Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 34(3): 1832–1856.
DOI: 10.3934/era.2026082
Received: 01 September 2025
Revised: 06 December 2025
Accepted: 14 February 2026
Published: 28 February 2026

Research article

Classification of systolic murmurs in heart sounds using multiresolution complex Gabor dictionary and vision transformer

Mahmoud Fakhry[1,*]**and Abeer FathAllah Brery**[2]

[1] CEIEC, Universidad Francisco de Vitoria, Pozuelo de Alarcón, Madrid 28223, Spain

[2] Departamento de Informática, Universidad Carlos III de Madrid, Leganés, Madrid 28911, Spain

* **Correspondence:** Email: mahmoud.fakhry@ufv.es.

Abstract: Systolic murmurs are extra heart sounds that occur during the contraction phase of the cardiac cycle, often indicating heart abnormalities caused by turbulent blood flow. Their intensity, pitch, and quality vary, requiring precise identification for the accurate diagnosis of cardiac disorders. This study presents an automatic classification system for systolic murmurs using a feature extraction module, followed by a classification model. The feature extraction module employs complex orthogonal matching pursuit to project single or multiple murmur segments onto a redundant dictionary composed of multiresolution complex Gabor basis functions (GBFs). The resulting projection weights are split and reshaped into variable-resolution time–frequency feature matrices. Processing multiple segments of a single recording using a shared dictionary mitigates murmur variability. This is achieved by learning the weights for each segment while enforcing that they correspond to the same set of basis functions in the dictionary, promoting consistent time–frequency feature matrices. The classification model is built based on a vision transformer to process multiple input matrices of different resolutions by passing each through a convolutional neural network for patch tokenization. All embedding tokens are then concatenated to form a matrix and forwarded to an encoder layer that includes multihead attention, residual connections, and a convolutional network with a kernel size of one. This integration of multiresolution feature extraction with transformer-based feature classification enhances the accuracy and reliability of heart murmur identification. An experimental analysis of four types of systolic murmurs from the CirCor DigiScope dataset demonstrates the effectiveness of the system, achieving a classification accuracy of 95.96%.

Keywords: cardiovascular disease; systolic murmurs; heart sounds; multiresolution complex Gabor dictionary; vision transformer

1. Introduction

Cardiovascular diseases (CVDs) are among the leading causes of mortality worldwide, highlighting the urgent need for automatic diagnostic tools that enable early monitoring and detection [1]. Recordings of heart sounds, or phonocardiogram (PCG) signals, offer a noninvasive and low-cost means of identifying abnormalities associated with CVDs [2]. Heart murmurs are abnormal heart sounds caused by turbulent blood flow, which provide critical indicators of potential cardiac problems, such as valve abnormalities or congenital heart conditions. The cardiac cycle of the PCG signals consists of two primary phases: diastole (expansion) and systole (contraction), as shown in Figure 1. Systolic heart murmurs occur during the contraction phase, between the first heartbeat (S_1) and second heartbeat (S_2), whereas diastolic heart murmurs manifest during the expansion phase, between S_2 and S_1.

The analysis of heart sound signals has evolved significantly, progressing from the use of traditional methods such as time- and frequency-domain techniques to more sophisticated approaches in the time–frequency domain [3]. Although time-domain analysis provides valuable insight into the temporal characteristics of signals, and frequency domain methods effectively uncover spectral patterns, both approaches may face limitations in handling the nonstationary and complex nature of heart murmurs. time–frequency techniques such as the short-time Fourier transform and wavelet transform offer a more comprehensive understanding by capturing the dynamic interplay between the temporal and spectral features. These methods are particularly effective for analyzing heart murmurs that vary over cardiac cycles, thereby enhancing their characterization.

Challenges such as individual variability and overlapping murmur components underscore the need for more advanced methods. Innovations in signal processing have provided new avenues for addressing these challenges, including dictionaries of predesigned basis functions such as Gabor functions [4]. By offering precise localization in both the time and frequency domains, these advanced tools enable a deeper understanding of the intricate features of heart murmurs, allowing the isolation of pathological characteristics from overlapping components [5, 6]. Integrating these techniques with deep learning models enhances the accuracy of feature extraction and classification, enabling efficient automated diagnosis of heart disease, especially in resource-limited settings.

Advances in deep learning have enabled the automated analysis of signals with increasing accuracy [7]. Recurrent neural networks (RNNs) and their variants, such as long short-term memory (LSTM) networks, are used to model sequential data. However, models built on RNNs often struggle with long-term dependencies, which hinders their performance for temporally long signals. Convolutional neural networks (CNNs) have become the dominant approach for analyzing various types of signals, including heart sound signals, due to their ability to extract hierarchical features. Despite their strengths, CNNs have limitations in capturing global dependencies across the entire signal and are susceptible to overfitting, particularly when trained on small datasets. Transformer models have recently emerged as powerful tools for signal and image processing [8]. Their exceptional performance is attributed to their ability to model long-range dependency and capture hierarchical representations. Transformers can capture the global context using self-attention mechanisms, thus addressing the limitations of CNNs and RNNs, which struggle with long-range dependencies. Unlike RNNs and their variants, transformers process entire sequences in parallel, enhancing their computational efficiency and making them well-suited for heart murmur analysis. A hybrid combination of CNNs for local feature extraction and transformers for global context analysis is expected to yield optimal results.

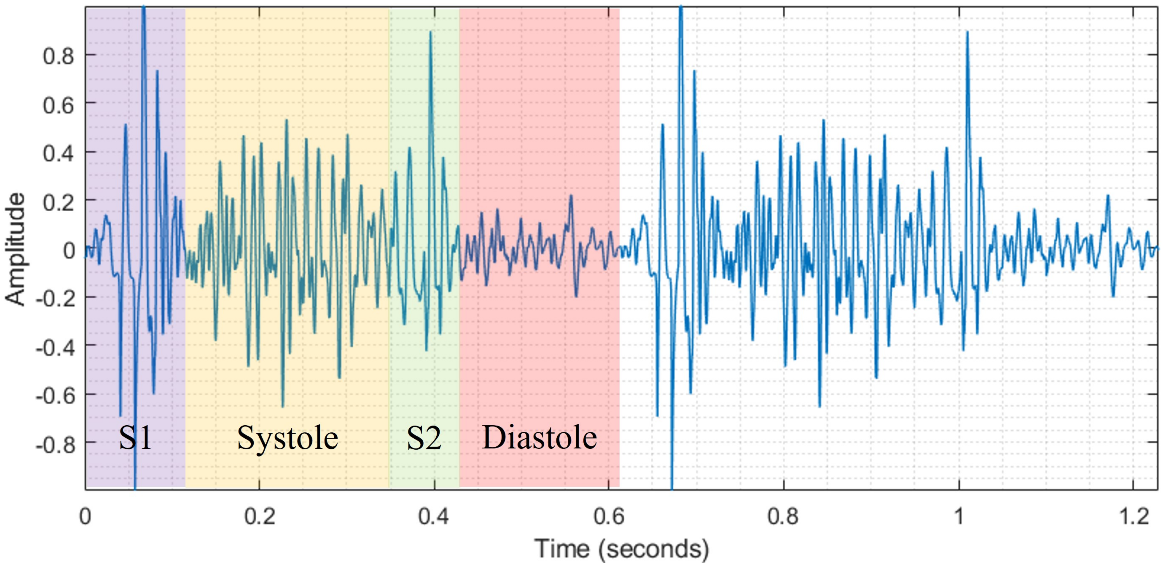

Figure 1. Two cardiac cycles of heart sound, with systolic murmurs of diamond type that are located between the first (S_1) and second (S_2) heartbeats.

Time–frequency representations of signals are commonly employed to transform one-dimensional (1D) signals into two-dimensional (2D) image-like matrices. These structured 2D representations offer a particularly suitable input format for vision transformers (ViTs), which are deep models designed for image classification using a transformer-based architecture applied to image patches [9]. By partitioning these representations into nonoverlapping patches, ViTs can efficiently extract localized spectral and temporal features while preserving the hierarchical structure of the data. This mechanism enables the model to capture complex patterns and subtle variations in heart sound signals, thereby enhancing its ability to identify pathological conditions with high sensitivity and specificity. Additionally, the high-dimensional features derived from these time–frequency representations provide rich contextual information, further improving the robustness and interpretability of ViT-based models in analyzing heart sound signals and murmurs.

In this study, we introduce a system for the automatic classification and identification of murmurs in abnormal heart sound signals. Based on the preceding discussion, the main novel contributions of this study are as follows:

1) Develop a feature extraction module to calculate multiresolution time–frequency features from murmur segments using the multiresolution complex Gabor dictionary and complex orthogonal matching pursuit.

2) Develop a classification model that uses multiresolution time–frequency features to identify different types of murmurs in abnormal heart sound signals through a hybrid combination of a CNN and a vision transformer.

3) Perform a comprehensive experimental analysis of the system on a publicly available dataset of sound recordings of pathological hearts, which includes four types of systolic heart murmurs.

Systolic murmurs are the most common in clinical practice and are the most consistently labeled in the CirCor DigiScope dataset [10] used in our research. They provide four well-defined morphological subclasses (crescendo, decrescendo, plateau, diamond), enabling rigorous supervised learning and benchmarking of the proposed time–frequency fusion model. In contrast, diastolic murmurs are scarce, lack consistent subclass definitions, and exhibit high intra-annotator uncertainties. For these reasons, systolic murmurs were selected as the primary target for methodological development.

The remainder of this paper is organized as follows: Section 2 introduces systolic heart murmurs and the dataset used to evaluate the developed system. Methods for the analysis of heart sound signals and related studies are presented in Section 3. In Section 4, we explain the developed systems in detail. The experimental analysis and results are discussed in Section 5. Finally, the conclusions of this study are summarized in Section 6.

2. Systolic heart murmurs and dataset

Systolic heart murmurs manifest as structural abnormalities, such as valvular stenosis or regurgitation, which disrupt normal unidirectional blood flow [11]. For instance, aortic stenosis produces harsh crescendo–decrescendo murmurs as blood is forcefully ejected through a narrowed aortic valve. Simultaneously, mitral regurgitation generates holosystolic murmurs due to backflow of blood into the left atrium during ventricular contraction. In some cases, systolic murmurs may also be caused by high-output states, such as hyperthyroidism, which increases blood flow velocity, or by congenital defects such as ventricular septal defects. These sounds vary widely in intensity, pitch, and quality, ranging from soft and blowing to loud and harsh, and often require careful auscultation and diagnostic techniques for their accurate identification and assessment. Understanding the characteristics and causes of systolic murmurs is critical in clinical practice, as they can serve as important clues for diagnosing and managing heart conditions. Below is a detailed description of the four common patterns of systolic, which are represented in Figure 2 as follows:

- **Diamond**: These murmurs increase in intensity during the early part of systole, reach a peak, and then decrease. They are often associated with aortic or pulmonic stenosis, in which turbulent blood flow occurs because of a narrow valve opening. The characteristics of these murmurs reflect changes in blood flow velocity during systole, providing clues for stenosis identification.
- **Plateau**: These murmurs maintain a relatively constant intensity throughout systole, resembling a flat or rectangular profile when visualized on a phonocardiogram. They are frequently associated with mitral regurgitation or ventricular septal defects, where consistent regurgitation flow or shunting occurs due to defective valve closure or septal abnormalities, respectively.
- **Decrescendo**: These murmurs start loudly at the beginning of systole and gradually decrease in intensity as systole progresses. They are typically associated with conditions such as aortic or pulmonic regurgitation, in which blood leaks backward into the ventricles during systolic contractions. This declining intensity reduces the pressure gradient as the ventricle empties.
- **Crescendo**: These murmurs begin softly and progressively increase in intensity, peaking either midsystole or late in systole. This pattern is associated with aortic stenosis or hypertrophic cardiomyopathy, in which the obstruction to blood flow progressively increases as the heart contracts. The crescendo shape highlights the increasing severity of turbulence during systole.

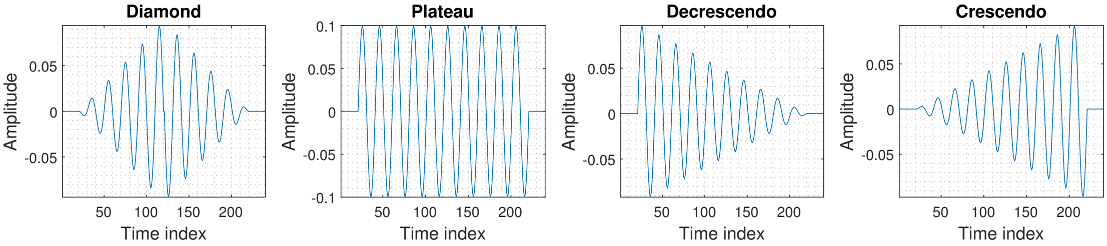

Figure 2. Shapes of different types of systolic heart murmurs in the time domain.

Table 1. Number of signal samples/number of segments for all types of murmurs.

Location	Patterns				Total
	Diamond	Plateau	Decrescendo	Crescendo	
AP	9/167	7/91	2/87	1/14	19/359
PP	10/283	36/787	14/420	1/13	61/1503
MP	4/32	30/750	8/293	0	42/1075
TP	8/164	38/979	10/257	0	56/1400
Total	31/646	111/2607	34/1057	2/27	178/4337

2.1. *Dataset*

The dataset used in the experimental analysis was extracted from the CirCor DigiScope dataset [10]. Heart sound recordings were acquired from four anatomical locations on the chest: the aortic (AP), pulmonary (PP), mitral (MP), and tricuspid points (TP). These recordings were sampled at a high resolution of 4000 Hz to capture the fine details of heart sound characteristics. They were normalized to the range $[-1, 1]$ to standardize the amplitude of the input signal. The normalized PCG signals were then automatically segmented using state-of-the-art algorithms to accurately determine the boundaries and positions of the heartbeat components S_1 and S_2 as well as the systole and diastole phases, ensuring precision in the extraction of clinically significant segments.

All heart sound recordings were systematically screened for murmurs at each auscultation site, with the majority defined as having systolic murmurs. Although diastolic murmurs have a lower prevalence in the dataset, they are clinically critical, often signaling severe conditions such as aortic regurgitation or mitral stenosis. To evaluate the feasibility of extending our framework, we examined the annotation structure of the dataset. Although 34 recordings (~8.9% of murmur-positive cases) were labeled as containing *diastolic murmurs*, none include morphological subtype annotations. In contrast, systolic murmurs have 4337 expert-annotated segments across four well-defined shape classes, enabling supervised shape-based classification.

Each recording was carefully segmented to isolate multiple systolic murmur segments, which were assigned a specific label corresponding to one of the four predefined murmur types. This structured labeling process ensures consistency and accuracy in data annotation, providing a robust framework for subsequent pattern recognition and classification processes. Table 1 summarizes the dataset composition, detailing the total number of signal samples, distribution of systolic segments with various murmur patterns, and specific locations where they were recorded. The average segment length is approximately 500 samples, highlighting the short but information-rich nature of these segments.

3. Analysis of heart sound signals and related studies

Traditional methods for analyzing heart sound signals include time, frequency, and time–frequency domain techniques. Time-domain analysis involves analyzing the amplitudes of the waveforms [12, 13]. This approach can identify transient features such as the onset and offset of murmurs. Frequency-domain methods, such as the Fourier transform, reveal the spectral components of signals [14, 15]. These methods excel in analyzing periodicity and dominant frequencies, which are often associated with structural anomalies, such as valve defects. However, they assume signal stationarity, which limits their applicability to nonstationary signals. time–frequency domain techniques, including the short-time Fourier transform and wavelet transforms, unify temporal and spectral information for the comprehensive analysis of nonstationary signals [16,17]. These methods enable the dynamic monitoring of frequency components over time, allowing for a more accurate characterization of murmurs that evolve during each cardiac cycle.

Efficient analysis of PCG signals, particularly in the presence of overlapping murmurs and other cardiac events, requires advanced techniques [18]. Recent advances in signal processing have introduced dictionaries of predesigned basis functions for signal analysis [19]. Among these, the Gabor dictionary, which comprises real Gaussian-windowed sinusoids localized in both time and frequency, is particularly suitable for analyzing heart murmurs [20, 21]. The complex Gabor dictionary extends the traditional dictionary by incorporating complex-valued functions and providing additional phase information. This added dimension of analysis helps distinguish between different feature patterns in complex signal data. In [22], we proposed an overcomplete multiresolution complex Gabor dictionary composed of complex Gabor functions with diverse time–frequency resolutions to create a hyperspectral space for detailed signal analysis. Using complex orthogonal matching pursuit, a systolic murmur segment is represented as a sparse, complex vector based on the dictionary. The calculated energy distribution across resolutions revealed that each type of systolic murmur exhibited unique patterns, enabling its differentiation.

In [23], the authors proposed a mathematical model for extracting crucial features from systolic murmurs using a Gabor dictionary. The extracted features were classified using a feedforward multilayer perceptron. In [24], the classification of heart murmurs was thoroughly explored using selected convolutional neural network models, such as VGGNet and ResNet. This study leveraged various signal representations, including spectrograms, Mel-frequency cepstral coefficients, and short-time Fourier transforms. A multiscale attention convolutional compression network was proposed in [25] to detect coronary artery disease. The network uses a multiscale convolution structure to capture comprehensive features and a channel attention module to enhance these features. Vision transformers (ViTs) were used to classify heart sound signals based on spectrograms in [26]. Similarly, ViTs have been applied to classify PCG signals using bispectrum-inspired features [27]. In [28], the authors proposed an ensemble approach for classifying PCG signals. This method leverages the complementary strengths of CNNs and ViTs using a mixture of expert frameworks. A hybrid convolution–transformer neural network was introduced in [29]. The study demonstrated the feasibility of using a transformer-based method to detect coronary heart disease in a clinical dataset of heart sound recordings. In [30], the authors introduced a parallel CNN transformer network to classify PCG signals. They applied a second-order spectral analysis and used a transformer along with two CNNs to extract hierarchical features, which were then fused for final classification.

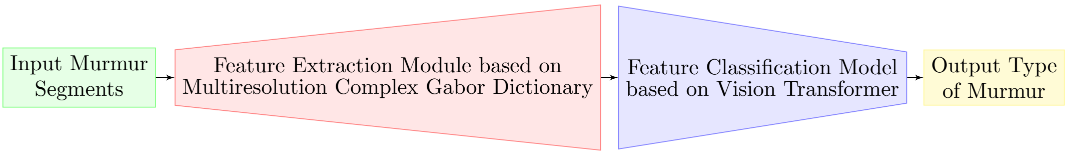
Input Murmur Segments
Feature Extraction Module based on Multiresolution Complex Gabor Dictionary
Feature Classification Model based on Vision Transformer
Output Type of Murmur

Figure 3. Block diagram of the developed system.

Expanding on our analysis of systolic heart murmurs presented in [22], we propose a novel feature extraction module that uses a multiresolution complex Gabor dictionary in conjunction with complex orthogonal matching pursuit to capture intrinsic time–frequency features with different resolutions. Furthermore, inspired by the widespread success of vision transformers in various domains, we introduce a hybrid feature classification model that integrates convolutional neural networks with a vision transformer to capture local feature patterns and enhance global contextual understanding.

4. Developed system

We present an innovative system tailored for the classification and identification of systolic heart murmurs by leveraging advanced signal processing and deep learning techniques. The system consists of two key components: a feature extraction module and a feature classification model. The feature extraction module systematically analyzes systolic heart murmurs and transforms them into representative time–frequency feature matrices using advanced decomposition and transformation techniques. These extracted features serve as inputs for the classification model, which employs deep learning models to achieve high accuracy and robustness. Each component is meticulously designed to enhance performance and ensure the reliable identification of systolic murmurs in diverse patient populations. Figure 3 shows a block diagram of the proposed system.

4.1. Feature extraction module

The feature extraction module is developed to capture the intricate time–frequency characteristics of heart murmurs, as shown in Figure 4. The module employs a multiresolution dictionary $\mathbf{D}$ composed of complex Gabor atoms that enables the representation of murmur signals across varying time–frequency resolutions. To achieve this, murmur segments are projected onto the dictionary, allowing the extraction of detailed signal features. The projection weights $\mathbf{a}$ are computed using complex orthogonal matching pursuit (COMP), which is a sophisticated algorithm optimized for sparse signal representation. The projection weights are divided into groups, each corresponding to a different resolution. These groups are then reshaped individually to construct 2D time–frequency feature matrices $\mathbf{A}$ that capture various temporal and spectral information scales. This approach ensures that the extracted features effectively capture the temporal and spectral variations in heart murmur signals. By providing multiple resolutions of time–frequency features, the module enhances the capability of the system to discern subtle differences between different types of heart murmurs. Figure 5 illustrates how a murmur segment is transformed into multiple time–frequency representations with varying resolutions. Figure 6 demonstrates that two segments from the same recording exhibit nearly identical support in the dictionary, although their complex weights differ in phase and magnitude. This enables the reshaped matrices to share identical spectrotemporal atom locations but with adaptive intensity and phase modulation. Such consistency is unattainable if phase is discarded, as the atoms selected per segment become decoupled.

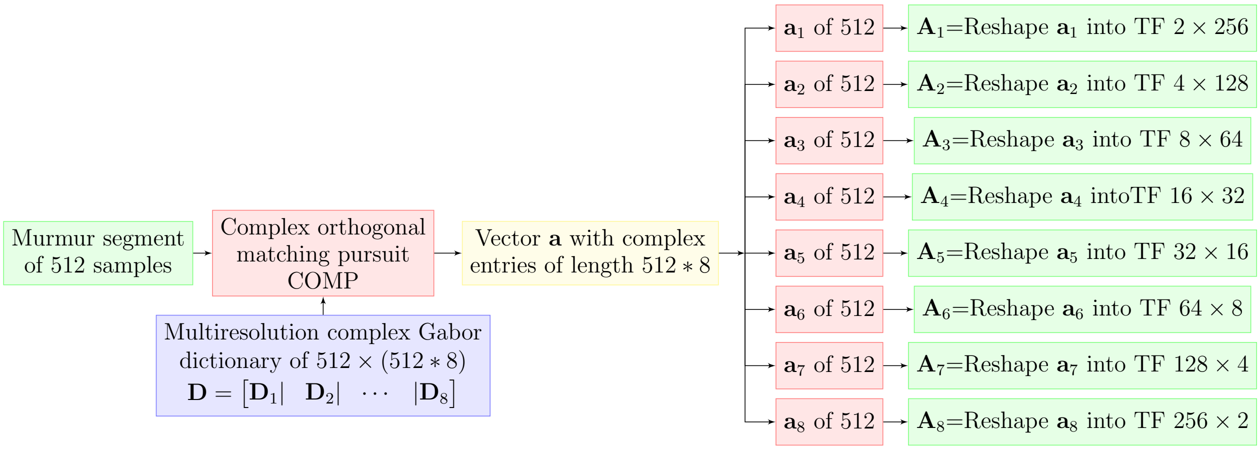
Murmur segment of 512 samples
Complex orthogonal matching pursuit COMP
Multiresolution complex Gabor dictionary of 512 × (512 ∗ 8)
$\mathbf{D} = [\mathbf{D}_1| \quad \mathbf{D}_2| \quad \cdots \quad |\mathbf{D}_8]$
Vector $\mathbf{a}$ with complex entries of length 512 ∗ 8
$\mathbf{a}_1$ of 512
$\mathbf{a}_2$ of 512
$\mathbf{a}_3$ of 512
$\mathbf{a}_4$ of 512
$\mathbf{a}_5$ of 512
$\mathbf{a}_6$ of 512
$\mathbf{a}_7$ of 512
$\mathbf{a}_8$ of 512
$\mathbf{A}_1$=Reshape $\mathbf{a}_1$ into TF 2 × 256
$\mathbf{A}_2$=Reshape $\mathbf{a}_2$ into TF 4 × 128
$\mathbf{A}_3$=Reshape $\mathbf{a}_3$ into TF 8 × 64
$\mathbf{A}_4$=Reshape $\mathbf{a}_4$ intoTF 16 × 32
$\mathbf{A}_5$=Reshape $\mathbf{a}_5$ into TF 32 × 16
$\mathbf{A}_6$=Reshape $\mathbf{a}_6$ into TF 64 × 8
$\mathbf{A}_7$=Reshape $\mathbf{a}_7$ into TF 128 × 4
$\mathbf{A}_8$=Reshape $\mathbf{a}_8$ into TF 256 × 2

Figure 4. Block diagram of the developed feature extraction module, mainly built based on complex orthogonal matching pursuit in conjunction with a multiresolution complex Gabor dictionary. The segment length is 512 samples, resulting in a Gabor dictionary with 8 different time–frequency resolutions.

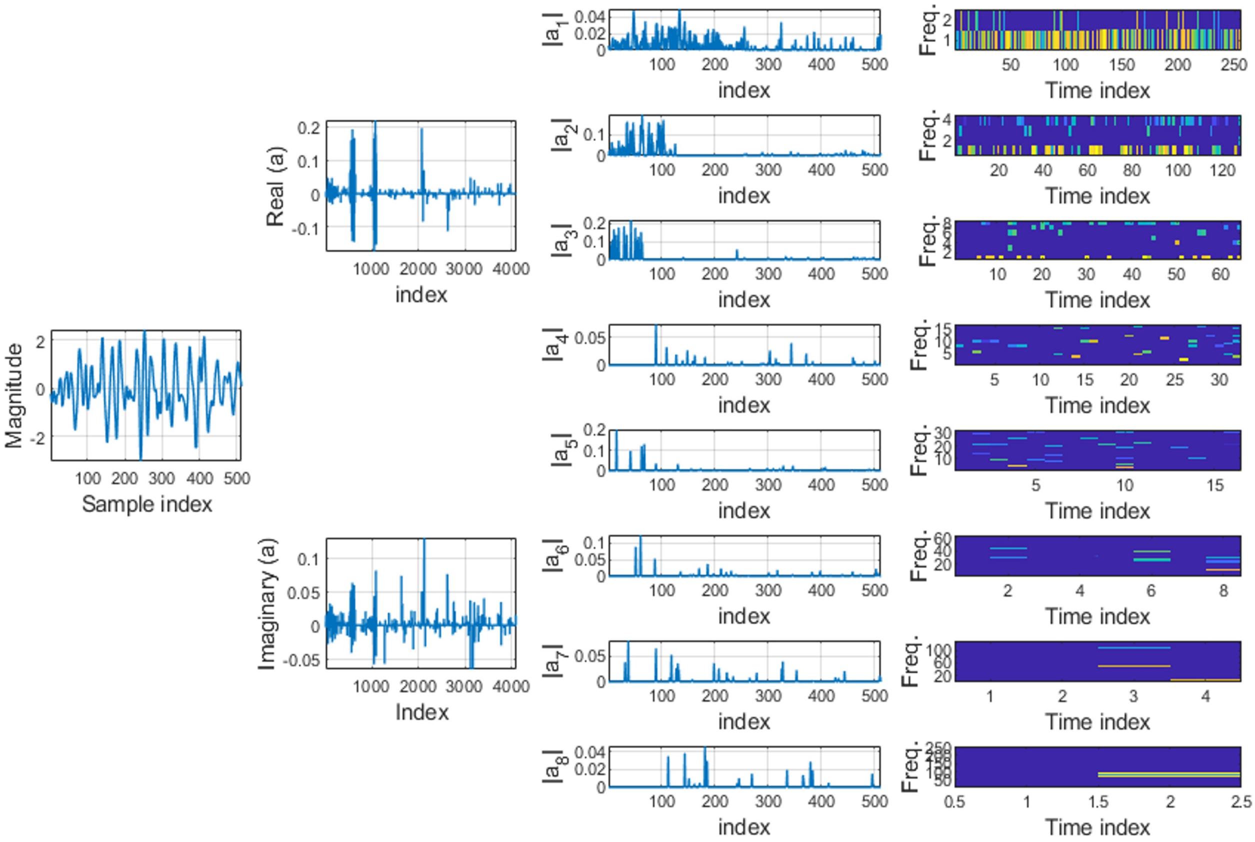

Figure 5. Feature extraction from a murmur segment of diamond type using COMP and the developed multiresolution complex Gabor dictionary. Columns from left to right: murmur segment, real and imaginary parts of the vector of projection weights $\mathbf{a}$, absolute value of $\mathbf{a}_j$, and reshaped time–frequency representations $\mathbf{A}_j$ of the absolute of $\mathbf{a}_j$.

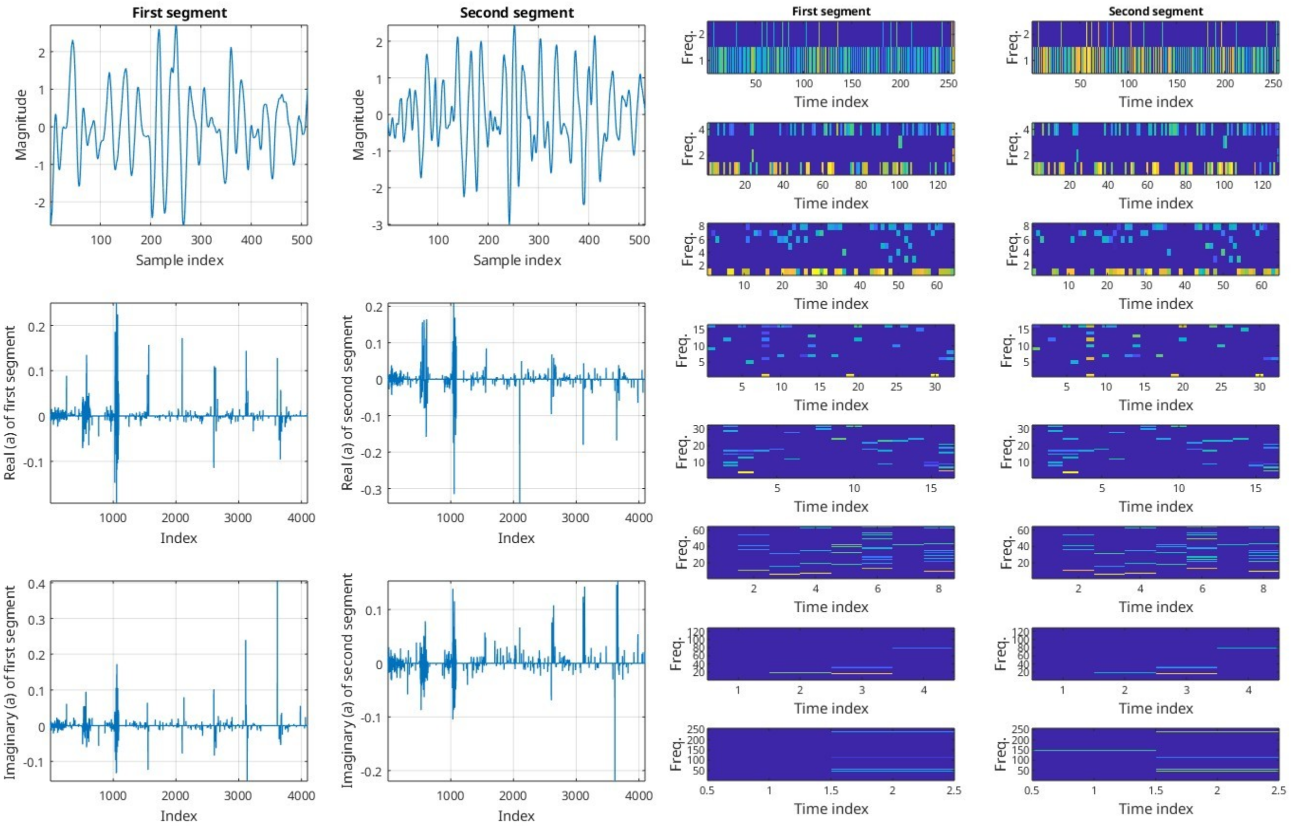

Figure 6. Feature extraction from two murmur segments from the same heart sound recording of diamond type decomposed together using COMP and the developed multiresolution complex Gabor dictionary. Columns from left to right: the first murmur segment and the real and imaginary of its decomposition, $\mathbf{a}$ followed by the second murmur segment and the real and imaginary of its decomposition $\mathbf{a}$, and the reshaped time–frequency representations $\mathbf{A}_j$ of the absolute of the corresponding $\mathbf{a}_j$ for each segment.

4.1.1. Complex Gabor dictionary

Unlike Fourier bases, Gabor atoms overlap, providing a flexible and adaptive representation that aligns better with the structures of real-world signals. This flexibility stems from the construction of Gabor atoms using a Gaussian window function, which is known for optimal time–frequency localization. The width of the Gaussian window can be scaled to allow fine-tuning of the resolution in the time and frequency domains. Because the window is correlated with the signal, the Gabor atoms capture localized features, making them effective for signals with transient or nonstationary properties.

In addition, modulating the Gaussian window generates oscillatory components, allowing the representation of quasiperiodic patterns in the signal. This combination of scaling, translation, and modulation creates a redundant but highly descriptive dictionary of atoms that can provide a compact and precise decomposition of complex signals. This adaptability makes Gabor transforms and related methods particularly valuable in applications such as speech processing, image analysis, and biomedical signal interpretation, where the ability to capture both temporal and spectral characteristics is crucial. Heart murmurs differ not only in amplitude envelopes, but also in micro-oscillation patterns and localized

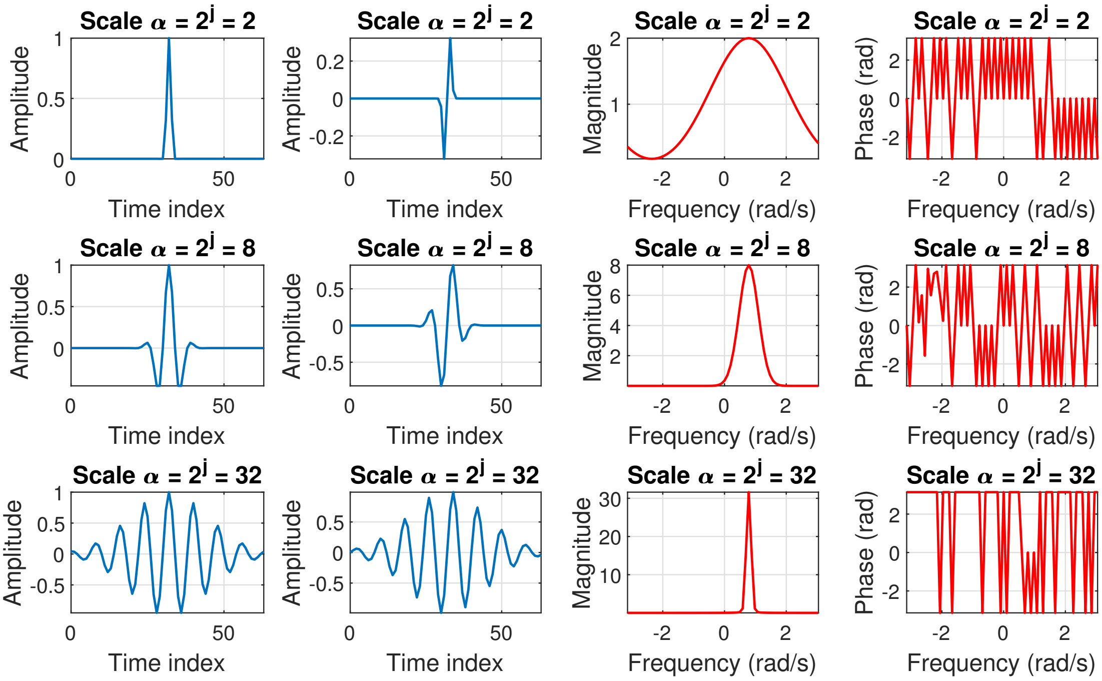

Figure 7. Examples of complex atoms with three different resolutions for $\alpha = 2^j$ values, with $j = 1, 3, 5$. The first two columns on the left show the real and imaginary parts of the atoms in the time domain, respectively. The third and fourth columns show the magnitudes and phases of the corresponding frequency representations, respectively. Plots are for $M = 64$, $\omega = \pi/2$, and $m_0 = 32$. Higher j increases spectral but reduces temporal resolution.

phase rotations caused by turbulent flow. Complex Gabor atoms encode these phase variations, which real-valued atoms cannot recover. The complex Gabor atom is defined as, per [22],

$$d_{(\alpha,m_0,\omega)}(m) = e^{-\pi(\frac{m-m_0}{\alpha})^2} e^{-i(\omega(m-m_0))}, \tag{4.1}$$

where $i = \sqrt{-1}$, and $m = 0, 1, 2, ..., M-1$. The Gaussian window limits the time extent of the complex Gabor function, thereby allowing localized analysis. To maintain the overlap among the Gabor atoms at 50%, the scale parameter α was first chosen as a power of 2, $\alpha = 2^j$ with $0 < j < \log_2(M)$. The translation m_0 and frequency ω parameters are then obtained as follows [31]:

$$m_0 = 2^j(0:1:2^{-j}M-1) \text{ and } \omega = 2^{-j}(0:1:2^j-1)2\pi. \tag{4.2}$$

The total number of frequency bins ω is equal to 2^j for each translation step m_0 for the total number of steps $2^{-j}M$. The number of atoms is $2^{-j}M \times 2^j = M$. The jth Gabor dictionary is built for a specific value of j by arranging the column atoms side-by-side in a matrix, that is, $\mathbf{D}_j = [\mathbf{d}_1\mathbf{d}_2...\mathbf{d}_m] \in \mathbb{C}^{M\times M}$. Figure 7 shows plots of three examples of complex atoms with their corresponding frequency representations.

The mathematical derivation of the Gabor dictionary parameters (α, ω, m_0), the justification of the 50% overlap criterion, and the validation of COMP sparsity levels have been comprehensively presented in our previously published work [22]. To avoid redundancy, we refer readers to that publication for the complete theoretical and experimental validation underlying the dictionary construction.

4.1.2. Multiresolution complex Gabor dictionary

In the above representation of the atoms, the resolution of the dictionary is intricately governed by parameter j, which determines the balance between the temporal and spectral resolutions. Specifically, as j increases, the real exponential function associated with the atom exhibits a slower decay around the central time m_0. This slower decay causes the duration of the atom to extend, resulting in a significant degradation of the temporal resolution. In other words, the atom becomes less precise when isolating transient events, leading to a smeared effect in the time domain. Moreover, the extended duration of an atom has a complementary effect on its spectrum. As the temporal span of the atom increases, its bandwidth decreases correspondingly, causing the frequency representation of the atom to become more narrowly focused around its center frequency ω. This bandwidth narrowing enhances the spectral resolution, making the dictionary more adept at capturing fine-grained details in the frequency domain. Consequently, the trade-off governed by j highlights the limitation of time–frequency representations, where improving the resolution in one domain inevitably comes at the expense of the resolution in the other. By carefully selecting j, the dictionary can be tailored to prioritize the temporal or spectral resolution, depending on the specific requirements of the signal analysis, underscoring the critical role of this parameter in constructing effective representations.

Because the time–frequency resolution varies with the variable j, the multiresolution Gabor dictionary $\mathbf{D} \in \mathbb{C}^{M \times JM}$ is formed by arranging the single-resolution dictionaries $\mathbf{D}_j \in \mathbb{C}^{M \times M}$ for $j = 1, \ldots, J$ next to each other in a large matrix such as

$$\mathbf{D} = \left[\mathbf{D}_1| \quad \mathbf{D}_2| \quad \cdots \quad |\mathbf{D}_J\right], \text{ with } J = \log_2(M) - 1. \tag{4.3}$$

4.1.3. Complex orthogonal matching pursuit (COMP)

The orthogonal matching pursuit (OMP) is a widely used greedy algorithm designed to obtain sparse solutions for linear systems by leveraging measurements and a predefined dictionary of potential basis elements. OMP operates iteratively by selecting the best-matching atoms, individual components, or dictionary columns that align most closely with the residual signal at each step. Starting with the original signal as the initial residual, the algorithm identifies the atom that maximizes its absolute correlation with the current residual, thereby capturing the most significant signal component in a given dictionary. Once an atom is selected, the OMP updates the approximation of the signal by solving a least-squares problem over the set of chosen atoms, ensuring that the resulting representation is optimal in terms of minimizing errors. The residual is then recomputed as the difference between the original signal and the current approximation, and the process is repeated until a specified stopping criterion is satisfied, such as a predefined sparsity level or an error threshold. This iterative selection and refinement process enables the OMP to construct a sparse approximation of the signal, making it particularly useful in applications such as compressed sensing, image processing, and feature selection, where sparsity is a desirable property for efficient representation and analysis.

Similar to the OMP, the complex orthogonal matching pursuit (COMP) finds a complex sparse vector $\mathbf{a} \in \mathbb{C}^{JM}$ given a measurement vector $\mathbf{x} \in \mathbb{R}^M$ and a dictionary $\mathbf{D} \in \mathbb{C}^{M \times JM}$, such that [32, 33]

$$\mathbf{x} \approx \mathbf{D}\mathbf{a}. \tag{4.4}$$

The best ζ-approximation vector $\mathbf{a}$ of the signal $\mathbf{x}$ is obtained by solving the following problem:

$$\arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 \;\text{ subject to }\; \|\mathbf{a}\|_0 \le \zeta, \tag{4.5}$$

Algorithm 1 Modeling single or multiple murmur segments using COMP

```
**Input:** single segment $\mathbf{x} \in \mathbb{R}^M$ or multiple segments $\mathbf{x}^1 \in \mathbb{R}^M, \mathbf{x}^2 \in \mathbb{R}^M, ..., \mathbf{x}^U \in \mathbb{R}^M$, dictionary $\mathbf{D} \in \mathbb{C}^{M \times JM}$, sparsity level $\zeta$
**Initialize:** residual $\mathbf{r}_0 = \mathbf{x}$ or $\mathbf{r}_0^1 = \mathbf{x}^1, \mathbf{r}_0^2 = \mathbf{x}^2, ..., \mathbf{r}_0^U = \mathbf{x}^U$, $\mathbf{a} \in \mathbf{0}^{JM}$ or $\mathbf{b}^1 \in \mathbf{0}^{JM}, \mathbf{b}^2 \in \mathbf{0}^{JM}, ..., \mathbf{b}^U \in \mathbf{0}^{JM}$, index set $\Lambda = \emptyset$, counter $c = 0$.
**while** $c < \zeta$ **do**
    $c = c + 1$
    $\mathbf{g} = |\mathbf{D}^H \mathbf{r}_{c-1}|$ or $= \sum_u |\mathbf{D}^H[\mathbf{r}_{c-1}^1, \mathbf{r}_{c-1}^2, ..., \mathbf{r}_{c-1}^U]|$    ▷ $.^H$ is the hermitian conjugate
    $i_c = \arg\max_i \mathbf{g}_i$    ▷ Find the index of a matched atom
    $\Lambda = \Lambda \cup \{i_c\}$    ▷ Update the index set of matched atoms
    $\mathbf{D}_\Lambda = [\mathbf{d}_{i_1}, \mathbf{d}_{i_2}, \ldots, \mathbf{d}_{i_c}]$    ▷ Form the sub-dictionary of matched atoms
    $\mathbf{a}_\Lambda = \mathbf{D}_\Lambda^\dagger \mathbf{x}$ or $[\mathbf{b}_\Lambda^1, \mathbf{b}_\Lambda^2, ..., \mathbf{b}_\Lambda^U] = \mathbf{D}_\Lambda^\dagger [\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^U]$    ▷ $.^\dagger$ is the pseudo inverse
    $\mathbf{r}_c = \mathbf{x} - \mathbf{D}_\Lambda \mathbf{a}_\Lambda$ or $[\mathbf{r}_c^1, \mathbf{r}_c^2, ..., \mathbf{r}_c^U] = [\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^U] - \mathbf{D}_\Lambda [\mathbf{b}_\Lambda^1, \mathbf{b}_\Lambda^2, ..., \mathbf{b}_\Lambda^U]$    ▷ Update
**end while**
**Output:** $\mathbf{a}$ or $[\mathbf{b}^1, \mathbf{b}^2, ..., \mathbf{b}^U]$ are with complex entries, $\mathbf{a}_\Lambda$ or $[\mathbf{b}_\Lambda^1, \mathbf{b}_\Lambda^2, ..., \mathbf{b}_\Lambda^U]$, respectively, at indices defined in $\Lambda$.
```

where $||\mathbf{x} - \mathbf{Da}||_2^2$ is the least-squares term that measures the reconstruction error between the original signal $\mathbf{x}$ and its approximation $\mathbf{Da}$, and $||\mathbf{a}||_0$ is the l_0-norm that promotes sparsity in $\mathbf{a}$. The problem in (4.5) can be solved using greedy methods, such as COMP. Recall that the main goal is to find the best ζ-term approximation of a given signal $\mathbf{x}$ by elements from the multiresolution complex Gabor dictionary $\mathbf{D}$ so that

$$\mathbf{x} \approx \mathbf{Da} = \begin{bmatrix} \mathbf{D}_1| & \mathbf{D}_2| & \cdots & |\mathbf{D}_J \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_J \end{bmatrix}. \tag{4.6}$$

The entries of vector $\mathbf{a}_j$ describe the contribution of atoms in the jth dictionary $\mathbf{D}_j$. For multiple murmur segments, the main goal is to find complex sparse multiple vectors $\mathbf{b}^1 \in \mathbb{C}^{JM}, \mathbf{b}^2 \in \mathbb{C}^{JM}, ..., \mathbf{b}^U \in \mathbb{C}^{JM}$ given multiple signals $\mathbf{x}^1 \in \mathbb{R}^M, \mathbf{x}^2 \in \mathbb{R}^M, ..., \mathbf{x}^U \in \mathbb{R}^M$ using the shared Gabor dictionary $\mathbf{D}$, so that

$$[\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^U] \approx \mathbf{D}[\mathbf{b}^1, \mathbf{b}^2, ..., \mathbf{b}^U] = \begin{bmatrix} \mathbf{D}_1| & \mathbf{D}_2| & \cdots & |\mathbf{D}_J \end{bmatrix} \begin{bmatrix} \begin{bmatrix} \mathbf{b}_1^1 \\ \mathbf{b}_2^1 \\ \vdots \\ \mathbf{b}_J^1 \end{bmatrix} \begin{bmatrix} \mathbf{b}_1^1 \\ \mathbf{b}_2^1 \\ \vdots \\ \mathbf{b}_J^1 \end{bmatrix}, ..., \begin{bmatrix} \mathbf{b}_1^U \\ \mathbf{b}_2^U \\ \vdots \\ \mathbf{b}_J^U \end{bmatrix} \end{bmatrix}. \tag{4.7}$$

The entries of vector $\mathbf{b}_j^u$ define the contribution of atoms in the j dictionary $\mathbf{D}_j$ to model the signal $\mathbf{x}^u$. Variable U denotes the number of segments, which varies depending on the number of segments in one recording. The COMP algorithm for modeling single and multiple murmur segments based on the multiresolution Gabor dictionary $\mathbf{D}_j$ is summarized in Algorithm 1.

Although OMP can be used with real atoms, it discards the Hermitian symmetry and phase of analytic complex Gabor atoms. Phase encodes instantaneous frequency drift and timing of peak turbulence, which are key discriminants for crescendo vs. diamond murmurs (Figure 2). The complex projection of COMP preserves this information in the weights $\mathbf{a} \in \mathbb{C}^{JM}$, which is leveraged implicitly in $||\mathbf{A}_j||^2$

and explicitly in the coherence modeling (Figure 6). COMP extends OMP to complex-valued atoms, preserving phase coherence, which is also critical for joint modeling of multiple segments.

Unlike conventional greedy algorithms such as fast matching pursuit (FastMP) [33], compressive sampling matching pursuit (CoSaMP) [34], or subspace pursuit (SP) [35], COMP supports joint sparse coding with a shared support Λ, enabling consistent time–frequency atom selection across variable murmur segments. This shared support is not achievable with real OMP or greedy alternatives without ad-hoc, post-hoc alignment. Moreover, FastMP, CoSaMP, and SP assume uniform, nonoverlapping supports, which are incompatible with the highly redundant multiresolution dictionary. Their iterative pruning mechanisms remove atoms from early resolutions, which degrades feature interpretability.

4.1.4. Reshaping vectors into matrices

To construct a time–frequency representation A_j, each coefficient vector a_j is reshaped into a matrix whose dimensions preserve the inherent structure of the atoms in the jth Gabor dictionary D_j. As defined in Eq (4.2), the dictionary at scale j contains 2^j frequency bins and 2^{8-j+1} translation steps, resulting in a total of $M = 512$ atoms. The complex coefficient vector $a_j \in \mathbb{C}^{512}$ produced by COMP thus naturally corresponds to a two-dimensional grid indexed by frequency (rows) and translation (columns) of the image. Reshaping a_j into a matrix $A_j \in \mathbb{R}^{2^j \times 2^{8-j+1}}$ preserves this intrinsic organization.

This aspect ratio reflects the resolution trade-off in the multiresolution dictionary: lower values of j provide high temporal and low spectral resolution, whereas higher values of j yield low temporal and high spectral resolution. By maintaining this structure, the reshaped matrices remain consistent with the scale-dependent time–frequency characteristics of systolic murmurs, capturing both the spectral concentration and temporal evolution typical of different murmur patterns. Figures 5 and 6 illustrate examples of the resulting representations for $M = 512$ and $J = 8$.

4.2. Feature classification model

The model is constructed using a transformer encoder by processing multiple time–frequency feature matrices of varying sizes, as shown in Figure 8. A feature matrix is obtained by calculating the squared magnitude of the entries of $|\mathbf{A}_j|^2$. Each matrix is then fed into a dedicated network branch comprising a two-dimensional convolutional neural network and a fully connected layer to reduce the dimensionality. The outputs from all branches are concatenated and normalized to form an embedding containing relevant information on systolic heart murmurs. The embedding computation process is detailed in Subsection 4.2.1. Subsequently, the embedding is passed to a multihead attention block to effectively model the interdependencies between all feature representations, as explained in Subsections 4.2.2 and 4.2.3. The output of the multihead attention is then normalized using a normalization layer to stabilize and accelerate training. The normalized output is forwarded to a feedforward network (FFN) consisting of a CNN with a kernel size of one, which enables channel-wise transformations while preserving the spatial dimensions, as explained in Subsection 4.2.4. Both the multihead attention and FFN are equipped with residual connections, as described in Subsection 4.2.5, which aim to improve the gradient flow. After processing through the transformer, the output is forwarded to a fully connected layer, followed by a softmax layer for final classification. This architectural design ensures that the model effectively integrates and processes diverse time–frequency feature matrices while leveraging the strengths of both convolutional operations and transformer mechanisms for a robust performance.

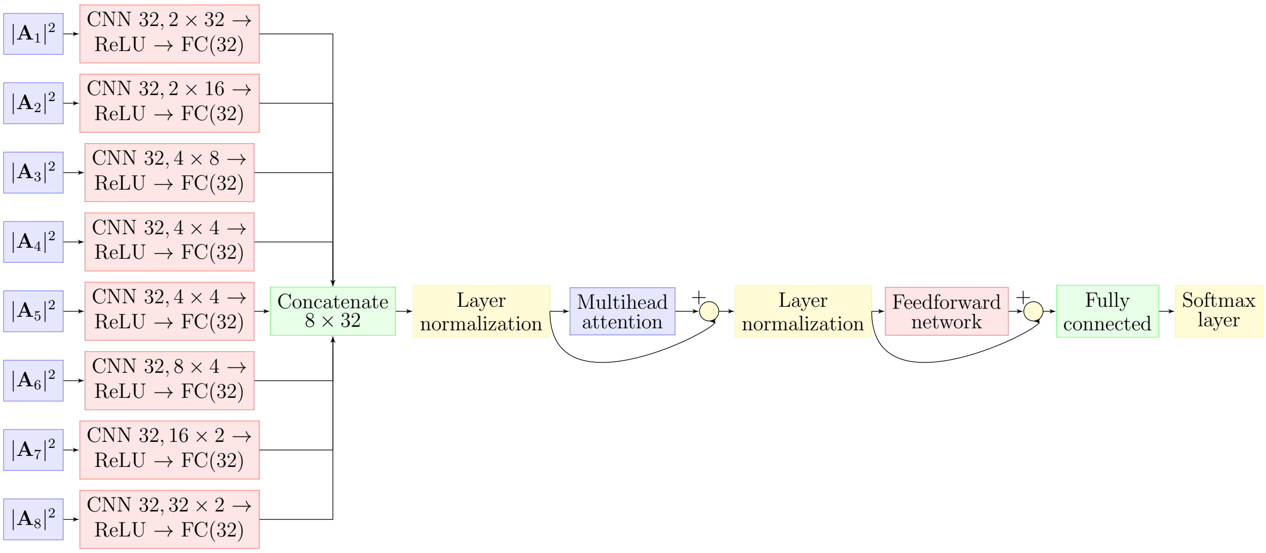

Figure 8. Block diagram of the developed vision transformer encoder-based model with multiple convolutional neural networks and one encoder layer.

The transformer architecture diverges from traditional recurrent networks, enabling parallel computation and achieving the highest performance for various tasks compared with other architectures. The transformer encoder, which is a critical component of this architecture, excels at modeling long-range dependencies within sequential data by leveraging self-attention mechanisms. This capability is indispensable for language modeling, time series analysis, and speech processing. The transformer encoder is a stack of identical layers, each comprising two main components: a multihead self-attention mechanism and a position-wise feedforward neural network. These components are augmented by residual connections and layer normalization to ensure stability and efficiency during training.

4.2.1. Patch embedding and tokenization

Patch embedding and tokenization processes involve dividing an input image into fixed-size patches, which are then tokenized into input tokens. Each patch is projected into a high-dimensional embedding space using a learnable linear projection implemented via 2D convolutional neural networks (CNNs), which enables compatibility with the transformer architectures. The output embedding for each patch is subsequently transformed into a 1D vector through a fully connected layer, where the output dimensionality corresponds to the number of filters used in the CNNs. All resulting token vectors are then concatenated to form a 2D matrix, where each row represents a token in a batch of length corresponding to the number of CNN filters. This matrix serves as an input representation for the transformer encoder. To enhance training stability, accelerate convergence, and reduce internal covariate shifts, a normalization layer is applied to the output of the embedding and tokenization stages.

4.2.2. Self-attention mechanism

The self-attention mechanism is a key feature of the transformer. Unlike recurrent networks, which process sequential inputs step by step, the self-attention mechanism simultaneously computes the attention scores for all input tokens. This enables the model to focus on the relevant parts of the input

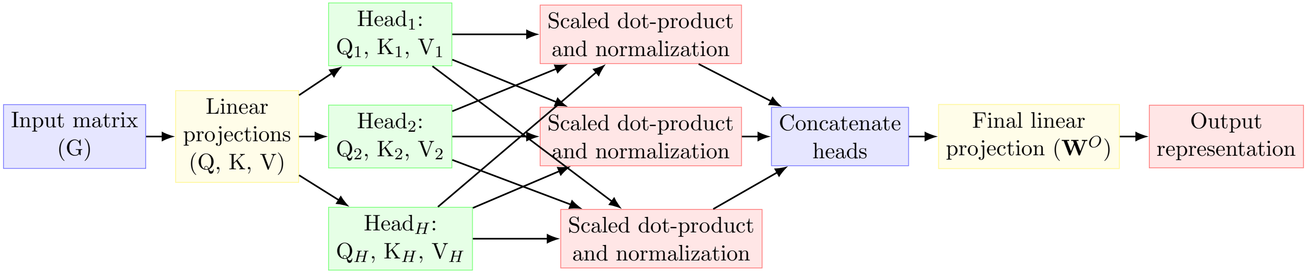
Input matrix (G)
Linear projections (Q, K, V)
Head$_1$: Q$_1$, K$_1$, V$_1$
Head$_2$: Q$_2$, K$_2$, V$_2$
Head$_H$: Q$_H$, K$_H$, V$_H$
Scaled dot-product and normalization
Scaled dot-product and normalization
Scaled dot-product and normalization
Concatenate heads
Final linear projection ($\mathbf{W}^O$)
Output representation

Figure 9. Block diagram of the multihead attention with H the total number of heads.

sequence regardless of their distance. For each token in the input sequence, the self-attention mechanism generates three vectors: query ($\mathbf{q}$), key ($\mathbf{k}$), and value ($\mathbf{v}$) vectors. These vectors are computed by applying learned weight matrices to the input embeddings. Given an input sequence represented as a matrix $\mathbf{G}$ (where each row is an input token embedding), the transformations are given by

$$\mathbf{Q} = \mathbf{G}\mathbf{W}_Q, \ \mathbf{K} = \mathbf{G}\mathbf{W}_K, \text{ and } \mathbf{V} = \mathbf{G}\mathbf{W}_V, \tag{4.8}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable weight matrices with dimensionality or number of channels equal to d_k. $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the matrices of query, key, and value for all input tokens, respectively. The transformed vectors are used to calculate attention scores by computing the scaled dot product between the query and key vectors. These scores are then normalized using a softmax function and applied to the value vectors, resulting in a weighted sum that captures the contextual representation of each token in the input sequence, as follows

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}. \tag{4.9}$$

4.2.3. Multihead attention

The multihead attention mechanism extends the single-head self-attention mechanism to capture various relationship patterns in data. It projects queries, keys, and values onto multiple subspaces (heads) using a learnable linear transformation. Each head independently computes its attention output, which is then concatenated and linearly transformed to obtain the final attention output. This diversity allows the model to address multiple aspects of the input simultaneously, thereby enhancing its representational capabilities. This process is expressed in (4.10).

$$\begin{aligned} \text{multihead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \text{Concatenate}(\text{head}_1, \ldots, \text{head}_H)\mathbf{W}^O \\ \text{head}_h &= \text{Attention}(\mathbf{Q}\mathbf{W}_Q^h, \mathbf{K}\mathbf{W}_K^h, \mathbf{V}\mathbf{W}_V^h), \end{aligned} \tag{4.10}$$

where $\mathbf{W}_Q^h$, $\mathbf{W}_K^h$, and $\mathbf{W}_V^h$ are learnable weight matrices for the corresponding h head. Finally, a linear transformation $\mathbf{W}^O$ combines the heads into the expected output dimension. Figure 9 illustrates the multihead attention mechanism.

4.2.4. Position-wise feedforward neural network

Following multihead attention, each token passes through an FFN that operates independently at every position. This FFN comprises a linear transformation using a CNN with a filter kernel of size one, followed by a ReLU, as expressed in (4.12). This enhances the capacity of the encoder to learn the complex transformations of the input embeddings.

4.2.5. Residual connections and layer normalization

To mitigate the vanishing gradient problem and enhance the gradient flow during backpropagation, the transformer encoder incorporates residual connections around both the multihead attention and FFN sublayers. These residual connections help to preserve information and facilitate more effective weight updates. Following each residual connection, layer normalization is applied to the outputs, as indicated in (4.11). This stabilizes training, accelerates convergence, and reduces internal covariate shifts. This combination of residual connections and layer normalization ensures smoother optimization, allowing the model to learn deeper representations.

$$\text{Output} = \text{LayerNorm}(\mathbf{z} + \text{sublayer}(\mathbf{z})). \tag{4.11}$$

$$\text{FFN}(\mathbf{y}) = \max(0, \mathbf{y}\mathbf{W} + \mathbf{b}). \tag{4.12}$$

5. Experimental analysis and results

The developed system was experimentally evaluated to measure its effectiveness in distinguishing between the four types of heart murmurs. The evaluation strategy involved assessing the performance of the system using standard classification metrics such as accuracy, specificity, and F1-score. Moreover, the confusion matrix effectively summarizes the classification performance of the model. The datasets were divided into 80% for training and 20% for testing, preserving the class distribution in both subsets to ensure that the model is trained and evaluated on representative examples of each murmur type. The evaluation results were carefully analyzed to identify strengths, limitations, and areas for improvement, guiding potential refinements to enhance performance.

5.1. System implementation

Table 2 presents the parameters governing the extraction of multiresolution time–frequency matrices derived from PCG signals along with the configuration details of the hybrid classification model. These parameters define the setup of the feature extraction module, including the size of the multiresolution complex Gabor dictionary and the dimensions of the resulting time–frequency matrices. It can be observed, for example, that the sparsity level ζ, is assigned a maximum value of 511, which matches the signal length (512) not because the method requires a full decomposition but because a large redundant dictionary is used (8 resolutions × 512 atoms each), which results in a minimum sparsity of $ratio = (4096 - 511)/4096 = 0.875$. Increasing such a value by reducing ζ results in high sparse models, which may underfit the training because the models are transformed into time–frequency matrices.

In addition, we evaluated several configurations of the model structure and training hyperparameters, and the parameter set reported in Table 2 corresponds to the combination that consistently produced the best performance and stable convergence. These optimal settings ensured effective learning without overfitting. The table outlines the key architectural elements of feature classification, such as the size of the convolutional kernels in the CNNs, which are assigned different values depending on the resolution of the input feature matrix. Moreover, it reports the number of multihead attention values and key channels used to implement the transformer encoder. It also details the essential training parameters, including the optimization algorithm (stochastic gradient descent with momentum (SGDM)) and the loss function (cross-entropy).

Table 2. Main parameters of the implemented system.

Feature extraction module		Feature classification model	
Parameter	Value	Parameter	Value
Murmur:		**CNN,** $j = 1, 2, ..., 8$**:**	
Segment length (samples)	512	no. kernel filters	32
Dictionary:		Filter size (j)	$2^{j/1.75} \times 2^{6-j/1.75}$
Resolution	$j = 1, 2, ..., 8$	Filter stride (j)	$2^{j/1.75} \times 2^{6-j/1.75}$
Size of Dictionary:	$512 \times (512 * 8)$	**Transformer:**	
COMP:		no. heads	$2, 4, 8, ...,$ or 256
no. iterations (sparsity level)	511	no. key channels	no. heads
Length of projection vector **a**	$512 * 8$	**Training:**	
Matrices $\mathbf{A}_j$**,** $j = 1, 2, ..., 8$**:**		Optimizer	SGDM
Size of matrix j	$2^j \times 2^{8-j+1}$	Loss function	crossentropy
		Batch size	150
		no. epochs	500
		no. classes	4

5.2. System computational complexity

The computational complexity of a neural network refers to the computational resources required for training and inference processes. This complexity depends on factors such as network depth (the number of layers and learnable parameters). Deeper networks involve more computations during forward and backward propagation, and a higher parameter count increases memory usage and training duration. Excessive complexity can hinder convergence and increase overfitting risks, necessitating an efficient architectural design and regularization to balance performance and computational efficiency. Table 3 lists the number of trainable parameters of the model in terms of the number of attention heads and dimensions of each head. Increasing the number of attention heads or the dimensionality of each head leads to a greater system complexity.

5.3. Evaluation metrics

To evaluate the performance of the proposed multiclass murmur classification system, we employed a set of metrics widely used in multiclass diagnostic applications. For each murmur class, we computed the true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN), from which class-specific performance measures were derived. However, due to the strong imbalance among the four murmur types—particularly the limited number of crescendo recordings—class-wise metrics alone do not provide a sufficiently fair overall assessment. Therefore, in addition to reporting per-class specificity, F1-score, and accuracy, we also include *macro-averaged* metrics, which treat all classes equally regardless of their sample sizes. These metrics are defined as follows:

$$\text{Macro-Specificity} = \frac{1}{C} \sum_{i=1}^{C} \frac{TN_i}{TN_i + FP_i}, \tag{5.1}$$

Table 3. The number of trainable parameters in thousands, with ”M” denoting a million, for a model of 44 layers, in terms of the number of attention heads and key or query channels.

Heads(H)	single-head	multihead							
Channels(d_k)	1	2	4	8	16	32	64	128	256
1 × no. heads	159.5	159.6	159.9	160.4	161.5	163.6	167.8	176.1	192.9
2 × no. heads	159.6	159.9	160.4	161.5	163.6	167.8	176.1	192.9	226.5
4 × no. heads	159.9	160.4	161.5	163.6	167.8	176.1	192.9	226.5	293.5
8 × no. heads	160.4	161.5	163.6	167.8	176.1	192.9	226.5	293.5	427.7
16 × no. heads	161.5	163.6	167.8	176.1	192.9	226.5	293.5	427.7	696
32 × no. heads	163.6	167.8	176.1	192.9	226.5	293.5	427.7	696	1.2 M
64 × no. heads	167.8	176.1	192.9	226.5	293.5	427.7	696	1.2 M	2.3 M
128 × no. heads	176.1	192.9	226.5	293.5	427.7	696	1.2 M	2.3 M	4.4 M

$$\text{Macro-F1} = \frac{1}{C}\sum_{i=1}^{C}\frac{2TP_i}{2TP_i + FP_i + FN_i}, \tag{5.2}$$

$$\text{Macro-Accuracy} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}, \tag{5.3}$$

where $C = 4$ denotes the number of murmur classes. These macro-averaged metrics provide an unbiased evaluation of the performance across all categories, irrespective of class imbalance. To further quantify the stability and robustness of the classifier, all reported metrics are presented as the mean ± standard deviation obtained from the stratified 5-fold cross-validation. This allows for a reliable comparison across different folds and reduces the impact of sample imbalance on the evaluation.

5.4. Training dynamics and convergence analysis

To examine the learning behavior of the proposed system and assess potential underfitting or overfitting, we monitored the evolution of the training and validation loss and accuracy across epochs. Figure 10 presents the corresponding curves for the best-performing configuration.

The training and validation accuracy curves exhibit similar increasing trends and converge toward high final values, with no observable divergence between the two. Similarly, the training and validation loss curves decrease smoothly and remain closely aligned throughout the optimization process. This consistent behavior indicates that the model generalizes well and does not overfit the training data, despite the inherent class imbalance and the limited number of crescendo samples.

Furthermore, the absence of oscillations or abrupt fluctuations in the validation curves suggests stable optimization dynamics. This stability is attributed to the combination of the SGDM optimizer and the multistage multiresolution Gabor projection, followed by the transformer encoder, which together promote robust feature learning. These observations confirm that the proposed architecture can effectively learn discriminative representations without suffering from underfitting or overfitting, even under challenging data conditions.

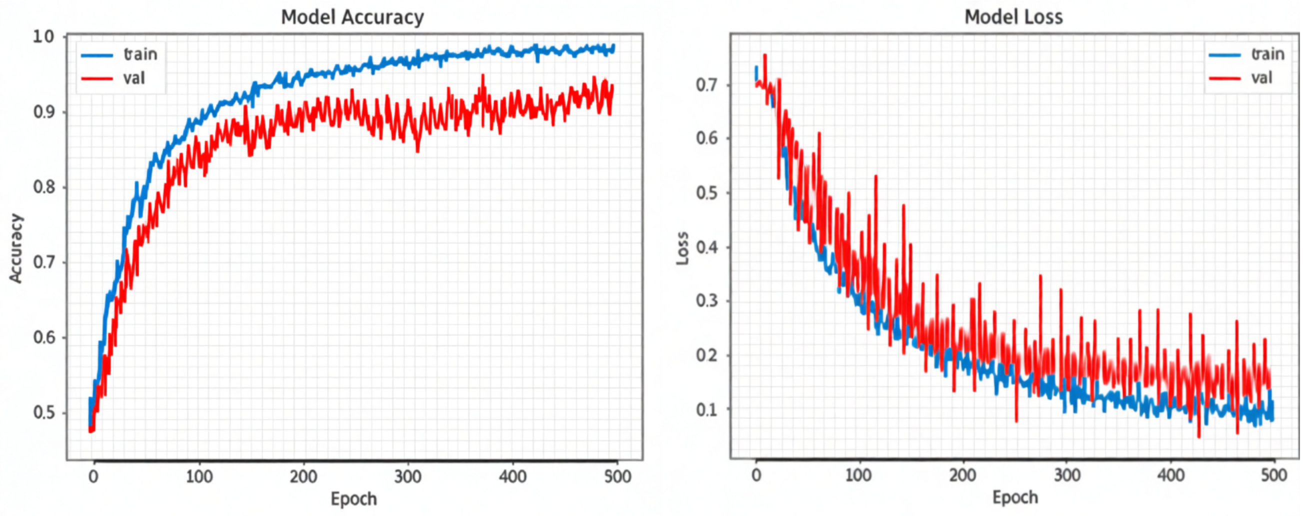

Figure 10. The evolution of the training and validation loss and accuracy across epochs.

5.5. Results

We conducted experiments using 5-fold cross-validation (5-fold CV) to assess system performance. In this process, the extracted feature matrices of murmur segments from each class were divided into five nonoverlapping subsets. The model was then trained and evaluated five times, each time using four subsets for training and the remaining subset for testing. This stratified approach ensures that all matrices are used for both training and validation across different folds, minimizing the risk of overfitting and providing a more reliable estimate of the generalization of the model.

Table 4 lists the average classification accuracy of the proposed system as a function of the number of attention heads (H) and the dimensionality of the output from each head, denoted by the number of channels (d_k) in the transformer encoder. These two parameters control the capacity of the model to capture multiresolution contextual relationships within the multisegment time–frequency representations, and modifying them effectively reflects the impact of simplifying or altering the attention mechanism.

Multihead attention improves classification accuracy over single-head configurations by allowing the model to capture a more diverse set of discriminative features from input representations. It can also be observed that a trade-off exists between the number of attention heads and the dimensionality of each head. Specifically, configurations with fewer attention heads can still achieve competitive or even improved performance when the dimensionality of each head is increased because the model gains a richer representation capacity in each attention stream. Conversely, a higher number of attention heads can compensate for a lower per-head dimensionality by allowing the model to attend to a wider range of features simultaneously. This complementary relationship suggests that optimal performance does not necessarily require the simultaneous maximization of both parameters; instead, a balanced configuration tailored to the task characteristics can yield superior results. These findings emphasize the flexibility of the transformer architecture and the importance of strategic hyperparameter tuning.

5.6. Comparison

To evaluate the performance of the developed system, we conducted a thorough comparative analysis by implementing and testing several systems based on well-established signal processing techniques

Table 4. The average classification macro-accuracy over 5-fold CV on the test set is reported as a function of the number of attention heads and the dimensionality of each head or the number of channels.

Heads(H) / Channels(d_k)	single-head	multihead							
	1	2	4	8	16	32	64	128	256
1 × no. heads	95.46	95.10	94.43	94.61	94.69	94.69	95.39	95.53	95.13
2 × no. heads	94.69	95.61	94.54	94.69	95.28	95.39	95.20	95.72	94.87
4 × no. heads	94.50	94.95	95.53	95.35	95.28	94.76	94.76	95.61	93.25
8 × no. heads	94.50	95.02	95.28	94.95	95.35	95.02	95.98	94.76	94.43
16 × no. heads	94.43	94.61	95.46	95.95	95.61	94.28	95.13	92.81	93.58
32 × no. heads	94.47	95.72	95.96	94.95	94.03	94.61	94.61	93.33	93.66
64 × no. heads	94.43	95.89	94.76	94.95	94.61	95.02	93.85	93.58	92.74
128 × no. heads	94.50	94.28	94.50	94.28	91.96	93.77	93.07	93.00	92.22

and deep learning models. Specifically, we built baseline systems using the short-time Fourier transform (STFT), continuous wavelet transform (CWT), and Wigner–Ville distribution (WVD) [15], in conjunction with residual network (ResNet) [36] and parallel CNNs-transformer (PCTMF-Net) [30].

These systems serve as representative benchmarks for both traditional signal analysis methods and modern deep learning approaches. The STFT extracts spectral features, the CWT captures multiresolution characteristics, and the WVD provides a high-resolution time–frequency representation of the input signals. Input representations were directly used as input to ResNet and PCTMF-Net, whereas they were standardized to 8 feature matrices (via resampling, tiling, or patching) for the proposed model. The extracted information is fed to a ResNet, which leverages its hierarchical feature learning capability to classify the inputs. The PCTMF-Net can process the extracted information through a two-way parallel CNN–transformer architecture with second-order spectral features. Table 5 lists the settings of the implementation parameters for STFT, CWT, and ResNet.

Table 6 presents a comparative evaluation of different feature extraction methods and classification models in terms of average classification accuracy, specificity, and F1-score, along with the corresponding standard deviations across the 5-fold CV. The results demonstrate the performance of two main classification architectures, ResNet and the proposed transformer-based model, when applied to different time–frequency representations of the input signals, namely the STFT, CWT, and resampled CWT. The number of attention heads and dimensionality of each head in the transformer encoder were set to the best-performing values obtained from the deep analysis results reported in Table 4, which corresponded to the case of using four heads with a dimensionality of 32.

The highest accuracy of 95.96% is achieved by the second variant of the proposed system, Proposed II (projection of multiple murmurs); it also attains the highest specificity of 98.43%, with the lowest standard deviations among all methods. This indicates superior performance in correctly identifying the types of murmur segments, particularly in minimizing false positives. The corresponding F1-score of 81.87% suggests a balanced trade-off between precision and recall for this configuration. These results underscore the effectiveness of projecting multiple murmur segments, which effectively mitigates murmur variability and overlap within a single recording by learning distinct weights that operate over consistent time–frequency distributions.

Table 5. Implementation of STFT, CWT, WVD, ResNet, and PCTMF-Net for comparison purposes.

Features		ResNet		PCTMF-Net	
Parameter	Value	Parameter	Value	Parameter	Value
STFT:		Initial filter size	4×4	Two-way CNN-transfomr	
Window length (samples)	32	No. initial filter	32	No. of CNNs	3
Window overlap (samples)	16	Initial stride	2×2	Initial filter size	3×3
STFT size (samples)	17×31	Stack depth	$[8, 4, 2]$	Initial stride	1×1
CWT:		No. filter	$[16, 32, 64]$	No. of heads	4
Wavelet name	Morse				
CWT size (samples)	61×512				
WVD:					
WVD size (samples)	64×512				

However, the discrepancy between the high accuracy (95.96%) and the lower macro F1-score (81.87%) for Proposed II arises from the intrinsic class imbalance in the dataset. Although accuracy is dominated by the performance of the majority classes, the macro F1-score more strongly penalizes misclassifications in minority murmur categories. The few errors made by the model occur primarily in these underrepresented classes, leading to a lower F1-score despite overall high classification accuracy. This behavior is consistent with imbalanced multiclass settings and does not indicate instability in the proposed method.

As shown in Table 6, the proposed transformer-based models consistently outperform both the ResNet and PCTMF-Net baselines across all feature representations. This indicates not only superior discriminative power but also enhanced stability across the cross-validation folds. In contrast, even the strongest baseline (CWT + ResNet) lags by ∼10% in accuracy and ∼6 points in the F1-score. The progressive gains from Proposed I (single-segment projection) to Proposed II (multisegment projection) confirm that enforcing consistent atom support across segments, thereby reducing intra-recording variability, decisively improves performance. These results validate that the synergy of multiresolution sparse modeling and global attention-based classification yields a more robust and clinically reliable framework for murmur shape identification than other methods. These results suggest that the integration of advanced feature projection techniques within a transformer-based architecture enhances the classification performance, offering a promising approach for the accurate and reliable diagnosis of heart disease.

Table 7 reports the per-class specificity and F1-score (mean (standard deviation) over 5-fold cross-validation) for the best-performing configuration (*Proposed II*). The results confirm high model reliability for the two majority classes, diamond and plateau, with specificity above 98% and F1-scores exceeding 93%. For the less frequent decrescendo class, the performance remains robust (specificity = 97.40%, F1 = 77.30%), albeit with higher variability. Critically, even for the severely underrepresented crescendo class (only 2 recordings, 27 segments), the model achieves a specificity of 95.80% and an F1-score of 59.10%, indicating nontrivial discriminative capability despite the extreme data imbalance—thus refuting concerns of complete misclassification (e.g., zero recall) and lending clinical credibility to the overall high accuracy.

Table 6. Comparison results: average classification accuracy and standard deviation.

Features	Classifier	Specificity	F1-score	Accuracy
STFT	ResNet	88.78 (0.82)	70.79 (2.03)	81.26 (1.19)
STFT	PCTMF-Net	89.75 (1.10)	71.93 (2.56)	83.31 (1.36)
CWT	ResNet	91.16 (0.74)	75.39 (1.64)	85.92 (1.16)
WVD	ResNet	89.31 (0.93)	73.05 (2.11)	83.64 (1.27)
CWT	PCTMF-Net	92.17 (0.85)	77.03 (2.07)	87.42 (1.04)
STFT	Proposed transformer	90.26 (1.34)	73.20 (3.38)	86.42 (1.06)
CWT	Proposed transformer	93.00 (1.34)	77.91 (2.84)	88.90 (0.91)
WVD	Proposed transformer	93.85 (0.999)	79.08 (2.08)	89.85 (1.60)
Proposed I (Projection of single murmur)		95.15 (0.86)	79.66 (1.89)	91.84 (0.97)
Proposed II (Projection of multiple murmurs)		98.43 (0.62)	81.87 (1.41)	95.96 (0.82)

Table 7. Per-class classification performance (mean(standard deviation)) for the best-performing configuration (Proposed II, 5-fold CV).

Class	Specificity	F1-score
Diamond	98.10 (1.10)	93.40 (1.90)
Plateau	99.20 (0.50)	96.90 (1.10)
Decrescendo	97.40 (1.50)	77.30 (4.90)
Crescendo	95.80 (2.30)	59.10 (10.60)

6. Conclusions and future work

This study presents a novel system for the automatic classification of systolic heart murmurs. Using sparse modeling through complex orthogonal matching pursuit, the system projects single or multiple murmur segments onto a shared multiresolution Gabor dictionary to effectively capture the time–frequency features while mitigating the variability and overlap between murmur segments within the same recording. These features were processed using a transformer encoder augmented with convolutional networks for tokenization and multiresolution integration. This model learns hierarchical representations from variable-resolution time–frequency matrices, thereby improving generalization and diagnostic accuracy.

Evaluated on the CirCor DigiScope dataset, the proposed system achieved high performance, outperforming conventional approaches, such as ResNet-based models, using standard time–frequency transforms, such as STFT and CWT. The results confirm that integrating multiresolution signal projection with transformer-based architectures significantly enhances the performance of medical audio classification tasks. This study demonstrates the potential of combining sparse representation techniques with deep learning models for the automated diagnosis of heart problems.

Future work will focus on extending this research to several key areas. Validating the system on larger, more diverse datasets, including multicenter and other real-world clinical recordings, will help assess its generalizability. Exploring its application to other medical audio tasks, such as the detection of diastolic murmurs or respiratory pathologies, could broaden its impact. Further architectural improvements, including optimized tokenization, specialized attention mechanisms, and enhanced interpretability, will improve performance and clinical trust. Finally, evaluating the system for real-time or point-of-care deployment will be explored.

Future work will also prioritize collaboration with clinicians to manually annotate diastolic murmur morphology (e.g., early vs. mid-diastolic, rumble vs. decrescendo) in existing datasets or prospective collections. This will enable supervised modeling of diastolic patterns and evaluation of cross-timing generalization (e.g., shared dictionary atoms for systolic vs. diastolic turbulence).

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

1. World Health Organization, *Cardiovascular Diseases CVDs*, 2021. Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).
2. H. Jung, L. S. Lilly, The cardiac cycle: Mechanisms of heart sounds and murmurs, in *Pathophysiology of Heart Disease: A Collaborative Project of Medical Students and Faculty*, 5th edition, Chapter 2, Philadelphia, (2011), 28–53.
3. B. Boashash, *time–frequency Signal Analysis and Processing: A Comprehensive Reference*, 2nd edition, Academic Press, Oxford, 2015.
4. S. Mallat, Z. Zhang, Matching pursuit with time–frequency dictionaries, *IEEE Trans. Signal Process.*, **41** (1993), 3397–3415. https://doi.org/10.1109/78.258082
5. X. Zhang, L. G. Durand, L. Senhadji, H. C. Lee, J. L. Coatrieux, Analysis-synthesis of the phonocardiogram based on the matching pursuit method, *IEEE Trans. Biomed. Eng.*, **45** (1998), 962–971. https://doi.org/10.1109/10.704865
6. X. Zhang, L. Durand, L. Senhadji, H. Lee, J. L. Coatrieux, time–frequency scaling transformation of the phonocardiogram based of the matching pursuit method, *IEEE Trans. Biomed. Eng.*, **45** (1998), 972–979. https://doi.org/10.1109/10.704866
7. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, Cambridge, MA, 2017.
8. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, (2017), 5998–6008.
9. Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, L. Wang, Vision transformers for image classification: A comparative survey, *Technologies*, **13** (2025), 32. https://doi.org/10.3390/technologies13010032
10. J. Oliveira, F. Renna, P. D. Costa, M. Nogueira, C. Oliveira, C. Ferreira, et al., The circor digiscope dataset: From murmur detection to murmur classification, *IEEE J. Biomed. Health. Inf.*, **26** (2022), 2524–2535. https://doi.org/10.1109/JBHI.2021.3137048

11. R. O. Bonow, D. L. Mann, D. P. Zipes, P. Libby, *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*, 9th edition, Elsevier Health Sciences, Philadelphia, 2011.

12. J. Singh, R. Anand, Computer aided analysis of phonocardiogram, *J. Med. Eng. Technol.*, **31** (2007), 319–323. https://doi.org/10.1080/03091900500282772

13. Y. Zheng, X. Guo, X. Ding, A novel hybrid energy fraction and entropy-based approach for systolic heart murmurs identification, *Expert Syst. Appl.*, **42** (2015), 2710–2721. https://doi.org/10.1016/j.eswa.2014.10.051

14. P. D. Stein, H. N. Sabbah, J. B. Lakier, S. R. Kemp, D. J. Magilligan, Frequency content of heart sounds and systolic murmurs in patients with porcine bioprosthetic valves: Diagnostic value for the early detection of valvular degeneration, *Henry Ford Hosp. Med. J.*, **30** (1982), 119–123.

15. M. Akay, *time–frequency and Wavelets in Biomedical Signal Processing*, Wiley-IEEE Press, New York, 1998.

16. M. Fakhry, A. Gallardo-Antolín, Variational mode decomposition and a light cnn-lstm model for classification of heart sound signals, in *IEEE EUROCON 2023 - 20th International Conference on Smart Technologies*, Torino, Italy, (2023), 295–300. https://doi.org/10.1109/EUROCON56442.2023.10199054

17. N. Atanasov, T. Ning, Isolation of systolic heart murmurs using wavelet transform and energy index, in *2008 Congress on Image and Signal Processing*, Sanya, China, (2008), 216–220. https://doi.org/10.1109/CISP.2008.758

18. A. Haghighi-Mood, N. Torry, time–frequency analysis of systolic murmurs, in *Computers in Cardiology 1997*, IEEE, (1997), 113–116. https://doi.org/10.1109/CIC.1997.647843

19. J. G. Daugman, Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, *J. Opt. Soc. Am. A*, **2** (1985), 1160–1169. https://doi.org/10.1364/JOSAA.2.001160

20. M. Fakhry, A. F. Brery, A. Gallardo-Antolín, Analysis of heart sound signals using sparse modeling with gabor dictionary, in *2022 IEEE International Symposium on Multimedia (ISM)*, Italy, (2022), 92–96. https://doi.org/10.1109/ISM55400.2022.00021

21. 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. https://doi.org/10.1016/j.engappai.2023.107406

22. M. Fakhry, A. Gallardo-Antolín, Analysis of systolic murmurs in heart sounds using multiresolution complex gabor dictionary, in *2024 International Conference on Computer and Applications (ICCA)*, IEEE, Cairo, Egypt, (2024), 783–787. https://doi.org/10.1109/ICCA62237.2024.10927981

23. S. Jabbari, H. Ghassemian, Modeling of heart systolic murmurs based on multivariate matching pursuit for diagnosis of valvular disorders, *Comput. Biol. Med.*, **41** (2011), 802–811. https://doi.org/10.1016/j.compbiomed.2011.06.016

24. M. Shabbir, X. Liu, M. Nasseri, S. Helgeson, Heart murmur classification in phonocardiogram representations using convolutional neural networks, in *The International FLAIRS Conference Proceedings*, **36** (2023). https://doi.org/10.32473/flairs.36.133189

25. C. Yin, Y. Zheng, X. Ding, Y. Shi, J. Qin, X. Guo, Detection of coronary artery disease based on clinical phonocardiogram and multiscale attention convolutional compression network, *IEEE J. Biomed. Health. Inf.*, **28** (2024), 1353–1362. https://doi.org/10.1109/JBHI.2024.3354832

26. J. Kim, G. Park, B. Suh, Classification of phonocardiogram recordings using vision transformer architecture, in *2022 Computing in Cardiology (CinC)*, IEEE, Tampere, Finland, **498** (2022), 1-4. https://doi.org/10.22489/CinC.2022.084

27. Z. Liu, H. Jiang, F. W. Zhang, W. B. Ouyang, X. Li, X. Pan, Heart sound classification based on bispectrum features and vision transformer mode, *Alexandria Eng. J.*, **85** (2023), 49–59. https://doi.org/10.1016/j.aej.2023.11.035

28. J. Han, A. Shaout, Enact-heart – ensemble-based assessment using cnn and transformer on heart sounds, preprint, arXiv:2502.16914. https://doi.org/10.48550/arXiv.2502.16914

29. W. Zhao, H. Ma, N. Jin, Y. Zheng, X. Guo, Detection of coronary heart disease based on heart sound and hybrid vision transformer, *Appl. Acoust.*, **230** (2025), 110420. https://doi.org/10.1016/j.apacoust.2024.110420

30. R. Wang, Y. Duan, Y. Li, D. Zheng, X. Liu, C. T. Lam, et al., Pctmf-net: heart sound classification with parallel CNNs-transformer and second-order spectral analysis, *Vis. Comput.*, **39** (2023), 3811–3822. https://doi.org/10.1007/s00371-023-03031-5

31. S. Qiu, H. G. Feichtinger, Discrete gabor structures and optimal representations, *IEEE Trans. Signal Process.*, **43** (1995), 2258–2268. https://doi.org/10.1109/78.469862

32. I. Rish, G. Grabarnik, *Sparse Modeling: Theory, Algorithms, and Applications*, CRC Press, 2014.

33. Z. Průša, N. Holighaus, P. Balázs, Fast matching pursuit with multi-Gabor dictionaries, *ACM Trans. Math. Software*, **47** (2021), 1–20. https://doi.org/10.1145/3447958

34. Z. Zhang, S. Wei, D. Wei, L. Li, F. Liu, C. Liu, Comparison of four recovery algorithms used in compressed sensing for ecg signal processing, in *2016 Computing in Cardiology Conference (CinC)*, IEEE, (2016), 401–404.

35. W. Dai, O. Milenkovic, Subspace pursuit for compressive sensing signal reconstruction, *IEEE Trans. Inf. Theory*, **55** (2008), 2230–2249. https://doi.org/10.1109/TIT.2009.2016006

36. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778.

AIMS Press