The University of Texas at Arlington

Project A: Speech Recognition

Joshua Catalan

CSE 3313-001

Dr. Jon Mitchell

19 November 2023

Problem

This project aims to address the challenge of speech recognition by developing a system capable of automatically categorizing audio samples containing spoken digits from 0 to 9. The dataset provided includes 40 samples for each digit. The approach for this task is to utilize the Discrete Fourier Transform (DFT) to extract features that are common to a particular digit. For instance, “one” uses lower frequencies and “six” uses higher frequencies. The task involves analyzing the DFTs of all speech samples for a given digit and identifying unique characteristics that distinguish it from other digits. The objective is to create an effective speech recognition system based on these distinct features.

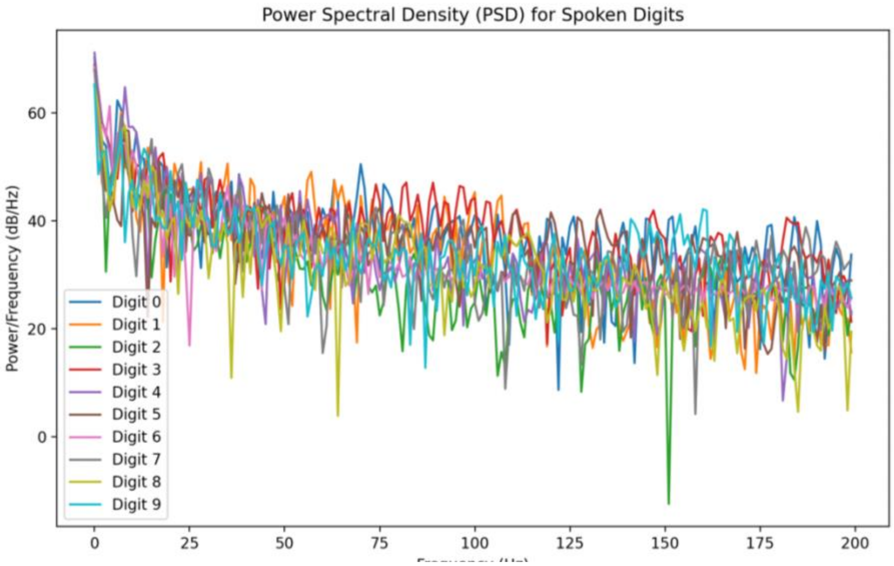
Approach

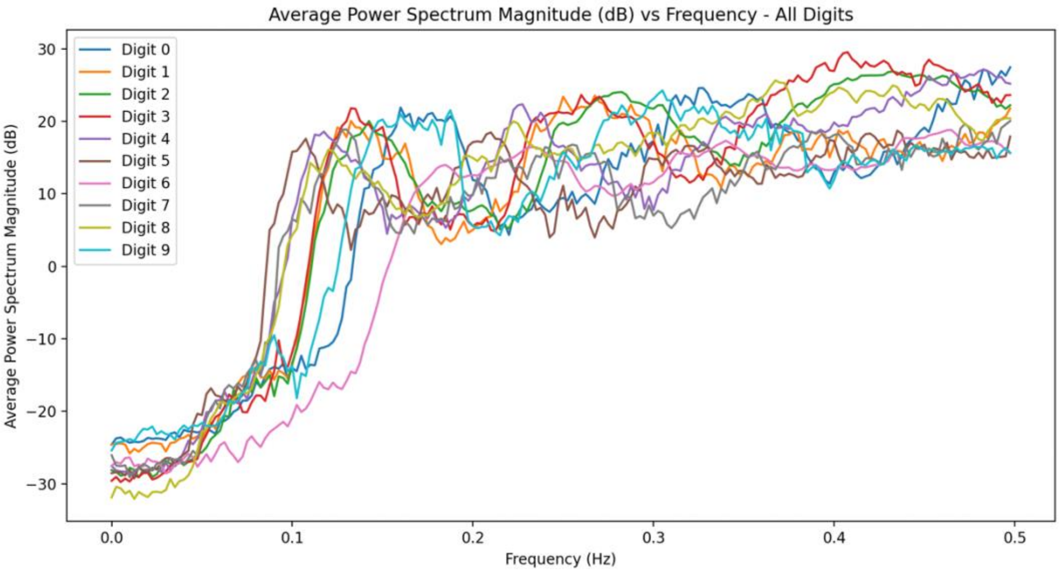
The approach I used to solve this problem involves a combination of feature extraction, spectral analysis, and machine learning. The key function, ‘extract_features’, loads the audio data using the librosa library and performs a Discrete Fourier Transform (DFT) to retrieve the magnitude of the frequency components. The magnitude is then adjusted to a targeted length to make the input data consistent for the machine learning model. The ‘calculate_ratio’ function computes the ratio of low to high DFT coefficients for the given feature which in this case is the magnitude.

I used a Support Vector Machine (SVM) with a linear kernel to train the classifier on the given dataset. I used SVM for its ability to effectively classify data into different categories and a linear kernel for a linear decision boundary for distinguishing between the different digits in the dataset. The Power Spectral Density (PSD) is visualized for each digit in the training set, and it provides frequency characteristics for the different digits. Additionally, I used a histogram to depict the occurrences vs the calculated ratio of low to high DFT coefficients for each digit.

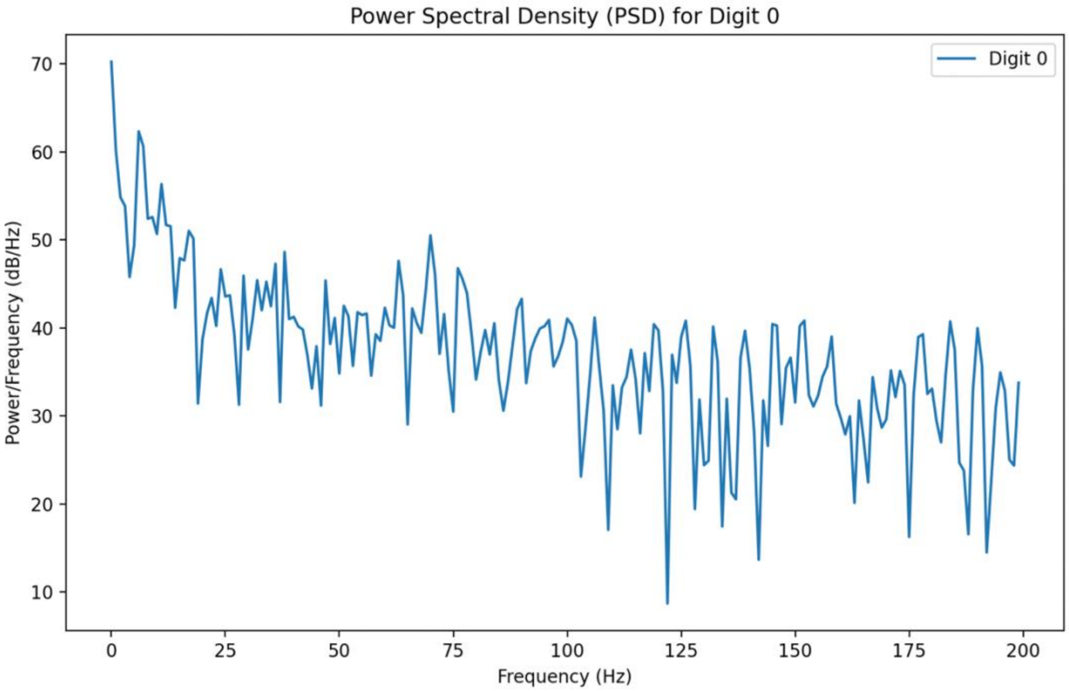
Once the classifier is trained with the provided dataset, the validation set is evaluated on it and the accuracy of the model is displayed. The digit is then predicted for the testing files using the trained classifier. Finally, PSD plots are generated for each test file, plotting the actual and predicted digits.

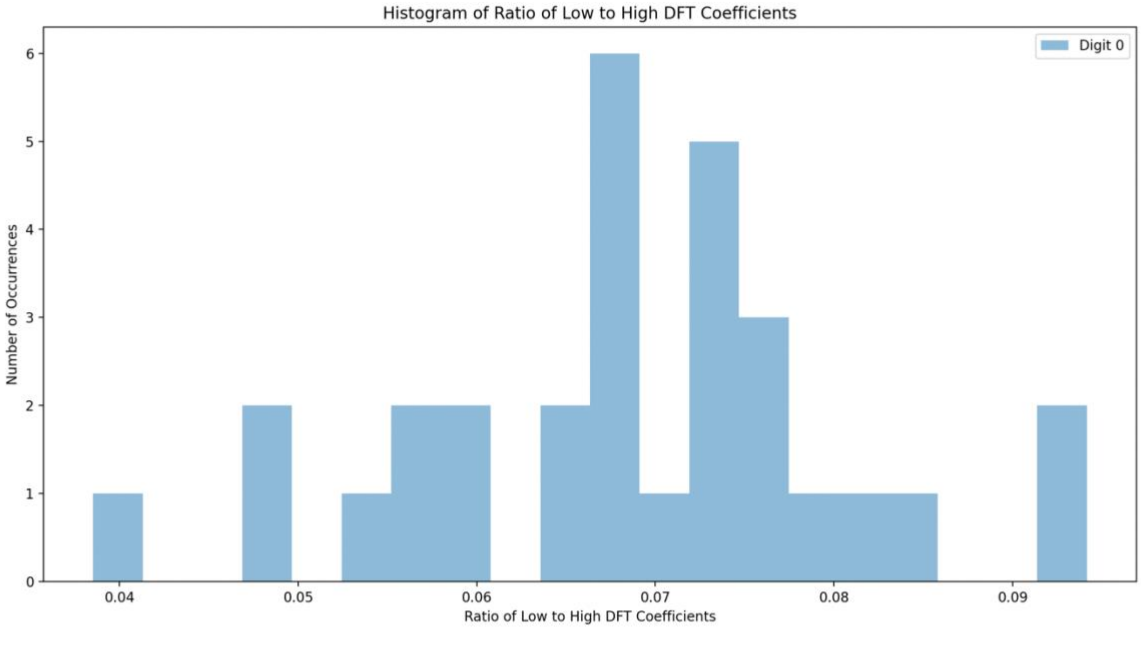
Results

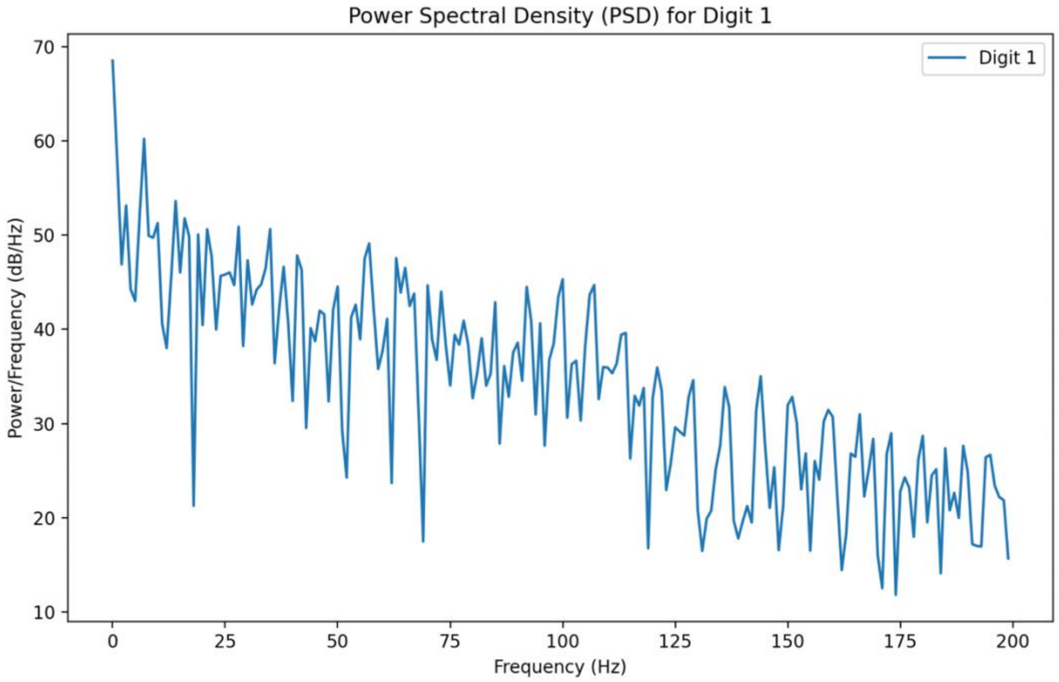


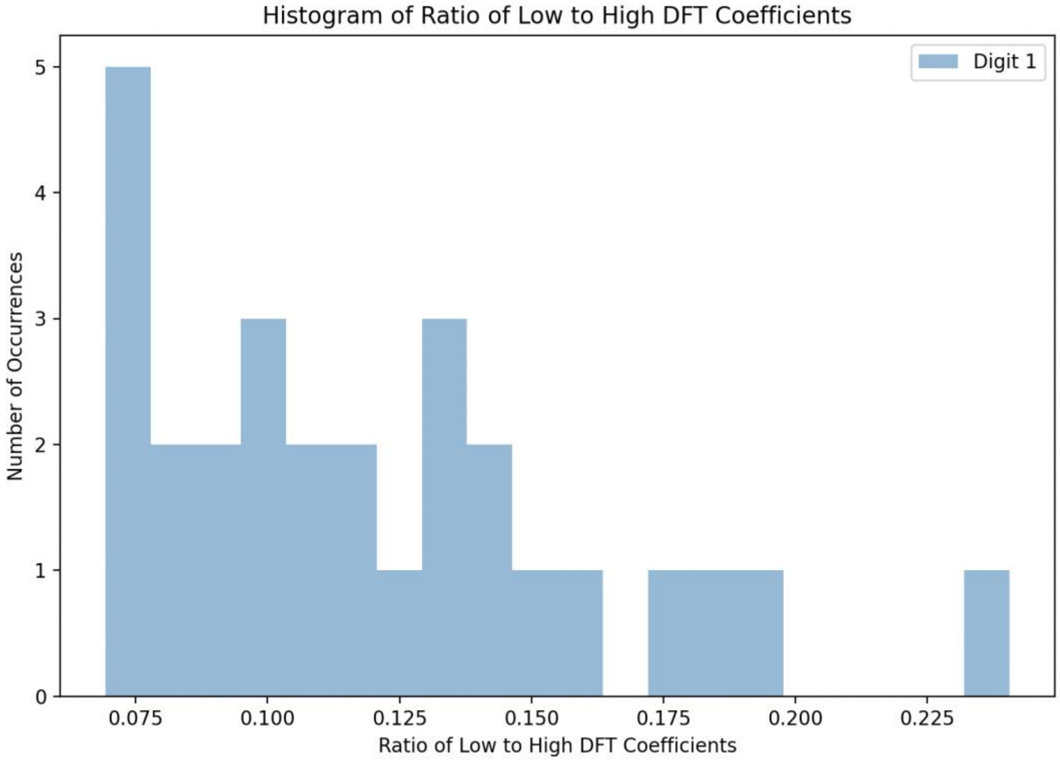


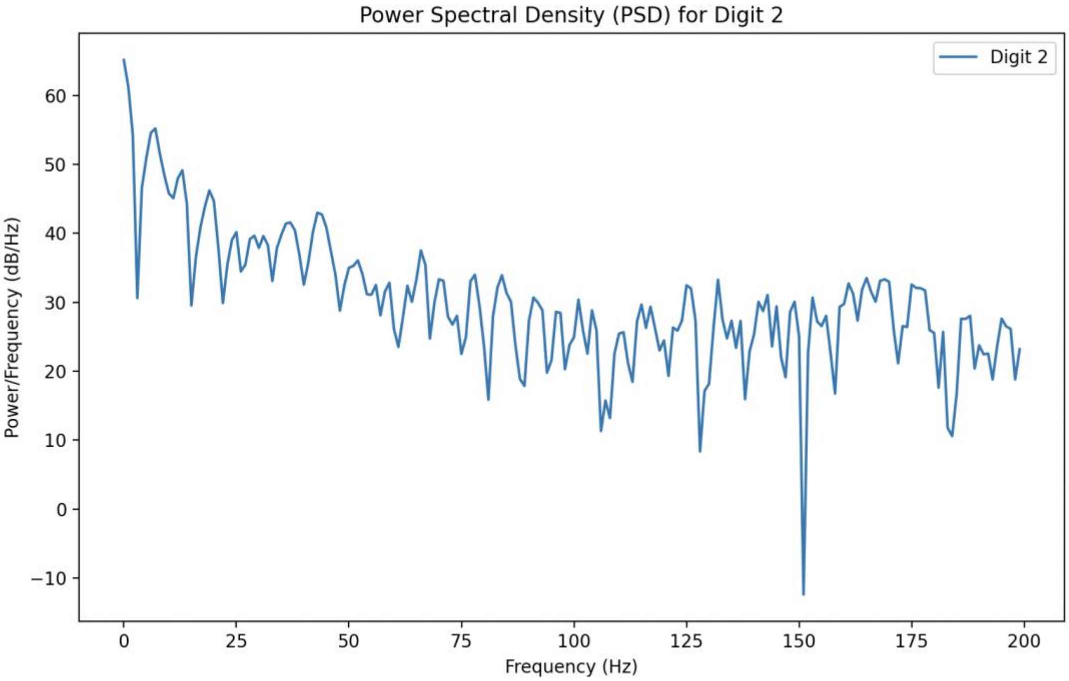
At first, I tried to plot the PSD for all the spoken digits together, but it wasn't easy to read, and you couldn't deduce much from it. So, I decided to individually plot the PSD of each token instead of plotting them together for visibility purposes. I also tried plotting the averages of each spoken digit, but it was still hard to read all together.

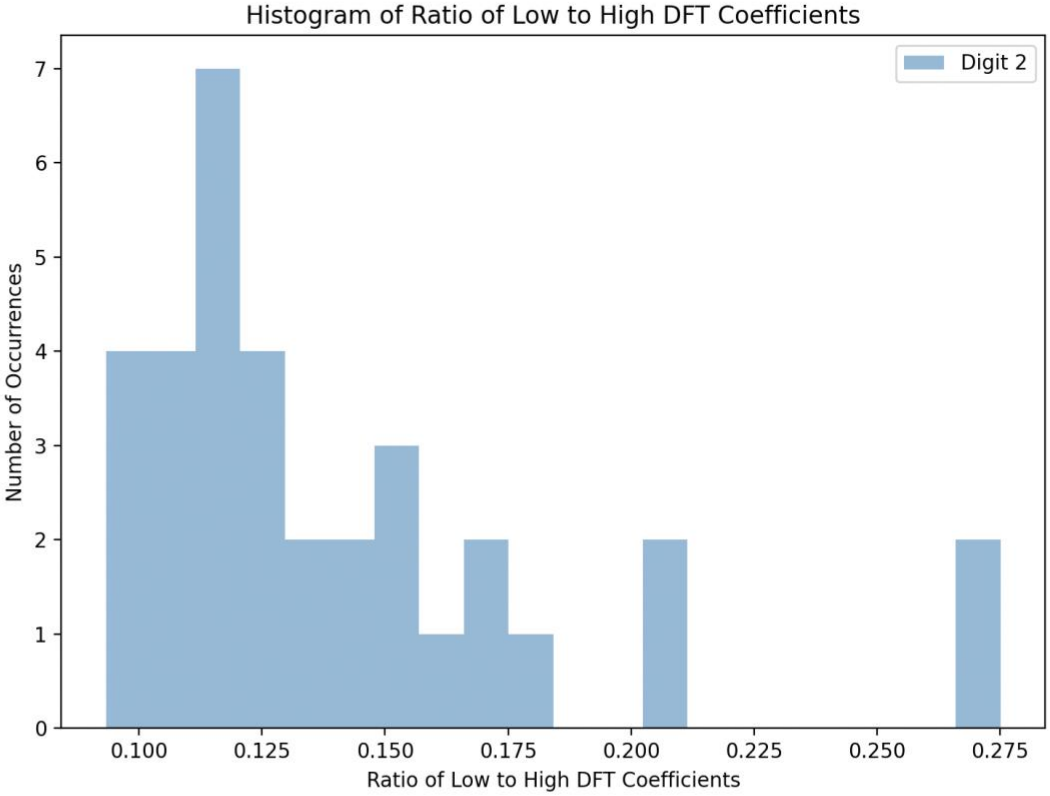


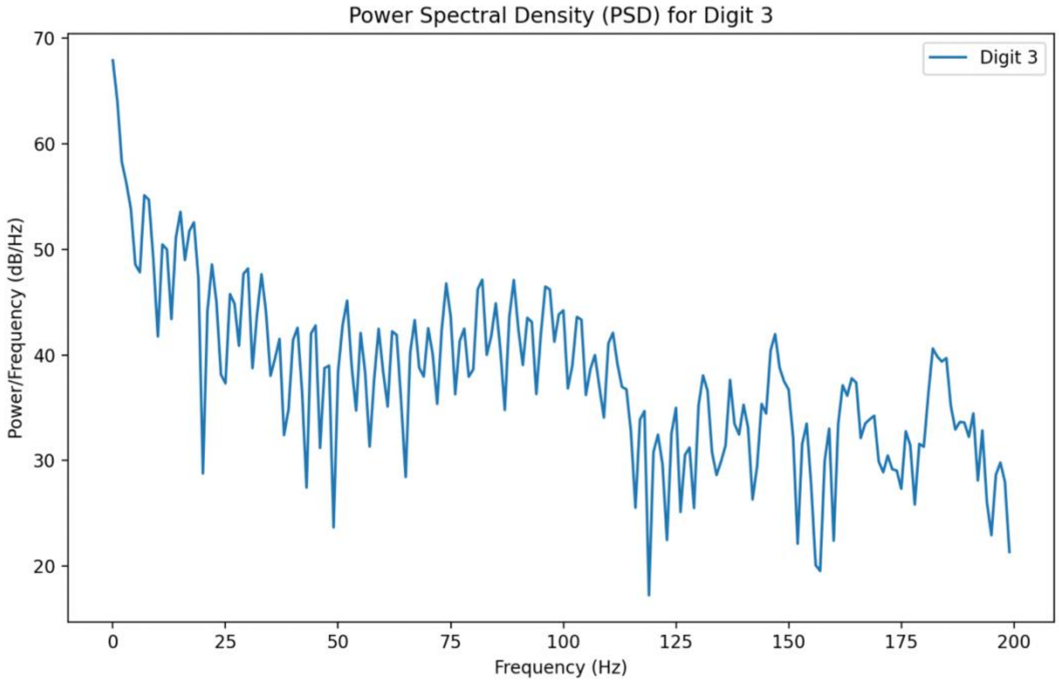


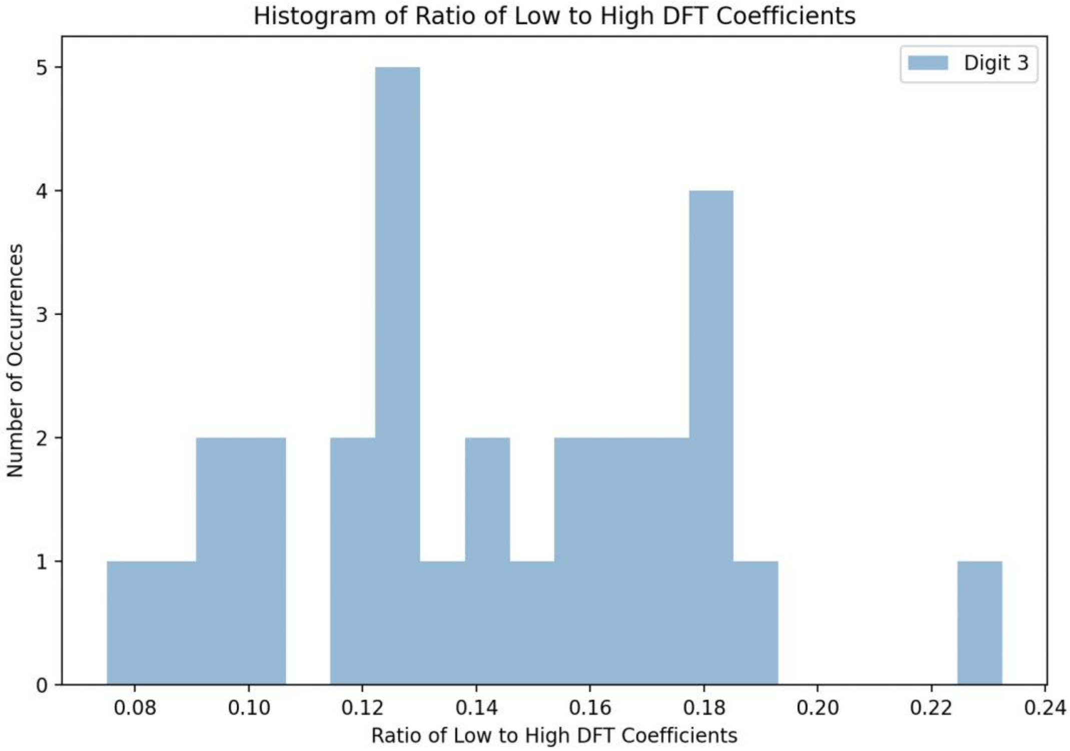

Power Spectral Density (PSD) for Digit 1
Digit 1
Power/Frequency (dB/Hz)
70
60
50
40
30
20
10
0
25
50
75
100
125
150
175
200
Frequency (Hz)

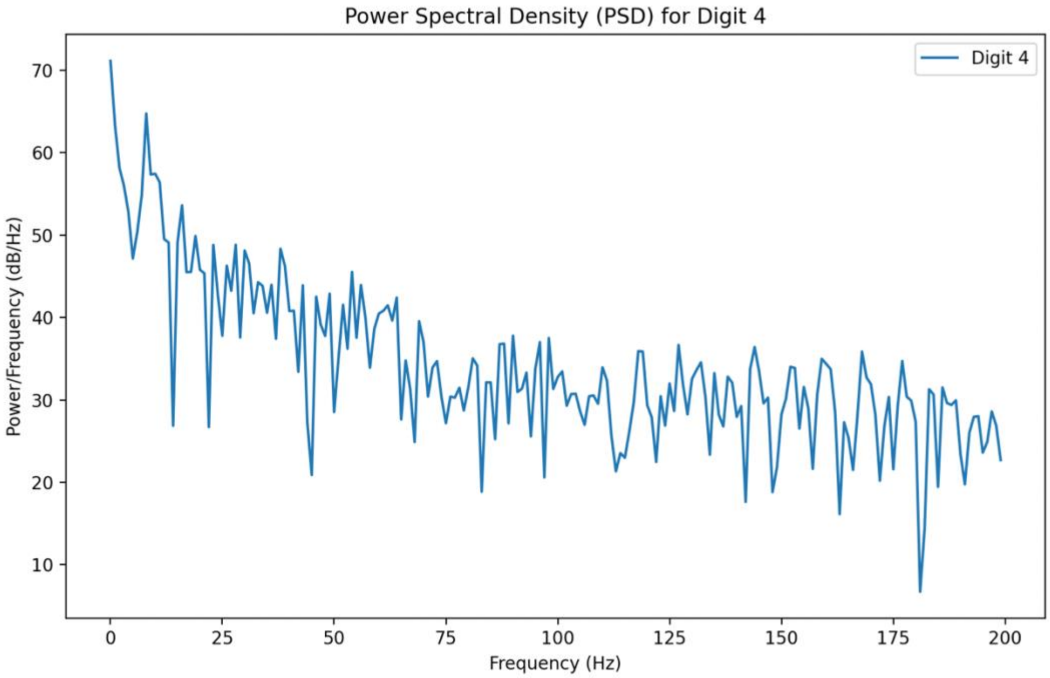

Histogram of Ratio of Low to High DFT Coefficients
Digit 1
Number of Occurrences
5
4
3
2
1
0
0.075
0.100
0.125
0.150
0.175
0.200
0.225
Ratio of Low to High DFT Coefficients

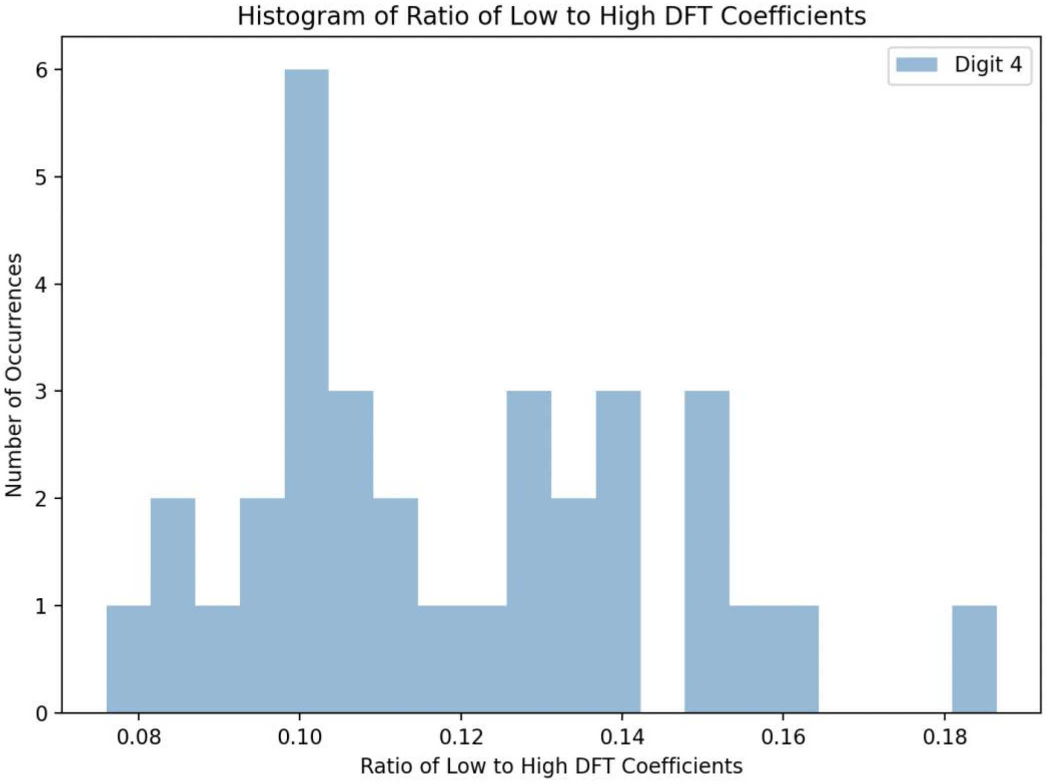

Power Spectral Density (PSD) for Digit 2
Digit 2
Power/Frequency (dB/Hz)
60
50
40
30
20
10
0
−10
0
25
50
75
100
125
150
175
200
Frequency (Hz)

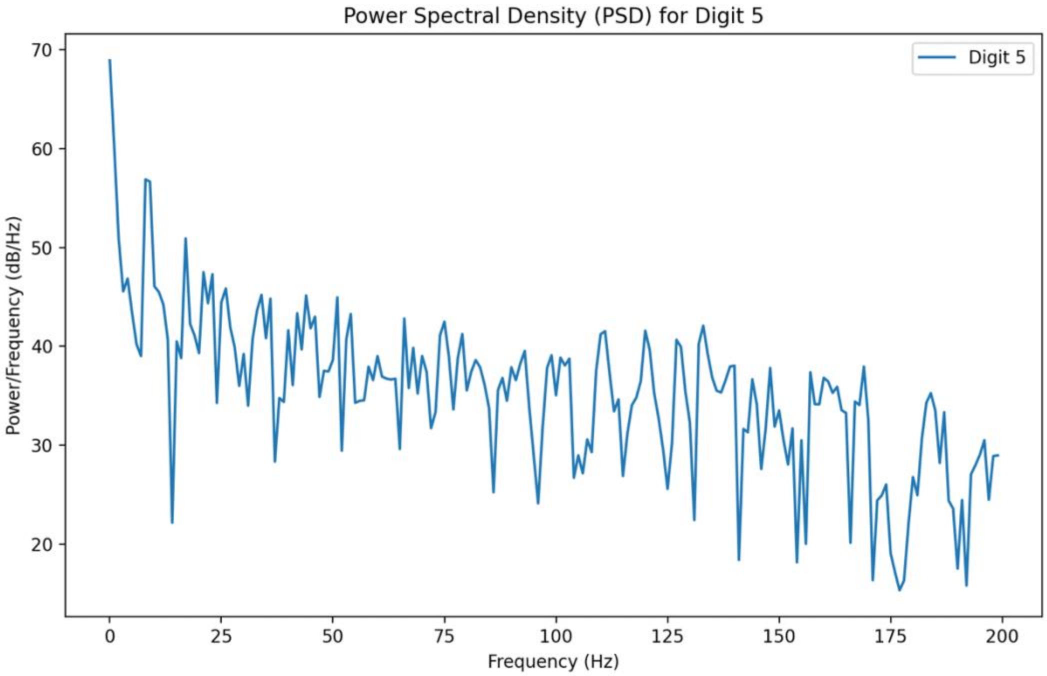

Histogram of Ratio of Low to High DFT Coefficients
Digit 2
Number of Occurrences
7
6
5
4
3
2
1
0
0.100
0.125
0.150
0.175
0.200
0.225
0.250
0.275
Ratio of Low to High DFT Coefficients

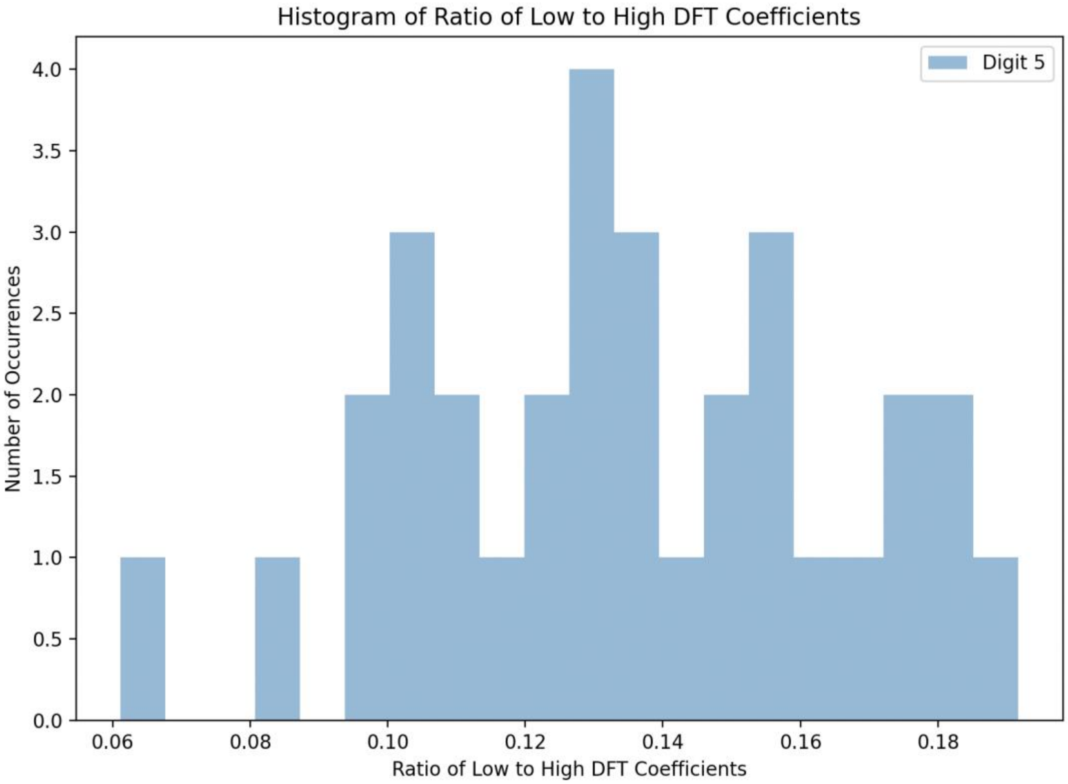

Power Spectral Density (PSD) for Digit 3
Digit 3
Power/Frequency (dB/Hz)
70
60
50
40
30
20
0
25
50
75
100
125
150
175
200
Frequency (Hz)

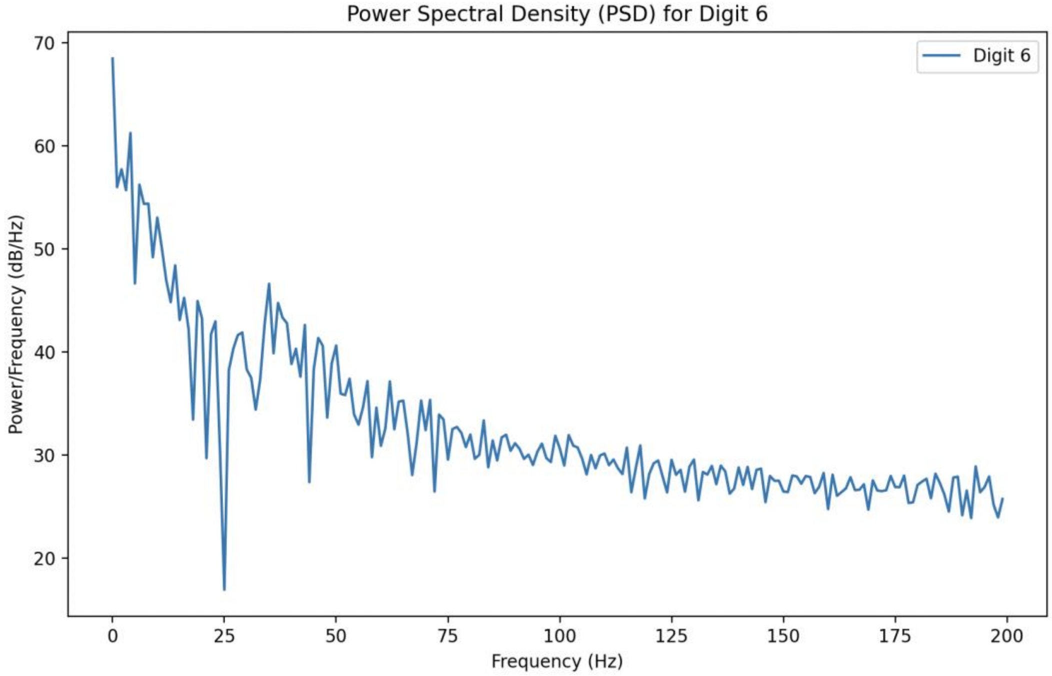

Histogram of Ratio of Low to High DFT Coefficients
Digit 3
Number of Occurrences
0
1
2
3
4
5
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.22
0.24
Ratio of Low to High DFT Coefficients

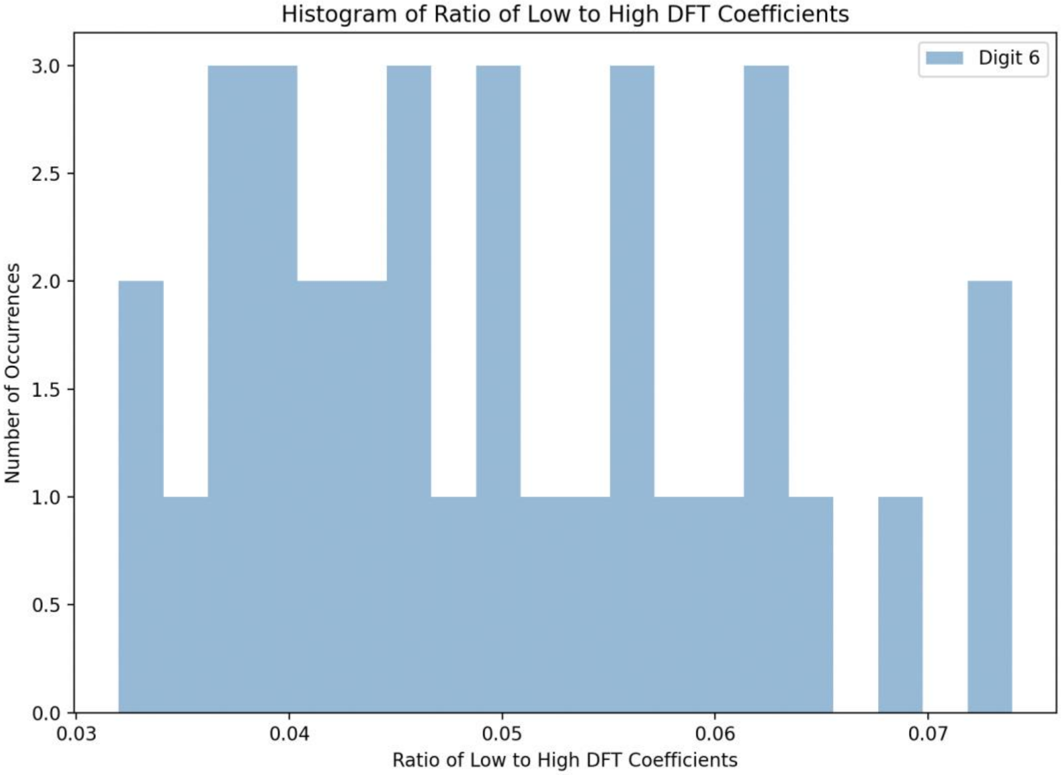

Power Spectral Density (PSD) for Digit 4
Digit 4
Power/Frequency (dB/Hz)
70
60
50
40
30
20
10
0
25
50
75
100
125
150
175
200
Frequency (Hz)

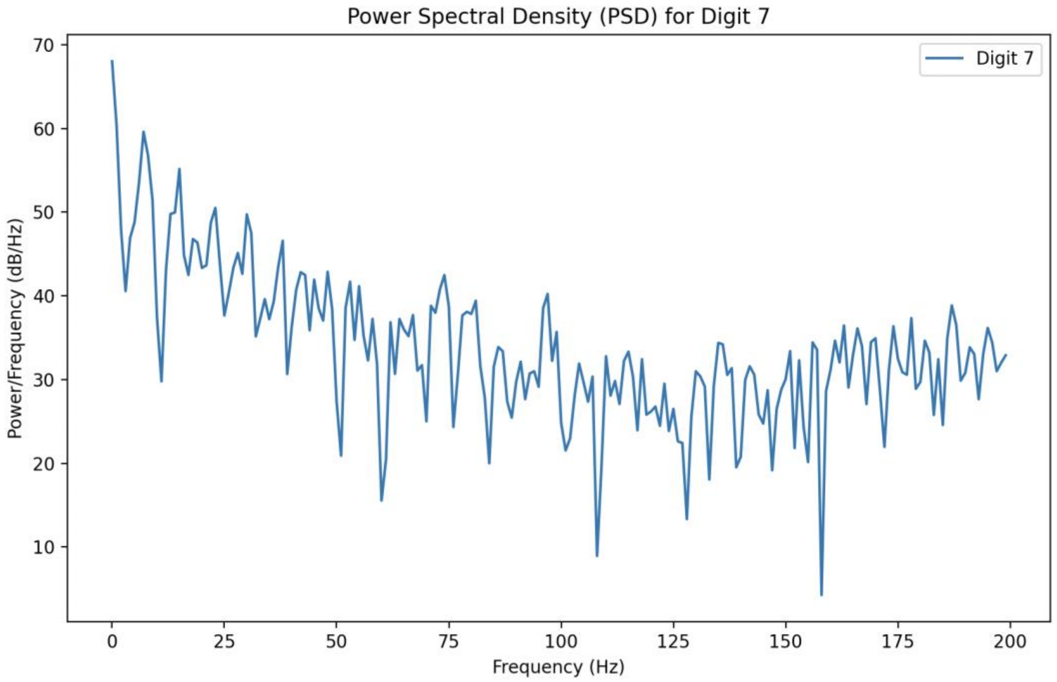

Histogram of Ratio of Low to High DFT Coefficients
Digit 4
Number of Occurrences
6
5
4
3
2
1
0
0.08
0.10
0.12
0.14
0.16
0.18
Ratio of Low to High DFT Coefficients

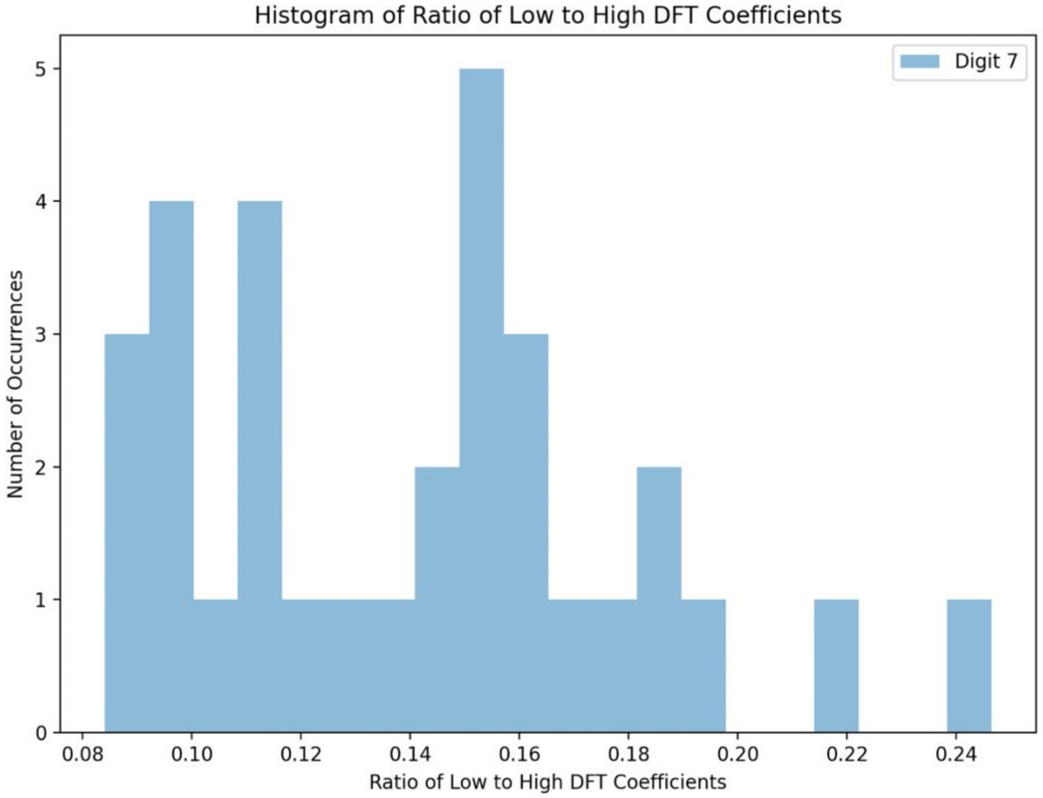

Power Spectral Density (PSD) for Digit 5
Digit 5
Power/Frequency (dB/Hz)
70
60
50
40
30
20
0
25
50
75
100
125
150
175
200
Frequency (Hz)

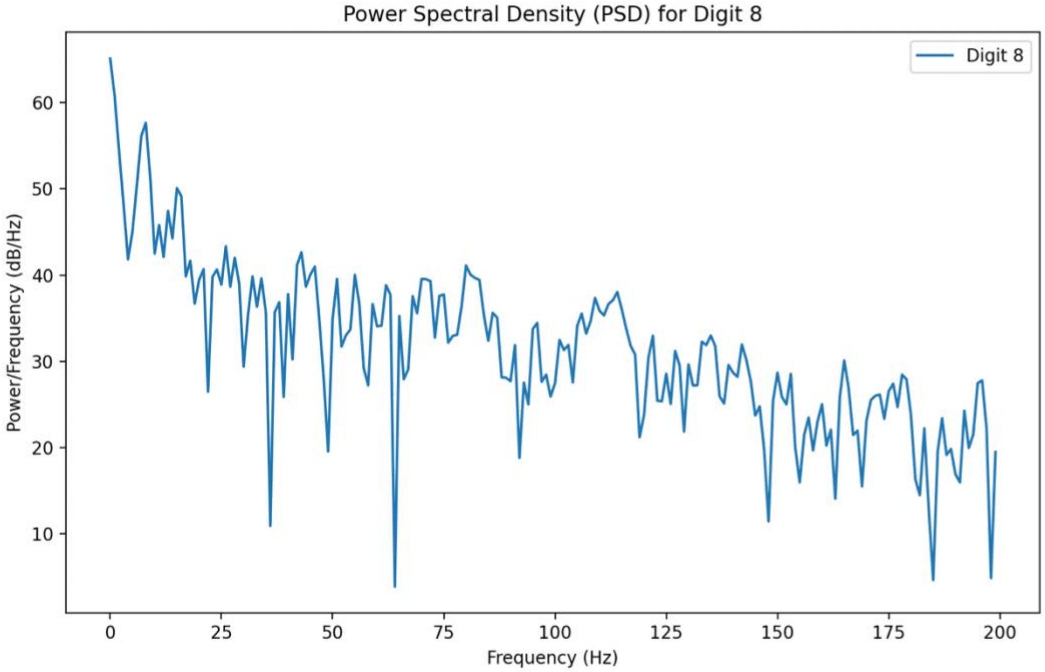

Histogram of Ratio of Low to High DFT Coefficients
Digit 5
Number of Occurrences
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.06
0.08
0.10
0.12
0.14
0.16
0.18
Ratio of Low to High DFT Coefficients

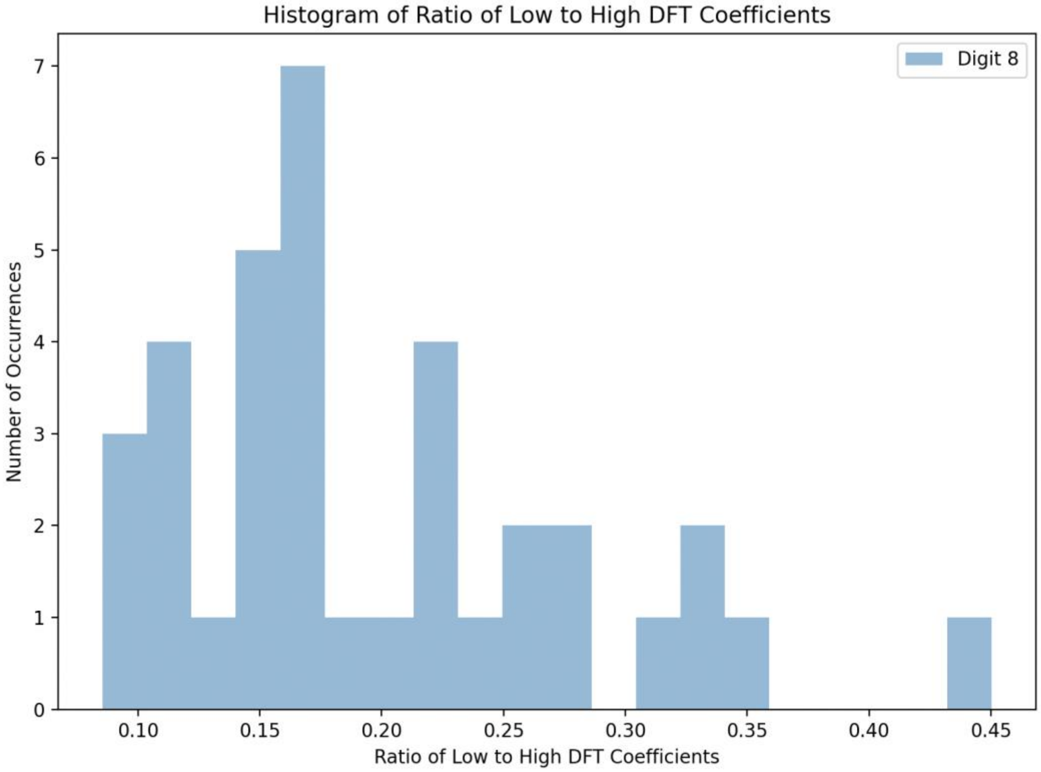

Power Spectral Density (PSD) for Digit 6
Digit 6
Power/Frequency (dB/Hz)
70
60
50
40
30
20
0
25
50
75
100
125
150
175
200
Frequency (Hz)

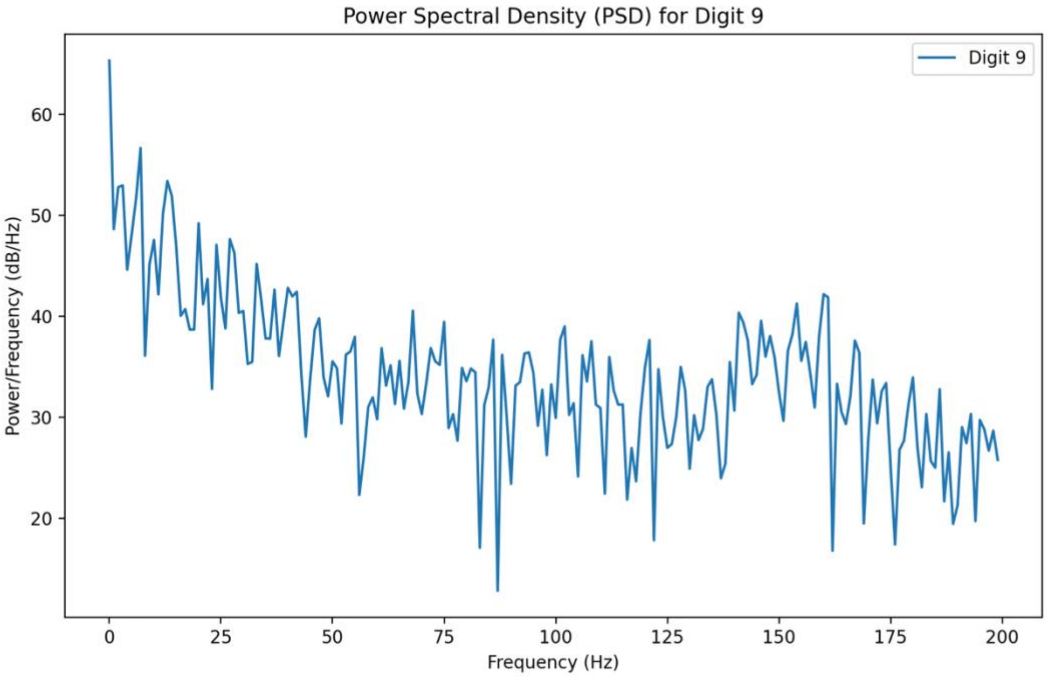

Histogram of Ratio of Low to High DFT Coefficients
Digit 6
Number of Occurrences
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.03
0.04
0.05
0.06
0.07
Ratio of Low to High DFT Coefficients

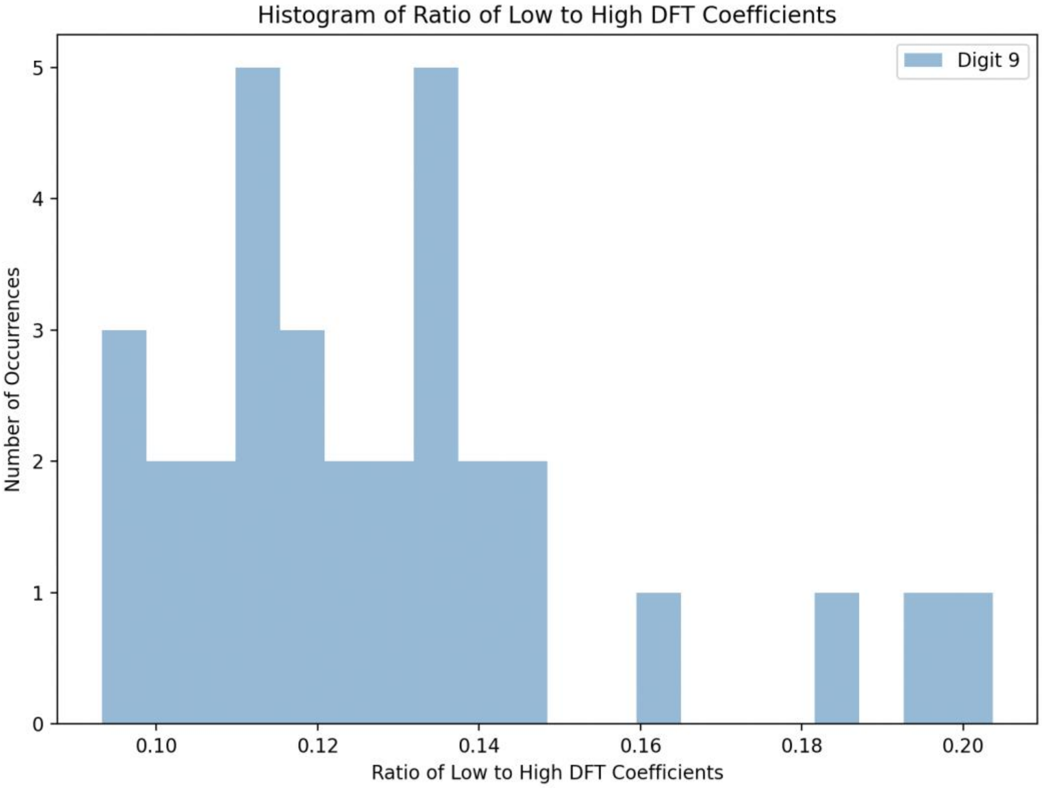

Power Spectral Density (PSD) for Digit 7
Digit 7
Power/Frequency (dB/Hz)
70
60
50
40
30
20
10
0
25
50
75
100
125
150
175
200
Frequency (Hz)


Histogram of Ratio of Low to High DFT Coefficients
Digit 7
Number of Occurrences
5
4
3
2
1
0
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.22
0.24
Ratio of Low to High DFT Coefficients


Power Spectral Density (PSD) for Digit 8
Digit 8
Power/Frequency (dB/Hz)
60
50
40
30
20
10
0
25
50
75
100
125
150
175
200
Frequency (Hz)


Histogram of Ratio of Low to High DFT Coefficients
Digit 8
Number of Occurrences
7
6
5
4
3
2
1
0
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
Ratio of Low to High DFT Coefficients

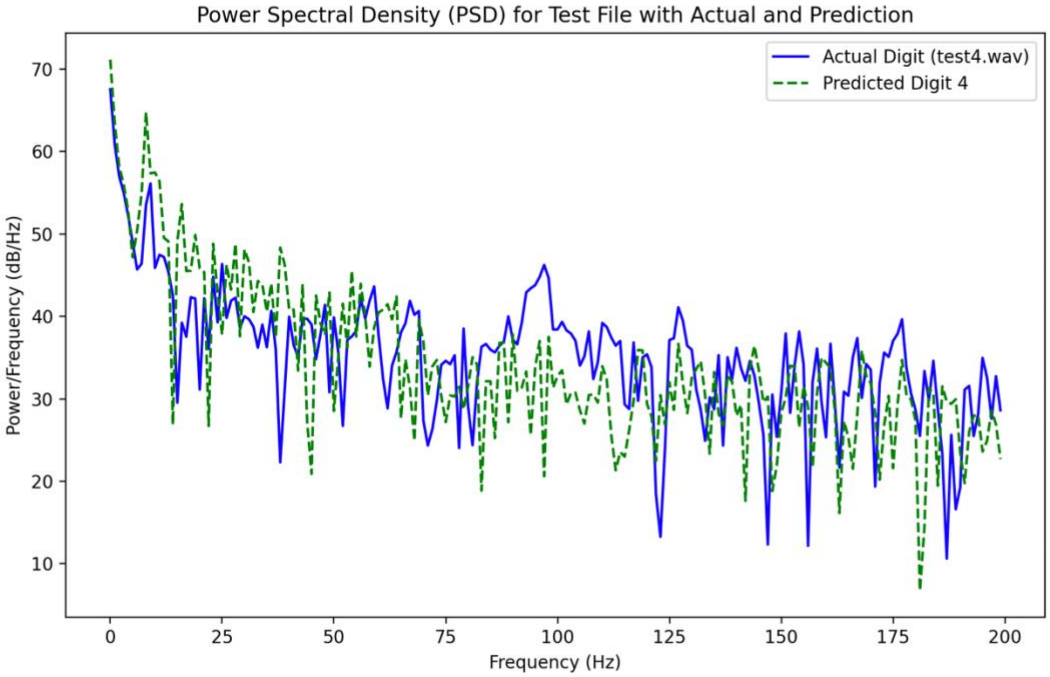

Power Spectral Density (PSD) for Digit 9
Digit 9
Power/Frequency (dB/Hz)
60
50
40
30
20
0
25
50
75
100
125
150
175
200
Frequency (Hz)

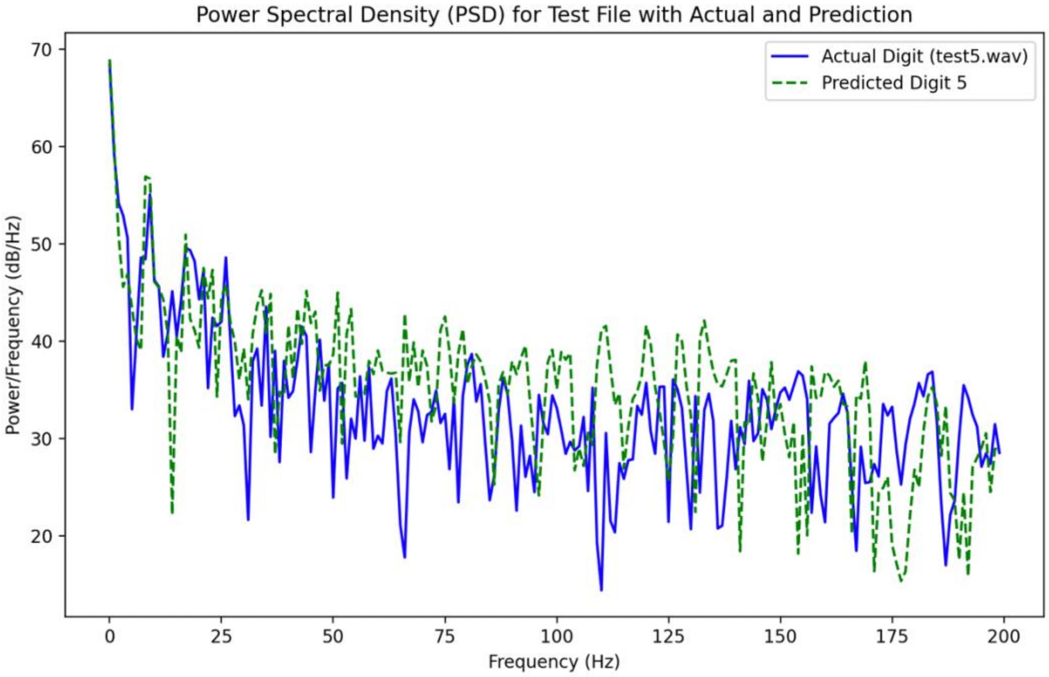

Histogram of Ratio of Low to High DFT Coefficients
Digit 9
Number of Occurrences
5
4
3
2
1
0
0.10
0.12
0.14
0.16
0.18
0.20
Ratio of Low to High DFT Coefficients

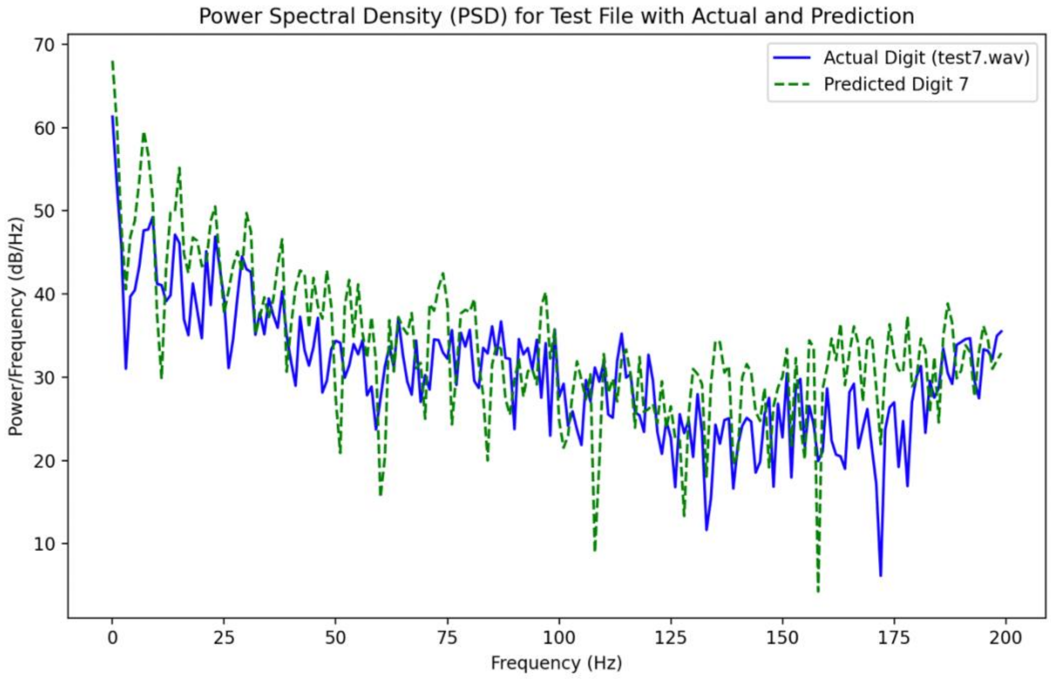
The Power Spectral Density (PSD) and Histogram of Ratio of Low to High DFT Coefficients are plotted above for each digit. I used one sample for each digit from the training set to plot these graphs. I only used one sample each for the purposes of complexity and for comparison. The PSD graph shows you the distribution of power across the different frequencies. This shows the unique frequency characteristics associated with each individual digit. The x-axis represents the frequency and shows which frequencies contribute the most to the spoken digit. The y-axis provides the amount of energy or power and is useful in showing which frequencies are more pronounced or emphasized. For instance, the digit ‘one’ has almost a constant linear decline while the digit ‘six’ has a high peak at the beginning and quickly dips and tailors off. This is a good indicator of how to distinguish between those two digits.

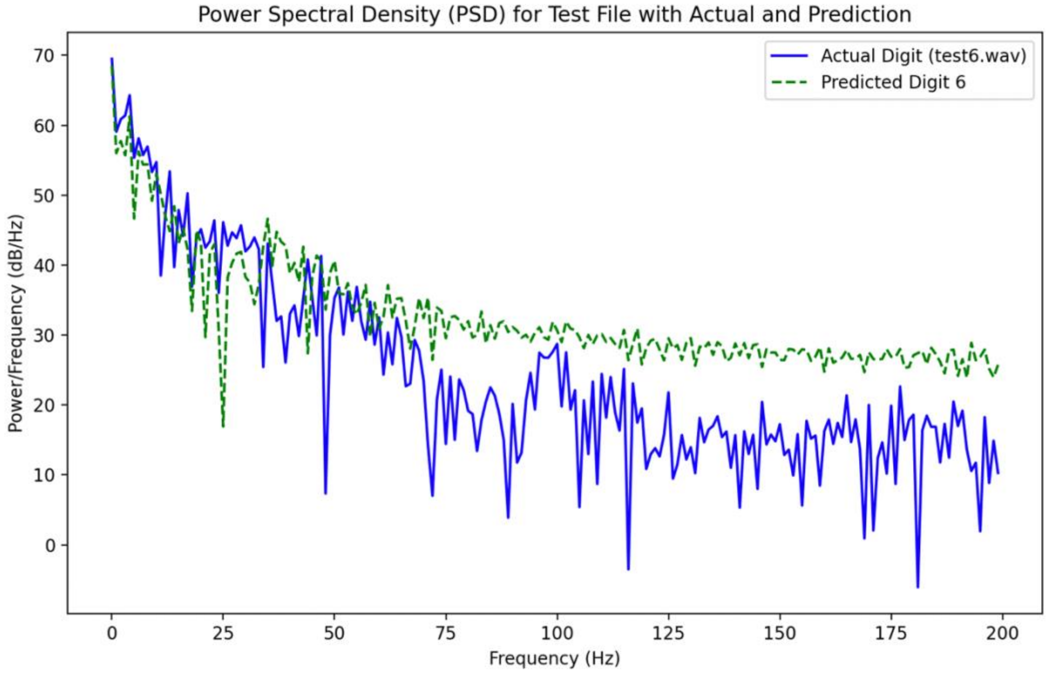
The histograms help observe whether certain digits tend to have specific ranges of ratio of low to high DFT coefficients than others. The peaks or clusters in the histogram show where the feature values are concentrated which provides a good distinguishable characteristic. Back to digits ‘one’ and ‘six’, you can see from the histogram that ‘one’ has a big cluster below 0.1625 and ‘six’ is more concentrated below 0.065 with much taller bars. The histogram is another way to distinguish between the digits by visually representing how the ratio of low to high DFT coefficients changes across the different digits.

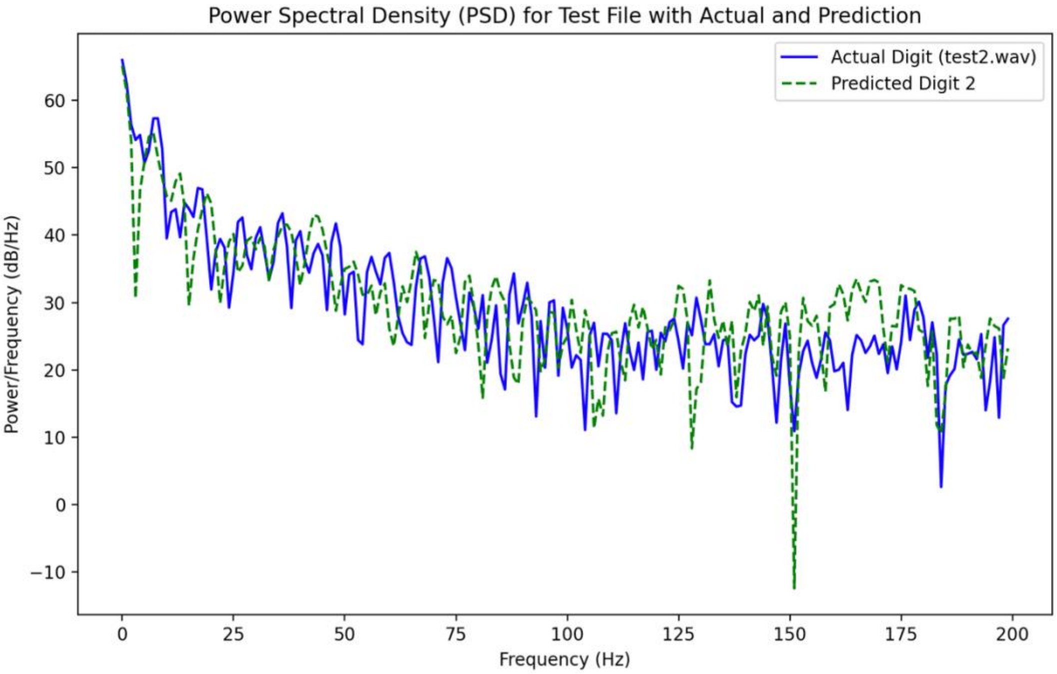
To test the trained classifier, I took 10 samples (one for each digit) from the provided dataset and set the apart. Therefore, I trained the classifier with 30 samples of each digit. Along with the digit prediction, I plotted the PSD of the actual test file and its predicted digit to visually see any similarities. For the predicted digit, I used the PSD of the associated digit in the training set.


Power Spectral Density (PSD) for Test File with Actual and Prediction
Actual Digit (test4.wav)
Predicted Digit 4
Power/Frequency (dB/Hz)
70
60
50
40
30
20
10
0
25
50
75
100
125
150
175
200
Frequency (Hz)

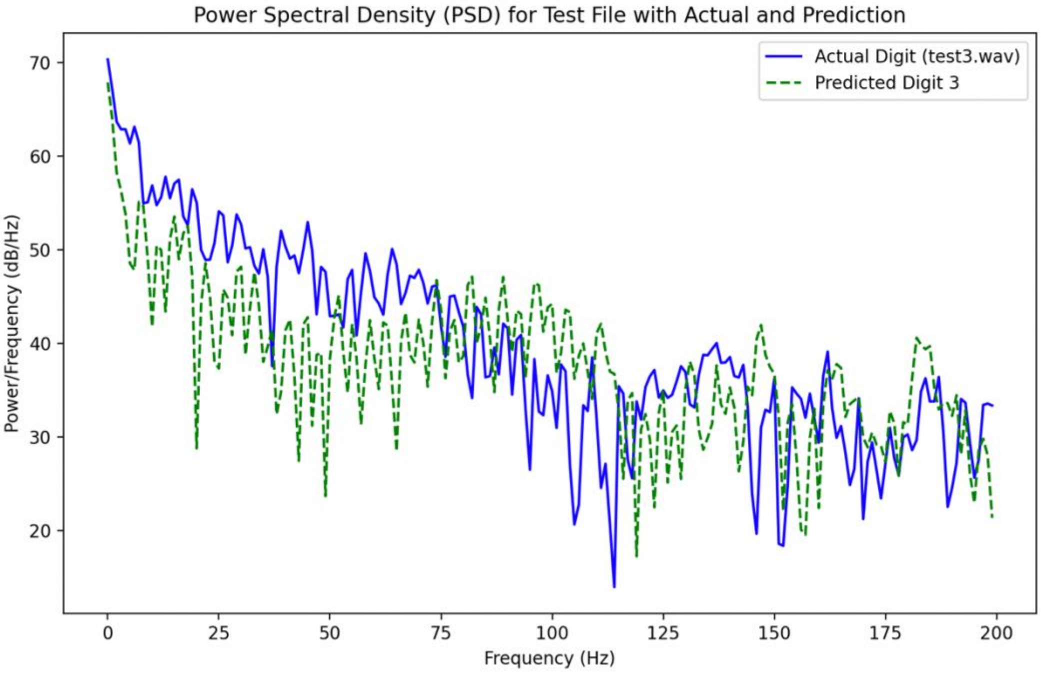

Power Spectral Density (PSD) for Test File with Actual and Prediction
Actual Digit (test5.wav)
Predicted Digit 5
Power/Frequency (dB/Hz)
70
60
50
40
30
20
0
25
50
75
100
125
150
175
200
Frequency (Hz)


Power Spectral Density (PSD) for Test File with Actual and Prediction
Actual Digit (test7.wav)
Predicted Digit 7
Power/Frequency (dB/Hz)
Frequency (Hz)
70
60
50
40
30
20
10
0
25
50
75
100
125
150
175
200

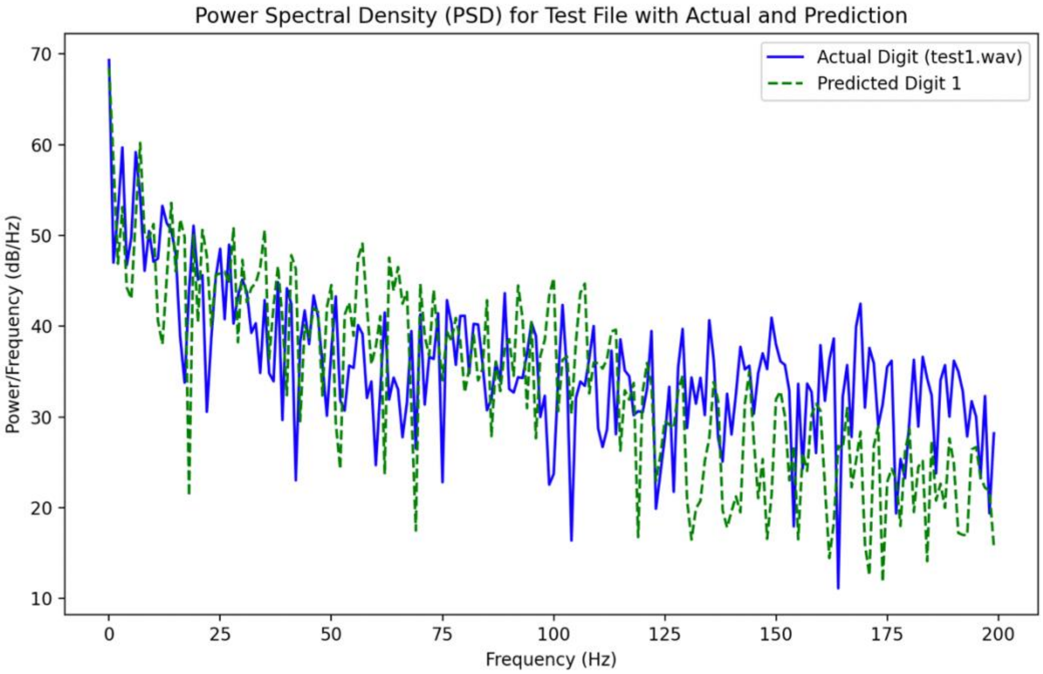

Power Spectral Density (PSD) for Test File with Actual and Prediction
Actual Digit (test6.wav)
Predicted Digit 6
Power/Frequency (dB/Hz)
Frequency (Hz)
70
60
50
40
30
20
10
0
0
25
50
75
100
125
150
175
200

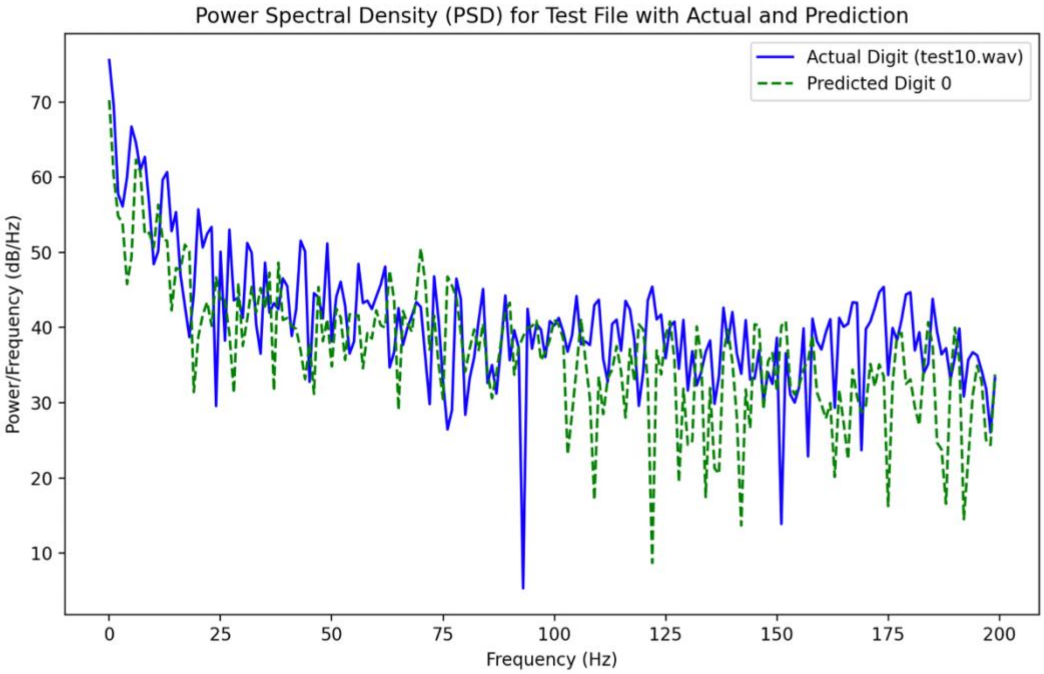


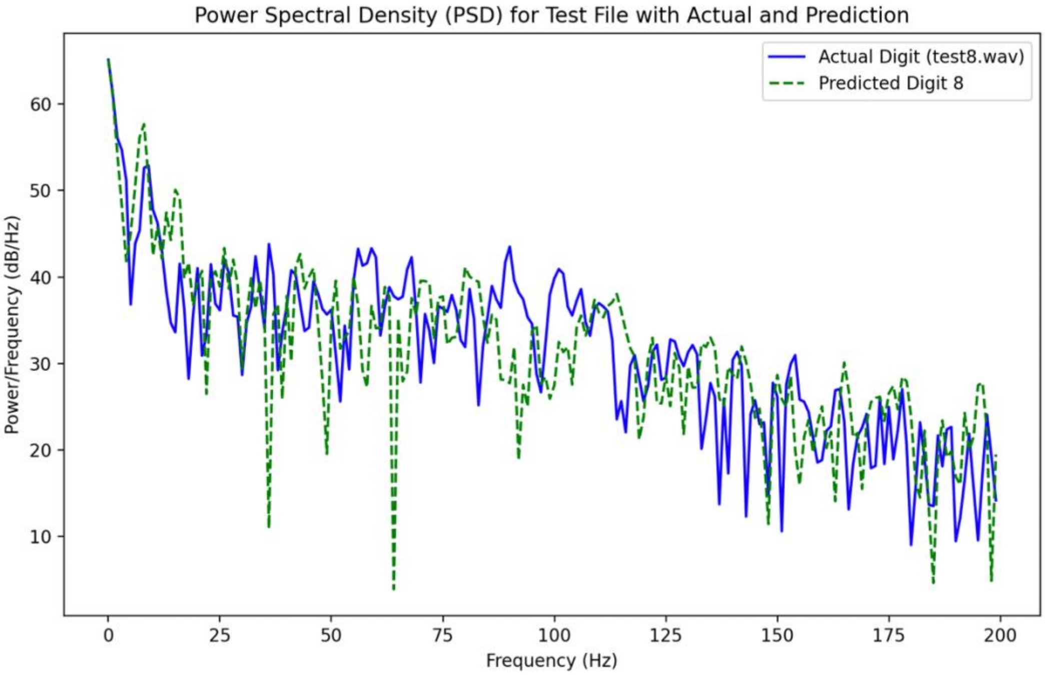
Power Spectral Density (PSD) for Test File with Actual and Prediction

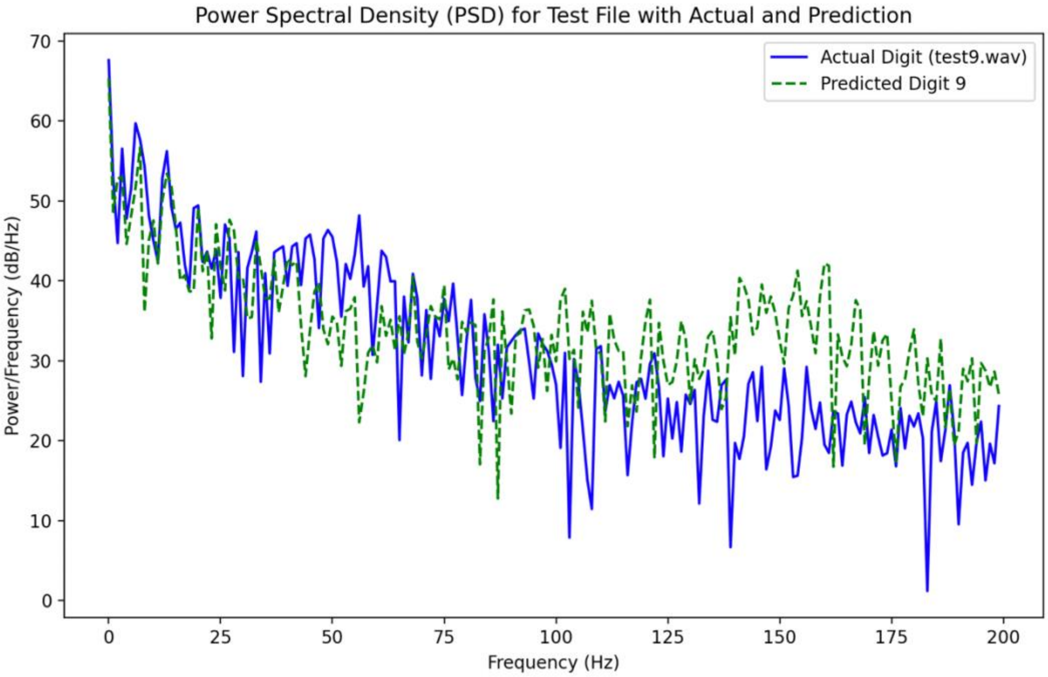


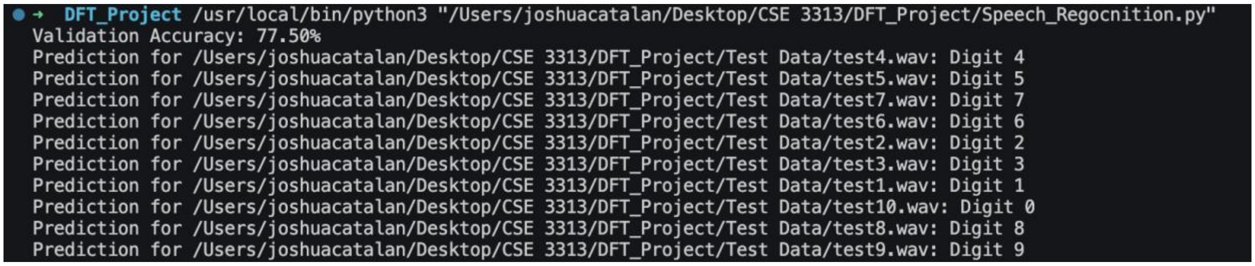
Power Spectral Density (PSD) for Test File with Actual and Prediction


Power Spectral Density (PSD) for Test File with Actual and Prediction
Actual Digit (test1.wav)
Predicted Digit 1
Power/Frequency (dB/Hz)
70
60
50
40
30
20
10
0
25
50
75
100
125
150
175
200
Frequency (Hz)


Power Spectral Density (PSD) for Test File with Actual and Prediction
Actual Digit (test10.wav)
Predicted Digit 0
Power/Frequency (dB/Hz)
70
60
50
40
30
20
10
0
25
50
75
100
125
150
175
200
Frequency (Hz)


Power Spectral Density (PSD) for Test File with Actual and Prediction
Actual Digit (test8.wav)
Predicted Digit 8
Power/Frequency (dB/Hz)
60
50
40
30
20
10
0
25
50
75
100
125
150
175
200
Frequency (Hz)


Power Spectral Density (PSD) for Test File with Actual and Prediction
Actual Digit (test9.wav)
Predicted Digit 9
Power/Frequency (dB/Hz)
70
60
50
40
30
20
10
0
0
25
50
75
100
125
150
175
200
Frequency (Hz)



```
● → DFT_Project /usr/local/bin/python3 "/Users/joshuacatalan/Desktop/CSE 3313/DFT_Project/Speech_Regocnition.py"
  Validation Accuracy: 77.50%
  Prediction for /Users/joshuacatalan/Desktop/CSE 3313/DFT_Project/Test Data/test4.wav: Digit 4
  Prediction for /Users/joshuacatalan/Desktop/CSE 3313/DFT_Project/Test Data/test5.wav: Digit 5
  Prediction for /Users/joshuacatalan/Desktop/CSE 3313/DFT_Project/Test Data/test7.wav: Digit 7
  Prediction for /Users/joshuacatalan/Desktop/CSE 3313/DFT_Project/Test Data/test6.wav: Digit 6
  Prediction for /Users/joshuacatalan/Desktop/CSE 3313/DFT_Project/Test Data/test2.wav: Digit 2
  Prediction for /Users/joshuacatalan/Desktop/CSE 3313/DFT_Project/Test Data/test3.wav: Digit 3
  Prediction for /Users/joshuacatalan/Desktop/CSE 3313/DFT_Project/Test Data/test1.wav: Digit 1
  Prediction for /Users/joshuacatalan/Desktop/CSE 3313/DFT_Project/Test Data/test10.wav: Digit 0
  Prediction for /Users/joshuacatalan/Desktop/CSE 3313/DFT_Project/Test Data/test8.wav: Digit 8
  Prediction for /Users/joshuacatalan/Desktop/CSE 3313/DFT_Project/Test Data/test9.wav: Digit 9
```

The classifier had a validation accuracy of 77.50%. From the 10 samples I set aside, I renamed them test and the corresponding spoken digit so that it's easier to tell if it was predicted correct or not. For test10, the corresponding digit is 0. From the prediction output for each test file, all the test files were accurately predicted. An accuracy of 77.50% isn't the best but it will predict accurately majority of the time. In the graphs with the predicted digit and the test file, you can see how similar they are. Some may have slight variations but overall, it shows enough to make a prediction at first glance.