HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 3 (2023)

Detecting Emotions in Audio Data of Patients with Post Traumatic Stress Disorder using Convolutional Neural Networks

Rohan Gupta

Gunn High School

ABSTRACT

As humans, we have an effortless ability and a high accuracy to identify another human's emotions through the tone, pitch and pace of their speech, even the emphasis and stress placed on each word. However, people with a traumatic experience suffering from Post Traumatic Stress Disorder (PTSD), 24.4 million people in the USA, can often repress emotions making it difficult for therapists to identify their patients genuine emotions and to treat them appropriately. Using an upcoming field of emotion detection in Artificial Intelligence (A.I.), I identified the human emotions from speech. Instead of using an audio transcription based model, I opted for a newer image based model, RESNET-18, which is widely used and utilizes spectrograms to preserve the subtleties in speech, critical in distinguishing emotions. To train the model, I used the RAVEDESS dataset which consists of wav files with eight different emotions. I was able to achieve an overall accuracy of 82% (greater than human detection by 25%). Specifically, I achieved 99% for no stress class (happiness), 97% for neutral class, (neutral, calm, and surprised), and 8\5% for stressed class (fearful, sadness, anger, disgust). I also found that the model got an accuracy of 87% when only trained on males, with continued training an overall accuracy above 90% is definitely achievable. In conclusion, it is possible to find the emotions of PTSD patients, and in the future, continued research can help improve the lives of people who are not able to express their true emotions.

Introduction

Post Traumatic Stress disorder (PTSD) was recognized as a real disorder in the 1980s, however, its history has much deeper roots. The first documented case of PTSD was 3000 years ago when Mesopotamian soldiers had slurred speech and depression after they came back from the battlefield. At the time, Mesopotamians thought that their soldiers were hexed by ghosts. PTSD was also documented in one of Shakespeare's monologues, Macbeth, where the main character speaks of her husband's inability to enjoy life after coming home from the battlefield. During World War I, PTSD was given its first name: shell shock. Due to the new horrifying inventions of World War I like poisonous gas and grenades, a large number of soldiers started to experience PTSD. The effects of PTSD continued, during the early 20th century, people with PTSD were often ostracized from their military peers and feared by society as a whole. However this would change after the Vietnam war when PTSD was recognized as a real disorder. Today, around 8 percent of the US population, 24.4 million people, experience Post Traumatic Stress Disorder at a given time. It has been estimated that the total cost of PTSD and other anxiety disorders is approximately 43.2 billion dollars in economic impact. Due to repressed emotions, people with PTSD can be given the wrong or no treatment which can be fatal to the patient. For example, stress can lead into depression and without accurate identification, it can lead to detrimental effects down the road. To solve this problem, I propose using AI to accurately derive emotions from people who have suppressed their emotions.

In the past, sentimental analysis algorithms were used on text that was transcribed from the given audio to obtain emotions from speech; however, there are a couple of major drawbacks. The first one being that words had to

be mapped to its corresponding emotion which is very time consuming. For example ”great” would often be mapped to the emotion of happiness and this needed to be done thousands of times for a multitude of words that describe emotions. Another major drawback is the omission of speech subtleties in audio data. For example, in the western cultures when people are confused and ask a clarifying question, they often raise the pitch of their voice at the end of their question. Not accounting for these subtleties leads to significantly lower emotion detection ability. Adding the two drawbacks above, someone could sarcastically say ”great job” and mean the opposite while the algorithm would think that the person was being positive.

To combat this loss of essential data, the emotion detection field of AI has added audio data with their transcriptions. Though effective, it is still not as efficient as the algorithm needs to map words in the transcribed text. This paper eliminates the need to use a transcript and uses audio data based on images which allows the model to train on the subtleties of speech.

Related Work

This paper will use the RAVDESS dataset to train the model as the dataset has the largest range of emotions, eight different emotions, out of all the available datasets such as SAVEE, or TESS [1]. There are several reasons why someone would use datasets other than RAVDESS. One being the fact that the alternate datasets have specific groups of people for location based research. For example, TESS is a dataset of english speaking Toronto citizens, whereas RAVDESS is a general dataset which includes english speaking people from around the globe. As the next step, I used Mel Spectrograms to pre process the audio data as it emulates human hearing and it performs better with Convolutional Neural Network (CNN); however, other research papers used another method of pre-processing called Mel Frequency Cepstral Coefficients (MFCCs) [1][2][3]. The difference between MFCCs and Mel Spectrograms is that M FCCs are a more processed version of Mel Spectrograms. Different types of AI models have also been used in the past for training. For example, some researchers have used K-Nearest Neighbors (KNNs) [4] while others used Support Vector Machines (SVMs) [5]. However, in this paper I use a more recent CNN model, RESNET-18, for image detection. Though the original use of RESNET was for computer vision and classifying images, this paper uses RESNET to identify emotions in audio data [3]. Previous papers have used emotion detection for real world applications. For example, the authors in [6] created a model that used emotion detection in HR recruitment. The authors in [7] use the same pretrained RESNET model, but they apply it for Computer vision recognition.

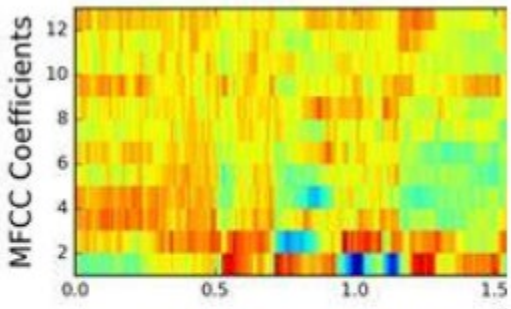
Methodology

Dataset

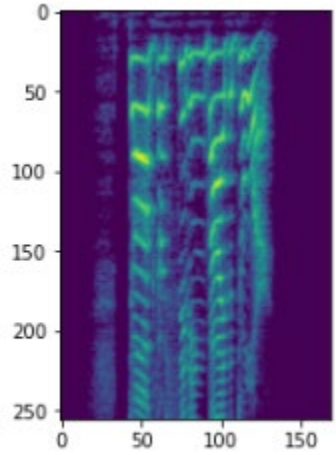
Though multiple datasets were considered, including a compiled dataset, the RAVDESS dataset was used as it has 15%-33% increase in range of emotions when compared to datasets like TESS. It contains three different data types which are audio only, video-only, and video and audio; however, only audio data was used as this paper focuses on emotions in audio data. The audio files were 16bit, 48kHz .wav files. In the audio data, there are 24 folders; one folder for each voice actor and there are 24 voice actors consisting of 12 males and 12 females. There are two statements, ”Kids are talking by the door” and ”Dogs are sitting by the door” , in 8 different emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. In addition, there are two levels of intensity: normal and intense. Each of the folders contain 60 audio files. In total, there are 1440 audio files in the dataset. The file tree below shows an example of how the data is formatted.

Preprocessing

The audio data from the RAVEDESS dataset has to be converted into an image as Resnet-18 [8] is an image based model. To change the audio data into images, this paper used the Mel Spectrogram function of the Librosa library.

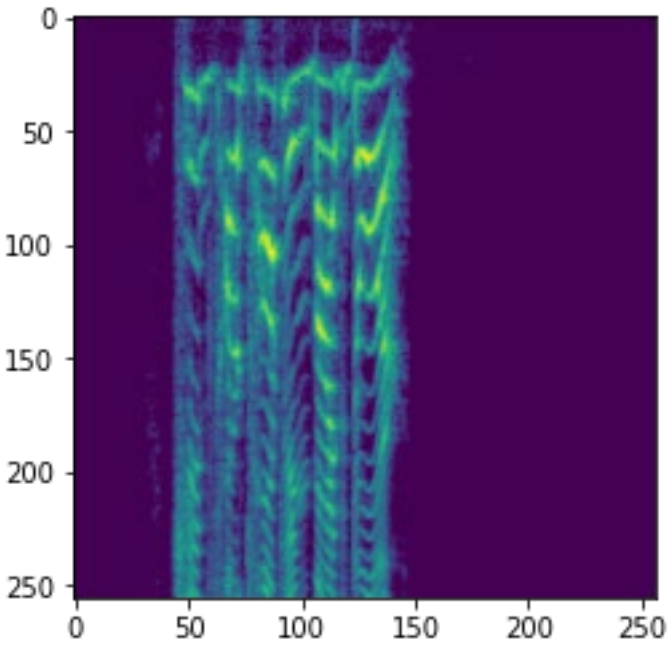


MFCC 1



Mel Spectrogram

In the figures above are a MFCC and a Mel Spectrogram. The main difference between the two converters is that the MFCC is the Discrete Cosine Transform (DCT) which takes the log of the Mel Frequency Spectrogram leading to more imposition and difficulty of training the CNN models. When creating the Mel Spectrogram, default parameters were used which included n_mels at 256, fmax at 4096, n_fft at 2048. The Mel Spectrograms produced spectrograms of different widths and sizes which was problematic while training the model, so the data was padded to an equal size of 256 by 256. To pad the spectrograms, the point at the bottom right of the image was copied and then used repeatedly to make the image a 256 by 256 rectangle.

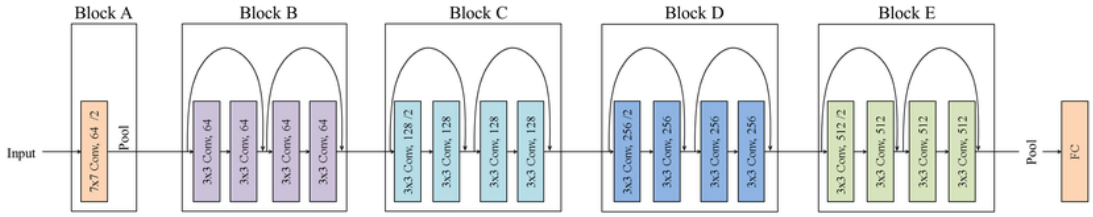


Resized Spectrogram

The reason for converting audio data into pre-processed image data was because AI models train better on images and return better accuracy than audio based models. Furthermore, there is a greater amount of research related to image based models than audio based models.

Model

In this paper, I used a prebuilt RESNET-18 data model [8]. RESNET is a computer vision model made up of X amount of layers where X is RESNET-X . For example, RESNET-18 has 18 layers. Alternate RESNET models with more layers were reviewed, but due to hardware limitations, RESNET-18 was chosen to be the optimal model for this research. Furthermore, as RESNET-18 was originally trained on images it was predicted to give higher accuracy over a pretrained model.



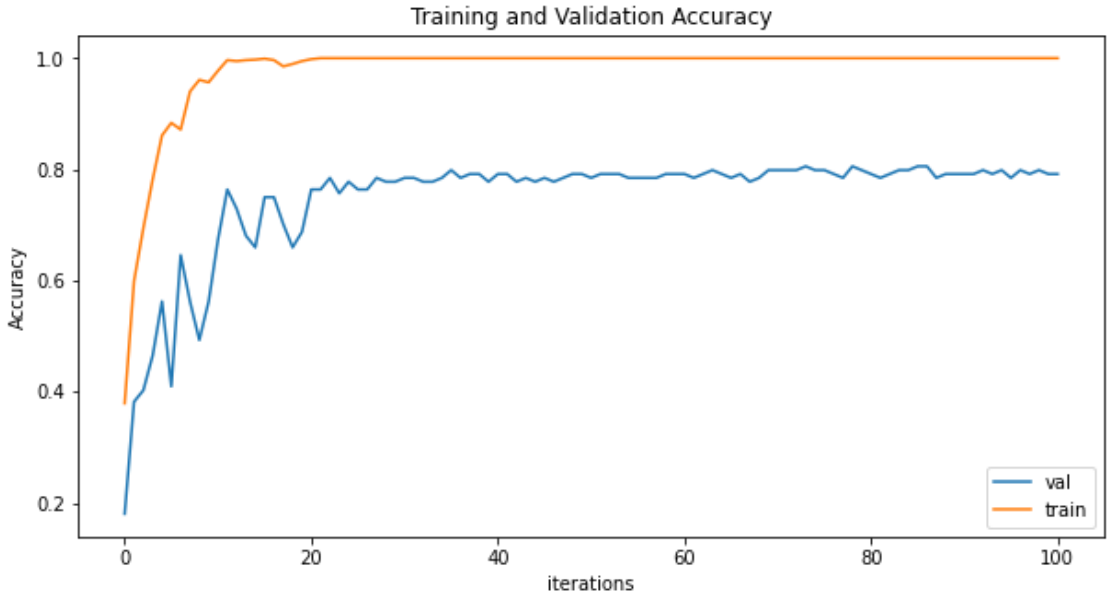
Resnet Model Visualized

Additional benefits include each layer consisting of residual blocks allowing -+RESNET to keep consistent performance compared to alternate models at the time. In addition, the model included shortcut connections allowing RESNET to keep the parameters from previous layers resulting in much better accuracy and performance. Resnet was groundbreaking as it didn't lose the accuracy with an increasing number of layers.

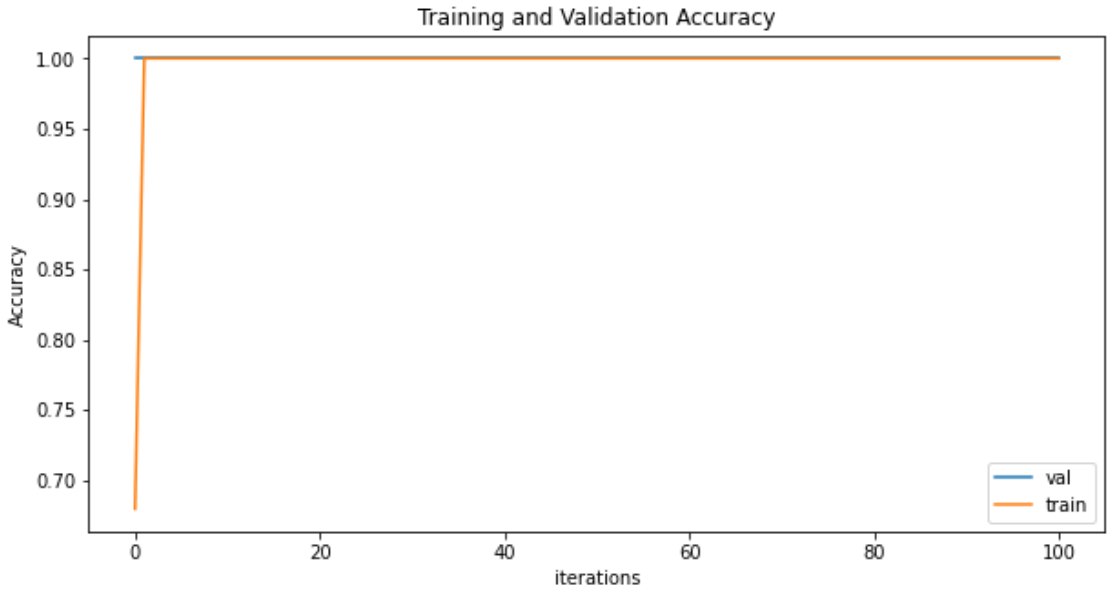
Results and Discussion

Hyperparameters & Results

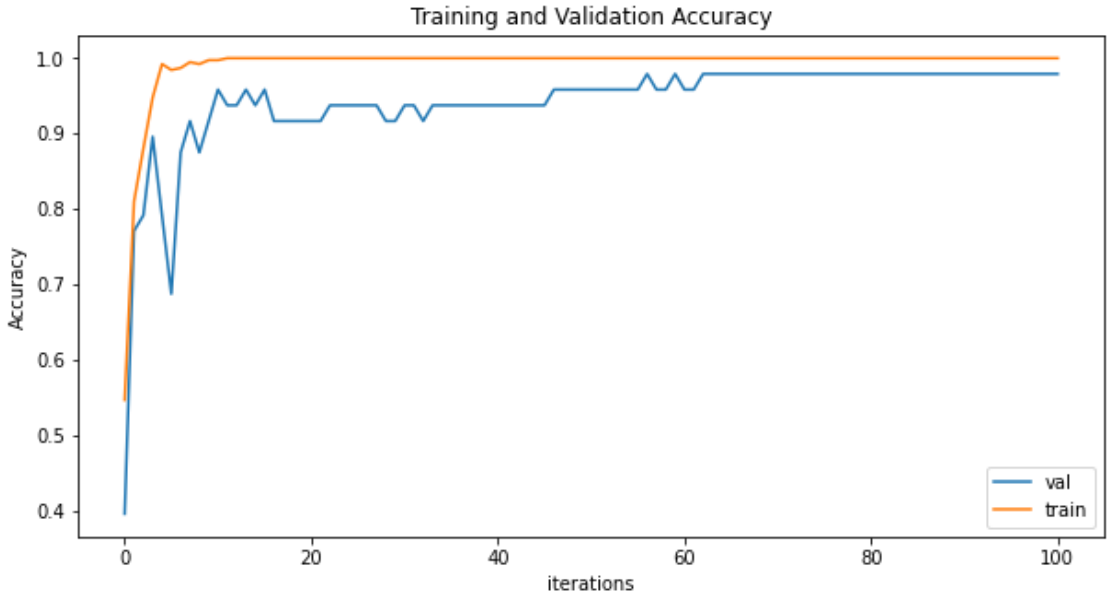
Hyperparameters are a set of parameters such as momentum, initial learning rate, etc…that are used to accurately train the model on the data. Getting the optimal value of the hyperparameters for training the model, based on trial and verify, was the most difficult aspect of the research. In this research I used two hyperparameters: learning rate and momentum, the values of which were set to 0.03 and 0.81 respectively. However with additional time more hyperparameters can be used in the research to further improve the accuracy.



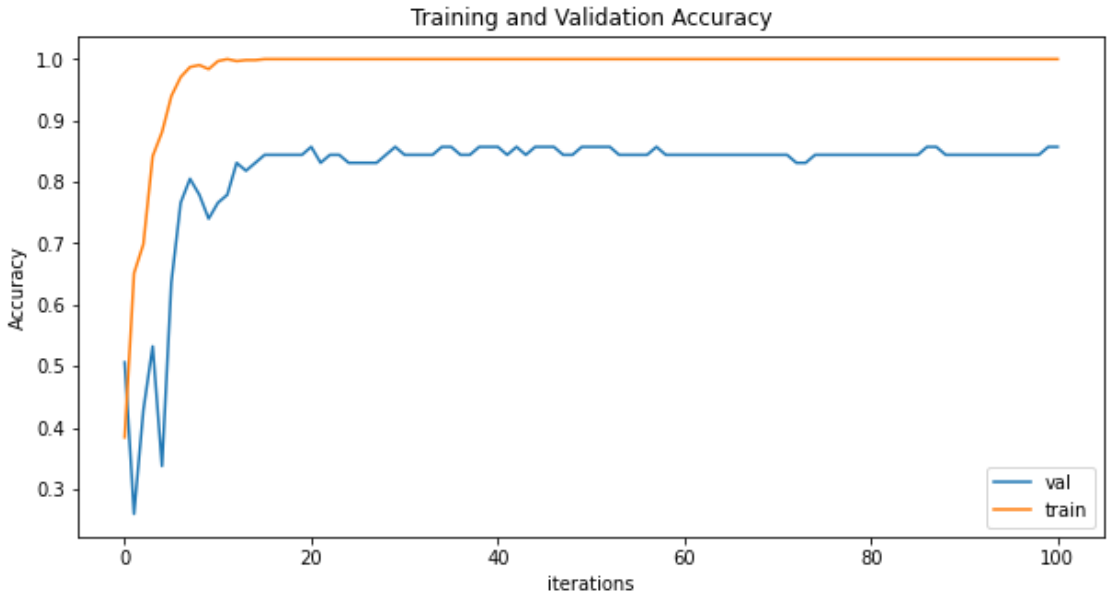
Best Val and Test accuracy



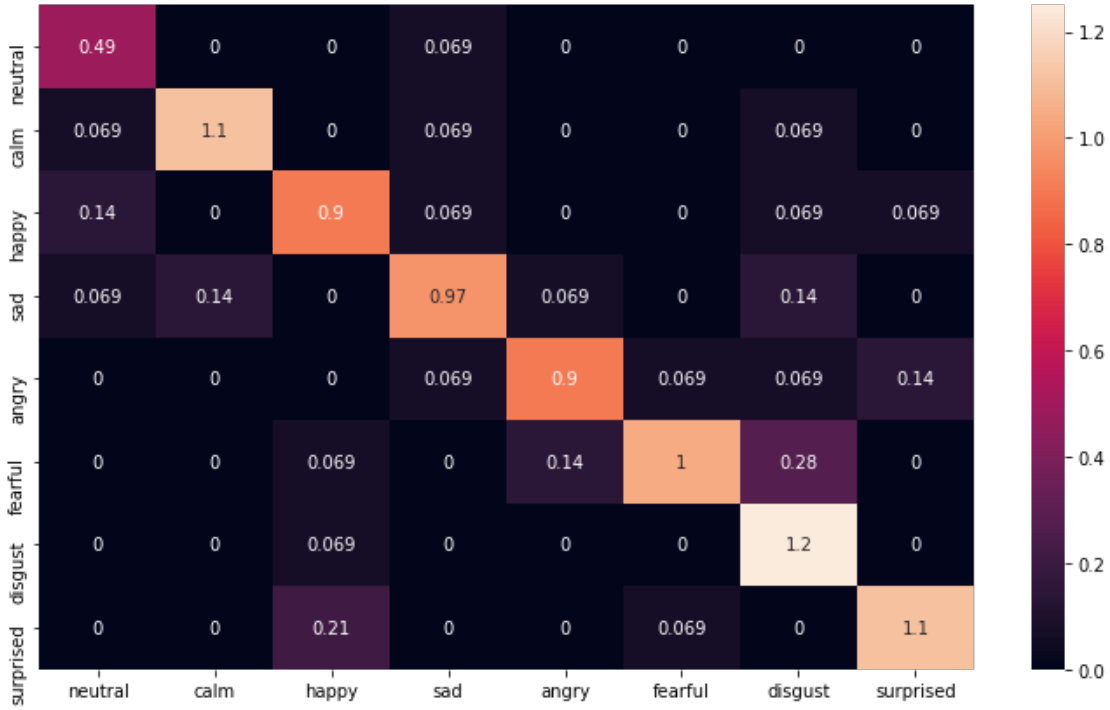
Best no stress accuracy



Best neutral accuracy



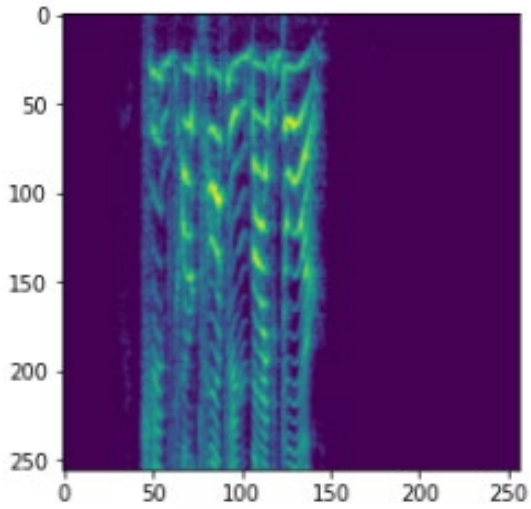
Best stress accuracy



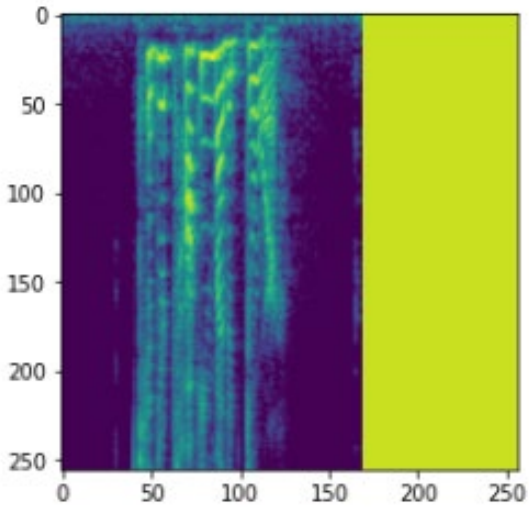
Confusion Matrix

Discussion

Overall results achieved, as shown in figure 3, were training accuracy of 100% while the validation accuracy was of 82% for the entire data set. The accuracy is further improved if the dataset is split into three emotional groups: no stress, neutral, and stress. I classified the eight emotions into the three different groups as follows: [happiness] in the no stress group, [neutral, calm, disgust, and surprised] in the neutral stress group, and [sad, angry, and fearful] in the stress group. I found that the no stress category has the highest accuracy at 100% as shown in Figure 4, compared to accuracy of the Neutral group with 97% as shown in Figure 5 and accuracy of the Stress group with 85% as shown in Figure 6. After analysis, the no stress group achieved the highest accuracy because it had only one emotion as compared to the other groups which had three or more emotions each. The reason for the stress group to do poorly is due to its inability to distinguish between neutral emotions and stress emotions. To help visualize the accuracy range, I created a confusion matrix as shown in Figure 7.The confusion matrix visualizes how the model performs very well on extreme emotions like happiness and disgust, while it degrades on calmer emotions like neutrality. In addition, the higher accuracy on the no stress group could be due to overfitting on the limited amount of no stress data. The validation accuracy achieved is on average 4% higher than the published research.



Mel 1



Mel with zeros

Conclusion

Human emotions under certain circumstances are hard to identify, especially stress related emotions. This is most evident with Post Traumatic Stress Disorder patients as they can have a challenging time expressing their emotions which in turn may lead to a misdiagnosis by medical personnel. This paper describes an image base model, RESNET-18, for emotion prediction in audio with high accuracy at 82%. With continued training, an accuracy above 90% is definitely achievable and will improve the lives of people who are not able to express their emotions. When dividing the emotions into emotion classes, the prediction accuracy was 99% for no stress class (happiness), 97% for neutral class, (neutral, calm, and surprised), and 85% for stressed class (fearful, sadness, anger, disgust). As part of the future work, the accuracy can be further improved by training on more datasets and using alternate methods of padding the image data.

Future Work

My aim for overall accuracy was at least 90%, the results achieved were 8% lower than the goal. For future work to improve the accuracy, the following three approaches can be implemented. Firstly,building a model from scratch or using a model built for audio classification specifically as opposed to RESNET as it was not built to classify emotions in audio. In addition, using different approaches to pad the images such as using only zeros for all of the Mel Spectrograms. Finally, to increase the accuracy, sourcing multiple datasets such as TESS as well as adding randomized data to emulate the noise in the real world from cars, other people, the oscillation of the A/C current coming from the power grid, etc…

References

[1] M. P. B. Andrew Huang, “Human vocal sentiment analysis,” arXiv preprint arXiv:1905.08632v1, 2019.
[2] M. B. Lindasalwa Muda and I. Elamvazuthi, “Voice recogni-

tion algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," arXiv preprint arXiv:1003.4083, 2010.
[3] S. T. Z. S. Ameya Ajit Mande, Sukrut Dani, "Emotion detection using audio data samples," ijarcs ijarcs: v10i6.6489, 2019.
[4] I. T. Meftah, "Emotion recognition using knn classification for user modeling and sharing of affect states."
[5] K. He, "Deep residual learning for image recognition."
[6] K. Tomba, J. Dumoulin, E. Mugellini, O. Abou Khaled, and S. Hawila, "Stress detection through speech analysis," Proceedings of the 15th International Joint Conference on e-Business and Telecommunications, 2018.
[7] J. Liang, "Image classification based on resnet," Journal of Physics: Conference Series, vol. 1634, p. 012110, 2020.
[8] A. Hassan and R. Damper, "Multi-class and hierarchical svms for emotion recognition," Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, pp. 2354–2357, 01 2010.