Systems and Soft Computing 7 (2025) 200391

Contents lists available at ScienceDirect

Systems and Soft Computing

journal homepage: www.elsevier.com/locate/soft-computing-letters

A hybrid neural network with multistage feature fusion for detecting heart failure and murmurs from time–frequency representations of phonocardiograms

Mahmoud Fakhry [a,*], Ascensión Gallardo-Antolín [b]

[a] CEIEC, Universidad Francisco de Vitoria, Ctra.M-515 Pozuelo-Majadahonda Km.1, 800, Pozuelo de Alarcón, Madrid, 28223, Spain
[b] Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Avenida de la Universidad, 30, Leganés (Madrid), 28911, Spain

ARTICLE INFO

Keywords:
Cardiovascular diseases
Phonocardiogram recordings
Adaptive multiresolution analysis
Hybrid neural networks
Multistage feature fusion

ABSTRACT

Heart failure produces abnormal sounds and murmurs due to weakened cardiac function and turbulent blood flow. This study presents a hybrid neural network model with interior multistage feature fusion to detect heart pathologies using the time–frequency analysis of phonocardiogram (PCG) recordings. The model combines convolutional neural networks with long short-term memory layers in a unique architecture to efficiently capture the spectro-temporal dependencies at multiple cascaded network stages. Moreover, a fusion mechanism is used to aggregate internal features from multiple stages to enhance pattern modeling. We investigated various time–frequency representations of PCG signals to extract relevant features for model training and evaluation. These representations were derived using multiresolution analysis (MRA) via the short-time Fourier transform or the continuous wavelet transform. Additionally, we examined representations obtained through adaptive multiresolution analysis (AMRA) by employing the Hilbert–Huang transform based on empirical mode decomposition, variational mode decomposition, or empirical wavelet transform. The classification performance of the model was evaluated using two separate datasets, showing that the fusion strategy increases the accuracy and that MRA is superior to AMRA, achieving a classification accuracy of 90.20% for the detection of heart murmurs. Compared with MRA, AMRA demonstrated high adaptability, achieving an accuracy of 99.30% in distinguishing five heart valvular conditions.

1. Introduction

Cardiovascular diseases (CVDs) are a significant global health concern and are among the leading causes of morbidity and mortality worldwide [1]. They comprise various diseases that affect the heart and blood vessels, such as rheumatic and coronary heart diseases. The search for medical attention in the automatic detection of CVDs is essential for heart health prevention and care. Consequently, automated tools are required to replace traditional human-based tools for monitoring and detecting CVDs.

The human heart comprises four chambers, two of which are the atria that form the upper part of the heart [2]. The lower part of the heart consists of two chambers called the ventricles. Deoxygenated blood arrives at the right atrium, and oxygenated blood returns to the left atrium. The mitral and tricuspid valves are located between the left atrium and ventricle and between the right atrium and ventricle, respectively. The left ventricle pumps blood to the aorta through the aortic valve and distributes it to all the organs in the body. Blood leaves the right ventricle through the pulmonary artery via the pulmonary valve and enters the lungs. The valves open and close as the heart muscle contracts and relaxes, allowing blood to flow alternately into the ventricles and atria.

Cardiologists examine heart health by hearing the sound of the heart [3]. This conventional examination strategy is time-consuming and requires extensive experience over several years. The need for an accurate diagnosis of heart problems has sparked the automatic analysis of heart sound signals. Phonocardiograms (PCG) have distinct advantages over electrocardiograms (ECG) and photoplethysmograms (PPG) because they record the acoustic properties of the heart [4,5]. PCG signals provide information on mechanical events within the heart, including valve opening and closing, blood flow patterns, and cardiac chamber functions. The cardiac cycle consists of two phases: systole, during which the ventricles contract to pump out blood, and diastole, during which they relax to fill with blood. For a healthy heart, this mechanical activity produces two audible heartbeats separated by two silent intervals: the first heartbeat (S_1) at the beginning of systole and the second heartbeat (S_2) at the beginning of diastole, as shown

* Corresponding author.
E-mail address: mahmoud.fakhry@ufv.es (M. Fakhry).

https://doi.org/10.1016/j.sasc.2025.200391
Received 14 April 2025; Received in revised form 24 August 2025; Accepted 16 September 2025
Available online 23 September 2025
2772-9419/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

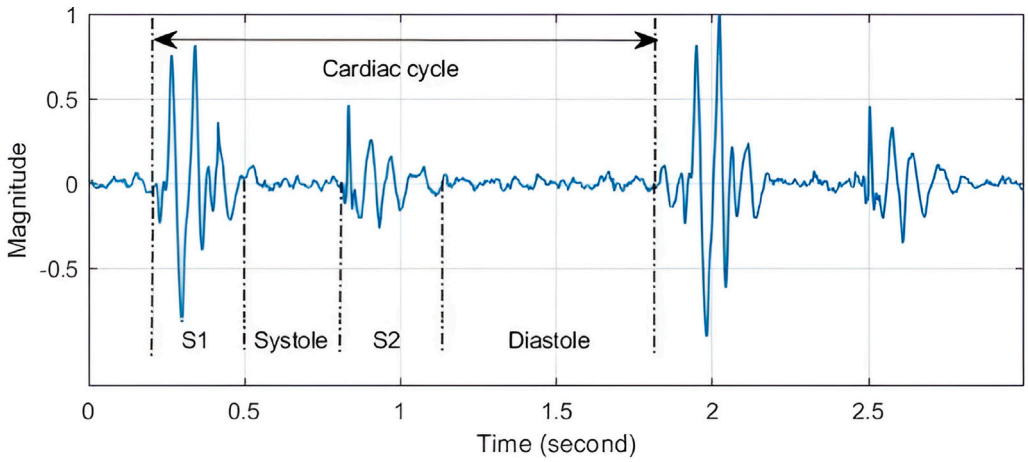


Fig. 1. Example of a heart sound signal.

in Fig. 1. Other audible heart sounds and murmurs associated with serious pathological conditions can occur when a valve experiences malfunctions, such as *regurgitation* or textit stenosis. Regurgitation occurs when the valve does not completely close, and blood flows back. Stenosis indicates that the valve is narrowed or damaged, restricting blood flow to the ventricles and atria.

Instantaneous changes in time and frequency can be observed in heartbeats S_1 and S_2 as well as in silent systolic and diastolic intervals in the presence of pathological abnormalities [6–8]. Regarding these changes, PCG signals can be analyzed in the time–frequency domain to obtain valuable features [9–11]. This can be achieved using multiresolution analysis (MRA) techniques, such as the short-time Fourier transform (STFT) [12] and continuous wavelet transform (CWT) [13, 14]. STFT obtains a time–frequency representation, called a spectrogram, by projecting the signal frame-by-frame onto a space spanned by complex exponential bases. Consequently, it is only feasible to determine the frequencies that exist within a time interval and not the instances [15,16]. Another relevant MRA technique is the CWT, which is used to analyze nonstationary signals by providing a simultaneous representation in both the time and frequency domains [17–19]. Using a variable-scale wavelet function, the CWT allows for a flexible and adaptive analysis of signals with changing frequency components over time. This process involves convolving the input signal with scaled and translated versions of a continuous wavelet to generate a two-dimensional representation, called a scalogram [20].

An alternative to MRA is the adaptive multiresolution analysis (AMRA) techniques. It is worth mentioning empirical mode decomposition (EMD), which, in contrast to the choice of wavelet functions and the setting of wavelet decomposition levels, separates the signals according to their characteristics without presetting the basis functions [21]. Therefore, a signal is adaptively divided into a finite number of intrinsic mode functions (IMFs), each containing different local time-scale signal characteristics. Another prevalent and effective AMRA technique is the variational mode decomposition (VMD) [22]. Compared to EMD, VMD exhibits excellent noise resistance, better decomposition performance, and stability. VMD formulates an optimization problem to estimate the IMFs with reduced total bandwidth. Finally, the empirical wavelet transform (EWT) constructs adaptive bandpass filters to decompose the signals and extract various functional components [23]. Because of its adaptability to learning, some studies have used EWT as part of complex models. Given the decomposition components, the Hilbert–Huang transform (HHT) is used to build time–frequency representations of the signals.

Deep neural networks are superior to traditional methods for the automatic diagnosis of cardiac problems [24,25]. Deep classifiers include convolutional neural networks (CNN) [24], recurrent neural networks (RNNs) [25,26], and hybrid combinations of both [27,28]. The convolutional layers extract representative information, and recurrent layers, such as long short-term memory (LSTM) layers, allow the output of some hidden nodes to affect the input of the same nodes, which is convenient for modeling PCG signals. Moreover, combining features from various layers or branches of layers is known as feature fusion, which is used to improve accuracy and is frequently implemented using simple operations, such as summation or concatenation. For example, deep network training is made possible within residual networks by merging identity-mapping features with residual learning using short-skip connections [29].

In this study, we propose a system to automatically detect heart failure and murmurs by classifying heart sound signals using a deep neural network. Based on the above discussion, this study makes the following contributions.

1. Develop a deep neural network model through a hybrid combination of multiple CNN and LSTM layers, incorporating a novel multistage feature fusion strategy.
2. Conduct a comparative study of five time–frequency representations of PCG signals derived from MRA and AMRA, using each individually to validate the model.
3. Validate the model for two tasks: detecting heart failure through the multiclass classification of PCG signals and identifying the presence of heart murmurs through the binary classification of PCG signals, using a separate dataset for each.

The proposed system achieves greater analytical depth by integrating deeper cascaded CNNs, parallel LSTM branches, and multistage feature fusion to capture the multiscale and multitemporal dynamics of PCG signals. The remainder of this paper is organized as follows. Section 2 presents recent works and studies related to this study. Brief introductions to methods of multiresolution analysis and adaptive multiresolution analysis of signals are presented in Sections 3 and 4, respectively. In Section 5, we explain the developed model in detail. The experimental analysis and results are discussed in Section 6. Finally, the conclusions of this study are summarized in Section 7

2. Related work

Machine learning and, more recently, deep learning have been widely used for tasks related to the analysis of PCG signals. Several studies have used MRA-based features (STFT and/or CWT) as inputs for deep models primarily based on CNNs. In [30–33] PCG signals were classified as normal or abnormal using complex STFT coefficients and CNN models. Time–frequency features were extracted from PCG signals using CWT, and a deep CNN model was built to classify the features as

Table 1
Comparison of STFT and CWT.

Feature	STFT	WT
Basis function	Fixed windowed sinusoidal functions	Scaled and shifted wavelets
Time–frequency resolution	Fixed	Adaptive
Frequency localization	Good for stationary signals	Suitable for nonstationary signals
Time localization	Limited due to fixed window size	High due to multiresolution approach
Computational complexity	Moderate	High
Limitation	Trade-off between time and frequency resolution	Selection of appropriate wavelet is crucial

normal or abnormal [34] and to discriminate multiple heart valvular disorders [35]. In [36] the authors proposed a lightweight CNN model to classify five categories of heart valvular conditions, in which the model was fed by features obtained from the CWT of the PCG signals. In [25,28] features extracted from the CWT and STFT of PCG signals were used to distinguish five valvular heart conditions. The classifiers consisted of a combination of CNN and LSTM in [28] and CNN and bidirectional LSTM in [25].

A multimodal residual neural network was developed to classify the extracted features based on EMD in [37]. The EMD was employed to decompose the ECG and PCG signals, and the IMF with the highest correlation with the original signals was selected to construct images that were classified using a residual network. In [38], the EWT was used in conjunction with deep learning models for the automatic recognition of PCG signals. Temporal envelope features were extracted from the EWT of the signals and were subsequently used to train and evaluate the models. In [39], the wavelet transform and VMD extracted key features from heart sound signals. The extracted features are used by deep neural networks to model, identify, and detect abnormal PCG signal dynamics associated with various heart diseases. The VMD and CNN-LSTM models discriminate between five heart valvular conditions. The PCG signals are decomposed into IMFs, and a weighted logarithmic operation is applied to the IMFs for feature extraction, which are classified using the model [40].

A multimodal CNN fusion architecture was developed to classify PCG signals in [41]. The architecture was trained and evaluated using features extracted from different domains, which were then merged to optimize the diverse features. In [42], the authors introduced a two-channel deep model to fuse 1D and 2D heart sound features, capturing both dynamic and time–frequency information. It also employs a dual attention mechanism with multi-head attention to focus on local and global relevance across channels. A fully convolutional fusion method for identifying the heartbeats S_1 and S_2 locations in PCG signals was presented in [43]. Moreover, a multimodal factorized bilinear pooling method is developed to merge the 1D envelope and 2D spectral heterogeneous features. In [44], the authors presented a CNN that fused features from different layers with varying resolution ratios and receptive-field sizes. Key features related to heart valve disease were weighted using a channel attention block in each layer. In [45], the authors presented a deep network for heart sound sequence labeling using physical signal features and a saliency-attentive network to suppress redundant information. The labeling results guided the design of a multichannel fusion network with a squeeze-excitation network that enhanced feature extraction.

Most prior works rely on single-stream CNN or CNN-LSTM models, apply shallow feature fusion only at the final layer, or perform limited multiscale analysis. In contrast, the proposed system is more comprehensive and biologically plausible because it captures the multiscale and multitemporal characteristics of heart sounds using deeper cascaded CNNs, parallel LSTM branches, and intermediate multistage feature fusion. Furthermore, this study explores multiple time–frequency feature extraction frameworks to enable a detailed comparison between fixed and adaptive resolution techniques for PCG analysis.

3. Multiresolution analysis (MRA)

In this section, we briefly describe multiresolution analysis (MRA) of signals, which employs techniques such as the STFT [46] and CWT [47] to decompose a signal into components that represent different frequencies. Table 1 compares STFT and CWT. The STFT of signal $x(t)$ is obtained as follows:

$$X(\omega,\tau)=\int_{-\infty}^{+\infty} x(t)w(\tau-t)e^{-j\omega t}dt, \tag{1}$$

where ω is the frequency and $w(t)$ is the windowing function. The magnitude component $|X(\omega,\tau)|$ is used to plot the spectrogram. The STFT is not able to represent abrupt changes because it assumes that the signal is stationary within a short time frame.

The CWT addresses the limitations of the STFT by decomposing the signal into localized wavelets. This process resembles that of the Fourier transform. However, instead of convolving it with sine waves, the wavelet transform convolves the signal with localized wavelets of varying scales and positions. The CWT of the processed signal is obtained as follows:

$$W(u,v)=\frac{1}{|u|^{\frac{1}{2}}}\int_{-\infty}^{+\infty} x(t)\overline{\psi}\left(\frac{t-v}{u}\right)dt. \tag{2}$$

Variables u and v are called the scaling and transition factors, respectively. The function $\psi(t)$ is called the mother wavelet, and $\overline{\psi}(t)$ is its complex conjugate. The Morlet wavelet is commonly used as a mother wavelet, which is characterized by

$$\psi(t)=e^{\frac{-t^2}{2\sigma^2}}e^{j\zeta t}, \tag{3}$$

where σ is the width of the Gaussian function, and ζ controls the time–frequency tradeoff. The magnitude components $|W(u,v)|$ are plotted to create a scalogram.

4. Adaptive multiresolution analysis (AMRA)

In this section, we explain techniques used in this study for adaptive multiresolution analysis (AMRA) of signals which involves the dynamic adjustment of resolution levels based on the characteristics of the signal being analyzed. In the context of time–frequency analysis, the AMRA adapts its resolution to capture the varying frequency components present in the signal. This adaptability is particularly beneficial when dealing with signals that exhibit nonstationary or time-varying characteristics. The three fundamental techniques under the umbrella of AMRA are EMD, EWT, and VMD. Table 2 compares STFT and CWT, and Fig. 2 shows an example of a heart sound signal with systolic murmurs and its decomposition into five IMFs.

4.1. Empirical mode decomposition (EMD)

An intrinsic mode function (IMF) must satisfy the following two conditions.

- The number of extrema (maxima and minima) and zero-crossings must be equal or differ by at most one throughout the entire signal.

Table 2
Comparison of EMD, EWT, and VMD.

Feature	EMD	EWT	VMD
Basis function	Data-driven	Adaptive wavelets	Variational optimization
Decomposition strategy	Iterative extrema-based filtering	Frequency-based segmentation	Mode separation using optimization
Mode extraction	Intrinsic Mode Functions	Wavelet sub-bands	Variational modes
Adaptability	High	Moderate	High
Noise sensitivity	High	Moderate	Low
Computational complexity	Low	Moderate	High
Frequency separation	Data-dependent	Predefined frequency bands	Automatically optimized
Reconstruction accuracy	Moderate	High	High
Robustness to mode mixing	Low	Moderate	High

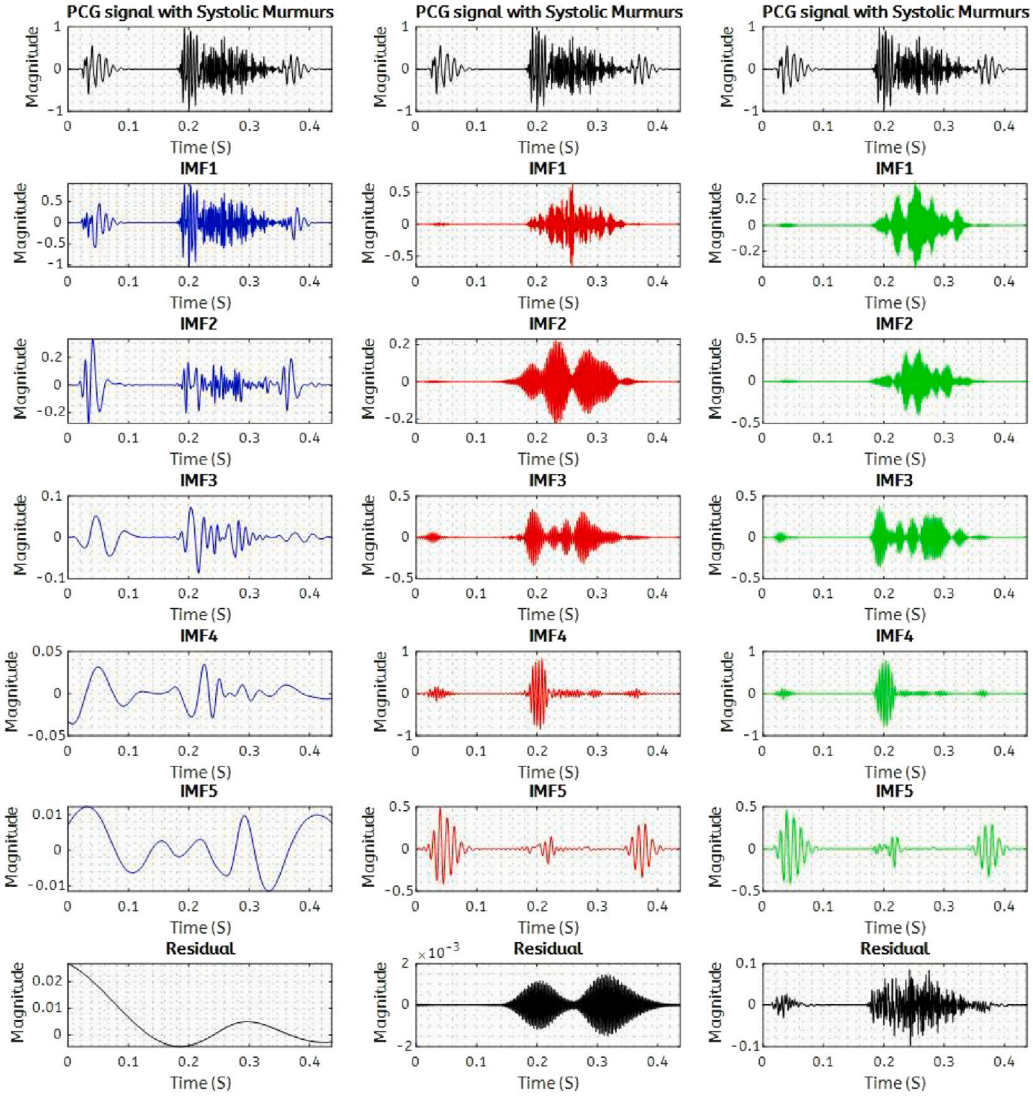


Fig. 2. Examples of decomposing a PCG signal into 5 IMFs using EMD at the left column, EWT at the center column, and VMD at the right column.

- At any point in the signal, the mean of the upper and lower envelopes (defined by local maxima and minima) should be equal or nearly equal to zero.

Once these extrema are identified, all local maxima are connected by a cubic spline line to form the upper envelope. Similarly, local minima are connected to create a lower envelope. The mean of these envelopes is denoted as $m_1(t)$, and the difference between the data and $m_1(t)$ constitutes the first component, $h_1(t)$, i.e.,

$$h_1(t) = x(t) - m_1(t). \tag{4}$$

If $h_1(t)$ satisfies these two conditions, it is defined as the mode function. Otherwise, the process has to be repeated k times until there $h_{1k}(t)$ is an IMF, which is

$$h_{1k}(t) = h_{1(k-1)}(t) - m_{1k}(t). \tag{5}$$

Then, the first IMF component $c_1(t) = h_{1k}(t)$. Overall, $c_1(t)$ should contain the finest-scale or shortest-period component of the signal. The IMF component $c_1(t)$ can be separated from the rest of the data by

$$r_1(t) = x(t) - c_1(t). \tag{6}$$

Given that the residual $r_1(t)$ retains information regarding the components of longer periods, the procedure is applied to all residuals $r_j(t)$, yielding the following result:

$$r_2(t) = r_1(t) - c_2(t), \ldots, r_k(t) = r_{k-1}(t) - c_k(t). \tag{7}$$

The process can be terminated based on predetermined criteria: when the component $c_k(t)$ or residual $r_k(t)$ diminishes to a magnitude below a predefined threshold of significance, or when the residual $r_k(t)$ transforms into a monotonic function, indicating that no further extraction of the IMFs is feasible. By aggregating the computed IMFs and the ultimate residual $r(t)$, the final result is obtained, such as

$$x(t) = \sum_{k=1}^{K} c_k(t) + r(t). \tag{8}$$

4.2. Empirical wavelet transform (EWT)

The wavelet transform acts as a filter bank, allowing the EWT to isolate distinct spectrum segments corresponding to modes with specific frequencies and compact support [23]. The Fourier supports $[0, \pi]$ are assumed to be segmented into N contiguous segments, as follows: We denote ω_n as the limit between segments (where $\omega_0 = 0$ and $\omega_N = \pi$). Each segment is denoted $\Lambda_n = [\omega_{n-1}, \omega_n]$, so it is easy to see $\cup_{n=1}^{N} \Lambda_n = [0, \pi]$. Located around each ω_n, we define the transition phase T_n with width $2\tau_n$. The simplest approach is to choose τ_n which is proportional to ω_n: $\tau_n = \gamma\omega_n$ where $0 < \gamma < 1$. Empirical wavelets are defined as bandpass filters for each Λ_n. The empirical scaling function $\hat{\phi}_n(\omega)$ and wavelet $\hat{\psi}_n(\omega)$ are defined as follows:

$$\hat{\phi}_n(\omega) = \begin{cases} 1, & |\omega| \leq (1-\gamma)\omega_n \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\gamma\omega_n}(|\omega| - (1-\gamma)\omega_n)\right)\right], & (1-\gamma)\omega_n \leq |\omega| \leq (1+\gamma)\omega_n \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

and

$$\hat{\psi}_n(\omega) = \begin{cases} 1, & (1+\gamma)\omega_n \leq |\omega| \leq (1-\gamma)\omega_{n+1} \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\gamma\omega_{n+1}}(|\omega| - (1-\gamma)\omega_{n+1})\right)\right], & (1-\gamma)\omega_{n+1} \leq |\omega| \leq (1+\gamma)\omega_{n+1} \\ \sin\left[\frac{\pi}{2}\beta\left(\frac{1}{2\gamma\omega_n}(|\omega| - (1-\gamma)\omega_n)\right)\right], & (1-\gamma)\omega_n \leq |\omega| \leq (1+\gamma)\omega_n \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

The function $\beta(x)$ is an arbitrary function such that

$$\beta(x) = \begin{cases} 0, & x \leq 0 \\ \beta(x) + \beta(1-x) = 1, & 0 < x < 1 \\ 1, & x \geq 1. \end{cases} \tag{11}$$

Given the number of segments N, it is necessary to identify $N+1$ boundaries, including the predefined boundaries 0 and π. To find $N-1$ additional boundaries, the local maxima in the spectrum were initially identified, excluding 0 and π. Assuming that there are M maxima, two scenarios can arise.

- $M \geq N$: the algorithm finds enough maxima to define the wanted number of segments, and then we keep only the first maxima.
- $M < N$: the signal has fewer modes than expected, then we keep all the detected maxima and reset to the appropriate value.

Equipped with the maxima plus 0 and π, the boundaries ω_n of each segment are defined as the centers between two consecutive maxima. The transition areas in which consecutive T_n values do not overlap have yet to be determined. For this purpose, γ is defined as

$$\gamma < \frac{\omega_{n+1} - \omega_n}{\omega_{n+1} + \omega_n}. \tag{12}$$

The EWT can be defined as $\mathcal{W}(n,t)$ in the same manner as the wavelet transform. The detailed coefficients of $x(t)$ are given by the inner products with the empirical wavelets and the approximation coefficients of the inner product with the scaling function,

$$\mathcal{W}(n,t) = \langle x(t), \psi_n(t) \rangle \quad \text{and} \quad \mathcal{W}(0,t) = \langle x(t), \phi_1(t) \rangle, \tag{13}$$

where $\psi_n(t)$ and $\phi_1(t)$ are the inverse Fourier transforms of $\hat{\psi}_n(\omega)$ and $\hat{\phi}_1(\omega)$. The empirical mode $z_k(t)$ is then calculated through the convolutional operation denoted by $*$, such as

$$z_0(t) = \mathcal{W}(0,t) * \phi_1(t) \quad \text{and} \quad z_k(t) = \mathcal{W}(k,t) * \psi_k(t). \tag{14}$$

Subsequently, the original signal $x(t)$ is obtained as follows:

$$x(t) = \sum_{k=1}^{K} z_k(t) + z_0(t). \tag{15}$$

4.3. Variational mode decomposition (VMD)

Variational mode decomposition (VMD) decomposes a PCG signal $x(t)$ into K mode functions $s_k(t)$ with specific sparsity properties as follows:

$$x(t) = \sum_{k=1}^{K} s_k(t) + r(t). \tag{16}$$

The sparsity of the kth mode $s_k(t)$ is determined by its bandwidth, which is concentrated around its center frequency ω_k. VMD aims to minimize the mode bandwidth shift to its center frequency via a complex harmonic model, resulting in an optimization process. The constrained problem can then be expressed as [22]

$$\min_{\{s_k(t)\},\{\omega_k\}} \left(\sum_{k=1}^{K} \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t}\right) * s_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right) \quad s.t. \quad \sum_{k=1}^{K} s_k(t) = x(t), \tag{17}$$

where $\delta(t)$ denotes the Dirac delta function and $\frac{1}{\pi t} * s_k(t)$ denotes the Hilbert transform. The problem in (17) is solved using the saddle point of the augmented Lagrangian in the alternating direction method of multipliers (ADMM). Following this formulation, the empirical mode $\hat{s}_k^{n+1}(\omega)$ is estimated iteratively, such as

$$\hat{s}_k^{n+1}(\omega) = \frac{\sum_{i=1}^{K} \hat{s}_i^n(\omega) - \sum_{i \neq k}^{K} \hat{s}_i^n(\omega) + \frac{\hat{\lambda}(\omega)}{2}}{1 + 2\gamma(\omega - \omega_k)^2}, \tag{18}$$

where γ denotes the Lagrangian multiplier, and $\lambda(\omega)$ denotes the quadratic penalty. The magnitude of the penalty term is inversely proportional to the inherent noise level of the data. The resulting IMF is recognized as a Wiener filter for the current estimate of $\hat{s}_k^{n+1}(\omega)$ with a signal prior to $1/(\omega - \omega_k)^2$. The central frequency refers to the frequency derived from the linear regression performed on the instantaneous phase observed within the mode, accomplished using the least-squares method as follows:

$$\omega_k^{n+1} = \frac{\int_0^{+\infty} \omega \, |\hat{s}_k^{n+1}(\omega)|^2 \, d\omega}{\int_0^{+\infty} |\hat{s}_k^{n+1}(\omega)|^2 \, d\omega}. \tag{19}$$

4.4. Hilbert-Huang transform (HHT)

The HHT calculates the instantaneous frequency of each IMF given the IMFs resulting from the application of AMRA to a signal, and the Hilbert–Huang transform (HHT) calculates the instantaneous frequency of each IMF. The instantaneous frequencies of all IMFs form a Hilbert spectrum (HS). The kth IMF $g_k(t) = c_k(t), z_k(t)$, or $s_k(t)$ has the Hilbert transform (HT), expressed as

$$q_k(t) = \mathrm{HT}(g_k(t)) = g_k(t) * \frac{1}{\pi t} = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{g_k(\tau)}{t - \tau} d\tau, \tag{20}$$

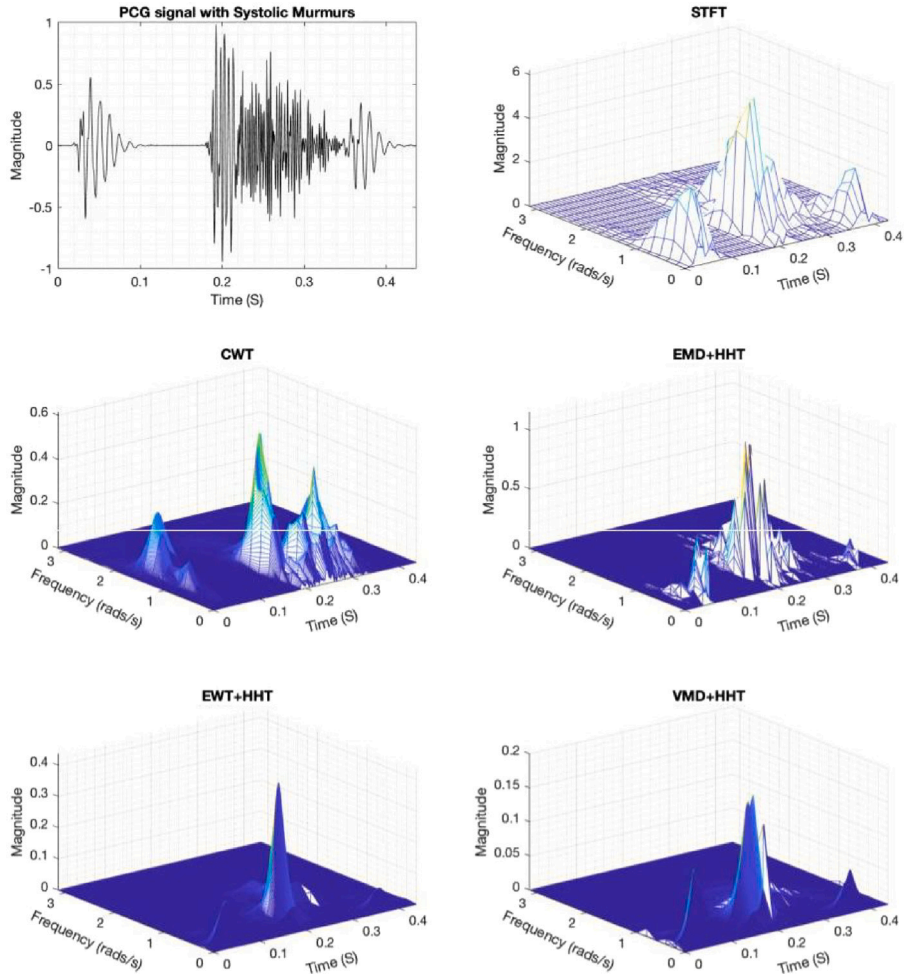


Fig. 3. Examples of MRA and AMRA time–frequency representation of PCG signal.

where the symbol * denotes a convolution operation, and the analytic signal is defined as

$$u_k(t) = g_k(t) + jq_k(t) = a_k(t)e^{j\theta_k(t)}, \text{ where } j = \sqrt{-1}, \tag{21}$$

$$a_k(t) = \sqrt{g_k^2(t) + q_k^2(t)} \quad \text{and} \quad \theta_k(t) = \arctan\left(\frac{q_k(t)}{g_k(t)}\right). \tag{22}$$

Here, $a_k(t)$ is the instantaneous amplitude, and $\theta_k(t)$ is the phase function. Consequently, the instantaneous frequency $\omega_k(t)$ is the first-order derivative of the phase function. By combining the instantaneous amplitude and frequency in a single vector of two functions, the Hilbert spectrum (HS) is obtained as $\text{HS}(g_k(t)) = [a_k(t), \omega_k(t)]$. After performing the HT on each component, the original data can be expressed as a real part in the following form

$$x(t) = \text{Real}\left[\sum_{k=1}^{K} a_k(t)e^{j\int \omega_k(t)dt}\right]. \tag{23}$$

The HHT is the HS ensemble of each IMF and residual, that is, each of the $K+1$ vectors has two functions. The underlying HHT of the signal is then obtained, such as [48]

$$\text{HHT}(\omega, t) = \sum_{k=1}^{K+1} a_k(t, \omega_k(t)), \tag{24}$$

where $a_k(t, \omega_k(t))$ combines the instantaneous amplitude $a_k(t)$ and the instantaneous frequency $\omega_k(t)$. Fig. 3 shows five different time-frequency representations of a heart sound signal with systolic murmurs.

5. Hybrid network with multistage feature fusion

This section describes the hybrid neural network model with a multistage feature fusion strategy developed in this study for the detection of heart failure and murmurs. Fig. 4 shows a block diagram of the proposed system. The model is trained and evaluated using feature matrices derived from time–frequency matrices obtained by applying AMRA and MRA to PCG signals. These matrices capture both the spectral and temporal dynamics of heart sound signals more effectively, thereby providing richer and more distinct classification features. The model architecture consists of three convolutional neural networks (CNNs), three rectified linear units (ReLUs), three average pooling layers (AVGP), three long short-term memory layers (LSTMs), a fusion layer, a fully connected layer, and a softmax classification layer. This architecture leverages the strengths of CNNs for the extraction of spatial features and LSTMs to capture their sequential dependencies. The AVGP layers reduce the size of the feature map and maintain important information. The fusion layer plays a critical role in integrating multistage encoded features at different network levels. This comprehensive integration of interior features is then passed to the fully connected layer for high-level decision-making, with the final classification performed by the softmax layer to output the prediction probabilities.

Stacking CNNs one after another in a cascade allows the model to progressively learn more abstract and high-level features. Early layers capture local, fine-grained patterns (e.g., murmurs, S_1/S_2 onset), whereas deeper layers extract global structures (e.g., rhythm,

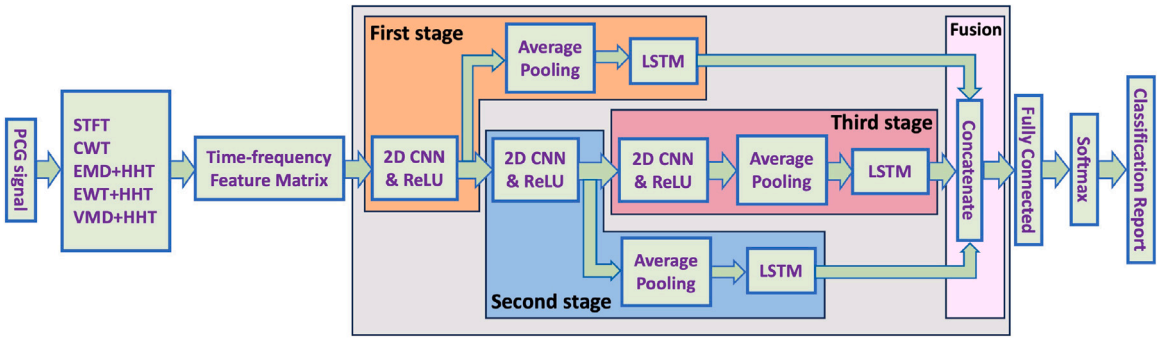


Fig. 4. Block diagram of the developed system with the multistage feature fusion network.

energy trends). By branching LSTM layers from different CNN stages, the model can analyze temporal dynamics at multiple feature depths. This design allows the system to detect both short-term anomalies (e.g., sharp murmurs) and longer-term rhythmic irregularities (e.g., gallop and arrhythmia). Moreover, fusing features from multiple intermediate stages allows the network to integrate low-level acoustic cues (e.g., onset, intensity), mid-level patterns (e.g., spectral transitions), and high-level abstractions (e.g., disease-specific sound profiles). The combination of cascaded CNNs, parallel LSTMs, and intermediate multistage fusion makes the model less sensitive to variability in PCG signals, such as different heart rates, recording conditions, and patient-specific variations.

The time–frequency representations of the PCG signals were obtained using AMRA through EMD, EWT, and VMD combined with HHT and MRA through STFT and CWT. Notably, this is the first time that EWT and VMD combined with HHT have been used for diagnostic applications. The choice of these analysis techniques depends on the characteristics of the signal under consideration and the specific requirements of the analysis. Each technique has its own advantages and limitations, making it suitable for a wide range of signals and applications. The STFT and CWT are widely used to analyze signals because of their simplicity and effectiveness. However, EMD, EWT, and VMD integrated with HHT are preferred for nonstationary signals, offering adaptability and improved mode separation compared with other methods. To train and test the model, we adopted the feature calculation technique presented in [5,40]. This involves applying a weighted logarithmic function to the time–frequency representations obtained from the MRA and AMRA before they are fed into the model.

The system was evaluated by completing two tasks using two separate datasets that were tailored to each task. The first evaluation task focused on detecting a diverse range of cardiac conditions. The dataset used for this task included PCG signals associated with five valvular heart conditions. The second evaluation task focused on predicting the presence of heart murmur. The dataset tailored for this task includes PCG signals with a mix of instances, some containing systolic heart murmurs and others devoid of murmurs.

5.1. Time–frequency feature matrix

The time–frequency representations generated by the MRA and AMRA serve as the basis for calculating the feature matrices required to train and evaluate the network. Given the time–frequency matrices of the PCG signals $|X(\omega,\tau)|$, $|W(u,v)|$, and HHT(ω,t), we calculate the feature matrices applying a weighted logarithmic function [5,40] such that

$$Y = -|X| \circ \log_{10}|X|, \tag{25}$$

where X refers to the time–frequency matrix with entries of values within the $[-1, 1]$ range. The symbol $\circ$ denotes the point-product operation. This weighted logarithmic function is a crucial step in the feature extraction process, contributing to the extraction of relevant information from the time–frequency representations of the PCG signals. It acts as a crucial preprocessing step, highlighting essential features while mitigating the effects of noise and irrelevant information. By incorporating this function into the analysis, the system ensures a robust and discriminative representation of PCG signals.

5.2. Multistage feature fusion

Feature fusion within neural networks refers to the process of combining multiple sets of features, which are often extracted from different stages, to enhance model generalizability. This fusion can occur at different levels of the network, such as early, mid, or late, depending on when the features are combined during the learning process. By integrating diverse information from different layer stages, feature fusion allows the network to capture a richer and more comprehensive representation of the input data, thereby improving its ability to accurately predict the target variable. This strategy is commonly implemented through elementary operations, such as summation or concatenation, and has been extensively explored in the literature [29]. This integration helps to train deep networks and address the challenges associated with the vanishing gradients.

In the multistage feature fusion network, features are computed through multiple branches and are subsequently fused in a concatenation layer. Each branch is composed of an AVGP followed by an LSTM. These branches receive inputs from consecutively connected two-dimensional CNNs, followed by ReLUs. This architecture enables the branches to encode heat maps with diverse structures and to capture intricate patterns and spatial relationships within the input data.

- The convolutional neural network (CNN) performs convolutional operations followed by ReLU. This sequence introduces nonlinearity, which improves the ability of the model to capture complex relationships in the data. This process takes important parts of the input and generates heat maps, which allow the detection of patterns in an input, regardless of their position. These patterns can be identified at higher levels through repeated convolutional layers, leading to the recognition of meaningful features. For a given 2D input feature matrix Y and 2D convolutional filter H, the output of the convolution operation at position (i,j) is

$$Z(i,j) = \sum_{m=0}^{H_f-1}\sum_{n=0}^{W_f-1} Y(i+m, j+n)H(m,n) + b, \tag{26}$$

where $Y(i,j)$ denotes the input feature map, $H(m,n)$ is the convolutional filter, and b is the bias. For an input Y of size $H_{\text{in}} \times W_{\text{in}} \times C_{\text{in}}$ and a filter H of size $H_{\text{f}} \times W_{\text{f}} \times C_{\text{f}}$ with stride $S_H \times S_W$ and

padding P, the output Z is of size given by:

$$H_{\text{out}} = \frac{H_{\text{in}} - H_{\text{f}} + 2P}{S_H} + 1, W_{\text{out}} = \frac{W_{\text{in}} - W_{\text{f}} + 2P}{S_W} + 1, \text{ and } C_{\text{out}} = C_{\text{f}}. \tag{27}$$

Here, the height H_{in} and width W_{in} of the feature maps Y decrease as the processing advances from one CNN stage to the next, whereas the number of channels C_{in} increases from one stage to the next. The filter height H_{f}, width W_{f}, and stride $S_H \times S_W$ are kept constant, and the number of filters C_{f} varies from stage to stage. The value of padding P was set to zero. The ReLU is a widely used activation function that introduces non-linearity and is applied element-wise to Z such as

$$Z' = \max(Z, 0). \tag{28}$$

- The average pooling layer (AVGP) calculates the average value for patches of a feature map and uses it to create a downsampled map. It is typically used after convolutional and rectified linear unit layers. This adds a small amount of translation invariance, which means that translating the feature map by a small amount does not significantly affect the pooled outputs. We use the AVGP to calculate the average of the multichannel 2D feature maps along the frequency axis, resulting in multichannel temporal sequences. For an input feature map Z' and a pooling window of size $H_p \times W_p$, the output at position (i, j) is

$$Z^p(i,j) = \frac{1}{H_p W_p} \sum_{m=0}^{H_p-1} \sum_{n=0}^{W_p-1} Z'(S_H^p i + m, S_W^p j + n), \tag{29}$$

where $Z'(i, j)$ is the input feature map, $S_H^p \times S_W^p$ and P^p are the stride and padding of the pooling operation, respectively, and H_p and W_p are the height and width of the pooling window, respectively. The output Z^p is of size given by

$$H^p = \frac{H_{\text{out}} - H_p + 2P^p}{S_H^p} + 1, W^P = \frac{W_{\text{out}} - W_p + 2P^p}{S_W^p} + 1 \text{ and } C^p = C_{\text{out}}. \tag{30}$$

Here, we choose the height S_H^p and width S_W^p of the stride to equal the height H_p and width W_p of the pooling window, respectively, and the padding P^p to equal zero. The height H_p of the pooling window is assigned a value equal to the height H_{out} of the input Z'. The width W_p of the pooling window is assigned a value of one. Accordingly, the output of the pooling layer Z^p is $1 \times W_{\text{out}} \times C_{\text{out}}$. A flattening layer is then used to obtain a flattened vector z^f of length W_{out} multiplied by C_{out}, which is the most suitable input for the LSTM layer.

- The long short-term memory layer (LSTM) is a recurrent neural network (RNN) that learns long-term dependencies in sequential data. A typical RNN has a single hidden state that propagates over time, making it difficult to learn long-term dependencies. LSTM overcomes this problem by introducing a memory cell that can store data for longer periods, thereby allowing the network to capture dependencies in long-term contextual information. Three gates regulate cell operations: input, forget, and output. These gates determine the information that is added to, removed from, and output by the memory cell. Given an input vector z^f of length W_{out} multiplied by C_{out}, the LSTM returns the last hidden state at the final step of the input for each unit (neuron), resulting in an output sequence of length equal to the number of LSTM units in the input sequence. Here, we use the same number of units for all LSTM layers found in different stages. Given the input sequence $z^f(n)$ with the entry index n, previous hidden state $h(n-1)$, and previous cell state $c(n-1)$, the LSTM equations are as follows:
  - Forget gate determines how much of the previous cell state to keep:

    $$f(n) = \sigma(W_f z^f(n) + U_f h(n-1) + b_f)$$

  - Input gate determines how much new information to store in the cell state:

    $$i(n) = \sigma(W_i z^f(n) + U_i h(n-1) + b_i)$$

  - Candidate cell state adds new candidate values to the cell state:

    $$\tilde{c}(n) = \tanh(W_c z^f(n) + U_c h(n-1) + b_c)$$

  - Update the cell state by combining old and new information:

    $$c(n) = f(n) \odot c(n-1) + i(n) \odot \tilde{c}(n)$$

  - Output gate controls what part of the cell state is sent to the hidden state:

    $$o(n) = \sigma(W_o x_t + U_o h(n-1) + b_o)$$

  - Hidden state update (Output of LSTM Layer)

    $$h(n) = o(n) \odot \tanh(c(n))$$

  where $h(n)$ is the hidden state at n, $c(n)$ is the cell state at n, W_f, W_i, W_c, W_o are the weight matrices for the forget, input, candidate cell, and output gates, respectively, U_f, U_i, U_c, U_o are the recurrent weight matrices for the hidden state, b_f, b_i, b_c, b_o are the bias terms, σ is the sigmoid activation function, tanh is the hyperbolic tangent function, and $\odot$ is element-wise multiplication.
- The multistage feature fusion network advances the conventional understanding of feature fusion by introducing a sophisticated architecture that takes advantage of multiple branches to compute and fuse features. This innovative approach contributes to state-of-the-art deep learning, particularly for tasks that require the integration of modeled spatial and temporal information for a comprehensive analysis of the input features. Considering the outputs from all the LSTM layers across different network stages, the fusion layer concatenates them into a single extended vector. The length of this vector is determined by multiplying the number of units in a single LSTM layer by the total number of fused stages. This process ensures comprehensive feature representation, enabling the model to leverage the temporal dependencies captured at multiple stages to enhance the classification performance.

6. Experimental analysis

In this section, we discuss the datasets used to evaluate the system, system implementation and computational complexity issues, evaluation metrics, and experimental results and comparison. The developed system was experimentally evaluated to measure its effectiveness in two unique tasks: diagnosing heart failure and predicting the presence of heart murmurs. Each activity uses a unique dataset customized for its goal. For both tasks, the evaluation approach involved assessing the performance of the system using standard classification metrics such as accuracy, specificity, and F1-score. A 5-fold cross-validation strategy was implemented to ensure the robustness of the evaluation results. The datasets were partitioned into training and testing sets to provide a reliable estimate of the generalization performance of the proposed model. To prevent overfitting, the script is designed to run enough epochs until saturation, and the best weights of the model, those that yield the lowest validation loss, are saved and restored, ensuring optimal generalization of unseen data. The results obtained from both evaluation tasks were meticulously analyzed to understand their strengths, limitations, and areas for further improvement. Any patterns or challenges identified during the analysis were considered for potential refinements or optimizations to improve the overall performance.

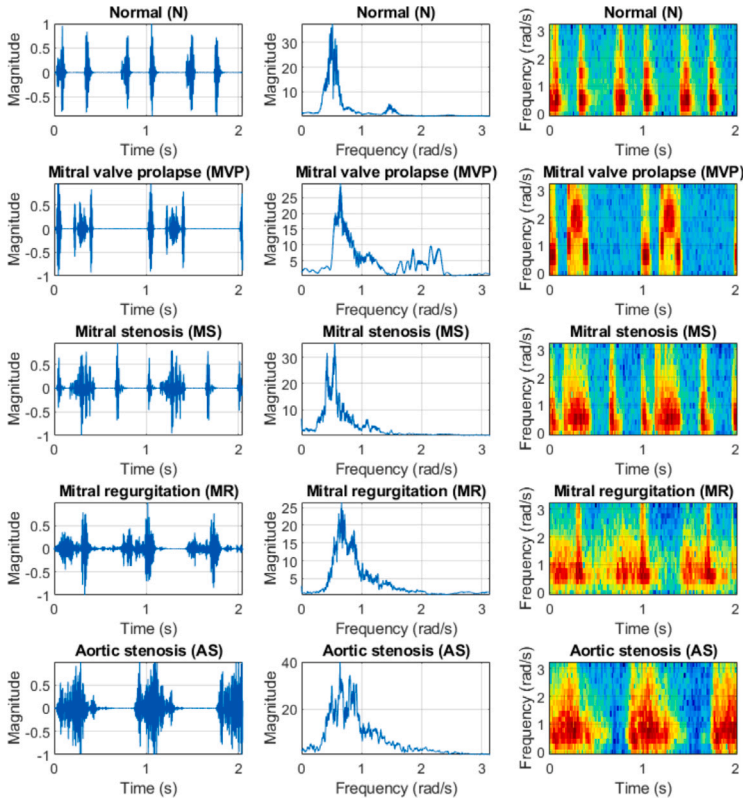


Fig. 5. PCG signals of a healthy heart and four with valvular conditions and their Fourier transforms and spectrograms at a sampling rate of $f_s = 1000$ Hz. These signals are from the first dataset.

6.1. First dataset

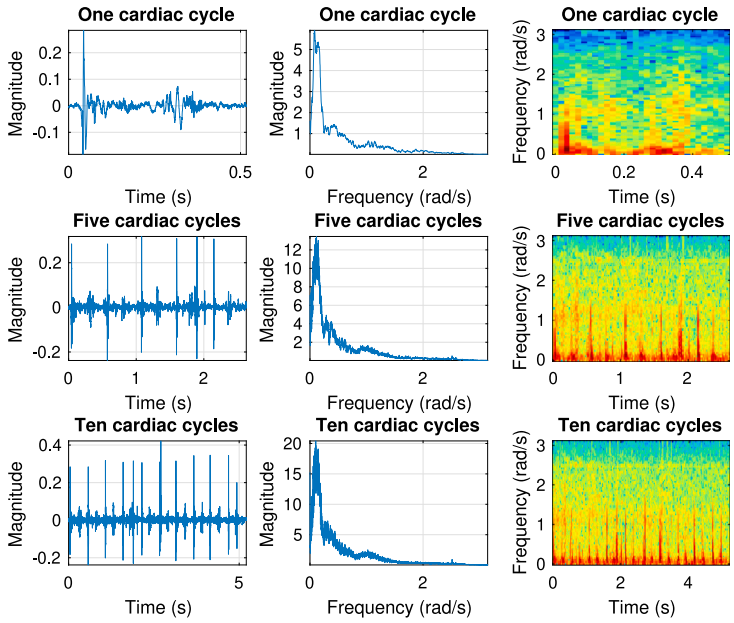
The first dataset comprises heart sound recordings categorized into five distinct groups, as detailed below [28,49]. These categories represent different cardiac conditions and provide a diverse set of samples for analysis.

- **Normal (N)** denotes PCG signals originating from healthy hearts. These signals exhibit a distinct pattern, beginning with the first heartbeat (S_1), followed by a silent systolic interval, the second heartbeat (S_2), and a silent diastolic interval.
- **Mitral valve prolapse (MVP)** refers to the systolic prolapse of the mitral leaflet into the left atrium. Although MVP is typically benign, it can lead to complications such as chordal rupture and mitral regurgitation. The presence of a midsystolic click and late systolic murmur indicates notable regurgitation.
- **Mitral stenosis (MS)** arises from an incomplete opening of the mitral valve that restricts blood flow from the left atrium to the left ventricle. This impediment leads to blood accumulation in the pulmonary circulation. Heart sounds reveal accentuated early S_1 in MS, with a soft S_1 in severe cases.
- **Mitral regurgitation (MR)** occurs when the mitral valve does not fully close, leading to a backflow of blood into the heart. In MR, S_1 may be soft or absent due to leaflet-valve sclerosis. Murmurs in MR typically begin after S_1 when leaflet absence occurs during systole, escalating to S_2.
- **Aortic stenosis (AS)** occurs due to a narrow or stiff aortic valve, leading to delayed closure of the aortic valve. The symptoms of AS include high-pitched diamond-shaped murmurs that are best detected at the upper right border of the sternum. In mild aortic coarctation, systolic murmurs peak during early systole.

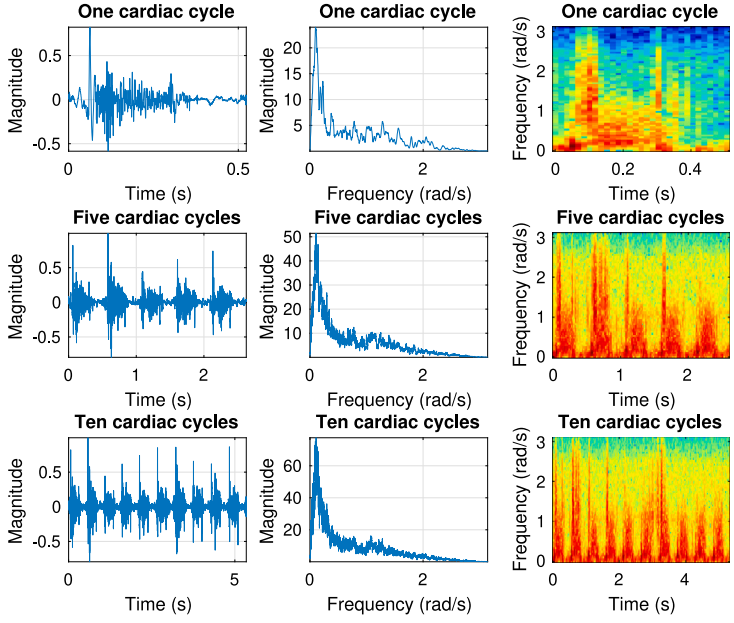
The total number of heart sound recordings was 1000 (200 audio files/category), with an average length of 20 000 samples or 2.5 s at a sampling rate of 8000 Hz. All the PCG signals were temporally clipped or zero-padded to the same length of 20 480 samples at a sampling rate of 8000 Hz. This step was followed by downsampling the signals by a factor of 10 to reduce the computational cost. Thus, the length of each preprocessed signal was $L = 2048$. Finally, each PCG signal was standardized to have a zero mean and unity standard deviation. Fig. 5 shows plots of heart sound signals and their corresponding Fourier transforms and spectrograms for a normal PCG signal of a healthy heart and PCG signals for four patients with different CVDs, in particular, mitral valve prolapse, mitral stenosis, mitral regurgitation, and aortic stenosis.

6.2. Second dataset

The second dataset used in our experiments was recently made available under the title CirCor DigiScope dataset [50]. Heart sound recordings were collected from four points on the chest: the aortic, pulmonary, mitral, and tricuspid valves. The signals were sampled at 4000 Hz and normalized within the [−1, 1] range. The normalized signals were automatically segmented and manually checked by cardiac experts to identify the boundaries between S_1 and S_2. If the expert disagreed with the automatic annotations, a new audio file of five cardiac cycles was created and saved with corresponding annotations. All recordings were labeled as normal or abnormal, and the presence of murmurs was screened at each auscultation location. Murmurs can be



(a) Healthy heart with no audible activity in silent intervals.



(b) Pathological heart with systolic murmurs

Fig. 6. PCG signals from the second dataset and their corresponding Fourier transforms and spectrograms at a sampling frequency of $f_s = 4000$ Hz.

systolic, diastolic, or systolic-diastolic, with the majority being systolic. A damaged heart valve typically produces a louder murmur in the corresponding auscultation area because of its proximity to the valve. In this dataset, the murmurs are most intense at the pulmonary point.

A balanced dataset was constructed, focusing on the most audible recordings among the four recordings collected from four points on the chest. This reduced dataset encompasses two distinct categories of labeled recordings: normal, with the absence of murmurs, and abnormal, with the presence of murmurs, which produced a total of 358 (179 audio files/category). Each file contains a minimum of five cardiac cycles with an average cycle length of 2300 samples. To streamline the subsequent processing, each cardiac cycle was resized to a uniform length of 2048 samples. Table 3 lists the total number of recordings for three different ranges of the number of cycles. Fig. 6 shows PCG signals with multiple cardiac cycles, highlighting cases with and without systolic heart murmurs, along with the corresponding Fourier transforms and spectrograms.

Table 3
Number of recordings for different ranges of number of cardiac cycles.

Number of cardiac cycles	Cycles = 5	10 > Cycles > 5	Cycles ≥ 10	Total
Number of recordings	6	25	327	358

Table 4
System implementation.

PCG signal		Multistage fusion network	
Parameter	Value	Parameter	Value
Signal length (samples):		**CNN:**	
First dataset	2048 × 1	Input size	TF × no.cycles
Second dataset	2048 × no. cycles	No. filters	64 × no. stages
Time–frequency (TF) for one cycle:		Filter size	3 × 16
STFT:		Filter stride	2 × 8
Window length (samples)	32	**LSTM:**	
Overlap (samples)	16	No. units	64
Size of spectrogram	17 × 127	**Training:**	
CWT:		Optimizer	ADAM
Mother wavelet	Morse	Loss function	MSE
Size of scalogram	17 × 2048	Learning rate	0.001
no. EMD modes	auto	Batch size	112
no. EWT modes	64	no. epochs	100
no. VMD modes	64		
Size of HHT	17 × 2048		

6.3. System implementation

Table 4 lists the parameters that govern the time–frequency matrices derived from the PCG signals and the configuration details of the CNN-LSTM classification network. The parameters encapsulating a spectrum of methodologies used in feature extraction, including STFT, CWT, EMD + HHT, EWT + HHT, and VMD + HHT, are characterized by a set of parameters that define the transformation process, facilitating the extraction of time–frequency features from the signals. It is worth noting that EMD automatically determines the number of modes. In contrast, for EWT and VMD, the number of modes is manually selected based on the best performance obtained from testing multiple values.

This table also provides information on the training parameters that are crucial for optimizing the developed model. These parameters include the optimization algorithm, adaptive moment optimization (ADAM), which governs the weight updates during training. Additionally, it specifies the learning rate, a critical hyperparameter that controls the magnitude of weight adjustments, along with the batch size, which determines the number of samples processed in each iteration.

6.4. System computational complexity

The computational complexity of neural networks refers to the resources required to train and evaluate a model, which is heavily influenced by the number of layers and the learnable parameters. As the depth of the network increases, the number of parameters increases, leading to higher computational costs in terms of time and memory usage. Deep networks with many layers can capture more complex patterns; however, they require more processing power and longer training times, particularly when they handle large datasets. Table 5 lists the number of layers and learnable parameters for the fused stages. The developed model contains between 1.3 and 3.4 million learnable parameters, classifying it as lightweight to moderately sized by modern standards. Compared to large models with tens, hundreds, or even billions of parameters, this compact size makes it particularly suitable for embedded clinical deployment on low-resource hardware.

6.5. Evaluation metrics

The effectiveness of the proposed system was evaluated by calculating four essential metrics: true positives (TP), which are the number of cases in which a specific valvular heart disease was correctly identified. False negatives (FN) are the number of cases in which the model incorrectly classifies a particular valvular heart condition as belonging to other conditions. False positives (FP) are the number of cases in which the model incorrectly classifies other conditions as belonging to a particular valvular heart condition. True negatives (TN) are the number of cases in which the other conditions were correctly identified.

Table 5
The number of layers and trainable parameters, with "M" denoting a million, for each architecture.

Fused stages	No. layers	No. of learnable parameters
First stage	13	1.3 M
First and second stages	17	2.4 M
Three stages	20	3.4 M

In addition to these metrics, additional measures can be calculated to provide deeper insights into the behavior of the model, offering a comprehensive understanding of its diagnostic capabilities.

- **Confusion matrix** summarizes the number of correct and incorrect predictions across different classes. It is structured as a table with actual class labels on one axis and predicted class labels on the other, highlighting TP, FP, TN, and FN.
- **Specificity** is the ratio between the number of correctly classified negative samples and the total number of negative samples.

$$\textbf{Specificity} = \frac{TN}{TN + FP}\%. \tag{31}$$

- **F1-score** is the harmonic mean of recall and precision.

$$\textbf{F1} - \textbf{score} = \frac{2TP}{2TP + FP + FN}\%. \tag{32}$$

- **Accuracy** is the number of correct predictions to the total number of predictions.

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}\%. \tag{33}$$

6.6. Results

We performed five-fold cross-validation, in which the recordings of each dataset were divided into five subsets. The model was trained and evaluated five times using four subsets as the training set and the remaining subset as the test set. This approach ensures that each signal is used for training and validation, thereby reducing overfitting and providing a comprehensive evaluation of model generalization.

6.6.1. Results of the first dataset

Table 6 and Fig. 7 illustrate the different performance metrics of the model for the first task, which involved classifying five heart valvular conditions. Regardless of the input features, the experimental results indicate that the model performance improved with an increase in the number of concatenated stages. This suggests that integrating multistage features enhances the ability of the model to capture relevant patterns in the data. Moreover, this improvement highlights the

Table 6
Average and standard deviation of performance metrics for distinguishing five heart conditions.

Fusion	Features	Metric	N	MVP	MS	MR	AS	Mean (std)
Three stages	STFT	Specificity	100	99.62	99.75	99.88	99.88	**99.83 (0.13)**
		F1-score	100	98.50	99.00	99.50	99.50	**99.30 (0.51)**
	CWT	Specificity	100	99.50	99.50	99.62	99.88	99.70 (0.20)
		F1-score	100	98.00	98.76	97.47	99.75	98.80 (0.97)
	EMD+HHT	Specificity	99.88	98.88	99.38	99.25	99.12	99.30 (0.33)
		F1-score	99.50	94.97	97.76	96.22	97.52	97.20 (1.53)
	EWT+HHT	Specificity	99.88	99.62	99.50	99.75	99.75	99.70 (0.13)
		F1-score	99.25	98.50	98.76	98.49	99.00	98.80 (0.30)
	VMD+HHT	Specificity	100	100	99.25	100	99.88	**99.83 (0.30)**
		F1-score	100	99.50	98.52	98.73	99.75	**99.30 (0.58)**
First and second stages	STFT	Specificity	99.88	99.62	99.50	99.50	99.88	**99.68 (0.17)**
		F1-score	99.75	97.99	98.51	98.00	99.25	**98.70 (0.70)**
	CWT	Specificity	100	99.75	99.38	99.62	99.12	99.58 (0.30)
		F1-score	100	96.94	98.51	97.73	98.28	98.30 (1.01)
	EMD+HHT	Specificity	99.75	99.50	99.50	99.38	99.50	99.53 (0.12)
		F1-score	99.25	96.97	98.76	97.24	98.25	98.10 (0.87)
	EWT+HHT	Specificity	99.88	99.88	99.50	99.50	99.38	99.63 (0.21)
		F1-score	99.50	98.48	99.01	97.49	98.01	98.50 (0.71)
	VMD+HHT	Specificity	100	99.88	99.50	99.75	99.25	**99.68 (0.27)**
		F1-score	99.75	97.97	99.01	98.49	98.27	**98.70 (0.63)**
First stage	STFT	Specificity	99.75	98.88	99.12	99.50	99.62	**99.38 (0.32)**
		F1-score	99.50	94.97	96.24	97.74	99.00	**97.50 (1.70)**
	CWT	Specificity	100	99.38	99.12	98.62	99.38	99.30 (0.45)
		F1-score	100	95.94	97.01	95.54	97.50	97.20 (1.57)
	EMD+HHT	Specificity	99.38	98.75	98.75	98.38	99.00	98.85 (0.33)
		F1-score	97.76	93.40	95.26	93.23	97.28	95.40 (1.89)
	EWT+HHT	Specificity	99.62	99.38	99.25	98.88	99.38	99.30 (0.24)
		F1-score	97.73	96.73	98.27	95.50	97.76	97.20 (0.98)
	VMD+HHT	Specificity	99.88	99.50	99.25	99.25	99.12	**99.40 (0.27)**
		F1-score	99.50	96.45	98.27	95.96	97.78	**97.60 (1.27)**

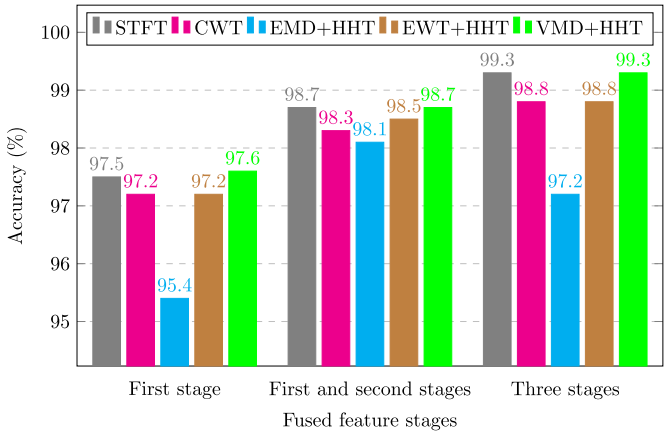


Fig. 7. Average accuracy of detection of five heart valvular conditions.

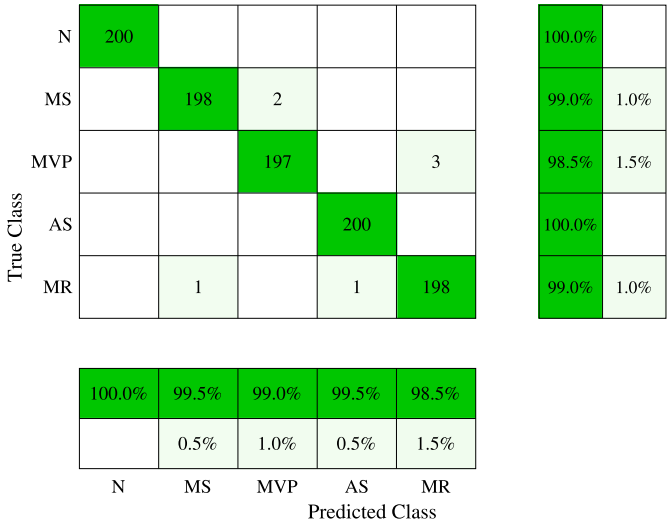


Fig. 8. The confusion matrix of the five heart valvular conditions for 5-fold CV obtained using the best-performing model. This model comprises three stages with internal feature fusion, and it is trained and tested on features extracted via STFT or VMD combined with HHT.

importance of feature fusion in enhancing the discriminative power of the model for an accurate classification.

Using time–frequency features obtained with AMRA based on EWT and VMD combined with HHT to train and evaluate the model demonstrated competitive performance in distinguishing the five heart valvular conditions compared with those obtained with MRA. Furthermore, the model architecture that fuses three-stage features achieves an impressive accuracy of 99.30% averaged over a 5-fold data split. This high level of accuracy underscores the effectiveness of combining multistage features to capture the complex nature of heart sounds. In contrast, when EMD was combined with HHT, the model yielded the worst accuracy among the methods tested. This could be attributed to the inherent limitations of the EMD in handling nonstationary signals, which may lead to less effective feature extraction and, consequently, poorer model performance.

Fig. 8 shows the confusion matrix for heart failure detection obtained using the best-performing model. This model consists of a three-stage structure with internal feature fusion and is trained and tested on features extracted from the STFT or VMD combined with HHT. The highest accuracy was achieved for healthy hearts (N) and aortic stenosis (AS), with all 200 examples correctly classified. In contrast, mitral valve prolapse (MVP) has the fewest correctly classified cases. The classification performance for mitral stenosis (MS) and mitral regurgitation (MR) is between that for MVP and both *N* and AS.

Table 7
Average of performance metrics for detecting the presence of heart murmurs.

Fusion	Features	Metric	Ten cycles		Five cycles		One cycle	
			Absent	Present	Absent	Present	Absent	Present
Three stages	STFT	Specificity	83.80	95.53	**83.24**	**94.97**	74.86	85.47
		F1-score	90.24	89.02	**89.70**	**88.43**	81.17	79.06
	CWT	Specificity	**88.27**	**92.18**	88.27	87.15	**77.09**	**87.15**
		F1-score	**90.41**	**90.03**	87.64	87.78	**82.98**	**81.18**
	EMD+HHT	Specificity	76.54	87.71	75.42	87.71	70.95	87.71
		F1-score	83.07	81.07	82.63	80.36	80.93	77.44
	EWT+HHT	Specificity	80.45	84.36	72.07	88.83	69.27	85.47
		F1-score	82.74	82.05	81.96	78.66	79.07	75.38
	VMD+HHT	Specificity	77.65	86.03	74.86	87.15	76.54	84.92
		F1-score	82.57	81.05	82.11	79.76	81.50	79.88
First and second stages	STFT	Specificity	82.68	91.06	**81.56**	**91.06**	**77.65**	**89.39**
		F1-score	87.40	86.30	**86.93**	**85.63**	**84.43**	**82.49**
	CWT	Specificity	**83.80**	**91.06**	81.56	88.83	77.09	83.24
		F1-score	**87.87**	**86.96**	85.71	84.64	80.76	79.54
	EMD+HHT	Specificity	79.33	86.59	73.18	91.06	67.60	85.47
		F1-score	83.56	82.23	83.59	80.37	78.46	74.23
	EWT+HHT	Specificity	77.09	87.71	70.95	77.93	72.63	82.68
		F1-score	83.29	81.42	77.37	76.28	78.72	76.47
	VMD+HHT	Specificity	75.42	92.18	65.36	88.83	68.72	87.71
		F1-score	85.05	82.32	79.50	74.05	80.10	75.93
First stage	STFT	Specificity	**81.01**	**88.83**	**80.45**	**93.30**	**73.18**	**87.71**
		F1-score	**85.48**	**84.30**	**87.66**	**85.97**	**81.77**	**78.92**
	CWT	Specificity	**81.01**	**88.83**	78.77	86.03	75.42	84.36
		F1-score	**85.48**	**84.30**	83.02	81.74	80.75	78.95
	EMD+HHT	Specificity	73.18	86.03	74.86	87.15	65.92	83.80
		F1-score	80.84	78.21	82.11	79.76	76.92	72.39
	EWT+HHT	Specificity	77.09	84.36	69.27	85.47	70.95	78.77
		F1-score	81.40	80.00	79.07	75.38	75.81	73.84
	VMD+HHT	Specificity	73.18	89.94	69.27	87.71	69.83	86.59
		F1-score	82.99	79.88	80.31	76.31	79.90	76.22

6.6.2. Results of the second dataset

Table 7 reports the performance metrics of specificity and F1-score for the same task. For recordings that contained fewer than ten cardiac cycles, we appended zero matrices to ensure uniformity across all datasets. Fig. 9 shows the accuracy of the system for detecting heart murmurs in the second task when trained using concatenated time–frequency feature matrices of ten, five, or one cardiac cycle. As observed, the values of all metrics improved with an increasing number of cardiac cycles and fused stages of internal features. Using time–frequency feature matrices derived from MRA based on both STFT and CWT to train and evaluate the model showed a notable improvement in the detection of heart murmurs compared with features obtained through AMRA. This performance can be attributed to the robustness of the MRA in capturing the essential characteristics of the PCG signals. The ability of MRA to capture data variability, which often challenges AMRA, plays a significant role in achieving high accuracy.

Fig. 10 shows the confusion matrix for murmur detection using the best-performing network model. This model employs a three-stage structure with internal feature fusion and is trained and tested on features extracted via CWT applied to ten cardiac cycles. The model successfully identified murmur presence and absence, with 164 and 159 correctly classified examples out of 179, respectively. These values highlight the potential of the model for the early detection and diagnosis of cardiac conditions.

The model employs a three-stage internal fusion of features, with each stage contributing spectro-temporal features, achieving an accuracy of 90.20% averaged over 5-fold CV when using features extracted based on CWT. This indicates that the multistage approach not only enhances the ability of the model to capture complex patterns, but also contributes to its overall robustness and generalizability across different datasets. Moreover, when the model was trained and evaluated using features extracted based on the STFT, the resulting performance was comparable to that achieved with CWT-based features. This consistency in high performance across different feature extraction techniques suggests that the model is well-suited for handling the nuances of PCG signals. Both the STFT and CWT offer complementary insights into the spectro-temporal representation of the signal, and the model effectively leverages these insights. Additionally, the multistage feature fusion approach likely enhances the ability of the model to differentiate subtle variations in heart conditions, leading to a more accurate classification across various heart valvular diseases. This adaptability across feature types underscores the versatility and robustness of the network architecture for analyzing PCG signals to detect heart murmurs.

6.7. Comparison results

Table 8 reports the experimental results of the proposed method compared with recently developed baseline methods that used the first dataset analyzed in this study.

The first baseline is based on raw PCG signals or their Fourier components (transformed PCG signals) using a deep CNN-LSTM classification model [28] that consists of 4 convolutional layers, 4 batch normalization layers, 4 ReLUs, 3 max-pooling layers, and 2 LSTM layers and is trained using ADAM with 10-fold cross-validation.

The second baseline method is based on VMD and a light CNN-LSTM model [40]. The PCG signals were decomposed into a finite set of IMFs using VMD, and the features were obtained by applying a weighted logarithmic operation to the IMFs. The model comprises one convolutional layer, one ReLU layer, and one LSTM layer and is trained and evaluated 1000 times separately, with random selection of the training and testing sets.

In the third baseline method [7], the authors introduced multiple deep networks based on CNNs and encoders. Each network processes inputs, including 1D PCG signals or 2D images, such as spectrograms, mel-spectrograms, and bispectral analysis. The best-performing network consists of 5 convolutional layers, 5 normalization layers, and 4 encoder layers. The final results were obtained by calculating the median of 5 training sessions.

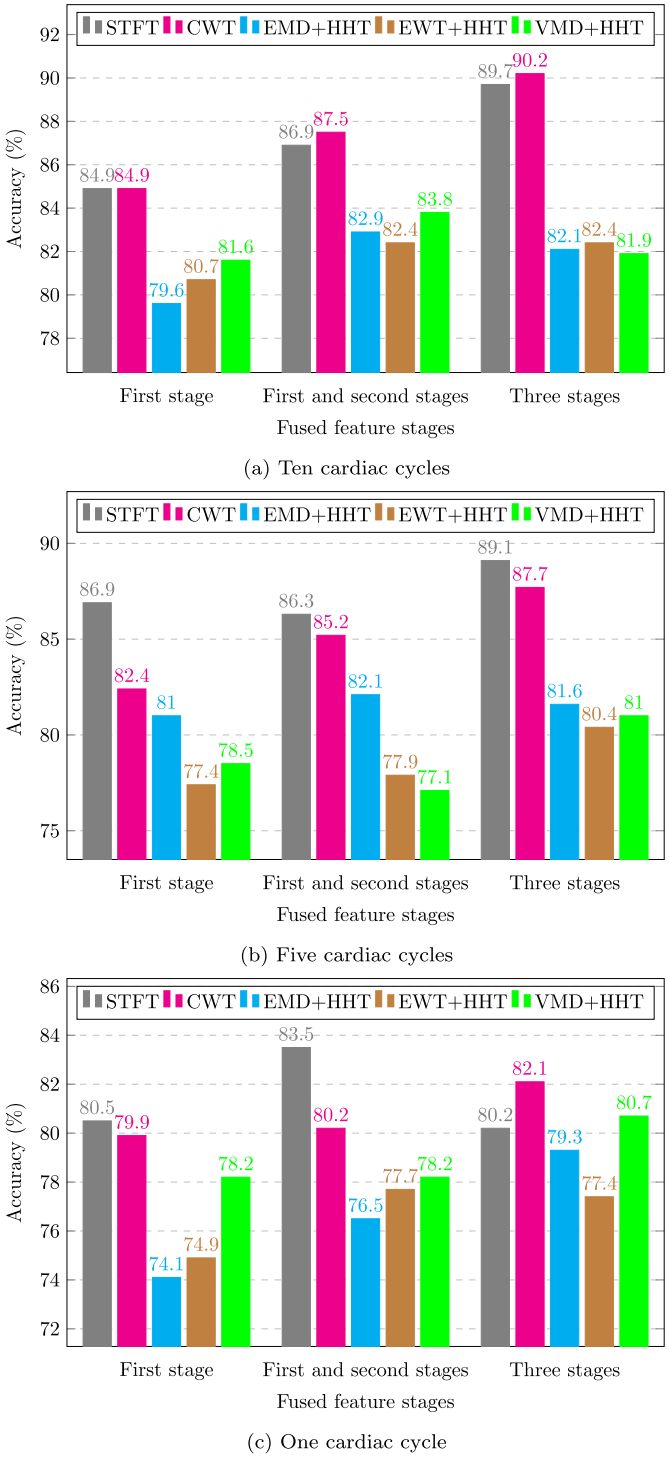


Fig. 9. Average accuracy of heart murmur detection using feature matrices of different numbers of cycles.

The fourth baseline method, which is based on the Gabor dictionary and 1D + 2D CNN-LSTM [5], estimates a coefficient vector through elastic net regularization by projecting the PCG signals onto the Gabor dictionary. The features were extracted using a weighted logarithmic operation on the time–frequency matrix derived from the coefficient vector. The model, consisting of 1D and 2D convolutional layers, 2 ReLUs, and an LSTM layer, was trained 1000 times using random training and testing data splits.

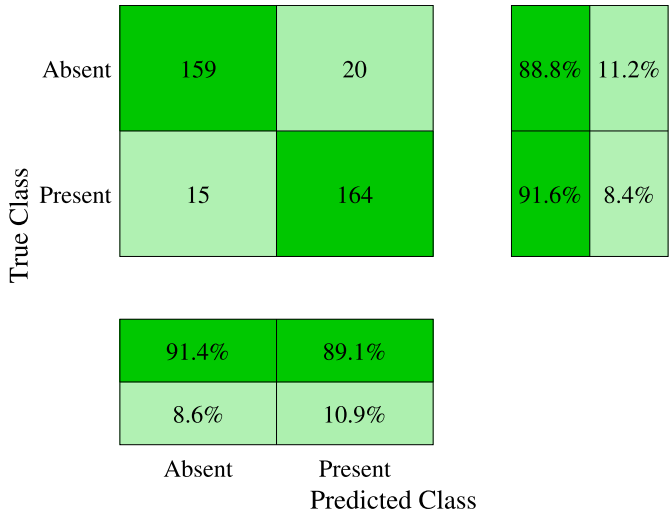


Fig. 10. The confusion matrix of murmur detection for all 5-fold CV obtained using the best-performing model. This model consists of three stages with internal feature fusion, and it is trained and tested on features extracted by applying CWT to ten cardiac cycles.

The fifth baseline is based on second-order spectral analysis and a parallel CNNs-transformer network [8]. They applied second-order spectral analysis to PCG signals and used a transformer and two CNNs to extract hierarchical features, which were then fused for the final classification. Each CNN is composed of three blocks, each containing convolutional, batch normalization, maximum pooling, and dropout layers. The transformer used four encoder modules, each containing four independent self-attention heads.

The table also presents the accuracy of the developed model, which employs the optimal feature extraction method (STFT or VMD combined with HHT) and a hybrid neural network with three-stage feature fusion. The results demonstrate a significant improvement in accuracy, outperforming the four baseline methods. This highlights the high efficiency of the model, which surpasses those of both lightweight and sequentially connected deep networks. The marked accuracy gain further underscores the robustness and adaptability of the model in capturing complex signal patterns.

The developed system achieves a minimum relative reduction in the classification error of 53.95% compared to the first baseline. This finding indicates that time–frequency features offer superior representations of PCG signals compared with raw PCG signals or their Fourier transformations. However, the relative error reduction for the second and third baselines is approximately 48.15%. This outcome suggests that generating time–frequency representations of IMFs using HHT after applying VMD to PCG signals yields more effective features than treating the IMFs as separate time signals. It also highlights that combining multiple time–frequency representations with deep models may lead to overfitting and reduced classification accuracy. Regarding the fourth baseline method, the developed system achieved a relative reduction in the classification error of 33.33%. This finding suggests that time–frequency features from the STFT, which uses complex exponential basis functions, outperform those from the elastic net regularization with the Gabor dictionary, which relies on scaled, translated, and modulated Gaussian window functions. The developed system achieved a classification accuracy comparable to that of the fifth method, even though the latter relies on a deep network with numerous layers, which requires significantly more computational resources and processing time.

While the proposed model demonstrates strong classification performance, it presents several limitations. First, the generalizability of

Table 8
Comparison of the proposed method with baseline methods.

Ref.	Features	Classifier	Accuracy
[28]	Raw PCG signals	Deep 2D CNN-LSTM	98.48
	Transformed PCG signals		95.40
[40]	VMD + weighted logarithmic	Light 1D CNN-LSTM	98.65
[7]	Raw PCG or 4 2D images	Deep CNN encoder	98.70
[5]	Gabor dictionary($\beta = 2^1$)+ elastic net ($\alpha = 0.1$)	1D+2D CNN-LSTM	98.95
[8]	MFCCs	Fusion of parallel 2 CNN	99.25
	Second-order spectral analysis	and transformer	99.36
Prop.	STFT+ weighted logarithmic	Multistage feature fusion	99.30
	VMD+HHT+ weighted logarithmic	of CNN and LSTM layers	99.30

the model to diverse clinical settings remains uncertain owing to the lack of external validation on unseen datasets. The system may also be sensitive to preprocessing quality, and the presence of class imbalance or label noise in the datasets may affect its robustness. Lastly, the high reported accuracy could indicate possible overfitting, especially if the datasets used for training and evaluation were small or not sufficiently diverse.

7. Conclusion and future work

This study introduced a deep hybrid network with an internal feature fusion strategy for detecting heart failure and murmurs based on phonocardiogram (PCG) signals. We thoroughly explored the application of adaptive multiresolution analysis (AMRA) and multiresolution analysis (MRA) to extract time–frequency features from PCG signals for network training and evaluation. Extensive experiments on two specialized datasets demonstrate the effectiveness of the proposed system for targeted diagnostic tasks.

The proposed network architecture utilizes three cascaded convolutional neural networks, each connected to its own branch, comprising average pooling and long short-term memory layers. The outputs from these branches are merged through a concatenation layer, effectively integrating their unique feature representations. This multistage fusion strategy captures complementary patterns at varying resolutions, enhancing the ability of the model to learn rich and discriminative features, leading to a notable improvement in classification performance.

Classical MRA techniques, such as the short-time Fourier transform and continuous wavelet transform, offer a robust framework for representing PCG signals, as confirmed by the exceptional classification performance of the model in detecting heart failure and murmurs using MRA-based time–frequency features. Although MRA effectively captures the dynamics of PCG signals, it is less adaptive to high variability compared to AMRA. In contrast, AMRA uses the Hilbert–Huang transform along with adaptive decomposition methods, such as the empirical wavelet transform or variational mode decomposition, which are based on signal-derived basis functions. This adaptive nature allows AMRA to excel in more complex classification tasks, particularly in distinguishing between five categories of heart valvular conditions.

To mitigate the limitations of the developed system and improve its robustness, future research should focus on three pivotal directions. First, we aim to investigate adaptive feature selection mechanisms within the hybrid network to dynamically prioritize the most relevant time–frequency features. Second, we plan to explore more sophisticated feature fusion techniques beyond simple concatenation, such as attention-based fusion and learnable fusion weights. Finally, we intend to develop a real-time PCG analysis system that can be deployed on mobile devices for point-of-care diagnosis and integrated with electronic health records to provide decision-making support for clinicians. These future directions offer a comprehensive roadmap for extending research and ultimately translating it into practical and clinical applications.

CRediT authorship contribution statement

Mahmoud Fakhry: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Ascensión Gallardo-Antolín:** Writing – review & editing, Supervision, Methodology.

Ethical approval

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

Funding information

This study did not receive any funding in any form.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] World health organization, Cardiovascular diseases (CVDs), 2019, URL https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.
[2] B. Phibbs, The Human Heart: A Basic Guide to Heart Disease, Lippincott Williams & Wilkins (LWW), 2007.
[3] N. Ranganathan, V. Sivaciyan, F.B. Saksena, The art and science of cardiac physical examination, in: Contemporary Cardiology, Humana Totowa, NJ, 2007.
[4] H.B. Sprague, P.A. Ongley, The clinical value of phonocardiography, Circulation 9 (1954) 127–134.
[5] M. Fakhry, A. Gallardo-Antolín, Elastic net regularization and Gabor dictionary for classification of heart sound signals using deep learning, Eng. Appl. Artif. Intell. 127 (2024) 107406.
[6] S.K. Ghosh, R.K. Tripathy, R.N. Ponnalagu, A study on time-frequency analysis of phonocardiogram signals, in: S. Goel (Ed.), Microelectronics and Signal Processing: Advanced Concepts and Applications, CRC Press, Boca Raton, 2021.
[7] J. Wang, J. Zang, S. Yao, Z. Zhang, C. Xue, Multiclassification for heart sound signals under multiple networks and multi-view feature, Measurement (2023).
[8] R. Wang, Y. Duan, Y. Li, D. Zheng, X. Liu, C.T. Lam, T. Tan, PCTMF-Net: heart sound classification with parallel CNNs-transformer and second-order spectral analysis, Vis. Comput. 39 (2023) 3811–3822.
[9] W. Wang, Z. Guo, J. Yang, Y. Zhang, L.-G. Durand, M. Loew, Analysis of the first heart sound using the matching pursuit method, Med. Biol. Eng. Comput. 39 (2001) 644–648.
[10] M. Fakhry, A.F. Brery, A. Gallardo-Antolín, Analysis of heart sound signals using sparse modeling with gabor dictionary, in: The 24th IEEE International Symposium on Multimedia (ISM), Naples, Italy, 5–7 December 2022, 2022, pp. 92–96.
[11] M. Fakhry, A.F. Brery, Comparison of window shapes and lengths in short-time feature extraction for classification of heart sound signals, Int. J. Electr. Comput. Eng. (IJECE) (2022).

[12] J.J. Lee, S.M. Lee, I.Y. Kim, H.K. Min, S.-H. Hong, Comparison between short-time Fourier and wavelet transform for feature extraction of heart sound, in: Proceedings of IEEE Region 10 Conference. TENCON 99, Vol. 2, 1999, pp. 1547–1550.
[13] A.K. Abbas, R. Bassam, R.M. Kasim, Mitral regurgitation PCG-Signal classification based on adaptive db-wavelet, 2008, IFMBE.
[14] F. Meziani, S.M. Debbal, A. Atbi, Analysis of phonocardiogram signals using wavelet transform, J. Med. Eng. Technol. 36 (2012) 283–302.
[15] O. Rioul, P. Flandrin, Time-scale energy distributions: a general class extending wavelet transforms, IEEE Trans. Signal Process. 40 (1992) 1746–1757.
[16] Z. Guo, L.-G. Durand, H. Lee, Comparison of time-frequency distribution techniques for analysis of simulated Doppler ultrasound signals of the femoral artery, IEEE Trans. Biomed. Eng. 41 (1994) 332–342.
[17] L. Senhadji, G. Carrault, J.-J. Bellanger, G. Passariello, Comparing wavelet transforms for recognizing cardiac patterns, IEEE Eng. Med. Biol. Mag. 14 (1995) 167–173.
[18] O. Bertrand, J. Bohorquez, J. Pernier, Time-frequency digital filtering based on an invertible wavelet transform: an application to evoked potentials, IEEE Trans. Biomed. Eng. 41 (1994) 77–88.
[19] L.H. Cherif, N. Benmessaoud, S.M. Debbal, Comparison between analysing wavelets in continuous wavelet transform based on the fast Fourier transform: application to estimate pulmonary arterial hypertension by heart sound, Int. J. Biomed. Eng. Technol. (2021).
[20] S. Patidar, R.B. Pachori, A continuous wavelet transform based method for detecting heart valve disorders using phonocardiograph signals, in: International Conference on Hybrid Information Technology, 2012.
[21] N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng, N. Yen, C.C. Tung, H.H. Liu, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 454 (1998) 903–995.
[22] K. Dragomiretskiy, D. Zosso, Variational mode decomposition, IEEE Trans. Signal Process. 62 (2014) 531–544.
[23] J. Gilles, Empirical wavelet transform, IEEE Trans. Signal Process. 61 (2013) 3999–4010.
[24] J. Gelpud, S. Castillo, M. Jojoa, B. Garcia-Zapirain, W. Achicanoy, D. Rodrigo, Deep learning for heart sounds classification using scalograms and automatic segmentation of PCG signals, 2021, IWANN.
[25] M. Alkhodari, L. Fraiwan, Convolutional and recurrent neural networks for the detection of valvular heart diseases in phonocardiogram recordings, Comput. Methods Programs Biomed. 200 (2021) 105940.
[26] M. Fakhry, A.F. Brery, A comparison study on training optimization algorithms in the biLSTM neural network for classification of PCG signals, in: 2022 2nd IRASET, 2022.
[27] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[28] Y. Al-Issa, A.M. Alqudah, A lightweight hybrid deep learning system for cardiac valvular disease classification, Sci. Rep. 12 (2022).
[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE CVPR, 2015, pp. 770–778.
[30] T. Li, C. Qing, X. Tian, Classification of heart sounds based on convolutional neural network, in: International Conference on Internet Multimedia Computing and Service, 2017.
[31] F. Demir, A. Şengür, V. Bajaj, K. Polat, Towards the classification of heart sounds based on convolutional deep neural network, Heal. Inf. Sci. Syst. 7 (2019) 1–9.
[32] J.S. Khan, M. Kaushik, A. Chaurasia, M.K. Dutta, R. Burget, Cardi-Net: A deep neural network for classification of cardiac disease using phonocardiogram signal, Comput. Methods Programs Biomed. 219 (2022) 106727.
[33] K.N. Khan, F.A. Khan, A. Abid, T. Olmez, Z. Dokur, A. Khandakar, M.E.H. Chowdhury, M.S. Khan, Deep learning based classification of unsegmented phonocardiogram spectrograms leveraging transfer learning, Physiol. Meas. 42 (2020).
[34] A. Meintjes, A. Lowe, M. Legget, Fundamental heart sound classification using the continuous wavelet transform and convolutional neural networks, in: 40th Annual International Conference of the IEEE EMBC, 2018, pp. 409–412.
[35] A.W. Sugiyarto, A.M. Abadi, S. Sumarna, Classification of heart disease based on PCG signal using convolutional neural network (CNN), TELKOMNIKA Telecommun. Comput. Electron. Control. 19 (2021).
[36] P.K. Jain, R.R. Choudhary, M.R. Singh, A lightweight 1-D convolution neural network model for Multi-class classification of heart sounds, 2022, pp. 40–44, 2022 (ICETCI).
[37] P. Qi, H. Xu, H. Zhang, J. Tong, S. Xia, Residual neural networks based on empirical mode decomposition for mitral regurgitation prediction, Biomed. Signal Process. Control. 86 (2023) 105265.
[38] K.A. Babu, B. Ramkumar, Automatic recognition of fundamental heart sound segments from PCG corrupted with lung sounds and speech, IEEE Access J. 8 (2020) 179983–179994.
[39] W. Zeng, J. Yuan, C. Yuan, Q. Wang, F. Liu, Y. Wang, A new approach for the detection of abnormal heart sound signals using TQWT, VMD and neural networks, Artif. Intell. Rev. 54 (2020) 1613–1647.
[40] M. Fakhry, A. Gallardo-Antolín, Variational mode decomposition and a light CNN-LSTM model for classification of heart sound signals, in: IEEE EUROCON 2023, Torino, Italy, 6 – 8 June, 2023, pp. 1–6.
[41] K.R. Ranipa, W. Zhu, M.N.S. Swamy, Multimodal CNN fusion architecture with multi-features for heart sound classification, in: 2021 IEEE ISCAS, 2021, pp. 1–5.
[42] J. Li, H. Liu, K. Li, K. Shan, Heart sound classification based on two-channel feature fusion and dual attention mechanism, in: 2024 5th ICCEA, 2024, pp. 1294–1297.
[43] Y. Jang, J. Jung, Y. Hong, J. Lee, H. Jeong, H. Shim, H.-J. Chang, Fully convolutional hybrid fusion network with heterogeneous representations for identification of S_1 and S_2 from phonocardiogram, IEEE J. Biomed. Heal. Inform. 28 (2024) 7151–7163.
[44] M. Li, Z. He, H. Wang, Heart sound classification based on Multi-Scale feature fusion and channel attention module, Bioengineering 12 (2025).
[45] W. Xiong, G. Zhang, D. Yan, L. Cao, X. Huang, D. Li, Multichannel feature fusion network-based technique for heart sound signal classification and recognition, Expert. Syst. Appl. 273 (2025) 126839.
[46] E. Sejdić, I. Djurović, J. Jiang, Time-frequency feature representation using energy concentration: An overview of recent advances, Digit. Signal Process. 19 (2009) 153–183.
[47] C. Torrence, G.P. Compo, A practical guide to wavelet analysis, Bull. Am. Meteorol. Soc. 79 (1998) 61–78.
[48] R. Yan, R.X. Gao, Hilbert–Huang Transform-Based vibration signal analysis for machine health monitoring, IEEE Trans. Instrum. Meas. 55 (2006) 2320–2329.
[49] Yaseen, G.-Y. Son, S. Kwon, Classification of heart sound signal using multiple features, Appl. Sci. 8 (12) (2018).
[50] J. Oliveira, F. Renna, P.D. Costa, M. Nogueira, C. Oliveira, C.A. Ferreira, A.M. Jorge, S. da Silva Mattos, T.d. Hatem, T. Tavares, A. Elola, A.B. Rad, R. Sameni, G.D. Clifford, M.T. Coimbra, The CirCor DigiScope dataset: From murmur detection to Murmur classification, IEEE J. Biomed. Heal. Inform. 26 (2021) 2524–2535.