Congenital Heart Disease Detection from Children's Heart Sounds using Vision Transformer-based Model

by

Damir Kabdualiyev

Submitted to the Department of Data Science
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

June 2025

Author .....
Department of Data Science
1

Certified by .....
Khalil Khan
Professor
Thesis Supervisor

Accepted by .....
Elizabeth Arkhangelsky
Dean, School of Engineering and Digital Sciences fix me

Congenital Heart Disease Detection from Children's Heart Sounds using Vision Transformer-based Model

by

Damir Kabdualiyev

Submitted to the Department of Data Science
on 1, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

Congenital heart disease (CHD) is one of the most common birth defects and has been a source of great morbidity and mortality among infants. Early diagnosis allows for early intervention and better patient outcomes. Traditional methods are effective yet have limitations such as high costs, dependency on expert knowledge, noise, etc. Recent developments in deep learning (DL) showcase the potential to address these limitations and automate diagnosis with high accuracies. This study proposes the use of a hybrid model of Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs) for the detection of CHD based on heart sounds. The proposed CNN-ViT-based architecture leverages both local feature extraction by CNN and global pattern modeling by ViT. Thus, the model is able to capture both local and global dependencies in heart sounds' spectrogram representations. The proposed approach enhances the accuracy of CHD detection by overcoming some of the limitations arising from traditional approaches and CNN based methods. This study utilized a recently released open-source ZCHSound dataset [1] of children's pediatric heart sound recordings. To evaluate this model we use accuracy as evaluation measure. The model showed exceptional performance on the binary the classification the task of clean heart, sounds, reaching an accuracy of 95% and adequate results for multi-class classification (74% accuracy). However, for multi-class classification, results are not convencing and some more research work will be needed. Nonetheless, the proposed approach set a new benchmark in CHD diagnostics, contributing to AI-driven healthcare solutions and furthering the application of deep learning in medical research.

Thesis Supervisor: Khalil Khan
Title: Professor

Acknowledgments

I would like to express my gratitude to my advisor, Prof. Khan for his guidance, support and feedback throughout the course of this Thesis. Also, I want to thank my family for their love, moral support and belief in me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Congenital heart disease (CHD) is the most common birth defect associated with a structural malformation of a heart or great vessels [3]. According to the World Health Organization (WHO), 60,000 infants are born with CHD. Some resources report that 1% of newborn children have this abnormality [4]. In 2017, a global estimate of CHD prevalence reached 11,998,283 people, which corresponds to an age-standardized prevalence rate of 170.6 cases per 100,000 population [3]. The highest incidence rates of CHD were observed in African or Asian countries (Burundi, Somalia, etc.) with low income where the rates reached over 30 cases out of 1,000 newborns and high-income countries (France, Portugal, Qatar) possess low incidence rates of CHD with less than 10 cases out of 1,000 [3].

Furthermore, this disease remains one of the leading causes of infant mortality due to birth defects. Approximately 4.5% of children with CHD die in utero, and 21% dies at or after birth. Survivals have a 9-fold increased risk of some disability [4]. Furthermore, CHD might be assymptomatic but cause mortality only after some time [4]. More than 97% of children with CHD survive beyond the age of 18 years, but they have 3-fold risk of dying before the age of 68 years than those without the abnormality [3].

Early detection of the heart defect might not only improve treatment but also reduce risks of disease progressions that lead to complications including lethal cases [5][4]. Thus, timely diagnosis of CHD is important.

There are different methods with the help of which CHD can be diagnozed. Some methods that have been reported in litrature and used by medical practitioners are Echocardiography (EKG)[6], Pulse Oximeter (POX), Chest X-Rays [7][8], Electrocardiography (ECG), Phonocardiography (PCG). However, they all possess some challenges, like equipment quality, knowledge base of a doctor, or noises in images or ultrasounds (US) [6].

Echocardiography is a screening method that uses ultrasounds to form a picture of the tissues and organs in a human chest. The formed picture helps seeing parts of the heart for defect analysis [9]. This is widely used for pre-natal [6] analysis, as well as for post-natal analysis [10]. While it is an effective tool, EKG is limited in terms of availability, high cost [10] and the need in specialized ultrasound specialist [7].

Electrocardiography record heart's electrical activity from the body surface and represent it with respect to time on a paper. While it provides important insights for CHD detection, it is also advantageous due to its cost-effectiveness and high efficiency [11].

Pulse oximetry measures blood's oxygen levels with pulse [12]. POX is cost-effective method for identifying critical CHD in newborns. Nonetheless, it is recommended to use it along with other screening techniques to ensure higher accuracy [10].

Despite that more advanced diagnostic imaging methods are available today, chest x-rays provide an easy, simple and cheap way to get essential information for the first step in identifying a heart disease and tracking its development [13].

Phonocardiography records heart sounds and is a cost-effective first-line screening method. However, identifying CHD from heart sounds required experienced physician knowledge [1].

While all screening methods possess some drawbacks and benefits in terms of accuracy, cost-effectiveness, availability and expertise, it is difficult to mark one as the best method. Among these, developing a DL model for screening assistance in resource-limited settings might reduce reliance on specialized expertise, lower cost and increase diagnostic speed; thus, expanding access to early detection.

Thus, this study aims at developing DL framework for CHD detection and classification using medical data such as phonocardiograms (PCGs). The goal is to improve the accuracy and efficiency of CHD diagnosis by leveraging advanced DL algorithms such as ViTs. Thereby, this study aims to address existing challenges in CHD detection and contribute to the broader field of AI-driven healthcare solutions.

Thus, this study aims to:

1. Develop a DL framework for CHD classification using PCG data.
2. Evaluate the performance of ViT-based architectures in detecting CHD from PCGs, and compare their performance with existing solutions.
3. Explore the potential of ViTs in addressing the challenges of CHD diagnosis.
4. Provide a scalable and automated solution to enhance the early detection of CHD in pediatric patients.

The structure of this study is organized as follows: In Section II, a detailed literature review will explore previous research on CHD diagnosis, existing diagnostic methods, and relevant DL approaches. Section III presents the Methodology: the dataset, preprocessing techniques, and a DL model employed for CHD classification. Section IV details the Results we obtained during this study. Finally, Section V provides Conclusion and Future Work.

1.1 Motivation

Early and accurate detection of CHD may significantly improve treatment outcomes, reduce complications, as well as enhance quality of life for significant subsets of the population. Thus, it is important to explore innovative approaches for automated diagnosis. Vision Transformers are worth investigating because they might outperform traditional methods by capturing complex patterns and long-range dependencies in PCG recordings, thereby providing explainability, which is important in decision-making or assistance.

Chapter 2

Related works

CHD can be diagnosed using different methods including Pulse Oximeter, Magnetic Resonance Imaging, ECG, EKG, etc. Each modality provides unique advantages from cost-effectiveness in Pulse Oximeter to detailed anatomical structures in CT scans. On the other hand, some diagnosis types possess some challenges and disadvantages such as need for expert knowledge, cost, or ineffectiveness in CHD types classification. Recently, machine learning (ML) has enhanced the accuracy and efficiency of CHD detection methods. In the following paragraphs each of these methods and then the corresponding work reported in litratures has been discussed.

2.0.1 CT Scans

Ecabert et al. [14] introduced one of the earliest model-based method for heart segmentation in CT scans. It comprised a 3D Generalized Hough Transform (GHT) for initial localization, piecewise affine transformation and deformable adaptation. This method achieved a mean surface-to-surface error of 0.82 mm which showed high accuracy for segmenting CT scans of a heart into its major components namely four chambers, myocardium and major vessels.

Later, Khan et al. [4] presented the Cardiac Deep Learning Model (CDLM) that employed a 3D U-Net architecture with graph matching for vessels classification. Overall, CDLM showed effectiveness in heart segmentation into chambers, blood pool

and vessels. CDLM demonstrated notable accuracy in aortic segmentation with high mean Dice scores across heart structures, surpassing earlier methods.

While CT scans may seem promising in CHD detection, collecting data for traning and affordability of equipment may suggest that it may not be the most effective way for heart defect classification.

2.0.2 Echocardiograms

In [15], researchers explored AI-assisted auscultation (AI-AA) for CHD detection using heart sound recordings. The model used a CNN to classify time-frequency representations of heart sounds and achieved a sensitivity of 97%, specificity of 89%, and an overall accuracy of 96%. This method aligns closely with traditional auscultation, offering accessible and effective remote CHD screening in resource-limited areas.

While the study mentioned above [15] focused on babies' heart sounds, Qiao et al. [6] introduced a novel model for fetal screening. The model integrated a Multi-Scale Gated Axial-Transformer Network (MSGATNet) along with residual learning to segment and parse the four chambers of the fetal heart in ultrasound images. The proposed method addressed such challenges of fetal screening as movements receiving precision of 95.92%, recall of 94%, and overall accuracy of 95%, surpassing 12 state-of-the-art models in fetal CHD detection. This demonstrated the potential of transformer-based networks in refining prenatal CHD diagnosis.

2.0.3 Electrocardiograms

Authors, in [11], developed CHDdECG, a DL model which integrates raw ECG waveforms with wavelet-transformed and human-concept features. The model achieved ROC-AUC scores of 0.915 (for internal tests) and 0.917/0.907 (external test sets). This model got a specificity of 0.881 on the internal test set, which might outperform traditional diagnostic approaches.

However, Du et al. [16] proposed a Residual of Residual (RoR) Network for CHD detection from ECG signals. The RoR model reached an accuracy of 92.45%,

sensitivity of 74.73%, and specificity of 94.07% using 10-second ECG segments, with potential for increased sensitivity through longer recordings.

2.0.4 Phonocardiograms

Hassanuzzaman et al. [17] proposed a DL model that integrates a one-dimensional CNN with an attention transformer that accepts raw PCG signals for the binary classification of CHD and non-CHD. They utilized a dataset of 484 patient recordings collected in clinical settings. Their model achieved promising results with an accuracy of 92.3%, a sensitivity of 98.3%, a specificity of 83.3%, f1-score of 93.9% and an AUC of 0.964.

In [18], authors introduced two lightweight CNN-based models for CHD detection from pediatric PCG data. The first model was dense-block architecture and the other one was clique-based architecture. Both models were utilizing raw PCG signals. This study used a pediatric heart sound dataset consisting of 528 recordings. The dense-block model showcased superior performance with accuracy of 96.21%, specificity of 98.08%, and precision of 91.67% and 98.77% for CHD and non-CHD cases respectively. The clique-based model showed worse, but still promising performance with accuracy of 91.68%, specificity of 89.17%, sensitivity of 93.40% and precision of 88.96% and 93.66% for CHD and non-CHD cases subsequently. Authors also revealed that both models do not required high-computational demands and thus can be used for real-time CHD diagnosis via IoT-enabled healthcare systems.

Yadav et al. [19] proposed an approach for CHD classification from PCG signal by leveraging Fourier Transform and Cepctrum analysis. The method employed down-sampling of recorded signals from 2000 Hz to 1000 Hz for faster processing and bandpass filtering to reduce noise. Then, a Fourier transform was performed for frequency domain analysis. The ratio of the mean and maximum values of the Fourier transform signal was used as the discrimination between CHD and non-CHD cases. Additionally, they used the mean values of cepstrum coefficients extracted from heart sounds. Authors utilized the 2016 PhysioNet/CinC Challenge database with a total of 100 heart sound recordings. A Support Vector Machine (SVM) was then trained

and tested on these statistical features. The classifier achieved an accuracy of 95% with sensitivity and specificity of 100% and 90% respectively.

Authors in [1] address the critical issue of accessible PCG datasets in this research area. They present a comprehensive effort to create a large-scale high-quality open-source dataset of PCG recordings for research purposes. The total number of participants is 1259 in the data collection process. This data was collected from 3 children's hospitals between 2020 and 2022. Participants ages range from newborns up to 14 years. To ensure correct diagnosis labellings, 2 cardiac experts reviewed the following heart sounds based on their ECG results, which provide more reliable diagnostic information for decision-making. They also aimed at establishing benchmark classification results with simple ML classifiers and hand-crafted features. The study extracted 84 time-domain and frequency-domain features with the usual 60-20-20 dataset split. The authors defined two main tasks, with two subtasks for each:

Table 2.1: Task Description Table [1]

Task Number	**Description**
Task 1-1	Binary classification of CHD vs. non-CHD in heart sounds.
Task 1-2	Multi-class classification of non-CHD vs. CHD subtypes in heart sounds.
Task 2-1	Binary classification of CHD vs. non-CHD in *low-quality* heart sounds.
Task 2-2	Multi-class classification of non-CHD vs. CHD subtypes in *low-quality* heart sounds.

Four different classifiers were evaluated:

1. Random Forest (RF)

2. Support Vector Machine (SVM)

3. K-Nearest Neighbors (KNN)

4. AdaBoost

The RF classifier achieved best results for high-quality heart sounds classification achieving accuracy of 90.3% for Task 1-1 and 93.4% for Task 1-2. For Tasks 2-1 and 2-2, KNN reached accuracy of 62.5% and 55.7% respectively (see Table 2.2).

Table 2.2: Best Classifiers' Results by [1]

Model	**Task**	**ACC**	**SE**	**SP**	**F1**
RF	1-1	0.903	0.876	0.924	0.903
RF	1-2	0.934	0.555	1.000	0.695
KNN	2-1	0.625	0.562	0.687	0.623
KNN	2-2	0.557	0.250	0.750	0.466

2.0.5 Vision Transformers for Heart Defect Analysis

Over the recent years, Vision Transformers have been steadily emerging as a likely alternative to CNNs not only in image classifications [20][21][22], but also in most current medical imaging and signal processing tasks. Apart from CNNs, ViTs are known for capturing long-range dependencies and global contexts which are quite critical in interpreting complex patterns across medical data.

As such, the study of Vaid et al. [23] has demonstrated that ViTs had higher performance in identifying heart diseases from ECG recordings with fewer datasets available than traditional models. They also found that ViTs provided explainability of a decision by highlighting the most important information for defect identification [23].

As such, Dong et al., [24], proposed a ViT-based model with deformable attention for the classification of arrhythmias. As expected, it outperformed traditional CNN-based methods. The results presented therein motivated us to explore ViT-based architectures for detecting CHD from heart sounds.

Table 2.3: Summary of Existing CHD Detection Methods and Results

Paper	**Data Modality**	**Methodology**	**Results**
[4]	CT Images	CDLM	Accuracy: 83.8%
[8]	Chest X-ray	Canny Edge + DSS	Accuracy: 93.3%
[15]	Heart Sounds	CNN	Sensitivity: 97%, Specificity: 89%, Accuracy: 96% ($\kappa = 0.84$)
[6]	Fetal Echocardiogram	SPReCHD	F1 Score: 94.95%
[16]	Electrocardiogram	RoR	Accuracy: 92.45%, Sensitivity: 74.73%, Specificity: 94.07%
[11]	Electrocardiogram	CHDdECG	Specificity: 88.1% to 93.7%
[17]	Phonocardiogram	CNN + Attention Transformer	Accuracy: 92.3%, Sensitivity: 98.3%, Specificity: 83.3%, AUC: 0.964
[18]	Phonocardiogram	Dense-block CNN	Accuracy: 96.21%, Specificity: 98.08%, Precision: 91.67% (CHD)
[19]	Phonocardiogram	Fourier Transform + Cepstrum + SVM	Accuracy: 95%, Sensitivity: 100%, Specificity: 90%
[1]	Phonocardiogram	Random Forest (RF)	Accuracy (Binary): 90.3%, Accuracy (Multi-class): 93.4%

Chapter 3

Methodology

3.0.1 Dataset

The study utilizes the ZCHSound dataset published in 2024 [1]. This is a clean and open-source pediatric heart sound database specially developed for the research of CHD. Data were prospectively obtained from 1259 patients in three of China's specialized children's hospitals between 2020-2022. All heart sound recordings were taken using a standard ChildCare G-100 smart stethoscope at an 8000 Hz sampling rate, and each recording was 11-30 seconds in duration and stored in standard .wav format. To provide reliability and consistency, all recordings were conducted on subjects in a supine position and underwent a stringent quality control procedure; trained clinicians manually screened the data to form two distinct subsets—a high-quality set of 941 clean recordings (both normal and various cases of CHD such as atrial septal defect (ASD), ventricular septal defect (VSD), patent ductus arteriosus (PDA), and patent foramen ovale (PFO)) and a low-quality set of 318 noisy recordings from neonates (see Table 3.1). Furthermore, confirmatory cardiac ultrasound examinations and professional opinions were employed to strictly allocate diagnostic labels, thus providing an accurate foundation upon which to construct and test wise auscultation algorithms. The age summary for clean and noisy audio recordings indicates that the mean age for the clean dataset is 3.16 years, while the mean age for the noisy dataset is 1.43 days (see Table 3.2). The gender distribution is adequately balanced both for

high-quality and low-quality sets (see Table 3.3).

Table 3.1: Diagnosis counts for clean and noisy data.

Diagnosis	**Clean Data Count**	**Noise Data Count**
NORMAL	533	160
VSD	187	14
ASD	119	102
PFO	70	35
PDA	32	7
Total Normal	533	160
Total Abnormal	408	158

Table 3.2: Age Summary for Clean Data (in years) and Noise Data (in days).

Statistic	**Clean Data (years)**	**Noise Data (days)**
Mean	3.16	1.43
Minimum	0.01	0.00
Maximum	14.57	6.00
Median	2.25	1.00

Table 3.3: Gender distribution in Clean and Noise datasets.

Gender	**Clean Data Count**	**Noise Data Count**
Male	468	166
Female	473	152

3.0.2 Tasks

Since the data is divided into different subsets, this study was split into 4 tasks.

- Task 1 is binary classification of high-quality heart sound recordings. For identifying the absence or presence of CHD (CHD vs. non-CHD).

- Task 2 is multiclass classification of high-quality audios by CHD types mentioned above. For classification of individual subtypes of CHD (non-CHD vs. CHD subtypes).

- Task 3 is binary classification of low-quality heart sound recordings.

- Task 4 is multiclass classification of low-quality heart sound recordings.

3.0.3 Pre-processing

We preprocessed the heart sounds into Mel spectrograms, which represent spectral features. A Mel spectrogram is a visual representation of sound where the x-axis represents time and the y-axis represents frequency in terms of the Mel scale for frequencies and decibels for amplitude, similar to how humans hear [25]. The Mel spectrograms are different from traditional spectrograms and are superior to audio DL because they preferentially emphasize areas of frequency and intensity in more human-hearing-relevant terms and thus extract more features for DL models [26]. Audio recordings were resampled at 16 kHz, normalized, and divided into fixed-length 5-second segments. Mel spectrograms were calculated with a window size of 1024 samples, a hop length of 512 samples, and 64 Mel frequency bins. The spectrograms were log-scaled and normalized to improve feature representation (see Fig 3-1).

3.0.4 Model Architecture

As discussed above, the architecture of the model was inspired by the work of [24] which combines CNN features and the Vision Transformer model. The model architecture can be seen in Fig. 3-2. The key modules are detailed below:

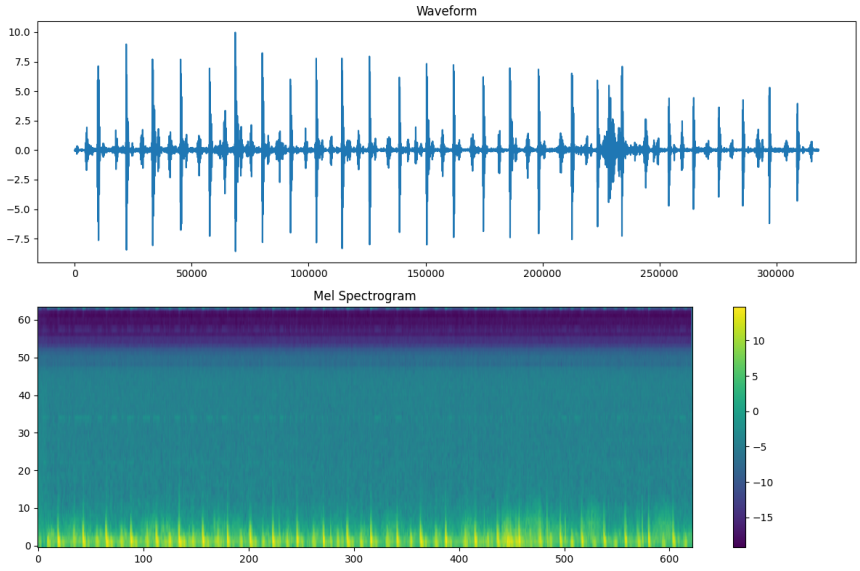

Figure 3-1: Waveform Representation vs. Mel Spectrogram

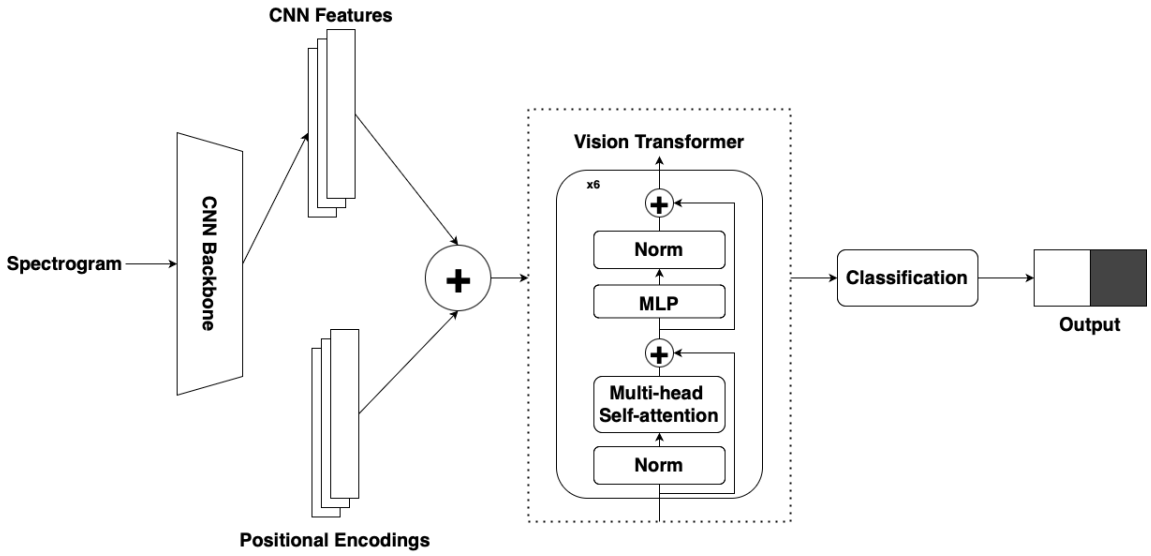

Figure 3-2: Model Architecture

CNN-Based Feature Extraction

The core of this model is a depthwise separable CNN that serves as the backbone for the extraction of spatial features from heart sound recordings. These convolutions independently process each channel to capture key local patterns in waveforms. The depthwise separable convolution is more parameter-efficient than standard convolu-

tion and therefore reduces computational cost.

- **Conv2D Layer:** The first convolutional layer has a kernel size 7×7 and is followed by batch normalization and ReLU activation.
- **Depthwise Separable Convolution Blocks:** These layers apply depthwise convolutions that operate on each input independently. They are followed by point-wise convolutions 1×1, while also using batch normalization and ReLU activation.
- **Max Pooling Layers:** These layers follow each convolution block to down-sample the feature map and reduce the spatial dimensions.

This CNN backbone produces a 2D feature map from spectrograms, which is flattened into a sequence of tokens for further processing by ViT.

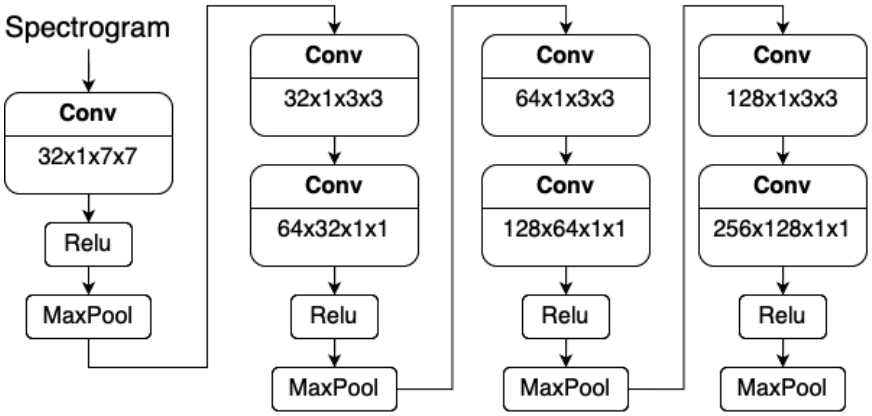

Figure 3-3: CNN Diagram

Vision Transformer

The CNN backbone makes 2D feature maps, which are then sent to the Vision Transformer module. This module uses self-attention mechanisms to pull out both local and global relationships in the input spectrogram. The transformer module employs multi-head attention mechanisms to allow the model capture complex relationships between different segments of the feature map.

It processes the features extracted from the CNN backbone:

- **Multi-Head Self-Attention Module:** This feature allows the model to attend to different parts of the feature map simultaneously, both locally and globally. Each attention head attends to a distinct feature of the input, increasing the ability of the model to detect composite structures.

- **Residual Connections:** To maintain consistency in the features and improve gradient flow during training.

- **Layer Normalization and MLP Blocks:** Both transformer layers are followed by a layer normalization step and a multi-layer perceptron block, which helps ensure training stability as well as the effectiveness of learning.

- **Positional Encodings:** Since transformers have no inherent sense of input token order, sinusoidal positional encodings are employed and appended to the input token sequence in order to provide it with this information. The positional encodings are learned and appended to the token embeddings before being fed into the transformer layers. The [CLS] token is also appended to the sequence, which is the token that encodes the global meaning of the entire sequence.

The CNN-ViT model is a standard transformer-based model, and to this end, the CNN features are first converted into sequence format, and the [CLS] token is added in the sequence. The [CLS] token is a special token that encapsulates the representation of the entire input sequence, which is ultimately used for classification.

Classification Layer

Following the processing of tokens through the transformer layers, the model uses the [CLS] token output for classification. A fully connected layer processes the final output, projecting the learned features onto the target classes. The layer generates the class predictions based on the learned representation.

3.0.5 Weighted Random Sampling

As shown in Table 3.1, the imbalance is not only for representation of subtypes but also for abnormal-normal samples. The imbalance creates enormous difficulty in ML tasks, affecting performance and generalization ability. To resolve the problem, Weighted Random Sampling technique was employed. It provides a good approach to assign sampling probability to minority classes. Particularly, the Weighted Random Sampling assigns a weight to each sample according to the inverse proportion of class frequencies in each class in the training data set.

$$w_i = \frac{1}{N_{y_i}} \tag{3.1}$$

In the above formula, the weight w_i of every sample x_i is set to be inversely proportional to the number of samples N_{y_i} in its own class y_i.

Because of this, less frequent class samples will have larger weights, thus an increased probability of selection when training. Through these normalized weights, the Weighted Random Sampler perfectly balances the data set such that all the classes are proportionately represented when being trained.

3.0.6 Training Process

The dataset was split using the 60-20-20 random split method, where 60% set was used for training, 20% for validation, and 20% for testing. We used the following configurations for training the model. Configurations are detailed in the Table 3.4.

3.0.7 Evaluation Metrics

The following metrics are considered in order to assess model performance:

- **Confusion Matrix:** When we want to evaluate multi-class classification models, a confusion matrix is used. It is a two-dimensional table, where one dimension corresponds to the true labels and the other dimension corresponds to the predicted labels. Figure 3-4 represents an example of a confusion matrix where

Table 3.4: Configurations for Model Training

Task	Loss Function	Optimizer and Hyperparameters
Task 1	Binary Cross-Entropy	Optimizer: Adam (lr: 0.0001)
		Batch size: 8
		Epochs: 20
		Early Stopping Patience: 5
Task 2	Cross-Entropy	Optimizer: Adam (lr: 0.0001)
		Batch size: 8
		Epochs: 30
		Early Stopping Patience: 5
Task 3	Binary Cross-Entropy	Optimizer: Adam (lr: 0.00005)
		Batch size: 8
		Epochs: 15
		Early Stopping Patience: 5
Task 4	Cross-Entropy	Optimizer: Adam (lr: 0.00005)
		Batch size: 8
		Epochs: 15
		Early Stopping Patience: 5
Regularization: Dropout layers in the transformer blocks to reduce overfitting.		

TP - True Positive (true label - Positive, predicted label - Positive), FN - False Negative (true label - Positive, predicted label - Negative), TN - True Negative (true label - Negative, predicted label - Negative), FN - False Positive (true label - Negative, predicted label - Positive).

- **Accuracy:** Overall correctness of CHD predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.2}$$

- **Precision, Recall, and F1 Score:** Trade-off between false positives and false

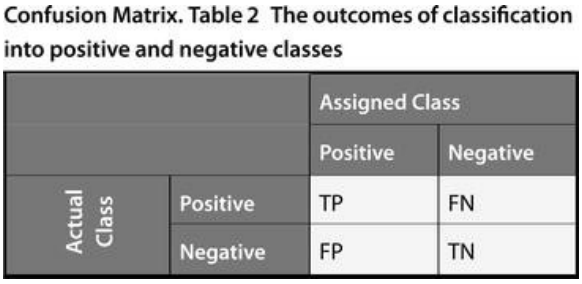

Figure 3-4: Confusion Matrix Definition [2]

negative results.

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.4}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{3.5}$$

- **ROC-AUC:** Discriminative capability for binary classification.

$$TruePositiveRate(TPR) = \frac{TP}{TP + FN} \tag{3.6}$$

$$FalsePositiveRate(FPR) = \frac{FP}{FP + TN} \tag{3.7}$$

3.0.8 Novelty and Contribution

- **Data Usage:** The study applies the CNN-ViTs model to CHD classification using PCG data which is able to identify complex patterns from the data.
- **Transformer Efficiency:** The attention module focuses on clinically relevant regions, offering a novel application for CHD detection.
- **Practicality:** The use of spectrogram data avoids preprocessing complexities of raw signals.

- **Explainability:** Since it focuses on clinically relevant regions, it might provide important and explicable insights to the decision-making process.

Chapter 4

Results

As discussed above, since the dataset used in this study has different sets (High-quality and Low-quality datasets with CHD subtypes such as PFO, PDA, ASD, and VSD), the model performance was tested on 4 different tasks.

4.0.1 Task 1 - Binary Classification of Clean Heart Sounds

Table 4.1 indicates the classification performance for clean heart sound binary classification between normal and abnormal heart conditions. The overall model performance was high, with an accuracy of 95%. Precision, recall, and F1-score of the normal class were 96%, 97%, and 96%, respectively, and for the abnormal class, these were similarly robust at 95% precision, 93% recall, and 94% F1-score. The confusion matrix (Figure 4-1) also depicts near-zero misclassification, which suggests that the model is able to discriminate perfectly between normal and abnormal states. Furthermore, the ROC-AUC analysis also revealed very high discriminative power, with a value of 0.97 (see Figure 4-2).

Table 4.1: Task 1 Classification Performance Metrics

Class	**Precision**	**Recall**	**F1-score**	**Accuracy**
Normal	0.96	0.97	0.96	0.95
Abnormal	0.95	0.93	0.94	

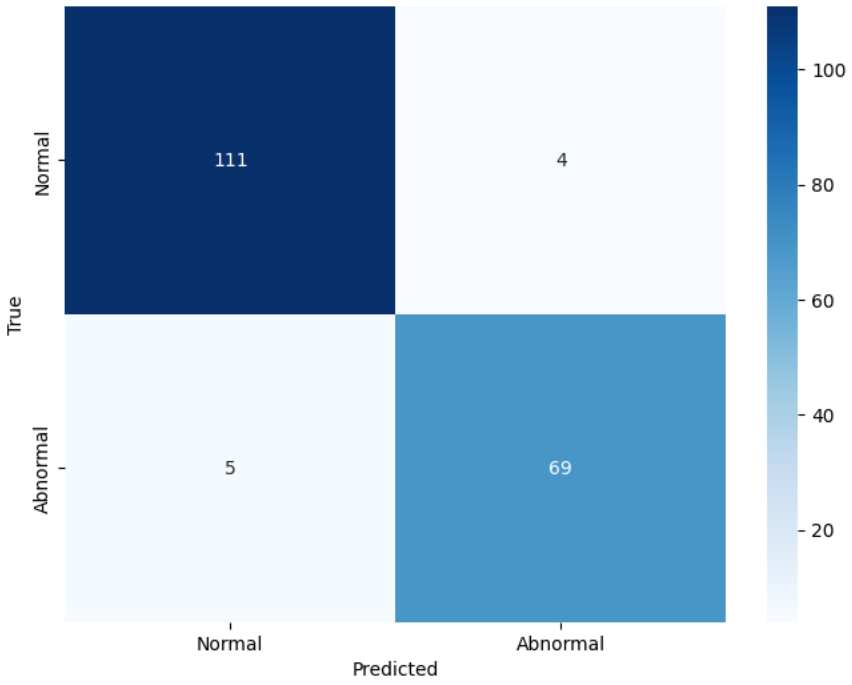

Figure 4-1: Task 1 Confusion Matrix

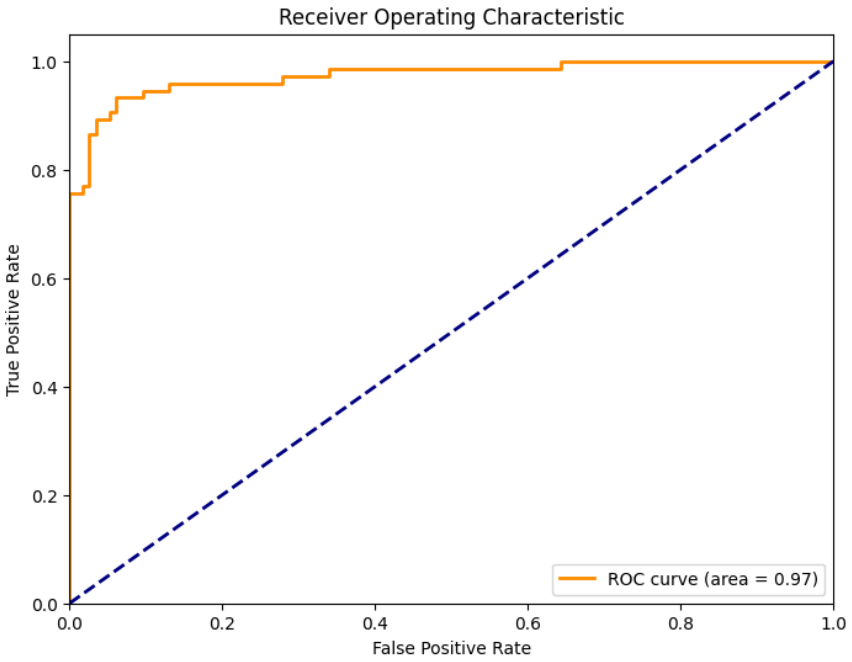

Figure 4-2: Task 1 ROC-AUC

4.0.2 Task 2 - Multi-Class Classification of Clean Heart Sounds

Table 4.2 indicates the classification performance for clean heart sound multi-class classification between CHD subtypes and normal conditions. The model performance decreased significantly for this task, with an accuracy of 74%. Precision, recall, and F1-score of the normal class were 96%, 88%, and 92%, respectively, while CHD subtype classes had considerably lower performance, with such classes as PDA achieving an F1 score of only 15%. The confusion matrix (Figure 4-3) shows that the model is able to discriminate almost perfectly between normal and abnormal states while failing on CHD subtype classification. However, analyzing the results with data distribution (see Table 3.1), we see that a limited number of samples for CHD subtypes might contribute to the worse results. For example, the class PDA with the least number of samples in the dataset (32 samples) showed the worst results, while the normal class with 533 samples in the dataset showed the best results. Thus, this pattern indicates that the classification's lower performance for Task 2 stems from insufficient data, which makes it difficult for the proposed model to effectively learn differences between CHD subtypes.

Table 4.2: Task 2 Classification Performance Metrics

Class	**Precision**	**Recall**	**F1-score**	**Accuracy**
NORMAL	0.96	0.88	0.92	
PFO	0.48	0.65	0.55	
ASD	0.26	0.33	0.53	0.74
VSD	0.68	0.55	0.61	
PDA	0.13	0.20	0.16	

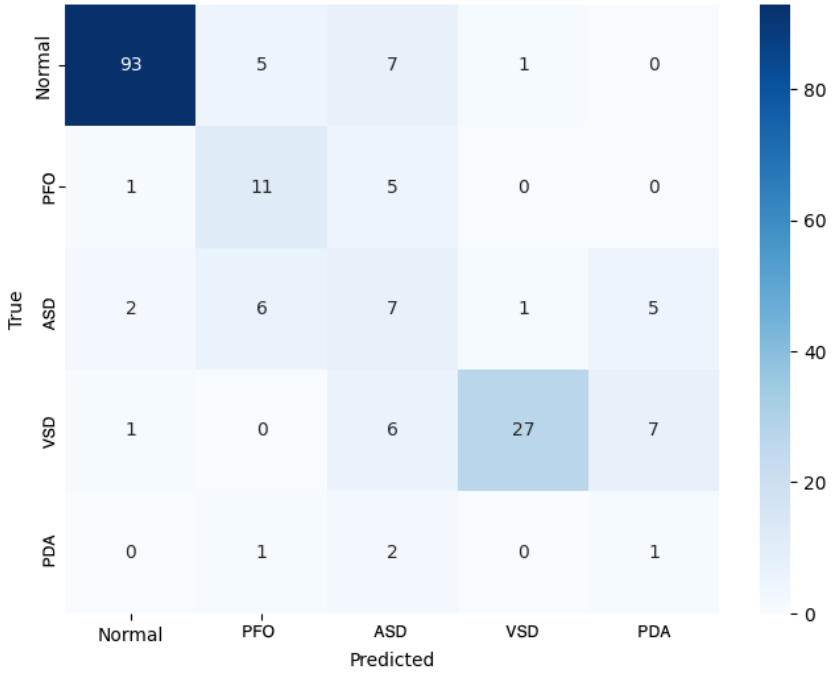

Figure 4-3: Task 2 Confusion Matrix

4.0.3 Task 3 - Binary Classification of Low-quality Heart Sounds

Table 4.3 indicates the classification performance for low-quality heart sound binary classification between normal and abnormal heart conditions. The overall model performance was very low, with an accuracy of only 48%. Precision, recall, and F1-score of the normal class were 48%, 69%, and 56%, respectively, and for the abnormal class, these were less robust at 47% precision, 27% recall, and 35% F1-score. The confusion matrix (Figure 4-4) also depicts the low performance of the model, with the abnormal class having worse performance. This means that the model was not able to discriminate between normal and abnormal states in low-quality, noisy heart sounds. Furthermore, the ROC-AUC analysis also revealed very low discriminative power, with a value of only 0.55 (see Figure 4-5). The poor classification accuracy reported for Task 3 can, to a great extent, be attributed to the intrinsic challenge of noisy and low-quality audio signals. The suggested model using CNNs and ViT relies on the ability to learn unique spatial-temporal acoustic features. However, noise tends to hide important acoustic features that are needed to tell the difference be-

tween normal and abnormal heart states. Additionally, the limited dataset size (see Table 3.1) might further restrict the model's generalizability and robust identification of pathological signatures from noisy recordings. Therefore, the combination of these problems—degradation due to noise and a small dataset—might explain the low classification performance (48%) and poor ROC-AUC (0.55) on this task.

Table 4.3: Task 3 Classification Performance Metrics

Class	**Precision**	**Recall**	**F1-score**	**Accuracy**
Normal	0.48	0.69	0.56	0.48
Abnormal	0.47	0.27	0.35	

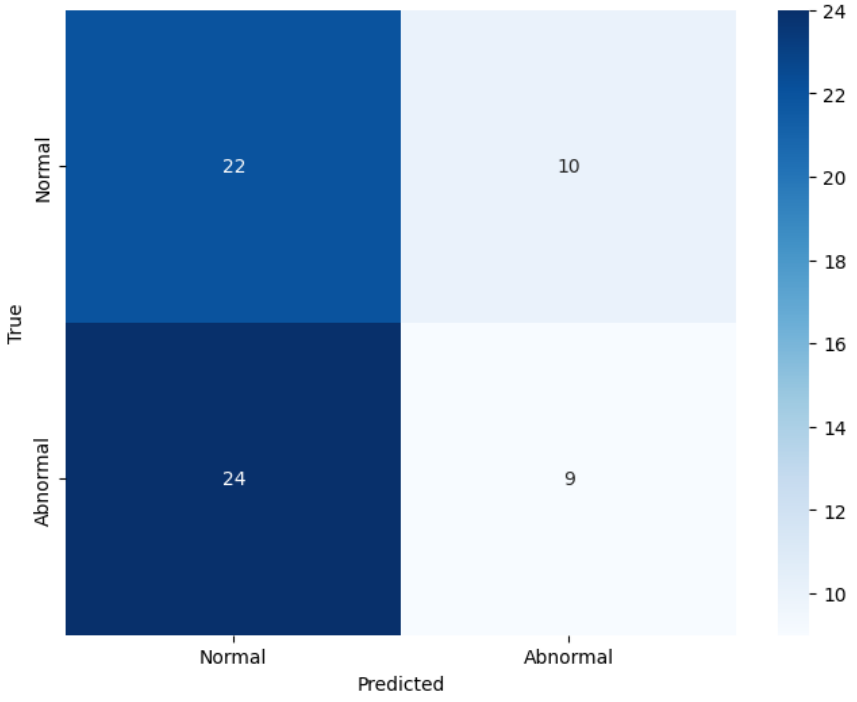

Figure 4-4: Task 3 Confusion Matrix

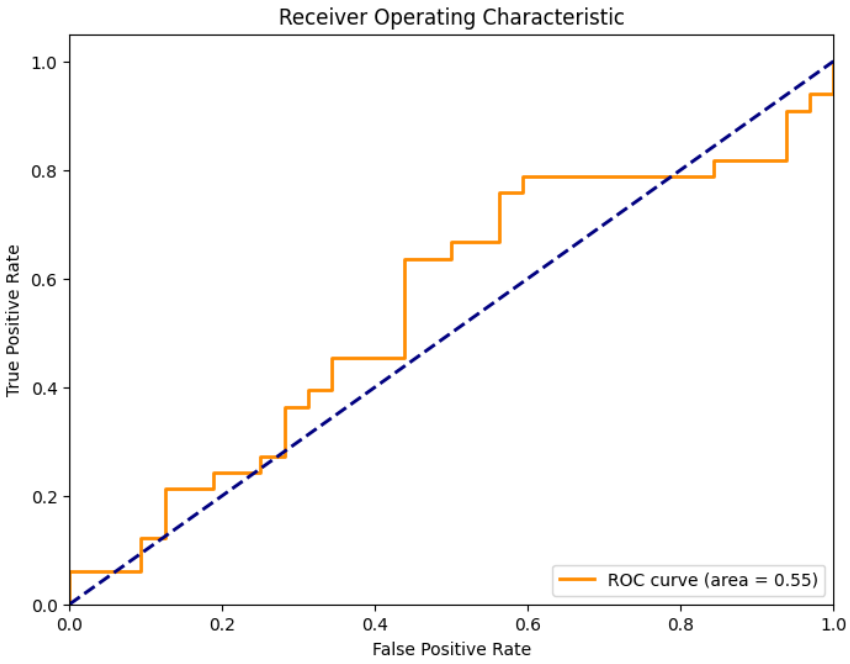

Figure 4-5: Task 3 ROC-AUC

4.0.4 Task 4 - Multi-class Classification of Low-quality Heart Sounds

Table 4.4 demonstrates the performance metrics on Task 4, that is, multi-class classification of poorly recorded heart sounds. The overall accuracy was very poor at 35%. The normal class had moderate precision (55%) and recall (53%), hence the F1 score was 54%, while PFO also enjoyed relatively higher recall (67%) but lower precision (44%). On the other hand, ASD, VSD, and PDA classes were far behind, with ASD scoring an F1 score of only 14%, while both VSD and PDA were not classified at all (F1 scores of 0%). The confusion matrix (Figure 4-6) also demonstrates the failure of the model to distinguish between the majority of the CHD subtypes in noisy conditions, indicating that noise and scarcity of samples significantly impair the CNN-ViT model's performance on this difficult task.

Table 4.4: Task 4 Classification Performance Metrics

Class	Precision	Recall	F1-score	Accuracy
NORMAL	0.55	0.53	0.54	
PFO	0.44	0.67	0.53	
ASD	0.20	0.11	0.14	0.35
VSD	0.00	0.00	0.00	
PDA	0.00	0.00	0.00	

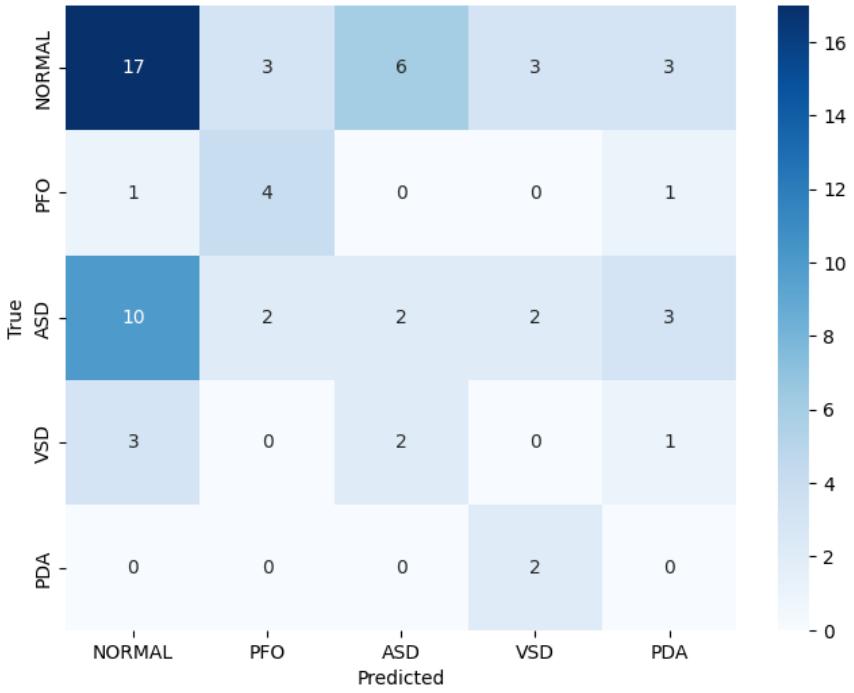

Figure 4-6: Task 4 Confusion Matrix

Table 4.5: Comparison of Existing CHD Detection Methods and Results with Our Results

Paper	**Data Modality**	**Methodology**	**Results**
[4]	CT Images	CDLM	Accuracy: 83.8%
[8]	Chest X-ray	Canny Edge + DSS	Accuracy: 93.3%
[15]	Heart Sounds	CNN	Sensitivity: 97%, Specificity: 89%, Accuracy: 96% ($\kappa = 0.84$)
[6]	Fetal Echocardiogram	SPReCHD	F1 Score: 94.95%
[16]	Electrocardiogram	RoR	Accuracy: 92.45%, Sensitivity: 74.73%, Specificity: 94.07%
[11]	Electrocardiogram	CHDdECG	Specificity: 88.1% to 93.7%
[17]	Phonocardiogram	CNN + Attention Transformer	Accuracy: 92.3%, Sensitivity: 98.3%, Specificity: 83.3%, AUC: 0.964
[18]	Phonocardiogram	Dense-block CNN	Accuracy: 96.21%, Specificity: 98.08%, Precision: 91.67% (CHD)
[19]	Phonocardiogram	Fourier Transform + Cepstrum + SVM	Accuracy: 95%, Sensitivity: 100%, Specificity: 90%
[1]	Phonocardiogram	Random Forest (RF)	Accuracy (Binary): 90.3%, Accuracy (Multi-class): 93.4%
This work	Phonocardiogram	Vision Transformer (CNN-ViT [24])	Accuracy (Binary): 95%, Accuracy (Multi-class): 74%

Chapter 5

Conclusion

CHD is still the leading global health concern, related to the high prevalence and associated mortality, as well as the long-term impacts on survivors. Precise diagnosis early in life is a critical determinant of outcomes and risk reduction for disease progression and complications. Although existing methods of diagnosis pose challenges in terms of access, expertise level, and diagnostic accuracy.

So, this study looked into how ViTs, specifically a CNN-ViT-based model, might be able to find CHD in PCG recordings. The proposed model leveraged the strength of CNNs for effective local features extraction with ViT's ability to capture complex global dependencies.

The experimental analysis was performed on ZCHSound [1] dataset under four different scenarios: binary (CHD vs. non-CHD) and multi-class (CHD types vs. non-CHD) for both clean and noisy heart sounds.

The proposed model displayed exceptional results for binary classification of clean heart sounds, achieving an accuracy of 95%, precision of 96%, recall of 97% and ROC-AUC of 0.97. This demonstrates the model's ability to discriminate between normal and abnormal heart sounds. However, multi-class classification of the model was less effective with an accuracy of 74%. On the other hand, this could be explained with a limited sample size for CHD subtypes. This means that the model requires more data to learn some features of CHD subtypes and discriminate them effectively. This limitation underscores the importance of dataset size and the balanced representation

required for successful DL model performance.

Nonetheless, a dramatic decline was observed with low-quality data. The proposed model revealed substantially lower accuracies both for binary and multi-class classification (48% and 35% respectively). These results highlight the effects of noise in pediatric heart sounds on DL models effectiveness. These noises presumably make the feature extraction process and learning process difficult for ViT-based models. Another possible factor is again sample size, which was low for the low-quality dataset.

Overall, this study contributes to the field by:

- Applying a sophisticated CNN-ViT-based model for CHD detection
- Highlighting potential of VitT-based architectures to identify complex temporal and spectral patterns crucial for cardiac diagnostics
- Demonstrating the feasibility and practicality of heart sounds' Mel spectrogram representation, simplifying pre-processing and ensuring scalability

Future work should address the identified shortcomings of this study. First, the proposed model should be tested on datasets with large sample sizes for CHD subtypes or propose advanced augmentation methods for heart sound generations. Further, the future work should address the issue with the proposed model being ineffective for low-quality heart sounds. In particular, one could apply advanced noise reduction techniques, which could possibly help get better results.

In conclusion, this study indicated that CNN-ViT hybrid models exhibit robust performance on children's CHD detection from pediatric heart sounds. However, further enhancements in noise resistance and class imbalance handling are necessary. Overall, the study provides a promising step towards accurate, automated, and accessible CHD diagnostics, which might foster improved clinical outcomes through earlier detection and intervention.

Bibliography

[1] Weijie Jia, Yunyan Wang, Renwei Chen, Jingjing Ye, Die Li, Fei Yin, Jin Yu, Jiajia Chen, Qiang Shu, and Weize Xu. Zchsound: Open-source zju paediatric heart sound database with congenital heart disease. *IEEE Transactions on Biomedical Engineering*, 71(8):2278–2286, 2024.

[2] Kai Ming Ting. *Confusion Matrix*, pages 260–260. Springer US, Boston, MA, 2017.

[3] D. Syssoyev et al. Epidemiology of congenital heart disease in kazakhstan: Data from the unified national electronic healthcare system 2014-2021. *Journal of Clinical Medicine of Kazakhstan*, 21(3):49–55, 2024.

[4] Khalil Khan, Taimur Ahmad, and Irfan Uddin. Cardiac deep learning model for congenital heart disease recognition. In *2023 15th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, pages 293–297, 2023.

[5] R. Deepika, P. Balaji Srikaanth, and R. Pitchai. Early detection of heart disease using deep learning model. In *2022 8th International Conference on Smart Structures and Systems (ICSSS)*, pages 1–4, 2022.

[6] Sibo Qiao, Shanchen Pang, Yi Sun, Gang Luo, Wenjing Yin, Yawu Zhao, Silin Pan, and Zhihan Lv. Sprechd: Four-chamber semantic parsing network for recognizing fetal congenital heart disease in medical metaverse. *IEEE Journal of Biomedical and Health Informatics*, 28(6):3672–3682, 2024.

[7] Li Zhixin, Luo Gang, Ji Zhixian, Wang Sibao, and Pan Silin. Chd-cxr: a deidentified publicly available dataset of chest x-ray for congenital heart disease. *Frontiers in Cardiovascular Medicine*, 11, 2024.

[8] S. Jyothi and K. Vanisree. Congenital heart septum defect diagnosis on chest x-ray features using neural networks. In *2016 Second International Conference on Computational Intelligence Communication Technology (CICT)*, pages 265–269, 2016.

[9] National Cancer Institute. Echocardiography, n.d. Accessed: 2025-04-05.

[10] Dalwinder Janjua, Japna Singh, and Amit Agrawal. Pulse oximetry as a screening test for congenital heart disease in newborns. *Journal of Mother and Child*, 26(1):1–9, 2022.

[11] Jintai Chen, Shuai Huang, Ying Zhang, Qing Chang, Yixiao Zhang, Dantong Li, Jia Qiu, Lianting Hu, Peng Xiaoting, Yunmei Du, Yunfei Gao, Danny Chen, Abdelouahab Bellou, Jian Wu, and Hui-Ying Liang. Congenital heart disease detection by pediatric electrocardiogram based deep learning integrated with human concepts. *Nature Communications*, 15, 02 2024.

[12] Ana Gotter. Pulse oximetry: Uses, readings, and how it works, 2024. Accessed: 2025-04-05.

[13] H. Kenneth Walker, W. Dallas Hall, and J. Willis Hurst. *Clinical Methods: The History, Physical, and Laboratory Examinations*. Butterworths, Boston, 3 edition, 1990.

[14] Olivier Ecabert, Jochen Peters, Hauke Schramm, Cristian Lorenz, Jens von Berg, Matthew J. Walker, Mani Vembar, Mark E. Olszewski, Krishna Subramanyan, Guy Lavi, and JÜrgen Weese. Automatic model-based segmentation of the heart in ct images. *IEEE Transactions on Medical Imaging*, 27(9):1189–1201, 2008.

[15] Jingjing Lv, Bin Dong, Hao Lei, Guocheng Shi, Hansong Wang, Fang Zhu, Chen Wen, Qian Zhang, Lijun Fu, Xiaorong Gu, Jiajun Yuan, Yongmei Guan, Yuxian Xia, Liebin Zhao, and Huiwen Chen. Artificial intelligence-assisted auscultation in detecting congenital heart disease. *European Heart Journal - Digital Health*, 2(1):119–124, 01 2021.

[16] Yunmei Du, Shuai Huang, Canhui Huang, Allam Maalla, and Huiying Liang. Recognition of child congenital heart disease using electrocardiogram based on residual of residual network. In *2020 IEEE International Conference on Progress in Informatics and Computing (PIC)*, pages 145–148, 2020.

[17] Md Hassanuzzaman, Nurul Akhtar Hasan, Mohammad Abdullah Al Mamun, Mohanad Alkhodari, Khawza I. Ahmed, Ahsan H. Khandoker, and Raqibul Mostafa. Recognition of pediatric congenital heart diseases by using phonocardiogram signals and transformer-based neural networks. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 1–4, 2023.

[18] Ding Chen, Weipeng Xuan, Yexing Gu, Fuhai Liu, Jinkai Chen, Shudong Xia, Hao Jin, Shurong Dong, and Jikui Luo. Automatic classification of normal–abnormal heart sounds using convolution neural network and long-short term memory. *Electronics*, 11(8), 2022.

[19] Anjali Yadav, Malay Kishore Dutta, Carlos M Travieso, and Jesus B. Alonso. Automatic classification of normal and abnormal pcg recording heart sound recording using fourier transform. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–9, 2018.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[21] Ali Farzipour, Omid Nejati Manzari, and Shahriar B. Shokouhi. Traffic sign recognition using local vision transformer. In *2023 13th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 191–196, 2023.

[22] Abdelhafid Berroukham, Khalid Housni, and Mohammed Lahraichi. Vision transformers: A review of architecture, applications, and future directions. In *2023 7th IEEE Congress on Information Science and Technology (CiSt)*, pages 205–210, 2023.

[23] Akhil Vaid, Joy Jiang, Ashwin Sawant, Stamatios Lerakis, Edgar Argulian, Yuri Ahuja, Joshua Lampert, Alexander Charney, Heather Greenspan, Jagat Narula, Benjamin Glicksberg, and Girish Nadkarni. A foundational vision transformer improves diagnostic performance for electrocardiograms. *npj Digital Medicine*, 6, 06 2023.

[24] Yanfang Dong, Miao Zhang, Lishen Qiu, Lirong Wang, and Yong Yu. An arrhythmia classification model based on vision transformer with deformable attention. *Micromachines*, 14:1155, 05 2023.

[25] S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 01 1937.

[26] Ketan Doshi. Audio deep learning made simple - why mel spectrograms perform better, 2023. Accessed: 2024-12-16.