Cardiovascular Auscultation Signal Analysis through Transformer Deep Learning

Abstract

The Philippines faces a dual burden of high disease prevalence and limited healthcare resources, leading to diagnostic delays and increased mortality rates. Auscultation, a non-invasive method for assessing heart and lung sounds, is often constrained by subjective interpretation and prone to human error. This study explores the effectiveness of Audio Spectrogram Transformer-to classify heart sounds using Digital Signal Analysis of one-dimensional heart audio signals. Model performance was assessed using precision, recall, F1-score, sensitivity, and specificity.

I. Introduction

According to the Philippine Institute for Development Studies, the Philippines faces a dual challenge: a high burden of disease alongside a critical shortage of healthcare resources, including limited medical personnel, equipment, and supplies [1]. These shortages contribute to delays in accessing basic healthcare services and lead to higher mortality rates across various demographic groups. Early detection is essential for improving the success rates of medical interventions. [2].

Auscultation, a non-invasive and efficient method used in many physical exams, helps assess sounds from the heart, lungs, and other organs. However, the interpretation of these sounds can be complex, particularly when diagnosing diseases related to the abdomen, cardiovascular system, and lungs, which poses a common challenge for healthcare professionals [3], [4], [5]. Therefore, many underserved and remote communities in local barangays experience delays in receiving accurate diagnoses and appropriate treatments, exacerbating health disparities and worsening their health conditions overtime.

Automated classification of heart sounds offers the potential to speed up the diagnostic process, allowing for more efficient and objective analysis of cardiac health. This technology could enhance the early detection accuracy of cardiovascular diseases (CVDs), helping reduce associated mortality rates. Consequently, a wide range of research efforts has been directed toward developing methods for the automatic classification of heart sounds.

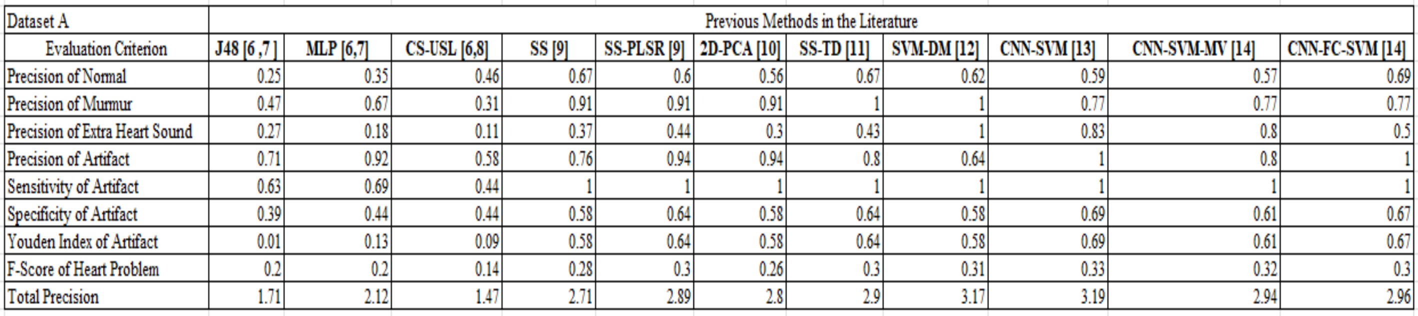
Recent research highlights the effectiveness of machine learning (ML) in classifying complex heart sounds, as demonstrated by the PASCAL Network of Excellence audio data competition [6]. Table 1 presents the results of previous classification methods using Dataset A, while Table 2 shows results with Dataset B.

ML-based methods offer certain advantages and challenges: they are typically more memory-efficient than deep learning models for small datasets, but the feature extraction process is often intricate and time-consuming. These systems rely on a range of feature extraction techniques, including statistical analysis of heart sound signal segmentation in the temporal domain and time-frequency analysis methods such as Cepstral transform, Fast Fourier Transform, and Continuous Wavelet Transform.

With the rise of deep learning, Transformer-based models have emerged as an alternative to traditional machine learning models for image classification, offering significant advantages through their self-attention mechanism, which captures both local and global dependencies in images. Unlike convolutional neural networks (CNNs) and support vector machines (SVMs), transformers process image patches independently and flexibly, allowing for a more comprehensive feature representation. This structure enables transformers to bypass the need for handcrafted feature engineering, learning complex patterns autonomously during training. Their parallel processing capability makes them highly scalable, facilitating training on large datasets while reducing computational time. Additionally, the widespread availability of pretrained transformer models allows for effective transfer learning, making them adaptable even to smaller datasets and multimodal tasks, where they can outperform traditional approaches and achieve state-of-the-art results in image classification [16].

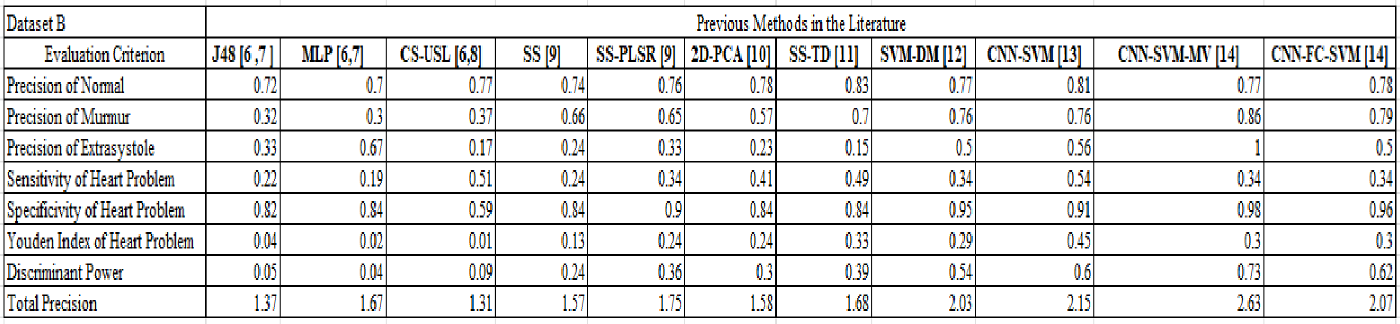
This paper investigates the effectiveness of classifying heart sounds using features extracted from images derived from 1D heart sound signals. The heart sound was transformed into 3 images: 2D viridis color map representation, Mel-scaled spectrograms and wavelet scalograms. The study utilizes pretrained transformer-based models to evaluate the preprocess images. The performance of the classifier is evaluated by precision, recall, discriminant power, F1-score, sensitivity, and specificity.

Table 1. Performance comparison of previous methods on Dataset A



Dataset A	Previous Methods in the Literature										
Evaluation Criterion	J48 [6 ,7]	MLP [6,7]	CS-USL [6,8]	SS [9]	SS-PLSR [9]	2D-PCA [10]	SS-TD [11]	SVM-DM [12]	CNN-SVM [13]	CNN-SVM-MV [14]	CNN-FC-SVM [14]
Precision of Normal	0.25	0.35	0.46	0.67	0.6	0.56	0.67	0.62	0.59	0.57	0.69
Precision of Murmur	0.47	0.67	0.31	0.91	0.91	0.91	1	1	0.77	0.77	0.77
Precision of Extra Heart Sound	0.27	0.18	0.11	0.37	0.44	0.3	0.43	1	0.83	0.8	0.5
Precision of Artifact	0.71	0.92	0.58	0.76	0.94	0.94	0.8	0.64	1	0.8	1
Sensitivity of Artifact	0.63	0.69	0.44	1	1	1	1	1	1	1	1
Specificity of Artifact	0.39	0.44	0.44	0.58	0.64	0.58	0.64	0.58	0.69	0.61	0.67
Youden Index of Artifact	0.01	0.13	0.09	0.58	0.64	0.58	0.64	0.58	0.69	0.61	0.67
F-Score of Heart Problem	0.2	0.2	0.14	0.28	0.3	0.26	0.3	0.31	0.33	0.32	0.3
Total Precision	1.71	2.12	1.47	2.71	2.89	2.8	2.9	3.17	3.19	2.94	2.96

Table 2. Performance comparison of previous methods on Dataset B



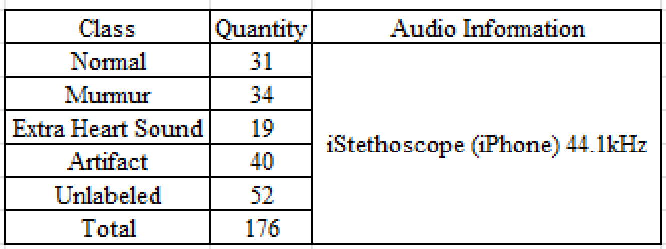
Dataset B	Previous Methods in the Literature										
Evaluation Criterion	J48 [6 ,7]	MLP [6,7]	CS-USL [6,8]	SS [9]	SS-PLSR [9]	2D-PCA [10]	SS-TD [11]	SVM-DM [12]	CNN-SVM [13]	CNN-SVM-MV [14]	CNN-FC-SVM [14]
Precision of Normal	0.72	0.7	0.77	0.74	0.76	0.78	0.83	0.77	0.81	0.77	0.78
Precision of Murmur	0.32	0.3	0.37	0.66	0.65	0.57	0.7	0.76	0.76	0.86	0.79
Precision of Extrasystole	0.33	0.67	0.17	0.24	0.33	0.23	0.15	0.5	0.56	1	0.5
Sensitivity of Heart Problem	0.22	0.19	0.51	0.24	0.34	0.41	0.49	0.34	0.54	0.34	0.34
Specificivity of Heart Problem	0.82	0.84	0.59	0.84	0.9	0.84	0.84	0.95	0.91	0.98	0.96
Youden Index of Heart Problem	0.04	0.02	0.01	0.13	0.24	0.24	0.33	0.29	0.45	0.3	0.3
Discriminant Power	0.05	0.04	0.09	0.24	0.36	0.3	0.39	0.54	0.6	0.73	0.62
Total Precision	1.37	1.67	1.31	1.57	1.75	1.58	1.68	2.03	2.15	2.63	2.07

II. Methodology

2.1 Dataset

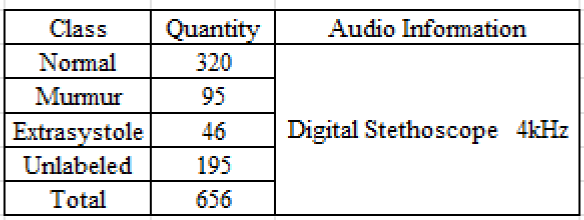
The datasets used in this study come from the Pascal Classifying Heart Sound Challenge and will be transformed into images representing heart sounds as inputs for the pretrained transformer model. Signal processing techniques will generate three types of images: Mel-scaled spectrograms, Wavelet scalograms, and 2D image in Viridis colormap. Dataset A comprises crowd-sourced audio recordings collected through the iStethoscope Pro app on iPhones, providing real-time filtering and amplification for audio quality that rivals or even surpasses digital stethoscopes. Dataset B, meanwhile, includes auscultation recordings obtained with the DigiScope Collector at the Maternal and Fetal Cardiology Unit of the Real Hospital Português (RHP) in Brazil. Tables 3 and 4 summarize the dataset's structure, detailing the number of recordings for each class label, as well as their sampling frequencies and origins.

Table 3. Dataset A Structure



Class	Quantity	Audio Information
Normal	31	iStethoscope (iPhone) 44.1kHz
Murmur	34	
Extra Heart Sound	19	
Artifact	40	
Unlabeled	52	
Total	176	

Table 4. Dataset B Structure

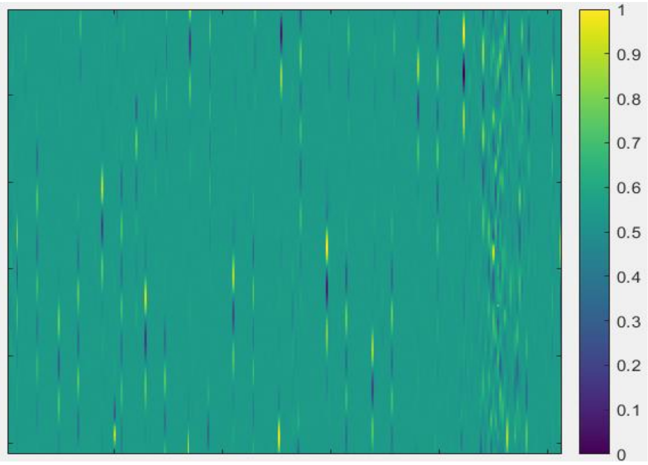


Class	Quantity	Audio Information
Normal	320	Digital Stethoscope 4kHz
Murmur	95	
Extrasystole	46	
Unlabeled	195	
Total	656	

2.2 2D image representation of 1D Heart Sound

The 2D image representation of the heart sound were generated by reshaping and interpolating the heart sound into 254*254-pixel image in the viridis color map domain. We chose the viridis color map because it improves the performance of CNN classifier. The Viridis color map is a sequential, perceptually uniform color scheme that transitions smoothly from blue through green to yellow. [17]. In figure 1, We reshape and interpolate the 1D Heart Sound into a 2D images using Matlab programming language.

Fig 1. The 2D Viridis Representation of Normal Heart Sound



2.3 Cepstrum of the Heart Sound

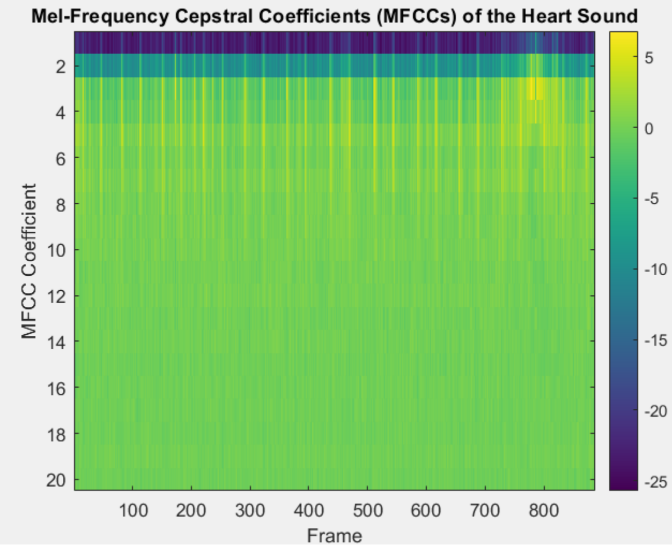
The Cepstrum of heart sounds closely replicates how the human auditory system perceives the intensity and frequency of sounds. Humans are particularly sensitive to small variations in pitch at lower frequencies, and this sensitivity is reflected in the Mel-Frequency Cepstral Coefficients scale, which is designed to be linear at lower frequencies and logarithmic at higher ones. In the MFCC algorithm, the audio signal is first segmented into smaller frames using a Hamming window, which reduces edge effects and smooths transitions between frames to minimize spectral leakage. This segmentation allows for precise analysis of short-term characteristics within each frame.

Once the signal is split into frames, a Fast Fourier Transform is applied to each one, converting the signal from the time domain into the frequency domain to capture the spectrum of each frame. This frequency spectrum is then passed through a Mel-scale filter bank given by equation (1), which consists of overlapping triangular filters spaced according to the Mel scale. Each filter emphasizes frequencies within its range, weighing the spectrum in a way that mirrors human auditory perception, particularly the heightened sensitivity to lower frequencies.

$$Mel\,(\,f\,) = 2595\,log\,(1 + \,f * 700\,) \qquad (1)$$

After filtering, logarithmic compression is applied to the output, which models the human ear’s logarithmic response to loudness and reduces dynamic range. Finally, a Discrete Cosine Transform (DCT) is applied to decorrelate the log-mel energies, producing a set of MFCCs that effectively capture the spectral characteristics of the audio signal in a compact representation. These MFCC vectors are crucial in applications like speech and audio analysis, as they provide a highly informative, perceptually relevant feature set that can capture the nuances of heart sounds for diagnostic or classification purposes [18]. In figure 2 shows the resulting mel-scale spectrogram by applying triangular filter banks to the rescale spectrogram and compute the discrete cosine function of the signal.

Fig 2. The Cepstrum Spectrogram of Normal Heart Sound



2.4 Wavelet Scalogram of the Heart Sound

The Fourier transform is a widely used method for analyzing data, as it reveals the frequency components of a signal. However, it does not provide information about the timing of the signal frequency. Both temporal and frequency information are crucial for effective analysis when signals have frequency content that changes over time, such as transient sound signals. The Windowed Fourier Transform addresses this by applying the Fourier Transform to fixed-size segments of the signal. However, fixed window sizes limit its ability to capture both high and low-frequency details simultaneously.

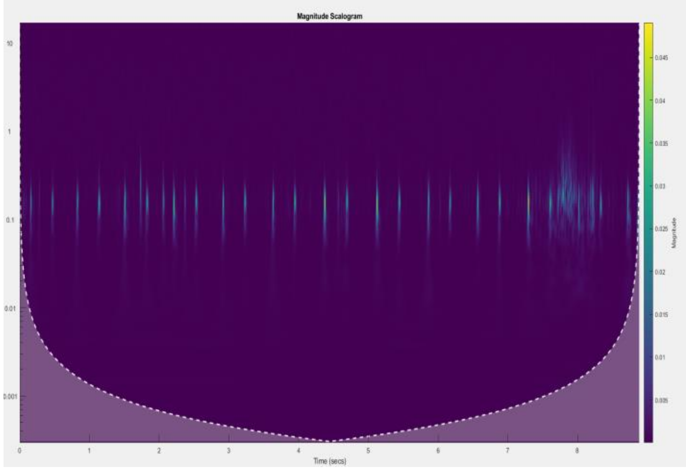
Wavelet Transform overcomes this limitation by using scalable and shiftable wavelet functions instead of sinusoids, allowing for better time-frequency resolution. It adapts to different frequency components, using finer scales for high frequencies and coarser scales for low frequencies.

The Discrete Wavelet Transform applies wavelets with discrete scaling and shifting values, enabling the efficient breakdown of a signal into its high-frequency and low-frequency components. Its capacity to analyze signals at various resolutions makes it highly effective for studying time-varying signals, such as heart sounds. These wavelet functions can be scaled and shifted, allowing the analysis of different frequency components with varying resolutions as shown in the equation (2).

$$\psi_{\gamma,\beta} \; = \; U(\omega)a_{\gamma,\beta}\,\omega^{\beta}e^{-\omega^{\gamma}} \qquad (2)$$

The families of morse wavelets will be the basis kernel of the Discrete Wavelet Transform, where U(ω) is the unit step function, $a_{\gamma,\beta}$ is the normalization constant, β and γ are the controlling parameters of Morse wavelet shape [19]. In figure 3 shows the viridis scalogram of the discrete wavelet transform using morse family of wavelets from the normal heart sound.

Fig 3. The Wavelet Scalogram of Normal Heart Sound

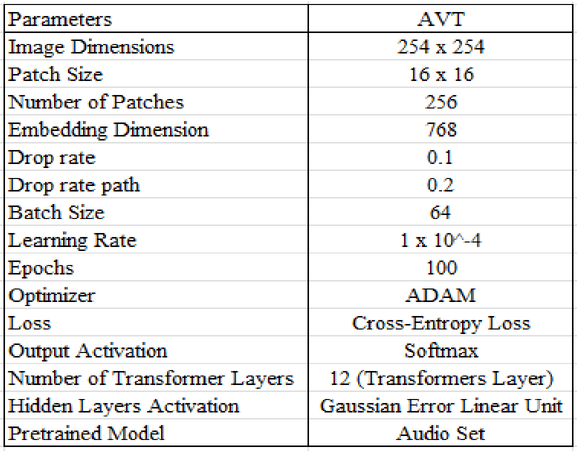


2.5 Pretrained Transformer as a classifier

The Audio Spectrogram Transformer (AST) is a deep learning model specifically designed for audio classification tasks, utilizing transformer-based architecture to process spectrograms—2D representations of audio signals that capture both time and frequency information [20]. Unlike traditional convolutional neural networks (CNNs), AST uses self-attention mechanisms to model long-range dependencies in the spectrograms, enabling the model to capture complex patterns and relationships across time and frequency. The model divides the spectrogram into patches, similar to image processing in vision transformers, and applies positional encodings to retain the time-frequency structure of the audio. AST has demonstrated strong performance in various audio classification tasks, such as speech recognition, sound event detection, and music genre classification, due to its ability to learn both local and global dependencies. Additionally, by leveraging pretraining on large datasets using Google AudioSet [21], AST can be fine-tuned for specific audio applications, making it a powerful and flexible tool for audio analysis.

We configure the parameters of AVT as a computer vision to classify our heart sound images. The parameters for our AVT model are presented in table 5

Table 5. Model Parameters of AVT



Parameters	AVT
Image Dimensions	254 x 254
Patch Size	16 x 16
Number of Patches	256
Embedding Dimension	768
Drop rate	0.1
Drop rate path	0.2
Batch Size	64
Learning Rate	1 x 10^-4
Epochs	100
Optimizer	ADAM
Loss	Cross-Entropy Loss
Output Activation	Softmax
Number of Transformer Layers	12 (Transformers Layer)
Hidden Layers Activation	Gaussian Error Linear Unit
Pretrained Model	Audio Set

2.6 Evaluation Criteria

We evaluate the performance of our classifier system by using the provided metrics specified by the PASCAL challenge. They are essentially based on precision, sensitivity and specificity. The computation of Youden's Index γ, a well-established measure for evaluating diagnostic accuracy, is defined as (3).

$$\gamma = sensitivity - (1 - specificity) \tag{3}$$

For Dataset A, γ was determined for the artifact class, while for Dataset B, it was applied to the problematic heartbeat class, which includes murmurs and extrasystoles. Furthermore, we calculated the F-Score with β = 0.9 for Dataset A, focusing on heart problem categories (murmurs and extra heart sounds combined). For Dataset B, the discriminant power was calculated to measure the algorithm's ability to distinguish between positive and negative samples.

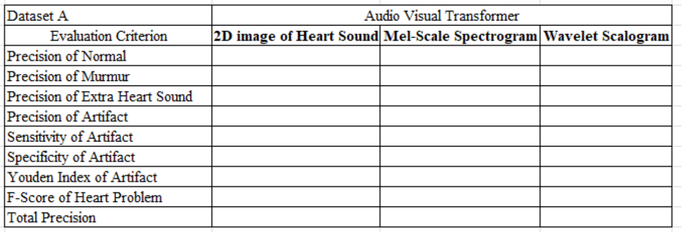
$$DP = \frac{\sqrt{3}}{\pi}(\log\left(\frac{sensitivity}{1-sensitivity}\right) + \log\left(\frac{specificity}{1-specificity}\right)) \tag{4}$$

A DP below 1 indicates poor discrimination, values between 1 and 3 reflect limited to fair performance, and values above 3 signify a strong algorithm. The evaluation script provided by the challenge in the form of an Excel spreadsheet was used for all these metrics calculation.

2.7 Experiment

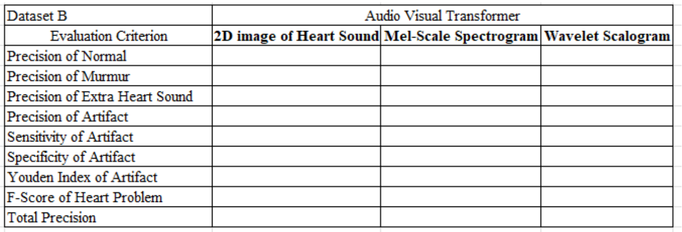
We train the AVT with the processed heart audio signal from signal processing methods separately and test them for its performance. Implementation and experiment are conducted in the open-source programming language Python for the classifier and license MATLAB for preprocessing the heart audio signal.

Table 6. Result for Dataset A



Dataset A	Audio Visual Transformer		
Evaluation Criterion	**2D image of Heart Sound**	**Mel-Scale Spectrogram**	**Wavelet Scalogram**
Precision of Normal			
Precision of Murmur			
Precision of Extra Heart Sound			
Precision of Artifact			
Sensitivity of Artifact			
Specificity of Artifact			
Youden Index of Artifact			
F-Score of Heart Problem			
Total Precision			

Table 7. Result for Dataset B



Dataset B	Audio Visual Transformer		
Evaluation Criterion	**2D image of Heart Sound**	**Mel-Scale Spectrogram**	**Wavelet Scalogram**
Precision of Normal			
Precision of Murmur			
Precision of Extra Heart Sound			
Precision of Artifact			
Sensitivity of Artifact			
Specificity of Artifact			
Youden Index of Artifact			
F-Score of Heart Problem			
Total Precision			

References

[1] V. G. T. Ulep and L. D. D. Casas, "Regional Health Integration and Cooperation in the Philippines".

[2] WHO. *WHO's Global Health Estimates: The Top 10 Causes of Death*; World Health Organization: Geneva, Switzerland, 2023

[3] A. H. Davidsen, S. Andersen, P. A. Halvorsen, H. Schirmer, E. Reierth, and H. Melbye, "Diagnostic accuracy of heart auscultation for detecting valve disease: a systematic review," BMJ Open, vol. 13, no. 3, p. e068121, Mar. 2023,

[4] Hafke-Dys, H., Bręborowicz, A., Kleka, P., Kociński, J., & Biniakowski, A. (2019). The accuracy of lung auscultation in the practice of physicians and medical students.

[5] B. M. Breum, "Accuracy of abdominal auscultation for bowel obstruction," *WJG*, vol. 21, no. 34, p. 10018, 2015,

[6] P. Bentley, G. Nordehn, M. Coimbra, and S. Mannor, "The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results,"

http://www.peterjbentley.com/heartchallenge/index.html.

[7] E. F. Gomes, P. J. Bentley, M. Coimbra, E. Pereira, and Y. Deng, "Classifying heart sounds: Approaches to the pascal challenge," in International Conference on Health Informatics, 2013, p. 337–340.

[8] E. F. Gomes and E. Pereira, "Classifying heart sounds using peak location for segmentation and feature construction," in Workshop Classifying Heart Sounds, 2012, p. 480–492.

[9] Y. Deng and P. J. Bentley, "A robust heart sound segmentation and classification algorithm using wavelet decomposition and spectrogram," in Workshop Classifying Heart Sounds, 2012, p. 1–6.

[10] S. Deng and J. Han, "Towards heart sound classification without segmentation via autocorrelation feature and diffusion maps," Future Generation Computer Systems, vol. 60, pp. 13–21, 2016.

[11] L. D. Avendano-Valencia, J. I. Godino-Llorente, M. Blanco Velasco, and G. Castellanos-Dominguez, "Feature extraction from parametric time-frequency representations for heart murmur detection," Annals of Biomedical Engineering, vol. 38(8), pp. 2716–2732, 2010.

[12] S. C. Oliveira, E. F. Gomes, and A. M. Jorge, "Heart sounds classification using motif based segmentation," in 18th International Database Engineering & Applications Symposium. Association for Computing Machinery, 2014, p. 370–371.

[13] W. Zhang, J. Han, and S. Deng, "Heart sound classification based on scaled spectrogram and tensor decomposition," Expert Systems with Applications, vol. 84, pp. 220–231, 2017.

[14] F. Demir, A. Sengur, V. Bajaj, et al., "Towards the classification of heart sounds based on convolutional deep neural network," Health Information Science and Systems, vol. 7(16), 2019.

[15] V. A. Rabello Landeira, J. O. Santos, and H. Nagano, "Comparing and Combining Audio Processing and Deep Learning Features for Classification of Heartbeat Sounds," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of: IEEE, Apr. 2024, pp. 7220–7224.

[16] A. Vaswani *et al.*, "Attention is All you Need".

[17] Freitag M, Amiriparian S, Cummins N, Gerczuk M, Schuller B. An 'End-toEvolution' hybrid approach for snore sound classifcation. In: Interspeech, Stockholm, Sweden, 2017, pp. 3507–11. 20.

[18] Nair, A.P.; Krishnan, S.; Saquib, Z. Mfcc based noise reduction in asr using kalman filtering. In Proceedings of the Conference on Advances in Signal Processing (CASP), Pune, India, 9–11 June 2016; IEEE: New York, NY, USA, 2016; pp. 474–478.

[19] J. M. Lilly, S. C. Olhede, "Generalized Morse wavelets as a superfamily of analytic wavelets", IEEE Transactions on Signal Processing, 60(11): 6036-6041, 2012.

[20] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," Jul. 08, 2021, *arXiv*: arXiv:2104.01778. doi: 10.48550/arXiv.2104.01778.

[21] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in ICASSP, 2017, pp. 776–780