Heart Murmur Classification in Phonocardiogram Representations Using Convolutional Neural Networks

Mehlam Shabbir, Xudong Liu, Mona Nasseri, Scott Helgeson
University of North Florida
John E. Mathews Jr. Computer Science
UNF Dr., Jacksonville, FL 32224
n01474570@unf.edu, xudong.liu@unf.edu, mona.nasseri@unf.edu, Helgeson.Scott@mayo.edu

Abstract

Heart murmurs are sounds made by rapid blood flow in the heart. Abnormal heart murmurs can be a sign of serious heart conditions such as arrhythmia and cardiovascular diseases. Therefore, heart murmur classification is crucial for early detection of such conditions. To this end, we study the heart murmur classification problem training selected convolutional neural network (CNN) models (such as VGGNet and ResNet) using various signal representations (such as spectrogram and mel-frequency cepstral coefficient (MFCC)) of the phonocardiograms in the public PASCAL CHSC dataset. Our preliminary results show that ResNet outperforms VGGNet across all metrics and representations, consistent with the recent published works we can find in literature. Unlike some of these works, however, we see MFCC and spectrogram images of larger hop length, in general to be more effective with higher test accuracies than spectrograms with reduced hop lengths across all CNN models. Looking forward, we propose to study other effective models (such as InceptionV3 and Vision Transformer) to predict heart murmur conditions in phonocardiogram representations including spectrogram and MFCC as well as others like Wigner Ville distribution.

Introduction

Cardiovascular disorders (CVDs) pose a significant threat to global health, with 17.9 million deaths in 2019 accounting for 31% of all mortalities(Benjamin et al. 2019). Normal hearts produce a similar pattern as a result of heart valve closure and opening, which can be observed in healthy individuals. Any deviation from this pattern is considered an abnormality, such as a murmur. A common public dataset many researchers have used for their studies, we also use in this work, is the PASCAL CHSC 2011 database (Gomes et al. 2013), which consists of two subsets totaling 832 audio snippets belonging to 5 classes (Artifacts, Normal, Murmur, Extrasystole, and Extra heart sounds).

Recent studies on PASCAL CHSC have employed emerging strategies to significantly enhance the accuracy of machine learning (ML) models. For instance, Bourouhou et al. applied ML models such as K-nearest neighbors, support vector machines, decision trees, and naive Bayes (Bourouhou et al. 2020). But this work considers the two subsets separately and results are improvable. Another study done by Almanifi et al., the most recent one we could find, conduct experiments to classify these heart audios into the five classes using VGGNet and ResNet for spectrogram and mel-frequency cepstral coefficient (MFCC) two signal representations (Almanifi et al. 2022). While this work provides insights into the performances of deep neural models using transfer learning, it lacks details about hyperparameter tuning, which makes repeating these results infeasible, and it has not considered other signal features such as Wigner Ville distribution (WVD).

To this end, our work intends to address these aspects. Specifically, the objectives of this research are: (1) to show effectiveness of different signal representations (i.e., variations of spectrograms and MFCC), (2) to demonstrate the performance of various transfer learning CNNs, such as VGGNet and ResNet, in terms of test accuracy and loss, and (3) to present the optimal hyperparameter values that result in optimal performance of these predictive models.

Signal Representations

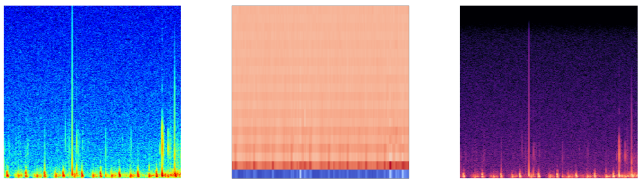
Signal representation methods are techniques used to represent signals like time-series information such as an audio file, in a concise and meaningful manner. The goal of signal representations is to extract the most important features of a signal, which can be used for further analysis and processing. We focus on two types of signal representation, the spectrogram and MFCC, while for spectrogram we tried two variations of different hop length and number of points in the Fast Fourier Transform (FFT). An example of each of them is included in Figure 1.

A spectrogram is a graphical representation of the frequency spectrum of a signal over time. It displays the distribution of frequencies present in a signal as it changes over time. Usually, the x-axis represents time, while the y-axis represents frequency, and the color intensity represents the amplitude or power of the signal at a given frequency.

MFCCs are a condensed representation of spectral information in speech and audio processing. Computed by transforming the signal from time-domain to frequency-domain using Fourier transform, mapping to the Mel-scale, and then back to time-domain using inverse Mel-frequency cepstral transform, retaining only the low-frequency coefficients. MFCCs are utilized in speech recognition, speaker identification, and music classification tasks. (Ghosal et al.

Table 1: Averaged performance across 5 runs for three CNN models trained on three signal representations

		VGG16			VGG19			ResNet50		
		Spec I	MFCC	Spec II	Spec I	MFCC	Spec II	Spec I	MFCC	Spec II
Categorical Accuracy	Train	90.48%	99.56%	99.86%	90.11%	76.90%	85.17%	100%	99.53%	99.93%
	Test	74.00%	76.00%	78.80%	70.80%	74.00%	72.80%	76.80%	76.40%	**80.80%**
Loss	Train	0.3344	0.0625	0.0359	0.3341	0.6862	0.5099	0.0675	0.0842	0.0664
	Test	1.5531	0.9832	3.7522	1.8805	0.9282	1.3075	0.9171	0.8327	0.9699



(a) Spectrogram I (b) MFCC (c) Spectrogram II

Figure 1: Signal representations of a PASCAL audio. Spectrogram I (hop length=128, no. fft=256), Spectrogram II (hop length=512, no. fft=2048)

2012).

The Short-time Fourier Transform (STFT) is a type of signal representation that provides a time-frequency representation of a signal (Sejdić, Djurović, and Jiang 2009). It involves dividing a signal into overlapping segments, or "windows," and computing the Fourier Transform of each segment to obtain its frequency content. By varying the size and overlap of the windows, the STFT allows for a detailed analysis of the signal's frequency content at different points in time. A smaller hop length value results in more overlap between frames, while a larger hop length value results in less overlap.

Experiment Setup and Tools

We used Python packages like Librosa for signal representations, Matplotlib for simple plotting, and TensorFlow/Keras for transfer learning models. The model was created by using a pre-trained feature extraction model from ImageNet, flattened and passed through a 20% dropout and a dense layer with 64 units, using a SeLU activation function and L2 regularization with the penalty term set to 0.0005. The final step was connecting the model to an output layer with a softmax activation function. The model was standardized with consistent layers and trained with different signal representation images.

Consistent with the most recent work (Almanifi et al. 2022), we only keep audio signals with at least 4 seconds of play time. This reduces the dataset to 404 samples. An 80:20 train-test split ratio was utilized to train and test the models. In Table 1, each experiment has been repeated 5 times for 50 epochs and the averages are reported.

Hyperparameter Tuning

Hyperparameter tuning is an important step in the process of training a deep learning model, including a transfer learning model. In this particular transfer learning model, the hyperparameters include the learning rate, batch size, image dimension, L2 regularization alpha value, and drop out layer ratio. After grid search, we set the learning rate to 0.00001, batch size to 50, image dimension to 400x400, L2 regularization alpha value to 0.0005, and drop out ratio to 0.2.

By carefully tuning these hyperparameters, we can ensure that the transfer learning model is able to generalize well to new data and achieve good performance on the target task. The values mentioned in the discussion are the ones that were utilized for the experiment performed, but they require further modification to attain more favorable outcomes.

Preliminary Results

The results of the preliminary experiments are summarized in Table 1. This table provides the performance metrics for three different convolutional neural network models trained on three different types of audio representations. The performance metrics reported are categorical accuracy and loss for both the training and testing sets. Overall, the results show that ResNet50 outperformed both VGGNet models in terms of both test accuracy and test loss values for all three audio features. This is in consistency with the most recent published results that we could find on the same dataset (Almanifi et al. 2022).

Furthermore, the results show that MFCC and Spec II representations are more effective with higher test accuracies than Spec I in all but one case. This is a different finding in our experiments to those aforementioned.

Additionally, our experiments consider Spec II the highest testing accuracy achieved by ResNet50 trained on STFT audio features of 80.80%.

Conclusion and Future Work

Preliminary results of the experiments shows great promise in a method that can utilize CNNs in modeling for predictions to classify for heartbeat murmurs. There are several signal representations that can be considered such as WVD. Further experiments needs to be conducted on further hyperparameter tuning with more extensive grid search as the preliminary results shows signs of overfitting. We also plan to study other neural network models such as InceptionV3 and Vision Transformer using other representations like WVD.

References

Almanifi, O. R. A.; Ab Nasir, A. F.; Razman, M. A. M.; Musa, R. M.; and Majeed, A. P. A. 2022. Heartbeat murmurs detection in phonocardiogram recordings via transfer learning. *Alexandria Engineering Journal* 61(12):10995–11002.

Benjamin, E. J.; Muntner, P.; Alonso, A.; Bittencourt, M. S.; Callaway, C. W.; Carson, A. P.; Chamberlain, A. M.; Chang, A. R.; Cheng, S.; Das, S. R.; et al. 2019. Heart disease and stroke statistics—2019 update: a report from the american heart association. *Circulation* 139(10):e56–e528.

Bourouhou, A.; Jilbab, A.; Nacir, C.; and Hammouch, A. 2020. Heart sound signals segmentation and multiclass classification.

Ghosal, A.; Chakraborty, R.; Dhara, B. C.; and Saha, S. K. 2012. Music classification based on mfcc variants and amplitude variation pattern: a hierarchical approach. *International Journal of Signal Processing, Image Processing and Pattern Recognition* 5(1):131–150.

Gomes, E. F.; Bentley, P. J.; Pereira, E.; Coimbra, M. T.; and Deng, Y. 2013. Classifying heart sounds-approaches to the pascal challenge. In *HEALTHINF*, 337–340.

Sejdić, E.; Djurović, I.; and Jiang, J. 2009. Time–frequency feature representation using energy concentration: An overview of recent advances. *Digital signal processing* 19(1):153–183.