RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt2567

# Optimized Vision Transformer Architecture for Cardiac Auscultation Classification using GAN augmented MFCC representations

**Divya Lalita Sri Jalligampala [a,*], Gangadhara Rao Kancharla [a], R.V.S. Lalitha [b]**

[a] Department of Computer Science Engineering, University College of Sciences, Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, Andhra Pradesh, 522510, India.

[b] Department of Computer Science Engineering, Aditya University, Suramplaem, Andhra Pradesh, 533447, India.

* Corresponding Author Email: lalitha517@gmail.com

DOI: https://doi.org/10.54392/irjmt2567

Received: 17-05-2025; Revised: 07-10-2025; Accepted: 23-10-2025; Published: 04-11-2025

**Abstract:** Heart auscultation is a key diagnostic tool for detecting cardiac abnormalities; however, human interpretation is subjective and prone to error. Classic machine learning algorithms like LSTMs and BiLSTMs have been employed for computer-aided heart sound classification but face challenges with handling acoustic variation, data sparsity, and long-range correlations in spectrograms. Solo Vision Transformers (ViT's) improve feature extraction but require large datasets to function best. This article introduces a hybrid model combining a Generative Adversarial Network (GAN) and a Vision Transformer (ViT) to address these issues, applying GAN-based data augmentation to enhance training diversity and leveraging ViT's self-attention mechanism to interpret spectrograms better. The data, accessed through the iStethoscope Pro app and clinical testing with DigiScope, comprised normal, murmur, and artifact classes. Preprocessing included silent cutting, resampling, and extraction of MFCCs, spectral contrast, chroma features, and RMSE. The proposed GAN+ViT model was compared to BiLSTM, LSTM, and standalone ViT. The performance showed that GAN+ViT outperformed all baseline models with 90% accuracy, 0.90 F1-score, 0.91 precision, and 0.89 recall, and AUC-ROC values of 0.92 for artifacts, 0.93 for murmurs, and 0.91 for normal sounds. On the other hand, BiLSTM (85%), LSTM (83%), and ViT (80%) were poor in their performance, particularly in discriminating between murmurs and normal sounds. The improved classification power of the hybrid model is due to complementary data augmentation and attention-based feature learning, thereby reducing misclassifications. This research recommends that GAN+ViT is a viable method for automated analysis of cardiac sounds, with high accuracy and generalizability for clinical applications. Future research could explore multimodal integration with ECG data and employ explainable AI methods to enhance diagnostic consistency.

**Keywords:** Heart sound classification, Generative Adversarial Network (GAN), Vision Transformer (ViT), Data augmentation, Mel-Frequency Cepstral Coefficients (MFCCs)

## 1. Introduction

Cardiovascular diseases (CVDs) have become more hazardous due to heightened industrialization, urbanization, and globalization, leading to a growing global death rate. Cardiovascular diseases (CVDs) include a variety of illnesses that impact the blood vessels and heart. These consist of deep vein thrombosis, cerebrovascular illness, and pulmonary embolism, coronary artery disease, peripheral artery disease, rheumatic heart disease, and congenital heart defects, among others. In 2016, cardiovascular diseases (CVDs) resulted in 17.9 million deaths, accounting for 31% of total worldwide mortality. Cardiovascular disorders were the predominant cause of mortality for 85% of this population [1]. Individuals in low- and middle-income nations bear an inequitable amount of the financial burden linked to cardiovascular diseases (CVDs), making early identification and intervention essential for reducing death rates. For almost 180 years, medical professionals have relied on cardiac auscultation, the use of a stethoscope to listen to the heart to identify cardiovascular problems, because of its simplicity, necessity, and effectiveness [2]. The early identification of cardiovascular diseases is crucial due to their non-invasive characteristics and precise representation of the circulatory system and cardiac mechanics.

The measurement of heart sounds, a type of physiological signal, is known as a phonocardiogram (PCG). The diastole and systole of the heart generate it,

providing physiological information about the atria, ventricles, and the functioning of the principal arteries [3]. The S1 and S2 are two of the basic heart sounds (FHSs). S1 usually occurs when the closed tricuspid and mitral valves abruptly reach their elastic limits at the onset of ventricular contraction, due to the rapid buildup of intraventricular stress. S2 is heard at the onset of diastole, as the pulmonic and aortic valves close. Another swooshing sound made by the rapid blood of your heart is called a heart murmur, distinct from the heartbeat sounds. To evaluate the possible hazard of a heart murmur or its association with cardiac diseases, physicians auscultate the heart and examine many characteristics, including intensity, location, frequency, timing, and variations in sound. Consequently, cardiac auscultation for heart murmurs may exclude individuals without heart disease or abnormalities [4]. The S1, systole, diastole, and S2 phases must be correctly identified and the FHSs must be segmented. The four stages of phonocardiogram (PCG) recording S1, systole, diastole, and S2 are shown in Figure 1, together with a real-time electrocardiogram (ECG) recording. When the QRS complex of an electrocardiogram (ECG) corresponds to the generation of cardiac sounds, the S1 and S2 sites may be identified. During subsequent diagnostic evaluations, FHSs provide essential preliminary indications for the assessment of cardiac illness.

When diagnosing heart conditions, extracting traits from every FHS division is necessary for quantitative analysis. However, not all frequency ranges are audible to the human ear, and cardiac auscultation requires considerable clinical expertise and skill. In this context, automated categorization of cardiac sounds has garnered more interest in recent decades.

A long-standing goal for researchers has been to achieve excellent precision in automatic heart sound classification methods. The two main approaches for classifying cardiac sound signals are deep learning and conventional machine learning. Deep learning (DL) methods for cardiac sound categorization are becoming increasingly popular due to recent advancements in medical big data and AI technologies [5]. DL algorithms have gained widespread use in the area of heart murmur studies [6-9]. Their ability to effectively process a large number of features and to solve feature selection issues enables them to achieve accuracy comparable to that of seasoned cardiologists [10]. However, interpretability issues with DL models, their complexity, and black-box characteristics have been raised, rendering it challenging to understand how they work. This could therefore render it challenging, particularly for medical professionals, to intervene in time [11]. This thus implies that, despite the huge progress. Early detection of cardiovascular disease (CVD) remains an area in need of improvement; more efficient, more reliable means are needed. This essay proposes a novel deep learning method for heartbeat classification that combines Vision Transformers (ViTs) and Generative Adversarial Networks (GANs) to make precise predictions of heart disease. Unlike common methods that use raw audio signals or manually designed features, this work combines multimodal data, PCG audio signals, and text annotations into a single learning pipeline. This work leverages Generative Adversarial Networks (GANs) to generate high-quality synthesized heart-sound samples to address the pervasive problem of class imbalance in clinical data. The generated samples are blended with real data for enhanced model fairness and performance. Concurrently, a classification model based on ViT with self-attention is used to learn rich temporal features from heart sounds to improve classification accuracy. Robust augmentation techniques, such as Pitch shifting and time stretching, and background noise, are utilized to enhance generalizability. Bayesian optimization is used to optimize the overall system, and stratified cross-validation is used to validate it, providing an unbiased and reliable evaluation process. Since heart sound signals are highly volatile and complex, our multimodal GAN-ViT model performs much better than previous models.

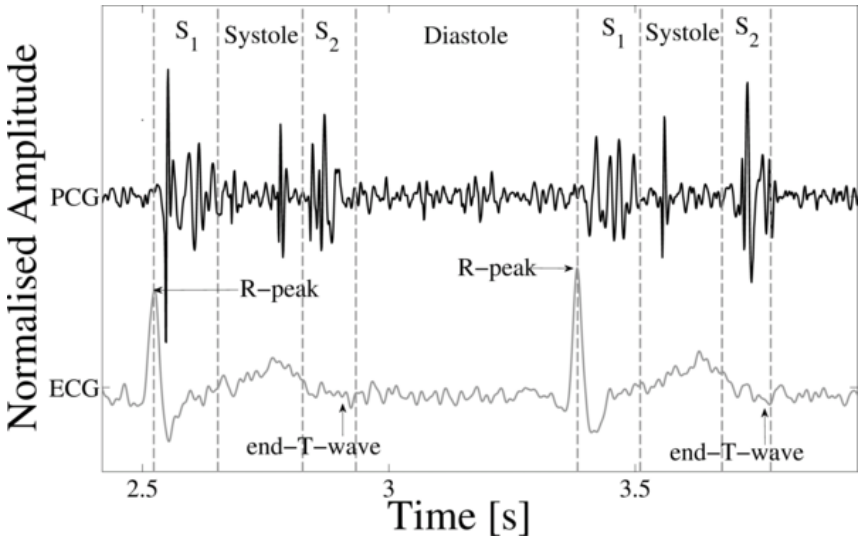


**Figure 1.** Heart Sound signals

This study's main findings are:

- Development of a cutting-edge deep learning architecture combining phonocardiogram (PCG) audio data with relevant textual metadata to perform precise heart sound classification.
- For utilizing Generative Adversarial Networks (GAN's) to synthesize high-quality heart sound samples, efficiently solving class imbalance and enhancing model fairness.
- To implement A structure based on the Vision Transformer (ViT) with self-awareness capabilities to record intricate spatial and structural patterns in heart sound signals.
- To improve model stability and applicability by using sophisticated data supplementation methods, including noise, pitch-shifting, and time-stretching-related additions.
- To optimize model performance through Bayesian hyperparameter tuning and stratified cross-validation, ensuring reliable, unbiased evaluation and improved diagnostic accuracy.

The following are the parts: An overview of the pertinent research is provided in Section 2, with identification of research gaps. Section 3 will provide background information on the proposed models. The approach for data collection and preprocessing, model designs, and assessment techniques is described in Section 4. Results are presented, performance measures are discussed, and the efficacy of the suggested models is compared in Section 5. Section 6 concludes the study, summarizing the main conclusions and offering recommendations for further study.

## 2. Literature Review

In several earlier studies, the heart sound, also known as Phonocardiograms (PCGs), has been studied and investigated as a potential marker for CVD diagnosis and personal identification. The automated, computer-assisted processing of PCG data is inherently non-invasive and low-cost [12]. Jia *et al.* [13] built fuzzy brain networks using features such as the discrete wavelet transform (DWT) and “Dispersion Entropy (DisEn)” to differentiate between normal and pathological sounds. The study's foundation was real data acquired by the researchers. A novel model [14] for heart sound classification was proposed using autocorrelation features and diffusion maps, eliminating the need for segmentation. An SVM classifier is trained using the framework's discrete wavelet decomposition of sub-band envelopes to extract autocorrelation characteristics. After that, it merges all the characteristics into a single representation using diffusion maps. Two publicly accessible datasets were utilised to evaluate the effectiveness of the PASCAL Classifying Heart Sounds Challenge framework [15], which used an ensemble neural network method. This research used the PhysioNet database. and an accuracy of 91.5% was achieved. Additionally, [16] achieved 94.24% classification accuracy using deep learning and the AdaBoost and CNN algorithms. A sample of 2575 normal and 665 aberrant noises was employed in this investigation. Zhang *et al.* [17] sought to categorize cardiac sounds using spectral images and a regression-based method on the PASCAL dataset with the SVM algorithm. The subsequent section of the research focused on the PhysioNet dataset, which was used for feature extraction via the tensor decomposition approach. The dataset was categorized with the SVM method. Another study examined a genetic algorithm that combined neural networks [18]. To improve the neural network's performance, they suggested raising its starting weights. They achieved 97% sensitivity, 92% specificity, and 93.85% accuracy with this approach. In a separate study [19], the authors distinguished between normal and abnormal noises before developing an autonomous method for detecting sound types using electrical circuits and sensors. An accuracy of about 97% was achieved. A deep convolutional neural network is used here. The heart sounds in this research were converted to sonograms and fed into deep learning algorithms. Using 303 samples with 14 characteristics, [20] used Naïve Bayesian (NBs) and KNN methods on heart attack data. The researchers separated the test findings into six categories, with KNN and NBs achieving the highest outcomes at 79.2% and 66.6%, respectively.

In [21], a 92% accuracy rate in classifying Recently, a brain network-based system for automatically identifying heart sounds was introduced by [7]. The most effective method was a hybrid of 1D and 2D convolutional neural networks (CNNs), achieving a sensitivity of 89.22% and an accuracy of 89.94%. Furthermore, a novel method for abnormal heart sound identification was proposed by [22], combining long-term and short-term memory and temporal quasi-periodic characteristics, without segmentation. The method detects dependency linkages using long short-term memory, calculates, and extracts the cardiac sound signal's spectrogram and temporal quasi-periodic characteristics. Utilising the dataset from the 2016 PhysioNet/Computing in Cardiology Challenge for outcome evaluation allows us to quantify the level of competition. Deng and Chen used temporal models for machine learning that used recurrent neural networks (RNNs) and CNNs [23-24]. Using a new deep learning approach called DsaNet, which combines discrete depth-wise convolution and attention [25], showed how to categorize PCG signals immediately without requiring a complex feature architecture. PhysioNet/CinC Challenge dataset performance was competitive for DsaNet, especially in unbalanced signal classification. Numerous investigations have examined combining domain-specific adaptations with machine learning. Combining clinical data with cardiac computed

tomography angiography (CCTA) data improves mortality prediction in patients with heart disease when ML models are used in tandem with, for instance, clinical measures alone [26].

The categorization of heart sounds into three categories was significantly improved using a deep neural network that employs attention mechanisms, as proposed by [27]. As feature extraction continues to advance, [28] have proposed equal-scale frequency cepstral coefficients (EFCC) as an improvement over conventional MFCCs for processing heart sound signals. Their CNN, RNN, and random forest layer-based model showed strong performance in real-time cardiac monitoring, especially when applied to new patient data. Similarly, [29] achieved 99.67% accuracy in multi-class heart-sound classification using log-mel spectra and hybrid deep learning models. This categorization covers both normal and diseased conditions. Further, [30] employed a combination of conventional machine learning techniques, transfer learning, and deep learning to switch the learning algorithms from heart sound classification to image classification. Their methodology, based on log-mel and log-power properties, showed promise in cardiac sound analysis, improving categorization accuracy by 6–10% compared to conventional approaches. The evolution of methods for analysing cardiac sounds was covered in depth in a study by [31], which also compared and contrasted deep learning with more conventional machine learning approaches and forecasted the field's future possibilities. [32] Showed that short-duration PCG signal analysis could be effectively accomplished using a hybrid method that included signal filtering, spectrogram synthesis, and a method of classification based on voting. Using the PASCAL dataset, our method achieved an accuracy above 95%. In a similar vein, the ChronicNet model uses PCG data analysis to improve early CHF identification by integrating ML and DL approaches [33]. ChronicNet outperformed conventional techniques in CHF identification, achieving a noteworthy accuracy of 98.84% by combining a CNN to identify significant CHF traits with MFCC for feature extraction. In conclusion, heart sound classification has significantly improved using the combination of deep learning models, domain-specific modifications, and time-frequency representations. The trend toward integrating deep learning models with numerous feature extraction approaches is leading to more accurate and reliable diagnoses with greater clinical relevance, especially in real-time monitoring systems.

### 2.1 Research Gap

Heartbeat classification has advanced significantly thanks to deep learning and machine learning techniques, however, there are still crucial issues that must be addressed. Current models largely focus on unimodal inputs and traditional architectures such as CNNs, LSTMs, and RNNs, without leveraging the power of multimodal data fusion and cutting-edge architectures like ViTs to improve contextual representation and global feature learning. Additionally, problems of class imbalance, limited real-world generalizability, and insufficient data continue to linger due to insufficient exploitation of data simulation methods such as GANs and poor augmentation methodologies. Most works also fail to use standard hyperparameter optimization and testing protocols, instead adopting rudimentary grid search and non-stratified data partitioning, which can lead to biased generalization. Addressing these limitations, this study formulates a new framework that merges PCG signals with associated metadata for multimodal analysis, leverages GANs for simulated pathological data, uses ViT- based models for handling intricate temporal relationships, features advanced audio augmentations for enhancing robustness, and utilizes Bayesian optimization with stratified cross-validation for sound performance estimation.

## 3. Background

### 3.1 Generative Adversarial Networks (GANs)

GANs represent a powerful deep learning framework in which the discriminator and the generator are two neural networks that compete with one another. Originally designed for image data synthesis [34], GANs have since been extended to various domains, including biomedical signal processing, such as PCG heart-sound synthesis. The discriminator and generator are the two main components of the GAN architecture (Figure 2), and they operate in complete opposition to one another. Differentiating between genuine and synthesized samples is the discriminator's responsibility, whereas the generator network's primary function is to produce data samples that closely match the original sample [35].

For this reason, GAN may be used for image, video, and audio synthesis. Within the standard GAN architecture, the network of discriminators is trained to distinguish between actual heart sound signals (from a source dataset) and synthetically generated signals, whereas the network of generators is trained to produce realistic heart sound signals that can mislead the discriminator. After learning to produce synthetic heart-sound signals, the generative network is trained to imitate the distribution of actual PCG recordings by first sampling random noise vectors. Optimization is posed as a two-player minimax game and modeled mathematically as:

$$\min_{\emptyset G} \max_{\emptyset D} \sum_{i=1}^{n_r} \log f(x_i; \emptyset_D) + \sum_{j=1}^{n_z} \log(1 - f(g(x_j; \emptyset_G); \emptyset_D \quad (1)$$

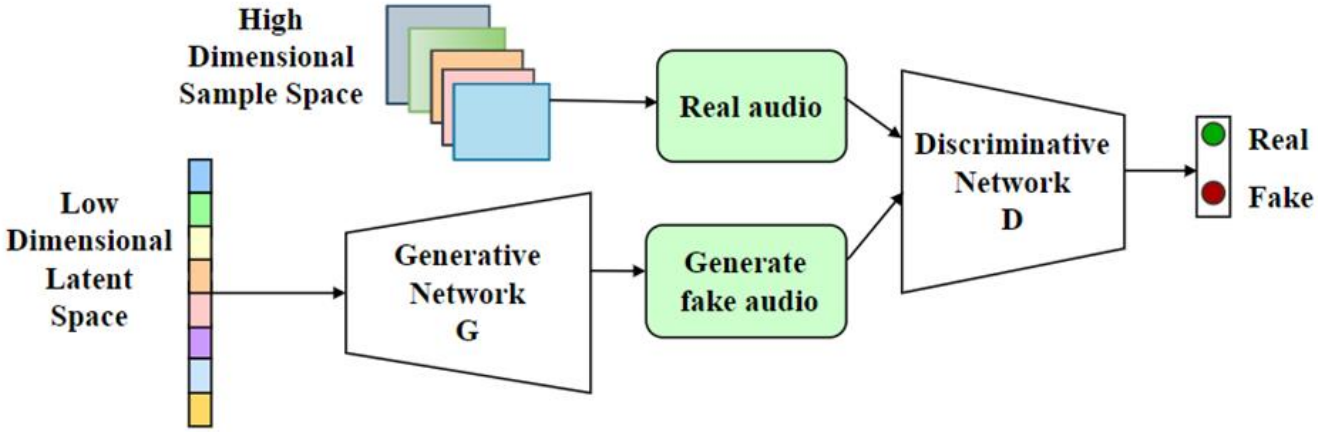


**Figure 2.** GAN architecture

Here $\emptyset_G$ means generator parameters, $\emptyset_D$ means discriminator parameters, $x_i$ carries original data, function g represents output of the generator, function f represents output of the discriminator etc. The training follows two alternating steps. In the first, the discriminator's parameters are adjusted to separate real heart sound signals from artificial ones properly. In the second, the generator's parameters are adjusted to generate signals that make it harder for the discriminator to distinguish them from real samples. This hostile instruction continues until the discriminator can no longer consistently separate real from generated heart sounds, indicating that the generator has learned the underlying distribution of the original heart sound data.

### 3.2 Vision Transformers (ViTs)

CNNs' long-standing supremacy in image processing has been challenged in recent years by ViTs, which have introduced a new approach [36]. To process picture data, ViTs use the Transformer architecture, which was first created for NLP. their flexible architecture has demonstrated significant potential in other domains, including biomedical signal processing. ViTs have shown potential in the analysis of PCG heart sound data, where precise categorization and diagnosis require global signal patterns and temporal relationships. ViTs are well-suited for problems requiring the categorisation of heart sounds because of their capacity to represent intricate, sequential connections, which permits the robust feature extraction from segmented audio patches. The components that make up the ViT architecture (Figure 3) are:

Embedding Layer: After splitting the input heartbeat signal into a collection of distinct patches, each patch is converted to a vector. To discover the input's characteristic images, the embedding layer receives the vector representations of each patch. Subsequently, the transform encoder receives the learnt values.

Transformer encoder: Transformer encoders are multi-level neural networks that include fully linked feed-forward multilayered perceptrons (MLPs) and multi-head self-attention (MSA) layers. Where l ranges from 1 to n, the encoder takes as input spot vector estimates, represented by z(l - 1). L is the sum of all encoder layers and also the index of a particular layer.

The description of the layer of encoders is:

$$\mathrm{Zl} = \mathrm{MLP}\,(\mathrm{MSA}\,(\mathrm{LN}\,(z^{l-1}))) + Z^{l-1}, \mathrm{l} = 1, \ldots, \mathrm{L} \qquad (2)$$

In this case, MLP stands for feed-forward multi-layered perception, MSA for multi-head self- awareness system, and LN for layer normalization. Normalization of layers improves network performance by normalizing input values across features, helping alleviate the issue with disappearing slopes in backpropagation. The MSA layer uses self-attention to focus on different regions of the heart sound signal, capturing interdependencies among segments. The result is then added back via a skip connection, preserving the input signal and enhancing gradient flow. The MLP's output is combined with its input via the MSA layer, using residual connections, further mitigating gradient vanishing and enabling deep feature learning. Every L encoder layer undergoes this procedure again. Each refines the heart sound representation to extract high-level, informative features.

MLP Head: Layer normalization is applied to the results from the last encoder level to produce the final heartbeat democracy, r:

$$r = LN(Z_L^0) \qquad (3)$$

To classify cardiac abnormalities, this model is then run through a single secret layer using an activated sigmoid in an MLP head.

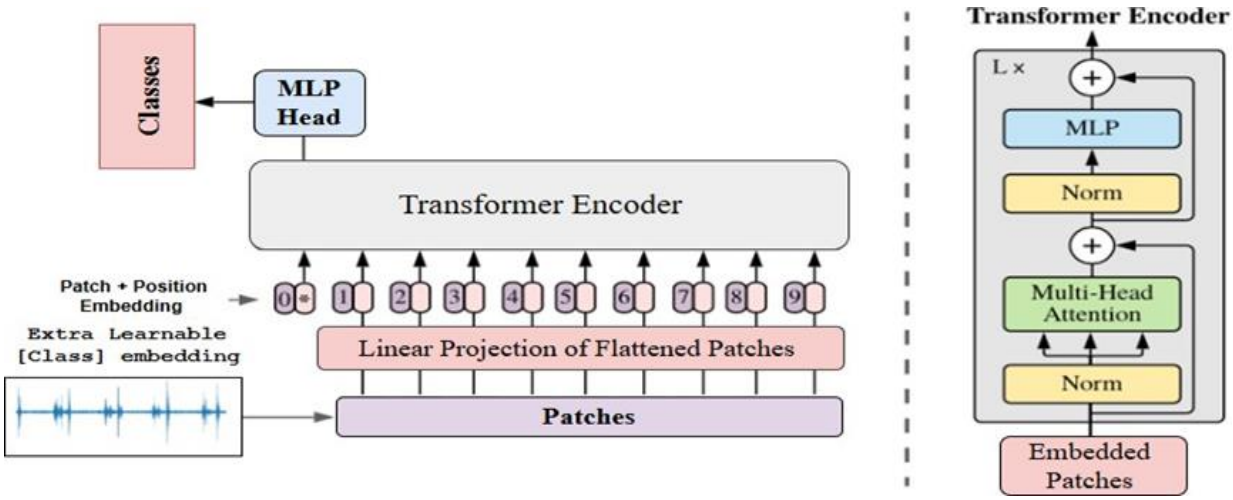


**Figure 3.** Vision Transformer ViT Architecture

This structure enables the model to learn discriminative features from PCG signals, making it particularly useful for applications involving heart-sound categorization.

## 4. Methodology

### 4.1 Data collection

The information came from two sources: (A) clinical research and (B) the general public using the iPhone app iStethoscope Pro at a hospital with the DigiScope digital stethoscope. This dataset contains recorded heart sounds from a range of people with both normal and pathological cardiac problems. To provide variation in auditory characteristics, the recordings come from a variety of sources, including wearable devices and clinical settings. An important feature of this dataset for classification problems is the ease with which normal and abnormal cardiac sounds can be distinguished. To access this dataset, kindly hold down the Ctrl key and click the link below.

Dataset link:
https://www.kaggle.com/datasets/kinguistics/heartbeat-sounds/data

### 4.2 Data Preprocessing

The initial phase in preprocessing heart sound data is loading the audio file with librosa. load (), which extracts the sample rate (sr) and transforms the signal into a NumPy array [37]. librosa. effects. trim () is employed to remove leading and trailing quiet to enhance signal clarity [38]. Resampling is performed if necessary to maintain a constant sampling rate for all recordings. Feature extraction ensues, including the estimation of Mel-Frequency Cepstral Coefficients to extract spectral contrast and timbral characteristics from the recording [39]. To find changes in frequency distribution, extracting chroma features for data related to pitch, and computing Root Mean Square Energy (RMSE) for measuring energy variations [40]. librosa.onset.onset_detect() is employed to perform onset detection, which yields important temporal information by identifying sudden amplitude and frequency changes [41].

Finally, to validate the effectiveness of preprocessing and ensure data quality before model training, waveforms, spectrograms, and feature plots are generated.

### 4.3 Model Building

It features a Vision Transformer (ViT) classification model, with data augmentation via a Generative Adversarial Network (GAN). It is composed of a discriminator and a generator. The generator is a sequential model that takes a latent vector and gradually transforms it through dense layers into an artificial sample that is very similar to the feature space of the source data. Another sequential model, the discriminator, determines whether a given input is synthetic or real. To enhance model generalization, the generator generates additional samples after training and adds them to the initial training dataset. A further split is made to separate the validation set from the training data. Labels are one-hot encoded to match the classification task after the GAN's augmented data is incorporated into the training set. A Vision Transformer model is then trained using this improved dataset. Each transformer block in the ViT model incorporates feed-forward dense layers, layer normalization, and multi-head self-attention mechanisms. During training, residual connections are employed to enhance gradient flow and stability. Before being expanded to fit the transformer input format, the input features are first projected into a higher-dimensional space. To locate the optimal parameter values, Bayesian optimization is used with Keras Tuner. The dimensions of the feed-forward layer, dropout rates, learning rates, the number of attention heads, and various head sizes span the search space across all multiple attention tiers. To choose the

optimal hyperparameters, the validation accuracy is employed. Applying K-fold cross-validation (with three splits) further ensures the modelis resilience.

## 4.4 Model Algorithm

***Algorithm for Data Loading & Preprocessing***

Step 1: Collection of heart sound recordings as the dataset.

Step 2: Loading audio dataset using librosa. Load () and preprocess the frequencies with librosa. Effect. Trim ().

Step 3: Resample audio and extract features by using MFCC, Spectrogram.

Step 4: Detect onsets using librosa. Onset_Detect () and generate visualizations.

***Algorithm for GAN Model for Data Augmentation***

Step 1: Construct GAN with Generator (G) and Discriminator (D).

Step 2: Train G and D using real and synthetic heart sound samples.

Step **3:** Generate augmented samples using the trained Generator.

Step 4: Divide your data into two parts: 80% for training and 20% for testing.

***Algorithm for Vision Transformer Model***

Step 1: One-hot encode lablels amd project features into higher –dimensional space.

Step 2: Build ViT using multi-head attention, normalization, and residual blocks.

Step 3: Use Bayesian optimization to tune hyper parameters (heads, dropout, learning rate).

Step 4: Apply 5-fold cross-validation to validate model performance.

***Algorithm for Compute Results***

Step 1: Evaluate ViT Model on the test dataset.

Step 2: Establish the F1-Score, recall, accuracy, and precise measures.

Step 3: Analyze the performance based on evaluation metrics.

Step 4: Report final results and model effectiveness.

### *4.4.1 Model Workflow*

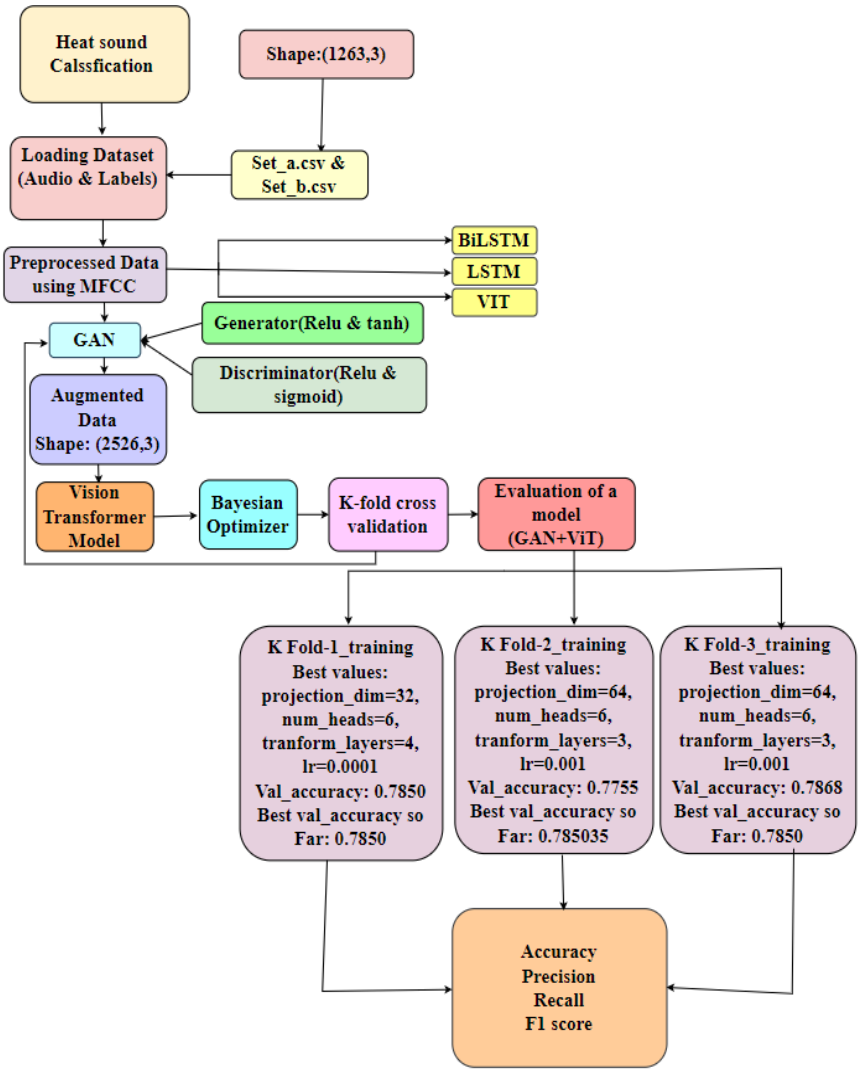


**Figure 4.** Functional View

The functional workflow of the proposed GAN–ViT-based cardiac auscultation classification system is shown in Figure 4, which includes pipelines for data preprocessing, GAN-based augmentation, ViT model training, and assessment. Evaluation of the Model

### 4.5 Model Evaluation

The process of "model evaluation" assesses a model's generalizability by testing its performance on new data. Accuracy, model performance is assessed using criteria such as F1 score, accuracy, recall, and dependability.

A confusion matrix resembles a table with actual class labels in one row and predicted class labels in the other. One can see how many records a classifier got right and how many it got wrong. One way to represent it is like this:

TP denotes "true positives," indicating the total number of tuples accurately identified as C1,

The number of tuples that were incorrectly categorized as C1 when they should have been C2 is known as "false positives," or FP for short.

FN - False negatives: the number of tuples that were mistakenly labeled as C1 when they should have been C2,

TN stands for "True Negatives" and is the sum of all the tuples that were correctly identified as C2.

**Table 1**. Depiction of the Confusion Matrix

| Actual / Predicted class label | Positive (YES) | Negative (NO) |
|---|---|---|
| Positive (YES) | True Positives | False Positives |
| Negative (NO) | False Negatives | True Negatives |

***Accuracy:*** The most straightforward way to gauge how often a classifier produces accurate predictions. Partitioning the percentage of successfully predicted favourable outcomes by all forecasts is another possible interpretation.

$$Accuracy = \frac{(TP+FP)}{(TP+FP+TN+FN)} \quad (4)$$

***Precision:*** As opposed to this ratio plus one minus it (1 – precision), which displays the proportion of false negatives, 1/Precision produces recall.

$$\text{Pr}ecision = \frac{(TP)}{(TP+FP)} \quad (5)$$

***Recall:*** In contrast, there exist what are known as false negatives with regard to true negatives.

$$\text{Re}call = \frac{(TP)}{(TP+FN)} \quad (6)$$

***F1-Score:*** A harmonic average is used to derive it from the recall and accuracy values.

$$F1 - Score = \frac{(2*\text{Precision}*\text{Recall})}{\text{Precision}+\text{Recall}} \quad (7)$$

## 5. Results and Discussion

The effectiveness of four neural network models—GAN + ViT, Bi-LSTM, LSTM, and ViT in categorizing heart sounds into three groups is thoroughly evaluated in this section. diagnostic classifications: normal, murmur, and artifact. Every model's efficacy is tested through meticulous evaluation using various metrics, including confusion matrices, ROC-AUC curves, F1-scores, recall, accuracy, and overall precision. The results show the relative advantages and disadvantages of each method, particularly highlighting the hybrid GAN + ViT architecture's enhanced classification performance.

### 5.1 Confusion Matrix

The figure depicts the confusion matrices for four different models. LSTM, GAN+ViT Bi-Directional LSTM, and ViT were used to classify cardiac sound data into three subsections: Murmur, artifact, and normal. The GAN+ViT model (Figure 5a) demonstrates superior performance, accurately recognizing 30 artifacts, 20 murmurs, and 242 normal cases, with negligible misclassifications across the categories..The Bi-Directional LSTM model (Figure 5b) demonstrates commendable performance, successfully classifying 97 artifacts, 107 murmurs, and 94 normal samples, but with marginally greater class confusion relative to GAN+ViT. The LSTM model (Figure 5c) accurately classifies 102 artifacts, 94 murmurs, and 95 normal cases; however, it shows a significant increase in misclassifying murmurs, with 18 examples incorrectly identified as normal. The ViT model (Figure 5d) has comparatively poorer accuracy in detecting artifacts and murmurs, accurately categorizing 91 artifacts, 88 murmurs, and 101 normal samples, with more confusion between the artifact and normal categories. The GAN+ViT model is the most accurate and trustworthy of the four, though the ViT model shows minor deficiencies in distinguishing among the various types of heart sounds.

### 5.2 ROC Curve

The figure displays the ROC curves for four different models. Cardiac sound classification into artifact, murmur, or normal is performed using GAN+ViT, Bi-Directional LSTM, LSTM, and ViT. Figures show the Area under the Curve (AUC) values for the false alarm rate and the number of positive results for each of the three groups, indicating how effectively the algorithms discriminate. The GAN + ViT model (Figure 6a) exhibits superior performance, achieving AUC values of 0.91,

0.93, and 0.92 for artefact, murmur, and artefact, respectively, on normal data, indicating robust classification across all three categories. The Bi-Directional LSTM model (Figure 6b) exhibits AUC values of 0.88 for artifact, 0.90 for murmur, and 0.88 for normal, reflecting commendable albeit marginally worse accuracy relative to GAN + ViT.

The LSTM model (Figure 6c) achieves AUC values of 0.89 for artifact, 0.84 for murmur, and 0.89 for normal, indicating modest classification performance, with a significant decline in murmur detection. Ultimately, the ViT model (Figure 7d) produces the lowest AUC scores— 0.85 for artifact, 0.85 for murmur, and 0.85 for normal—indicating relatively inferior class classification. The ROC curves corroborate the results from the confusion matrices: the GAN + ViT model exhibits the highest efficacy in categorizing heart sounds, succeeded by the Bi-Directional LSTM, LSTM, and ViT models in that order.

## 5.3 Performance Metrics

The table 2 and the bar plot depicting the performance metrics for several models employed in heart sound classification identify the Hybrid GAN+ViT model as the highest performer, as shown in figure 7. With a maximum accuracy of 0.91, precision of 0.89, accuracy of 0.90, and F1-score of 0.90, classification ability is optimally balanced and extremely good. Bi-Directional LSTM is a very good model for sequential modelling, with an F1-score of 0.83 and precision, recall, and accuracy of 0.85, 0.85, and 0.84, respectively. With an F1-score, recall, accuracy, and precision of 0.82, all are descriptors of a standard LSTM model. Its performance is somewhat worse, but comparable.

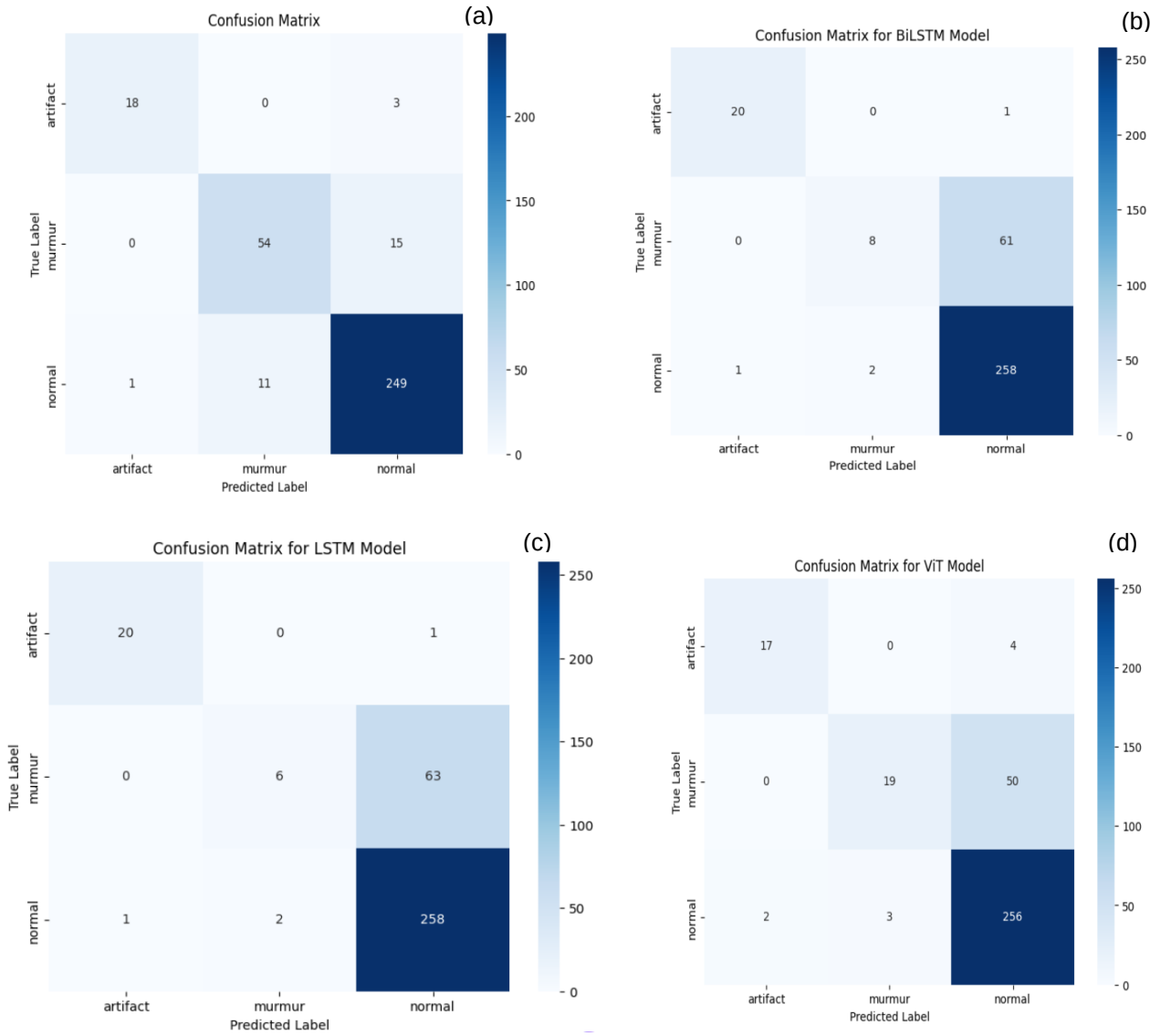


**Figure 5.** The confusion matrix of several classifiers **a**. A GAN + ViT Confusion Matrix, **b)** A Bi-LSTM Confusion Matrix, **c)** The LSTM Model's Confusing Matrix, **d)** A ViT Model Confusion Matrix,

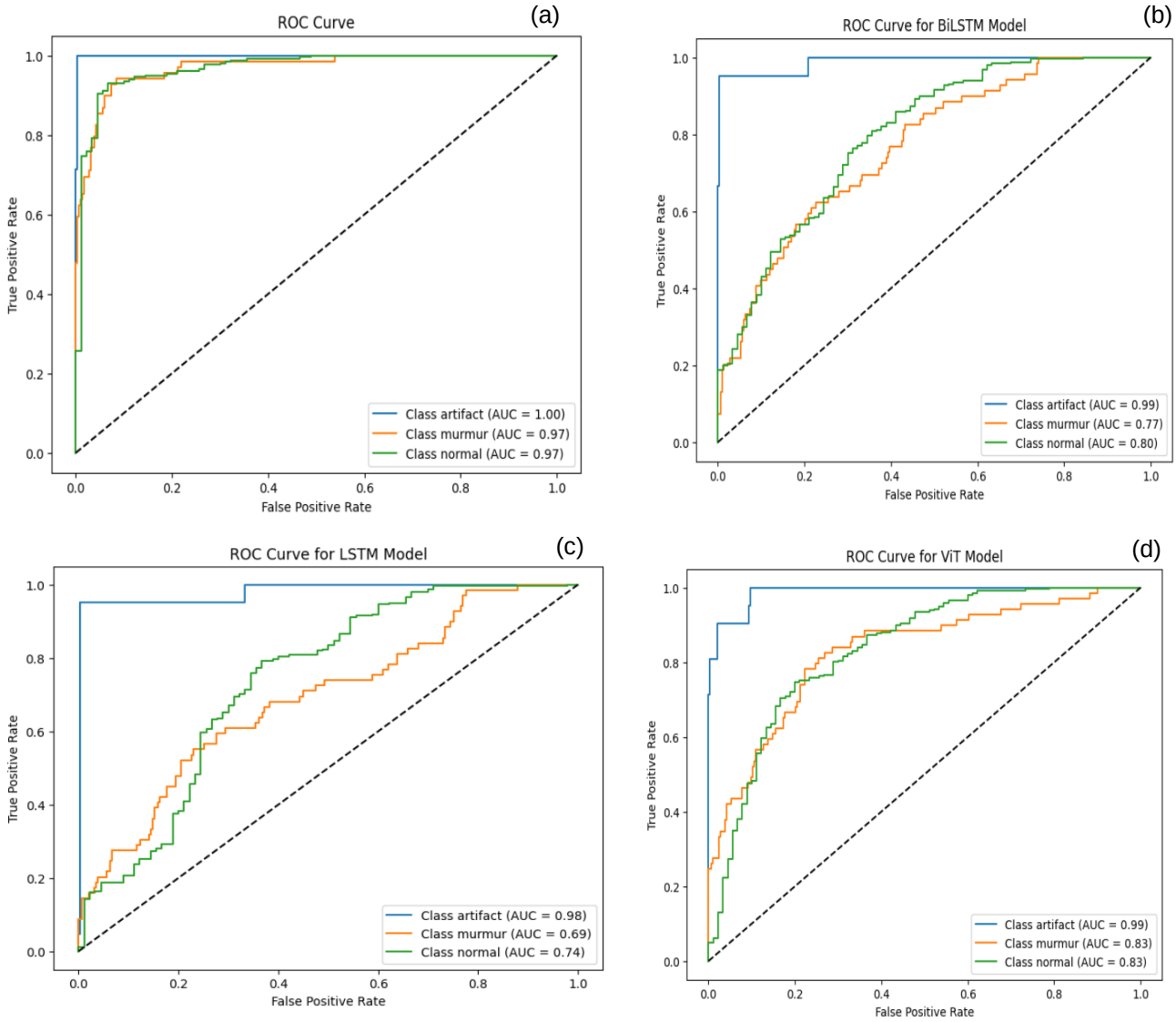


**Figure 6.** ROC curve for several classifiers, **a)** The GAN+ViT ROC curve, **b)** The ROC Curve of Bi-LSTM, **c)** ROC Curve for LSTM, **d)** The ViT ROC Curve

**Table 2.** Models' Comparative Performance

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| **Hybrid GAN+ ViT** | 0.900 ± 0.011 | 0.910 ± 0.013 | 0.890 ± 0.015 | 0.900 ± 0.012 |
| **Bi-LSTM** | 0.850 ± 0.014 | 0.840 ± 0.018 | 0.850 ± 0.016 | 0.830 ± 0.014 |
| **LSTM** | 0.830 ± 0.017 | 0.820 ± 0.020 | 0.830 ± 0.015 | 0.820 ± 0.018 |
| **ViT** | 0.800 ± 0.016 | 0.800 ± 0.017 | 0.790 ± 0.019 | 0.800 ± 0.016 |

The ViT model is the poorest of the four, with reliability, precision, and recall scores of 0.80, 0.79, and 0.79, respectively.

This comparison reveals that hybrid deep learning algorithms, particularly the GAN-ViT combination, generally perform better at categorising heart sound signals.

This pattern is supported by the comparison findings in Table 3, which show that the proposed GAN+ViT model outperforms earlier benchmark models in terms of accuracy, recall, and F1 Scores across a variety of datasets.

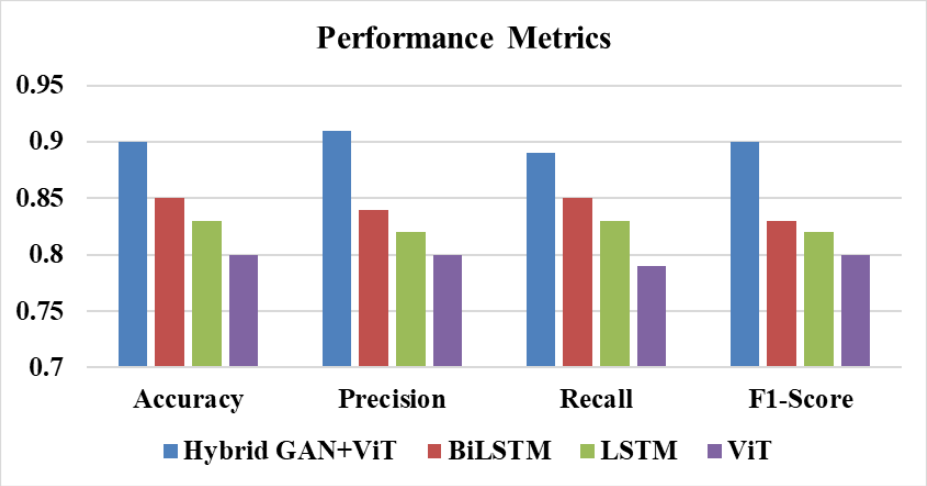


**Figure 7.** Performance Metrics

**Table 3.** Comparative Performance of Heart Sound Classification Models Across Studies

| Reference | Model | Dataset/Task | Accuracy | Precision | Recall | F1-Score | Remarks |
|---|---|---|---|---|---|---|---|
| [42] | CNN + Ensemble | PhysioNet heart sounds | 0.86 | 0.85 | 0.84 | 0.84 | Baseline deep learning approach for heart sounds |
| [43] | Bi-LSTM | Heart sound anomaly detection | 0.84 | 0.83 | 0.84 | 0.82 | Good temporal modeling, but struggles with noise |
| [44] | Attention-CRNN | Murmur classification | 0.88 | 0.87 | 0.88 | 0.87 | Attention boosts interpretability and accuracy |
| [45] | Hybrid CNN-RNN | PCG classification | 0.87 | 0.86 | 0.87 | 0.86 | The hybrid model improves robustness under variability |
| [46] | GAN + ViT | Heart sound classification | 0.90 | 0.91 | 0.89 | 0.90 | Best performance with augmentation + attention |

## 5.4 Discussion

The study provides an thorough evaluation and comparison of four deep learning architectures GAN+ViT, Bi-LSTM, LSTM, and ViT for Heart sounds could be classified as artifacts, murmurs, or normal sounds Among the hybrid models, GAN+ViT exhibits superior performance, achieving the maximum accuracy (90%), precision (0.91), recall (0.89), and F1-score (0.90). Outstanding area under the receiver operating characteristic (ROC) curve (AUC-ROC) scores (0.92 for artifact, 0.93 for murmur, and 0.91 for normal) and strong discriminative power demonstrate its outstanding performance. The low misclassifications in the GAN+ViT confusion matrix indicate that It accurately discriminates between normal and pathological cardiac auscultatory sounds. Generative Adversarial Networks and Vision Transformers collaborate well to achieve the success of this approach. Generative Adversarial Networks (GANs) enhance training data by incorporating artificial examples, which mitigates overfitting and improves generalization, whereas Vision Transformers (ViT) efficiently encode distant interactions in spectrogram representations of cardiac sounds using a self-attention system. The above hybrid ensemble guarantees ensure the model is appropriately processed and trained using high-quality augmented data spatial-spectral information to make it very apt for use in medical audio classification tasks. The BiLSTM model, as efficient with a recall of 0.85 and accuracy of 85%, lags behind the hybrid model due to its limited performance in uncertain categorization

scenarios. The standard LSTM model registered an accuracy of 83%, particularly struggling in murmur classification by incorrectly labeling 18 murmur instances as normal, and exhibiting its weakness in dealing with sophisticated acoustic variabilities. The need for augmentation or hybrid models to more effectively leverage the strengths of attention-based architectures in biomedical audio analysis is highlighted by the standalone ViT model's suboptimal performance, which achieved only 80% accuracy in separating artefacts from normal heart sounds.

The findings underscore the importance of hybrid architectures in biomedical signal processing, as GAN+ViT improves performance while enhancing model stability, making it a strong candidate for clinical applications in real-world settings. Physicians can leverage machine-based heart sound analysis with such architectures to early diagnose cardiovascular diseases, reducing diagnostic delays. Apart from accuracy, clinical implementation of AI models is determined by interpretability and transparency. In this context, Explainable AI (XAI)—specifically, attention visu,alisation in Vision Transformers—is especially important. Attention maps may highlight regions of the spectrogram that most significantly contributed to predictions and enable doctors to monitor and confirm the AI's diagnostic outputs. Interpretability is critical in healthcare applications, where black-box models have the potential to erode trust. Saliency heatmaps and class activation mapping (CAM) tools promote confidence, traceability, and regulatory compliance, making it possible to introduce them securely into diagnostic pipelines.

The recommended GAN+ViT is also very easy to set up on edge devices such as mobile phones, tablets, or digital stethoscopes since it is lightweight and adaptable. This is especially useful when it's hard to see an experienced cardiologist, such as in areas that can't be reached remotely or lack sufficient resources. The model can be integrated into telemedicine systems or mobile health applications for real-time, AI-enhanced cardiac screening during virtual consultations. It has a short inference time and supports TensorFlow Lite and ONNX runtimes. This kind of integration might make it much easier to detect diseases early, make healthcare more equitable, and enable large-scale screening programs in rural or primary care clinics.

The anticipated GAN+ViT model demonstrates improved classification performance; however, its clinical use is based on both accuracy and transparency. In this regard, Explainable AI (XAI), particularly with regard to visualization in Vision Transformers, is of great relevance. Attention maps identify the segments of heart sound spectrograms that most strongly influenced the final categorization, thereby allowing physicians to understand the rationale for the AI's decision. Model opacity can lead practitioners to be hesitant or doubtful, making interpretability crucial in high-risk sectors such as healthcare. Attention-based saliency heatmaps or class activation mapping (CAM) make the model's focus at the time of distinction between murmurs and normal sounds understandable. Enhanced trust in automated systems, simpler clinical verification, auditability, and compliance with regulation are all outputs of AI visual interpretability, which then makes it safer and more reliable to use in actual diagnostic environments. The study highlights the potential of GAN+ViT and other hybrid deep learning frameworks for enhancing automated cardiac auscultation analysis. These frameworks offer interpretability, usability, and high accuracy in the contemporary AI-based healthcare framework.

## 6. Conclusion

GAN + ViT has demonstrated outstanding performance in differentiating normal, murmur, and artifact cardiac sounds. Vision Transformers and GANs are two parts of Generative Adversarial Networks. With the best score of 0.90 for F1-score, 0.89 for recall, 90% for accuracy, and 0.91 for precision. when compared to all architectures that were experimented upon, i.e., Bi-LSTM, LSTM, and stand-alone ViT. Its outstanding discriminative ability was also well supported by high AUC-ROC values (0.92 for artifact, 0.93 for murmur, and 0.91 for normal), with insignificant misclassifications within the matrix of confusion. The efficacy of the GAN + ViT model lies in the synergistic combination of GAN-based data augmentation, this hybrid approach improves generalization and ensures efficient feature extraction, making it highly suitable for medical audio classification tasks. Compared to sequential models (Bi-LSTM and LSTM), GAN + ViT showed better ability in handling small acoustic variations, particularly in distinguishing murmurs from normal noises. While Bi-LSTM achieved 85% accuracy, the hybrid model surpassed it, and the standard LSTM demonstrated serious shortcomings in murmur detection. The standalone ViT model, as promising as it was, shown poor performance (80% accuracy), highlighting the necessity of GAN augmentation to improve its classification effectiveness. These results demonstrate the effectiveness of hybrid deep learning models for automated cardiac diagnosis, providing a reliable tool for early identification of cardiac anomalies. Future work can explore multi-modal fusion, for example, combining ECG and phonocardiogram signals with explainable AI techniques to enhance interpretability for healthcare applications. In general, GAN + ViT offers a state-of-the-art heart sound detection solution that enhances accuracy, efficiency, and scalability in AI-driven healthcare applications.

## References


[1] WHO, (2020) Cardiovascular Diseases, Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] C. Liu, D. Springer, Q. Li, B. Moody, R.A. Juan, F.J. Chorro, F. Castells, J.M. Roig, I. Silva, A.E. Johnson, Z. Syed, S.E. Schmidt, C.D. Papadaniil, L. Hadjileontiadis, H. Naseri, A. Moukadem, A. Dieterlen, C. Brandt, H. Tang, M. Samieinasab, M.R. Samieinasab, R. Sameni, R.G. Mark, G.D. Clifford, An open access database for the evaluation of heart sound algorithms. Physiological measurement, 37(12), (2016) 2181. https://doi.org/10.1088/0967-3334/37/12/2181

[3] C. Liu, A. Murray, Applications of complexity analysis in clinical heart failure. In Complexity and Nonlinearity in Cardiovascular Signals, Springer, (2017) 301–325. https://doi.org/10.1007/978-3-319-58709-7_11

[4] D.H. Peters, A. Garg, G. Bloom, D.G. Walker, W.R. Brieger, M. Hafizur Rahman, Poverty and access to health care in developing countries. Annals of the New York Academy of Sciences, 1136(1), (2008) 161–171. https://doi.org/10.1196/annals.1425.011

[5] A.K. Dwivedi, S.A. Imtiaz, E. Rodriguez-Villegas, Algorithms for automatic analysis and classification of heart sounds–a systematic review. IEEE Access, IEEE, 7, (2018) 8316–8345. https://doi.org/10.1109/ACCESS.2018.2889437

[6] J.S. Chorba, A.M. Shapiro, L. Le, J. Maidens, J. Prince, S. Pham, M.M. Kanzawa, D.N. Barbosa, C. Currie, C. Brooks, B.E. White, Deep learning algorithm for automated cardiac murmur detection via a digital stethoscope platform. Journal of the American Heart Association, 10(9), (2021) e019905. https://doi.org/10.1161/JAHA.120.019905

[7] F. Noman, C.M. Ting, S.H. Salleh, H. Ombao, (2019) Short-segment heart sound classification using an ensemble of deep convolutional neural networks. In ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, Brighton, UK, 1318–1322. https://doi.org/10.1109/ICASSP.2019.8682668

[8] J. Lee, T. Kang, N. Kim, S. Han, H. Won, W. Gong, I.Y. Kwak, Deep learning based heart murmur detection using frequency-time domain features of heartbeat sounds. In 2022 Computing in Cardiology (CinC), IEEE, Tampere, Finland, 498, (2022) 1–4. https://doi.org/10.22489/CinC.2022.071

[9] H. Lu, J.B. Yip, T. Steigleder, S. Grießhammer, M. Heckel, N.V.S.J. Jami, B. Eskofier, C. Ostgathe, A. Koelpin, A lightweight robust approach for automatic heart murmurs and clinical outcomes classification from phonocardiogram recordings. In 2022 Computing in Cardiology (CinC), IEEE, (2022) 1–4. https://doi.org/10.22489/CinC.2022.165

[10] G.B. Lim, AI used to detect cardiac murmurs, Nature Reviews Cardiology, 18(7), (2021) 460. https://doi.org/10.1038/s41569-021-00567-8

[11] M. Zha, G. Meng, C. Lin, Z. Zhou, K. Chen. (2019) RoLMA: a practical adversarial attack against deep learning-based LPR systems. In International conference on information security and cryptology, Springer, 101–117. https://doi.org/10.1007/978-3-030-42921-8_6

[12] K. Phua, J. Chen, T.H. Dat, L. Shue, Heart sound as a biometric. Pattern Recognit, 41(3), (2008) 906–919. https://doi.org/10.1016/j.patcog.2007.07.018

[13] L. Jia, D. Song, L. Tao, Y. Lu, Heart sounds classification with a fuzzy neural network method with structure learning. In International Symposium on Neural Networks, Springer, (2012) 130–140. https://doi.org/10.1007/978-3-642-31362-2_15

[14] S.W. Deng, J.Q. Han, towards heart sound classification without segmentation via autocorrelation feature and diffusion maps. Future Generation Computer Systems, 60, (2016) 13–21. https://doi.org/10.1016/j.future.2016.01.010

[15] M. Zabihi, A.B. Rad, S. Kiranyaz, M. Gabbouj, A.K. Katsaggelos, Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. In 2016 computing in cardiology conference (CinC), IEEE, (2016) 613–616. https://doi.org/10.22489/CinC.2016.180-213

[16] C. Potes, S. Parvaneh, A. Rahman, B. Conroy, Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. In 2016 computing in cardiology conference (CinC), IEEE, (2016) 621–624. https://doi.org/10.22489/CinC.2016.182-399

[17] W. Zhang, J. Han, S. Deng, Heart sound classification based on scaled spectrogram and partial least squares regression. Biomedical Signal Processing and Control, 32, (2017) 20–28. https://doi.org/10.1016/j.bspc.2016.10.004

[18] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, A.A. Yarifard, Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Computer methods and programs in biomedicine, 141, (2017) 19–26. https://doi.org/10.1016/j.cmpb.2017.01.004

[19] J.P. Dominguez-Morales, A.F. Jimenez-Fernandez, M.J. Dominguez-Morales, G.

Jimenez-Moreno, Deep neural networks for the recognition and classification of heart murmurs using neuromorphic auditory sensors. IEEE transactions on biomedical circuits and systems, IEEE, 12(1), (2017) 24–34. https://doi.org/10.1109/TBCAS.2017.2751545

[20] F. Zulfiqar, U.I. Bajwa, Y. Mehmood, Multi-class classification of brain tumor types from MR images using EfficientNets. Biomedical Signal Processing and Control, 84, (2023) 104777. https://doi.org/10.1016/j.bspc.2023.104777

[21] M. Hamidi, H. Ghassemian, M. Imani, Classification of heart sound signal using curve fitting and fractal dimension. Biomedical Signal Processing and Control, 39, (2018) 351–359. https://doi.org/10.1016/j.bspc.2017.08.002

[22] W. Zhang, J. Han, S. Deng, Abnormal heart sound detection using temporal quasi-periodic features and long short-term memory without segmentation. Biomedical Signal Processing and Control, 53, (2019) 101560. https://doi.org/10.1016/j.bspc.2019.101560

[23] M. Deng, T. Meng, J. Cao, S. Wang, J. Zhang, H. Fan, Heart sound classification based on improved MFCC features and convolutional recurrent neural networks. Neural Networks, 130, (2020) 22–32, 2020. https://doi.org/10.1016/j.neunet.2020.06.015

[24] W. Chen, Q. Sun, X. Chen, G. Xie, H. Wu, C. Xu, Deep learning methods for heart sounds classification: A systematic review. Entropy, 23(6), (2021) 667. https://doi.org/10.3390/e23060667

[25] G. Tian, C. Lian, Z. Zeng, B. Xu, Y. Su, J. Zang, Z. Zhang, C. Xue, Imbalanced heart sound signal classification based on two-stage trained dsanet. Cognitive Computation, 14(4), (2022) 1378–1391. https://doi.org/10.1007/s12559-022-10009-3

[26] W. Xu, K. Yu, J. Ye, H. Li, J. Chen, F. Yin, J. Xu, J. Zhu, D. Li, Q. Shu, Automatic pediatric congenital heart disease classification based on heart sound signal. Artificial intelligence in medicine, 126, (2022) 102257. https://doi.org/10.1016/j.artmed.2022.102257

[27] Z. Ren, K. Qian, F. Dong, Z. Dai, W. Nejdl, Y. Yamamoto, B.W. Schuller, Deep attention-based neural networks for explainable heart sound classification. Machine Learning with Applications, 9, (2022) 100322. https://doi.org/10.1016/j.mlwa.2022.100322

[28] X. Chen, H. Li, Y. Huang, W. Han, X. Yu, P. Zhang, R. Tao, Heart sound classification based on equal scale frequency cepstral coefficients and deep learning. Biomedical Engineering/Biomedizinische Technik, 68(3), (2023) 285–295. https://doi.org/10.1515/bmt-2021-0254

[29] M.T. Nguyen, W.W. Lin, J.H. Huang, Heart sound classification using deep learning techniques based on log-mel spectrogram. Circuits, Systems, and Signal Processing, 42(1), (2023) 344–360. https://doi.org/10.1007/s00034-022-02124-1

[30] M. Xiang, J. Zang, J. Wang, H. Wang, C. Zhou, R. Bi, Z. Zhang, C. Xue, Research of heart sound classification using two-dimensional features. Biomedical Signal Processing and Control, 79, (2023) 104190. https://doi.org/10.1016/j.bspc.2022.104190

[31] Z. Ren, Y. Chang, T.T. Nguyen, Y. Tan, K. Qian, B.W. Schuller, A comprehensive survey on heart sound analysis in the deep learning era. IEEE Computational Intelligence Magazine, IEEE, 19(3), (2024) 42–57. https://doi.org/10.1109/MCI.2024.3401309

[32] S. Ismail, B. Ismail, I. Siddiqi, U. Akram, PCG classification through spectrogram using transfer learning. Biomedical Signal Processing and Control, 79, (2023) 104075. https://doi.org/10.1016/j.bspc.2022.104075

[33] M. Bahreini, R. Barati, A. Kamali, Cardiac sound classification using a hybrid approach: MFCC-based feature fusion and CNN deep features. EURASIP Journal on Advances in Signal Processing, 2025(1), (2025) https://doi.org/10.1186/s13634-025-01203-0

[34] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets. Advances in neural information processing systems, 27, (2014).

[35] A.M. Shaker, M. Tantawi, H.A. Shedeed, M.F. Tolba, Generalization of convolutional neural networks for ECG classification using generative adversarial networks. IEEE access, 8, (2020) 35592–35605. https://doi.org/10.1109/ACCESS.2020.2974712

[36] J.N. Mogan, C.P. Lee, K.M. Lim, M. Ali, A. Alqahtani, Gait-CNN-ViT: Multi-model gait recognition with convolutional neural networks and vision transformer. Sensors, 23(8), (2023) 3809. https://doi.org/10.3390/s23083809

[37] S. Li, (2024) Audio Feature Extraction Algorithms and Implementation Technologies Analysis. In 2024 5th International Conference on Information Science, Parallel and Distributed Systems (ISPDS), IEEE, Guangzhou, China, 507–516. https://doi.org/10.1109/ISPDS62779.2024.10667490

[38] X. Fang, G. Wei, Research on entertainment creation robot based on artificial intelligence speech recognition in the process of music style analysis. Entertainment Computing, 51, (2024) 100739. https://doi.org/10.1016/j.entcom.2024.100739

[39] A. Tajik. (2025). Beyond Voice Recognition: Integrating Alexa's Emotional Intelligence and ChatGPT's Language Processing for EFL Learners' Development and Anxiety Reduction-A Comparative Analysis. https://doi.org/10.21203/rs.3.rs-5989702/v1

[40] A.J. Benjamin, K. Siedenburg, Effects of spectral manipulations of music mixes on musical scene analysis abilities of hearing-impaired listeners. PLoS One, 20(1), (2025) e0316442. https://doi.org/10.1371/journal.pone.0316442

[41] J. Shi, L. Liu, Construction and Implementation of Content-Based National Music Retrieval Model under Deep Learning. International Journal of Information System Modeling and Design, 15(1), (2024) 1–17. https://doi.org/10.4018/IJISMD.343631

[42] S. Chakraborty, P. Kochhar, S. Patil, K. Kotecha, S. Gite, G. Selvachandran, S. Das, Generative adversarial network augmented data for improved heart sound abnormality detection. Computers in Biology and Medicine, 195, (2025) 110623. https://doi.org/10.1016/j.compbiomed.2025.110623

[43] S.U.R. Khan, Z. Khan, Detection of Abnormal Cardiac Rhythms Using Feature Fusion Technique with Heart Sound Spectrograms. Journal of Bionic Engineering, (2025) 1–20.

[44] E. Partovi, A. Babic, A. Gharehbaghi, A review on deep learning methods for heart sound signal analysis. Frontiers in Artificial Intelligence, 7, (2024) 1434022. https://doi.org/10.3389/frai.2024.1434022

[45] A.O. Ige, M. Sibiya, (2024). State-of-the-art in 1d convolutional neural networks: A survey. IEEE Access,IEEE, 144082 – 144105. https://doi.org/10.1109/ACCESS.2024.3433513

[46] M.T. Ahad, S.A. Preanto, B. Song, Y. Li, Gan-Generated Spectrogram Detection and Classification for Heartbeat Classification Using a Vision Transformer. SSRN 4892869.

## Acknowledgments

The authors are deeply grateful to the Acharya Nagarjuna University Computer Science and Engineering Department and fellow peers for their overall academic and technical support during this research study.

## Authors Contribution Statement

Divya Lalita Sri Jalligampala: Conceptualization, Data curation, Methodology, Investigation, Software, Writing – Original Draft. Gangadhara Rao Kancharla: Supervision, Formal analysis, Validation, Resources, Writing – Review & Editing. R.V.S. Lalitha: Validation, Visualization, Writing – Review & Editing, Methodology, Statistical analysis, Supervision. All authors contributed to the manuscript and approved the final version of the work.

## Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

## Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

## Data Availability

This study utilized a publicly available dataset which was downloaded from Kaggle website. iStethoscope Pro software and DigiScope instruments were utilized for recording the heart sounds. To get the dataset, click the following link:
https://www.kaggle.com/datasets/kinguistics/heartbeat-sounds/data
No personally identifiable patient information was provided, and recordings are anonymized, so the dataset is suitable for open research utilization.

## Has this article screened for similarity?

Yes

## About the License