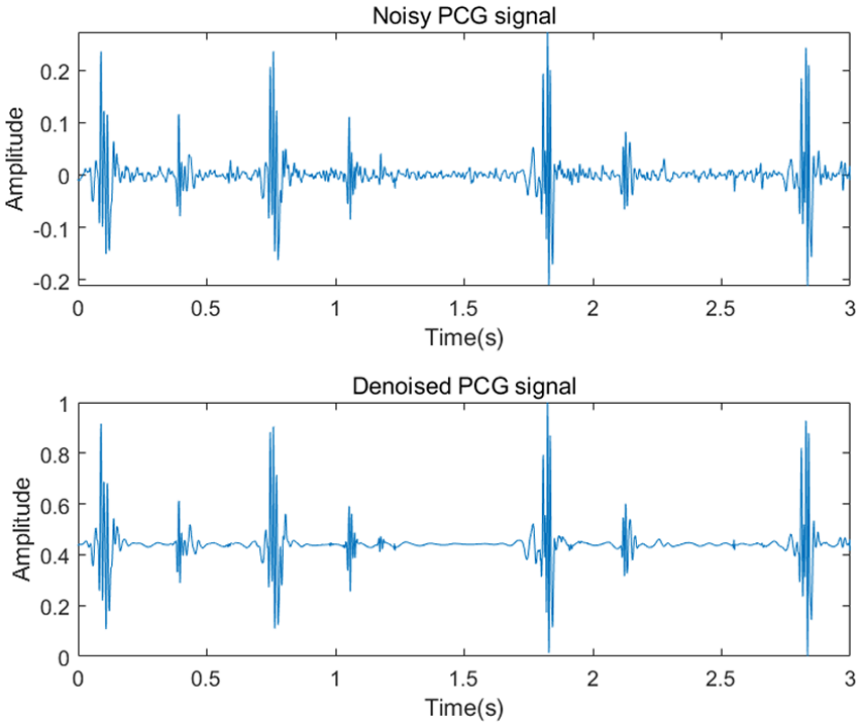

Figure 1. Comparison of PCG signals before and after preprocessing. PCG, phonocardiogram.

processing technologies have made multimodal technology an emerging direction, as the two modalities together provide complementary information [7].

Chakir et al. extracted ten time-domain statistical features from ECG and PCG [8]. Li et al. divided the PCG signal into four frequency bands and used convolutional neural networks and long short-term memory for feature extraction [9]. Sun et al. improved coronary artery disease detection by decoupling ECG and PCG to create coupled signals, and extracted entropy features and recursive graph depth features [10]. Li et al. enhanced the Dempster-Shafer evidence theory to fuse classification results, which improved the accuracy of the final decision [11]. Zhang et al. proposed an end-to-end Co-learning-assisted Progressive Dense Fusion Network with a three-branch interleaved architecture, which showed strong performance on both public and private datasets [12]. Zhu et al. proposed Dual-Scale Deep Residual Network, which uses a dual-scale feature aggregation module to merge low-level features from different scales [13]. Liu et al. used the Vision Transformer to detect coronary heart disease [14]. However, several challenges remain, including low positive detection rates for medical conditions, limited feature representations, and complex network architectures.

This study proposes a cardiac function state recognition method based on dual-modality time–frequency representations. The time–frequency maps provide a comprehensive view of the cardiac signals. A neural network module then automatically extracts features from these maps, and finally, a machine learning model performs classification.

2 MATERIALS AND METHODS

2.1 Overall architecture

To address the aforementioned issues, this study proposes a cardiac function state recognition model based on dual-modality PCG and ECG data.

The proposed model framework consists of four main modules: the signal preprocessing module, the time–frequency information conversion module, the feature vector extraction module, and the feature fusion and classification module.

2.2 Data preprocessing

Given the characteristics of the signals and the presence of noise, this study first performed preprocessing and denoising on the PCG and ECG signals separately.

Noise in PCG signals usually comes from friction sounds caused by a subject's breathing or movements [15]. To denoise the PCG signals, this study adopted the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise proposed in the literature [16]. First, the method was used to extract multi-timescale intrinsic mode function (IMF) components. Next, detrend fluctuation analysis metrics were calculated for these IMF components. Finally, adaptive dual-threshold wavelet analysis was applied to the IMF components based on these metrics to perform denoising and signal reconstruction, producing the cleaned PCG signal. Following preprocessing, the PCG signal was normalized to the 0–1 range.

Figure 1 shows the PCG signals before and after noise reduction. The noise was effectively removed without altering the original waveform shape.

The main sources of noise in ECG signals are baseline drift caused by a subject's respiration and power-line interference from electrical equipment [17]. To remove baseline drift, this study applied a fourth-order Butterworth high-pass filter with a cutoff frequency of 0.5 Hz. To eliminate 60 Hz power-line interference and its 120 Hz harmonic, a second-order infinite impulse response comb filter was used on the ECG signal. **Figure 2** shows the ECG signals before and after noise reduction. The noise was also eliminated.

The total duration of normal sample recordings was 3,770 seconds, while abnormal samples totalled 8,897 seconds. To bal-

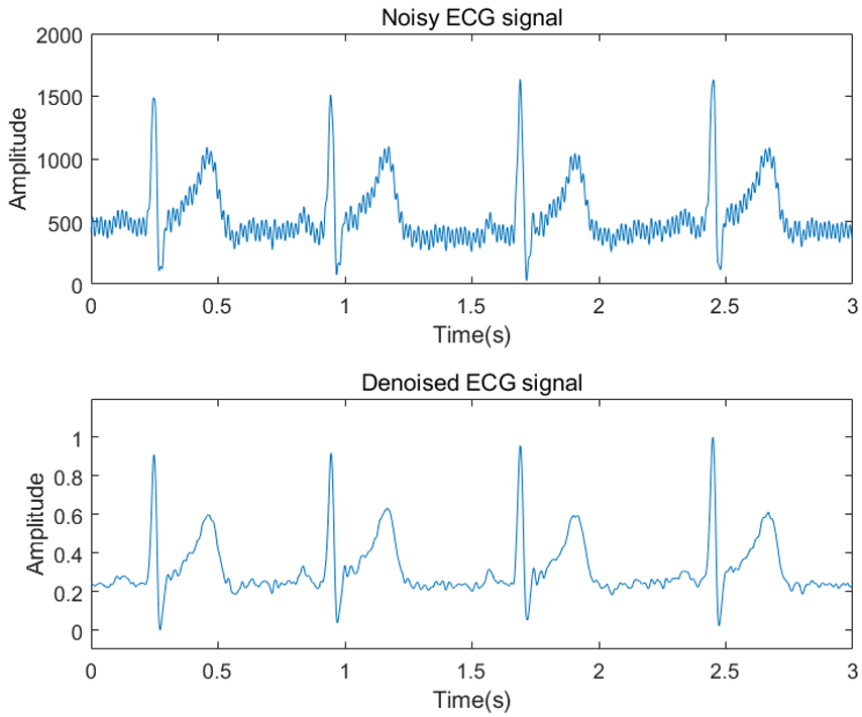

Figure 2. Comparison of ECG signals before and after preprocessing. ECG, electrocardiogram.

ance the number of positive and negative samples, this study used overlapping sliding windows to segment the raw signals. Each window spanned 3 seconds, with a stride of 1.25 seconds for normal samples. After segmentation, there were 2,809 normal segments and 2,786 abnormal segments.

2.3 Time-frequency information conversion module

A spectrogram is a time–frequency representation of a signal, converting a one-dimensional time-domain signal into a two-dimensional image that contains both time and frequency information. This allows for a more comprehensive analysis of non-stationary signals. Compared with one-dimensional time-series signals, time–frequency plots preserve both temporal and frequency details, helping models capture the dynamic evolution of signals more effectively. Unlike traditional spectral analysis, which provides only global frequency information, time–frequency plots show when specific frequencies occur or change, offering greater flexibility and precision.

The time–frequency plot shows time on the horizontal axis and frequency on the vertical axis. Color or brightness represents the signal's power or energy, with brighter colors indicating higher energy at a specific time and frequency. This representation provides neural networks with input that carries rich physical meaning.

2.3.1 Mel-frequency cepstral coefficients (MFCC) transformation of PCG signals

PCG signals are non-stationary, with frequencies that change over time. This study uses MFCC as the time–frequency representation for one-dimensional PCG signals. MFCC has the advantage of simulating the human ear's sensitivity to low frequencies using Mel filter banks, which map the linear spectrum onto the Mel scale.

The logarithmic Mel-spectrogram is created by first filtering a signal through a set of Mel-scale filters, followed by logarithmic compression applied to the resulting power spectrum.

First, the power spectral density of the signal $P(k)$ is calculated:

$$P(k) = \frac{1}{N}\left| X(k) \right|^2 \quad (1)$$

where $X(k)$ denotes the frequency component of the signal in the k band, and N represents the number of sample points.

Next, the Mel filter bank is calculated.

The Mel-scale frequency intervals for the Mel filter bank are:

$$\Delta\phi = (\phi_{max} - \phi_{min})/(M+1) \quad (2)$$

where M denotes the number of Mel filters. The Mel filter bank covers a minimum frequency of ϕ_{min} and a maximum frequency of ϕ_{max}.

Therefore, the centre frequencies of the Mel scale for the Mel filters in Group M are:

$$\phi_c(m) = m\Delta\phi, \quad 0 \le m \le M \quad (3)$$

The formula for converting frequency from hertz to the Mel scale is:

$$\phi = 2595\log_{10}(1 + \frac{f}{700}) \quad (4)$$

The formula for converting frequency from the Mel scale back to hertz is:

$$f = 700(10^{\frac{\phi}{2595}} - 1) \quad (5)$$

The frequency response formula for an equal-height triangular filter bank is:

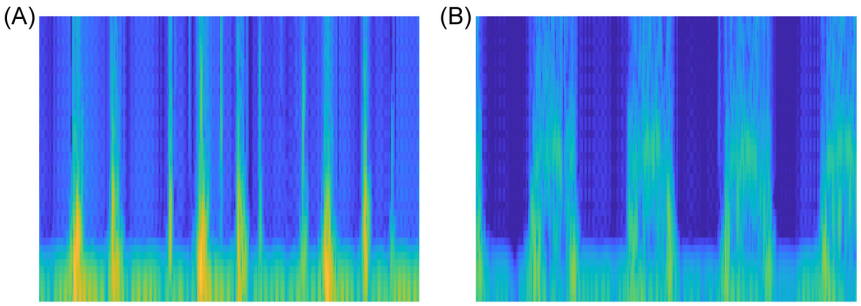

Figure 3. Logarithmic Mel-spectrograms of normal and abnormal PCG. (A) Normal PCG; (B) Abnormal PCG. ECG, electrocardiogram; PCG, phonocardiogram.

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) < k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \quad (6)$$

where $f(m)$ denotes the centre frequency of the m-th filter bank in hertz.

Next, the logarithmic spectral energy distribution of the Mel frequencies for each frame is calculated:

$$S(m) = \log_e\left(\sum_{k=0}^{N-1} P(k) \cdot H_m(k)\right), \quad 0 \le m \le M \quad (7)$$

In this study, the MFCC window length was set to 25 milliseconds to balance temporal and frequency resolution. The overlap length was set to 15 milliseconds to reduce information loss between frames and to smooth time–frequency variations. The number of Mel filters was set to 40, providing higher resolution in the low-frequency range while still covering the high-frequency range. The fast Fourier transform length was set to 512 to balance efficiency and accuracy. Since PCG signals have frequency components between 0 and 500 Hz, the frequency range was set to 0–500 Hz [18].

The MFCCs of normal and abnormal PCG signals are shown in **Figure 3A** and **3B**, respectively.

The spectrogram of normal PCG signals shows higher energy in the low-frequency range and displays more distinct frequency bands.

2.3.2 Fourier synchrosqueezing transform (FSST) transformation of ECG signals

ECG signals are also non-stationary. FSST allows visualization of their frequency information over time. When ECG signals have abrupt changes or transient frequency components, FSST can identify their instantaneous frequency characteristics [19]. By synchronously compressing the results of the short-time Fourier transform (STFT), FSST concentrates dispersed energy toward the instantaneous frequency curve, producing a clearer time–frequency representation.

First, the STFT of the input signal is computed:

$$V(t, f) = \int_{-\infty}^{+\infty} x(\tau) \cdot g(\tau - t) \cdot e^{-j2\pi f\tau} d\tau \quad (8)$$

where f represents frequency, τ represents the time parameter, and $g(\tau-t)$ represents the window function.

The STFT is limited by the principle of time–frequency uncertainty. Signal energy spreads in the time–frequency plane due to the window function, causing frequency blurring. To estimate instantaneous frequency, the partial derivative with respect to time is calculated:

$$\hat{\omega}(t, f) = \mathrm{Re}\left(\frac{1}{2\pi i} \times \frac{\partial_t(V(t, f))}{V(t, f)}\right) \quad (9)$$

The coefficients $V(t,f)$ obtained from the STFT are first mapped along the frequency axis to the estimated instantaneous frequency $\hat{\omega}(t, f)$. Synchronous compression is then applied to redistribute the energy, producing the amplitude of the FSST:

$$T(t, \omega) = \frac{1}{g(0)} \int_{-\infty}^{+\infty} V(t, f)\,\delta(\omega - \hat{\omega}(t, f))\,dt \quad (10)$$

where ω represents the frequency after synchronous compression, $g(0)$ denotes the window function at time zero, and δ represents the Dirac delta function.

The FSST tracings for normal and abnormal ECG signals are shown in **Figure 4A** and **4B** respectively.

Normal ECG signals show cleaner and more concentrated energy in the FSST, whereas abnormal ECG signals display chaotic and dispersed energy across the spectrum.

In this study, the FSST used a Kaiser window function with a length of 64. Since ECG signals contain frequency components between 0 and 250 Hz, the frequency range was set to 0–250 Hz accordingly [14].

2.4 Feature vector extraction module

2.4.1 PCG feature vector extraction module

This study uses VGG16 as the baseline model for fine-tuning and modification. The network architecture and parameters are shown in **Table 1** and **Figure 5**.

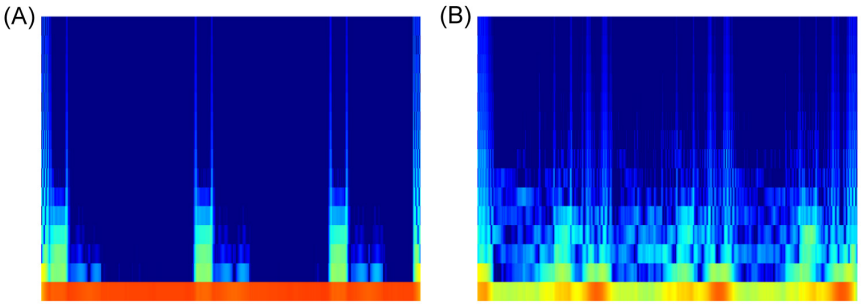

Figure 4. FSST plots of normal and abnormal ECG signals. (A) Normal ECG signal; (B) Abnormal ECG signal. ECG, electrocardiogram; PCG, phonocardiogram.

Table 1. Parameters of the PCG feature vector extraction module

Layer	Structure
Conv	3×3, Cin=Cout, S=1
Conv1	3×3, Cin=3, Cout=16, S=1
Conv2	3×3, Cin=16, Cout=32, S=1
Conv3	3×3, Cin=32, Cout=64, S=1
Conv4	3×3, Cin=64, Cout=256, S=1
Conv5	3×3, Cin=256, Cout=512, S=1
Dropout1	p=0.1
Dropout2	p=0.5
MaxPool	2×2, S=2
DA Module	3×3, d_model=512, H=8, W=8
Linear	In=512×4×4, Out=64
Fully connected	In=64, Out=2

Note: PCG, phonocardiogram; Cin, the number of input channels for the convolutional layer; 3×3, the size of the convolutional kernel; Cout, the number of output channels; p, the dropout probability for the Dropout layer; S, stride; d_model, the number of feature channels; H, the height of the feature map; W, width; In, the input dimension size for the linear layer; Out, the output dimension size; DA Module, Dual Relation-Aware Attention Network Module; MaxPool, max pooling layer; Conv, convolutional layer; Linear, linear layer; Fully Connected, fully connected layer; Dropout, dropout layer.

Additionally, this study incorporates the Dual Relation-Aware Attention Network Module (DA Module) to improve classification accuracy [20]. The architecture of the DA Module is shown in **Figure 6**.

The DA Module contains three components: the Position Attention Module, the Channel Attention Module, and the additive fusion layer. Both the input size and the output size are C×H×W.

The structures of the Position Attention Module and Channel Attention Module are shown in **Figure 7A** and **7B** respectively.

Within the Position Attention Module, the input feature map A first passes through a convolutional layer to extract features. The module then replicates these features into three components: Query, Key, and Value. Each component is reshaped before the positional attention calculation begins. After reshaping, the Query feature map B is transposed to size HW×C, while the Key feature map C is reshaped to size C×HW. The attention scores between these two maps are then computed, reducing the feature map size to HW×HW. The module multiplies this encoded positional attention by the Value feature map *D*, and finally reshapes the result back to C×H×W. The computational process is as follows:

$$K = \text{reshape}(D \cdot \text{Softmax}(B^T \cdot C)) \quad (11)$$

where $B \in \mathbb{R}^{C \times HW}$, $C \in \mathbb{R}^{C \times HW}$, $D \in \mathbb{R}^{C \times HW}$, $K \in \mathbb{R}^{C \times H \times W}$.

After obtaining the feature map K modulated by the positional attention mechanism, the module multiplies it by the weighting coefficient α. It then fuses this result with the original input feature map A through addition:

$$E = \alpha K + A \quad (12)$$

where α is set to 0.5.

Similarly, within the Channel Attention Module, the input feature map A is first reshaped to C×HW. The module then multiplies A by its transpose A^T, producing a matrix of size C×C. After that, the feature is normalised using the Softmax function to generate attention scores between channels. These attention scores are then multiplied by A and finally reshaped back to C×H×W. The computational process is as follows:

$$U = \text{reshape}(\text{Softmax}(A \cdot A^T) \cdot A) \quad (13)$$

where $A \in \mathbb{R}^{C \times HW}$ and $U \in \mathbb{R}^{C \times H \times W}$.

After obtaining the feature map U modulated by the channel attention mechanism, the module multiplies it by the weighting coefficient β and fuses it with the input feature map A through summation:

$$E = \beta U + A \quad (14)$$

where β is set to 0.2.

The outputs from the two attention modules are fused through summation, enriching the spatial contextual information and strengthening the discriminative power of the channel features.

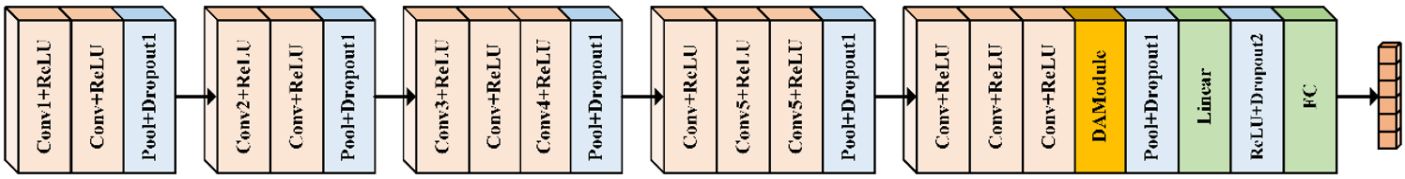

Figure 5. Backbone architecture of the PCG feature vector extraction module. PCG, phonocardiogram; Conv, convolutional layer; ReLU, Linear rectification function; Pool, pooling layer; Dropout, dropout layer; DA Module, Dual Relation-Aware Attention Network Module; Linear, linear layer; FC, fully connected layer.

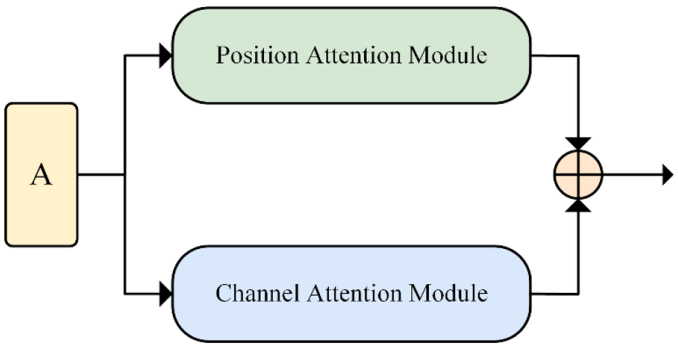

Figure 6. DA Module structural diagram. DA Module, Dual Relation-Aware Attention Network Module.

2.4.2 ECG feature vector extraction module

This study uses ResNet18 as the baseline model for fine-tuning and modification. The network architecture and parameters are shown in **Table 2** and **Figure 8**.

Here, SC represents the residual connections, DSConv represents depthwise separable convolutions, and MidConv denotes the convolutional layers where the input and output dimensions are identical.

To reduce the number of parameters and computational load, this study replaces the convolutional layers within residual blocks with depthwise separable convolutions. Depthwise separable convolutions decompose standard convolution operations into two lighter, decoupled steps: depthwise convolution and pointwise convolution. The depthwise convolution uses fixed-size kernels to independently convolve each channel and stack the resulting feature maps. The subsequent pointwise convolution employs a 1×1 kernel to aggregate information from the depthwise stage, enabling linear combinations across channels. To further enhance inter-channel information fusion, a MidConv layer is introduced to strengthen feature interactions among different channels.

Additionally, this study incorporates the Selective Kernel Module (SK Module) to improve classification accuracy [21]. The architecture of the SK Module is shown in **Figure 9**.

The SK Module includes three components: Split, Fuse, and Select.

In the Split section, two convolutional kernels of different sizes are first applied to the input feature map X, producing $\tilde{U}$ and $\hat{U}$, respectively. These two outputs are then added together to generate U. The computational process is as follows:

$$\tilde{U} = \tilde{F}_{3\times3}(X) \tag{15}$$

$$\hat{U} = \tilde{F}_{5\times5}(X) \tag{16}$$

$$U = \tilde{U} + \hat{U} \tag{17}$$

where $X,\tilde{U},\hat{U} \in \mathbb{R}^{H\times W\times C}$.

The Fuse component uses global average pooling to extract key information from the overall context:

$$S = F_{gp}(U) \tag{18}$$

where $S \in \mathbb{R}^{1\times1\times C}$.

Subsequently, a fully connected layer is used to compress the key information S into a compact representation Z:

$$Z = \mathrm{ReLU}(\mathrm{BatchNorm}(F_{fc}(S))) \tag{19}$$

where $Z \in \mathbb{R}^{1\times1\times d}$ and $d<C$.

The Select component then processes Z usingtwo fully connected layers of size $d\times C$. The resulting outputs are concatenated and normalized via the softmax function to generate the weights $W \in \mathbb{R}^{2\times1\times C}$:

$$W_a = \tilde{F}_{fc}(Z) \tag{20}$$

$$W_b = \hat{F}_{fc}(Z) \tag{21}$$

$$W = \mathrm{softmax}(\mathrm{Concat}(W_a, W_b)) \tag{22}$$

where $W_a, W_b \in \mathbb{R}^{1\times1\times C}$.

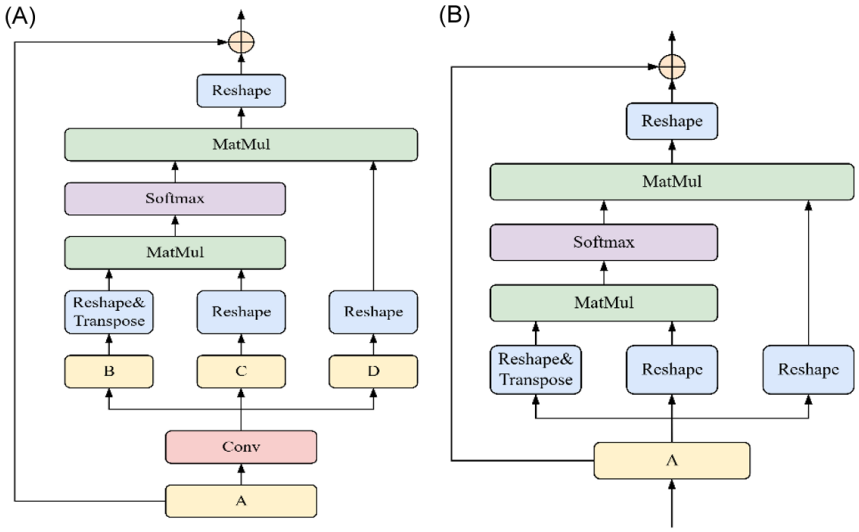

Figure 7. Submodule architecture of the DA Module. (A) Architecture of the position attention module; (B) Architecture of the channel attention module. DA Module, Dual Relation-Aware Attention Network Module.

Table 2. Parameters of the ECG feature vector extraction module

Layer	Structure
Conv	7×7, Cin=3, Cout=64, S=2
DSConv1	3×3, Cin=Cout=64, S=1
DSConv2	3×3, Cin=Cout=128, S=1
DSConv3	3×3, Cin=Cout=256, S=1
DSConv4	3×3, Cin=Cout=512, S=1
MidConv	3×3, Cin=Cout
Dropout1	p=0.4
Dropout2	p=0.7
Pool	AdaptiveAvgPool (1, 1)
SK Module	Kernels=[1, 3, 5, 7], Reduction=8, L=32
Linear	In=512, Out=64
Fully connected	In=64, Out=2

Note: ECG, electrocardiogram; Conv, convolutional layer; Cin, the number of input channels for the convolutional layer; Cout, number of output channels; S, stride; DSConv, depthwise separable convolution; MidConv, intermediate convolution; Dropout, dropout layer; Dropout, dropout layer; p, the dropout probability for the Dropout layer; AdaptiveAvgPool, adaptive average pooling; (1, 1), pooling the feature map to a spatial size of (1, 1); SK Module, Selective Kernel Module; Kernels, the convolutional kernel sizes used in the SK Module; Reduction, the compression ratio; L, the minimum dimension.

Finally, the first dimension is multiplied by weight W and applied to $\tilde{U}$, while the second dimension is multiplied by weight W and applied to $\hat{U}$. This process produces the feature map V, which is modulated by the attention mechanism with selectable kernels:

$$V = W[0] \odot \tilde{U} + W[1] \odot \hat{U} \quad (23)$$

where $Z \in \mathbb{R}^{H \times W \times C}$.

2.5 Feature fusion and recognition

This study employed a feature vector extraction module to obtain 64-dimensional features from each modality. These features were then concatenated to form a 128-dimensional feature vector. Finally, classification was performed using a Support Vector Machine (SVM) model.

3 EXPERIMENTAL SETUP

3.1 Experimental data

To validate the effectiveness of the proposed method, data from the publicly available Physio+Net/Computing in Cardiology Challenge 2016 dataset were used. The *training*-a subset of this dataset contains 409 samples. After excluding samples without paired ECG recordings and those officially annotated as unclassifiable due to excessive noise, 388 samples remained. Each sample includes synchronously recorded PCG and ECG signals and is labeled as either normal or abnormal. The sampling frequency is 2,000 Hz. Statistical analysis indicates that the sample durations in the *training*-a set range from 9 to 36 seconds. The 388 samples were divided into training and test sets at a ratio of 9:1. To prevent data leakage, segments derived from the same recording were assigned exclusively to either the training set or the test set. To better reflect real-world conditions, record-level classification results were used instead of segment-level results. Specifically, record-level predictions were obtained by averaging the segment-level output probabilities.

3.2 Experimental setup

The model was trained on a system equipped with a single NVIDIA RTX 4060 GPU (16 GB VRAM). The implementation was based on the PyTorch deep learning framework (version 2.5.1). During training, the Adam optimizer was employed for 60 epochs with a learning rate of 1×10^{-5}. The batch size was set to 32, and the cross-entropy loss function was used.

3.3 Evaluation criteria

This study uses accuracy (Acc), sensitivity (Sen), specificity (Spe), and F1 score to evaluate model performance. The definitions of these four metrics are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (24)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (25)$$

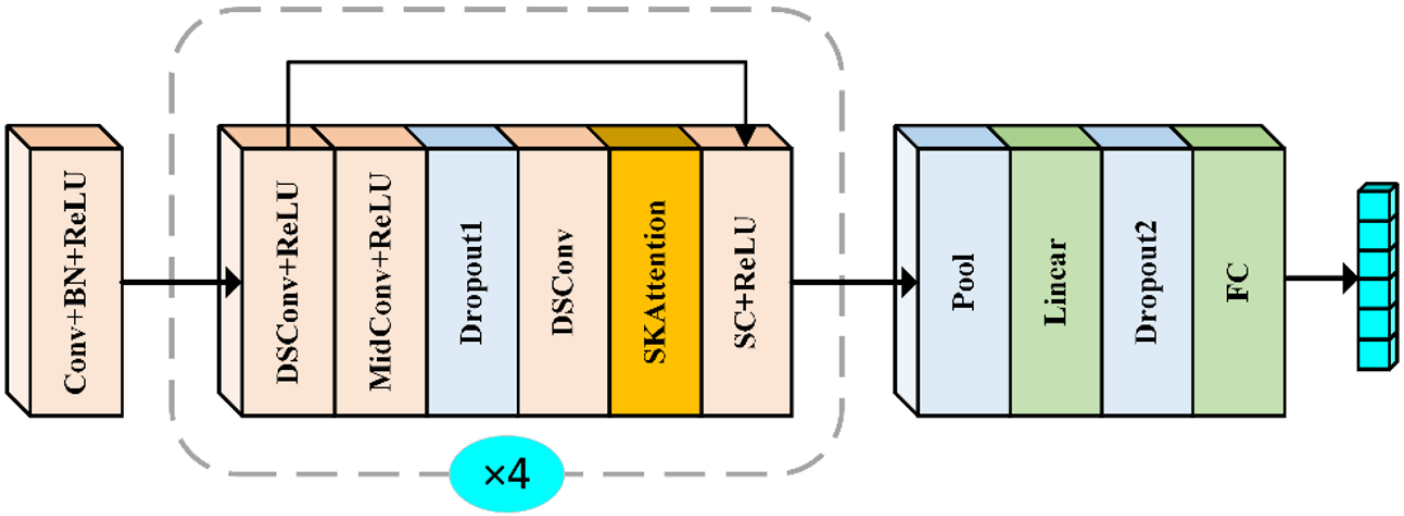

Figure 8. Backbone architecture of the ECG feature vector extraction module. ECG, electrocardiogram; Conv, convolutional layer; BN, batch normalization; ReLU, Linear rectification function; MidConv, intermediate convolution; Dropout, dropout layer; DSConv, depthwise separable convolutions; SK Attention, Selective Kernel Module; SC, Skip Connection; Linear, linear layer; Pool, pooling layer; FC, fully connected layer.

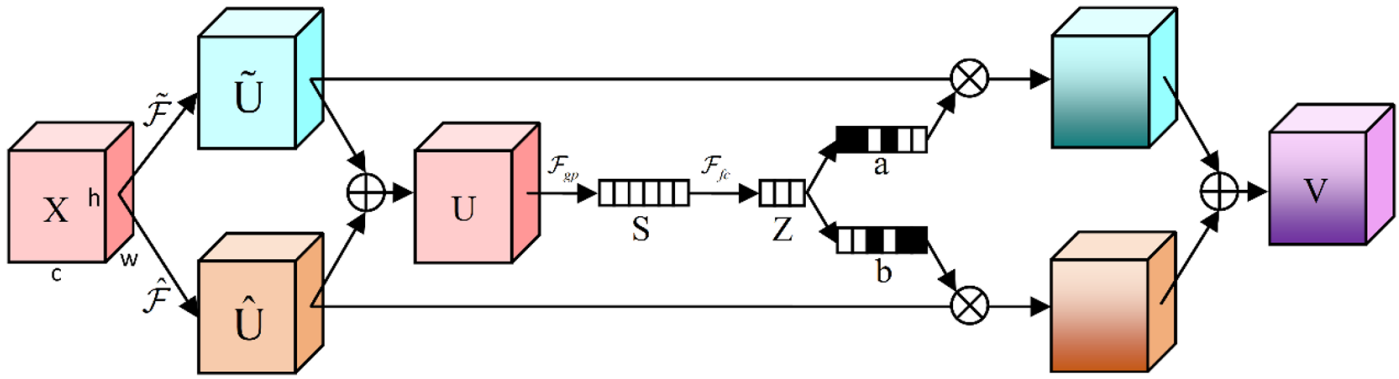

Figure 9. Structure of the SK Module. SK Module, Selective Kernel Module.

$$Specificity = \frac{TN}{TN + FP} \quad (26)$$

$$F1 = 2 \times \frac{Sen \times Pre}{Sen + Pre} \quad (27)$$

where *TP*, *FP*, *TN*, and *FN* represent the number of true positives, false positives, true negatives, and false negatives in the confusion matrix, respectively.

In binary classification, sensitivity measures the proportion of actual positive samples correctly predicted as positive. Higher sensitivity indicates a stronger ability to identify positive samples, resulting in fewer missed detections. Specificity measures the proportion of actual negative samples correctly predicted as negative. Higher specificity indicates a stronger ability to exclude negative examples, resulting in fewer false positives. Precision measures the proportion of true positives among all samples predicted as positive, reflecting the accuracy of the model's predictions. The F1 score is the harmonic mean of sensitivity and precision, providing a balanced metric for overall model performance.

4 RESULTS AND DISCUSSION

As shown in **Table 3**, the improved PCG model achieved higher Acc, Sen, Spe, and F1 score. The PCG model with the Axial Attention Module achieved the highest Spe. However, its overall Acc remained lower than that of the model with the DA Module, which reached 85.3%. As shown in **Table 4**, the improved ECG model incorporating the SK Module achieved the highest Acc, Sen, and F1 scores. **Table 5** presents a comparative assessment of dual-modality versus single-modality results. The PCG and ECG dual-modality network outperformed the single-modality approach across all four metrics, with particularly strong complementary improvement in specificity. While time–frequency images effectively detect abnor-

Table 3. Evaluation metrics for the baseline single-modality PCG model and models with different modules

	Acc	Sen	Spe	F1
Baseline VGG16	0.827	0.890	0.681	0.878
Benchmark	0.840	0.901	0.698	0.888
Benchmark+CBAM	0.822	0.879	0.690	0.874
Benchmark+AA	0.802	0.790	0.828	0.848
Benchmark+DA	0.853	0.901	0.741	0.896

Note: PCG, phonocardiogram; Acc, accuracy; Sen, sensitivity; Spe, specificity; F1, F1 score; VGG16, Visual Geometry Group 16-layer network; CBAM, Convolutional Block Attention Module; AA, Axial Attention; DA, Dual Relation-Aware Attention Network.

Table 4. Evaluation metrics for the baseline single-modality ECG model and models with different modules

	Acc	Sen	Spe	F1
Baseline ResNet18	0.820	0.820	0.819	0.864
Benchmark	0.840	0.853	0.810	0.882
Benchmark+CBAM	0.796	0.846	0.681	0.853
Benchmark+AA	0.874	0.879	0.862	0.907
Benchmark+SK	0.876	0.930	0.750	0.913

Note: ECG, electrocardiogram; Acc, accuracy; Sen, sensitivity; Spe, specificity; F1, F1 score; ResNet18, Residual Network 18-layer; CBAM, Convolutional Block Attention Module; AA, Axial Attention; SK, Selective Kernel.

Table 5. Evaluation metrics for single-modal and dual-modal classification

	Acc	Sen	Spe	F1
PCG-Only	0.853	0.901	0.741	0.896
ECG-Only	0.876	0.930	0.750	0.913
PCG+ECG	0.954	0.974	0.905	0.967

Note: Acc, accuracy; Sen, sensitivity; Spe, specificity; F1, F1 score; PCG, phonocardiogram; ECG, electrocardiogram.

mal cardiac function, the dual-modality approach increases the model's confidence in negative predictions and reduces false alarms. The "Benchmark" in **Tables 3** and **4** refers to the backbone network without any additional modules introduced.

Table 6 compares different classifiers. The decision tree model achieved the highest specificity, but its overall accuracy was relatively low. The SVM achieved the highest Acc, Sen, and F1 score.

Table 7 compares metrics across studies. This research achieved an Acc of 95.4% and a Sen of 97.4%, outperforming other studies. However, there is still room to improve Spe and the F1 score.

5 CONCLUSION

This study proposed a cardiac function state recognition model based on dual-modality time–frequency representations

Table 6. Evaluation metrics for the ablation experiments of the replacement classifier

	Acc	Sen	Spe	F1
KNN	0.915	0.960	0.810	0.941
DT	0.879	0.864	0.914	0.909
GBDT	0.943	0.971	0.879	0.960
XGBoost	0.951	0.974	0.897	0.965
SVM	0.954	0.974	0.905	0.967

Note: Acc, accuracy; Sen, sensitivity; Spe, specificity; F1, F1 score; KNN, K-Nearest Neighbor; DT, Decision Tree; GBDT, Gradient Boosting Decision Tree; XGBoost, eXtreme Gradient Boosting; SVM, Support Vector Machine.

Table 7. Evaluation metrics for comparative experiments

	Acc	Sen	Spe	F1
Li P et al. [9]	0.873	0.903	0.845	0.874
Zhang et al. [12]	0.946	0.949	0.940	0.974
Li J et al. [11]	0.864	0.850	0.931	0.911
Hettiarachchi et al. [22]	0.904	0.947	0.750	0.939
Chakir et al. [8]	0.925	0.923	0.929	0.941
Approach in this study	0.954	0.974	0.905	0.967

Note: Acc, accuracy; Sen, sensitivity; Spe, specificity; F1, F1 score.

of PCG and ECG signals. First, both PCG and ECG signals underwent preprocessing for noise reduction and segmentation. Next, the one-dimensional time-series signals were converted into time–frequency plots to better capture pathological information. For these two-dimensional time–frequency plots, distinct automatic feature extraction neural networks were designed. Accuracy and sensitivity were improved by refining the baseline model and integrating attention modules. Since minimizing missed diagnoses is crucial in disease screening, this study prioritized high sensitivity. Finally, the 64-dimensional feature vectors from each modality were concatenated and input into an SVM, achieving a classification accuracy of 95.4% and a sensitivity of 97.4%.

Compared with state-of-the-art methods, the approach developed in this study achieved higher Acc and Sen on public datasets. In future work, this research will focus on refining the methodology and models to improve their generalization capabilities, allowing them to be applied in scenarios with missing modalities.

DECLARATIONS

Author contributions

Mingzhi Zhang was responsible for data processing, time–frequency conversion, model construction, and manuscript writing. Piding Li made important contributions to data analysis.

Funding

This research received no external funding.

Data availability

This study utilised only publicly available datasets.

Ethics approval and consent to participate

Not applicable.

Consent for publication

The authors agree to the publication of this manuscript in both print and electronic formats in ZENTIME and grant the publisher permission to make necessary editorial and formatting revisions.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors confirm that this manuscript is original, unpublished, and not under consideration elsewhere. All authors have approved the manuscript and confirm adherence to ethical guidelines, including disclosure of potential conflicts of interest.

REFERENCES

[1] World Health Organization. WHO reveals leading causes of death and disability worldwide: 2000-2019 [Internet]. Geneva (Switzerland): WHO; 2020 [updated 2020 Dec 9; cited 2025 Nov 20]. Available from: https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019

[2] National Center for Cardiovascular Diseases, The Writing Committee of the Report on Cardiovascular Health and Diseases in China. Report on cardiovascular health and diseases in China 2023: An updated summary. Chin Circ J. 2024 Jul 27;39(7):625-660. https://doi.org/10.3969/j.issn.1000-3614.2024.07.001

[3] Karhade J, Dash S, Ghosh SK, Dash DK, Tripathy RK. Time–frequency-domain deep learning framework for the automated detection of heart valve disorders using PCG signals. IEEE Trans Instrum Meas. 2022 Mar 29;71:1-11. https://doi.org/10.1109/TIM.2022.3163156

[4] Houssein EH, Ibrahim IE, Neggaz N, Hassaballah M, Wazery YM. An efficient ECG arrhythmia classification method based on manta ray foraging optimization. Expert Syst Appl. 2021 Nov 1;181:115131. https://doi.org/10.1016/j.eswa.2021.115131

[5] Liu C, Springer D, Li Q, Moody B, Juan RA, Chorro FJ, et al. An open access database for the evaluation of heart sound algorithms. Physiol Meas. 2016 Dec;37(12):2181. https://doi.org/10.1088/0967-3334/37/12/2181

[6] Han H, Huang X, Chang H, Fan H, Chen P, Chen J. Review of self-supervised learning methods in field of ECG. J Front Comput Sci Technol. 2024;18(7):1683-1704. https://doi.org/10.3778/j.issn.1673-9418.2310043

[7] Cosoli G, Poli A, Scalise L, Spinsante S. Measurement of multimodal physiological signals for stimulation detection by wearable devices. Measurement. 2021 Nov;184:109966. https://doi.org/10.1016/j.measurement.2021.109966

[8] Chakir F, Jilbab A, Nacir C, Hammouch A. Recognition of cardiac abnormalities from synchronized ECG and PCG signals. Phys Eng Sci Med. 2020 Jun;43(2):673-677. https://doi.org/10.1007/s13246-020-00875-2

[9] Li P, Hu Y, Liu ZP. Prediction of cardiovascular diseases by integrating multi-modal features with machine learning methods. Biomed Signal Process Control. 2021 Apr;66:102474. https://doi.org/10.1016/j.bspc.2021.102474

[10] Sun C, Liu X, Liu C, Wang X, Liu Y, Zhao S, et al. Enhanced CAD detection using novel multi-modal learning: Integration of ECG, PCG, and coupling signals. Basel. 2024 Oct 30;11(11):1093. https://doi.org/10.3390/bioengineering11111093

[11] Li J, Ke L, Du Q, Chen X, Ding X. Multi-modal cardiac function signals classification algorithm based on improved D-S evidence theory. Biomed Signal Process Control. 2022 Jan;71(Part A):103078. https://doi.org/10.1016/j.bspc.2021.103078

[12] Zhang H, Zhang P, Lin F, Chao L, Wang Z, Ma F, et al. Co-learning-assisted progressive dense fusion network for cardiovascular disease detection using ECG and PCG signals. Expert Syst Appl. 2024 Mar 15;238(Part F):122144. https://doi.org/10.1016/j.eswa.2023.122144

[13] Zhu J, Liu H, Liu X, Chen C, Shu M. Cardiovascular disease detection based on deep learning and multi-modal data fusion. Biomed Signal Process Control. 2025 Jan;99:106882. https://doi.org/10.1016/j.bspc.2024.106882

[14] Liu X, You L, Lv C, Chen M, Wei L, Zheng Y, et al. Integrating ECG and PCG signals through a Dual-Modal ViT for coronary artery disease detection. IEEE Journal of Biomedical and Health Informatics 2025. https://doi.org/10.1109/JBHI.2025.3589257

[15] Xu W, Yu K, Ye J, Li H, Chen J, Yin F, et al. Automatic pediatric congenital heart disease classification based on heart sound signal. Artif Intell Med. 2022 Apr;126:102257. https://doi.org/10.1016/j.artmed.2022.102257

[16] Lu G, Tang T, Qi J, Wang Y, Zhao Y. Heart sound denoising based on CEEMDAN and wavelet analysis with adaptive dual thresholds. J Nanjing Univ Posts Telecommun (Nat Sci). 2025 Aug;45(4):36-47. https://doi.org/10.14132/j.cnki.1673-5439.2025.04.005

[17] Kabir MA, Shahnaz C. Denoising of ECG signals based on noise reduction algorithms in EMD and wavelet domains. Biomed Signal Process Control. 2012 Sep;7(5):481-489. https://doi.org/10.1016/j.bspc.2011.11.003

[18] Suryady Z, Rahman AWA, Oktarina D. Human emotion classification based on phonocardiography signals (PCG). AIP Conf Proc. 2024 Feb 7;2729(1):020002. https://doi.org/10.1063/5.0168132

[19] Daubechies I, Lu J, Wu HT. Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool. Appl

Comput Harmon A. 2011 Mar;30(2):243-261. https://doi.org/10.1016/j.acha.2010.08.002

[20] Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, et al. Dual attention network for scene segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. https://doi.org/10.1109/cvpr.2019.00326

[21] Li X, Wang W, Hu X, Yang J. Selective Kernel Networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 June 15-20; Long Beach, CA, USA. IEEE. p. 510-519. https://doi.org/10.1109/CVPR.2019.00060

[22] Hettiarachchi R, Haputhanthri U, Herath K, Kariyawasam H, Munasinghe S, Wickramasinghe K, et al. A novel transfer learning-based approach for screening pre-existing heart diseases using synchronized ECG signals and heart sounds. 2021 IEEE International Symposium on Circuits and Systems (ISCAS); 2021 May 22-28; Daegu, Korea. IEEE; 2021. p. 1-5. https://doi.org/10.1109/ISCAS51556.2021.9401093