DYSARTHRIA DETECTION IN AUDIO FILES

BY

DIEGO DEL CARPIO

Senior Thesis in Computer Engineering

University of Illinois Urbana-Champaign
Advisor: Thomas Moon

Spring 2024

ABSTRACT

Motor speech disorders such as Dysarthria affect the way people can speak and communicate with others. Damage to the nervous systems causes one to lose control over their tongue and voice box which brings upon difficulty in the pronunciation of words. Voices have characteristics that can help distinguish between normal and abnormal voices. This research aims to take features that have previously worked and incorporate more features to further improve the robustness of identifying whether an audio file contains a person speaking with dysarthria.

Key Words: Dysarthria, Fundamental Frequncy, GNR, Formants, Feature Extraction, Machine Learning, Multi-Layer Percerpton, Confusion Matrix

To my parents, for their love and support.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Over the last couple years, machine learning has been more frequently use to aid in the detection of disorders. Patients with dysarthria benefit from these advancements. This could help potentially detect dysarthria earlier and prevent the situation to advance or become worse.

Dysarthria is a voice disorder that affects the motor functions of the lips, larynx, and jaw. Pronunciation of words make it difficult for patients who are diagnosed with the voice disorder. Research exploring the application of machine learning in the medical field has been expanding. The goal of this research is to extract relevant features in order to make a robust model that can identify dysarthria in audio files.

The research methodology will involve feature selection, feature extraction from a dataset, model training, and validation using metrics such as accuracy, precision and confusion matrices.

The structure of this paper is as follows: Section II provides a literature review on the extracted features and machine learning models that have excelled in previous research. Section III outlines the method used. Section IV presents the results, followed by conclusions.

CHAPTER 2

LITERATURE REVIEW

It was important to get familiar with acoustic analysis and the features that have been used in machine learning models to get accurate classification for dysarthria. An overview of findings related to voice disorder detection and will be discussed in this section of the paper.

2.1 Extracted Features

Acoustic analysis is the study of sound waves, more specifically analyzing properties in audio signals in our situation. With acoustic analysis, one is able to extract features such as fundamental frequencies, harmonic to noise ratio, and mel-frequency cepstral coefficients (MFCC).

The findings by Hedge et al. [1] and Hossan et al. [2] showed that MFCC was a great feature to extract. MFCC is a representation of the short-term power spectrum of a sound. It is based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. MFCC passed through a gaussian mixture model was able to reach accuracies above 79% for vocal disorders and pattern recognition in previous studies.

Features such as fundamental frequency (f0), shimmer, jitter, harmonic to noise ratio (HNR), and glottal to noise excitement ratio (GNE) were features to also be used after investigating the research done by Jalali-najafabadi et al. [3], Lopes et al. [4], and Teixeira et al. [5]. Autocorrelation is the correlation between a signal and a delayed copy of itself. This digital signal processing technique is used to find periodic signals obscured by noise which returns the fundamental frequency. An example of the fundamental frequncies throughout a sample from the dataset is seen in Figure 2.1.

Perturbations in voices can be analyzed through the features shimmer and jitter. Shimmer refers to the variations in the cycle-to-cycle amplitude

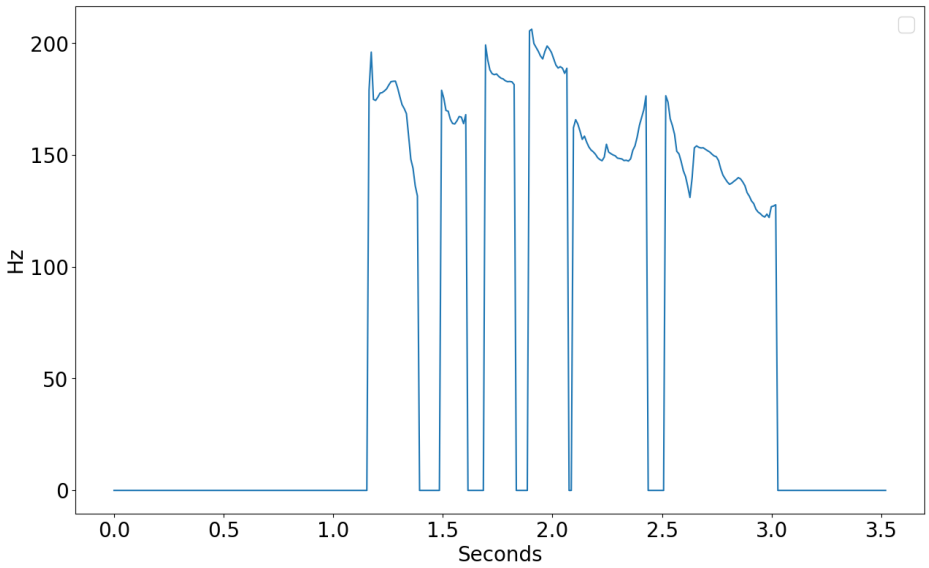

Figure 2.1: Fundamental frequencies of a sample

of vocal fold vibrations while jitter refers to the variation in fundamental frequency. These can be seen in Figure 2.2 Jitter can be represented with the percentage of variation and difference in milliseconds. The same applies to shimmer with it being represented as a percentage or change in decibels. These perturbations can be more apparent with patients that are diagnosed with dysarthria which make it an excellent feature to look at. Harmonic to noise ratio is the ratio that quantifies the amount of additive noise in the voice signal. This feature can confirm the presence of a voice disorder when the signal is unclear due to the impact on vocal clarity. Glottal to noise excitement is the ratio between amount of voice excitation by vocal fold oscillations versus excitation by turbulent noise. Some patients may not control their vocal cords well resulting in more breathy pronunciation. Having quality of vocal fold vibration at the glottal level is beneficial when determining a voice disorder.

Formants are resonant frequencies that are from the vocal tract that shape the quality of speech sounds. Vowel and consonant pronunciation can be affected when a disorder is present, resulting in apparent abnormal frequencies. The first three formants were examined, where f1 conveys information about vowel height, f2 about vowel backness, and f3 about lip rounding. In the figure 2.3 below, the three formants are displayed on top of a spectrogram. The formants exhibit smoother characteristics during speech production and

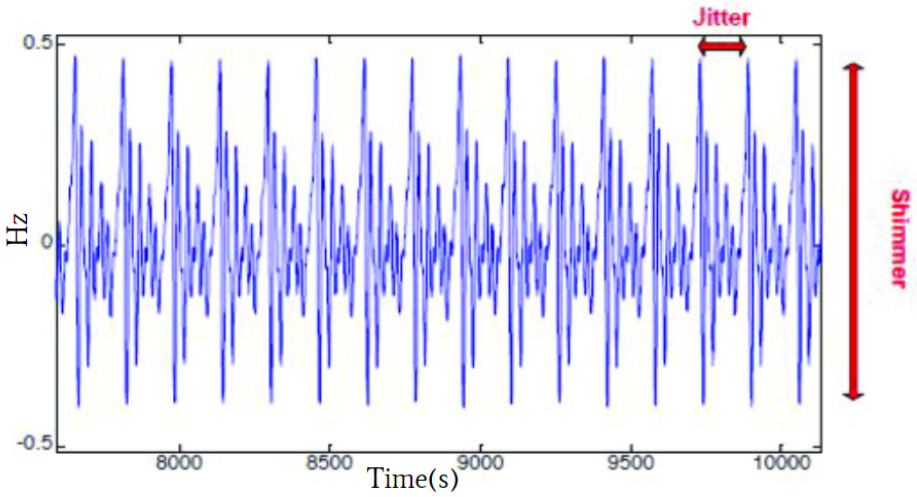

Figure 2.2: Jitter and shimmer diagram

become noisier in the absence of speech.

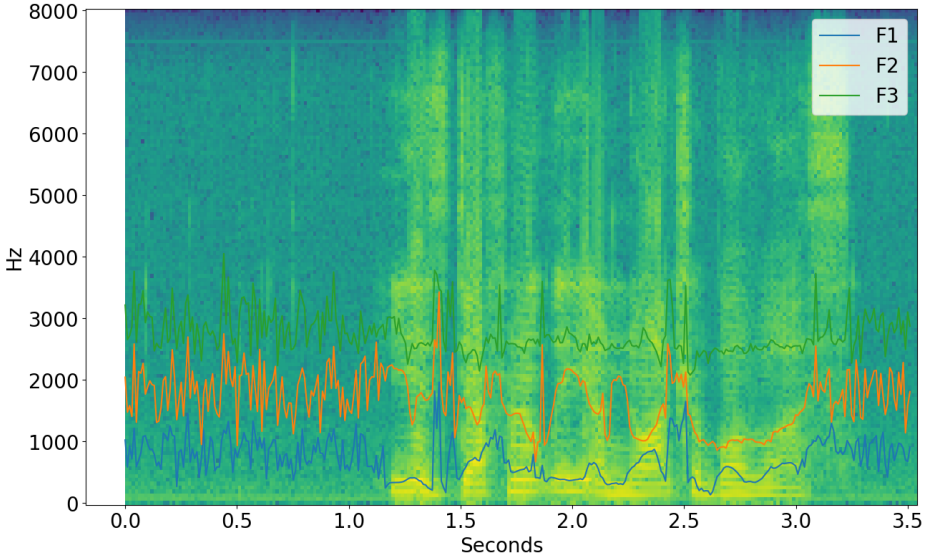

Figure 2.3: Formant plot on spectogram

2.2 Machine Learning Models

Machine learning models are programs that learn a pattern and can compute a prediction based on the pattern recorded. It is not at all previously programmed and learned from the data that it is given. Support vector machines (SVM) and artificial neural network (ANN) or multi-layer perceptron

(MLP) were looked at [1].

A Support vector machine is a supervised learning algorithm used for classification and regression tasks. They are commonly used in binary classification where a hyperplane separates the features of different classes. SVMs excel in handling high-dimensional data and can be adjusted to achieve either a strict or soft margin of separation. Figure 2.4 displays binary classification where the data points are being separated into two classes by the line. In previous reaserch, features passed through an SVM resulting in accuracies above 80%.

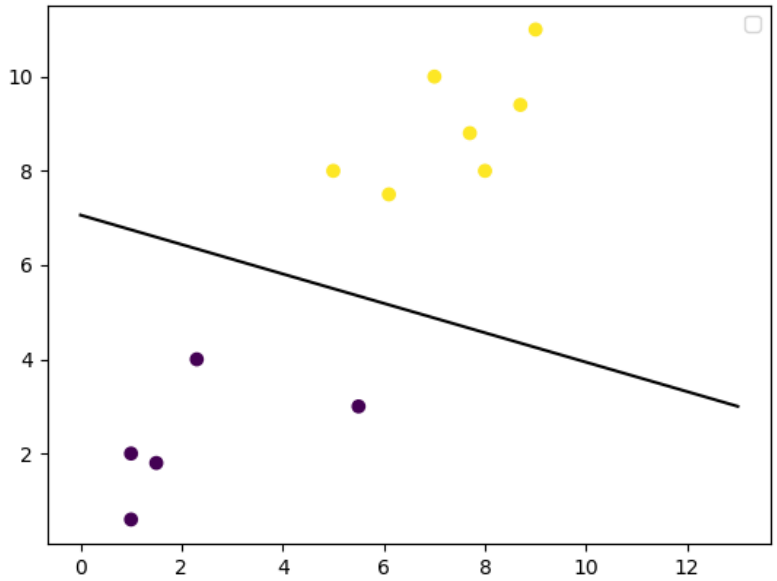

Figure 2.4: Support Vector Machine

A multi-layer perceptron algorithm is modeled on the concepts of neural networks in humans. It consists of multiple layers of nodes (neurons), organized in an input layer, one or more hidden layers, and an output layer. Each node in one layer is connected to every node in the subsequent layer, forming a network of interconnected nodes as seen in Figure 2.5. During training, the network adjusts the weights of connections between nodes to minimize the difference between the predicted output and the actual output. Like SVM, previous finding showed that MLPs could get accuracies over 80%

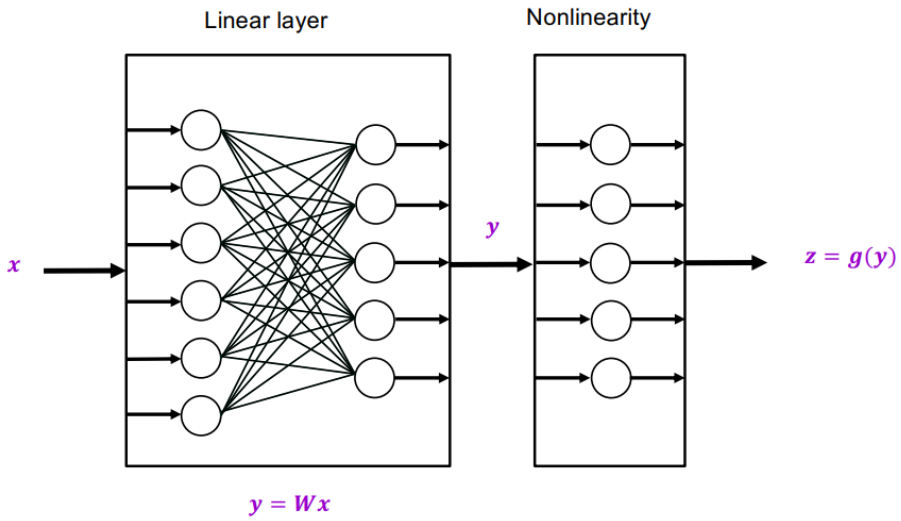

Figure 2.5: Multi-Layer Perceptron

CHAPTER 3

PROPOSED METHOD USED

The proposed method to detect dysarthria from a .wav file follows the model for an automatic voice disorder detection (AVDD) [1]. The dataset from the TORGO database included samples from both male and female patients. It consists of individuals with dysarthria and those without dysarthria. Data cleaning involved samples without a noticeable signal to extract features from. The samples are pre-labeled so the focus was on feature extraction and evaluating the machine learning model.

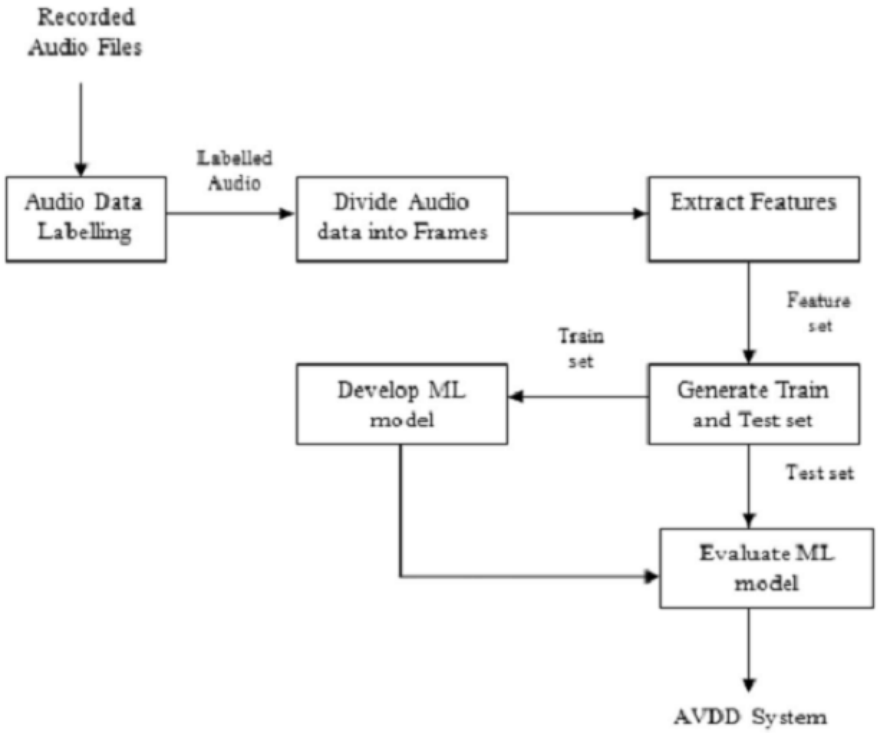

Figure 3.1: Automatic Voice Disorder Detection (AVDD)

Mel frequency cepstral coefficient was the first feature to be extracted. This was done by using digital signal processing techniques [2]. Pre emphasis and padding was the only form of pre preocessing done on the signal from

the samples. Padding was applied to insure that all the signals had the same length for MFCC extraction. The signal was cut into frames of size of 25ms with a step size of 10ms. A hanning window was then applied to the windowed signal.

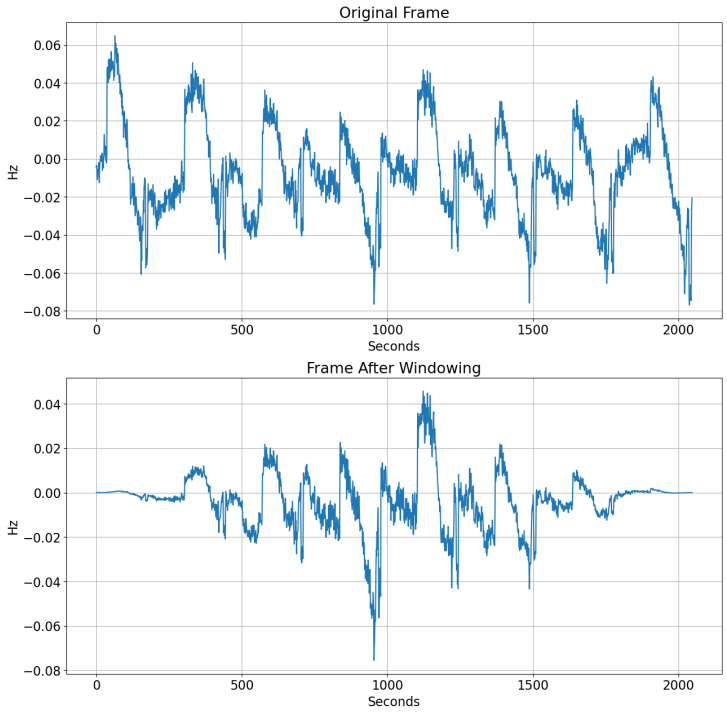

Figure 3.2: Windowing Signal

Fast Fourier transform (FFT) was applied to obtain the signal in the frequency domain. Mel filter banks, shown in Figure 3.4, were calculated by obtaining evenly spaced center points and creating the filters from them. The mel-frequency cepstral coefficients are obtained by taking the dot product of the mel filter banks with the power of the FFT windowed signal. This results in a matrix of frequency values which can be passed through a machine learning model.

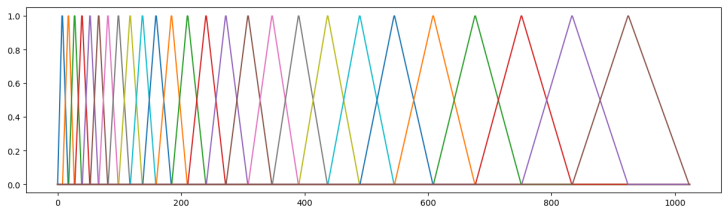

Figure 3.3: Automatic Voice Disorder Detection (AVDD)

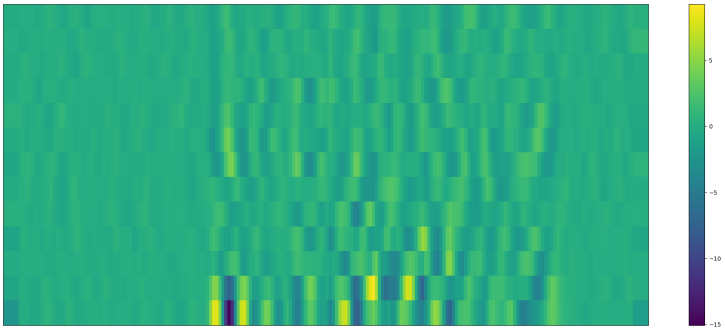

Figure 3.4: Automatic Voice Disorder Detection (AVDD)

Pre-existing libraries simplified much of the feature extraction process. Parselmouth was the python library utilized to accomplish the extraction of the fundmental frequency, shimmer, jitter, harmonic to noise ratio, and formants. Praat is a freeware program that is used in the analysis and reconstruction of acoustic speech signals. Parselmouth allows Praat to be ported so that it can be used in Python. Sections without noise were used for the average fundamental frequency and average formants as that data was relevant. The correlation between the formants was then obtained after by also looking at sections without noise.

Glottal to noise excitement ratio was not able to be recovered like the other features so digital signal processing techniques were used just like in MFCC extraction [6]. Linear prediction analysis was part of the process in obtaining the GNE. In order to obtain twelve linear prediction coefficients (a_k), Equation (3.1) is is used. Autocorrelation is used to obtain a vector of values. The coefficients are then calculated using Equation (3.2) where r is the vector calculated in Equation (3.1) and R is the toeplitz matrix (Figure 3.5). P in our case is twelve resulting in a 12x12 matrix. The predicted signal uses the twelve coefficients calculated to obtain the predicted signal in equation(3.3). Equation (3.4) shows the original signal subtracted by the predicted signal, represented by $\hat{s}(n)$, returning the error signal. This gives us narrow pulses where hilbert envelopes can be calculated from. Three envelopes are analyzed centered at 500Hz, 1500Hz, and 2500Hz. The GNE is finally determined by selecting the maximum correlation value from each pair and then choosing the maximum among these maximum values.

$$R(i) = \sum_{n=1}^{N-1} s(n)s(n-i) \tag{3.1}$$

$$R = \begin{bmatrix} R(1), R(2), R(3), ..., R(P) \\ R(2), R(1), R(2), ..., R(P-1) \\ R(3), R(2), R(1), ..., R(P-2) \\ . \\ . \\ R(P), R(P-1), R(P-2), ..., R(1) \end{bmatrix}$$

Figure 3.5: Toeplitz matrix

$$A = -R^{-1} \cdot r \tag{3.2}$$

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k \cdot s(n-k) \tag{3.3}$$

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k \cdot s(n-k) \tag{3.4}$$

Once all the features we desire are extracted, those features are then passed through a machine learning model for classification. The main focus for this part was to evaluate and refine the machine learning model. Data was split 70/30 for training and testing purposes respectively. The hyperparameters analyzed for fine-tuning included batch size, epoch count, learning rate, and the dysarthria-to-non-dysarthria data ratio. Once the parameters that showed a correlation with increased accuracy were identified, they were then utilized to achieve the desired level of accuracy.

CHAPTER 4

DESCRIPTION OF RESEARCH RESULTS

Figure 4.1 displays the confusion matrix of the first attempt for the dysarthria detector. The first model involved the use of MFCC and was passed through a support vector machine. An accuracy of 63% was achieved but that was not enough for robust model. The model was able to identify samples without dysarthria better than samples with dysarthria. This was mainly due to the imbalance of data since there were more samples that did not have dysarthria. Another issue that rose was that it was difficult to differentiate MFCC for samples with dysarthria and without. Other features were researched becuase of these results.

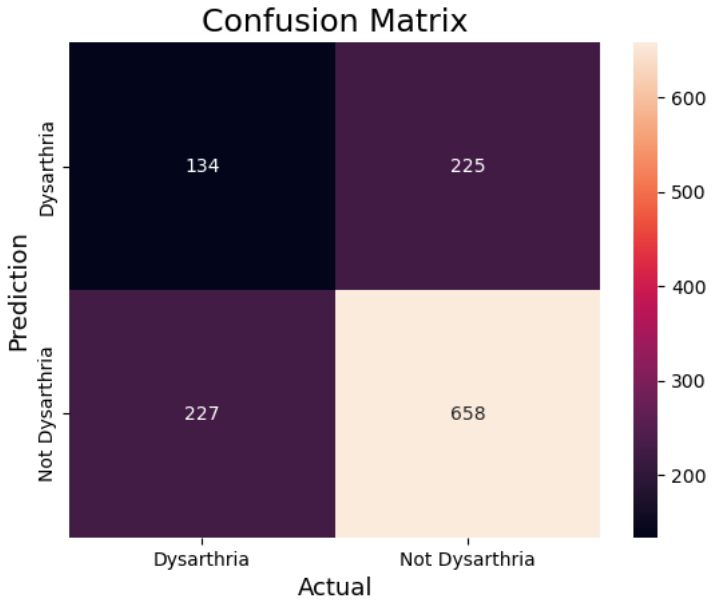

Figure 4.1: MFCC SVM Confusion Matrix

The second version of the detector consisted of 7 features extracted and again passed through a support vector machine. This achieved slightly better results but the accuracy was not able to go beyond 70%. A plot of fundamen-

tal frequency vs jitter(%) was created to illustrate the data (Figure 4.2),with '1' indicating audio samples containing dysarthria and '0' representing the absence of dysarthria. The problem of finding a distinct difference between the two classes was still not apparent. Many data points overlapped causing it difficult to find a plane using SVM.

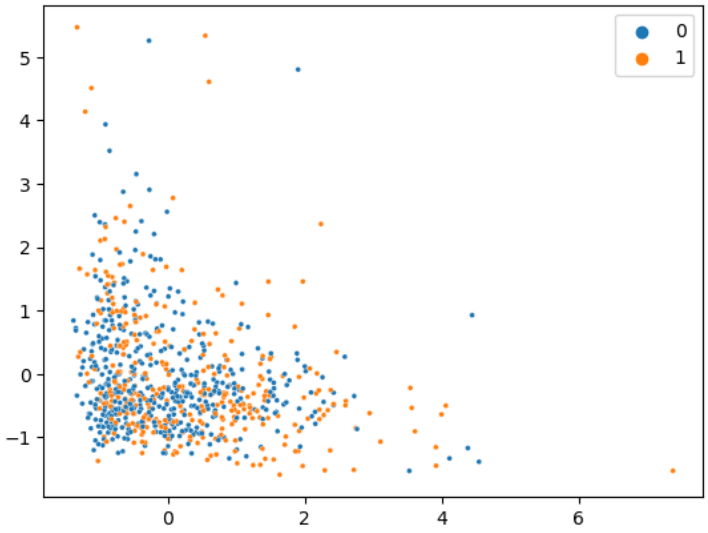

Figure 4.2: Fundamental Frequency vs Jitter(%) Test Data

It was then that the model switched from an SVM to a multi layer perceptron. To balance the data, data for non dysarthria was halved so that the dysarthria-to-non-dysarthria data ratio was closer to 1:1. This adjustment ensured that the accuracy was correctly classifying dysarthria and achieving an inflated accuracy due to accurately identifying many non-dysarthria samples. These modifications contributed to the accomplishment of reaching a 78.03% accuracy for classification (Figure 4.3).

It was important to refine the model by looking at the batch size, epoch count, learning rate, and the dysarthria-to-non-dysarthria data ratio. It was apparent from figure 4.4 that a higher learning rate resulted in better accuracy. Batch size and epoch count had some influence but were not defining parameters. The dysarthria-to-non-dysarthria data ratio being close to 1:1 had a lower accuracy than the imbalanced data. It was decided to keep the ratio closer to 1:1 as classification accuracy would be more meaningful. 78.53% was the highest achieved accuracy with the best parameters used (Figure 4.5). It was a minuscule improvement but an accuracy passing the 80% threshold was desired.

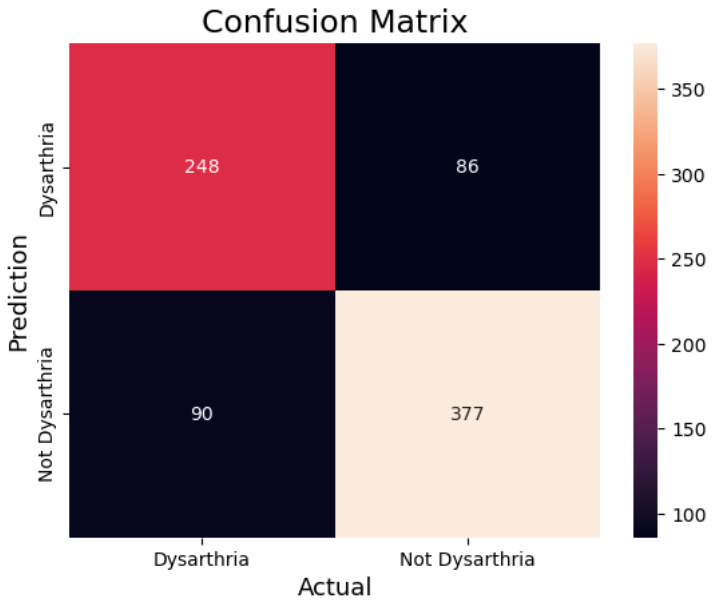

Figure 4.3: Neural Network Confusion Matrix

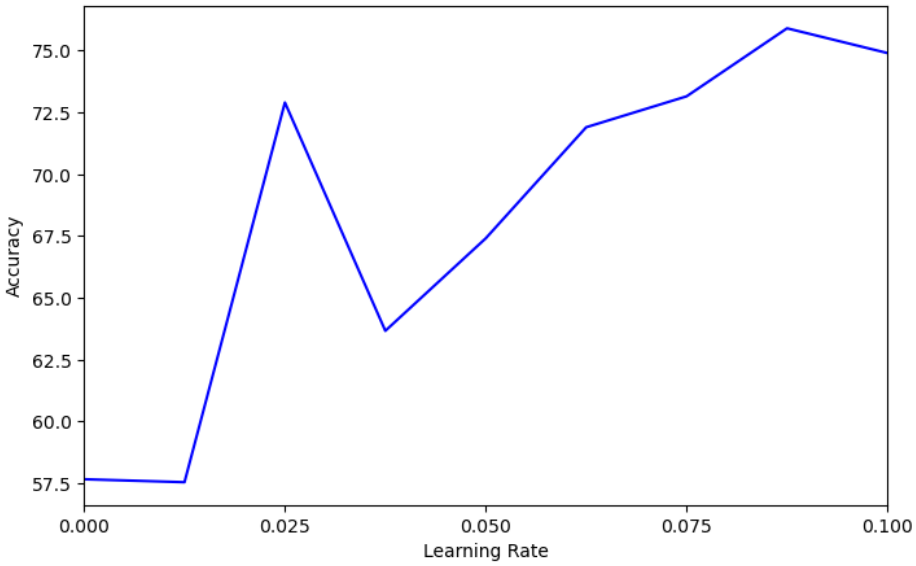

Figure 4.4: Learning rate vs accuracy

Epochs	Batch Size	Accuracy (AVG)	Run 1	Run 2	Run 3	Run 4	Run 5
1000	16	0.7416	0.7353	0.7341	0.7403	0.7541	0.7441
2000	16	0.7620	0.7478	0.7391	0.7853	0.7665	0.7715
3000	16	0.7376	0.7553	0.7341	0.7166	0.7266	0.7553
1000	32	0.7316	0.7241	0.7316	0.7503	0.7116	0.7403
2000	32	0.7228	0.7228	0.7029	0.7416	0.7203	0.7266
3000	32	0.7336	0.7066	0.7278	0.7441	0.7378	0.7516
1000	64	0.7381	0.7516	0.7441	0.6891	0.7578	0.7478
2000	64	0.7278	0.7291	0.7066	0.7091	0.7291	0.7653
3000	64	0.7386	0.7541	0.7690	0.7179	0.7403	0.7116
1000	128	0.7323	0.7179	0.7341	0.7141	0.7403	0.7553
2000	128	0.7216	0.7303	0.7166	0.6841	0.7378	0.7391
3000	128	0.7403	0.7266	0.7191	0.7815	0.7428	0.7316

Figure 4.5: Table of parameter tuning

Formants were introduced in the hopes to aid in the improvement of classification due to apparent differences seen between the correlation between formant 1 and 3. In total, there were now 13 features extracted from the samples with the dysarthria-to-non-dysarthria data ratio being close to 1:1. Figure 4.6 visualized the model being able to achieve detection of dysarthia at an accuracy of 81.14%. The model is able to distinguish between dyarthria and non-dysarthria pretty well, surpassing its performance at the beginning of this analysis (Figure 4.7).

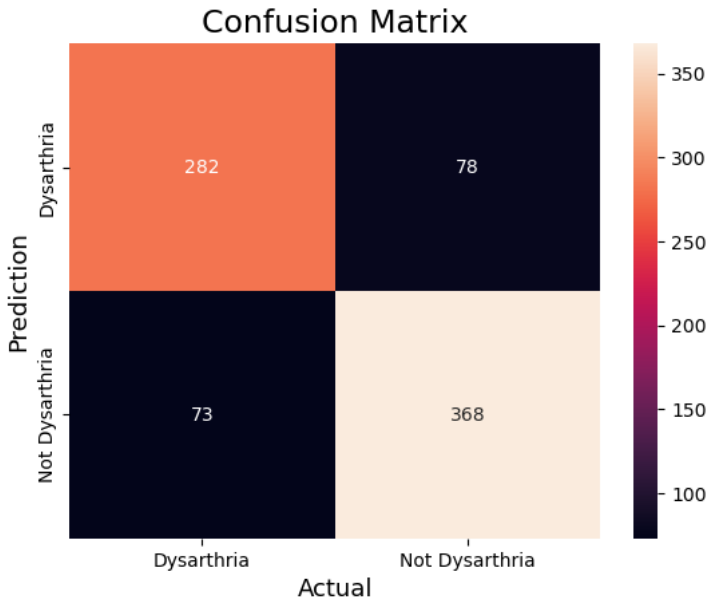

Figure 4.6: Final Model Confusion Matrix

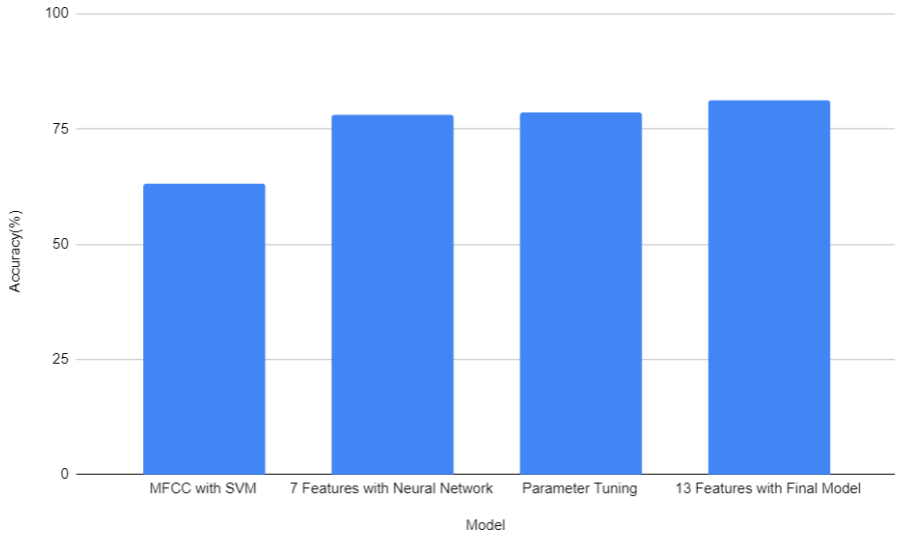

Figure 4.7: Model vs Accuracy Bar Graph

CHAPTER 5

CONCLUSION

This study has further explored the features and models crucial for the detection of dysarthria in audio files. The findings found contribute to the ongoing research done in voice disorder detection in the medical field. Notably, formants emerged as the defining feature in this study to help achieve the goal of a classification accuracy above 80%. This encapsulates the robustness of the model due to the limitations encountered with the dataset.

The research utilized a dataset involving male and female patients with and without dysarthria. As mentioned before much cleaning was necessary for feature extraction with many samples not having significant information or containing too much noise to the point where the signal was not detectable. It was important to work with what was given in order to make the model as robust as possible.

The significance of feature extraction was highlighted in this study. While Mel-frequency cepstral coefficients (MFCC) were examined, they did not provide substantial valuable data. Exploration of additional features was done to effectively distinguish between samples with and without dysarthria. The effectiveness of the machine learning model is based on the data it is given.

This research hopes to aid in the finding of what goes into making detection for disorders more viable for patients facing this problem. As technology continues to advance, integrating real-time analysis could revolutionize the diagnosis and management of dysarthria and other voice disorders, improving the lives of individuals affected by these conditions

REFERENCES

[1] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, "A survey on machine learning approaches for automatic detection of voice disorders," *Journal of Voice*, vol. 33, no. 6, pp. 947.e11–947.e33, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199718301437

[2] M. A. Hossan, S. Memon, and M. A. Gregory, "A novel approach for mfcc feature extraction," in *2010 4th International Conference on Signal Processing and Communication Systems*, 2010, pp. 1–5.

[3] F. Jalali-najafabadi, C. Gadepalli, D. Jarchi, and B. M. Cheetham, "Acoustic analysis and digital signal processing for the assessment of voice quality," *Biomedical Signal Processing and Control*, vol. 70, p. 103018, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421006157

[4] L. Lopes, V. Vieira, and M. Behlau, "Performance of different acoustic measures to discriminate individuals with and without voice disorders," *Journal of Voice*, vol. 36, 08 2020.

[5] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis – jitter, shimmer and hnr parameters," *Procedia Technology*, vol. 9, pp. 1112–1122, 2013, cENTERIS 2013 - Conference on ENTERprise Information Systems / ProjMAN 2013 - International Conference on Project MANagement/ HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212017313002788

[6] D. Michaelis, T. Gramß, and H. W. Strube, "Glottal-to-noise excitation ratio - a new measure for describing pathological voices," *Acustica*, vol. 83, pp. 700–706, 1997. [Online]. Available: https://api.semanticscholar.org/CorpusID:10631351