2022 5th International Conference on Advances in Science and Technology (ICAST)

Hybrid CNN-LSTM network to detect Dysarthria using Mel-Frequency Cepstral Coefficients

Kush Vora
Department of Computer Engineering
K.J Somaiya College of Engineering
Mumbai, India
kush.v@somaiya.edu

Dishant Padalia
Department of Electronics and Telecommunication
K.J Somaiya College of Engineering
Mumbai, India
dishant.padalia@somaiya.edu

Darshil Mehta
Department of Computer Engineering
K.J Somaiya College of Engineering
Mumbai, India
darshil05@somaiya.edu

Deepak Sharma
Department of Computer Engineering
K.J Somaiya College of Engineering
Mumbai, India
deepaksharma@somaiya.edu

2022 5th International Conference on Advances in Science and Technology (ICAST) | 978-1-6654-9263-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAST55766.2022.10039514

***Abstract* — Dysarthria is a speech problem acquired at birth due to cerebral palsy (CP) or developed after severe brain damage. Dysarthria affects more than 70% of Parkinson's patients and 10% to 65% of people with traumatic brain injury. It is critical to detect dysarthria and other voice speech difficulties early to diagnose the underlying cause. Intelligent systems capable of identifying dysarthria with incredible precision have been developed using audio processing techniques and various deep learning models. This paper presents a hybrid CNN-LSTM model for classifying patients with dysarthria using audio recordings. The CNN-LSTM combination helps capture spatial and temporal information where CNN acts as a feature extractor while LSTM functions as a classifier. The proposed model was trained on the publicly available 9184 audio recordings from the TORGO dataset, and various audio augmentation techniques were employed to generate synthetic data. A total of 128 features were extracted using Mel Frequency Cepstral Coefficients (MFCC) and fed into the architecture as inputs. The K-fold cross-validation technique was used to avoid overfitting and increase the generalization capability of the model. The proposed architecture achieved a state-of-the-art 99.59% accuracy on the dataset. The presented work will minimize the workload of speech pathologists and help them detect dysarthria precisely and effectively.**

***Keywords* — Dysarthria, Audio Processing, Feature Extraction, Convolution Neural Network, Long Short Term Network**

I. INTRODUCTION

Speech is one of the most basic ways that humans interact with one another, and it is essential in everyday life. Like other markers of proper growth and development, speech develops gradually with age and time. These are circumstances in which a person has difficulty producing or generating the speech sounds required to communicate with others. These disorders can make the language unintelligible and incomprehensible. There are various speech impediments: stuttering, apraxia , dysarthria, and articulation error. These disorders can be caused by various factors, including brain damage from a head accident, damaged vocal cords, throat or mouth cancer, degenerative diseases such as Parkinson's disease, or weak muscles. Communication difficulties may affect 5% to 10% [1] of Americans, costing the country between $154 and 186 billion a year. Stuttering alone affects more than 3 million. Speech impairments are more common in children aged 2 to 7. Almost 7.7% of children in the United States have a speech, voice, language, or swallowing issue. Dysarthria is a form of speech disorder that occurs when a person struggles to control the muscles required for speaking (the brain and nervous system control these muscles). Dysarthria symptoms vary from person to person and are often determined by the underlying cause. Symptoms include sluggish or rapid speech, inability to talk loudly (can only whisper), difficulties moving face muscles or tongue, odd or uneven speech rhythm, and a monotonous voice. It can be either acquired during birth (cerebral palsy) or developed (the result of a brain tumor or severe brain damage). According to C. Mitchell et al. [2], 64% of the 88,974 stroke survivors reported communication problems, with 24% having dysarthria. Dysarthria is classified according to the part of the brain affected. One example is flaccid dysarthria, which is caused by cranial and spinal nerve injury. Other forms include spastic dysarthria (damage to the portion of the brain that controls movement) and ataxic dysarthria (damage to the connection between the cerebellum and other parts).To determine the cause, a speech-language pathologist (SLP) assesses the patient's speech. The neurologist then uses this cause to treat the underlying condition. The SLP examines the patient's lips, tongue, mouth, and breathing patterns when they pronounce words, phrases, and paragraphs. Electromyography is another technique used to diagnose Dysarthria (EMG). It is used to evaluate the condition of muscles as well as the nerve cells that govern them (motor neurons). Clinicians inject thin needles through the skin and into the muscles. When the patient moves the needles, electrodes on the ends measure muscle activity. Doctors use EMG to identify accidents, muscle illness, and neuromuscular disorders. It aims to expand the quantity of knowledge about the myofascial system that patients can use to improve their musculature and speech articulation.

Our research proposes a hybrid CNN-LSTM model that speech-language pathologists can adopt to mass screen patients with dysarthria effectively and with high precision. The proposed architecture takes advantage of both Convolution Neural Networks (CNN) and Long Short Term Memory Networks (LSTM) to diagnose dysarthric patients accurately. The CNN layers in the architecture serve as feature extractors, while the LSTM assists in learning long-term dependencies and serves as a classifier.

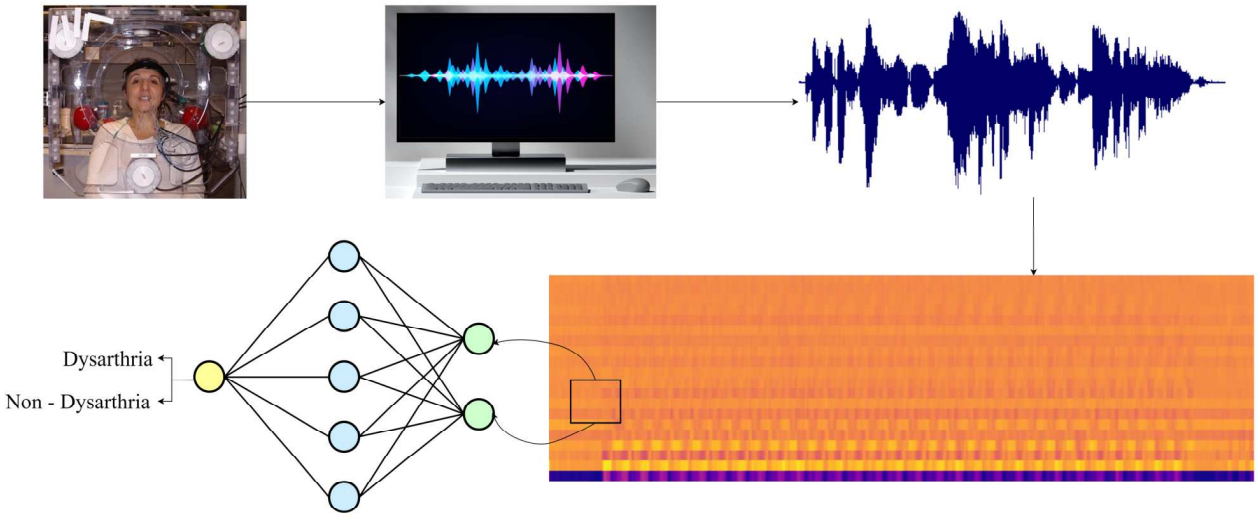

Fig. 1 Concept Diagram of the proposed system

The architecture consists of 25 layers, 3,267,841 parameters, and achieved a state-of-the-art 99.59% accuracy in detecting dysarthria in the TORGO dataset. Figure 1 represents the concept diagram of the proposed system where an electromagnetic articulography (EMA) system with two microphones is used to collect the movement data of the muscles and the time-aligned sound data when a person pronounces sentences and words displayed on a screen. Post capturing the audio, Mel-frequency cepstral coefficients (MFCC) are used to extract a total of 128 features from the audio recordings. These features are fed to the CNN-LSTM architecture to predict if the patient is dysarthric or not.

II. Literature Review

Previous research on this topic has used a variety of machine learning and deep learning techniques to detect dysarthria from audio samples. These studies have experimented with various datasets, model architectures, feature extraction methods, and pre-processing strategies. D. Korzekwa et al. [3] presented a deep learning architecture that detects dysarthria and reconstructs unintelligible speech. The mel-spectrogram of an audio signal is fed to the model as an input, and the model is trained to detect dysarthria and rebuild normal speech from dysarthric speech. The proposed model comprises two output networks that are trained together using a shared encoder. The audio and text encoders generate a low-dimensional dysarthric latent space and encode the input text sequentially. From a dysarthric latent space and encoded text, the audio decoder recreates the input mel-spectrogram, and finally, the logistic classification model predicts the chance of a dysarthric speech from the latent space.

S. Prakash [4] modifies D. Korzekwa's [3] CNN model by replacing the Gated Recurrent Unit (GRU) layer with a Long Short Term Memory layer (LSTM). Before feeding information to the model, the given work employs various audio processing techniques. Short-Time Fourier Transform (STFT), Mel-Filterbank, Spectrogram, and Mel-Spectrogram are the methods used. The audio signals were trimmed (or padded if too short) to a 5-second duration. The audio signals converted to the Mel-spectrogram achieved the highest accuracy of 68% S. Shahamiri [5] employed a point-to-point deep learning-based method to address the issues in dysarthria detection caused by phoneme alternation and inaccuracy, a paucity of dysarthric audio data, and phoneme labeling imprecision. UA-Speech dataset by the University of Illinois was used in this research. To address the scarcity of dysarthric data problems, they employed augmentation techniques to create more voice grams by methods like shifting the width of voice grams, sheering, and zooming through them. Additionally, normal-speech generation techniques were used to produce synthetic dysarthric samples. Using a deep 2D Spatial Convolutional Neural Network, this study created a Speech Vision (SV) model that learns to comprehend the structure of words produced by dysarthric people and recognizes individual words (S-CNN). Transfer learning was also used to improve the model's performance by learning from healthy speech visuals and transferring that comprehension to dysarthric speech data. The proposed SV model achieved an absolute average word recognition accuracy of 64.71%. Deep learning applications powered by transfer learning have achieved tremendous results in a number of domains. In transfer learning, a pre-trained network is used as a starting point for learning a new task. This typically speeds up and simplifies the network compared to starting from zero and training a network with randomly initialized weights. S R Sekhar et al. [6] leveraged this transfer learning approach to detect dysarthria. The TORGO dataset, which included 9094 speech samples of male and female participants with and without dysarthria, was used in this study. Furthermore, the audio clips were converted to a Mel-spectrogram to extract the relationship between amplitude, frequency, and time. The data was split in a 70:30 ratio for training to testing sets; additionally, 10% of the training set was used for validation. The ResNet-50 model was trained on the pre-processed data and achieved a test accuracy of 97.73%. Effective feature extraction from an audio file enhances

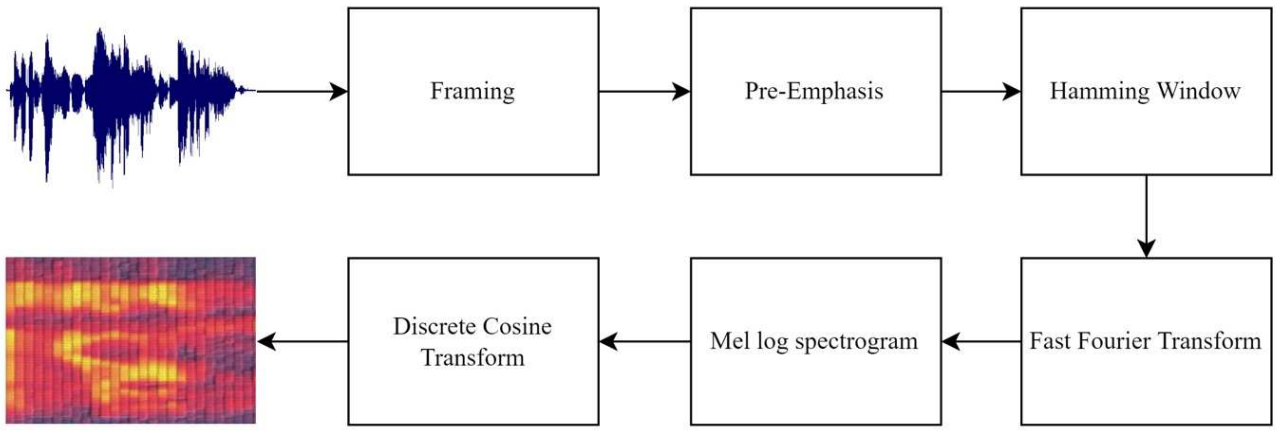

Fig. 2 Data preprocessing pipeline

the system's performance. Linear predictive analysis (LPC), linear predictive cepstral coefficients (LPCC), Mel-frequency cepstral coefficients (MFCC), Mel scale cepstral analysis (MEL), and perceptual linear predictive coefficients are a few of the notable extraction techniques used (PLP). Various studies in the field have employed and contrasted these feature extraction strategies to improve overall accuracy.

BN Suhas et al. [7] compared MFCC along with log Mel-spectrograms and concluded that spectrograms outperformed MFCCs in the multi-class classification of Amyotrophic Lateral Sclerosis (ALS), Parkinson's Disease (PD), and Healthy Control (HC). Better performance of log Mel-spectrograms was observed for both 3-class and 5-class ALS severity classifications. Following research performed classification on audio from 5 different sources. Furthermore, overlapping windows of different lengths between 0.8 to 3 seconds were fed to the CNN. The multi-class classification for ALS vs. PD vs. HC performed the best for window length equal to 1 second. AUC scores obtained after feeding spectrograms to the CNN were 94.6%, 98.8%, and 93% for ALS, HC, and PD, respectively. While the AUC scores obtained using MFCCs were 85.9%, 88.8%, and 77% for ALS, HC, and PD, respectively. A. Benba [8] proposed a method to detect patients who have Parkinson's Disease (PD) using MFCCs and SVM with different kernels. The acquired dataset contains 17 healthy and 17 samples of patients suffering from PD. Several features were extracted from MFCC in the range of 1 to 20 and were then passed through a Support Vector Machine (SVM). The study used SVM with different kernels, viz. linear, RBF, and polynomial. The first 12 coefficients of the MFCC by linear kernel SVM achieved an accuracy of 91.17%. Amula Anna Joshy et al. [9] conducted a thorough investigative study that included several deep learning architectures and audio processing approaches. The study also includes a second-level feature learning on i-vectors using Deep Neural Networks. Mel-frequency cepstral coefficients (MFCCs) had the lowest computational complexity of any classifier, while DNNs with MFCC and i-vectors had the highest accuracy. On the UA Speech dataset, the architecture achieved an accuracy of 93.97% in the speaker-dependent scenario and 49.22% in the speaker-independent scenario. P. Dumane et al. [10] have used several speech features such as zero crossing rates, MFCCs, spectral roll-offs, and spectral centroids to train a CNN model. The proposed system obtained an accuracy of 93.87% on the TORGO dataset.

III. Methodology and Material

A. Dataset

The TORGO [11] dataset contains aligned acoustics and articulatory characteristics recorded from patients that have cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS). The patients were made to read English texts from a screen. To capture movement and time-aligned sound data, a 3D AG500 electro-magnetic articulography (EMA) device with fully automated calibration was used. Two microphones were used to acquire all acoustic data. An Acoustic Magic Voice Tracker microphone was used to record audio at 44.1 kHz and was located 61 cm away from the patient. The second head-mounted microphone recorded audio at 22.1 kHz. Each participant with dysarthria was assessed by a speech-language pathologist who used standardized measures of speech-motor function. These tests were performed in accordance with the Frenchay Dysarthria Assessment (FDA) [12]. The TORGO dataset consists of 9184 audio recordings from one head-mounted and one directional microphone from seven individuals (four females and three males).

B. Data augmentation

Data augmentation helps to overcome the problem of data scarcity and helps the model generalize well by producing synthetic data. Noise injection, time shifting, pitch shifting, and increasing/decreasing the volume of an audio file are all processing techniques that can be used to augment an audio database. We used two strategies to expand the TORGO dataset for the proposed task. First, the audio files' speeds were changed by stretching the time series at a constant rate. The audio speed was increased and decreased by 2x and 0.5x, respectively. In addition, the audio files' loudness was boosted and decreased by 15 decibels.

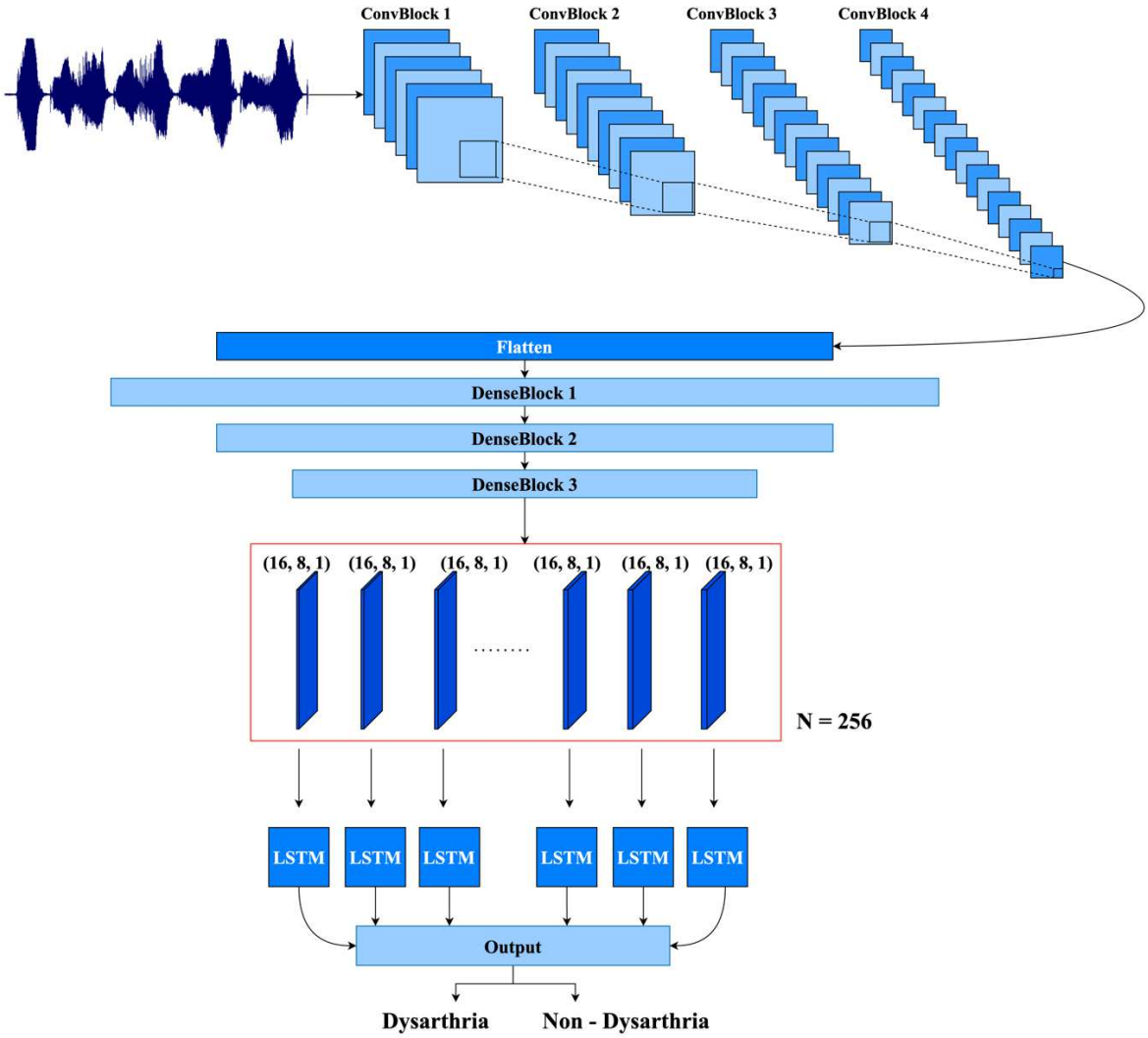

Fig. 3 Architecture diagram of the proposed CNN-LSTM network

C. *Feature Extraction*

For developing an efficient speech recognition system it is necessary to determine the components of the audio signal that are essential for detecting the linguistic content and discard other information like noise, emotion, etc. The proposed work uses Mel-frequency cepstral coefficients (MFCC) to extract features from dysarthric and non-dysarthric patients' audio files (.wav). Mel-Frequency cepstral coefficients was used to extract 128 features from audio files in the proposed work. A series of steps were performed to extract these features. The complete pipeline for extracting Mel-frequency cepstral coefficients from an audio file can be seen in Fig 2.

1) Framing:

A voice signal varies over time and is not static; thus, it is essential to divide these voice signals into small fragments. The voice fragments are usually in the 15ms to 40ms range since the voice signal is deemed stable in this range. A window length of 25ms was used in the presented work. These voice fragments overlapped each other after a specific distance less than the window length, leading to a total of N samples of the length of 25ms.

2) Pre-emphasis:

Pre-emphasis is a preprocessing step that is used to emphasize higher frequencies. It smoothes or balances out the instant steep transitions to higher frequencies.

3) Hamming window:

A Fourier transform converts audio fragments from the time domain to the frequency domain. Before conducting the Fourier transform on each window, the voice signals should be smooth and free of ripples. This is achieved by using a hamming window. Hamming windows improve harmonics, smoothen edges, reduce edge effects, and eliminate signal discontinuities.

4) Fast Fourier Transform:

A fast Fourier transform (FFT) computes the discrete Fourier transform (DFT) and inverse DFT of a sequence. It is used since it is computationally efficient and does not lose any relevant information.

5) Filter Bank Analysis:

Once the audio signal is converted from the time domain to the frequency domain, it gets converted to a Mel spectrum. Mel is a unit used to calculate the frequency the human ear perceives. The Mel frequency

scale has been observed to be linearly spaced below 1 kHz and logarithmically spaced above 1 kHz.

6) Discrete Cosine Transform:

Finally, the estimated log filter bank amplitudes are subjected to a discrete cosine transform. Because the previously estimated signal is symmetric along the y-axis, Discrete Cosine Transform (DCT) is utilized instead of Inverse Fast Fourier Transform (IFFT). While IFFT has a more complicated mathematical implementation, DCT performs the same purpose as IFFT but is more efficient because it takes advantage of signal redundancy performs the same purpose as IFFT but is more efficient because it takes advantage of signal redundancy

D. Proposed Architecture

Neural networks, a subtype of machine learning, are used in deep learning approaches. Convolutional neural networks (CNN) [13] differ significantly from other neural networks because they perform better with image, voice, or audio signal inputs. CNN architectures extract local features from high-layer inputs and then transfer these features to the lower layers in order to get more intricate features. Convnets have proven to be extremely effective in different domains, including image classification, object identification, and medical image analysis [14].

The proposed architecture was developed using a CNN and a LSTM network. LSTMs are a form of Recurrent Neural Network (RNN) capable of learning long-term dependencies. LSTM uses memory blocks rather than typical RNN [15] units to overcome the problem of vanishing and exploding gradients. The cell state is crucial for LSTMs. The LSTMs can delete/add information to these cell states using gates. Gates are a means of enabling information to flow through selectively. The LSTM comprises three different gates viz. forget gate, input gate and the output gate. The forget gate determines what should be discarded from the cell state during the first stage of the LSTM.

The subsequent step is to decide what should be stored in the cell state. This is realized in two stages: first, the sigmoid layer selects the values that need to be updated and then a tanh layer generates an array of values that can be added to the cell state. The old cell states are then updated to the new cell states. Finally, the output of the LSTM is calculated by the output gate. This is done by utilizing a sigmoid layer to determine which portions of the cell state are sent to the output. These are then multiplied by the tanh of the cell state to get the output.

This study employs a novel hybrid CNN-LSTM model to detect dysarthria automatically from audio samples. In this architecture, CNN extracts complex features from the audio sample, and LSTM is used as a classifier. The network has 25 layers of which 4 are convolutional layers, 4 are max-pooling layers whereas 3 are fully connected (FC) layers. Other layers in the architecture include normalization and dropout layers, along with one input, LSTM, and output layer. Each convolutional layer has ReLU [16] activation with padding as "same" so that the layer's output has the same shape as the input. There are four convolutional blocks, each comprising a convolutional, max pooling, and batch normalization layer. The output from the last ConvBlock is passed to a flatten layer to compress the multi-dimensional tensor to a single dimension. The output of the flatten layer is passed to three blocks of dense layers, each containing a fully connected batch normalization and dropout layer. The function map is forwarded to the LSTM layer with 256 units in the final section of the architecture to extract time information. Finally, the output from the LSTM layer is passed to the output layer with sigmoid activation to classify the input audio sample as dysarthric or non-dysarthric. Fig 3. shows the architecture diagram of the proposed CNN-LSTM network. In this study, Adam (adaptive moment estimation) optimizer (learning rate of 0.0001) was used to converge the loss to minima.

Aside from CNN-LSTM, the hybrid CNN-SVM model also performed admirably for this task. SVM [17] is a supervised machine learning (ML) technique for solving classification and regression problems. The SVM algorithm seeks a hyperplane in an N-dimensional space that classifies the input points definitively. In the CNN-SVM model, SVM is used as a binary classifier, replacing the sigmoid layer output layer. Instead of the typical binary cross entropy loss function, the CNN-SVM model employs Hinge Loss.

In this study, we also trained and tested other popular machine learning algorithms for this problem, such as the random forest [18], decision tree [19], and XGBoost [20]. The target class is predicted using a decision tree classification technique that learns simple decision rules from the dataset. Random forest is a more advanced variant of the decision tree algorithm; decision trees give all possible outcomes of a decision, whereas random forests show only the set of decisions that work. Another decision-tree-based ensemble technique that incorporates a gradient boosting framework is XGBoost.

IV. Results and Discussion

In the presented work, a number of ML and deep learning (DL) algorithms were trained and compared using three different ways.

A. Training on original dataset and testing on augmented data

The original TORGO dataset consists of 9,183 audio files randomly divided in an 80-20 ratio to produce the training and testing sets. Of the 9,183 files, 1,837 were used as the testing set, while the remaining 7,346 were used to train the architectures. The testing set was augmented to imitate real-life scenarios with the methodologies defined in data augmentation section. After audio augmentation, the testing set contained 3,674 files. Five distinct models were trained and tested: a hybrid of CNN and LSTM, CNN-SVM, Decision Tree, Random Forest, and XGBoost. The CNN-LSTM and CNN-SVM fared equally well on the testing set, with 98.31% and 98.39% accuracy, respectively. Decision tree calculations can be much more complex than other algorithms at times. As shown in Table 1, it achieves 100% accuracy on the train set, but because it is highly receptive to outliers, it fails to generalize on the unknown test set, achieving just 87% accuracy.

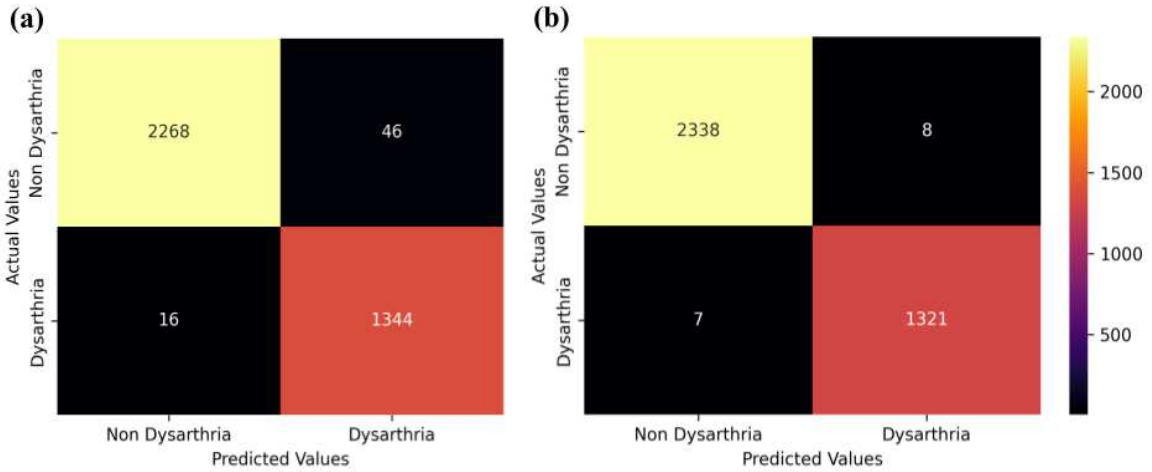

Fig. 4 Confusion matrices of the proposed model. (a) The model was trained on the original dataset and evaluated on the augmented data, (b) The model was trained and evaluated on the augmented data.

TABLE 1 PERFORMANCE OF MODELS TRAINED ON THE ORIGINAL DATASET

Model	Train Acc (in %)	Test Acc (in %)
CNN-LSTM	99.33	98.31
CNN-SVM	99.50	98.39
Random Forest	99.59	97.71
Decision Tree	100	87.53
XGBoost	99.48	97.90

B. Training and testing on augmented data

The training set was augmented along with the testing set using the procedures outlined in data augmentation section to improve the model performance. The enhanced training set acts as a regularizer, reducing overfitting when training a machine learning model. The TORGO dataset's 9183 audio files were augmented to 18366 files, separated in an 80-20 ratio (14692 files in the training set and 3674 files in the testing set). The CNN-LSTM outperformed the CNN-SVM with an accuracy of 99.59% when trained and tested using augmented data. Table 2 shows the performance of the five different models when trained on the augmented data. Figure 5 shows the two confusion matrices of the CNN-LSTM model when trained in two different settings – on the original dataset and on the augmented dataset.

TABLE 2 PERFORMANCE OF MODELS TRAINED ON THE AUGMENTED DATA.

Model	Train Acc (in %)	Test Acc (in %)
CNN-LSTM	**99.98**	**99.59**
CNN-SVM	99.82	99.53
Random Forest	99.71	97.76
Decision Tree	100	92.86
XGBoost	99.04	97.52

C. Training and testing on augmented data using K-fold cross validation

K-Fold cross validation is a technique for evaluating model competence and efficacy. With K-Fold validation every observation from the dataset gets a chance of appearing in the training and test sets, resulting in a less biased and less optimistic estimate of the model's performance than a conventional train-test split. In addition to training and testing the models on the augmented dataset, the models were also trained using the k-fold cross validation technique with a fold value of 5. The CNN-LSTM model achieved a mean test accuracy of 99.901% with a standard deviation of 0.00190. Table 3 presents the testing accuracy with standard deviation of the five models using k-fold cross validation.

TABLE 3 PERFORMANCE OF MODELS TRAINED USING K-FOLD

Model	Train Acc (in %)	Test Acc (in %)
CNN-LSTM	**99.97 ± 0.000374**	**99.90 ± 0.0017**
CNN-SVM	99.976 ± 0.000479	99.87 ± 0.00117
Random Forest	99.86 ± 0.000227	98.75 ± 0.00352
Decision Tree	1 ± 0	93.57 ± 0.00448
XGBoost	1 ± 0	99.17 ± 0.00151

Table 4 indicates that the proposed hybrid architecture outperforms other architectures in the literature. The comprehensive pre-processing pipeline aids in obtaining the best 128 features from the audio samples, and the highly fine-tuned CNN-LSTM yields better classification accuracy than the existing ConvNets and machine learning algorithms. Figure 6 represents a quantitative comparison of the five different models trained in this research work in three different settings – training on the original data, training on the augmented data, and training on the augmented data with 5-fold cross-validation.

TABLE 4 COMPARISON WITH OTHER ARCHITECTURES

Method	Accuracy (in %)
S R Sekhar et al. [6]	97.73
S. Prakash [4]	68
P. Dumane et al. [10]	93.87
Proposed work	**99.59**

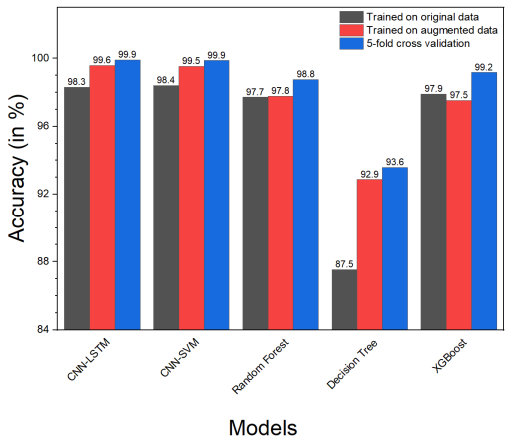

Fig. 5 A quantitative comparison of 5 different models trained in 3 different settings.

V. CONCLUSION

Dysarthria does not impair the level of intelligence or understanding, but a person with the disorder may experience difficulties during social interaction, work, and education. Dysarthria in children is typically developmental, and if not addressed promptly, it can progress to severe speech impairment. Furthermore, dysarthria often results from a major underlying condition, such as Parkinson's disease or acute brain damage, which might endanger a person's life. Hence, it is necessary to detect dysarthria as early as possible in order to begin the proper course of treatment. In the world, approximately one in every twelve children (7.7%) and 7.6% of adults are suspected of having a speech or language issue. A speech pathologist's duty is burdensome, with many patients suffering from diverse speech impairments. Moreover, it takes profound expertise and time to precisely diagnose each of these speech problems to avoid the wrong diagnosis. The presented study proposes a hybrid CNN-LSTM architecture trained on features extracted using the Mel cepstral coefficients (MFCC). The solution has achieved state-of-the-art results in this discipline and can be used by speech-language pathologists to detect dysarthria rapidly and correctly.

REFERENCES

[1] Ruben, R.J., 2000. Redefining the survival of the fittest: communication disorders in the 21st century. The Laryngoscope 110, 241–241.

[2] Mitchell, C., Gittins, M., Tyson, S., Vail, A., Conroy, P., Paley, L., Bowen, A., 2021. Prevalence of aphasia and dysarthria among inpatient stroke survivors: describing the population, therapy provision and outcomes on discharge. Aphasiology 35, 950–960.

[3] Korzekwa, D., Barra-Chicote, R., Kostek, B., Drugman, T., Lajszczak, M., 2019. Interpretable deep learning model for the detection and reconstruction of dysarthric speech. arXiv preprint arXiv:1907.04743.

[4] Prakash, S., 2020. Deep learning-based detection of dysarthric speech disability.

[5] Shahamiri, S.R., 2021. Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. IEEE Transactions on Neural Systems and Rehabilitation Engineering 29, 852–861.

[6] Sekhar, S.M., Kashyap, G., Bhansali, A., Singh, K., 2021. Dysarthricspeech detection using transfer learning with convolutional neural networks. ICT Express.

[7] Suhas, B. N., Mallela, J., Illa, A., Yamini, B. K., Atchayaram, N., Yadav, R., ... & Ghosh, P. K. (2020, July). Speech task based automatic classification of ALS and Parkinson's Disease and their severity using log Mel spectrograms. In 2020 international conference on signal processing and communications (SPCOM) (pp. 1-5). IEEE.

[8] Benba, A., Jilbab, A., Hammouch, A., Sandabad, S., 2015. Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson's disease, in: 2015 International conference on electrical and information technologies (ICEIT), pp. 300–304.

[9] Joshy, A.A., Rajan, R., 2022. Automated Dysarthria Severity Classification: A Study on Acoustic Features and Deep Learning Techniques. IEEE Transactions on Neural Systems and Rehabilitation Engineering 30, 1147–1157.

[10] Dumane, P., Hungund, B., Chavan, S., 2021. Dysarthria Detection Using Convolutional Neural Network.

[11] 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, 523–541.

[12] Enderby, P., 1980. Frenchay dysarthria assessment. British Journal of Disorders of Communication 15, 165–173.

[13] O'Shea, Keiron, and Ryan Nash. "An introduction to convolutional neural networks." arXiv preprint arXiv:1511.08458 (2015).

[14] Tajbakhsh, Nima, et al. "Convolutional neural networks for medical image analysis: Full training or fine tuning?." IEEE transactions on medical imaging 35.5 (2016): 1299-1312.

[15] Sherstinsky, Alex. "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network." Physica D: Nonlinear Phenomena 404 (2020): 132306.

[16] Agarap, Abien Fred. "Deep learning using rectified linear units (relu)." arXiv preprint arXiv:1803.08375 (2018).

[17] Hearst, Marti A., et al. "Support vector machines." IEEE Intelligent Systems and their applications 13.4 (1998): 18-28.

[18] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[19] Quinlan, J. Ross. "Induction of decision trees." Machine learning 1.1 (1986): 81-106.

[20] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.