



Review

Generative artificial intelligence for heart sound analysis: From model architectures to clinical applications

Mahmoud Fakhry^a, Abeer FathAllah Brery^b^a CEIEC, Universidad Francisco de Vitoria, Ctra.M-515 Pozuelo-Majadahonda Km.1, 800, Pozuelo de Alarcón, Madrid, 28223, Spain^b Department of Computer Science, Universidad Carlos III de Madrid, Avenida de la Universidad, 30, Madrid, 28911, Spain

ARTICLE INFO

Keywords:

First study

Cardiac auscultation

Heart sound signals

Phonocardiogram recordings

Generative artificial intelligence

ABSTRACT

Cardiac auscultation involves listening to heart sounds to diagnose cardiac conditions. Accurate interpretation of these sounds requires considerable experience, which limits consistency among clinicians and across settings. Although deep learning offers a path toward automation, data scarcity, class imbalance, and signal complexity pose challenges for its application to heart sound analysis. This study presents a pioneering comprehensive examination of how generative artificial intelligence (generative AI) models, including generative adversarial networks, variational autoencoders, diffusion models, and transformer-based architectures, can address key challenges in heart sound analysis. Specifically, these models mitigate core limitations by synthesizing high-fidelity heart sound signals for data augmentation, improving signal quality through advanced denoising, and enabling semi-supervised and unsupervised detection of rare pathological conditions. In addition, we discuss their roles in advanced feature learning and cross-modal translation. Despite notable progress, we highlight future challenges, including ensuring physiological fidelity, enhancing model interpretability, and validating clinical utility. In conclusion, generative AI addresses a key translational challenge by enabling accessible, accurate, and scalable expert cardiac evaluation, thereby supporting the implementation of precision cardiology.

1. Introduction

The mechanical expansion and contraction of the heart muscle, coupled with the opening and closing of the heart valves, drives blood flow between the cardiac chambers and through systemic circulation (Libby et al., 2021). This dynamic activity, along with the associated turbulent blood flow, generates vibrations within the heart and adjacent vascular structures. These vibrations are transmitted to the chest wall as audible signals known as heart sounds, which can be recorded and visualized as time-series waveforms referred to as phonocardiogram (PCG) (Rangayyan & Lehner, 1987; Shariat Panah, Hines, McKeever, & McKeever, 2022). A typical cardiac cycle in a PCG recording consists of the first and second heartbeats (S_1 and S_2), corresponding to atrioventricular and semilunar valve closure, respectively. Under pathological or specific physiological conditions, additional sounds such as S_3 , S_4 , clicks, and murmurs may be present. These acoustic events often reflect abnormal ventricular filling, turbulent blood flow, or structural defects and therefore play a crucial role in the diagnosis of cardiovascular diseases, including valvular disorders, congenital heart defects, and heart failure. Accurate interpretation depends not only on the presence

of these sounds but also on subtle variations in their timing, frequency, and intensity (Fakhry & Brery, 2022; Fakhry & Gallardo-Antolín, 2025).

Traditionally, heart sounds have been assessed through manual auscultation using a stethoscope, relying on the auditory skills and experience of the clinicians. However, this method has limitations in terms of reproducibility, objectivity, and sensitivity, particularly for high-frequency components that may not be perceptible to the human ear. To address these challenges, phonocardiography employs specialized sensors, such as piezoelectric or microphone-based transducers, to record heart sounds with high fidelity, thereby enabling detailed analyses. Although PCG signals theoretically span a wider frequency range (approximately 20–2000 Hz), many practical systems apply narrower band-pass filtering (commonly 20–400 Hz or up to 500 Hz) to suppress noise and retain diagnostically relevant components, as adopted in several recent studies.

Advanced signal processing techniques, including time–frequency analysis, wavelet transforms, and machine learning (ML) algorithms, have been increasingly applied to PCG data for segmentation, feature extraction, classification, and anomaly detection purposes. These methods allow the identification of subtle patterns in heart sounds

* Corresponding author.

E-mail addresses: mahmoud.fakhry@ufv.es (M. Fakhry), abeer_brery@aswu.edu.eg (A.F. Brery).<https://doi.org/10.1016/j.iswa.2026.200637>

Received 3 November 2025; Received in revised form 3 February 2026; Accepted 10 February 2026

Available online 14 February 2026

2667-3053/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

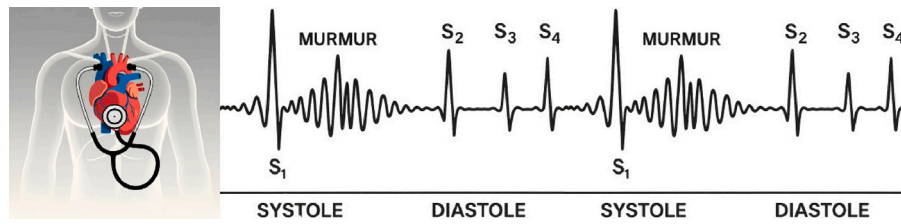


Fig. 1. Two cardiac cycles of a phonocardiogram recording, PCG signal, with systolic murmurs located between the first (S_1) and second (S_2) heartbeats. Extra heart sounds, S_3 and S_4 , are present during diastole.

that may correlate with specific pathologies, thereby supporting the early diagnosis and continuous monitoring of cardiovascular diseases (CVDs). Deep learning (DL) has demonstrated significant potential for automating diagnostic tasks by learning complex patterns directly from data (Liu, Lv, Cao, & Guo, 2025; Vimalajeewa, Lee, & Vidakovic, 2025; Zhao, Ma, Jin, Zheng, & Guo, 2025). Numerous studies have reviewed the application of deep learning models in heart sound analysis (Bi, Tang, Xiao, Li, Gao et al., 2025; Chen, Guo, et al., 2024; Partovi, Babic, & Gharehbaghi, 2024; Zhao et al., 2024), highlighting that although these models show promise, they continue to face significant challenges. These challenges arise from the inherently complex and non-stationary nature of heart sounds, limited availability of large, well-annotated datasets, and high susceptibility of PCG signals to noise, motion artifacts, and variations in recording conditions. Recent advances in generative artificial intelligence have opened up promising new avenues for addressing these challenges.

Generative artificial intelligence (Generative AI) refers to AI models capable of creating new data that mimic patterns in existing datasets (Chen, Yan, & Zhu, 2024). Unlike traditional deep learning models, which focus on recognition or classification, generative models learn the underlying distribution of a dataset to produce novel outputs, including text, images, audio, video, and structured data such as code or molecular structures. Key technologies include generative adversarial networks (GANs) (Goodfellow et al., 2014), variational autoencoders (VAEs) (Kingma & Welling, 2019), diffusion models (Sohl-Dickstein, Weiss, Maheswaranathan, & Ganguli, 2015), and transformer-based models, such as generative pre-trained transformer (GPT) for text (Han et al., 2021) and DALL·E for images (Ramesh et al., 2021). GANs are widely used for high-resolution image synthesis and quality enhancement, whereas VAEs efficiently represent latent spaces for generation and anomaly detection. Diffusion models generate detailed and coherent content by learning to reverse a gradual noise process. Transformer-based models excel in coherent text generation, image synthesis from descriptions, music composition, and applications in drug discovery and protein structure prediction.

This study makes three key contributions. First, it is a pioneering comprehensive survey focused exclusively on the application of generative AI models (including GANs, VAEs, diffusion models, and transformers) to heart sound analysis, addressing a gap left by previous reviews, as detailed in Table 1. This is based on a systematic search of major academic databases, from which 30 key studies were selected and categorized by the generative paradigm. Second, it presents a framework that links methodological advances, such as synthetic data generation, unsupervised anomaly detection, and cross-modal translation, to clinical impact by addressing data scarcity, class imbalance, and noise. Third, it provides in-depth coverage of emerging models, highlighting their transformative potential in biomedical signal processing. Finally, this study outlines future directions, including physiologically constrained modeling, explainability, personalization, and clinical validation, to guide the development of trustworthy AI in precision cardiology.

The remainder of this paper is organized as follows: Section 2 provides an overview of heart sounds. The procedures and architectures of the generative models are described in Section 3. Section 4 introduces

the proposed search method, and Section 5 presents the results. The datasets and data representations explored are briefly explained in Sections 6 and 7, respectively. Section 8 discusses the clinical applications of revised models. Sections 9 and 10 outline discussion and future research directions, respectively. Finally, Section 11 concludes the paper.

2. Heart sound signals

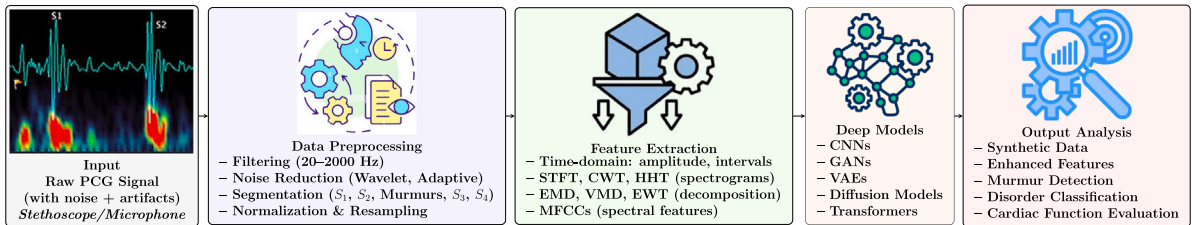
Heart sound signals originate from vibrations produced by the rapid movement and sudden deceleration of blood, as well as the opening and closing of heart valves (tricuspid, aortic, pulmonary, and mitral), which generate distinct sounds. Fig. 1 illustrates two cardiac cycles with the components of PCG signals. The closure of the mitral and tricuspid valves at the start of ventricular systole generates the first heartbeat S_1 . Conversely, the closure of the aortic and pulmonary valves at the onset of semilunar diastole generates the second heartbeat S_2 . In addition to these two fundamental heartbeats, third (S_3) and fourth (S_4) heart sounds may be present under specific pathological conditions. S_3 is often associated with increased ventricular filling during early diastole and can indicate heart failure. S_4 typically reflects forceful atrial contraction against a stiff or noncompliant ventricle, commonly seen in conditions such as hypertension or hypertrophic cardiomyopathy. Murmurs, extra heart sounds, clicks, and rubs further enrich the complexity of PCG signals and often serve as critical diagnostic clues for structural abnormalities, such as valvular stenosis or regurgitation, septal defects, or pericardial inflammation. These abnormal sounds arise due to turbulent flow through narrowed or leaking valves, abnormal shunting of blood between chambers, or friction between inflamed layers of the pericardium. For instance, systolic murmurs may suggest aortic stenosis or mitral regurgitation, whereas diastolic murmurs are often indicative of aortic regurgitation or mitral stenosis. Clicks are frequently linked to mitral valve prolapse, and pericardial friction rubs have high specificity for acute pericarditis. Understanding the timing, location, intensity, and quality of these sounds is essential for accurate clinical interpretation and differential diagnoses.

Deep learning techniques applied to heart sound analysis typically involve preprocessing raw signals, extracting relevant features, and deploying deep models for classification or detection (Fig. 2). Raw PCG recordings are typically contaminated with various types of noise and artifacts, including environmental sounds, respiratory noise, and motion-induced disturbances. To address these issues, preprocessing commonly begins with filtering, targeting a particular frequency range to eliminate irrelevant low and high frequency components. Techniques such as wavelet denoising and adaptive filtering can further enhance signal clarity by isolating the key components of heart sounds while suppressing noise (Gavrovskaya, Slavkovic, Reljin, & Reljin, 2013). Furthermore, PCG signals are often segmented into cardiac cycles or heart sound phases (systole, diastole, and components such as S_1 , S_2 , murmurs, S_3 , and S_4) using methods such as hidden Markov models, envelope detection, or energy thresholding (Li et al., 2024; Rieke, Povinelli, & Johnson, 2005), enabling the isolation of diagnostically relevant signal segments. Amplitude normalization and resampling were applied to ensure consistency across recordings, particularly when data were obtained from different sensors or setups.

Table 1

Comparison between the proposed study and related surveys in the literature. CNNs, RNNs, LSTMs, SVMs, DSP, and HMMs refer to convolutional neural networks, recurrent neural networks, long short-term memory networks, support vector machines, digital signal processing, and hidden Markov models, respectively.

Feature	Proposed study	Ref. Partovi et al. (2024)	Ref. Zhao et al. (2024)	Ref. Chen, Guo, et al. (2024)	Ref. Bi et al. (2025)
Core Focus	Generative AI models for creation, augmentation, and enhancement of heart sounds.	General deep learning methods for analysis and classification.	Deep learning techniques and their clinical applications.	AI classification methods, including traditional ML and DL.	Detection and segmentation methods, with experimental insights.
Key Models Covered	GANs, VAEs, Diffusion Models, Transformer-based generative models.	CNNs, RNNs, LSTMs, Autoencoders, Hybrid models.	CNNs, RNNs, Transformers, Attention mechanisms.	SVMs, Random Forests, CNNs, RNNs, Transformers.	DSP, HMMs, SVM, CNNs, and segmentation networks.
Primary Emphasis	Solving data-centric challenges via generation; linking methods to clinical translation.	Technical review of DL models and performance in classification.	Survey of DL techniques→ clinical outcomes; applied cardiology.	Comparative review of classification accuracy across AI paradigms.	Systematic evaluation of performance in detection/segmentation.
Addresses Data Scarcity/ Imbalance?	Yes—central theme. Focus on synthetic data generation, augmentation, denoising.	Limited –focuses on model performance using existing datasets.	Partially –may discuss augmentation but not as primary generative focus.	Indirectly –via traditional augmentation, not generative AI.	No –focuses on algorithmic processing of available signals.
Covers Clinical Applications?	Yes –Dedicated section on diagnostics, denoising, anomaly detection, multimodal tools.	Limited –primarily technical model analysis.	Yes –Strong focus on clinical use cases and deployment.	Moderate –links classification to diagnostic outcomes.	Minimal –focuses on algorithmic rather than clinical pipelines.
Highlights Future Directions?	Yes–Detailed roadmap for physiological fidelity, explainability, personalization, validation, real-time use.	Likely includes technical DL trends.	Presents clinical and technical future directions.	Suggests improvements in classification accuracy and robustness.	Focuses on algorithmic improvements and benchmarking.
Unique Contribution	First dedicated study on generative AI for heart sounds; bridges data generation with clinical cardiology.	DL overview; useful for non-generative heart sound analysis.	Strong clinical-translational perspective; connects DL to patient care.	Comprehensive classification-method survey across AI eras.	Systematic experimental review of detection/segmentation methods.

**Fig. 2.** The processing flow of heart sound signals using deep learning.

Following preprocessing, feature extraction transforms the cleaned PCG signals into representations suitable for deep learning models. Time-domain features, such as signal amplitude, heart sound duration, inter-beat intervals, and systolic/diastolic timing, offer basic descriptive information but often lack sufficient complexity for such models. Therefore, time–frequency analysis methods, such as the short-time Fourier transform, continuous wavelet transform, and Hilbert–Huang transform, are commonly employed to generate 2D time–frequency feature maps that preserve temporal and spectral characteristics (Fakhry & Gallardo-Antolín, 2024a, 2024b; González-rodríguez, Alonso-Arévalo, & García-Canseco, 2023). These representations serve as rich inputs for convolutional neural networks (CNNs) or hybrid models that combine CNNs with recurrent architectures. Alternatively, advanced decomposition techniques such as variational mode decomposition or empirical wavelet transform can be used to isolate intrinsic signal modes before constructing feature matrices. In some frameworks, Mel-frequency cepstral coefficients (MFCCs) are adapted to PCG signals to capture perceptually significant spectral information (El-Dahshan & Bassiouni,

2020). When these feature extraction techniques are combined, deep learning architectures can learn discriminative representations of PCG signals, which makes it easier to perform tasks such as valvular disorder classification, cardiac function evaluation, and murmur detection.

3. Generative AI models

Generative AI models are sophisticated computational frameworks designed to generate new data that closely emulate the characteristics of real-world data. In contrast to traditional machine and deep learning models, which focus primarily on discrimination or prediction, generative AI models learn the underlying data characteristics and use this knowledge to generate new content. By modeling complex high-dimensional patterns from training data, these systems can generate realistic and controllable outputs that resemble observed samples while extending the data space. This capability makes them particularly valuable in domains where data scarcity, imbalance, and variability limit the performance and generalizability of analytical models.

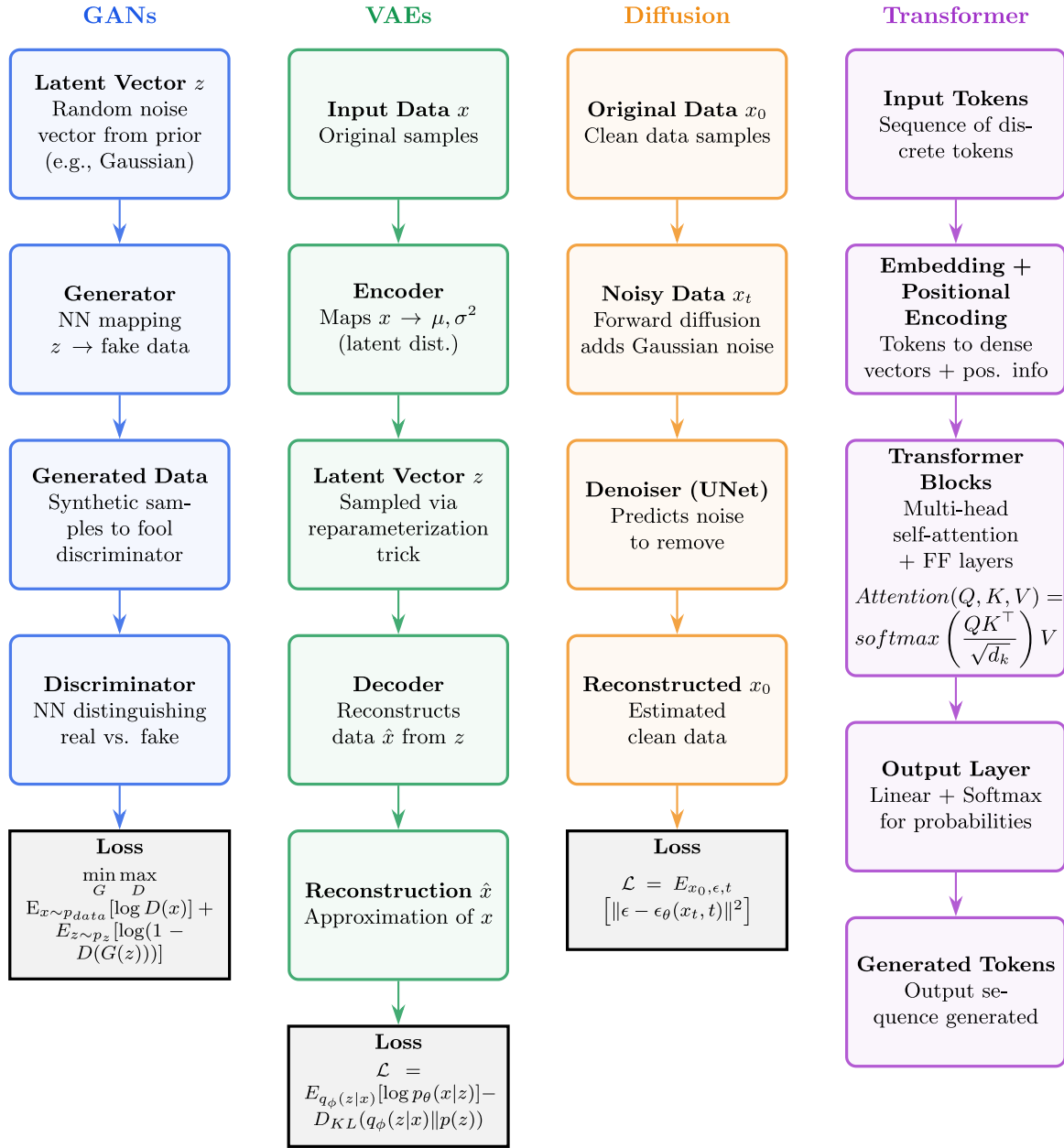


Fig. 3. Procedure steps and key formulas of four generative AI models.

Recently, generative AI has become a vital tool in various fields, including creative media, healthcare, and biomedical signal processing. In the healthcare domain, generative AI contributes to critical tasks, such as data augmentation, in which synthetic medical signals, such as PCG signals, are generated to enrich limited datasets, thereby improving the training of diagnostic models. Additionally, generative AI enables the creation of realistic simulations of physiological phenomena, facilitating system testing and algorithm development. Beyond data generation, generative AI models are increasingly employed for signal denoising, feature enhancement, and interpretive modeling, providing valuable support for clinical decision-making. As these technologies continue to evolve, generative AI is expected to play a transformative role in automating healthcare diagnostics and treatment planning processes. Fig. 3 shows the procedure steps and key formulas of the four generative models, Fig. 4 illustrates the layer construction of each model, and Table 2 holds a comparison of them.

Generative adversarial networks (GANs) are a class of generative AI models that consist of two competing neural networks: a generator and

a discriminator. The generator creates synthetic data (such as images, audio signals, or time-series data), whereas the discriminator evaluates whether the data it receives is real (from the original dataset) or fake (produced by the generator). Through this adversarial process, the generator progressively improves its ability to create realistic data by learning to deceive the discriminator in successive training iterations. GANs are well-suited for generating high-fidelity synthetic signals, particularly when data augmentation is required to address limited dataset sizes. However, their training instability and risk of mode collapse must be carefully managed during their application.

Variational autoencoders (VAEs) combine probabilistic graphical models with deep learning concepts. The two primary components of VAEs are an encoder that maps input data to a latent space represented by a probability distribution (typically a Gaussian distribution) and a decoder that reconstructs data from samples drawn from the probability distribution. Unlike conventional autoencoders, VAEs add a regularization term during training to ensure a smooth and continuous latent space, facilitating the generation of meaningful information.

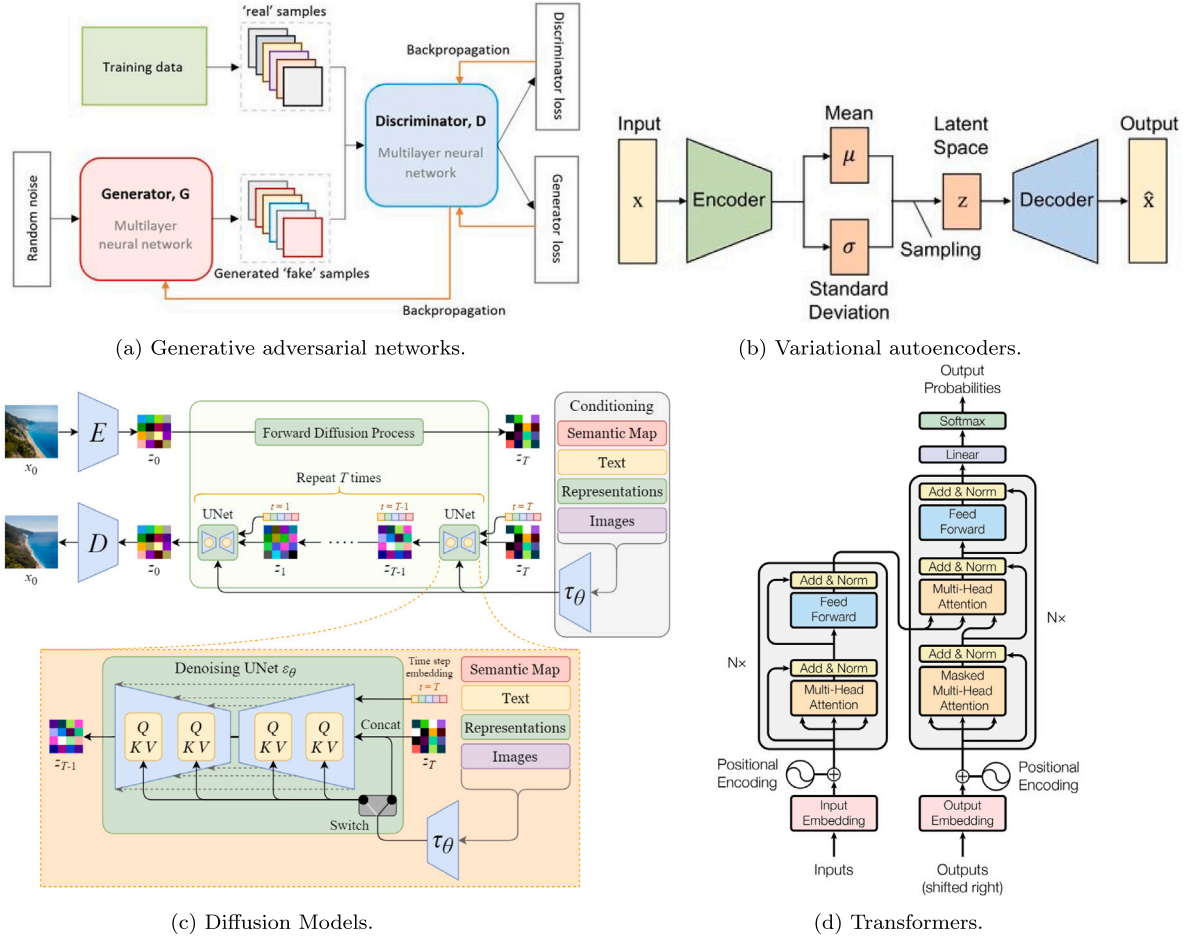


Fig. 4. Detailed block diagrams of generative AI models.

This property makes VAEs well-suited for applications such as signal modeling, image synthesis, and anomaly detection. Although VAEs generally produce outputs with lower perceptual quality than GANs, they benefit from more stable training dynamics and interpretable latent representations, which are advantageous for feature learning and analysis tasks.

Diffusion models generate data by learning to reverse the process of progressive noise corruption. During training, noise is incrementally added to the original data over multiple steps, and the model learns to recover the clean signal by iteratively denoising it. Compared with adversarial generative models, diffusion models are noted for their stable optimization behavior and capacity to produce highly diverse and detailed samples. Recent advances, such as Stable Diffusion (Rombach, Blattmann, Lorenz, Esser, & Ommer, 2021) and WaveGrad (Chen, Zhang, Zen, Weiss, Norouzi et al., 2020), have demonstrated strong performance in text-to-image synthesis and speech generation, respectively. Diffusion models have the potential to improve performance in data-scarce environments, enhance datasets for training algorithms, and generate synthetic training data. Although diffusion models often require longer times due to their iterative procedures, they frequently surpass GANs in terms of generation quality, making them increasingly attractive for synthetic data generation and data augmentation in data-scarce environments.

Transformer-based models employ attention mechanisms to efficiently process sequential data by modeling the global dependencies across entire sequences. Unlike recurrent architectures, transformers operate in parallel, enabling scalable training and improved handling of long-range contextual relationships in the data. Consequently, they underpin many state-of-the-art systems in natural language processing

and computer vision, including GPT for text generation, bidirectional encoder representations from transformers (BERT) for language understanding (Devlin, Chang, Lee, & Toutanova, 2019), DALL·E for text-to-image synthesis, and MusicLM for music generation (Agostinelli et al., 2023). Although transformers are not inherently generative in their original formulation, they can be adapted for sequence-to-sequence generation or integrated into more complex generative frameworks. Their ability to capture contextual and temporal dependencies renders them particularly effective for conditional generation tasks and multimodal applications that require attention-driven analysis.

4. Method

4.1. Search strategy

Fig. 5 shows a flowchart of the proposed search methodology. To create a summary of the most recent research findings on heart sound analysis using generative AI, we searched PubMed, IEEE Xplore, Web of Science, and Google Scholar for the terms “generative artificial intelligence” or “generative adversarial networks” or “GANs” or “variational autoencoders” or “VAEs” or “diffusion models” or “transformer-based models” in combination with “cardiac auscultation” or “heart sounds” or “heart murmurs” or “phonocardiogram” or “phonocardiography” or “PCG”. We broadened our search to encompass all article fields to prevent missing papers that did not specifically use these keywords in their titles. This strategy was designed to ensure the inclusion of relevant studies regardless of how the concepts were indexed or described.

Peer-reviewed journal articles and conference proceedings published recently were considered. Special attention was paid to studies

Table 2
Comparative overview of key generative AI models.

Feature	GANs	VAEs	Diffusion	Transformers
Core Idea	Adversarial game: Generator vs. Discriminator	Probabilistic latent space+ Reconstruction	Gradual noise addition/removal (reverse diffusion)	Autoregressive token prediction via self-attention
Architecture	-Generator (G) -Discriminator (D)	-Encoder -Decoder -Latent space	-Forward diffusion -Reverse denoising U-Net	Encoder–decoder stacks with self-attention
Training Process	Min–max game: G fools D, D distinguishes real/fake	Maximize ELBO (Evidence Lower Bound)	Learn iterative denoising steps	Maximize likelihood of next token prediction
Strengths	-High-fidelity samples -Fast generation	-Stable training -Meaningful latent space	-State-of-the-art quality -Training stability	-Long-range dependencies -Multi-modal flexibility
Weaknesses	-Mode collapse -Training instability -No density estimate	-Blurred outputs -Limited sample diversity	-Slow sampling -High computations	-Sequential sampling slowness -Context window
Output Quality	Sharp, high-fidelity (esp. images)	Blurry/smooth	State-of-the-art photorealism	Highly coherent sequences
Speed	Fast (single pass)	Fast (single pass)	Slow(100–1000 steps)	Slow(token-by-token)
Probabilistic Foundation	Implicit density	Explicit latent distribution	Explicit likelihood (approximated)	Explicit sequence likelihood
Applications	-Image synthesis -Style transfer -Data augmentation	-Anomaly detection -Dimensionality reduction	-Photorealistic image generation -Video synthesis	-Text generation -Translation -Code generation
Examples	DCGAN, StyleGAN, CycleGAN	β -VAE, VQ-VAE	WaveGrad, Stable Diffusion	GPT, DALL·E, BERT, MusicLM

that used generative techniques for anomaly detection, data augmentation, denoising, heart sound synthesis, and diagnosis support. Duplicate studies and those lacking clear relevance to generative modeling or heart sound analysis were excluded. The resulting collection of literature provides a foundation for understanding how generative AI is currently being used in the investigation of heart sound signals, highlighting emerging trends, methodological innovations, and existing challenges in this rapidly evolving field of research.

4.2. Study screening and selection

We considered only published, peer-reviewed original research articles and excluded reviews, editorials, conference abstracts without full papers, and studies unrelated to heart sound analysis or artificial intelligence. Given that this study specifically focuses on generative AI, we further excluded studies with a narrow emphasis on traditional deep learning or discriminative models outside the generative AI framework. After applying the inclusion and exclusion criteria, 30 original articles were selected for detailed analysis. These studies can be broadly classified into four major categories based on the generative modeling approach employed: GANs (12+1*), VAEs (8), diffusion models (2+1*), and transformer-based models (7). The asterisk (*) indicates studies that integrate or hybridize multiple generative paradigms.

5. Results

Fig. 6 shows the distribution of references over time. Fig. 7 shows the number of publications that referred to the use of generative AI for heart sound analysis each year. Table 3 lists the publications that used generative AI for heart sound analysis. GANs are widely adopted generative models for heart sound signal analysis. Their primary application lies in data augmentation, where synthetic PCG signals are generated to address the issues of class imbalance and limited dataset sizes. Beyond augmentation, GANs have been explored, although less frequently, for signal enhancement and denoising to improve signal quality and robustness in noisy environments. Following the success of GANs, VAEs, and their variants, β -VAEs have been employed in tasks

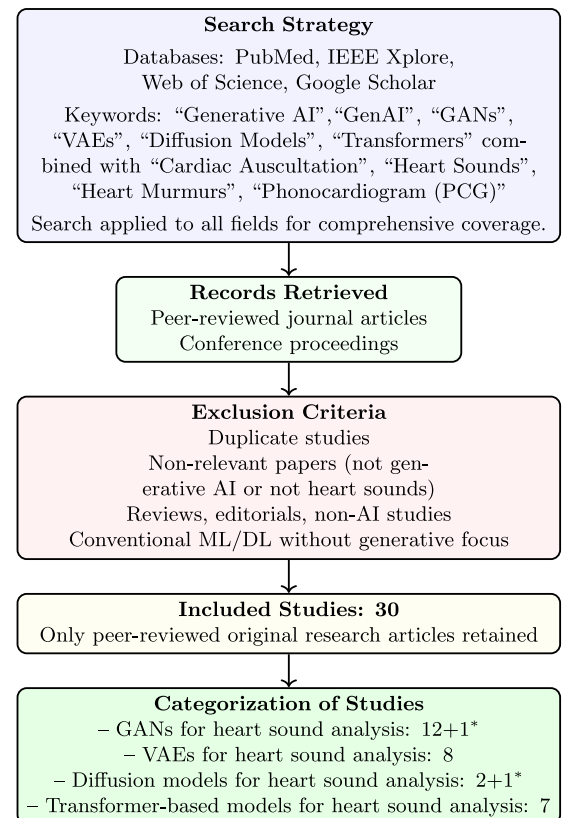


Fig. 5. Flowchart of search methodology. 1* indicates that the GANs and diffusion models were investigated in the same article.

such as dimensionality reduction and anomaly detection in various fields. By learning compact latent representations, these models enable the efficient characterization of normal and pathological heart sounds, thereby facilitating diagnostic support.

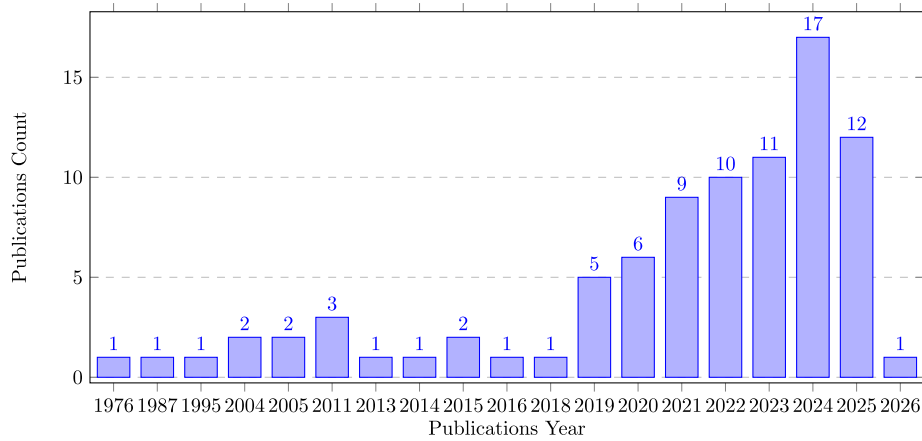


Fig. 6. Number of referenced publications per year.

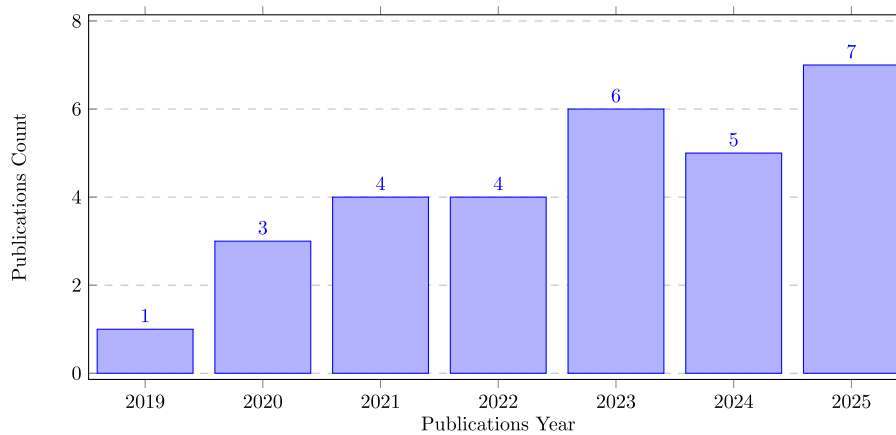


Fig. 7. Number of referenced publications of generative AI used for heart sound analysis per year.

Diffusion models, a recent family of generative techniques, have been marginally investigated in this domain, primarily for data augmentation. Despite their demonstrated success in other biomedical signal and image applications, their potential for high-fidelity heart sound synthesis and denoising remains to be explored. In parallel, transformer-based architectures have gained popularity in recent years. By leveraging attention mechanisms, transformers capture long-range dependencies within sequential data, making them well-suited for complex time-series data, such as PCG signals. Current research has primarily applied transformers to anomaly detection and murmur classification, demonstrating improvements over traditional deep learning models. Their flexibility further suggests opportunities for integration with generative models, which could enhance representation learning and improve diagnostic interpretability of the models. In the following subsections, we describe the different architectures used for heart sound analysis and those developed based on the generative AI models introduced in this study. Following this descriptive analysis, we present a critical synthesis of the characteristics and performance of the reviewed studies using the generative paradigm.

5.1. GANs-based models

Generative adversarial networks (GANs) have emerged as powerful tools for addressing data limitations in signal analyses. Their primary application is synthetic data generation to mitigate class imbalances and dataset scarcity, which are common obstacles in training robust models. This is achieved using various architectures that are tailored to specific technical requirements. For instance, progressive Wasserstein

GANs (Chakraborty et al., 2025) and inception-based GANs (Gawande et al., 2025) have been used to generate realistic pathological heart sounds. Similarly, deep convolutional GANs (DCGANs) and conditional GANs (CGANs) have been effectively used to augment datasets for class-specific synthesis of signals such as PCG and electrocardiography (ECG) signals (Narváez & Percybrooks, 2020; Rayavarapu, Prasanthi, Kumar, Rao, & Singham, 2023; Yang et al., 2021).

The quality of the synthetically generated signals is validated using a suite of objective metrics that assess both fidelity and utility. Common evaluation metrics include the Fréchet audio distance (FAD), structural similarity index (SSIM), and cross-correlation for signal quality (Chakraborty et al., 2025; Rayavarapu, Prasanthi, Kumar, Rao, & Prashanti, 2023; Rayavarapu, Prasanthi, Kumar, Rao, & Singham, 2023). Other metrics, such as the maximum mean discrepancy (MMD) and Jensen–Shannon divergence (JSD), are used for distributional similarity (Jamshidi et al., 2024), and mel-cepstral distortion (MCD) is used for audio realism (Narváez & Percybrooks, 2020). The ultimate validation is performance-based, in which synthetic data are used either individually or in conjunction with real data to train and evaluate classical and deep classifiers (e.g., CNNs, VGGNet, Bi-LSTM, and ResNet18 networks; (Takezaki & Kishida, 2022; Vinay et al., 2024)), successfully measured by significant improvements in model accuracy.

In addition to data augmentation, GANs facilitate advanced signal enhancement and translation. Hybrid models combining GANs with techniques such as parameterless scale-space boundary detection (PSBD) and empirical wavelet transform (EWT) are highly effective for denoising, significantly improving the signal-to-noise ratio (SNR), and reducing the root mean square error (RMSE) in challenging low-SNR conditions (Hu et al., 2025; Liang et al., 2022). Furthermore, GAN

Table 3
Generative AI models for heart sound analysis.

Ref.	Model	Task class	Datasets	Represent.	Key findings/Performance
Chakraborty et al. (2025)	GANs	Data augmentation	PhysioNet, PASCAL, Michigan	Raw signal	FAD of 1.43 and 2.23. The CAD detection sensitivity 99% and specificity 98% using a 1D-CNN classifier.
Gawande, Goyal, and Sankhla (2025)	GANs	Data augmentation	PhysioNet	Raw signal	CHF detection accuracy 97% and F1-score 96% using a CNN+autoencoder classifier.
Yang, Guo, Wang, and Zheng (2021)	GANs	Data augmentation	Private dataset	Spectrogram	LVDD detection accuracy 98.70%, sensitivity 98.60%, and specificity 98.80% using a CNN classifier.
Rayavarapu, Prasanthi, Kumar, Rao, and Singham (2023)	GANs	Data augmentation	PhysioNet	Raw signal	Cross-correlation coefficient 0.9638 and SSIM 0.9732 between real and generated signals using DCGAN.
Narváez and Percybrooks (2020)	GANs	Data augmentation	PhysioNet, PASCAL	Raw signal	Improved MCD distortion metric of synthesized normal heart sounds and 100% classification accuracy of SVM.
Rayavarapu, Prasanthi, Kumar, Rao, and Prashanti (2023)	GANs	Data augmentation	PhysioNet, PASCAL	Raw signal	SSIM of 0.9805 between real and generated signals using CycleGAN.
Jamshidi, Arif, Kalhoro, and Gelbukh (2024)	GANs and diffusion	Data augmentation	CirCor DigiScope	Raw signal	JSD and MMD of Doppelganger GAN are 0.014 and 0.0001, and of DiffWave are 0.015 and 0.0001, respectively.
Vinay, Vidyasagar, Rohith, Pruthviraja, Supreeth et al. (2024)	GANs	Data augmentation	PASCAL	HHT with EMD	Improved cardiovascular disease recognition accuracy of 96% using an multi-layered Bi-LSTM classifier
Takezaki and Kishida (2022)	GANs	Data augmentation	PhysioNet	Spectrogram	Improved prediction accuracy of 92.6% using a CNN classifier.
Hu, Chen, Yang, Shen, and Gao (2025)	GANs	Denoising	PASCAL, Michigan, Washington	Raw signal	Combination of PSBD and EWT with GANs to enhance SNR of noisy PCG recordings (e.g., from -5 dB to 9.26 dB for Gaussian noise.)
Liang, Hu, Zhou, and Xiao (2022)	GANs	Denoising	Michigan, Washington	Raw signal	Learns deep features from clean and noisy heart sound signals using GANs, which increases SNR by about 1 dB.
Dissanayake, Fernando, Denman, Sridharan, and Fookes (2023)	GANs	Data augmentation and modality transfer	PhysioNet	Raw signal	The objective evaluation was done using a CNN classifier, achieving 89% accuracy on synthetic data and 96% on real data.
Unnikrishnan, Singh, Yang, and Chua (2020)	GANs	Anomaly detection	PhysioNet	Mel-spectrogram	Semi-supervised GANs reduce label dependency and outperform supervised baselines with 88.50% accuracy.
Li, Tian, and Wang (2021)	VAEs	Anomaly detection	PhysioNet	Mel-spectrogram	Unsupervised β -VAE achieves 0.924 AUC at $\beta = 0.01$, indicating that minimal KL regularization enhances sensitivity.
Banerjee and Ghose (2020)	VAEs	Anomaly detection	PhysioNet, PASCAL, Michigan	Log-spectrogram	Semi-supervised VAE with anomaly score calculated using KL-divergence, outperforming supervised baselines with sensitivity between 89% and 99%.
Li and Tian (2021)	VAEs	Analysis of latent space	PhysioNet	Mel-spectrogram	Density-based β -VAE produces highly non-Gaussian distributions through estimating distribution in the latent space.
Colt, Várady, Volpi, and Malagò (2020)	VAEs	Feature extraction	PASCAL	Raw signal	WaveNetVAE with MMD extracts features from PCG signals, classified by SVM, achieving 96% specificity.
Buchh (2024)	VAEs	Dimensionality reduction and anomaly detection	PASCAL	Fourier transform	Outperforms methods (PCA, KPCA, ICA) in learning non-linear manifolds for effective clustering of heart sounds
Adiban, BabaAli, and Shehnepoor (2019)	VAEs	Dimensionality reduction and feature enhancement	PhysioNet	i -vector	i -vector and VAE embeddings with GMM-SVM achieved 97.34% accuracy- 16% over the baseline.
Zhang, Li, Guo, and Liu (2023)	VAEs	Data augmentation	Yaseen et al.	Bi-spectrum	VAE-generated synthetic data combined with real data boosted the BDF classification to 99.94% accuracy.
Jamal, Ramasamy, and Abdullah (2025)	VAEs	Signal separation in the latent space	PhysioNet	Raw signal	Separating heart and lung sounds, then classifying latent features with a CNN, achieved 91% accuracy.
Tan et al. (2025)	Diffusion	Data augmentation	PhysioNet	Mel-spectrogram	VAEs with a U-Net diffusion model for heart sound generation with FID= 18.39 enhances WideResNet semi-supervised learning, achieving 93.4% accuracy.
Abbott, Marocchi, Fynn, Rong, and Nordholm (2026)	Diffusion	Data augmentation	PhysioNet, SMPECG	Mel-spectrogram	Generates PCG conditioned on ECG using WaveGrad and DiffWave, improving in-distribution and out-of-distribution MCC by 0.066 and 0.297, respectively.
Wang, Hu, Du, Yuan, Xie et al. (2025)	Transformer	Anomaly detection	Yaseen et al. Private	Wavelet coefficients	Combines CNN for feature extraction with transformer for disease diagnosis with 99.98% and 91.75% accuracies.
Han and Shaout (2025)	Transformer	Anomaly detection	PASCAL	Spectral centroid	MoF fuses predictions from VIT and CNN operating on complementary representations, yielding 97.52% accuracy.
Liu, Liu, Gu, and Yang (2024)	Transformer	Anomaly detection	PhysioNet	Mel-spectrogram	DCv-swin transformer combines convolutional embedding with dual-channel attention, achieving 93.4% accuracy
Jamil and Roy (2023)	Transformer	Anomaly detection	Yaseen et al.	MFCCs	Combines deep CNNs, particle swarm optimization, and a genetic algorithm with VIT for the detection of valvular heart disease, achieving 99.90% accuracy.
Yang et al. (2023)	Transformer	Anomaly detection	Yaseen et al.	Spectrogram	A transformer is used for classifying PCG signals, achieving 98.74% and 100% accuracy for 5- and 10-CV, respectively.
Hassanuzzaman et al. (2023)	Transformer	Anomaly detection	Private dataset	Raw signal	A 1D CNN with an attention transformer to recognize pediatric congenital heart diseases, achieving 92.30% accuracy.
Alkhodari, Azman, Hadjileontiadis, and Khandoker (2022)	Transformer	Murmur detection	CirCor DigiScope	Wavelet coefficients	An ensemble transformer detects heart murmurs for the prediction of clinical outcomes, achieving 75.70% accuracy.

architectures, such as CycleGAN, Wasserstein autoencoder GAN (WAE-GAN), and Pix2Pix-style models, enable not only data generation but also cross-modal translation (e.g., between ECG and PCG) and R-peak-independent generation, offering pathways for data harmonization and privacy-preserving synthetic data sharing (Dissanayake et al., 2023; Rayavarapu, Prasanthi, Kumar, Rao, & Prashanti, 2023).

Finally, GANs enable semi-supervised learning, in which the ability to generate realistic samples allows models to leverage abundant, unlabeled data. This approach reduces the need for large, fully annotated datasets while maintaining high performance in anomaly detection (Unnikrishnan et al., 2020).

5.2. VAEs-based models

Variational autoencoders (VAEs) have emerged as a powerful framework for addressing key challenges in signal analysis, including unsupervised anomaly detection. This capability is primarily realized through several sophisticated mechanisms. The first is reconstruction-based detection using a beta variational auto-encoder (β -VAE), where anomalies exhibit high reconstruction errors as the decoder struggles to process signals that are not encountered during training (Li et al., 2021). The second is latent space divergence, wherein the Kullback-Leibler (KL) divergence between the latent distribution of signals and a prior Gaussian distribution serves as a robust anomaly score, with

elevated values indicating abnormality (Banerjee & Ghose, 2020). The third mechanism involves structured latent space formulations, such as density-based VAEs (DBVAE), which replace the standard Gaussian prior with a density-based loss to explicitly cluster normal representations. This approach significantly enhances the detection performance of conventional VAEs (Li & Tian, 2021). Furthermore, advanced architectures, such as WaveNetVAE, incorporate temporal dependencies and employ maximum mean discrepancy regularization to prevent posterior collapse, substantially improving the modeling of complex signal patterns (Colt et al., 2020).

In addition to anomaly detection, VAEs are highly effective for nonlinear dimensionality reduction and feature enhancement. They outperform linear techniques, such as principal component analysis (PCA) and independent component analysis (ICA), by learning latent spaces that capture the complex, nonlinear manifolds of signal data. This enables effective clustering of signal types and the distinction of anomalies from normal patterns (Buchh, 2024). VAEs can also be integrated into pipelines to create fixed-length, enhanced feature embeddings that improve the performance of downstream classifiers, such as support vector machines (SVMs) and Gaussian mixture models (GMMs) (Adiban et al., 2019) and best discrepancy forest (BDF) (Zhang et al., 2023).

Ultimately, VAEs enable data augmentation and signal separation, thereby addressing the practical constraints of data analysis. They are used to generate synthetic signals to mitigate class imbalances in limited datasets, enabling the development of more accurate, lightweight classifiers for deployment on low-cost hardware. Their disentanglement capabilities are also leveraged to separate overlapping signals in noisy recordings, which is a critical step in improving data clarity and analytical value (Jamal et al., 2025; Zhang et al., 2023).

5.3. Diffusion models

Diffusion models represent a cutting-edge advancement in generative AI for addressing the pervasive challenges of data scarcity and class imbalance in signal databases. These models operate through a learned denoising process to synthesize high-fidelity signals, demonstrating significant potential for augmenting datasets and enhancing analytical algorithms. A key strength of this approach lies in its capacity for conditional generation, which enables the targeted synthesis of specific signal classes using various conditioning mechanisms. Text prompts or categorical labels can guide latent diffusion models to produce various types of signals (Tan et al., 2025). Denoising diffusion probabilistic models (DDPMs), such as DiffWave, can be conditioned on complementary data sources to ensure congruent signal generation (Abbott et al., 2026; Jamshidi et al., 2024).

The fidelity and utility of the synthetic data are rigorously validated using a multifaceted evaluation framework. Quantitative metrics, including MMD and JSD, confirm that the generated signals closely match the statistical distribution of the real data (Jamshidi et al., 2024). The Fréchet inception distance (FID) is also typically used to evaluate the quality of the generation model (Tan et al., 2025). Furthermore, synthetic data consistently outperform those from traditional models on objective benchmarks (Tan et al., 2025). The ultimate validation lies in downstream applications. Training models on datasets augmented with diffusion-generated samples leads to substantial improvements in robustness and generalization performance, as evidenced by significant gains in key performance indicators, including accuracy, balanced accuracy, and Matthew's correlation coefficient (MCC), in both in-distribution and out-of-distribution test sets (Abbott et al., 2026). This underscores the enhanced analytical capability facilitated by these synthetic samples.

5.4. Transformer-based models

Transformer-based architectures have been adapted to overcome challenges in signal analysis, demonstrating significant prowess in capturing both long-range contextual dependencies and intricate local patterns within signals for anomaly detection. A predominant trend is the development of hybrid models that synergistically combine CNNs and transformers. For instance, some models integrate specialized kernels into a modified transformer to jointly learn local and global features without positional encoding (Wang et al., 2025). Similarly, other frameworks employ a mixture of experts (MoE) to combine different architectures and process various signal representations for complementary feature learning (Han & Shaout, 2025). Other approaches use convolutional embedding and transform methods to enhance feature extraction before applying self-attention (Liu et al., 2024).

A critical consideration is the input representation of data. Many studies have converted raw, variable-length signals into structured formats suitable for Transformer processing. A common and effective strategy is to transform signals into time–frequency representations, which are then treated as images and classified using vision transformers (ViTs) (Han & Shaout, 2025; Jamil & Roy, 2023; Yang et al., 2023). Alternatively, some models directly process raw signals by first using convolutional layers for feature extraction and embedding before applying the self-attention mechanism of the transformer to the resulting sequence (Hassanuzzaman et al., 2023). Several studies have incorporated specialized training strategies to address the practical challenges of data analysis. Techniques such as focal loss have been adopted to mitigate class imbalance (Liu et al., 2024), whereas synthetic minority oversampling (SMOTE) has been used to augment minority classes before training ensemble models (Alkhodari et al., 2022). Furthermore, some models leverage large-scale pretraining to capture complex, high-dimensional patterns, improving generalization in downstream analytical tasks (Yang et al., 2023).

5.5. Meta-analytical synthesis

To move beyond the above descriptive cataloging toward critical synthesis, we quantitatively aggregated the methodological characteristics and performance claims across the 30 studies reviewed (Table 4). This analysis reveals substantial heterogeneity that fundamentally limits cross-study comparability and clinical interpretability of reported metrics. Key findings expose critical limitations in the current evidence base:

- **Dataset scale and imbalance:** Median training set size was 832 recordings (IQR: 490–2141). Critically, 22/30 studies (73%) relied exclusively on PhysioNet/CinC 2016, a composite dataset with severe class imbalance (89% normal vs. 11% pathological) and heterogeneous recording conditions across seven source databases (Liu et al., 2016). Only three studies utilized the larger CirCor DigiScope dataset ($\#recordings = 5272$; 35.9% murmurs) (Oliveira et al., 2022), limiting generalizability assessments.
- **Metric inflation through class imbalance:** GAN-augmented classifiers reported median accuracy of 96.5% (range: 89–99.9%), but 18/21 studies (86%) used accuracy on imbalanced test sets without reporting balanced accuracy, F1-score, or per-class sensitivity. For example, an 98% accuracy on PhysioNet may reflect 99.5% sensitivity for normal sounds but < 70% for diastolic murmurs (prevalence < 2% in the dataset)—a clinically catastrophic failure mode masked by aggregate metrics.
- **Physiological fidelity gap:** None of the 30 studies validated synthetic heart sounds against biomechanical or hemodynamic ground truth. While 19 studies (63%) reported statistical fidelity metrics (SSIM > 0.96, FID < 20), none of the studies verified whether synthetic murmurs preserved physiologically plausible

Table 4

Meta-analytical summary of reviewed studies by generative paradigm. Values represent medians (interquartile range) where applicable. Accuracy ranges are not directly comparable due to heterogeneous datasets, class definitions, and evaluation protocols.

Metric	GANs	VAEs	Diffusion	Transformers
# articles	13	8	3	7
Median dataset size (records)	1050 (490–2141)	832 (490–1841)	3132 (2000–5272)	1000 (832–2141)
Reported accuracy (%)	96.5 (89–99.9)	94.2 (88.5–99.94)	93.4	97.5 (75.7–99.98)
Balanced accuracy reported	2/13	3/8	1/3	2/7
Clinician validation of synthetics	0/13	1/8	1/3	0/7
Physiological fidelity validation (S1–S2 timing, spectral realism)	0/13	0/8	0/3	0/7
XAI techniques implemented (attention maps, latent traversal, counterfactuals for <i>generation</i>)	0/13	0/8	0/3	0/7
Synthetic data governance addressed (provenance, privacy, bias auditing)	0/13	0/8	0/3	0/7
Code publicly available	2/13	1/8	1/3	0/7
Single-dataset evaluation	11/13	7/8	2/3	6/7

timing relative to S1/S2 intervals, matched spectral characteristics with expected turbulence patterns (> 200 Hz for stenotic murmurs), or correlated with simulated pressure gradients. Statistical similarity does not guarantee clinical validity—a synthetic murmur achieving SSIM = 0.98 may exhibit non-physiological timing that would mislead clinicians.

- **Interpretability deficit:** Despite 12/30 studies (40%) acknowledging “explainability” as a future direction in their conclusions, zero studies implemented or evaluated XAI techniques for generative processes. No study has employed attention rollout maps to visualize which cardiac cycle segments (S_1 , systole, S_2 , diastole) drive murmur synthesis, latent traversal with physiological constraints to map dimensions to hemodynamic parameters, or counterfactual generation to isolate pathological features. One transformer study (Wang et al., 2025) included attention visualization for classification (not generation), but no study implemented XAI for generative processes. This opacity prevents clinicians from auditing why a synthetic murmur was generated at a specific temporal location.
- **Regulatory and ethical oversight:** All 30 studies (100%) omitted discussion of synthetic data governance despite generating medically sensitive signals. Critical dimensions absent from every study include (a) data provenance tracking (immutable logs linking synthetic samples to generator version, training subset, and random seed), (b) privacy risk assessment against membership inference attacks, (c) bias auditing for demographic representation in synthetic outputs, and (d) liability allocation frameworks for diagnostic errors arising from models trained on synthetic data.
- **Validation gap:** 19/30 studies (63%) validated synthetic samples using statistical metrics only (SSIM > 0.96, FID < 20, MMD < 0.001). Only two studies (Abbott et al., 2026; Tan et al., 2025) included clinician evaluations of synthetic murmurs, and none demonstrated physiological grounding (e.g., correlation between synthetic murmur timing and simulated valve pressure gradients). Statistical fidelity does not guarantee clinical validity—a synthetic murmur may achieve SSIM = 0.98 yet exhibit non-physiological timing relative to S1/S2.
- **Reproducibility crisis:** Only four studies (13%) released training code; seven (23%) shared synthetic datasets. This opacity impedes the independent validation of performance claims such as “99% CAD detection sensitivity” (Chakraborty et al., 2025), which may

reflect dataset-specific overfitting rather than generalizable capability. Notably, transformer-based studies showed the poorest reproducibility (0/7 released codes).

- **Evaluation protocol weaknesses:** 24/30 studies (80%) used accuracy as the primary metric, despite class imbalance. Only six reported confusion matrices stratified by pathology type. Three studies (Dissanayake et al., 2023; Rayavarapu, Prasanthi, Kumar, Rao, & Prashanti, 2023; Rayavarapu, Prasanthi, Kumar, Rao, & Singham, 2023) used CycleGAN for cross-modal translation (ECG→PCG) but evaluated classifiers on synthetic PCGs generated from real ECGs, effectively leaking label information through the conditioning signal and inflating reported performance.

6. Datasets

Heart sound datasets are essential resources for advancing automated detection of cardiovascular diseases. The datasets described here constitute a diverse, international collection of PCG recordings, each with distinct characteristics. They vary significantly in size, recording quality, sampling rates, and class distribution, with many exhibiting the common challenge of a severe imbalance between normal and pathological cases. Large public benchmark repositories, such as the PhysioNet and CirCor challenges (Liu et al., 2016; Oliveira et al., 2022), provide valuable research resources. In addition, more specialized and clinically annotated libraries provide complementary data for focused investigations. Together, these datasets provide essential raw materials for developing, training, and evaluating deep models to improve non-invasive cardiac prescreening. Fig. 8 shows the number of times each dataset was used in the reviewed articles.

6.1. PhysioNet/computing in cardiology (CinC) challenge 2016 dataset

The physioNet/computing in cardiology (CinC) challenge 2016 dataset comprises a comprehensive collection of heart sound data, consisting of nine smaller databases gathered from seven different countries, resulting in a diverse and international set of audio recordings (Abbott et al., 2026; Adiban et al., 2019; Banerjee & Ghose, 2020; Chakraborty et al., 2025; Dissanayake et al., 2023; Gawande et al., 2025; Jamal et al., 2025; Jamshidi et al., 2024; Li & Tian, 2021; Li et al., 2021; Liu et al., 2024, 2016; Narváez & Percybrooks, 2020; Rayavarapu, Prasanthi, Kumar, Rao, & Prashanti, 2023; Rayavarapu,

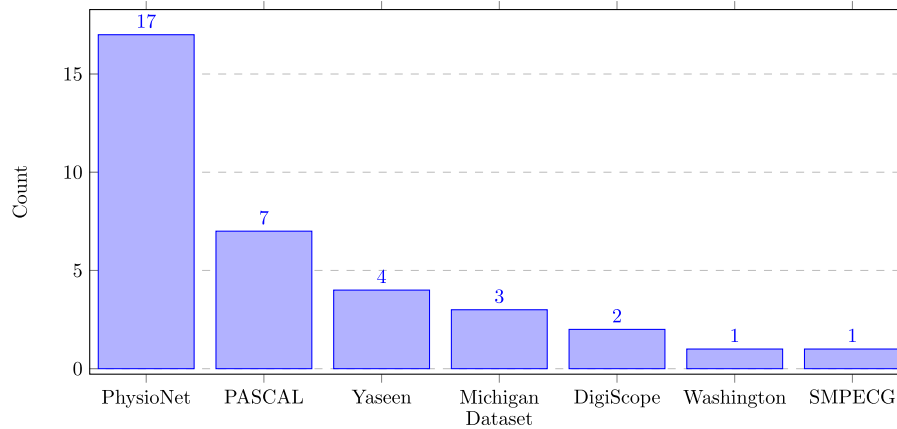


Fig. 8. Number of usages of each dataset in the reviewed articles.

Prasanthi, Kumar, Rao, & Singham, 2023; Takezaki & Kishida, 2022; Tan et al., 2025; Unnikrishnan et al., 2020). This composite dataset contains 3200 recordings obtained from healthy individuals and patients with various heart conditions across different age groups. Common sampling rates found in the dataset include 2000 Hz, 4000 Hz, and 44.1 kHz. This publicly available dataset has become a benchmark for developing and testing automated heart sound analysis algorithms. However, it is characterized by a significant class imbalance, with normal heart sounds vastly outnumbering pathological cases.

The Aalborg University heart sound database is a sub-database within PhysioNet that was specifically leveraged for coronary artery disease (CAD) (Chakraborty et al., 2025). It consists of 490 heart sound audio recordings. Exploratory analysis revealed a severe class imbalance within AADHSDB, with nearly 79% of the recordings belonging to healthy individuals and only 104 audio files (21%) representing CAD cases. The Dalian University of Technology heart sound database is another sub-database from PhysioNet and serves as a major source of data owing to its larger size (Chakraborty et al., 2025). It contains 2141 heart sound recordings, making it substantially larger than the AADHSDB. However, it exhibits an even more pronounced class imbalance, with an overwhelming 91.5% of recordings (1958 files) from healthy individuals and only 8.5% (183 files) from CAD patients.

6.2. PASCAL heart sounds database

The PASCAL heart sounds database includes recordings of clinically annotated murmurs that can be indicative of coronary artery disease (CAD) (Banerjee & Ghose, 2020; Bentley, Nordehn, Coimbra, Mannor, & Getz, 2011; Buchh, 2024; Chakraborty et al., 2025; Colt et al., 2020; Han & Shaout, 2025; Hu et al., 2025; Vinay et al., 2024). It contains 832 recordings collected from healthy participants and patients with cardiac abnormalities. The recordings in the database were divided into two main sets: Set A with 176 recordings (sampled at 8 kHz) and Set B with 656 recordings (sampled at 44.1 kHz). The use of this dataset alongside others helped ensure that the generative model was trained on acoustically diverse examples of abnormal cardiac sounds, improving the potential generalizability of the synthetic samples.

6.3. Michigan heart sounds and murmurs library

The Michigan heart sounds and murmurs library comprises a specialized collection of expert-annotated cardiac auscultation recordings (Banerjee & Ghose, 2020; Chakraborty et al., 2025; Hu et al., 2025; Judge, 2015). This clinically refined resource contains over 200 high-quality heart sound recordings, including clear examples of systolic and diastolic murmurs, valvular abnormalities, and other pathological acoustic signatures. The library is particularly valued for its inclusion of

rare diastolic murmurs, often associated with conditions such as aortic regurgitation or mitral stenosis, which are systematically documented with clinical context and hemodynamic correlates. All recordings were sampled at 44.1 kHz, preserving critical acoustic details for human interpretation and machine learning applications.

6.4. The CirCor DigiScope dataset

The CirCor DigiScope dataset is a large, publicly available collection of 5272 heart sound recordings from 3120 patients (Alkhodari et al., 2022; Jamshidi et al., 2024; Oliveira et al., 2022). It was specifically designed to develop algorithms for detecting clinically significant heart murmurs and triaging clinical outcomes. This dataset contained a much larger volume of data, with the prevalence of murmurs reported as 35.9% (1894 recordings). It was formally split into a public training set (3132 recordings), a hidden validation set (662 recordings), and a hidden test set (1478 recordings) to rigorously evaluate the generalizability of the models to new patient populations, which is a key challenge in the field.

6.5. The Washington university heart sounds dataset

The Washington university heart sounds dataset is a publicly available collection of digital PCG recordings, often used in early research on automated heart sound analysis (Hu et al., 2025; The public heart sound dataset of the University of Washington, 2022). It comprises 1000 recordings collected from patients, including both normal and pathological cases, such as murmurs and other cardiac abnormalities. The recordings were typically made using electronic stethoscopes or similar transducers, often at a sampling rate of 8 kHz or 11 kHz, and include annotations indicating the presence and timing of fundamental heart sounds (S_1 and S_2) and any murmurs. The dataset is particularly noted for its use in validating segmentation algorithms and testing preliminary classification models in the pre-deep learning era.

6.6. Yaseen et al.

This well-known heart sound dataset, frequently employed as a benchmark for multi-class classification tasks, comprises a total of 1000 audio records meticulously balanced across five distinct categories: normal, aortic stenosis, mitral stenosis, mitral regurgitation, and mitral valve prolapse, with exactly 200 records per class (Jamil & Roy, 2023; Wang et al., 2025; Yang et al., 2023; Yaseen, Son, & Kwon, 2018; Zhang et al., 2023). Compiled from multiple sources to ensure diversity, all recordings were standardized through resampling at a frequency rate of 8000 Hz and converted to a mono-channel format to ensure consistency for algorithmic processing. The structured and balanced design of the dataset has made it particularly valuable for developing and evaluating machine learning models that aim to distinguish specific valvular heart conditions beyond the binary normal-abnormal classification.

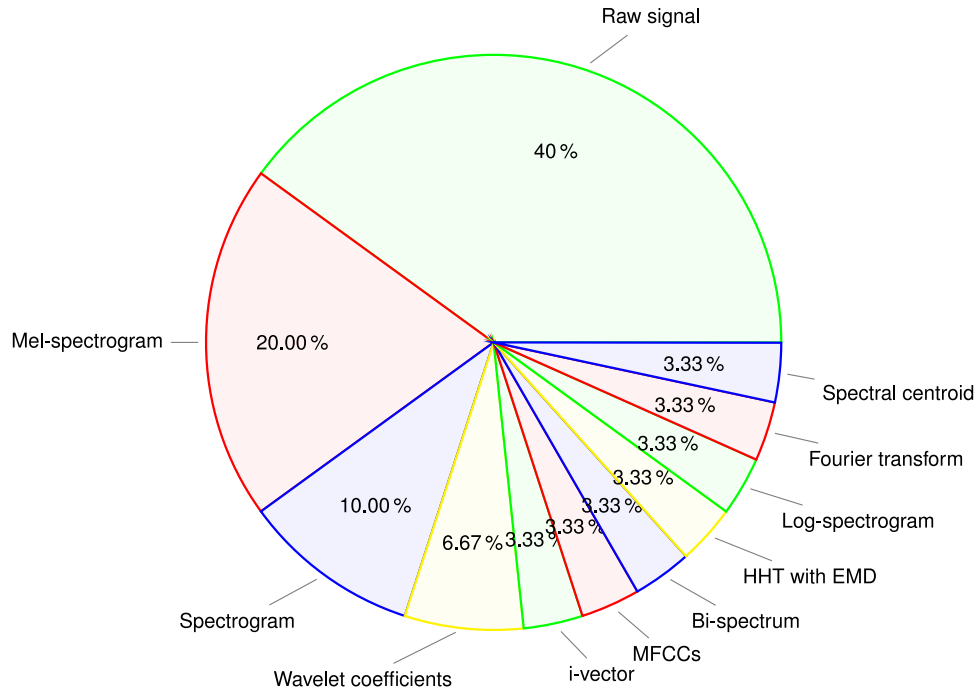


Fig. 9. Distribution of different data representation methods used in the reviewed articles.

6.7. Synchronized multichannel PCG and ECG (SMPECG) dataset

The synchronized multichannel PCG and ECG (SMPECG) dataset consists of synchronized multi-channel PCG and ECG recordings collected using a wearable vest equipped with six digital stethoscopes and a three-lead ECG sensor (Abbott et al., 2026; Rong, Fynn, & Nordholm, 2004). The stethoscopes were placed at clinically relevant positions on the chest to capture heart sounds from different valves and arteries. Using this vest, data were collected from 105 participants, some of whom 46 were diagnosed with coronary artery disease. The signals were sampled at 2 kHz and high-pass filtered at 20 Hz. The dataset was designed to support the development of machine learning models for pre-screening coronary artery disease (CAD) by providing high-integrity, noise-resistant, temporally aligned cardiac signals. This is particularly valuable for research on adaptive noise cancellation, feature extraction, and multimodal classification.

7. Data representation

The effective analysis of physiological and acoustic signals, such as heart sounds, speech, and environmental audio, requires robust data representation techniques that can capture the essential temporal, spectral, and statistical characteristics. Raw signals often contain rich information but are typically high-dimensional and sensitive to noise. The field of signal processing has developed a rich suite of methods for transforming and representing signals, revealing hidden structures, and creating meaningful feature sets. These transformations project the signal into new domains, such as frequency or time-frequency, where specific properties become more salient and easier to interpret for both humans and machines. The choice of representation is fundamental, as it directly influences the performance of downstream tasks, such as machine learning, pattern recognition, anomaly detection, and data compression. Fig. 9 shows the distribution of different data representations in the reviewed articles. The following sections briefly explain the data representation methods employed by the models for heart sound data.

7.1. Raw signal

The raw signal is the most fundamental representation that captures the amplitude of a phenomenon that varies with time. It is the direct output of a sensor, such as a microphone or stethoscope, that is sampled at a fixed rate (Chakraborty et al., 2025; Colt et al., 2020; Dissanayake et al., 2023; Gawande et al., 2025; Hassanuzzaman et al., 2023; Hu et al., 2025; Jamal et al., 2025; Jamshidi et al., 2024; Lee, Su, Hsieh, Huang, Lee et al., 2023; Liang et al., 2022; Narváez & Percybrooks, 2020; Rayavarapu, Prasanthi, Kumar, Rao, & Prashanti, 2023; Rayavarapu, Prasanthi, Kumar, Rao, & Singham, 2023). Although it contains complete information on the original event, its features (such as frequency content) are not explicitly available. For complex signals, such as PCG signals, the raw waveform is highly non-stationary and difficult for machines to interpret directly, making it a challenging starting point for many analysis tasks, despite its simplicity.

7.2. Fourier transform (FT)

The Fourier transform is a foundational mathematical operation that decomposes a signal from its native time domain into its constituent frequencies (Buchh, 2024; Yoganathan, Gupta, Corcoran, Udawadia, Sarma et al., 1976). It converts the raw signal into a frequency spectrum, revealing the amplitude (and phase) of each frequency component present throughout the entire duration of the signal, with the fast Fourier transform (FFT) being an efficient algorithm. The critical limitation of the standard FT is that it is entirely global; it shows the frequencies that are present but provides no information about when they occur, making it unsuitable for analyzing signals whose frequency content changes over time.

7.3. Spectrogram

A spectrogram addresses the limitation of FT by providing a time-frequency representation. It is created by computing the FFT on short, consecutive, and often overlapping windowed segments of the signal

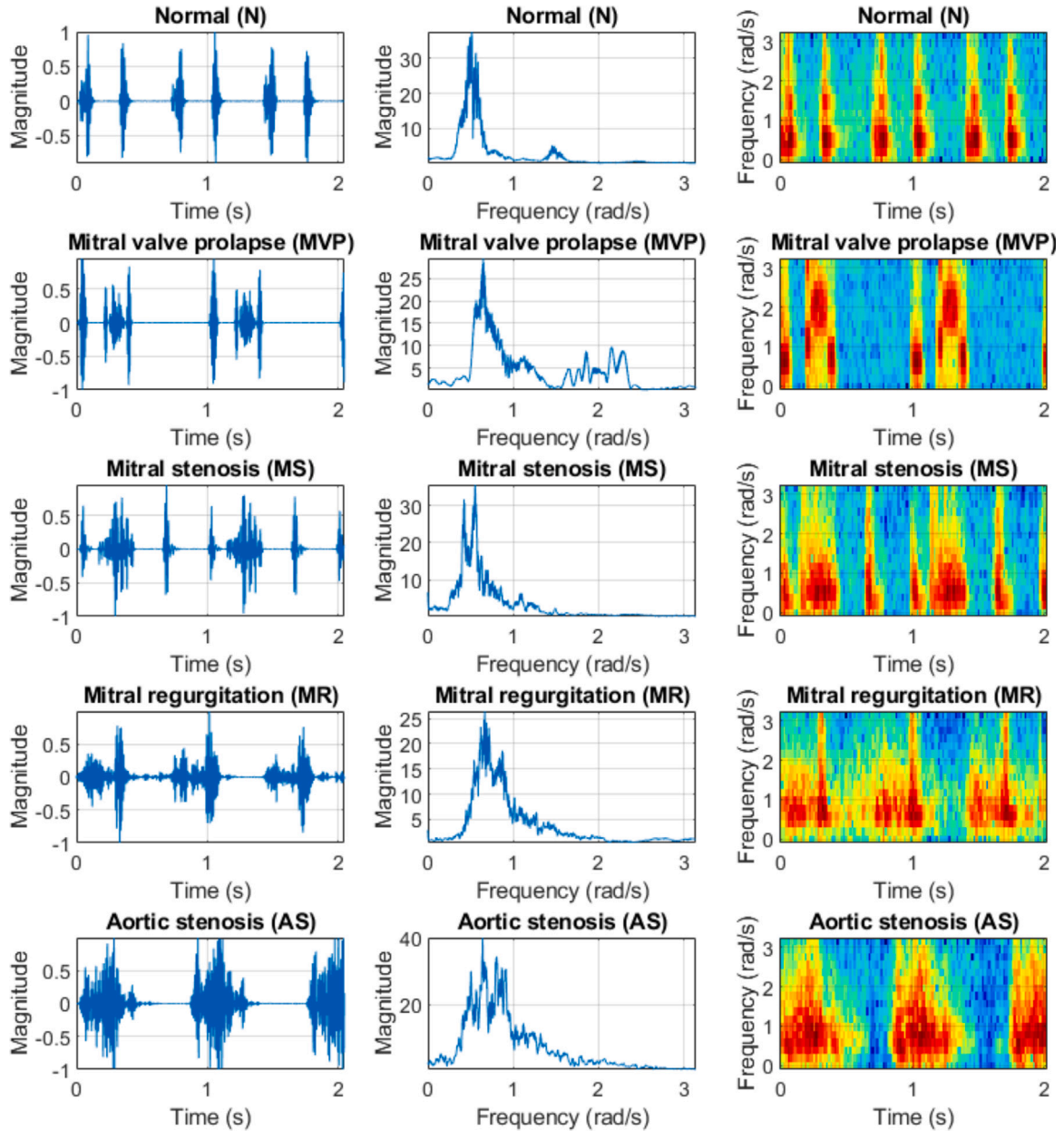


Fig. 10. PCG signals of a healthy heart and four with valvular conditions and their Fourier transforms and spectrograms at a sampling rate of 1000 Hz. These signals were extracted from the Yaseen et al. dataset (Yaseen et al., 2018).

and stacking the spectra over time (Lei, 2004; Takezaki & Kishida, 2022; Yang et al., 2021, 2023). The resulting 2D plot has time on the x -axis, frequency on the y -axis, and intensity represented by color. This allows an analyst to see how the spectral content of a signal, such as a bird song or a spoken word, evolves, making it one of the most intuitive and widely used tools for audio analysis. Fig. 10 shows plots of PCG signals and corresponding Fourier transforms and spectrograms.

7.4. Mel-spectrogram

A mel-spectrogram is a perceptually inspired version of the standard spectrogram (Abbott et al., 2026; Li & Tian, 2021; Li et al., 2021; Liu et al., 2024; Tan et al., 2025; Unnikrishnan et al., 2020; Zhou, Chen, & Chien, 2021). It applies the Mel scale, a nonlinear scale based on human auditory perception, to the frequency axis. This scale approximates the

human ear's response: it is more discriminative at lower frequencies (where humans are more sensitive to small changes in pitch) and compressive at higher frequencies. By warping the frequency bins, the Mel-spectrogram creates a representation that aligns more closely with how humans hear, which is why it is the dominant input feature for modern deep learning models in speech and music processing.

7.5. Log-spectrogram

A logarithm of the amplitude (or power) values is created in a standard or mel-spectrogram (Banerjee & Ghose, 2020; Nguyen, Lin, & Huang, 2022). Human loudness perception follows a logarithmic scale: doubling the sound power does not double the perceived loudness. Applying a log transform (e.g., converting power to decibels) compresses the dynamic range of amplitude values. This enhances quieter

spectral components, which are often perceptually important. Louder components can overshadow them in a linear spectrogram, leading to improved performance in audio analysis and machine learning.

7.6. Mel-frequency cepstral coefficients (MFCCs)

MFCCs are among the most significant features used for speech recognition (Abdul & Al-Talabani, 2022; Jamil & Roy, 2023). The extraction process involves calculating the mel-spectrogram, taking the logarithm of the mel-band energies (log-mel-spectrogram), and applying a discrete cosine transform (DCT) to decorrelate the energies and compress the information into a small number of lower order coefficients. The resulting MFCCs effectively represent the spectral envelope (the “shape” of the spectrum) of a short-time frame, which characterizes the vocal tract filter in speech production. They are compact, computationally efficient, and highly effective for speech modeling.

7.7. Bi-spectrum

The bi-spectrum is a higher-order spectral analysis tool that goes beyond the power spectrum (FT magnitude squared) (Ergen & Tatar, 2005; Zhang et al., 2023). It analyzes the relationships between three frequency components by capturing phase coupling between them. Whereas the power spectrum characterizes a Gaussian process, the bi-spectrum can detect nonlinear interactions and deviations from Gaussianity. This makes it valuable for specialized applications, such as fault detection in mechanical systems (e.g., identifying specific gear faults) and analyzing biomedical signals.

7.8. Spectral centroid

The spectral centroid is a single-value feature that represents the “center of mass” of a spectrum (Ferdoushi, Paul, & Fattah, 2019; Han & Shaout, 2025). It is calculated as the weighted mean of the frequencies present in the signal, weighted according to their magnitude. Perceptually, it is strongly correlated with the brightness or sharpness of a sound. A high spectral centroid indicates a sound rich in high-frequency energy, whereas a low spectral centroid indicates a duller, bass-heavy sound. It is a fundamental descriptor of audio signal processing for retrieving music information and classifying sounds.

7.9. Wavelet coefficients

Wavelet analysis provides a multi-resolution time–frequency representation (Alkhodari et al., 2022; Delyon & Juditsky, 1995; Wang et al., 2025). Unlike the fixed window size of the STFT (used for spectrograms), wavelets utilize basis functions (“wavelets”) that can be stretched (for low-frequency analysis) and compressed (for high-frequency analysis). This allows for good frequency resolution at low frequencies and good time resolution at high frequencies. The resulting wavelet coefficients are ideal for analyzing signals with transient sharp events or features occurring at different scales, such as ECG heartbeat signals.

7.10. Hilbert-Huang transform with empirical mode decomposition (HHT with EMD)

The HHT is a two-step adaptive method for analyzing nonlinear and nonstationary data (Ompokov & Boronoev, 2019; Vinay et al., 2024). First, empirical mode decomposition (EMD) adaptively decomposes a signal into intrinsic mode functions (IMFs), which are simple oscillatory functions. Then, the Hilbert Transform is applied to each IMF to obtain instantaneous frequency and amplitude data. Unlike Fourier or wavelet analysis, which uses predefined basis functions, HHT derives its basis from the data itself, making it highly effective for analyzing complex real-world signals such as ocean waves, structural vibrations, and physiological data.

7.11. Identity vector (i-vector)

An *i*-vector is not a direct signal transformation but a high-level feature representation primarily used in speaker recognition (Adiban et al., 2019; Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2011). The core idea is to model a variable-length speech utterance (e.g., represented by many MFCC frames) as a single, fixed-length, low-dimensional vector. This vector is extracted from a statistical model (Gaussian mixture or universal background models). It represents the vocal characteristics of the signal source and other factors, such as the channel and environment. The “*i*” stands for “identity”, as these vectors summarize the identity of the source in a compact form that is suitable for machine learning classifiers.

8. Clinical applications

Cardiac auscultation remains a fundamental diagnostic technique in cardiovascular medicine; however, its effectiveness is limited by inter-rater variability, background noise, and the subtlety of many pathological sounds. The development of automated systems for heart sound analysis has been hindered by several data-related challenges, including the scarcity of labeled pathological samples, class imbalance, signal quality issues, and privacy concerns that restrict data sharing. Recent advances in generative AI, particularly GANs, VAEs, diffusion models, and transformers, offer promising solutions to these limitations. This study organizes and analyzes the clinical applications of these technologies, as demonstrated in recent literature, providing a structured framework for understanding how generative AI is transforming cardiac diagnostics. Fig. 11 shows block diagrams of the clinical applications in the reviewed literature.

8.1. Data augmentation for precision diagnostics

Generative models address the critical challenge of class imbalance in medical datasets by generating high-fidelity synthetic pathological heart sound data. This approach enables the development of more accurate and generalizable classifiers for specific cardiac conditions than those in previous studies. For the detection of coronary artery disease (CAD), augmented data using GANs have significantly improved detection accuracy, facilitating the development of scalable noninvasive screening tools (Chakraborty et al., 2025; Takezaki & Kishida, 2022). Similarly, for chronic heart failure (CHF), data augmented using GANs improves transfer learning models, creating robust automated detection systems (Gawande et al., 2025). In the diagnosis of left ventricular diastolic dysfunction (LVDD), deep convolutional GANs create synthetic spectrograms that train CNNs, establishing PCG analysis as a viable, low-cost alternative to echocardiography (Yang et al., 2021). For the diagnosis of valvular heart disease (VHD), synthetic data augmentation using GANs combined with vision transformers enables efficient and accurate detection from spectrogram representations (Jamil & Roy, 2023).

Beyond standard GANs, advanced generative models, such as latent diffusion models and denoising diffusion probabilistic models, enable precise conditional generation. These models can be guided by pathological labels (Tan et al., 2025) or synchronous ECG signals (Abbott et al., 2026) to produce clinically relevant synthetic data that are tailored to specific diagnostic objectives. Empirical evidence confirms the efficacy of this approach; a diffusion-based generative model with a U-Net backbone trained on mel-spectrograms of the PhysioNet dataset achieved FID=18.39 for data augmentation. Incorporation of the synthesized data improved the semi-supervised classification accuracy of PCG signals to 93.4%, demonstrating that high-fidelity synthetic data can directly enhance diagnostic performance in data-scarce clinical environments (Tan et al., 2025).

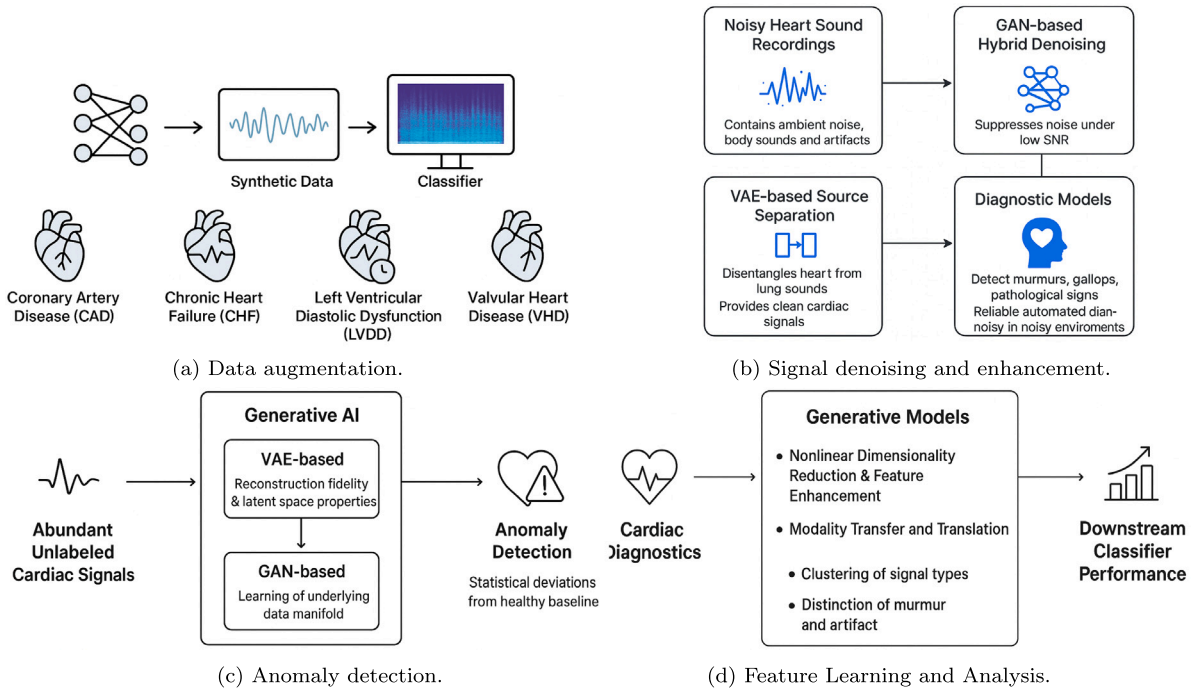


Fig. 11. The clinical applications tackled in the reviewed articles.

8.2. Signal denoising and enhancement for improved diagnostic clarity

Generative models can be used to improve signal quality, which is essential for reliable diagnosis of arrhythmias. Heart sound recordings are often corrupted by ambient noise, internal body sounds, and measurement artifacts, which can obscure clinically relevant information. Hybrid denoising models built using GANs integrated with advanced signal processing techniques effectively suppress noise under low SNR conditions (Hu et al., 2025; Liang et al., 2022). For example, the model (Hu et al., 2025) combined GAN with parameterless scale-space boundary detection and empirical wavelet transform (EWT) and was evaluated on noisy PCG recordings from the PASCAL, Michigan, and Washington datasets. Under Gaussian noise, the method improved the SNR from -5 dB to 9.26 dB. This enhancement enables clearer heart sound representations and supports reliable automated diagnosis in low-quality recording conditions.

VAEs have been successfully used for source separation, specifically disentangling heart sounds from lung sounds in auscultation recordings (Jamal et al., 2025). This capability serves as a critical preprocessing step, ensuring that subsequent analyses are performed on clean cardiac signals and dramatically improving the detection accuracy of murmurs, gallops, and other pathological indicators. These enhancement techniques form a vital foundation for building robust cardiac diagnostic systems that can be deployed in acoustically challenging clinical settings.

8.3. Unsupervised and semi-supervised anomaly detection

Generative AI enables robust anomaly detection despite extreme data scarcity for pathological conditions, operating through semi-supervised or unsupervised paradigms that learn representations of normal cardiac physiology from abundant unlabeled data. Pathologies are identified as statistical deviations from learned healthy baselines. VAE-based detection frameworks utilize reconstruction fidelity and latent space properties, where models trained on normal sounds exhibit a high reconstruction error or large Kullback-Leibler (KL) divergence when processing anomalous signals (Banerjee & Ghose, 2020; Li & Tian, 2021; Li et al., 2021). For instance, a semi-supervised β -VAE

was applied to log-spectrogram representations of PCG signals from the PhysioNet, PASCAL, and Michigan datasets, with anomaly scores derived from KL divergence in the latent space (Banerjee & Ghose, 2020). The model achieved sensitivities ranging from 89% to 99%, demonstrating the clinical utility of VAE-based anomaly screening and early detection of rare or underrepresented cardiac abnormalities.

Semi-supervised approaches using GANs leverage their ability to learn the underlying data manifold of normal sounds, using abundant unlabeled data to improve feature learning and discrimination capabilities while reducing the dependency on expensive curated datasets (Unnikrishnan et al., 2020). These strategies overcome significant barriers to medical AI by enabling accurate diagnostic models for rare conditions that are poorly represented in clinical datasets, thus facilitating early detection and equitable healthcare outcomes.

8.4. Advanced feature learning and multi-modal analysis

Generative models advance cardiac diagnostics by learning compressed feature representations and integrating information across various data modalities. In nonlinear dimensionality reduction and feature enhancement, VAEs significantly outperform traditional linear methods by learning latent spaces that capture the complex nonlinear manifolds of heart sound data. This enables more effective clustering of signal types and improves the distinction between pathological murmurs and artifacts, thus improving the downstream classifier performance (Adiban et al., 2019; Buchh, 2024).

Generative models also facilitate modality transfer and translation, bridging different types of physiological data. Frameworks such as CycleGAN enable cross-modal translation between ECG and PCG signals (Dissanayake et al., 2023; Rayavarapu, Prasanthi, Kumar, Rao, & Prashanti, 2023; Rayavarapu, Prasanthi, Kumar, Rao, & Singham, 2023). This allows data harmonization and presents opportunities for privacy-preserving data sharing through realistic synthetic substitutes for sensitive patient data. These advancements have been realized in hybrid deep learning architectures that combine recurrent networks with GANs (Vinay et al., 2024) and integrate wavelet transforms with transformers (Wang et al., 2025) and ensemble vision transformers with CNNs (Han & Shaout, 2025). These sophisticated systems achieve the

Table 5
Comparative analysis of generative AI models for heart sound analysis.

Feature	GANs	VAEs	Diffusion	Transformers
Primary Application	Data augmentation, signal enhancement, cross-domain synthesis.	Unsupervised anomaly detection, dimensionality reduction, latent feature learning.	High-fidelity data augmentation, conditional signal synthesis, denoising.	Attention-based murmur classification, anomaly detection, long-range dependency modeling.
Key Strengths	Generates sharp, realistic synthetic samples; effective for semi-supervised learning and domain translation.	Stable training; structured and interpretable latent space; suitable for anomaly scoring (e.g., KL divergence).	Superior sample quality and diversity; stable optimization; denoising-oriented formulation.	Exceptional ability to model global context; adaptive attention to clinically relevant heart sound segments; flexible integration with models.
Limitations & Challenges	Training instability and mode collapse; lack of explicit likelihood estimation; sensitive to hyperparameter tuning; limited interpretability for generative process; requires careful validation of physiological plausibility.	Tendency toward overly smooth reconstructions; limited realism compared to adversarial models; difficulty capturing fine temporal details; latent space may not align with clinical features.	High computational cost; slow inference due to iterative sampling process; clinically validation limited to statistical metrics; lack of physiological grounding.	Not inherently generative in standard form; computationally demanding; typically requires large-scale datasets; opaque attention mechanisms without clinical correlation; limited XAI integration for generation tasks.
Clinical Impact	Enhances diagnostic robustness through realistic synthetic data; improves performance under noisy acquisition conditions; but may lack physiological validity.	Facilitates early detection of rare or abnormal cardiac patterns with minimal labeled data; enriches feature representations; useful for anomaly screening but limited in perceptual quality.	Produces highly plausible synthetic heart sounds to improve generalization and robustness of diagnostic models; promising but not yet validated in clinical trials.	Improves classification accuracy for complex cardiac conditions by leveraging global temporal and contextual information; yet interpretability and clinical trust remain barriers.

best performance in critical diagnostic tasks, including the identification of congenital heart diseases in pediatric populations (Hassanuzzaman et al., 2023) and heart murmur detection (Alkhodari et al., 2022). For instance, the ensemble transformer model presented in Alkhodari et al. (2022) leveraging wavelet-derived PCG representations from the CirCor DigiScope dataset achieved 75.7% murmur detection accuracy. The study indicates that advanced attention-based feature learning can serve as a scalable foundation for multimodal cardiac triage and outcome-oriented clinical decision support.

9. Discussion

GANs, VAEs, diffusion models, and transformer-based architectures represent distinct yet complementary paradigms of generative modeling in heart sound analysis. Each model family addresses key challenges in processing PCG signals, such as data scarcity, noise contamination, and complex temporal dependencies, from different theoretical and computational perspectives. Understanding their relative strengths and limitations is essential for selecting appropriate architectures for clinical and research-oriented applications. Table 5 compares the generative AI models for heart sound analysis.

GAN-based methods are mainly used in PCG analysis for high-fidelity data augmentation and signal enhancement. By learning an adversarial mapping between real and synthetic signals, they generate realistic heart sounds that improve classifier robustness, particularly under limited data and for rare pathological classes. However, training

instability, mode collapse, and the absence of explicit likelihood estimation limit clinical validation and reproducibility, requiring careful training and comprehensive evaluation strategies.

In contrast, VAEs focus on stable probabilistic representation learning by encoding heart sounds into structured latent spaces, making them well-suited for unsupervised or minimally labeled settings. However, despite their interpretability and consistent training behavior, VAEs typically produce over-smoothed reconstructions that lack fine temporal details, reducing their effectiveness for high-fidelity data augmentation compared to adversarial models.

Recently, diffusion models have been adopted for high-quality heart sound generation by learning to reverse a gradual noise corruption process, aligning naturally with denoising and reconstruction tasks in PCG analysis. They offer superior sample realism and diversity with stable training, often surpassing GANs, but their high computational cost and slow iterative sampling limit real-time and large-scale clinical deployment.

Transformer-based models are not inherently generative but excel at modeling long-range temporal dependencies via attention. In heart sound analysis, they capture global context across cardiac cycles to improve murmur classification and anomaly detection. Moreover, in hybrid generative frameworks, they guide conditional synthesis or enhance clinically salient features. However, their reliance on large datasets and high computational cost limits standalone use in data-scarce clinical settings.

The introduced meta-analytical synthesis reveals critical limitations that temper current enthusiasm and highlight gaps between technical

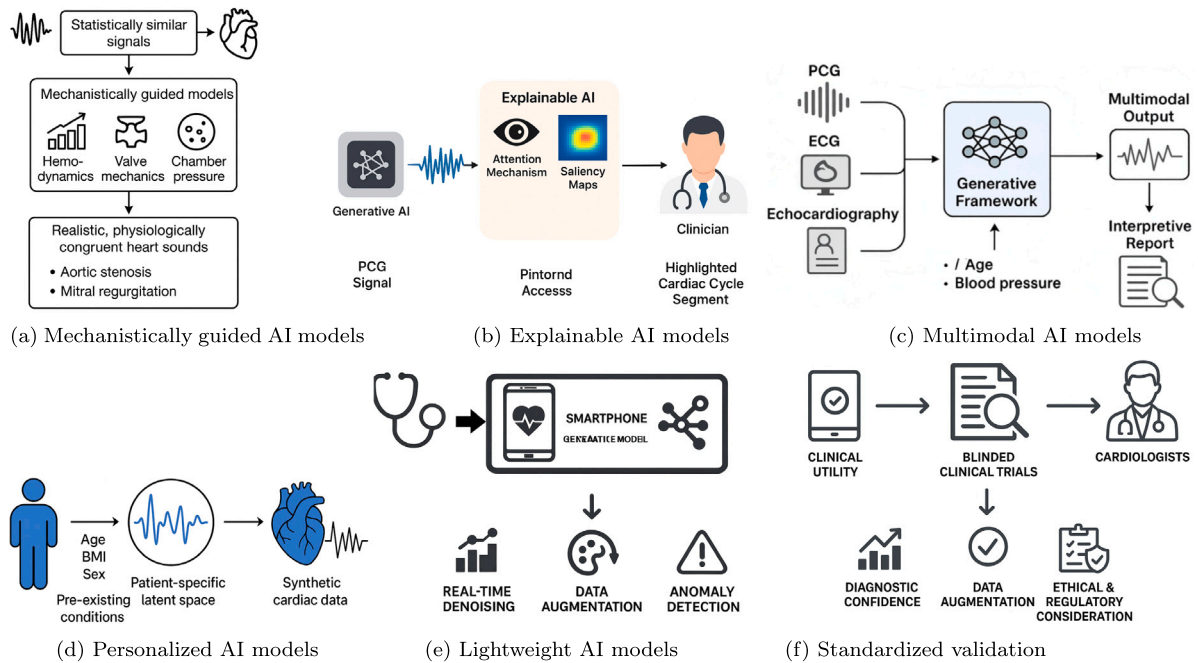


Fig. 12. Future research directions.

performance and clinical readiness. First, the heavy reliance on imbalanced public datasets, notably PhysioNet/CinC 2016, has led to inflated accuracy metrics that mask poor performance in rare pathological classes. Second, the physiological fidelity gap remains unaddressed; none of the reviewed studies validated synthetic heart sounds against biomechanical or hemodynamic ground truths.

Moreover, the interpretability deficit is pronounced. Despite frequent calls for explainable AI in future studies, no study has implemented XAI techniques, such as attention maps or latent space traversal, to clarify how synthetic sounds are generated or which signal segments drive diagnostic decisions. This opacity limits clinical trust and hinders model auditing processes. In addition, issues of reproducibility and governance have been largely overlooked. Few studies released code, and none discussed synthetic data provenance, privacy risks, bias auditing, or liability frameworks—essential considerations for regulatory approval and ethical deployment.

The clinical applications reviewed in this study demonstrate promising use cases (data augmentation, denoising, anomaly detection, and multimodal integration); however, their translational impact remains constrained by methodological and validation gaps.

10. Future research directions

Despite significant progress, generative AI for heart sound analysis must now transition from technical demonstration to clinical adoption. Key challenges persist, including limited physiological fidelity, poor model interpretability, computational inefficiency, and a lack of standardized clinical validation. Therefore, future research must prioritize the creation of physiologically accurate, interpretable, and efficient tools validated for real-world use, ultimately transforming this technology into an accessible and actionable diagnostic aid that enhances patient care. Fig. 12 illustrates key future research directions.

10.1. Enhancing the physiological accuracy and clinical relevance of the generated signals

Current generative AI models produce signals that are statistically similar to real data. However, they may lack precise physiological congruence with the human body. To move beyond pattern replication and

toward a true digital simulation of cardiac acoustics, research should prioritize mechanistically guided generative models. This approach involves incorporating cardiovascular principles, such as hemodynamics, valve mechanics, and chamber pressure, directly into the generative process (Almasi, Shamsollahi, & Senhadji, 2011; Sel et al., 2024; Šeman, Stephens, Kaye, Gregory, & Stub, 2024). For example, integrating the differential equations governing blood flow or biomechanical simulations of valve closure can constrain the model. This ensures that synthetic heart sounds are realistic and accurately reflect specific, quantifiable pathophysiological states. Moreover, guided generative AI, such as the generation of specific valvular abnormal heart sounds, addresses label reliability, physiologically grounded interpretability, and the cautious use of generated labels for pre-training or regularization.

10.2. Improving the explainability of generative models for their adoption in clinical settings

Clinicians must trust and understand the results of generative AI models to make better decisions. Future models should provide not only predictions or synthetic data but also clinically meaningful explanations for their results. Research could explore advanced explainable artificial intelligence (XAI) using attention mechanisms and saliency maps. These tools should be specifically designed for the temporal structure of PCG signals (Majhi & Kashyap, 2024). They can visually highlight specific segments of the cardiac cycle, such as the crescendo of a systolic murmur or the timing of the S_3 gallop, which are the most influential factors in classification. For generative tasks, the models can indicate which physiological features are prioritized for synthesizing pathological sounds. This approach bridges the latent space of the model with the diagnostic reasoning of clinicians.

10.3. Multimodal integration for creating comprehensive cardiac assessment tools

Heart function is best understood by synthesizing information from multiple complementary data streams. Future research should focus on developing unified generative frameworks. These frameworks would simultaneously process PCG signals with other modalities, such as ECG, echocardiography (ultrasound), and clinical metadata, including age

and blood pressure (Li, Ding, Huang, Zhuang, & Grau, 2023), to generate synchronized multimodal outputs. For example, they can create a synthetic PCG trace that corresponds to a given ECG abnormality or translate between modalities to enrich the data. Furthermore, the integration of large language models can bridge the gap between signal analysis and clinical workflows. These models can generate interpretive narrative reports from PCG signals, directly aiding in documentation and diagnosis (Goh et al., 2024).

10.4. Personalized generative models to move from population-level to patient-specific tools

Cardiovascular physiology varies among individuals based on factors such as age, body mass index, sex, and pre-existing conditions (Srikanthan, Horwich, Press, Gornbein, & Watson, 2021). Future research should focus on developing generative models that can adapt to the individual characteristics of patients. This could involve creating patient-specific latent spaces or using conditional generation to produce synthetic data that reflect the unique physiological baseline of each patient. Such personalized models can simulate the acoustic effects of disease progression and therapeutic interventions in specific patients. This approach paves the way for highly tailored diagnostic and prognostic applications.

10.5. Advancing real-time generative applications in resource-limited settings

Generative AI shows promise for heart sound synthesis; however, its real-world use in resource-limited settings is hindered by high computational costs, particularly for diffusion models. However, heart sounds are low-bandwidth and short, which means that clinically useful PCG generation may require less complex models. Because models are trained centrally, efficiency is crucial for on-device inference using tools such as portable stethoscopes or wearables. Efficient generation enables real-time denoising, augmentation, and enhancement while reducing latency, protecting privacy, and improving the robustness of offline systems. Thus, future work should focus on lightweight diffusion variants, model distillation, and task-specific designs for short PCG segments to enable practical, scalable clinical solutions (Moslemi, Briskina, Dang, & Li, 2024; Xie & Du, 2022).

10.6. Establishing standardized validation for the clinical translation of generative models

Moving beyond performance metrics, such as the signal-to-noise ratio or Fréchet inception distance, future studies must develop evaluation protocols that assess clinical utility, including structured validation through blinded clinical trials (Betzalet, Penso, & Fetaya, 2024). In these trials, cardiologists can diagnose conditions using both real and synthetic datasets. Future studies should also develop metrics to quantify the diagnostic confidence and accuracy. Concurrently, research must address the ethical and regulatory considerations associated with the use of synthetic medical data. Standards for its generation, curation, and deployment are required to ensure patient safety and data protection.

11. Conclusion

This study systematically charts the transformative impact of generative artificial intelligence on heart sound analysis. We have detailed how foundational models – generative adversarial networks (GANs), variational autoencoders (VAEs), diffusion models, and transformer-based architectures – address perennial challenges that have hindered automated cardiac auscultation. The discussion reveals a clear paradigm shift: generative AI is transitioning from a state of data

scarcity to a state of data abundance through the generation of high-fidelity synthetic signals. This is not merely a technical achievement; it is a practical solution that enables robust training of diagnostic models for rare pathologies, mitigates severe class imbalances, and provides a pathway for privacy-preserving data sharing. Furthermore, the application of these models to signal denoising and enhancement directly addresses the problem of poor recording quality, promising to deliver cleaner and more interpretable signals to both clinicians and algorithms in the future.

However, integrating these technologies into clinical practice is challenging. This discussion highlights a significant disparity between statistical realism and physiological precision. A synthetic heart sound may be indistinguishable from a real one from a deep model based on learned features; however, it must also be physiologically plausible and representative of a specific, quantifiable pathophysiological state that cardiologists can rely on. The “black box” nature of many complex models remains a significant barrier to adoption, as clinical decision-making requires model transparency and interpretability. Future progress hinges on the development of explainable AI techniques that can highlight the segments of a sound, such as the timing or shape of a murmur, that contribute to a diagnostic output. Finally, the computational burden of state-of-the-art models, such as diffusion models, currently limits their utility in real-time point-of-care applications, necessitating research into more efficient and lightweight architectures.

In conclusion, generative AI has emerged as a cornerstone technology in the field of cardiovascular diagnostics. By providing powerful tools for data augmentation, signal enhancement, and unsupervised anomaly detection, deep learning paves the way for more accurate, accessible, and scalable heart sound analyses. The journey from the algorithm to the bedside now depends on concerted interdisciplinary efforts. Cardiologists, computer scientists, and biomedical engineers must collaborate to ensure that future generative models are grounded in cardiovascular physiology, are transparent in their reasoning, and are rigorously validated in clinical trials. By bridging the gap between algorithmic innovation and clinical necessity, generative AI has the profound potential to democratize expert-level cardiac evaluation, ultimately leading to earlier diagnosis, more personalized treatment, and improved patient outcomes.

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

No data was used for the research described in the article.

References

- Abbott, Leigh, Marocchi, Milan, Fynn, Matthew, Rong, Yue, & Nordholm, Sven (2026). Generative deep learning and signal processing for data augmentation of cardiac auscultation signals: Improving model robustness using synthetic audio. *Biomedical Signal Processing and Control*, 112, Article 108469. <http://dx.doi.org/10.1016/j.bspc.2025.108469>.
- Abdul, Zrar Kh., & Al-Talabani, Abdulbasit K. (2022). Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, 10, 122136–122158. <http://dx.doi.org/10.1109/ACCESS.2022.3223444>.
- Adiban, Mohammad, BabaAli, Bagher, & Shehnepoor, Saeedreza (2019). Statistical feature embedding for heart sound classification. *Journal of Electrical Engineering*, 70, 259–272.
- Agostinelli, Andrea, Denk, Timo I., Borsos, Zsolt, Engel, Jesse, Verzett, Mauro, Cailon, Antoine, Huang, Qingqing, Jansen, Aren, Roberts, Adam, Tagliasacchi, Marco, Sharifi, Matthew, Zeghidour, Neil, & Frank, Christian Havnø (2023). MusicLM: Generating music from text. ArXiv, [abs/2301.11325](https://arxiv.org/abs/2301.11325).
- Alkhodari, Mohamad, Azman, Syafiq Kamarul, Hadjileontiadis, Leontios J., & Khandoker, Ahsan Habib (2022). Ensemble transformer-based neural networks detect heart murmur in phonocardiogram recordings. In *Computing in cardiology (cinC)*: vol. 498, (pp. 1–4). <http://dx.doi.org/10.22489/CinC.2022.035>.

- Almasi, Ali, Shamsollahi, Mohammad Bagher, & Senhadji, Lotfi (2011). A dynamical model for generating synthetic phonocardiogram signals. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 5686–5689). <http://dx.doi.org/10.1109/IEMBS.2011.6091376>.
- Banerjee, Rohan, & Ghose, Avik (2020). A semi-supervised approach for identifying abnormal heart sounds using variational autoencoder. In *IEEE international conference on acoustics, speech and signal processing* (pp. 1249–1253).
- Bentley, Peter, Nordehn, Glenn, Coimbra, Miguel, Mannor, Shie, & Getz, Rita (2011). Classifying heart sounds challenge (PASCAL dataset). <http://www.peterjbentley.com/heartchallenge/>.
- Betzalel, Eyal, Penso, Coby, & Fetaya, Ethan (2024). Evaluation metrics for generative models: An empirical study. *Machine Learning and Knowledge Extraction*, 6(3), 1531–1544. <http://dx.doi.org/10.3390/make6030073>.
- Bi, Xiuli, Tang, Shizhan, Xiao, Bin, Li, Weisheng, Gao, Xinbo, & Liò, Pietro (2025). A systematic review of heart sound detection algorithms: Experimental results and insights. *IEEE Transactions on Instrumentation and Measurement*, 74, 1–16.
- Buchh, Irshad A. (2024). Benchmarking variational autoencoders for dimensionality reduction and application to medical heartbeats data. In *IEEE international conference on technology management, operations and decisions* (pp. 01–07).
- Chakraborty, Shaunak, Kochhar, Prishita, Patil, Shrutika A., Kotecha, Ketan V., Gite, Shilpa, Selvachandran, Ganeshsree, & Das, Swagatam (2025). Generative adversarial network augmented data for improved heart sound abnormality detection.. *Computers in Biology and Medicine*, 195, Article 110623.
- Chen, Junxin, Guo, Zhihuan, Xu, Xu, Jeon, Gwanggil, & Camacho, David (2024). Artificial intelligence for heart sound classification: A review. *Expert Systems*, 41.
- Chen, Yunhao, Yan, Zihui, & Zhu, Yunjie (2024). A comprehensive survey for generative data augmentation. *Neurocomputing*, 600, Article 128167.
- Chen, Nanxin, Zhang, Yu, Zen, Heiga, Weiss, Ron J., Norouzi, Mohammad, & Chan, William (2020). WaveGrad: Estimating gradients for waveform generation. *ArXiv*, abs/2009.00713.
- Colt, Robert-George, Várady, Csongor-Huba, Volpi, Riccardo, & Malagò, Luigi (2020). Automatic feature extraction for heartbeat anomaly detection. In *European conference on machine learning and principles and practice of knowledge discovery in databases*.
- Dehak, Najim, Kenny, Patrick, Dehak, Réda, Dumouchel, Pierre, & Ouellet, Pierre (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19, 788–798.
- Delyon, Bernard, & Juditsky, Anatoli (1995). *Estimating Wavelet Coefficients* (pp. 151–168). New York, NY: Springer New York, http://dx.doi.org/10.1007/978-1-4612-2544-7_10.
- Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, & Toutanova, Kristina (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American chapter of the association for computational linguistics*.
- Dissanayake, Theekshana, Fernando, Tharindu, Denman, Simon, Sridharan, Sridha, & Fookes, Clinton (2023). Generalized generative deep learning models for biosignal synthesis and modality transfer. *IEEE Journal of Biomedical and Health Informatics*, 27(2), 968–979.
- El-Dahshan, El-Sayed A., & Bassiouni, Mahmoud M. (2020). Intelligent methodologies for cardiac sound signals analysis and characterization in cepstrum and time-scale domains. *Computational Intelligence*, 36, 427–458.
- Ergen, B., & Tatar, Y. (2005). Characterization of phonocardiogram signals using bispectral estimation. In *Proceedings of the eighth international symposium on signal processing and its applications*, 2005.: vol. 1, (pp. 203–206). <http://dx.doi.org/10.1109/ISSPA.2005.1580231>.
- Fakhry, Mahmoud, & Brery, Abeer FathAllah (2022). Comparison of window shapes and lengths in short-time feature extraction for classification of heart sound signals. *International Journal of Electrical and Computer Engineering (IJECE)*.
- Fakhry, Mahmoud, & Gallardo-Antolín, Ascensión (2024a). Elastic net regularization and gabor dictionary for classification of heart sound signals using deep learning. *Engineering Applications of Artificial Intelligence*, 127, Article 107406. <http://dx.doi.org/10.1016/j.engappai.2023.107406>.
- Fakhry, Mahmoud, & Gallardo-Antolín, Ascensión (2024b). Analysis of systolic murmurs in heart sounds using multiresolution complex gabor dictionary. In *2024 international conference on computer and applications* (pp. 1–5). <http://dx.doi.org/10.1109/ICCA62237.2024.10927981>.
- Fakhry, Mahmoud, & Gallardo-Antolín, Ascensión (2025). A hybrid neural network with multistage feature fusion for detecting heart failure and murmurs from time-frequency representations of phonocardiograms. *Systems and Soft Computing*, 7, Article 200391. <http://dx.doi.org/10.1016/j.sasc.2025.200391>.
- Ferdoushi, Munia, Paul, Mala, & Fattah, Shaikh Anowarul (2019). A spectral centroid based analysis of heart sounds for disease detection using machine learning. In *2019 IEEE international WIE conference on electrical and computer engineering (WIECON-ECE)* (pp. 1–6). <http://dx.doi.org/10.1109/WIECON-ECE48653.2019.9019975>.
- Gavrovskaa, Ana, Slavkovic, Marijeta, Reljin, Irini, & Reljin, Branimir (2013). Application of wavelet and EMD-based denoising to phonocardiograms. In *International symposium on signals, circuits and systems ISSCS*.
- Gawande, Namrata, Goyal, Dinesh, & Sankhla, Kriti (2025). Generative adversarial network based improved transfer learning models for chronic heart failure detection. *Ingénierie Des Systèmes D Information*.
- Goh, Ethan, Gallo, Robert, Hom, Jason, Strong, Eric, Weng, Yingjie, Kerman, Han-nah, Cool, Joséphine A., Kanjee, Zahir, Parsons, Andrew S., Ahuja, Neera, Horvitz, Eric, Yang, Daniel, Milstein, Arnold, Olson, Andrew P. J., Rodman, Adam, & Chen, Jonathan H. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10), <http://dx.doi.org/10.1001/jamanetworkopen.2024.40969>, e2440969–e2440969.
- González-rodríguez, Cristóbal, Alonso-Arévalo, Miguel A., & García-Canseco, Eloísa (2023). Robust denoising of phonocardiogram signals using time-frequency analysis and U-nets. *IEEE Access*, 11, 52466–52479.
- Goodfellow, Ian J., Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron C., & Bengio, Yoshua (2014). Generative adversarial networks. *Communications of the ACM*, 63, 139–144.
- Han, Jiho, & Shaout, Adnan (2025). ENACT-heart - ENsemble-based assessment using CNN and transformer on heart sounds. *ArXiv*, abs/2502.16914.
- Han, Xu, Zhang, Zhengyan, Ding, Ning, Gu, Yuxian, Liu, Xiao, Huo, Yuqi, Qiu, Jiezhong, Zhang, Liang, Han, Wentao, Huang, Minlie, Jin, Qin, Lan, Yanyan, Liu, Yang, Liu, Zhiyuan, Lu, Zhiwu, Qiu, Xipeng, Song, Ruihua, Tang, Jie, rong Wen, Ji, Zhu, Jun (2021). Pre-trained models: Past, present and future. *ArXiv*, abs/2106.07139.
- Hassanuzzaman, Md, Hasan, Nurul Akhtar, Abdullah-Al-Mamun, Mohammad, Alkhodari, Mohanad, Ahmed, Khawza Iftekhar Uddin, Khandoker, Ahsan Habib, & Mostafa, Raqibul (2023). Recognition of pediatric congenital heart diseases by using phonocardiogram signals and transformer-based neural networks. In *45th annual international conference of the IEEE engineering in medicine & biology society* (pp. 1–4).
- Hu, Jianqiang, Chen, Lin, Yang, Miao, Shen, Shigen, & Gao, Xiao-Zhi (2025). PSBD-EWT-EGAN: Heart sound denoising using PSBD-EWT and enhancement generative adversarial network. *Computer Science Information Systems*, 22, 239–257.
- Jamal, Arshad, Ramasamy, R. Kanesaraj, & Abdullah, Junaiddi Bin (2025). Generative AI respiratory and cardiac sound separation using variational autoencoders (VAEs). In *International conference on sustainable computing and green technologies*.
- Jamil, Sonain, & Roy, Anisha (2023). An efficient and robust phonocardiography (PCG)-based valvular heart diseases (VHD) detection framework using vision transformer (ViT). *Computers in Biology and Medicine*, 158, Article 106734.
- Jamshidi, Ainaz, Arif, Muhammad, Kalhor, Sabir Ali, & Gelbukh, Alexander F. (2024). Synthetic time series data generation for healthcare applications: A PCG case study. In *59th annual conference on information sciences and systems* (pp. 1–6).
- Judge, Richard D. (2015). Heart sound and murmur library. <http://open.umich.edu/education/med/resources/heart-sound-murmur-library/2015>.
- Kingma, Diederik P., & Welling, Max (2019). An introduction to variational autoencoders. *Foundations and Trends Machine Learning*, 12, 307–392.
- Lee, Shuenn-Yuh, Su, Po-Han, Hsieh, Yi-Ting, Huang, Sheng-Hsin, Lee, I-Pei, & Chen, Ju-Yi (2023). Intelligent stethoscope system and diagnosis platform with synchronized heart sound and electrocardiogram signals. *IEEE Access*, 11, 47420–47431. <http://dx.doi.org/10.1109/ACCESS.2023.3275015>.
- Lei, Yang (2004). Short-time fourier transform analysis of the phonocardiogram signal. *Journal of Chongqing University*.
- Li, Lei, Ding, Wangbin, Huang, Liqin, Zhuang, Xiahai, & Grau, Vicente (2023). Multi-modality cardiac image computing: A survey. *Medical Image Analysis*, 88, Article 102869. <http://dx.doi.org/10.1016/j.media.2023.102869>.
- Li, Liuying, Huang, Min, Dao, Lin, Feng, Xixi, Liu, Yifeng, Wei, Changyou, Liu, Fang-fang, Zhang, Jing, & Xu, Fan (2024). Construction and validation of a method for automated time label segmentation of heart sounds. *Frontiers in Artificial Intelligence*, 6.
- Li, Shengchen, & Tian, Keyu (2021). Unsupervised phonocardiogram analysis with distribution density based variational auto-encoders. *Frontiers in Medicine*, 8.
- Li, Shengchen, Tian, Keyu, & Wang, Rui (2021). Unsupervised heart abnormality detection based on phonocardiogram analysis with beta variational auto-encoders. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8353–8357.
- Liang, Mingfeng, Hu, Jian-Qiang, Zhou, Xiaobao, & Xiao, Sikai (2022). A GAN-based heart sound denoising model. In *12th international conference on information technology in medicine and education* (pp. 666–670).
- Libby, Peter, Bonow, Robert O., Mann, Douglas L., Tomaselli, Gordon F., Bhatt, Deepak L., Solomon, Scott D., & Braunwald, Eugene (2021). *Braunwald's heart disease: A textbook of cardiovascular medicine: 2 vol set: vol. 1–2*, (12th ed.). Elsevier.
- Liu, Zhihai, Liu, Wen, Gu, Zheng, & Yang, Fengkai (2024). Efficient anomaly detection algorithm for heart sound signal. *IEEE Access*, 12, 139225–139236.
- Liu, Xu, Lv, Chengcong, Cao, Linchun, & Guo, Xingming (2025). Detection of coronary artery disease using a triplet network and hybrid loss function on heart sound signal. *Biomedical Signal Processing and Control*, 104, Article 107601. <http://dx.doi.org/10.1016/j.bspc.2025.107601>.
- Liu, Chengyu, Springer, David B., Li, Qiao, Moody, Benjamin, Juan, Ricardo Abad, Chorro, Francisco Javier, Castells, Francisco, Roig, José Millet, Silva, Ikaro, Johnson, Alistair E. W., Syed, Zeeshan, Schmidt, Samuel Emil, Papadaniil, Chrysa D., Hadjileontiadis, Leontios J., Naseri, Hosein, Moukadem, Ali, Dieterlen, Alain, Brandt, Christian, Tang, Hong, Clifford, Gari D. (2016). An open access database for the evaluation of heart sound algorithms. *Physiological Measurement*, 37, 2181–2213.

- Majhi, Babita, & Kashyap, Aarti (2024). Explainable AI-driven machine learning for heart disease detection using ECG signal. *Applied Soft Computing*, 167, Article 112225. <http://dx.doi.org/10.1016/j.asoc.2024.112225>.
- Moslemi, Amir, Briskina, Anna, Dang, Zubeka, & Li, Jason (2024). A survey on knowledge distillation: Recent advancements. *Machine Learning with Applications*, 18, Article 100605. <http://dx.doi.org/10.1016/j.mlwa.2024.100605>.
- Narváez, Pedro, & Percybrooks, Winston S. (2020). Synthesis of normal heart sounds using generative adversarial networks and empirical wavelet transform. *Applied Sciences*.
- Nguyen, Minh Tuan, Lin, Wei Wen, & Huang, Jin Huang (2022). Heart sound classification using deep learning techniques based on log-mel spectrogram. *Circuits, Systems, and Signal Processing*, 42, 344–360.
- Oliveira, Jorge, Renna, Francesco, Costa, Paulo Dias, Nogueira, Marcelo, Oliveira, Cristina, Ferreira, Carlos, Jorge, Alípio, Mattos, Sandra, Hatem, Thamine, Tavares, Thiago, Elola, Andoni, Rad, Ali Bahrami, Sameni, Reza, Clifford, Gari D., & Coimbra, Miguel T. (2022). The CirCor DigiScope dataset: From murmur detection to murmur classification. *IEEE Journal of Biomedical and Health Informatics*, 26(6), 2524–2535. <http://dx.doi.org/10.1109/JBHI.2021.3137048>.
- Ompokov, V. D., & Boronoev, V. V. (2019). Mode decomposition and the Hilbert-Huang transform. In *2019 Russian open conference on radio wave propagation: vol. 1*, (pp. 222–223). <http://dx.doi.org/10.1109/RWP.2019.8810217>.
- Partovi, Elaheh, Babic, Ankica, & Gharehbaghi, Arash (2024). A review on deep learning methods for heart sound signal analysis. *Frontiers in Artificial Intelligence*, 7.
- Ramesh, Aditya, Pavlov, Mikhail, Goh, Gabriel, Gray, Scott, Voss, Chelsea, Radford, Alec, Chen, Mark, & Sutskever, Ilya (2021). Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092.
- Rangayyan, R. M., & Lehner, R. J. (1987). Phonocardiogram signal analysis: a review.. *Critical Reviews in Biomedical Engineering*, 15 3, 211–236.
- Rayavarapu, Swarajya Madhuri, Prasanthi, Tammineni Shanmukha, Kumar, Gottapu Santosh, Rao, G. Sasibhushana, & Prashanti, Gottapu (2023). A generative Model For Deep fake augmentation Of phonocardiogram and electrocardiogram SIGNALS USING LSGAN AND CYCLE GAN. *Informatyka, Automatyka, Pomiary W Gospodarce I Ochronie Środowiska*.
- Rayavarapu, Swarajya Madhuri, Prasanthi, Tammineni Shanmukha, Kumar, Gottapu Santosh, Rao, G. Sasibhushana, & Singham, Aruna (2023). Employing generative networks for synthetic phonocardiogram and electrocardiogram signal creation: A privacy-ensured approach to data augmentation in heart diagnostics. *Ingénierie Des Systèmes D Information*.
- Ricke, Anthony D., Povinelli, Richard J., & Johnson, Michael T. (2005). Automatic segmentation of heart sound signals using hidden Markov models. *Computers in Cardiology*, 953–956.
- Rombach, Robin, Blattmann, A., Lorenz, Dominik, Esser, Patrick, & Ommer, Björn (2021). High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10674–10685).
- Rong, Y., Fynn, M., & Nordholm, S. (2004). In A. Al-Jumaily, P. Crippa, A. Mansour, & C. Turchetti (Eds.), *Non-Invasive Health Systems Based on Advanced Biomedical Signal and Image Processing*. Taylor & Francis.
- Sel, Kaan, Osman, Deen, Zare, Fatemeh, Shahrabab, Sina Masoumi, Brattain, Laura J., Hahn, J-O, Inan, Omer T., Mukkamala, Ramakrishna, Palmer, Jeffrey, Paydarfar, David, Pettigrew, Roderic I., Quyyumi, Arshed A., Telfer, Brian A., & Jafari, Roozbeh (2024). Building digital twins for cardiovascular health: From principles to clinical impact. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease*, 13.
- Šeman, Michael, Stephens, Andrew F., Kaye, David M, Gregory, Shaun D., & Stub, Dion (2024). Computational modelling of valvular heart disease: haemodynamic insights and clinical implications. *Frontiers in Bioengineering and Biotechnology*, 12.
- Shariat Panah, Davoud, Hines, Andrew, McKeever, Joseph A., & McKeever, Susan (2022). An audio processing pipeline for acquiring diagnostic quality heart sounds via mobile phone. *Computers in Biology and Medicine*, 145, Article 105415. <http://dx.doi.org/10.1016/j.combiomed.2022.105415>.
- Sohl-Dickstein, Jascha Narain, Weiss, Eric A., Maheswaranathan, Niru, & Ganguli, Surya (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *ArXiv*, abs/1503.03585.
- Srikanthan, Preethi, Horwich, Tamara B, Press, Marcella A Calfon, Gornbein, Jeffrey, & Watson, K. (2021). Sex differences in the association of body composition and cardiovascular mortality. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease*, 10.
- Takezaki, Shumpei, & Kishida, Kazuya (2022). Data augmentation and the improvement of the performance of convolutional neural networks for heart sound classification. *IAENG International Journal of Computer Science*, 49(4), 1033–1040.
- Tan, Yang, Zhang, Haojie, Chang, Yi, Wu, Qingrong, Qian, Kun, Hu, Bin, Schuller, Björn W., & Yamamoto, Yoshiharu (2025). A latent diffusion model for heart sound synthesis. In *4th international symposium on computer applications and information technology* (pp. 1314–1319).
- The public heart sound dataset of the university of washington. (2022). <https://depts.washington.edu/physdx/heart/demo.html>.
- Unnikrishnan, Balagopal, Singh, Pranshu Ranjan, Yang, Xulei, & Chua, Matthew Chin Heng (2020). Semi-supervised and unsupervised methods for heart sounds classification in restricted data environments. *ArXiv*, abs/2006.02610.
- Vimalajeewa, Dixon, Lee, Chihoon, & Vidakovic, Brani (2025). Multiscale analysis of heart sound signals in the wavelet domain for heart murmur detection. *Scientific Reports*, 15.
- Vinay, N. A., Vidyasagar, K. N., Rohith, S., Pruthviraja, Dayananda, Supreeth, S., & Bharathi, S. H. (2024). An RNN-BiLSTM-based multi decision GAN approach for the recognition of cardiovascular disease (CVD) from heart beat sound: A feature optimization process. *IEEE Access*, 12, 65482–65502.
- Wang, Suiyan, Hu, Junhui, Du, Yanwei, Yuan, Xiaoming, Xie, Zhongliang, & Liang, Pengfei (2025). WCFormer: An interpretable deep learning framework for heart sound signal analysis and automated diagnosis of cardiovascular diseases. *Expert Systems with Applications*, 276, Article 127238.
- Xie, Pengtao, & Du, Xuefeng (2022). Performance-aware mutual knowledge distillation for improving neural architecture search. In *2022 IEEE/CVF conference on computer vision and pattern recognition* (pp. 11912–11922). <http://dx.doi.org/10.1109/CVPR52688.2022.01162>.
- Yang, Yang, Guo, Xingming, Wang, Hui, & Zheng, Yi-Neng (2021). Deep learning-based heart sound analysis for left ventricular diastolic dysfunction diagnosis. *Diagnostics*, 11.
- Yang, Dong, Lin, Yi-Chun, Wei, Jianwen, Lin, Xiongwei, Zhao, Xiaobo, Yao, Yingbang, Tao, Tao, Liang, Bo, & Lu, Sheng-Guo (2023). Assisting heart valve diseases diagnosis via transformer-based classification of heart sound signals. *Electronics*.
- Yaseen, Son, Guiyoung, & Kwon, Soonil (2018). Classification of heart sound signal using multiple features. *Applied Sciences*.
- Yoganathan, Ajit P., Gupta, Ramesh, Corcoran, William Harrison, Udawadia, Firdaus E., Sarma, Radha J., & Bing, Richard J. (1976). Use of the fast Fourier transform in the frequency analysis of the second heart sound in normal man. *Medical and Biological Engineering*, 14, 455–460.
- Zhang, Miao, Li, Min, Guo, Liang, & Liu, Jianya (2023). A low-cost AI-empowered stethoscope and a lightweight model for detecting cardiac and respiratory diseases from lung and heart auscultation sounds. *Sensors (Basel, Switzerland)*, 23.
- Zhao, Qinghao, Geng, Shijia, Wang, Boya, Sun, Yutong, Nie, Wenchang, Bai, Baochen, Yu, Chao, Zhang, Feng, Tang, Gongzheng, Zhang, Deyun, Zhou, Yuxi, Liu, Jian, & Hong, Shenda (2024). Deep learning in heart sound analysis: From techniques to clinical applications. *Health Data Science*, 4.
- Zhao, Wenhao, Ma, Hongwen, Jin, Ni, Zheng, Yineng, & Guo, Xingming (2025). Detection of coronary heart disease based on heart sound and hybrid vision transformer. *Applied Acoustics*, 230, Article 110420. <http://dx.doi.org/10.1016/j.apacoust.2024.110420>.
- Zhou, George, Chen, Yunchan, & Chien, Candace (2021). On the analysis of data augmentation methods for spectral imaged based heart sound classification using convolutional neural networks. *BMC Medical Informatics and Decision Making*, 22.