HEART MURMUR CLASSIFICATION: UTILIZING MEL-SPECTROGRAMS WITH MACHINE LEARNING TECHNIQUES

Syed Mahdi

M.S. Computer Science
College of Natural Sciences

The University of Texas at Austin

Table of Contents

ABSTRACT

Cardiac auscultation is a technique that varies among medical professionals, with results dependent on individual skill and experience. Machine learning has seen advances in its applications in the medical field and could also be applied for audio classification of heartbeats. Using the Heart Sounds dataset by Ed King from Kaggle, this paper will explore the various methods for analyzing and interpreting audio data. By comparing Convolutional Neural Networks and Long Short-Term Memory neural networks, the research demonstrates a foundation for machine learning assisted cardiac auscultation, along with avenues for further work and improvement.

1. INTRODUCTION

Heart murmurs are irregular or abnormal sounds that are usually detected via cardiac auscultation using tools such as stethoscopes [1]. Doctors are able to discern variations in pitch, volume, and rhythm to identify murmurs and changes in blood flow. While murmurs can be benign, they can often be indicators of something more serious, such as septal defects and other valvular pathologies [2]. Their detection and precise interpretation are vital for patients to receive appropriate care, especially considering, according to the World Health Organization, cardiovascular disease is the leading cause of death globally [3].

Traditional cardiac auscultation, the method of listening to heart sounds to diagnose conditions, is a skill that varies a great deal among medical professionals. While it remains a cornerstone in cardiac diagnostics, its effectiveness depends on an individual doctor's experience and auditory acuity. This variance in diagnostic capabilities was highlighted in a study where a group of 78 healthcare professionals at a conference on cardiac auscultation were tested on their ability to identify heart sounds, including heart murmurs, and had resulted in a post-conference score of 44.7% [4].

The medical industry has benefited greatly from the improvements in machine learning, ranging from use cases in natural language processing with healthcare research to gaining meaningful insights from health data and informatics [5][6]. To address some of the challenges mentioned earlier with detecting murmurs, we can apply machine learning to identify patterns that may otherwise be difficult to ascertain using traditional methods. Doing so could help medical professionals catch potentially serious cardiac conditions more effectively, ultimately improving the quality of care that patients receive. This paper explores methods of representing, processing, and analyzing heartbeat audio data, focusing on the application of various machine learning models and techniques to classify heartbeats and detect murmurs using Mel-spectrograms.

2. RESEARCH BACKGROUND

Prior to implementing machine learning towards heartbeat audio, this paper will first explore the various ways of representing audio.

2.1. AMPLITUDE WAVEFORM

In the real world, sound waves can be characterized as continuous, mechanical waves. This differs clearly from computer use cases, where data is required to be discrete and digital [7]. This makes converting the mechanical waves into a digital representation key for any audio analysis. By sampling fluctuations in the air pressure at a microphone's position, electrical output can be converted into a digital format [8]. This ultimately allows us to chart the waveforms of the signals; in figure 2.1.1 a waveform is charted by plotting the amplitude against the time of the audio, indicating the strength of the signal at a given point in time.

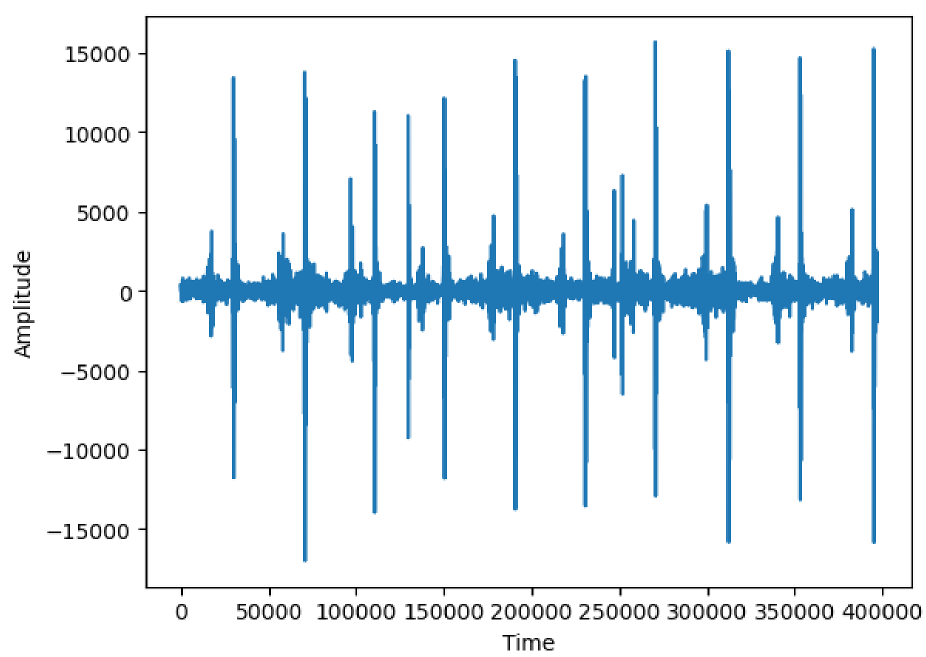


Figure 2.1.1: Waveform generated for a heart murmur audio file.

2.2. SHORT TIME ENERGY

The short time energy (STE) is typically used for speech processing, as it expresses the energy associated with a given audio segment. While it's widely used in speech processing, it can also be relevant with non-speech related tasks as well [9]. Heartbeat audio would fall under this domain, as the "lub-dubs", or the S1 and S2 sounds, in a heartbeat audio sample can result in various auditory energy levels. The STE for an audio clip can be defined as the sum of the squares for each signal within a sample [10]:

$$E = \sum x^2(m)$$

Figure 2.2.1 shows the STE for the same audio clip of a murmur used in previous figures.

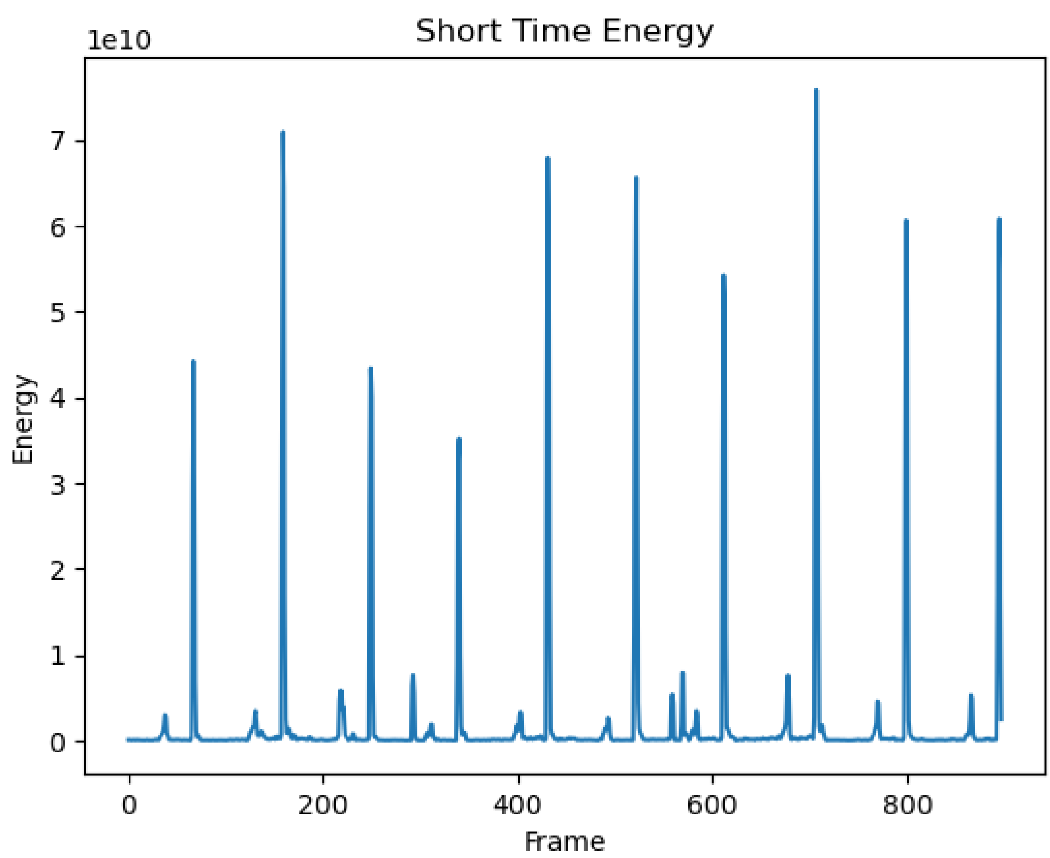


Figure 2.2.1: Short Time Energy (STE) for murmur

2.3. ZERO-CROSSING RATE

The zero-crossing rate (ZCR) can be defined as the rate at which a given region of a signal goes from positive to negative or back [11]. This value can be calculated for a region in each signal as the number of times a zero-crossing occurs, divided by the samples within that region [12]:

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} 1_{\mathbb{R}<0}(S_t S_{t-1})$$

Compared to other audio features discussed in the paper, information such as amplitude, frequency, and energy across time are not conveyed with the zero-crossing rate. While it is a useful metric, it wouldn't be sufficient on its own for the purposes of classifying heartbeat audio.

Figure 2.3.1 shows the zero-crossing rate for the murmur audio clip as used in previous figures.

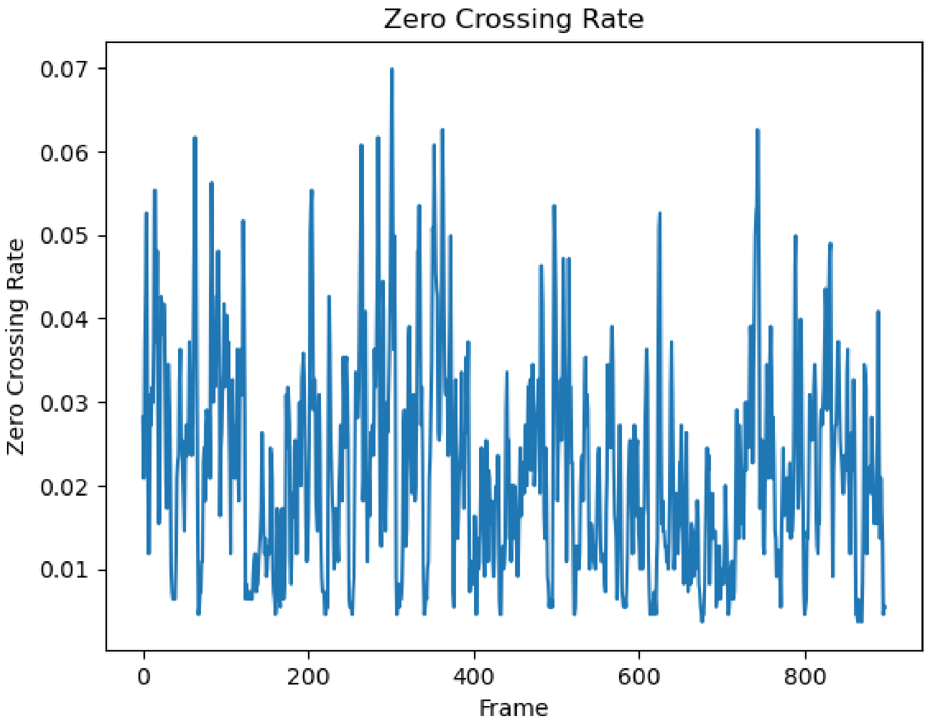


Figure 2.3.1: Zero-crossing rate for murmur

2.4. MEL-SPECTROGRAM

Spectrograms have been used in audio pattern recognition, and as Convolutional Neural Networks (CNN) have become more popular, the features that spectrograms can provide them have proven to be quite effective [13]. According to the Weber-Fechner law, humans perceive sound logarithmically relative to stimulus intensity [14]. Essentially, this means that humans are able to discern changes at lower frequencies with much higher sensitivity than they can with higher frequencies. This leads way to the Mel scale, which logarithmically transforms the frequencies of a given audio input to better match human perception [15]. O'Shaughnessy provides the following formula that allows one to convert hertz (*f*) into mels (m) [16]:

$$m = 2595\ log_{10}(1 + \frac{f}{700})$$

The Mel-spectrogram is a spectrogram that results from converting frequencies into the Mel scale, meaning that they're logarithmically displayed to better align with human auditory perception [17]. Since the Mel scale will emphasize the frequencies most significant to human hearing, and the Mel-spectrogram offers rich information regarding frequency and decibel levels across time, it can prove to be an effective feature to classify heartbeat audio with.

Figure 2.3.1 shows the Mel-spectrogram for the murmur audio clip as used in previous figures. Note that the Y-axis, which represents the frequency in hertz, is logarithmically scaled, and that the color bar found on the right indicates decibel levels.

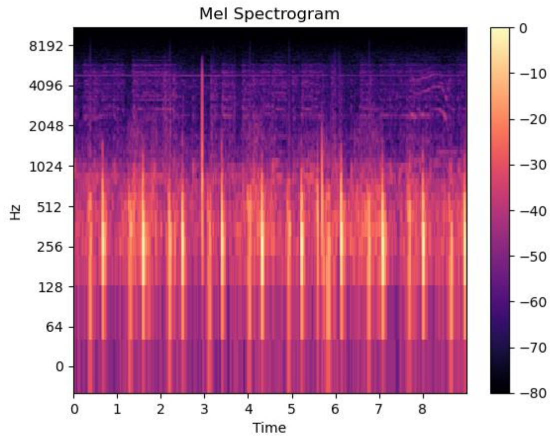


Figure 2.4.1: Mel-spectrogram for murmur

For the reasons mentioned above, this paper will explore the Mel-spectrogram feature as the input the machine learning models.

3. MATERIALS / DATA / SOURCES

To train a classifier to identify murmurs within heartbeats, we will use the Heartbeat Sounds dataset from Kaggle created by Ed King. This dataset contains heartbeat audio samples from two sources, the first source being the general public through the iStethoscope app, and the second source being from a clinic trial in hospitals [18].

3.1. DATA ANALYSIS

The data is presented in two separate sets, set A and set B, where each set corresponds to one of the two sources the data was retrieved from. The audio files contained within both sets vary in duration, ranging from 0.76 seconds to 27.87 seconds, as can be seen in figure 3.1.1. To ensure that our audio samples contain multiple “lub-dubs”, or S1 and S2 sounds, of a heartbeat, our first step is to filter out any audio samples that are less than 3 seconds in duration.

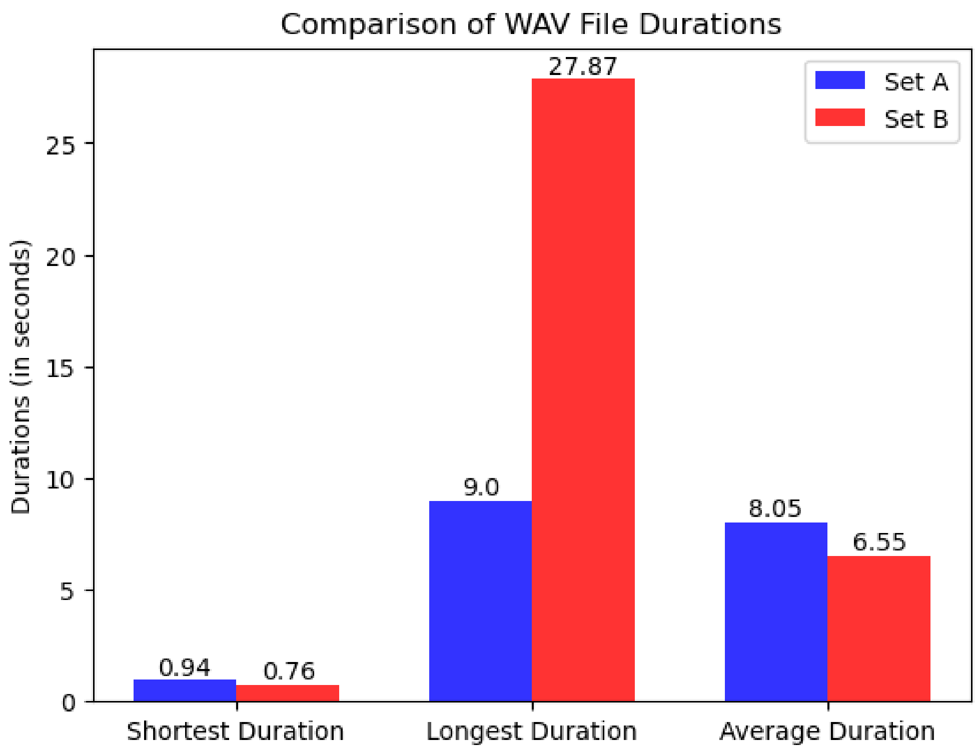


Figure 3.1.1: Comparison of WAV file durations

Additionally, the dataset also contains unlabeled data, which can't be used for training or testing. We further filter down the dataset by removing the unlabeled audio samples, as it would be unreasonable to manually label them without a trained medical professional to help. After ensuring that we have relevant data to train and test with, we're left with 462 audio samples distributed across five categories: normal, murmur, extrastole, extrahls, and artifact.

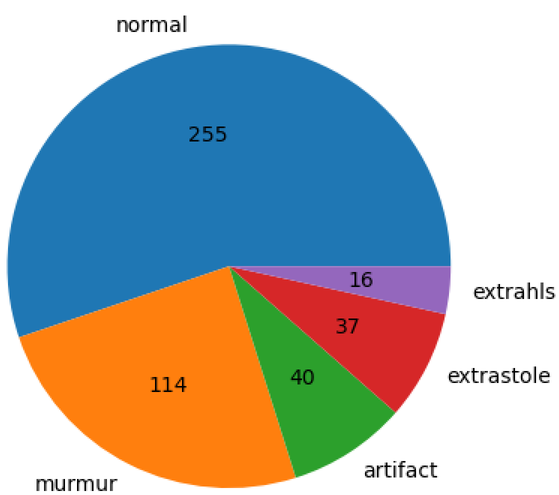


Figure 3.1.2: Breakdown of heartbeat categories within the dataset

3.2. NORMAL HEARTBEATS

The normal category indicates that there is nothing wrong with the given heartbeat, and can be characterized by constant, distinct "lub-dubs" [19]. The first heart sound that's heard is known as S1, which corresponds to the "lub" sound in a heartbeat. This is then followed by the second heart sound, known as S2, which corresponds to the "dub" sound in a heartbeat. We can see the waveform of a normal heartbeat in figure 3.2.1, where we can clearly distinguish the distinct heartbeat sounds as amplitude increases at a steady rhythm with minimal noise in between them. We can also see the Mel-spectrogram in figure 3.2.2 of this normal heartbeat audio, where we see similar patterns as the waveform, along with patterns of how decibel levels change across the frequency band.

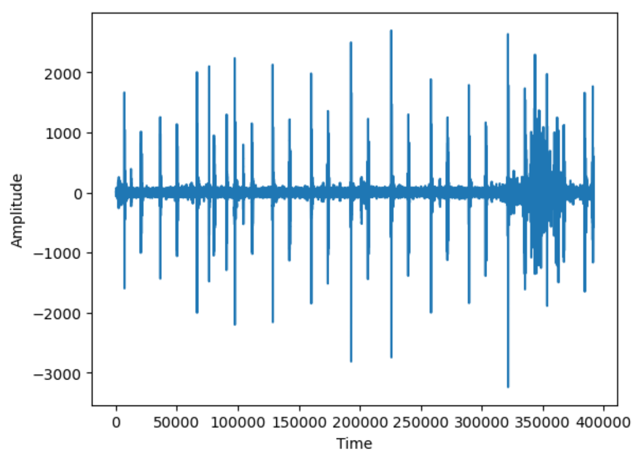


Figure 3.2.1: Waveform of a normal heartbeat

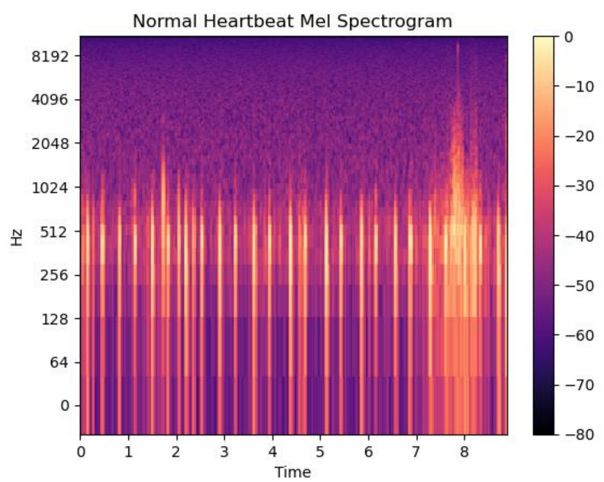


Figure 3.2.2: Mel-spectrogram of a normal heartbeat

3.3. MURMURS

As mentioned earlier in this paper, murmurs are abnormal sounds within a heartbeat, which can either be innocent or indicators of more serious problems with the heart. They're a result of turbulence with blood flow within the heart and can be identified by characteristics such as timing and noise [21]. Looking at the waveform and Mel-spectrogram of a murmur heartbeat in figures 3.3.1 and 3.3.2, we can see the distinct peaks for the heart's S1 and S2 sounds are not as cleanly separated. Instead, especially when compared to normal heartbeats, we find areas of slight noise around those peaks.

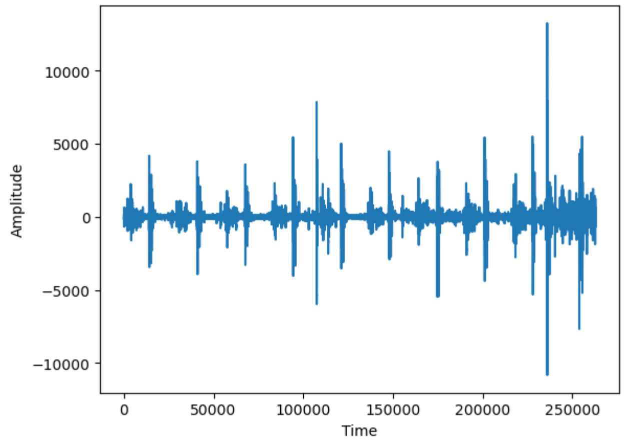


Figure 3.3.1: Waveform of a murmur heartbeat

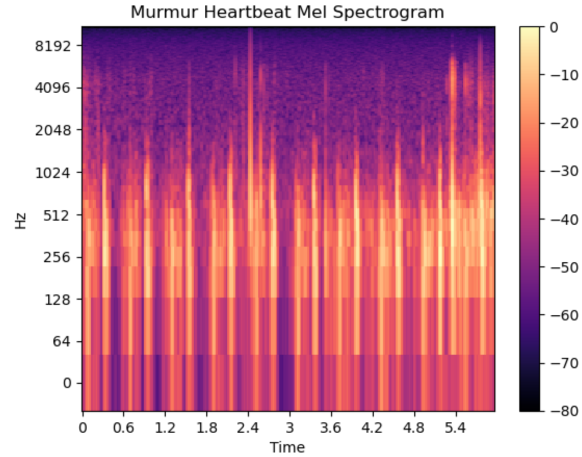
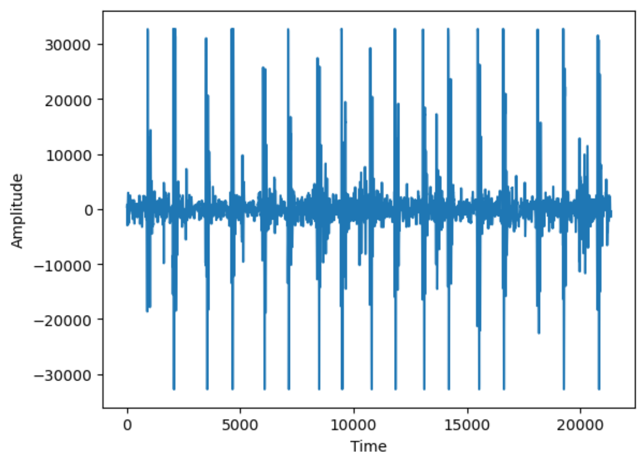


Figure 3.3.2: Mel-spectrogram of a murmur heartbeat

3.4. EXTRASYSTOLES

Our dataset also contains samples of heartbeats with extrasystoles within them, labeled as extrastole within the data. An extrasystole is an additional heartbeat which occurs outside of a heart's normal rhythm. These additional heartbeats can also lead to serious symptoms within patients and are heard faintly next to the regular S1 and S2 sounds of a heartbeat [22]. The waveform and Mel-spectrograms of a heartbeat with extrasystoles in figures 3.4.1 and 3.4.2 are somewhat similar to murmurs in that they contain extra sounds in between the peaks caused by the S1 and S2 sounds, though in this case they're caused by the faint extra heartbeats. Another interesting observation is that the peaks are at irregular intervals, which is a result of the extra faint heartbeats.



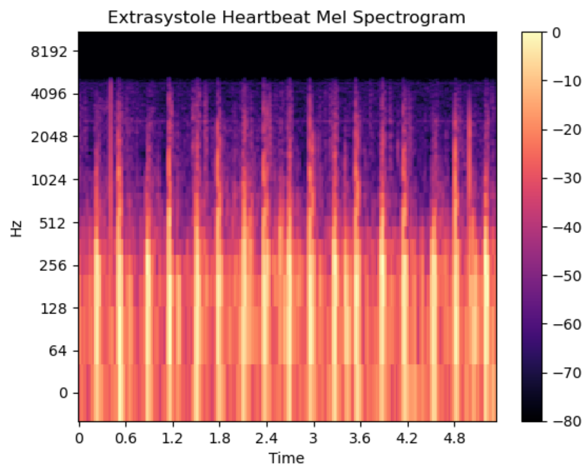


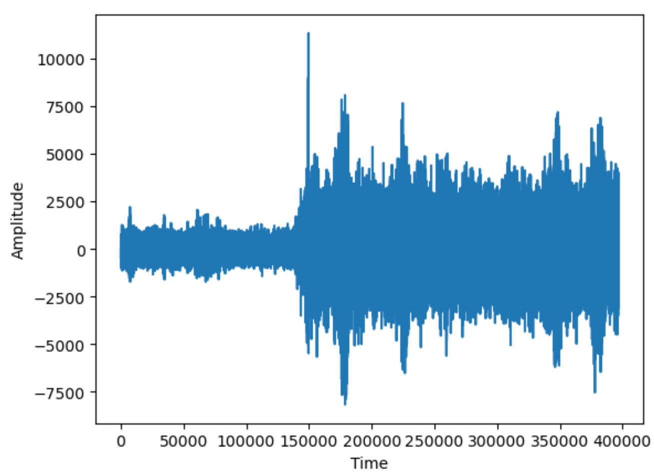
Figure 3.4.1: Waveform of extrasystoles

Figure 3.4.2: Mel-spectrogram of extrasystoles

Training a model to classify extrasystoles within heartbeats would also be an important aid for medical professionals. Since they tend to be faint, machine learning can help medical professionals better identify them.

3.5. ARTIFACTS AND EXTRA HEART SOUNDS

Artifacts and extra heart sounds are important to classify since they'd indicate to the user that they need to try recording the audio sample again and reattempting classification. Artifacts are non-heart sounds and clearly cannot be used for any heartbeat classification. Likewise, extra heart sounds, labeled as extrahls within the dataset, are heart sounds with extra sounds in the background, making labeling and classification difficult.



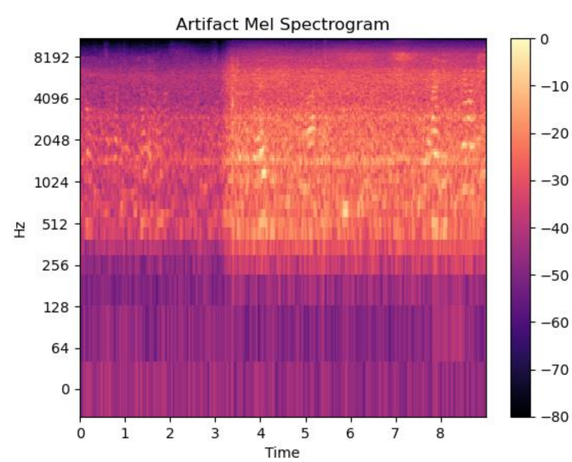


Figure 3.5.1: Waveform for artifact

Figure 3.5.2: Mel-spectrogram for artifact

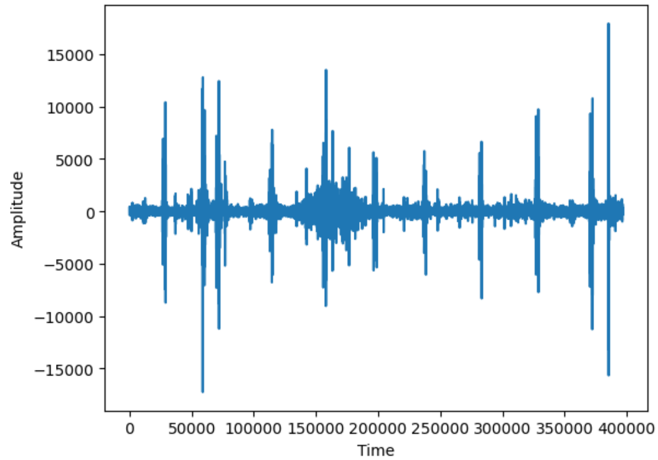


Figure 3.5.3: Waveform for extra heart sounds

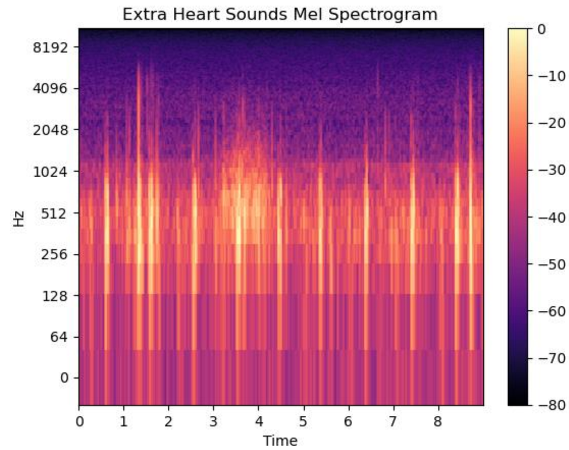


Figure 3.5.4: Mel-spectrogram for extra heart sounds

4. METHODS

Prior to training, we split our filtered dataset by setting aside 80% of our data for training and 20% for testing. We then extract the Mel-spectrogram features for our training and testing data.

4.1. FEATURE SCALING

As the extracted Mel-spectrogram data is numeric, Min-Max normalization was performed to rescale the data within a range of 0 and 1, and is a linear transformation that can be defined with the following formula [23]:

$$X_{new} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

By normalizing the data in this way, we can lessen the effects that outliers would have on the training of our models. Normalization during the data preprocessing stage can help improve the performance of machine learning algorithms and has been applied here to help boost the performance of the models we'll be testing with [24].

4.2. CONVOLUTIONAL NEURAL NETWORK (CNN)

Convolutional neural networks (CNN) have been used for audio-based machine learning, such as for tasks like music classification using spectrograms [25]. Seeing the success of CNNs with audio classification suggests similar results can be achieved in the case of heartbeat audio classification.

The CNN model used was comprised of two 2-dimensional convolutional layers. Both convolutional layers have 32 filters applied, where the first layer takes in the input features and the second layer takes in the output of the first set of layers. Batch normalization is applied after the convolutional layers to stabilize the inputs between the layers, maintaining the mean output near 0 and the output standard deviation near 1 [26]. This is then followed by a 2-dimensional max pooling layer for both our convolutional layers. A regular densely connected neural network layer with 128 units is then applied and subsequently flattened, followed by a dropout layer with a rate of 0.5. The model is finished off with a dense output layer with 5 units and softmax activation for classification, and results in 479,973 parameters.

Training was done using the Adam optimizer, and techniques such as learning rate schedulers and early stopping were applied to supplement training. The results of the best-case training achieved after hyperparameter tuning are shown in figure 4.2.1.

Figure 4.2.1: CNN accuracy and loss results from training and validation

4.3. LONG SHORT-TERM MEMORY (LSTM)

Like CNNs, Long Short-Term Memory (LSTM) models have also been successfully used for audio classification purposes [28]. Since audio signals are inherently sequential, and LSTM networks are designed to process sequences of data, they're well suited for classifying heartbeat audio [28].

The LSTM model used begins with a bidirectional LSTM layer with 64 units, along with a dropout and recurrent dropout rate of 0.1 to help prevent overfitting. Additionally, this bidirectional LSTM layer also employs L2 regularization to combat overfitting during training [29]. We then apply batch normalization, followed by a dropout layer with a rate of 0.5. The model then includes a regular densely connected neural network layer with 32 units, which is also regularized with L2 regularization, followed by one more layer of batch normalization. A flatten layer is applied to transform the data for the final output layer, which is a dense layer with 5 units, one per classification category, and softmax activation. This model ends up with a total of 279,205 parameters, significantly less than what our CNN model had.

Training was performed in a similar manner to our CNN model, where we used the Adam optimizer and techniques such as learning rate scheduling and early stopping. The results of the best-case training achieved after hyperparameter tuning are shown in figure 4.3.1.

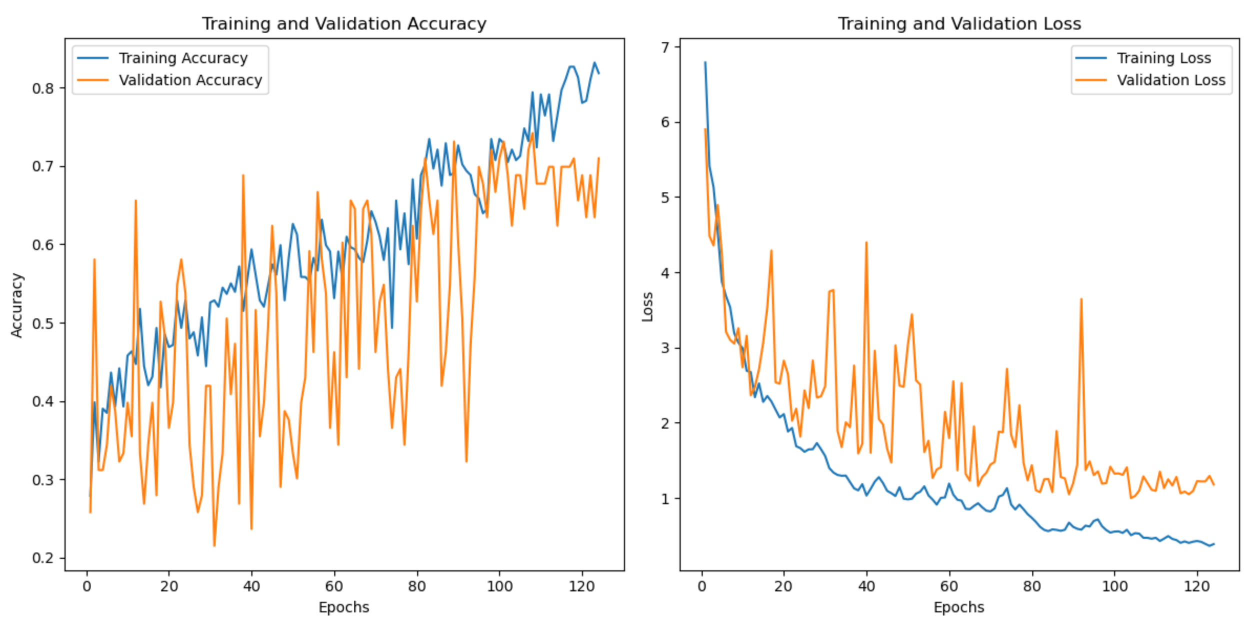


Figure 4.3.1: LSTM accuracy and loss results from training and validation

5. RESULTS

Examining the result of the training for both the CNN and LSTM models, we can see that there's an issue of overfitting. The training accuracy and loss for both models perform well, but both models struggled to generalize during validation. One of the most likely reasons for this overfitting issue with the training of both is the class imbalance, as can be seen in figure 3.1.2. There's an overwhelming amount of data labeled as normal, while certain classes such as artifact, extrastole, and extrahls are not represented nearly as much. Both achieved validation scores of nearly 70%, though considering both models' performances and parameter sizes, the LSTM model was much more computationally efficient since it had significantly fewer parameters than the CNN model.

Two additional metrics we will use to examine the results of how well the models performed classification are sensitivity and specificity. Sensitivity is a measure of true positives that are correctly identified by the model, where a higher sensitivity indicates the model's ability to catch positives and not miss them. Sensitivity can be calculated as follows, where TP represents the number of true positives and FN represents false negatives [30]:

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity on the other hand is a measure of true negatives that are correctly identified by the model, where a higher specificity indicates the model's ability to avoid false alarms. Specificity can be calculated as follows, where TN represents the number of true negatives and FP represents false positives [30]:

$$Specificity = \frac{TN}{TN + FP}$$

These are both important metrics to perform well in, as a medical professional using a model such as the ones that we tested would want to ensure that actual cases of disease are caught (through a high sensitivity score), and that healthy patients are not falsely diagnosed with a condition they don't have (through a high specificity score).

In figure 5.1, the CNN model showed varying degrees of sensitivity across the different classes. Sensitivity varied throughout the classes, most notably showing around a 0.76 sensitivity score for normal heartbeats and 0.62 for murmurs. The CNN model had a poor sensitivity score of 0.12 for extrasystoles however, indicating a weakness at catching these within heartbeats. Specificity remained high throughout the classes, indicating effectiveness at ruling out conditions when they aren't present.

As shown in figure 5.2, the LSTM model resulted in superior sensitivity for normal heartbeats with a result of 0.8, though it performed slightly worse at identifying murmurs with a sensitivity score of 0.54. Similar to the CNN model, extrasystoles classification was equally weak for the LSTM model with a sensitivity of 0.12. Specificity was also scored highly across the categories for the LSTM model.

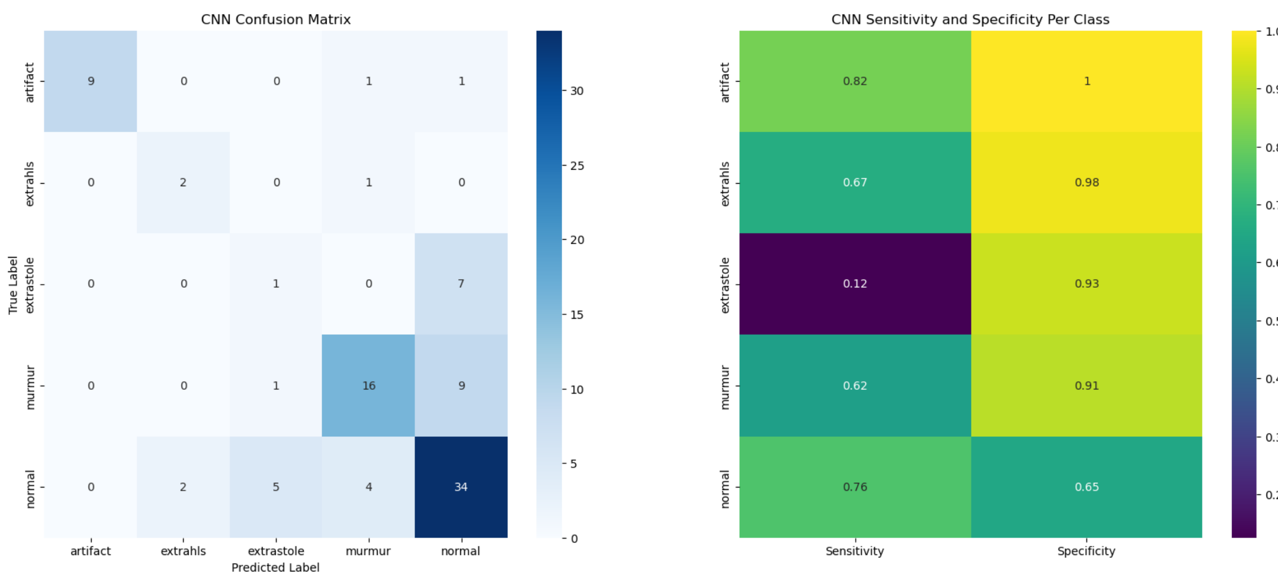


Figure 5.1: CNN confusion matrix and sensitivity-specificity per class

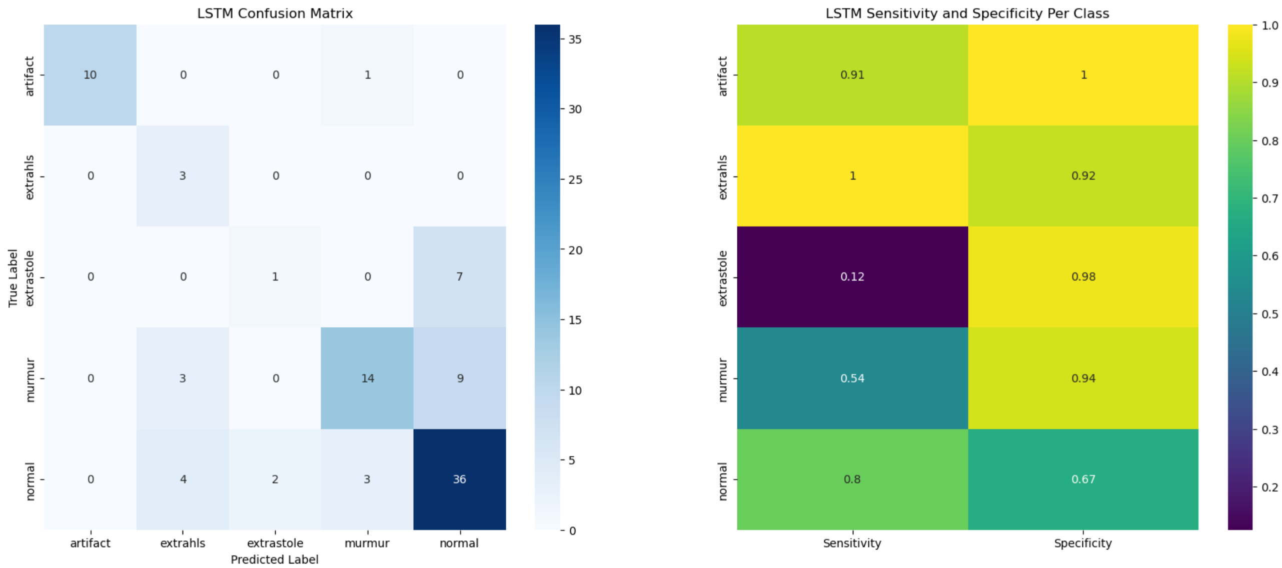


Figure 5.1: LSTM confusion matrix and sensitivity-specificity per class

6. DISCUSSION & CONCLUSION

This paper aimed to harness machine learning, particularly Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, to classify heartbeat audios and identify heart murmurs. We explored and compared the various methods of representing audio for use with training these models, ultimately deciding on leveraging Mel-spectrograms for use during training. Both the CNN and LSTM models demonstrated a fair degree of accuracy in classifying heart murmurs and normal heartbeats. They both exhibited a proficient ability to recognize true negatives as evidenced from their high specificity scores across the classes, which is critical in clinical settings to avoid potential false alarms. While they performed similarly, the LSTM model was able to achieve its results in a more computationally efficient manner. However, with the current results they are not fit for actual clinical use and require further tuning and improvements. Overfitting remained a challenge for the CNN and LSTM models, as they were not able to generalize their training performances. This was likely a result of the class imbalance in the dataset, where less represented categories generally resulted in poorer performance.

The research presented in this paper demonstrates the abilities of employing machine learning techniques to support cardiac auscultation. While the current models require further improvements, they offer a foundation for developing more robust tools to aid medical professionals with identifying murmurs and other cardiac conditions. Future work would focus on improving the dataset to overcome the class imbalance, which would lead to more accurate and efficient models. Exploring model performance using different input features, such as the ones explored in this paper, could also yield better results.

REFERENCES

1. Elola, A., Aramendi, E., Oliveira, J., Renna, F., Coimbra, M. T., Reyna, M. A., Sameni, R., Clifford, G. D., and Rad, A. B. (2023). Beyond heart murmur detection: Automatic murmur grading from phonocardiogram. IEEE Journal of Biomedical and Health Informatics, 27(8), 3856–3866. https://doi.org/10.1109/jbhi.2023.3275039
2. Thomas SL, Heaton J, Makaryus AN. Physiology, Cardiovascular Murmurs. [Updated 2023 Jul 17]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK525958/
3. World Health Organization. (2022). Cardiovascular Diseases. World Health Organization. https://www.who.int/health-topics/cardiovascular-diseases
4. March, S. K., Bedynek, J. L., & Chizner, M. A. (2005). Teaching Cardiac Auscultation: Effectiveness of a Patient-Centered Teaching Conference on Improving Cardiac Auscultatory Skills. *Mayo Clinic Proceedings*, *80*(11), 1443–1448. https://doi.org/10.4065/80.11.1443
5. Tan Yen Kheng. (2021). Smart Manufacturing. BoD – Books on Demand.
6. Aziz Sheikh, & Savill, J. (2017). Key advances in clinical informatics: transforming health care through health information technology. Academic Press.
7. mlearnere. (2021, April 14). Learning from Audio: Wave Forms. Medium. https://towardsdatascience.com/learning-from-audio-wave-forms-46fc6f87e016
8. Bhanu Prasad, & S R Mahadeva Prasanna. (2008). Speech, audio, image and biomedical signal processing using neural networks. Springer.
9. Short Term Time Domain Processing of Speech (Theory) : Speech Signal Processing Laboratory : Biotechnology and Biomedical Engineering : Amrita Vishwa Vidyapeetham Virtual Lab. (n.d.). Vlab.amrita.edu. Retrieved November 28, 2023, from https://vlab.amrita.edu/?sub=3&brch=164&sim=857&cnt=1
10. Jalil, M., Butt, F. A., & Malik, A. (2013, May 1). Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals. IEEE Xplore. https://doi.org/10.1109/TAEECE.2013.6557272
11. Chen, C. H. (1988). Signal processing handbook. Dekker.
12. Fabien Gouyon, François Pachet, & Olivier Delerue. (2000). ON THE USE OF ZERO-CROSSING RATE FOR AN APPLICATION OF CLASSIFICATION OF PERCUSSIVE SOUNDS.
13. Md. Istiaq Ansari, & Hasan, T. (2022). SpectNet : End-to-End Audio Signal Classification Using Learnable Spectrograms. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2211.09352

14. Varshney, L. R., & Sun, J. Z. (2013). Why do we perceive logarithmically? *Significance*, *10*(1), 28–31. https://doi.org/10.1111/j.1740-9713.2013.00636.x
15. Volkmann, J., Stevens, S. S., & Newman, E. B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, *8*(3), 208–208. https://doi.org/10.1121/1.1901999
16. O'Shaughnessy, D. (n.d.). *Speech Communications: Human And Machine (ieee)*. Universities Press.
17. Roberts, L. (2020, March 14). *Understanding the Mel Spectrogram*. Medium. https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53
18. King, E. (2016). Heartbeat Sounds: Classifying heartbeat anomalies from stethoscope sounds. Retrieved October, 8 2023, from https://www.kaggle.com/datasets/kinguistics/heartbeat-sounds
19. *Heart Murmurs*. (n.d.). Texas Heart Institute. https://www.texasheart.org/heart-health/heart-information-center/topics/heart-murmurs
20. *Cardiac Cycle*. (n.d.). Www.utmb.edu. https://www.utmb.edu/pedi_ed/CoreV2/CardiologyPart1/CardiologyPart12.html
21. Pelech, A. N. (1998). THE CARDIAC MURMUR: When to Refer? *Pediatric Clinics of North America*, *45*(1), 107–122. https://doi.org/10.1016/S0031-3955(05)70585-X
22. *What are extrasystoles (palpitations)? | CardioSecur*. (n.d.). Www.cardiosecur.com. https://www.cardiosecur.com/magazine/specialist-articles-on-the-heart/what-are-extrasystoles-palpitations
23. Henderi, H. (2021). Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. *IJIIS: International Journal of Informatics and Information Systems*, *4*(1), 13–20. https://doi.org/10.47738/ijiis.v4i1.73
24. Kelsy Cabello-Solorzano, de, O., Peña, M. A., Correia, L., & Tallón-Ballesteros, A. J. (2023). The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis. *Lecture Notes in Networks and Systems*, 344–353. https://doi.org/10.1007/978-3-031-42536-3_33
25. Costa, Y. M. G., Oliveira, L. S., & Silla, C. N. (2017). An evaluation of Convolutional Neural Networks for music classification using spectrograms. *Applied Soft Computing*, *52*, 28–38. https://doi.org/10.1016/j.asoc.2016.12.024
26. *Keras documentation: BatchNormalization layer*. Keras.io. https://keras.io/api/layers/normalization_layers/batch_normalization/

27. Banuroopa, K., & Shanmuga Priyaa, D. (2021). MFCC based hybrid fingerprinting method for audio classification through LSTM. *International Journal of Nonlinear Analysis and Applications, 12*(Special Issue), 2125-2136
28. Jose, G. V. (2019, September 27). *Time Series Forecasting with Recurrent Neural Networks*. Medium. https://towardsdatascience.com/time-series-forecasting-with-recurrent-neural-networks-74674e289816
29. Anuja Nagpal. (2017, October 13). *L1 and L2 Regularization Methods*. Medium; Towards Data Science. https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c
30. Baratloo A, Hosseini M, Negida A, El Ashal G. *Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity*. Emerg (Tehran). 2015 Spring;3(2):48-9. PMID: 26495380; PMCID: PMC4614595.