Heart Sound Classification using Vision Transformer Models

by

Zhanat Adilkhanuly

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

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

November 2023

Author .....
Department of Computer Science
November 23, 2023

Certified by .....
Dr. Meiram Murzabulatov
Assistant Professor
Thesis Supervisor

Accepted by .....
Dr. Yelyzaveta Arkhangelsky
Dean, School of Engineering and Digital Sciences

Heart Sound Classification using Vision Transformer Models

by

Zhanat Adilkhanuly

Submitted to the Department of Computer Science
on November 23, 2023, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

The automatic heart sound classification is an integral part of the early diagnosis of cardiovascular diseases(CVDs). Even though advances in medical technologies allow us to diagnose many CVDs, it remains one of the leading causes of death worldwide due to its absence of symptoms at the initial stages. Thus, there is a huge demand to develop other methods of identifying heart sound abnormalities that are less expensive, simple, and applicable. Several audio feature extraction methods, in combination with classification models, have been developed over time. However, existing feature extraction methods are sensitive to noise, which negatively impacts the performance of the heart sound classification model. In addition, there is a strong need to develop models more sensitive to heart sound abnormalities in patients. In this work, we address the limitations of extracted features by using spectrogram images that are taken from Discrete Fourier Transform, and introducing them to Vision Transformer Model. Results of our experiments on the benchmark of PhysioNet Heart Sound Dataset show that the proposed method outperforms existing methodologies with an accuracy of 0.925 and with a sensitivity score of 0.955

Thesis Supervisor: Dr. Meiram Murzabulatov
Title: Assistant Professor

Acknowledgments

I want to express my sincere gratitude to Dr. Meiram Murzabulatov for his invaluable guidance, unwavering support, and insightful feedback throughout the entirety of my thesis work. His expertise and dedication have played a crucial role in shaping the quality and depth of this research.

I extend my heartfelt appreciation to my family and friends for their unconditional love, encouragement, and companionship during the challenging phases of my academic journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background Information

Cardiovascular diseases (CVDs) are a class of disorders that affect the heart and blood vessels and include coronary artery disease, heart failure, arrhythmias, and strokes. According to the WHO, these are among the leading causes of death globally, claiming an estimated 17.9 million lives each year. Early diagnosis to manage and prevent complications of CVDs is a crucial part of fighting against cardiovascular diseases so that the patients can get all the care they need such as lifestyle modifications, medication, and, in severe cases, medical procedures [1].

The early diagnosis of CVDs usually starts by checking patients' heart conditions. There are several medical procedures available including medical imaging procedures, ECG(electrocardiogram), and echocardiography. However, they can be inconvenient and costly. Heart sound auscultation provides a practical alternative to assess heart health. The heart's primary function is to supply oxygen and nutrients to various organs while removing metabolic byproducts, allowing cells to maintain their physiological balance. The heart comprises four chambers: the left atrium, left ventricle, right atrium, and right ventricle, each playing a specific role in the circulation process. The rhythmic contraction and relaxation of the heart generate heart sounds. However, early-stage abnormalities might be missed during the heart sound auscultation due to human errors [6].

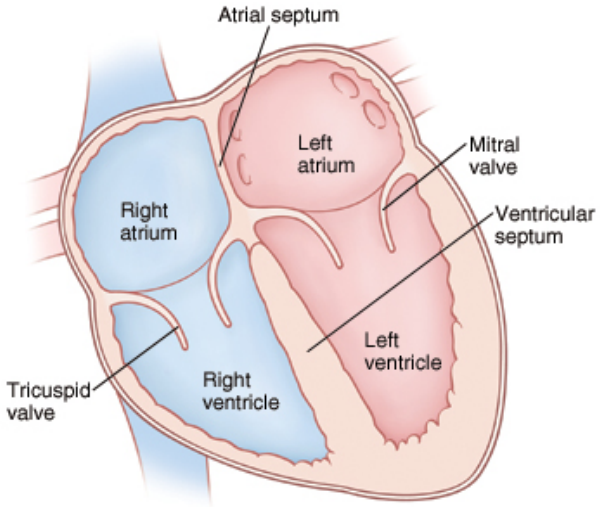

Figure 1-1: A human heart structure[11]

One heartbeat or a cardiac cycle has four sounds in turn. Each stage represents the condition of different chambers and valves in between. The first stage(S1) marks the closure of the tricuspid and mitral valves that are between the atriums and ventricles. It is a loud sound with high intensity. The second sound(systole) begins with the contraction of the heart atrium with less intensity than the first sound. The third sound(S2) does occur due to the closure of pulmonary and aortic valves. This is followed by the contraction of the ventricles (diastole) right before the next cardiac cycle. Usually, there is no diastole sound. However, the existence of even a tiny sound that humans don't hear during the last stage can be pathological [10]. An example of heart signals is shown in Figure 1-2 below.

As has been mentioned, understanding heart sounds can be challenging due to their low intensity and nearby dominant frequencies. In addition, they might often fall near the lower threshold of human hearing. Therefore, interpreting these sounds accurately requires extensive training and experience in auscultation [2]. This is where heart sound classifier models come in, offering valuable assistance in detecting different abnormalities accurately. The development of machine learning models in heart sound classification can be significant since it can increase the performance of diagnosing CVDs at early stages.

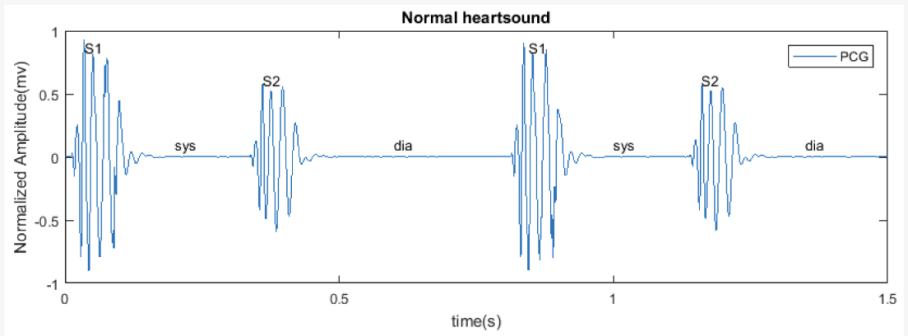

Figure 1-2: A waveform representation of heart sounds with S1, systole, S2, and diastole stages

Most previous studies approach heart sound classification problems, starting by segmenting each heart sound into four parts and extracting features from four parts separately [2, 3, 4, 5]. It has been claimed that it improves the accuracy of the classification. However, this classification method adds a complex layer called segmentation to the task. Only recent studies have started addressing this problem without segmenting the audio signals into four parts before extracting features from them[6, 7, 8, 9].

1.2 Related Works

1.2.1 Heart Sound Segmentation

As mentioned, heart sound segmentation before extracting features from audio sounds has received significant research attention. To solve the classification problem, Bozkurt [2] developed a model that focuses on the segmentation and time-frequency representation components of the CNN(Convolutional Neural Networks)-based design. In their work, they used the period synchronic and asynchronous segmentation methods and got S1 and S2 sounds from audio. Next, Mel-frequency cepstral coefficients(MFCCs), Mel-frequency Spectral coefficients(MFSCs), and sub-band envelopes as time series representation features were extracted from those audio files. By the end, they had implemented a two-dimensional convolutional layer model with an accuracy of 0.75. Humayun[3] implemented another model with a similar segmentation

method but used time series signals as a feature of the time convolutional units, including finite impulsive response filters. Even though the model has been simplified, there was no total improvement in the performance of the proposed model.

To further improve the performance of the classification, Norman[4] has introduced another methodology that uses a dimensional convolutional layer for extracting features from audio combined with another two-dimensional layer as a classification model. During the feature extraction part, they used the Butterworth filter between 25 Hz and 400 Hz to solve the background noise problem. As a result, they have achieved an accuracy of 0.892 and a sensitivity of 0.899. Latif [5] implemented this model using Residual Convolutional Layers, such as LSTM, and bLSTM making it state-of-the-art in performance with an accuracy of 0.96.

1.2.2 Without Heart Sound Segmentation

The other half of the researchers have developed models without segmenting the audio files into separate audios. One of the works introduced the improved MFCC feature extraction method to minimize noise from the dataset [6]. In addition, to radically improve the performance of heart sound classification, they have merged three datasets (PhysioNet, Pascal, and Yaseen Datasets) and implemented the Deep Residual Learning model. An improved MFCC feature is the first and second derivatives of MFCC. As a result, they have achieved an accuracy of 0.94, outperforming most of the models developed after segmenting the audio files in datasets.

Another work developed by Fatih [7], has introduced another type of feature extraction method called the Fractional Fourier Transform, which indicates the rotation of the signal and a generalization of the Fourier transform. They have also concluded that extracted performed well with traditional classifiers such as kNN SVM. They achieved an accuracy of 0.92 at a sensitivity of 0.874.

Dominquez [8] tried to improve the performance of the classification model by implementing the work on a large number of datasets and using nine different datasets. They have implemented the resulting features into AlexNet and got a performance accuracy of 0.94 at a sensitivity of 0.93

1.3 Motivation for this work

Even though there have been a lot of improvements made in the sphere of heart sound classification, most of them require complex structured feature extraction methods together with computationally expensive models. In a real-time environment, these implemented models might be used on computers with less computational power(e.g., edge devices near electronic stethoscopes). For this reason, there is a need to use either fewer feature extraction methods or computationally less expensive classification models.

Another important factor to consider in heart sound classification is to have models with high-sensitivity results. This is needed to ensure that the model will miss as few patients with heart abnormalities as possible. It is something that most of the research papers didn't take into account in their works. These reasons can be considered as a viable direction for future research.

Chapter 2

Methodology

2.1 Dataset

In this work, the PhysioNet Heart Sound Dataset was used as a benchmark [12]. These are the heart sound recordings of normal/abnormal sounds that are recorded in a clinical and nonclinical environment using an electronic stethoscope. PhysioNet is a resource platform for complex physiological signal research managed by the MIT Computational Physiology Laboratory [12].

This dataset contains five databases with a total of 3126 recordings. Audio files are labeled as normal and abnormal. The length of audio files range from 5s to 120 seconds. Normal recordings are from patients with healthy heartbeats, while abnormal recordings are from patients with different types of heart problems. This includes mitral valve prolapse, mitral regurgitation, aortic stenosis, and valvular surgery defects. In addition, some of the patients have coronary heart disease. These diseases are all labeled as abnormal and there is no information on a case-by-case label for abnormal recordings. All the recordings are re-sampled into 2000 Hz and saved in a .wav format [12]. More detailed information about the dataset is given in Figure 2.1.

Filename	Normal	Abnormal
Training-a	117	292
Training-b	386	104
Training-c	7	24
Training-d	27	28
Training-e	1958	183
Training-f	80	34
Total	**2575**	**665**

Table 2.1: PhysioNet Heart Sound Dataset. Number of normal/abnormal recordings in each database

2.2 Data Preprocessing

Initially, since one heart cycle falls into less than 3 seconds, the first 5 seconds of the data were taken from each dataset.

First Preprocessing part is to calculate the Discrete Fourier Transform from each audio file. This is calculated using the formula

$$\hat{f}_k = \sum_{j=0}^{n-1} f_j e^{-i2\pi jk/n} \ , \tag{2.1}$$

where $\hat{f}_k$ is the magnitude/power of the k-th frequency [13]. An example of a visual representation is shown in Figure 2-1.

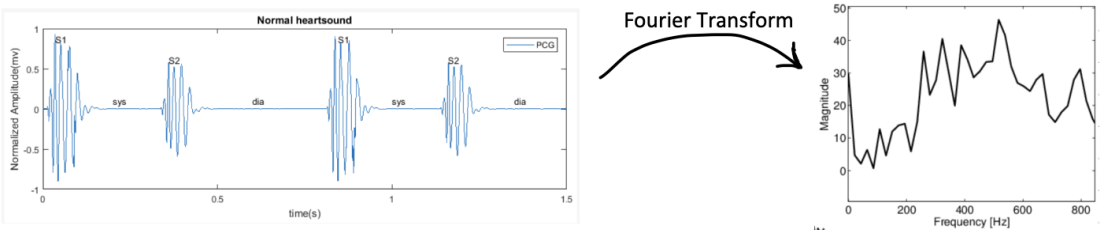

Figure 2-1: A representation of time-frequency graph and resulting frequency-magnitude/power graph after Discrete Fourier Transform.

However, the limitation of the resulting vector is that we need to know when the following frequencies occurred in the audio. Since we are working on classifying sounds that follow four sounds, we must keep the information about when those

frequencies happened in the audio. Here we can use the Gabor Transform method to save information about when the frequencies occurred [13]. The Gabor transform's general formula is

$$G(f) = \hat{f}_g(t, w) = \int_{-\infty}^{\infty} f(\tau) e^{-j\omega\tau} g(t - \tau)\, d\tau \ , \tag{2.2}$$

where τ is coming from the weight Gabor function, $g(\tau)$. It is the frequency domain function that is calculated by the Fourier transform and weighted by the Gabor function [13]. The visual representation is shown in Figure 2-2. As we can see from the figure, the Gabor function is sliding across in time through the time-frequency graph, helping us to preserve the time when the frequencies occurred in the audio.

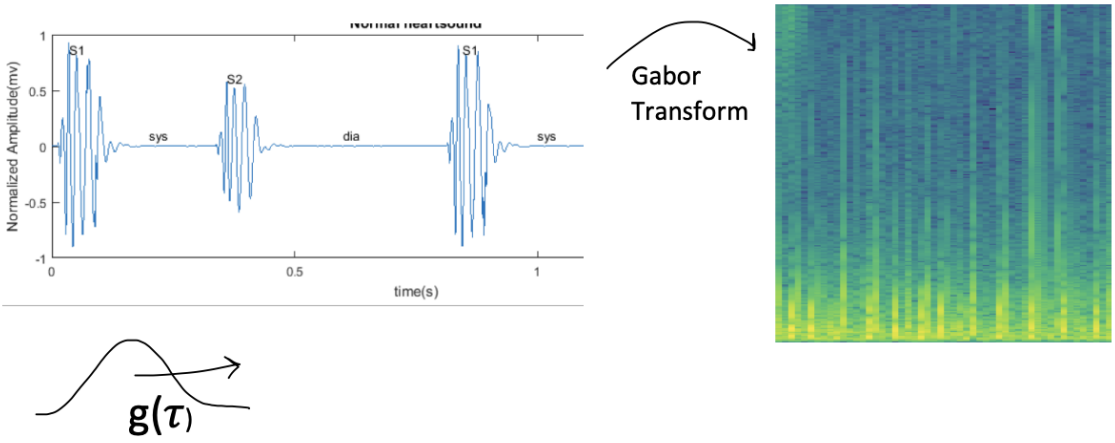

Figure 2-2: A representation of time-frequency graph and resulting spectrogram image after Gabor Transform has been applied.

So, only the spectrogram images were used as extracted features. All audio files went through the steps above, and spectrogram images were ready to be put into the proposed model.

2.3 Proposed Model

As can be seen from Figure 2-3, all the images are initially saved in 224x224 pixels. Then, they are patched into smaller pieces in 16x16 pixels, separating each image into 14 patches. Since transformer encoding accepts only one-dimensional vectors, all

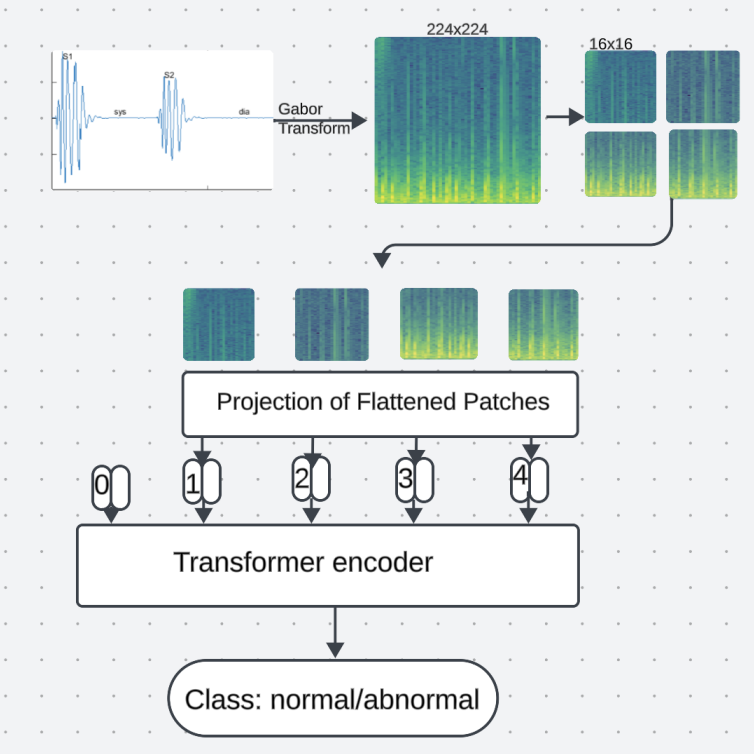

Figure 2-3: A proposed Vision Transformer model

patches are flattened via projection, resulting in one-dimensional vectors. Numbers are connected to each embedding to save the position information of each patch.

Since our Transformer Model is pre-trained, the encoder saves its inside architecture. This is how the model works: after projecting flattened patches, we apply the convolutional layer to them to get low-level features from images. Next, we use the softmax function for low-level features to get an attention layer. After that, we apply transformers, which are several visual tokens(group pixels) to model the relationship between features. These steps are repeated several times to reach the best accuracy with minimum loss[14].

The output of those projections is called patch embeddings. So, those embeddings serve as the input to the encoder [14]. The output is expected to be either normal or abnormal.

2.4 Evaluation Metrics

Four different evaluation metrics have been used to compare the performance of the proposed and existing models.

2.4.1 Accuracy

One of the effective ways of comparing the performance of classification models is by calculating the accuracy of predictions. This is calculated by the ratio of the number of predictions we got to the total number of predictions in a test set. So, the formula is given as:

$$Accuracy = \frac{total\,number\,of\,correct\,predictions}{total\,number\,of\,predictions}.$$

However, when we are talking about medication data, solely looking at the accuracy of the classification model is not enough.

2.4.2 Sensitivity

Another essential evaluation metric for classifying heart sound data is sensitivity. This is the ratio of true positives to the sum of people who truly have abnormalities. In other words, sensitivity gives us information on the percentage of people correctly identified to have abnormalities by the test.

$$Sensitivity = \frac{True\,Positive}{True\,Positive\,+ False\,Negative},$$

where false negative is the number of people who have abnormalities, but the test showed normal results. The goal is to have as small a false negative as possible, giving us high sensitivity.

2.4.3 Specificity

Another evaluation metric for classifying heart sound data is specificity. This is the ratio of true negative to the sum of people with normal heart sounds. In other words,

specificity gives us information on the percentage of people correctly identified to have normal heart sounds by the test.

$$Specificity = \frac{True\,Negative}{False\,Positive\,+True\,Negative},$$

where false positive is the number of people with normal heart sounds, but the test showed abnormal results.

2.4.4 Negative Predicted Value

Another evaluation metric that has special importance is a negative predictive value. It gives us information on the ratio of correct normal heart sound prediction to the sum of all heart sound recordings that tested normal. The formula is the following:

$$Negative\,Predicted\,Value(NPV) = \frac{True\,Negative}{False\,Negative\,+True\,Negative}.$$

The performance of all three metrics is between 0 and 1, or 0% and 100%.

Chapter 3

Experiments and Results

3.1 Experimental Setup

The proposed model with spectrogram image features is compared with other existing models using all performance evaluation metrics described in Section 2.4, namely Accuracy, Sensitivity, and Specificity. The data is split into train/validation/test sets with a 70/20/10 ratio. We have split the labeled data separately to have an equal proportion of normal and abnormal sounds. A train set contains 1545 normal and 400 abnormal heart sounds, a validation set of 515 normal, 268 abnormal, and a test set of 262 normal and 132 abnormal heart sounds.

The proposed model was implemented using Transformers ViTForImageClassification. The training set was 100 epochs with a batch size of 16 for the weight update. To prevent overfitting from happening, the dropout was applied through the network. In addition, another method of early stopping with a patience of 25 epochs was applied as an alternative method of preventing overfitting. So, if there is no improvement in the validation set for 25 epochs, it will automatically stop predicting.

In addition to the proposed method, other methodologies were also implemented to look for better performance and speed of the prediction. They are the following. They are the following feature extraction and methodology combinations:

- **[SIF-VGG16]** Spectral Image Features in combination with VGG16 pre-trained

Model	Accuracy	Sensitivity	Specificity	NPV
SIF-VGG16	0.85	0.8	0.9	0.81
MS-eCNN	0.85	0.8	0.88	0.9
MFCC-YAMNet	0.63	0.65	-	-
S-ViT	**0.927**	**0.955**	**0.91**	**0.97**

Table 3.1: Accuracy, Sensitivity, Specificity, and NPV of the test set for different models that we have tried

Paper	Segmentation	Accuracy	Sensitivity	Specificity	NPV
Bozkurt[2]	yes	0.81	0.845	0.785	-
Humayun[3]	yes	0.81	0.86	0.76	-
Latif [4]	yes	0.96	0.97	0.95	-
Norman [5]	yes	0.88	0.89	0.86	-
Feng [6]	no	0.94	0.92	0.95	-
Abduh [7]	no	0.91	-	-	-
Dominquez [8]	no	0.94	0.93	0.95	-
Xiao [9]	no	0.93	0.86	0.95	-
S-ViT	**no**	**0.927**	**0.95**	**0.91**	**0.97**

Table 3.2: Accuracy, Sensitivity, Specificity, and NPV of the test set compared to other existing models

CNN model.

- **[MS-eCNN]** Mel Spectrogram with ensemble CNN.
- **[MFCC-YAMNet]** Mel frequency cepstral coefficients with YAMNet pre-trained model.
- **[S-ViT]** Proposed Model. Spectrogram with Vision Transformers.

3.2 Results

Tables show the resulting performances of models based on accuracy, sensitivity, and specificity. Table 3-1 gives different combinations that have been implemented to improve the performance and complexity of the classification model.

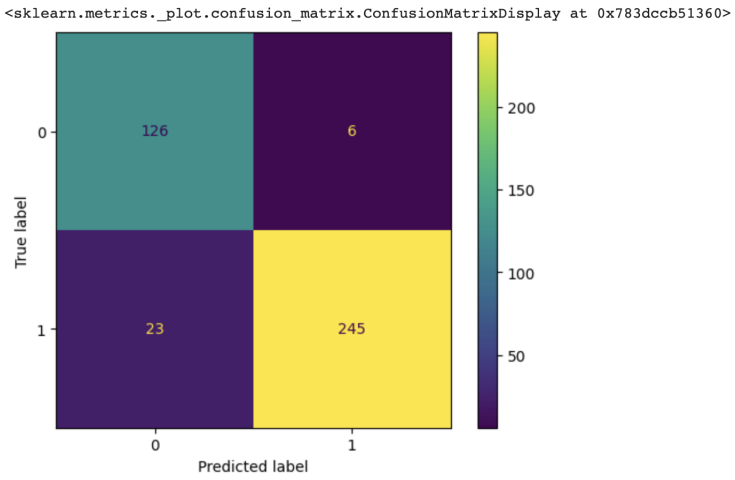

Figure 3-1: A confusion matrix of the proposed model, Vision Transformers. 0-abnormal, 1-normal

Table 3.1 indicates that spectrogram images with vision transformers perform better than other feature-model combinations with an accuracy of 0.927 at a sensitivity of 0.955. Initially, we extracted mel spectrogram features from the dataset. We implemented them into ensemble CNN, where we used one-dimensional CNN and two-dimensional CNN together, where the features are put into both models. By the end, the models were combined into a Late fusion model as a meta-classifier and were used to get the classification results[15]. However, the results were almost close to the existing models and didn't outperform in terms of simplicity or performance metrics. So, there was a need to look for a simpler model and less extracted features. That is why using an audio spectrogram as an image came into the scene.

As explained in the Methodology part, spectrogram images were taken by applying the Gabor transform into time-frequency domain audio data, and image features were put into a model. In the next model, spectrogram image features were used together with model VGG16 [16] with fine-tuning. VGG16 was pre-trained with the ImageNet dataset[17]. We have predicted that the results will be better than other simple machine learning models. However, the accuracy and sensitivity of the data were much less than the state-of-the-art classification models. The reason might be that the pre-trained model's training images do not include spectrogram images. Therefore, the model architecture doesn't work well with heart sound spectrogram images.

Since there was a need to implement a model that is better than the VGG16, Vision transformers were implemented(Figure 3-1). The same spectrogram images were put into the Vision Transformers pre-trained models trained on the ImageNet21k[18] database. After seven epochs, we reached the training accuracy of 0.90 with minimum validation loss. After that, we reached an accuracy of 0.927 at a sensitivity of 0.955.

Now, compared to other existing models, the accuracy of the proposed model is almost similar to models that didn't use segmentation. As the data we are working on is healthcare data, the main goal is not to miss patients with heart abnormalities. This is done by getting high sensitivity and negative predictive value results. So, we need to reach the goal of including as many patients with heart abnormalities as possible to diagnose their condition early. As we can see from the table, the sensitivity score is greater than other models, at least for 2%.

The only work that performs better than the proposed is Latif[4], with an accuracy of 0.96 at a sensitivity of 0.97. It has used the segmentation method before extracting features and has used MFCC features. Compared to their work, the proposed method skips the segmentation step and doesn't spend time extracting MFCC, which is sensitive to noise. Even though ViT might be structurally more complex than ResNet, it will perform faster and computationally efficiently to predict new data after building a model. However, ViT is still a complex architecture and might be computationally expensive relative to less powerful computers.

Humayun[3] implemented another simpler model with less computationally expensive model and features. In their work, times series signals(raw audio files) are used as features, and 1D-CNN is used for modeling. However, they have used Hidden Markov models for heart sound segmentation which adds another complexity to their methodology. In addition, they applied the Fast Fourier Transform to time signals and got accuracy and sensitivity scores of 0.81 and 0.86. Our proposed model outperforms in terms of skipping the segmentation section and increased accuracy of 0.92 and a sensitivity of 0.95.

Another motivation to develop the heart sound classification research was to implement a model compatible with computers with less computational power since

paramedics and other healthcare professionals could use the implemented model in different environmental areas where powerful computational tools may not be available. That's why another model was tested to be implemented. The audio files were immediately put into a YAMNet model. YAMNet model is a modified version of MobileNet with an additional layer in front to extract features from the audio files. This means that it has a built-in layer to extract features. So, YAMNet is a 54-layered deep convolutional neural network with an additional layer for feature extraction. The advantage of using YAMNet models is that it has less computationally expensive and satisfies the resource constraints of most mobile devices. However, the results have shown significantly lower performance than other models. This might be because the YAMNet was trained on environmental sounds such as barks, sirens, etc. That's why it has not been chosen as a proposed model. However, the idea of using less computationally expensive models remains actively relevant.

Chapter 4

Conclusion

In this study, a machine learning model Vision Transformers for heart classification was proposed and implemented, which patches spectrogram images into several smaller patches and uses a transfer encoder to classify data. In addition, other different types of models were tested that address the issue of using complex feature extraction methods in combination with complex machine learning models.

The performance of the proposed model was examined using the PhysioNet Heart Sound Dataset containing more than 3000 sounds heart sound audio files. The implemented model then was compared with other existing models such as ensemble CNN(1D-CNN+2D-CNN), RNN, SVM with MFCC, and time series signal features. The results have shown that in terms of sensitivity, the proposed model outperforms most of the models.

Results show better performance compared to existing methodologies via evaluation metrics. However, Vision Transformers might still be computationally expensive relative to less powerful computers. So, the need to develop simpler models that ideally use raw audio files remains an open research question.

Bibliography

[1] WHO, Cardiovascular diseases on WHO, 2017, last updated Oct. 2022. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab1 (Accessed October 10, 2023).

[2] Bozkurt B., Germanakis I., Stylianou Y. A study of time-frequency features for CNN-based automatic heart sound classification for pathology detection. Comput. Biol. Med. 2018;100:132–143. doi: 10.1016/j.compbiomed.2018.06.026.

[3] Humayun A.I., Ghaffarzadegan S., Ansari I., Feng Z., Hasan T. Towards Domain Invariant Heart Sound Abnormality Detection Using Learnable Filterbanks. IEEE J. Biomed. Health Inform. 2020;24:2189–2198. doi: 10.1109/JBHI.2020.2970252.

[4] Latif S., Usman M., Rana R., Qadir J. Phonocardiographic sensing using deep learning for abnormal heartbeat detection. IEEE Sens. J. 2018;18:9393–9400. doi: 10.1109/JSEN.2018.2870759.

[5] Noman F., Ting C.-M., Salleh S.-H., Ombao H. Short-segment heart sound classification Using an ensemble of deep convolutional neural networks; Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Brighton, UK. 12–17 May 2019; pp. 1318–1322.

[6] Li F, Zhang Z, Wang L, and Liu W. Heart sound classification based on improved mel-frequency spectral coefficients and deep residual learning. Front. Physiol. 13:1084420, 2022, doi: 10.3389/fphys.2022.1084420

[7] Abduh Z., Nehary E.A., Wahed M.A., Kadah Y.M. Classification of heart sounds using fractional fourier transform based mel-frequency spectral coefficients and traditional classifiers. Biomed. Signal Process. Control. 2019;9:1–8. doi: 10.1016/j.bspc.2019.101788.

[8] Dominguez-Morales J.P., Jimenez-Fernandez A.F., Dominguez-Morales M.J., Jimenez-Moreno G. Deep Neural Networks for the Recognition and Classification of Heart Murmurs Using Neuromorphic Auditory Sensors. IEEE Trans. Biomed. Circuits Syst. 2018;12:24–34. doi: 10.1109/TBCAS.2017.2751545.

[9] Xiao B., Xu Y., Bi 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 Things J. 2020;7:1994–2004. doi: 10.1109/JIOT.2019.2961132.

[10] Anderson RH, Brown NA. The anatomy of the heart revisited. Anat Rec. Sep, 1996; 246(1):1-7. doi: 10.1002/(SICI)1097-0185(199609)246

[11] Furst, John. What are the Four Chambers of the Heart? April, 2020. https://www.firstaidforfree.com/what-are-the-four-chambers-of-the-heart/ (Accessed Nov 2023)

[12] G. D. Clifford et al., "Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in Cardiology Challenge 2016," 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 2016, pp. 609-612.

[13] Brunton S. L., Kutz J. N. Data-driven science and engineering: Machine learning, dynamical systems, and control. – Cambridge University Press, 2022.

[14] Demir F., Şengür A., Bajaj V., Polat K. Towards the classification of heart sounds based on convolutional deep neural network. Health Inf. Sci. Syst. 2019;7:1–9. doi: 10.1007/s13755-019-0078-0.

[15] Oztavli E., Aptoula E. Effect of early and late fusion on heart sound classification. 2018 26th Signal Processing and Communications Applications Conference (SIU). IEEE, 2018

[16] Krizhevsky A., Sutskever I., Hinton G.E. ImageNet classification with deep convolutional neural networks; Proceedings of the Neural Information Processing Systems Foundation; Lake Tahoe, NV, USA. 3–6 December 2012; pp. 1090–1105.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.

[18] Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M., et al. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis. 2015;115:211–252. doi: 10.1007/s11263-015-0816-y