Analyzing Speech Impairments: A Machine Learning Approach to Dysarthria Detection

Milan Gupta
*Francis Parker High School San Diego, California
*milangupta2026@gmail.com

Abstract

Dysarthria includes dysfunction in the nerves and muscles controlling speech, leading to unclear spoken words. While many studies have been carried out to examine speech impairment, the variation of this problem among people with a similar dysarthria diagnosis has necessitated the need for more research in this area. The particular type and severity of the impairment are essential to monitor the progress of dysarthria and make effective therapeutic interventions. This project describes a Convolutional Neural Network (CNN) model for dysarthria detection, where several acoustic features are extracted in the form of zero crossing rates, Mel Frequency Cepstral Coefficients (MFCCs), spectral centroids, and spectral roll-off. Using the TORGO database of speech signals, training the model, and testing it for its efficiency has shown much promise in the early diagnosis of dysarthric speech. The numerical results indicate that the model design provides an efficiency of nearly 95%, which is higher than previous model architectures. This model aims to identify the condition early and help improve the management of dysarthria through timely and accurate diagnosis.

Introduction

Dysarthria is a complex motor speech disorder resulting from neurological impairments that affect the muscles used in speech production. It is characterized by slurred, slow, and unpredictable speech. Other features range from abnormally loud or soft volumes to distorted vocal qualities. This involves a disruption of one or several subsystems, including respiration, phonation, resonance, articulation, and prosody. Such disorders have their etiologies in various pathologies of the nervous system and lead to a varied array of speech impairments.

The physiological bases of dysarthria are highly integrated and complicate the isolation of speech functions affected. While respiration relies on controlled activity by the muscles in the abdomen and thorax and diaphragmatic action, phonation relies on the laryngeal mechanisms. Similarly, resonance implicates the pharyngeal, oral musculature, soft palate, and nasal cavities, and the movements of the tongue, jaw, and lips control articulation. Dysarthria is severe to a degree, and its nature depends upon the site and severity of neurological insult, resulting in various forms of dysarthria: flaccid, spastic, ataxic, hypokinetic, hyperkinetic, and mixed.

Classification of dysarthria, as well as the determination of its severity, is essential for carrying out efficient management and therapeutic planning. This paper insists on using Convolutional Neural Networks (CNNs) and establishes them as a more advanced tool for detecting dysarthria at an earlier stage than traditional techniques. We aim to integrate sophisticated speech-processing techniques into neural network models to enhance patient diagnosis. The proposed manuscript consists of the following structure. First, this paper describes the methodology for developing a CNN-based classification model (Section 2). We present the results in Section 3, discuss the results in Section 4, and conclude with an overview of our study and future directions in Section 5.

The economic and health care costs that Dysarthria can cause lead to significant financial impacts for those with the condition. A study examining the Lee Silverman Voice Treatment for Parkinson's disease-associated dysarthria suggests that the tool may not always be cost-effective depending on the patient's outcome.

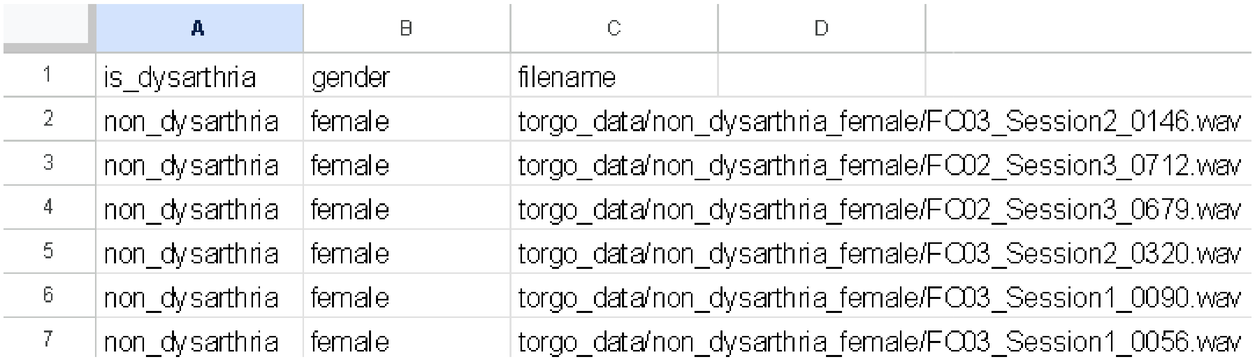
Proposed Methodology

Data Collection and Preparation

This study uses the Universal Access Dysarthic speech corpus and the TORGO database. These databases include the dysarthric and non-dysarthric speech samples and their respective data from each of the male and female speakers in the study. The speaking data comprises `.wav` files, capturing the acoustics of each subject. The sample data files and their respective details are shown in Figure 1. The data was cleaned and modified effectively with background noises by utilizing a Wiener filter. This effectively cleaned the data by minimizing the mean square error between the estimated random process and the desired signal.

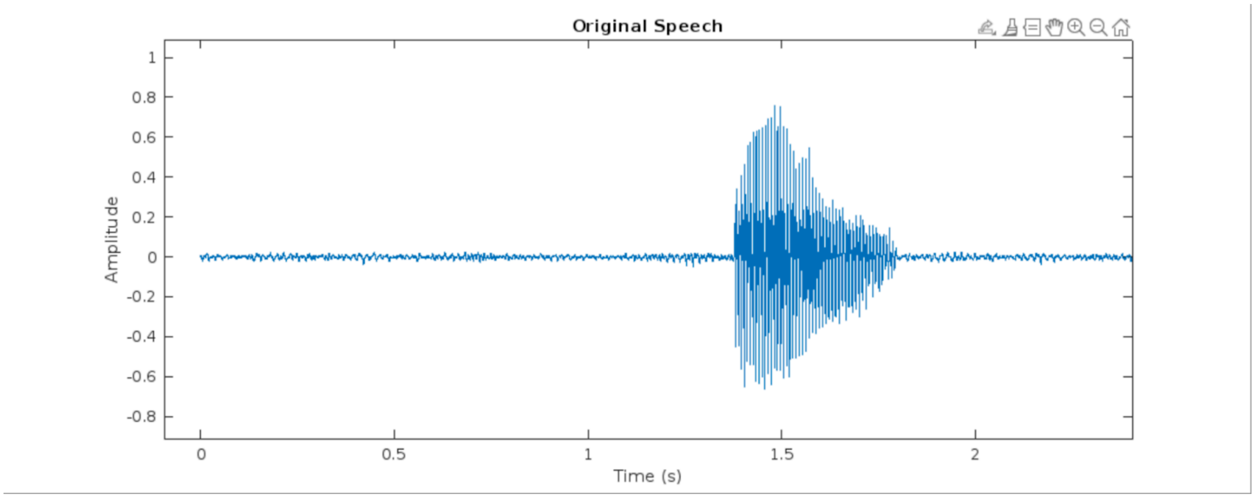
Data Analysis

In addition to auditory speech evaluation, several signal processing techniques were employed to analyze the speech signals, focusing on changes in vowel formants, fundamental frequency (f0), duration of the speech signal, amplitude variations, prolonged vowel duration, voice onset time, and variations in speech tempo. Key features extracted for analysis included the Short-Time Fourier Transform (STFT), Mel Frequency Cepstral Coefficients (MFCC), spectral centroid, spectral bandwidth, spectral roll-off, and zero-crossing rate. These features were extracted from both dysarthric and non-dysarthric speech samples, with detailed feature plots for dysarthric female speakers provided in Figure 2.

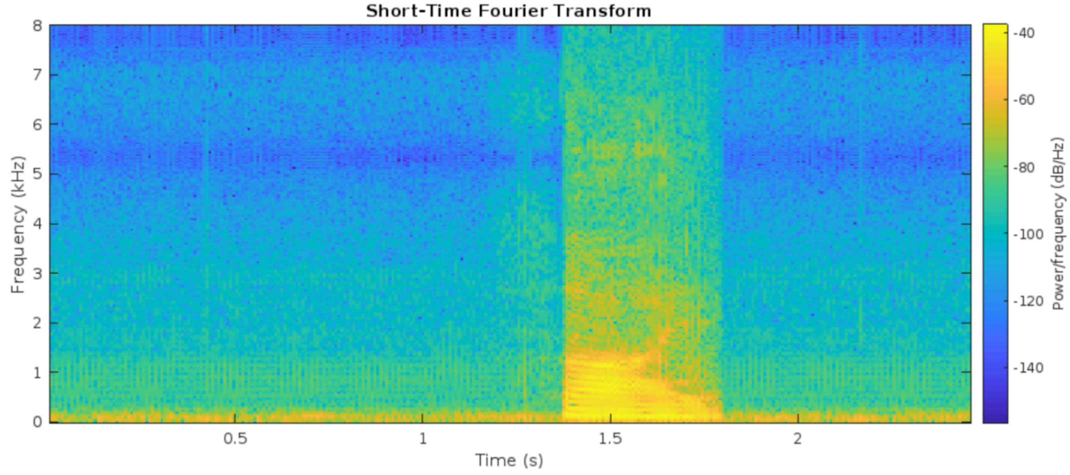


	A	B	C	D	
1	is_dysarthria	gender	filename		
2	non_dysarthria	female	torgo_data/non_dysarthria_female/FC03_Session2_0146.wav		
3	non_dysarthria	female	torgo_data/non_dysarthria_female/FC02_Session3_0712.wav		
4	non_dysarthria	female	torgo_data/non_dysarthria_female/FC02_Session3_0679.wav		
5	non_dysarthria	female	torgo_data/non_dysarthria_female/FC03_Session2_0320.wav		
6	non_dysarthria	female	torgo_data/non_dysarthria_female/FC03_Session1_0090.wav		
7	non_dysarthria	female	torgo_data/non_dysarthria_female/FC03_Session1_0056.wav		

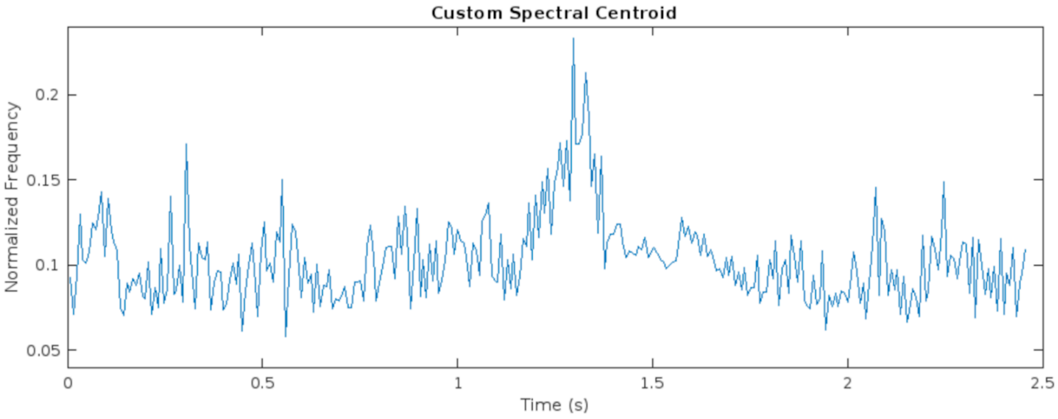
Fig. 1 Block schematic of methodology.


Original Speech
Amplitude
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
0
0.5
1
1.5
2
Time (s)

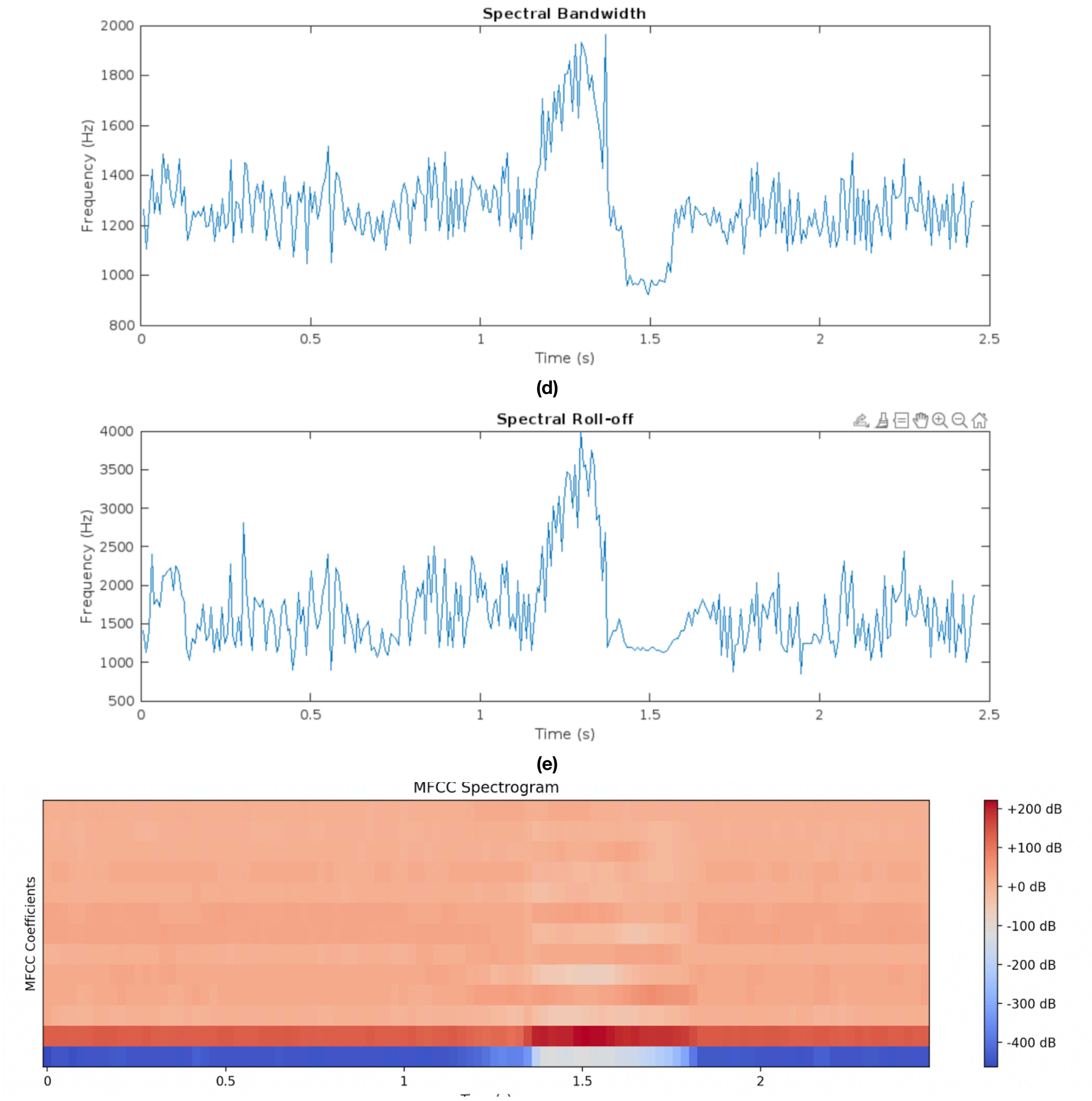
(a)


Short-Time Fourier Transform
Frequency (kHz)
8
7
6
5
4
3
2
1
0
0.5
1
1.5
2
Time (s)
Power/frequency (dB/Hz)
-40
-60
-80
-100
-120
-140

(b)


Custom Spectral Centroid
Normalized Frequency
0.2
0.15
0.1
0.05
0
0.5
1
1.5
2
2.5
Time (s)

(c)



(d)

(e)

(f)

Fig. 2 Speech analysis (Type 1: Male without Dysarthria) a) Original speech b) Short-term Fourier transform (STFT) c) Spectral centroids d) Spectral bandwidth (p = 1, p = 2, p = 3) e) Spectral roll-off f) Mel-frequency cepstral coefficients (MFCC) spectrogram [Phrase: Thigh]

Data Pre-Processing

The preprocessing phase involved converting the unstructured audio data into numeric features and standardizing the sample rate to 128 features per second. The audio signals were transformed into Mel Spectrograms to facilitate feature extraction. To prepare the data for model training, the feature matrix was normalized to ensure consistency. The dysarthria status of each speaker was encoded as a categorical variable. Subsequently, the dataset was split into 80% training data and 20% testing data.

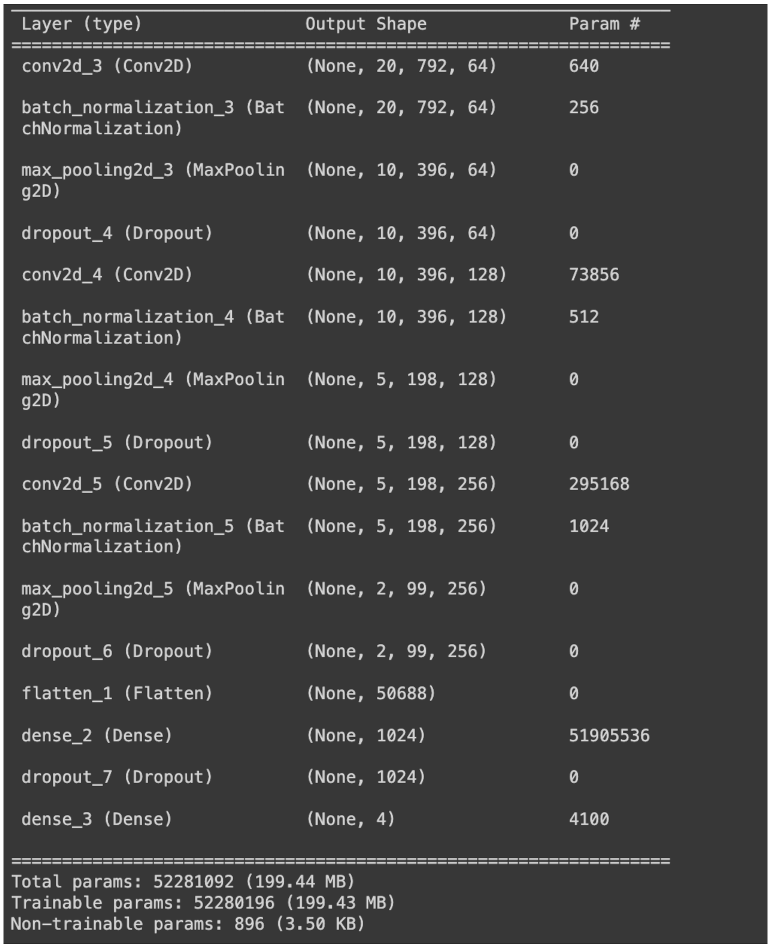
Data Modeling

The preprocessed data was transformed into Mel-frequency cepstral coefficients (MFCCs) extracted from each audio file. These were then used as input into a design to identify patterns associated with Dysarthria. This input data is formatted into the shape that matches the input layer of the model, specifically (20, 792, 1) for each audio sample, where 20 represents the number of MFCC coefficients, and 792 represents the maximum number of frames after padding.

The first layer of the CNN contains 64 filters with a kernel size of 3 × 3 to capture basic features from the input data, such as edges and simple speech-related textures. A kernel size, here, specifies the dimensions of the filter used in each convolutional layer, which directly influences the input data during the process of feature extraction. The output from this layer is batch-normalized to enable the activation normalization and increase the stability of the network. Subsequently, a max-pooling layer with a window of size 2 × 2 reduces the spatial dimensions for all feature maps by half. This helps reduce computational complexity and further prevents overfitting through feature abstraction. The architecture proceeds to a second convolutional layer of 128 filters, all of size 3 × 3. This layer delves more deeply into the details of the extracted features, learning more intricate and higher-level patterns that aid in the recognition of dysarthria. It is followed by a batch normalization and max-pooling sequence to further refine and reduce the feature maps.

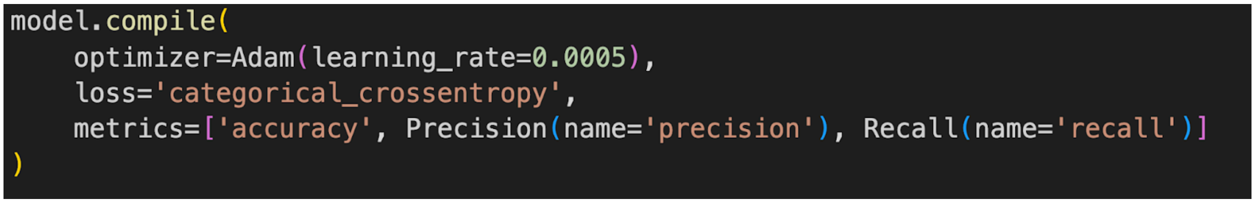
To prevent overfitting, a dropout layer with a rate of 0.3 is added after the pooling stages to make the model resistant. Overfitting can cause the model to capture unrelated noise that is irrelevant to the detection, leading to a lower testing accuracy. This dropout layer will randomly shut off a fraction of neurons during training, thereby strengthening the model by making it less dependent on neuron weights. The output from the convolutional and pooling layers will be a multidimensional tensor; this is flattened into a one-dimensional array. The flattened data feeds into a dense layer with 1024 neurons, allowing the network to learn from the extensive feature set developed in the previous steps. A following dropout layer with a rate of 0.4 prevents overfitting and promotes generalization.

The final layer of the model contains four neurons. These represent the classes into which an individual can be classified: dysarthric male, non-dysarthric male, dysarthric female, and nondysarthric female. This layer has a softmax activation, providing the probability regarding every class in which it is located, where higher probabilities are more robust predictions for each category. It uses the Adam optimizer with a learning rate of 0.0005. The learning rate is a tuning parameter that aims to reach a minimum loss function. However, model training varies many hyperparameters to tune performance; similarly, evaluation on another test set confirms the model's accuracy, precision, and recall.



```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_3 (Conv2D)           (None, 20, 792, 64)       640

 batch_normalization_3 (Bat  (None, 20, 792, 64)       256
 chNormalization)

 max_pooling2d_3 (MaxPoolin  (None, 10, 396, 64)       0
 g2D)

 dropout_4 (Dropout)         (None, 10, 396, 64)       0

 conv2d_4 (Conv2D)           (None, 10, 396, 128)      73856

 batch_normalization_4 (Bat  (None, 10, 396, 128)      512
 chNormalization)

 max_pooling2d_4 (MaxPoolin  (None, 5, 198, 128)       0
 g2D)

 dropout_5 (Dropout)         (None, 5, 198, 128)       0

 conv2d_5 (Conv2D)           (None, 5, 198, 256)       295168

 batch_normalization_5 (Bat  (None, 5, 198, 256)       1024
 chNormalization)

 max_pooling2d_5 (MaxPoolin  (None, 2, 99, 256)        0
 g2D)

 dropout_6 (Dropout)         (None, 2, 99, 256)        0

 flatten_1 (Flatten)         (None, 50688)             0

 dense_2 (Dense)             (None, 1024)              51905536

 dropout_7 (Dropout)         (None, 1024)              0

 dense_3 (Dense)             (None, 4)                 4100

=================================================================
Total params: 52281092 (199.44 MB)
Trainable params: 52280196 (199.43 MB)
Non-trainable params: 896 (3.50 KB)
```

Fig. 3 The Convolutional Neural Network Model Architecture

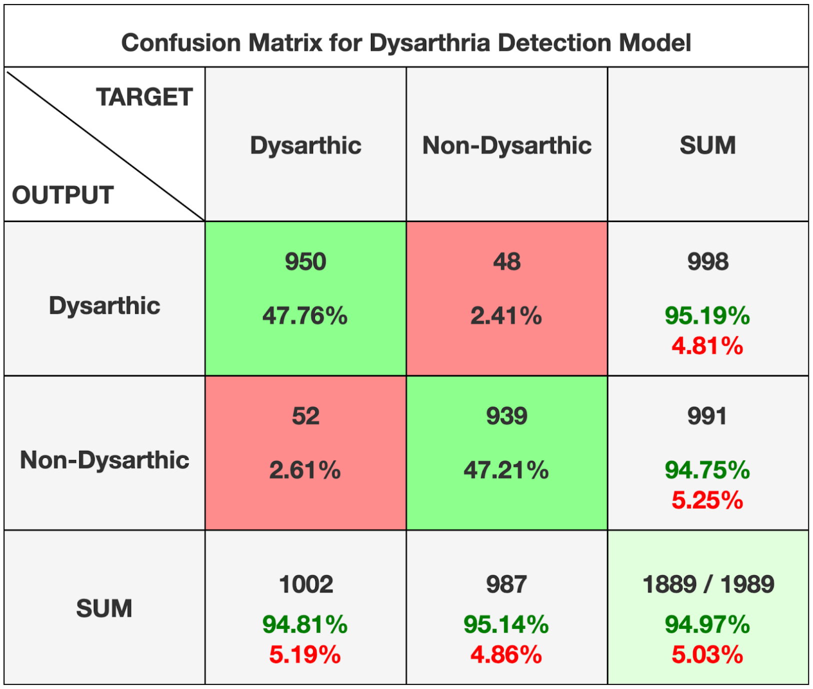


```
model.compile(
    optimizer=Adam(learning_rate=0.0005),
    loss='categorical_crossentropy',
    metrics=['accuracy', Precision(name='precision'), Recall(name='recall')]
)
```

Fig. 4 Parameters of the Convolutional Neural Network

Results

After about 20 epochs, the training accuracy began to halt from an exponential gain, transitioning to a linear increase. It can also be observed that the training loss at around 28 epochs began to hold a linear decrease. From epoch 28 and onward, the loss stopped its decrease and fluctuated between two and one-thousandths. The model utilized a call-back function to terminate the training process if there was no significant or notable improvement in the loss, effectively preventing the model from overfitting and ensuring good performance. Furthermore, after 53 epochs of training, the model yielded a training accuracy of 96.46% and a testing accuracy of 94.97%. This data shows that the CNN model effectively identifies the unique features within the voice data to classify a patient with dysarthria with an accuracy rate of 94.97%.



Confusion Matrix for Dysarthria Detection Model

TARGET OUTPUT	Dysarthic	Non-Dysarthic	SUM
Dysarthic	950 47.76%	48 2.41%	998 95.19% 4.81%
Non-Dysarthic	52 2.61%	939 47.21%	991 94.75% 5.25%
SUM	1002 94.81% 5.19%	987 95.14% 4.86%	1889 / 1989 94.97% 5.03%

Fig. 5 Confusion Matrix

Parameters	Test Accuracy	Test Precision	Test Recall
Score (%)	94.97%	95.28%	95.10%

Fig. 6 The Convolutionalconvolutional Neural Network's Performance Evaluation

As shown in the confusion matrix in Figure 5, out of the 1989 audio files used, 1889 were accurately predicted, while 100 were not accurately predicted. The testing accuracy closed the gap at 94.97% (Figure 6), proving the model's strong ability for generalization. For individuals with dysarthria, the model effectively evaluated 950 audio files correctly. On the other hand, it misclassified 52 audio files to be non-dysarthic. The model also accurately identified 939 non-dysarthic audio files but missed 48 audio files, classifying them as dysarthic. A possible cause for these errors was that the audio files used short phrases or words at times, making the speech challenging to classify.

Discussion

With continued research on this topic, the data should be based on a richer dataset with phrases that will get more diversified in their linguistic characteristics once a proper phonetic analysis has been established. Further steps in this direction will result in better improvements that can be transformed into increased robustness of models applicable to broad domains such as healthcare. It is further utilized to establish the groundwork of speech models for multi-irregularities in different languages and dialects that can allow diagnostic precision and intervention approaches. Embedding these machine learning models into real-time speech processing applications will revolutionize treatment approaches by giving immediate feedback and changes at the therapy session. Such models allow improvements at the clinical level and in patients' home-care systems and will be in this line of research. The earlier model observed had an accuracy as high as 93.97%, which**,** when compared to the 94.97% accuracy presented, is less accurate.

Conclusion

Deep learning technologies are making a paradigm shift in decision support systems in medical diagnostics, including the management of dysarthria. Our study on speech disorders, which complicate personal expression and inflict social and psychological challenges on the affected, was managed with a diagnostic accuracy of 94.97% using convolutional neural networks. Such precision thus refines therapeutic strategies and brings betterment to patient outcomes. Although the accuracy rate is high, some misclassifications suggest that model improvement and further exploration of other diagnostic features should be undertaken to assess the severity of dysarthria more accurately, guiding future research into increasing model diagnosis capabilities and extending clinical applicability.

References

1. The TORGO Database (2020) Acoustic and articulatory speech from speakers with dysarthria, https://www.cs.toronto.edu/compling-web/data/TORGO/torgo.html
2. Rudzicz, F., Namasivayam, A.K., Wolff, T. (2012) The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Language Resources and Evaluation, 46(4), pages 523--541.
3. Ambuj Mehrish, Navonil Majumder, Rishabh Bharadwaj, Rada Mihalcea, and Soujanya Poria. 2023. A review of deep learning techniques for speech processing. Inf. Fusion 99, C (Nov 2023). https://doi.org/10.1016/j.inffus.2023.101869
4. Lim, Jae S. *Two-Dimensional Signal and Image Processing*. Englewood Cliffs, NJ: Prentice Hall, 1990. pp. 536-540.
5. Jingdong Chen, J. Benesty, Yiteng Huang and S. Doclo, "New insights into the noise reduction Wiener filter," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 4, pp. 1218-1234, July 2006, doi: 10.1109/TSA.2005.860851.
keywords: {Noise reduction;Wiener filter;Speech enhancement;Acoustic distortion;Working environment noise;Acoustic noise;Signal to noise ratio;Degradation;Microphones;Automatic speech recognition;Microphone arrays;noise reduction;speech distortion;Wiener filter},
6. Jayaraman DK, Das JM. Dysarthria. [Updated 2023 Jun 5]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK592453/
7. Page AD, Yorkston KM. Communicative Participation in Dysarthria: Perspectives for Management. Brain Sci. 2022 Mar 22;12(4):420. doi: 10.3390/brainsci12040420. PMID: 35447952; PMCID: PMC9031517.
8. Schröter-Morasch H, Ziegler W. Rehabilitation of impaired speech function (dysarthria, dysglossia). GMS Curr Top Otorhinolaryngol Head Neck Surg. 2005;4:Doc15. Epub 2005 Sep 28. PMID: 22073063; PMCID: PMC3201013.
9. C. Clarke, Z. Abdali, S. Jowett, C. Rick, M. Brady, R. Woolley, C. Burton, S. Patel, P. Masterson-Algar, A. Nicoll, C. Smith, N. Ives, G. Beaton, S. Dickson, R. Ottridge, H. Nankervis, C. Sackley. Cost-effectiveness of Lee Silverman Voice Treatment LOUD (LSVT®) versus NHS Speech and Language Therapy versus control for dysarthria in Parkinson's disease: An economic evaluation alongside the PD COMM trial [abstract]. *Mov Disord.* 2023; 38 (suppl 1). https://www.mdsabstracts.org/abstract/cost-effectiveness-of-lee-silverman-voice-treatment-loud-lsvt-versus-nhs-speech-and-language-therapy-versus-control-for-dysarthria-in-parkinsons-disease-an-economic-evaluation-alongs/. Accessed August 12, 2024.
10. B. Vimal, M. Surya, Darshan, V. S. Sridhar and A. Ashok, "MFCC Based Audio Classification Using Machine Learning," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 1-4, doi: 10.1109/ICCCNT51525.2021.9579881. keywords: {Support vector machines;Training;Emotion recognition;Machine learning algorithms;Speech recognition;Feature extraction;Classification algorithms;RAVDESS dataset;Emotion recognition;Decission Tree;Support Vector Machine (SVM);Random Forest},
11. Rezapour Mashhadi MM, Osei-Bonsu K. Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. PLoS One. 2023 Nov 21;18(11):e0291500. doi: 10.1371/journal.pone.0291500. PMID: 37988352; PMCID: PMC10662716.
12. G. Hinton et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," in IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, Nov. 2012, doi: 10.1109/MSP.2012.2205597.

keywords: {Automatic speech recognition;Speech recognition;Hidden Markov models;Training;Gaussian processes;Acoustics;Neural networks;Data models},

13. Jingdong Chen, J. Benesty, Yiteng Huang and S. Doclo, "New insights into the noise reduction Wiener filter," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 4, pp. 1218-1234, July 2006, doi: 10.1109/TSA.2005.860851.

keywords: {Noise reduction;Wiener filter;Speech enhancement;Acoustic distortion;Working environment noise;Acoustic noise;Signal to noise ratio;Degradation;Microphones;Automatic speech recognition;Microphone arrays;noise reduction;speech distortion;Wiener filter},

14. Benesty, J., Chen, J., Huang, Y.(., Doclo, S. (2005). Study of the Wiener Filter for Noise Reduction. In: Speech Enhancement. Signals and Communication Technology. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-27489-8_2