Deep Learning for Heart Sound Abnormality of Infants: Proof of Conceptual Study of 1D and 2D Representations

Eashita Wazed , Jimin Lee , Hieyong Jeong *

Posted Date: 7 August 2025

doi: 10.20944/preprints202508.0454.v1

Keywords: audio data processing; audio feature extraction; deep learning in cardiology; heart sound; heart function in infant; stethoscope

 doi:10.20944/preprints202508.0454.v1

Article

Deep Learning for Heart Sound Abnormality of Infants: Proof of Conceptual Study of 1D and 2D Representations

Eashita Wazed [1], Jimin Lee [2] and Hieyong Jeong [1,*]

[1] Department of Artificial Intelligence Convergence, Chonnam National University
[2] SW Convergence Education Institute, Chosun University
* Correspondence: h.jeong@jnu.ac.kr; Tel.: +82-062-530-3427

Abstract

Timely diagnosis and treatment of Congenital Heart Defects (CHDs) in pediatric patients are critical, as approximately 1% of neonates worldwide are affected by these anomalies. Traditional stethoscope auscultation relies on the clinician's skills, which can lead to missed subtle symptoms. This study introduces a deep-learning framework for the early diagnosis of congenital heart disease, utilizing time-series data from cardiac auditory signals captured via stethoscopes. The audio signals were transformed using Mel Frequency Cepstral Coefficients (MFCC) to create time-frequency representations. Our novel architecture combines Convolutional Neural Networks (CNN) for feature extraction with Long Short-Term Memory (LSTM) networks to capture temporal dependencies. This model achieved an impressive accuracy of 98.91% in early disease detection. While methods such as Electrocardiograms (ECG) and Phonocardiograms (PCG) are necessary for confirming diagnoses, previous AI-driven studies have largely focused on ECG and PCG datasets. Our approach emphasizes the potential of using cardiac acoustics for the early diagnosis of CHDs, enhancing clinical outcomes for infants.

Keywords: audio data processing; audio feature extraction; deep learning in cardiology; heart sound; heart function in infant; stethoscope.

1. Introduction

Research and clinical findings indicate that heart disease can be inherited through genetic predispositions from parents [1]. Various genetic disorders may contribute to the transmission of heart disease across generations within a familial lineage. Early heart disease detection and diagnosis are critical for improving patient outcomes and longevity [2]. It is not uncommon for pediatric patients to inherit cardiac conditions without parental knowledge, and due to their limited communication abilities, these issues often remain undiagnosed until adulthood.

The standard modalities for diagnosing heart disease typically include electrocardiograms (ECG) and echocardiograms (Echo), which can be resource-intensive in terms of time and cost. For this reason, ECG and Echo are widely used for diagnosing heart disease [3]. However, these examination methods are only available in specialized hospitals, are costly, and in many cases, ECG detects heart abnormalities only after the disease has significantly progressed. In contrast, heart murmurs, stenosis, and valvular insufficiencies can be easily detected through heart sounds. In recent years, substantial research has focused on analyzing heart sounds, particularly utilizing Phonocardiogram (PCG) signals [4].

Studies have explored various machine learning architectures, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. Additionally, methodologies involving Fast Fourier Transform (FFT) and mel-frequency cepstral Coefficient (MFCC) have been employed for feature extraction in these models. Future directions in this research area will focus on

 doi:10.20944/preprints202508.0454.v1

augmenting existing datasets, integrating a wider array of features, and referencing key works [1] ∼ [3,5]. Some studies have leveraged the Wavelet Transform (DWT) [2]. as a practical feature extraction technique and discussed the Short-Term Fourier Transform (STFT) [6] for signal analysis.

While numerous concise reviews exist on the application of deep learning models to ECG and PCG signals [7,8,10] ∼ [12] et al., there remains a notable gap concerning the analysis of heart sounds concerning audio data. The delayed diagnosis of myocardial infarctions often results in increased mortality rates due to the lack of timely medical intervention. This lag in treatment is frequently attributed to the high costs and labor-intensive nature of conventional ECG and Echo testing. Consequently, exploring the utilization of audio datasets presents a promising alternative in the quest for more accessible and efficient diagnostic solutions.

A stethoscope is a fundamental tool for assessing the cardiopulmonary health of neonates by detecting heart sounds. Figure 1 shows a concept of an AI-based assistance system that classifies whether a heart condition is normal or abnormal using heart audio-type data. Suppose it is possible to distinguish between normal and abnormal heart disease from a stethoscope that is always used for regular health checks in infants. In that case, heart disease can be helpful for early detection. ECG or PCG is a test performed when a detailed examination is necessary at the request of the medical doctor in charge, so it is possible when the disease has worsened. Clinicians leverage this auditory data to diagnose various cardiac abnormalities. However, relying on the practitioner's clinical acumen and experience can lead to potential oversights in early pathology detection, particularly in cases exhibiting subtle clinical manifestations.

This study aimed to develop an AI-driven support system that utilizes heart sound data to enhance diagnostic accuracy and expedite patient management in hospital settings. Electrocardiography (ECG) and Phonocardiography (PCG) are the most reliable modalities for diagnosing cardiovascular disorders. It is crucial to emphasize that our proposed methodology and findings are optimized for early detection, as the bulk of existing AI research in this domain predominantly utilizes ECG and PCG datasets for predictive modelling.

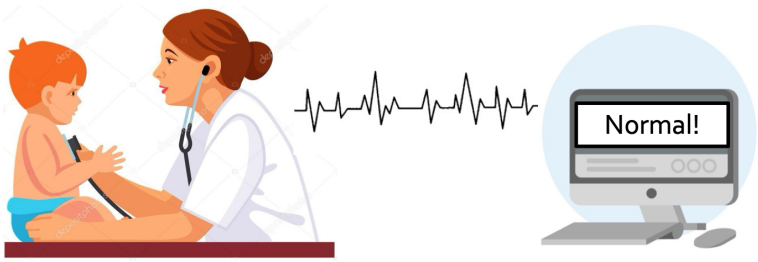


Figure 1. Concept of an AI-based assistance system that classifies whether heart condition is normal or abnormal using heart audio type data.

2. Related Works

Table 1 shows the previously proposed models on Heart Sound research. Electrocardiogram (ECG) and phonocardiogram (PCG)-based studies have been used in clinical settings for decades as reliable medical technologies for diagnosing heart diseases. ECG records the heart's electrical activity, making it practical for detecting arrhythmias, myocardial infarctions, and other electrical abnormalities, whereas PCG is useful for identifying physical abnormalities such as heart murmurs. Recent research has focused on leveraging deep learning and AI techniques to automate ECG and PCG data analysis and develop models for disease prediction.

Xiao developed an AI system combining CNN and LSTM to enable early detection of congenital heart disease (CHD) in children [1]. This study integrated IoT technology to facilitate remote analysis of heart sounds, demonstrating the feasibility of incorporating smart medical devices. Islam developed an electronic stethoscope and recorded children heart sound, they proposed SVM (Support Vector Machine) model to classify the data [2,9]. They did not describe the model architecture, number of parameters of the model, training cost etc. And SVM has many limitations, therefore we feel more better

 doi:10.20944/preprints202508.0454.v1

model should be proposed. Tao, Zihan developed a model that converts PCG signals into spectrogram images and applies Vision Transformer (ViT) and Convolutional Recurrent Neural Network (CRNN) to detect heart disease [3]. Lu, Kailong proposed an advanced method for predicting heart diseases by integrating temporal and textual information for more precise analysis [4]. Additionally, Wang introduced a hybrid model combining CNN and Transformer in a parallel structure to analyze heart sound data, applying frequency transformation techniques to improve accuracy over existing models [5]. However, ECG-based analysis faces real-time processing challenges and requires sensor attachment, which limits its use for immediate screening.

Table 1. Previously proposed models on heart sound research.

Reference	Year	Dataset	Age	Data type	Model	Accuracy [%]
Ref.[1]	2019	Pediatric	Children	Signal	CNN	96
Ref.[2]	2019	Mendelay	Children	**Audio**	SVM	**94.2**
Ref.[7]	2024	CirCor	Adult	PCG	CRNN	99.7
Ref.[8]	2023	PASCAL	Adult	PCG	RNN	90
Ref.[10]	2023	CirCor	Adult	PCG	CNN/LSTM	99
Ref.[11]	2012	Proposed	Adult	PCG	Diagnose	90
Ref.[14]	2010	-	Adult	ECG	MLP	93
Ref.[17]	2023	CirCor	Adult	PCG	CNN	91
Ref.[18]	2006	Proposed	Children	ECG	Analysis	93
Ref.[20]	2015	Proposed	Adult	PCG	KNN	93.3
Ref.[21]	2021	ECG	Adult	ECG	HMM	99
Ref.[19]	2001	Proposed	Children	Signal	ANN	100
Ref.[22]	2020	PhysioNet	Adult	ECG/PCG	HSMM	96
Ref.[24]	2024	Proposed	Adult	**Audio**	Pre-traind	**58.0**
Ref.[15]	2020	Not specified	Adult	ECG	CNN	97
Ref.[13]	2017	PhysioNet	Adult	PCG	CNN	83
Ref.[25]	2020	PhysioNet	Adult	PCG	CNN/MLP	98
Ref.[26]	2023	PhysioNet	Adult	PCG	CNN	71

PCG-based research provides a more cost-effective and accessible alternative compared to traditional ECG and echocardiography. Recent advancements focus on integrating AI techniques to automate heart sound analysis for early disease detection. Radha, Kodali developed a CRNN (Convolutional Recurrent Neural Network) model that directly analyzes raw heart sound data, outperforming conventional MFCC- and spectrogram-based approaches [7]. Habijan, Marija proposed a CNN-GRU (Gated Recurrent Unit) model for heart sound classification, utilizing various signal processing techniques to remove noise and enhance accuracy [8]. Nguyen, Minh Tuan introduced a CNN-LSTM model incorporating log-mel spectrograms, achieving higher performance compared to conventional time-frequency transformation methods [10]. Emmanuel described the signal processing technique for heart sound analysis in clinical diagnosis [11]. Deep learning-based computer-aided heart sound analysis in children has been introduced by Liu and Rubin [12,13]. Pediatric heart sound segmentation without using the ECG has been introduced by Sepehri and Li [14,15]. Deep learning framework based on spectrograms for heart sound classification is proposed by Chen [16,17]. Children's heart sounds at a distance with digital recordings has been proposed [18,19]. Many open access databases for the evaluation of heart sound, (16-18), are available but children heart sound dataset is only proposed by Mendeley [9]. Heart sound segmentation approach is also proposed [20] ∼ [24,25]. Chen proposed Log-Mel Spectrum Features [26]. Ren proposed Time and time-frequency features integrated CNN model [6].

However, several limitations remain in heart sound analysis. Noise and environmental factors can significantly affect accuracy, and diagnosing all heart diseases solely based on PCG signals remains a challenge. Additionally, PCG-based analysis still requires clinical interpretation, and to enhance diagnostic reliability, an automated system integrating smart stethoscopes and AI technology is essential. Such a system would enable heart disease detection without direct involvement of medical professionals, making PCG analysis more practical for real-world applications.

 doi:10.20944/preprints202508.0454.v1

3. Methods

3.1. Dataset

In this study, we utilized a meticulously curated heart sound dataset from pediatric patients at Khulna Shishu Hospital and Khulna Fortis Hospital. Heart sounds were recorded with a stethoscope under the supervision of a qualified pediatric cardiologist, ensuring high medical standards.

The recordings were averaged over six-second intervals, resulting in 60 distinct heart sounds: 30 associated with congenital and acquired cardiac anomalies (e.g., Ventricular Septal Defect, Atrial Septal Defect, Patent Ductus Arteriosus, Tetralogy of Fallot, Pulmonary Stenosis, Aortic Stenosis) and 30 categorized as normal. This normal baseline is critical for identifying deviations that may indicate cardiac disorders.

The dataset's classification relied on phonocardiographic (PCG) analysis to differentiate normal from abnormal heart sounds. Healthy PCG signals are characterized by the clear presence of the two fundamental heart sounds, S1 and S2. Abnormal signals, however, display distinctive traits, such as systolic murmurs at the left upper sternal border or variations in timing relative to S1 and S2. These nuances in heart sound patterns are essential for understanding pediatric cardiac health.

3.2. Architecture

Figure 2 presents a proposed architecture for classifying heart conditions from audio-typed data. The left depicts the workflow, while the right side details the processing involved. We employ three feature extraction techniques and compare their effectiveness alongside three deep learning models to determine the optimal approach for audio classification.

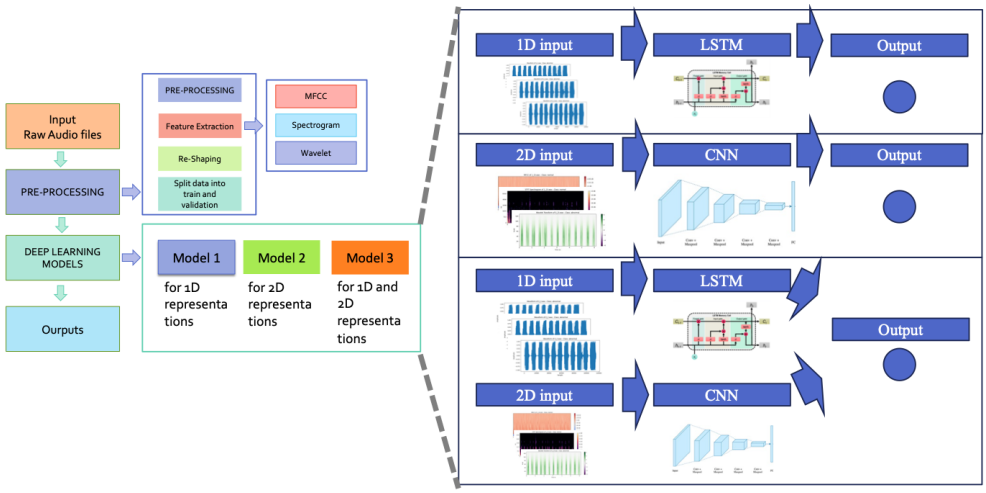


Figure 2. A proposed architecture for classifying cardiac conditions based on audio-type data: the left side represents the workflow, and the right represents detailed processing. We use three types of feature extractions and compare which feature extractor is suitable for the audio type. In addition, we apply three different deep learning models and compare the results to see which deep learning model is ideal for learning audio data types.

The first model utilizes recurrent neural networks, incorporating gated recurrent units (GRUs) and long short-term memory (LSTM) networks to capture temporal dependencies in raw signals. It also uses a 1D convolutional neural network (CNN) for analysis. The second model transforms raw audio data into spectrograms and employs a specialized 2D CNN for classification. The final model combines 1D and 2D data streams, assessing the advantages of multimodal deep learning for improved classification.

This study contributes significantly to applying deep learning in audio signal classification, providing valuable insights for future investigations.

 doi:10.20944/preprints202508.0454.v1

3.3. Feature Extraction

Figure 3 illustrates 1D and 2D representations of data, with (a) showing time-series data from audio signals and (b) to (d) depicting the transformed 2D images.

Cardiovascular diseases (CVD) are a significant cause of global mortality, necessitating improved diagnostic methods. Electrocardiograms (ECG) and heart sound analysis are essential non-invasive tools for assessing the heart's electrical and mechanical functions. However, manual interpretation of these signals, as seen in Figure 3(a), is complex and laborious, creating challenges for clinicians. Automated analysis systems offer a solution, providing faster and more accurate evaluations.

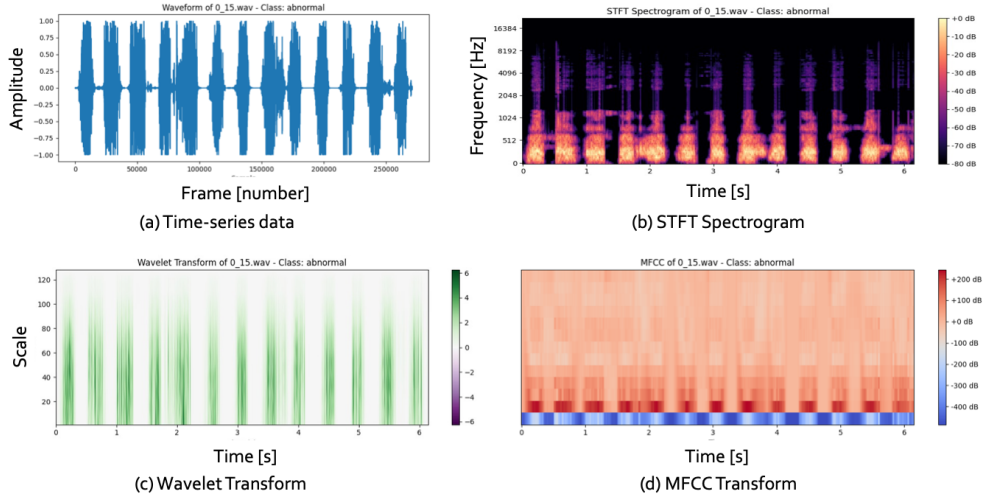


Figure 3. An example of 1D and 2D representations: the 1D representation in (a) represents time-series data, and the 2D representations in (b) ∼ (d) represent the results of transformation from 1D time-series data to the image.

Recent studies have focused on transforming 1D time-series data into 2D representations, as shown in Figure 3(b) to (d). Convolutional Neural Networks (CNNs) have emerged as powerful tools for image processing and have been employed for heart sound classification in CVD diagnostics. Common transformation methods include the Short Time Fourier Transform (STFT), the Wavelet transform, and the Mel-Frequency Cepstral Coefficient (MFCC), which result in 2D time-frequency spectrograms.

For example, Huang et al. developed a 2D CNN for classifying five arrhythmias from ECG data, achieving 99.00% accuracy compared to 90.93% for a traditional 1D CNN. Their findings suggest manual preprocessing techniques, such as signal filtering and feature selection, are unnecessary when using 2D CNNs for ECG classification.

3.4. Models

In our experimental framework, we implemented three distinct modelling approaches. The first utilized a Long Short-Term Memory (LSTM) architecture for anomaly detection in raw audio data, which yielded an accuracy of 66% on a dataset of children's heart sounds. The second model was based on a Convolutional Neural Network (CNN), employing Mel-Frequency Cepstral Coefficients (MFCC), Spectrogram, and Wavelet transforms for feature extraction. Our evaluation indicated that MFCC outperformed the other extraction techniques, as detailed in Table 2. The third approach integrated a hybrid architecture that combined CNN and LSTM, explicitly employing a 2D-CNN in conjunction with LSTM layers. This hybrid model, which we are particularly proud of, diverged from conventional architectures that mainly utilize 1D-CNNs, showcasing superior performance metrics compared to existing models in the relevant literature.

 doi:10.20944/preprints202508.0454.v1

3.4.1. Model 1 for 1D Representations

Our models are developed using Keras, leveraging TensorFlow as the backend. Model 1, as shown in Figure 2, employs a multi-layer LSTM architecture with three LSTM layers designed for classification tasks using three distinct window sizes. The input shape is configured at (64, 64, 3).

The architecture begins with an LSTM layer comprising 128 units, followed by a dropout layer set at a rate of 0.2 to address potential overfitting. The second layer mirrors the first, featuring another 128-unit LSTM and a dropout layer with the same 0.2 rate. The third LSTM layer maintains identical specifications, containing 128 units followed by a dropout layer at 0.2, before proceeding to fully connected layers.

We utilized the Adam optimizer for optimization, while the loss function employed during training was categorical cross-entropy. The model was trained for 50 epochs with a batch size of 32. The output layer incorporates a softmax activation function and includes two neurons, facilitating the classification of two distinct categories of children's heart sounds.

3.4.2. Model 2 for 2D Representations

The architecture of Model 2, meticulously designed for efficiency, consists of a series of convolutional layers followed by fully connected layers optimized for processing input images of size $128 \times 128 \times 3$. The model begins with a 2D convolutional layer that employs 32 filters, with a kernel size of 3 and a stride of 1. This is succeeded by a max pooling layer configured with a pool size of 2 and a stride of 2. Batch normalization is applied after max pooling to enhance the stability and speed of the training process.

The subsequent convolutional layer increases complexity by utilizing 64 filters, maintaining the kernel size of 3 and stride of 1. This is followed again by a max pooling layer and batch normalization. A third convolutional layer, featuring 128 filters and the same kernel and stride parameters, is introduced next, followed by another round of max pooling and batch normalization.

After processing through the convolutional stack, the resultant feature maps are flattened and fed into fully connected layers containing 128 units. ReLU activation functions across all convolutional layers is a key design choice, introducing non-linearity into the model and ensuring its robustness. A dropout layer, with a dropout rate of 50%, follows the fully connected layers to mitigate the risk of overfitting.

The Adam optimizer, a state-of-the-art choice for training deep learning models, is utilized for the training procedure. It is paired with the categorical cross-entropy loss function to gauge performance metrics. The model is trained over 100 epochs with a batch size of 32. A softmax activation function is applied to the output layer consisting of two units, enabling classification between the two classes of pediatric heart sounds.

3.4.3. Model 3 for 1D and 2D Representations

The architecture of Model 3 combines several convolutional layers and LSTM units, complemented by fully connected layers and dropout mechanisms, optimized for performance through rigorous training. It starts with a 2D convolutional layer featuring 32 filters with a 3×3 kernel, accepting input dimensions of (64, 64, 3) and utilizing ReLU activation. This is followed by a max-pooling layer with a pool size of 2 and a stride of 2, along with batch normalization to improve stability and convergence.

The model progresses to a second convolutional layer with 64 filters, maintaining a 3×3 kernel and a stride of 1, again followed by another max-pooling layer and batch normalization. The resulting feature maps are flattened before entering two sequential LSTM layers with 64 and 128 units, showcasing the model's advanced architecture. A dropout layer is applied post-LSTM with a 0.2 rate to mitigate overfitting. The architecture is completed with fully connected layers.

Training is executed using the Adam optimizer, with categorical cross-entropy as the loss function. The model was trained over 50 and 100 epochs with a batch size of 32. Finally, it outputs predictions

 doi:10.20944/preprints202508.0454.v1

through a softmax layer with two units, showcasing its ability to accurately distinguish between normal and abnormal heart sounds in pediatric patients, demonstrating its strong diagnostic potential.

3.5. Model Evaluation

To evaluate the classification strategies, we utilized ten-fold cross-validation. In each fold, we computed four key performance metrics: accuracy (ACC), sensitivity (Se), specificity (Sp), and modified accuracy (MAcc). These metrics were determined using the standard definitions:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Se = \frac{TP}{TP + FN} \tag{2}$$

$$Sp = \frac{TN}{FP + TN} \tag{3}$$

$$MAcc = \frac{Sp + Se}{2} \tag{4}$$

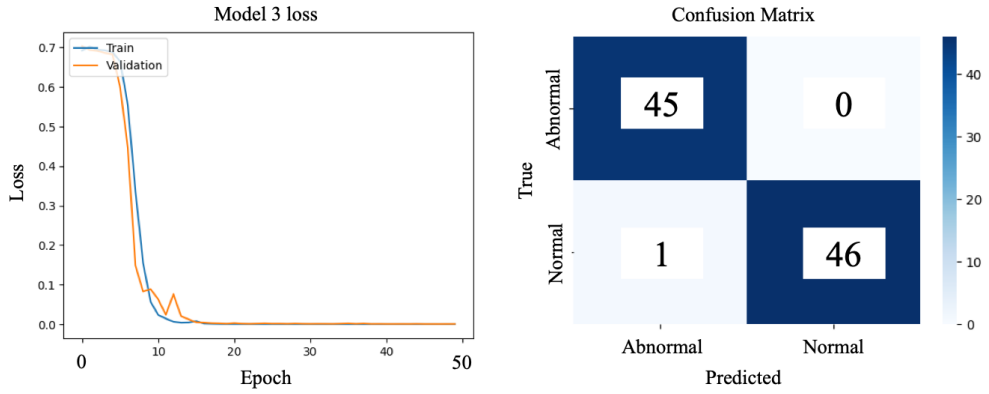
- TP (true positives): the number of patients correctly identified as patients.
- FN (false negatives): the patients incorrectly classified as healthy.
- FP (false positives): healthy individuals misclassified as patients.
- TN (true negatives): healthy subjects accurately classified as healthy.

After completing the cross-validation, we calculated the average values of these parameters across the 10 folds. The entire process was repeated 50 times, with each iteration randomly selecting a set of 218 normal signals to concatenate with 218 abnormal heart sounds. This random selection process ensures that our results are free from bias. The reported average results across these 50 iterations are presented in Section 4.

4. Results

4.1. Results of Performance

Figure 4 illustrates the learning outcomes of Model 3. In Figure 4(a), we observe the evolution of the loss metric, while Figure 4(b) presents the confusion matrix outcomes. Notably, the loss trajectory converges towards zero smoothly, indicating no adverse effects on learning mechanisms, such as overfitting. Performance evaluations reveal a high rate of accurate predictions across the test data, instilling confidence in the model's performance.



(a) Results of loss function of the train and validation for Model 3.

(b) Results of the confusion matrix of the train for Model 3.

Figure 4. Results of train and validation for Model 3: the plot (a) represents the loss function, and (b) represents the confusion matrix.

 doi:10.20944/preprints202508.0454.v1

Table 2 shows the comparison of our proposed model with recently introduced models. Initial analysis of the 1D representation—achieved by feeding raw time series data directly to the model—indicates superior performance from the two LSTM models operating in tandem compared to a sole LSTM implementation. Conversely, results from the third model utilizing SVM provide promising classification outcomes. However, it warrants caution regarding potential overfitting risks inherent with nonlinear classifiers. The standalone LSTM model with Model 1 exceeds the performance of the second LSTM model but does not match the aggregate efficacy of the first two LSTMs.

Table 2. Comparison of our proposed model with recently introduced models.

Representation	Models	Dataset	Data type	Accuracy [%]
1D representations	Ref.[8] LSTM, RNN	PASCAL	Audio	90.0
	Ref.[24] LSTM	Proposed	Audio	38.0
	Ref.[2] SVM	Mendelay	Audio	94.1
2D representations	Ref.[3] 2D Vit1D CRNN	PhysioNet	Audio	97.3
	Ref.[24] ResNet 50	Proposed	Audio	56.0
	Ref.[5] PCTMF-Net	PhysioNet	Audio	93.0
Our proposed models	Model 1 for 1D representations	Mendelay	Audio	**66.7**
	Model 2 for 2D representations	Mendelay	Audio	**91.7**
	Model 3 for 1D and 2D representations	Mendelay	Audio	**98.9**

Advancing to the 2D representation, where time series data is transformed into image formats for model input, we witness a significant enhancement in overall performance compared to the raw time series method. This improvement is primarily due to the meticulous and rigorous hyperparameter tuning and variations in the preprocessing techniques. Our approach with Model 2 involved three distinct preprocessors combined with straightforward CNN architectures, leading to significant performance disparities within the same model framework, as detailed in Table 3. This underscores the critical nature of thorough hyperparameter optimization and network architecture design, where strategies to mitigate overfitting and configurations of CNN and pooling layers profoundly influence the results.

Lastly, this study delves into the effectiveness of a multimodal learning strategy that independently trains and integrates two modalities. The results for Model 3 demonstrate enhanced performance metrics, such as higher accuracy and lower loss, over benchmarks from prior research. Furthermore, Model 3 consistently surpasses Models 1 and 2, maintaining robust and reliable performance under identical experimental conditions.

Table 3. Different feature extractor performance over the Model.

Feature Extractor	Precision	F1-score	Test Accuracy [%]
MFCC	(Ab)0.88 (N)0.80	(Ab)0.82 (N)0.84	**0.83**
Wavelet	(Ab)0.75 (N)0.70	(Ab)0.71 (N)0.74	0.72
STFT	(Ab)0.60 (N)0.62	(Ab)0.63 (N)0.59	0.61

4.2. Results of Cost for Training

Table 4 offers an in-depth comparative analysis of resource expenditure associated with training each model, conducted within a consistent environment that aligns with established research methodologies. This analysis employs our proposed evaluation metrics to ensure accuracy and relevance. Prior studies have shed light on specific performance indicators, yet they also highlighted notable gaps in parameter specifications that hindered comprehensive comparisons. In our examination, Model 1, Model 2, and Model 3 were meticulously evaluated under identical conditions to facilitate a valid comparison.

 doi:10.20944/preprints202508.0454.v1

Table 4. Comparison of our proposed model with recently introduced models with different evolution parameter.

Models	Acc	Sp	Se	MAcc	Number of Parameters
Ref.[3] 2D Vit-1D CRNN	0.9733	0.9731	0.9735	0.9733	-
Ref.[6] TTFI-CNN	0.9715	0.9713	0.9717	0.9715	-
Our Model 1	0.6666	0.5000	0.5000	0.5000	331,010
Our Model 2	0.9167	0.8333	**1.000**	0.9167	9,914,309
Our Model 3	**0.9891**	**0.9894**	0.9894	**0.9894**	**3,346,370**

Model 1, which utilizes a one-dimensional representation, consistently demonstrates lower performance across all evaluated metrics compared to its counterparts. However, it stands out for its remarkably reduced parameter count during training. This simplification correlates with significantly shorter training durations, making it highly efficient and capable of successful training even on hardware with limited specifications.

In contrast, Model 2 employs a two-dimensional representation and delivers a balanced performance that matches contemporary studies. Despite this, the lack of prior research outlining its parameter requirements leaves us with an incomplete understanding of its scalability. It is worth noting that Model 2's parameter count is approximately thirty times greater than that of Model 1, representing a substantial increase in resource demands, albeit in exchange for enhanced performance levels.

Model 3, utilizing a multi-modal approach, achieves a slight performance boost compared to earlier research findings. Interestingly, the parameter count for Model 3 is reduced by about one-third compared to Model 2. This noteworthy reduction suggests a potential decrease in computational burden while still maintaining competitive performance capabilities, paving the way for innovative applications in the future.

These findings illuminate a crucial insight: direct modelling of time series data does not automatically translate into superior outcomes. However, using image-converted data for training can significantly enhance performance metrics. Furthermore, the multi-modal learning framework adeptly minimizes the computational costs associated with training, thereby presenting a promising direction for future developments in the field.

5. Discussion

A heart murmur, an auscultatory finding indicative of turbulent blood flow within the cardiovascular system, is a common occurrence in the pediatric population during outpatient evaluations. It is the leading cause of referrals from primary care to pediatric cardiology clinics. The prevalence of asymptomatic heart murmurs in children is significant, ranging from 24% to 97.5%, with a peak incidence observed in the 8 to 12-year age group. In newborns, the incidence is reported to be approximately 40-50 per 1,000 live births, while the prevalence in school-aged children and adolescents is estimated at 75-80%.

It's crucial to understand that only a subset of these murmurs is associated with structural heart disease; most are classified as innocuous. The responsibility of accurate differentiation between pathological murmurs, which may indicate underlying heart disease, and benign murmurs—commonly observed in healthy children—is significant during physical examination. Despite the growing reliance on echocardiography, which can diminish the necessity for auscultation, proficiency in evaluating heart murmurs remains vital. Clinicians must ascertain whether a patient should be referred to a pediatric cardiologist for further assessment or if the likelihood of significant heart disease is minimal, thus reassuring the patient and guardians.

Heart murmurs can frequently be detected in pediatric patients. While many are non-pathological, they can occasionally serve as the sole indicator of severe cardiac pathology, necessitating a careful and urgent approach. Murmurs classified as pathological typically present with specific characteristics: they may occur during systole or diastole, exhibit grade III intensity or more significant, have a coarse quality, feature an abnormal second heart sound (S2) or systolic click, or are exacerbated by postural

 doi:10.20944/preprints202508.0454.v1

changes such as standing. In instances where a pathological murmur is suspected, immediate further evaluation with echocardiography is indicated to confirm the presence of any underlying cardiac abnormalities.

While auscultation findings are paramount, external factors such as clinic congestion, physician fatigue, and ambient noise can compromise diagnostic accuracy. Given the critical decision-making in determining the necessity for echocardiographic evaluation, the auxiliary system proposed in this study will enhance clinicians' ability to make timely and accurate assessments.

6. Limitations

This study introduces an AI-driven heart disease prediction model leveraging an intelligent stethoscope. However, it's important to note that the model's performance may be compromised due to the constrained dataset used for training. This could potentially lead to a lack of generalizability across diverse patient populations. A dataset skewed towards a specific demographic may adversely affect real-world clinical applicability. To address this, future research will involve the development of a comprehensive, large-scale dataset encompassing a wide range of age groups and pathological conditions. Importantly, we will seek multi-institutional collaborations to enhance the robustness and reliability of the model, recognizing the value of collective efforts in improving healthcare technology.

Another significant limitation is the AI model's insufficient interpretability, which challenges clinician confidence in its outputs. Deep learning algorithms' black-box nature renders their diagnostic reasoning opaque. Consequently, forthcoming studies will employ interpretability frameworks such as Grad-CAM, SHAP, and LIME to generate visual insights into the AI system's decision-making pathways. Collaborating with medical professionals, including you, to refine these interpretability techniques will also be a priority to foster trust among end-users.

Furthermore, the absence of comparative analyses with established diagnostic modalities like ECG and echocardiography underscores a critical gap in assessing the clinical utility of the intelligent stethoscope. To address this, future investigations will systematically compare the performance of the AI-based analytical model with traditional diagnostic approaches. Rigorous experimentation with medical practitioners will be conducted to validate the model's efficacy in authentic clinical settings. Through these initiatives, the intelligent stethoscope aims not to replace but to complement conventional diagnostic processes, facilitating applications in remote healthcare and primary care environments.

7. Conclusions

In this study, we investigated the methodology for analyzing heart sound data to facilitate early detection of cardiovascular abnormalities in infants. We examined the efficacy of one-dimensional (1D) signal processing alongside two-dimensional (2D) image transformations applied to various heart sound signals, and we compared the classification performance of both image-based and signal-based deep learning (DL) models. Additionally, we assessed the validity and effectiveness of a multimodal fusion approach that integrates both 1D and 2D representations.

Our findings indicate that 1D heart sound signal networks exhibit a lower training cost than those based on 2D signal representations while outperforming the latter in classification tasks. This disparity in performance can be attributed to the 2D image transformation's ability to encapsulate richer information, enabling simultaneous learning of temporal dynamics and frequency components. We confirmed that multimodal fusion models can effectively reduce training costs while sustaining robust performance metrics.

In conclusion, our study substantiates the viability of a multimodal approach for the early diagnosis of cardiac abnormalities in infants. This approach, which integrates both 1D and 2D representations, holds significant promise for the future of early diagnosis in this vulnerable population. It contributes valuable insights to the existing literature. It paves the way for future advancements in

 doi:10.20944/preprints202508.0454.v1

heart sound signal processing techniques aimed at prompting the identification of cardiac diseases in infants.

Author Contributions: Conceptualization, J.L. and H.J.; methodology, E.W.; software, E.W.; validation, E.W., J.L. and H.J.; formal analysis, E.W.; investigation, H.J.; resources, H.J.; data curation, E.W.; writing—original draft preparation, E.W.; writing—review and editing, H.J.; visualization, E.W.; supervision, H.J.; project administration, H.J.; funding acquisition, H.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea grant, funded by the Ministry of Education (No. RS-2021-NR066151), the Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government(MSIT), and the Institute of Civil-Military Technology Cooperation, funded by the Defense Acquisition Program Administration and the Ministry of Trade, Industry and Energy of the Korean government under grant No. 23-CM-DI-11.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the public dataset.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is a public dataset that anyone can download and use for model training.

Acknowledgments: This research was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea grant, funded by the Ministry of Education (No. RS-2021-NR066151), the Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government(MSIT), and the Institute of Civil-Military Technology Cooperation, funded by the Defense Acquisition Program Administration and the Ministry of Trade, Industry and Energy of the Korean government under grant No. 23-CM-DI-11.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MFCC	Mel Frequency Cepstral Coefficients
ECG	Electrocardiograms
PCG	Phonocardiograms
CHD	Congenital Heart Disease
LSTM	Long Short-Term Memory
CNN	Convolutional Neural Networ
FFT	Fast Fourier Transform
DWT	Discrete Wavelet Transform
STFT	Short-Term Fourier Transform
CVD	Cardiovascular diseases
MAcc	Mean Accuracy

References

1. Xiao, B; Xiuli Bi, Y. X.; Li, W.; Ma, Z.; Zhang, J.; Ma, X. Follow the Sound of Children's Heart: A Deep-Learning-Based Computer-Aided Pediatric CHDs Diagnosis System. *IEEE Internet of Things Journal* **2020**, 7(3), 1994–2004.
2. Islam, R.; Hassan, M.; Raihan, M.; Datto, S. K.; Shahriar, A.; More, A. A Wireless Electronic Stethoscope to Classify Children Heart Sound Abnormalities. *22nd International Conference on Computer and Information Technology (ICCIT)* **2019**, 18–20.

 doi:10.20944/preprints202508.0454.v1

3. Tao, Z.; Ren, Z.; Yang, X.; Liang, Y.; Shi, X. 2D ViT and 1D CRNN-based Heart Sound Signals Detection Model. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* **2023**, 05–08.
4. Lu, K.; Zhao, F.; Gu, P.; Wang, H.; Zang, T.; Wang, H. TetraCVD: A Temporal-Textual Transformer based Model for Cardiovascular Disease Diagnosis. *IEEE International Conference on Bioinformatics and Biomedicine* **2023**, 2129–2132.
5. Wang, R.; Duan, Y.; Li, Y.; Zheng, D.; Liu, X.; Lam, C. T.; Tan, T. PCTMF-Net: heart sound classification with parallel CNNs-transformer and second-order spectral analysis. *Journal The Visual Computer,Springer* **2023**, 39(8), 3811—3822.
6. Ren, Z.; Qiao, Y.; Yuan, Y.; Zhou, Y.; Liang, Y.; & Shi, X. Time and time-frequency features integrated cnn model for heart sound signals detection. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM).* **2022**, 1138–1143.
7. Radha, K.; Bansal, M.; Sharma, R. Raw Waveform-Based Custom Scalogram CRNN in Cardiac Abnormality Diagnosis *IEEE Access* **2024**, 12, 13986–14004.
8. Habijan, M.; Galić, I.; Pizurica, A. Heart Sound Classification using Deep Learning. *IEEE 8th International Conference on Smart and Sustainable Technologies (SpliTech)* **2023**, 1—6.
9. Islam, Md Riadul; rakib , mahadi hassan ; Raihan, M. Children Heart Sound- Normal & Abnormal. *Mendeley Data* **2023**, v1.
10. Nguyen, M. T.; Lin, W. W.; Huang, J. H. Heart sound classification using deep learning techniques based on log-mel spectrogram. *Journal Circuits, Systems, and Signal Processing* **2023**, 42(1), 344–360.
11. Emmanuel, B. S.A Review of signal processing techniques for heart sound analysis in clinical diagnosis. *Journal of medical engineering & technology* **2012**, 36(6), 303–307.
12. Liu, J.; Wang, H.; Yang, Z.; Quan, J.; Liu, L.; Tian, J. Deep learning-based computer-aided heart sound analysis in children with left-to-right shunt congenital heart disease.*International Journal of Cardiology* **2022**,348, 58–64.
13. Rubin, J.; Abreu, R.; Ganguli, A., Nelaturi; S., Matei, I.; Sricharan, K. Recognizing abnormal heart sounds using deep learning. *journal arXiv preprint arXiv:1707* **2017**, 04642.
14. Sepehri, A. A.; Gharehbaghi, A.; Dutoit, T.; Kocharian, A.;Kiani, A. A. Novel method for pediatric heart sound segmentation without using the ECG. *Computer methods and programs in biomedicine* **2010**, 99(1), 43–48.
15. Li, Y.; Pang, Y.; Wang, K.; Li, X. Toward improving ECG biometric identification using cascaded convolutional neural networks. *journal Neurocomputing, Elsevier* **2020**, 391, 83–95.
16. Zhao, Q.; Geng, S.; Wang, B.; Sun, Y.; Nie, W.; Bai, B.; Yu, C.; Zhang, F.; Tang, G.; Zhang, D.; Zhou, Y.; Liu, J.; Hong, S. Deep Learning for Heart Sound Analysis: A Literature Review. *Journal medRxiv, Cold Spring Harbor Laboratory Press)* **2023**, 09.
17. Chen, J.; Guo, Z.,;Xu, X.; Zhang, L. B.; Teng, Y.; Chen, Y.; More; Wang, W. A robust deep learning framework based on spectrograms for heart sound classification .*IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2023**, 21(4), 936–947.
18. Finley, J. P.; Warren, A. E.; Sharratt, G. P.; Amit, M. Assessing children's heart sounds at a distance with digital recordings. *journal Pediatrics,American Academy of Pediatrics)* **2006**, 118(6), 2322–2325.
19. DeGroff, C. G.; Bhatikar, S.; Hertzberg, J.; Shandas, R.; Valdes-Cruz L.; Mahajan, R. L.Artificial neural network–based method of screening heart murmurs in children. *journal Circulation, Am Heart Assoc* **2023**, 103(22), 2711–2716.
20. Randhawa, S. K.;Singh, M. Classification of heart sound signals using multi-modal features .*journal Procedia Computer Science,Elsevier* **2023**, 58, 165–171.
21. Altuve, M.;Monroy, N. F. Hidden Markov model-based heartbeat detector using electrocardiogram and arterial pressure signals .*Biomedical Engineering Letters, Springer* **2021**, 11(3), 249–261.
22. Shukla, S.; Singh, S. K.; Mitra, D. An efficient heart sound segmenta tion approach using kurtosis and zero frequency filter features. *Biomedical Signal Processing and Control* **2020**, 57, 101762.
23. Sound, M. H. Murmur Library. *University of Michigan Heart Sound and Murmur Library* **2012**.
24. Qian, K.; Bao, Z.; Zhao, Z.; Koike, T.; Dong, F.; Schmitt, M.;More; Yamamoto, Y. Learning Representations from Heart Sound: A Comparative Study on Shallow and Deep Models. *Cyborg and Bionic Systems, AAAS* **2024**, 5, 0075.
25. Dissanayake, T.; Fernando, T.; Denman, S.; Sridharan, S.; Ghaemmaghami, H.; & Fookes, C. Understanding the importance of heart sound segmentation for heart anomaly detection. *Clinton, journal arXiv preprint arXiv:2005* **2020**, 10480.
26. Chen, W.; Zhou, Z.; Bao, J.; Wang, C.; Chen, H.; Xu, C.; Wu, H. Classifying heart-sound signals based on cnn trained on melspectrum and log-melspectrum features. *Journal Bioengineering* **2023**, 10(6), 645.

 doi:10.20944/preprints202508.0454.v1

27. Liu, C.; Springer, D.; Li, Q.; Moody, B.; Juan, R. A.; Chorro, F. J.; Castells, F.; Roig, J. M.; Silva, I.; Johnson, A.; Syed, Z.; Schmidt, S. E.; Papadanill, C. D.; Hadjiileontiadis, L.; Naseri, H.; Moukadem, A.; Dieterlen, A.; Brandt, C.; Tang, H.; Samieinasab, M.; Samieinasab, M. R.; Sameni, R.; Mark, R. G.; Clifford, G. D. An open access database for the evaluation of heart sound algorithms.*journal Physiological measurement* **2016**, 37(12), 2181.
28. PhysioBank, P.Physionet: components of a new research resource for complex physiologic signals. *journal Circulation* **2000**, 101(23), e215–e220.
29. Chakir, F.; Jilbab, A.; Nacir, C.; Hammouch, A. Phonocardiogram signals processing approach for PASCAL classifying heart sounds challenge. *Signal, Image and Video Processing,Springer* **2018**, 12(6), 1149–1155.