

Compact CNN-Based Detection of Dysarthric Speech Using MFCC Spectrograms

Patina Tejaswara Rao¹, Karri Venugopal², Y Jagadeesh Kumar³
^{1,2,3}*Sri Sivani College of Engineering, Chilakapalem Jn, Srikakulam*

Abstract—Dysarthria is an apraxia of speech that significantly affects speech clarity and intelligibility, frequently resulting from neurological impairments. Accurate and early detection of dysarthric speech is essential for effective diagnosis and therapeutic intervention. This work proposes a Model based on compact convolutional neural network (CNN) for the autonomous detection of dysarthria using spectrogram representations of speech signals. Speech recordings of 8000 samples from UA speech dataset are pre-processed and converted into 2D spectrograms using Mel-frequency Cepstral Coefficients (MFCC) which seizure essential time-frequency attributes of the signal. Around 128 features have been extracted to train the Random Forest ML-based algorithm and obtained an accuracy of 93% with the test dataset. For further improvement, these spectrograms are applied to train the compact convolutional frame work with rmsprop optimizer in 4:1 ratio and obtained an accuracy of 96% and F1 score of 96% which is better when compared to other state-of the art techniques.

Index Terms—Dysarthria, Convolution Neural Network, Mel-frequency Cepstral Coefficients.

I. INTRODUCTION

Dysarthria is a neuro-motor speech disorder characterized by impaired articulation, resonance, phonation, and prosody, primarily neurological injury of motor-speech system. Individuals affected by dysarthria often exhibit slurred, slow, or effortful speech, which significantly hinders effective verbal communication. Early and precise detection of dysarthric speech is critical for enabling seasonable clinical participation and personalized therapy. Traditional machine learning (ML) approaches have been widely explored for this task, typically relying on hand-crafted attributes such as Mel-frequency cepstral coefficients (MFCCs), prosodic characteristic, or articulatory cues. These features are applied to various

Machine Learning (ML) classifiers or Gaussian Mixture Models (GMM) to distinguish dysarthric from healthy speech.

Advances in deep learning and Convolutional Neural Networks (CNNs) have appeared as a powerful option by mechanically learning stratified representations from raw or spectrogram-based speech inputs. CNNs are particularly effective in capturing local time-frequency patterns in spectrograms, enabling more robust and scalable classification. Unlike traditional ML methods that depend heavily on manual feature design, CNN-based approaches learn features directly from the data, resulting in improved generalization across speakers and speech conditions. Recent research has shown that CNNs, especially when combined with distance-based or attention mechanisms, can significantly outperform classical ML techniques in dysarthric speech identification tasks. The structure of Paper is as follows: section 2 had related work discussion, section 3 holds proposed methodology with appropriate block diagram with MFCC based feature extraction and spectrogram conversions and Compact CNN architecture, section 4 discussed about the Experimental results and section 5 gives over all conclusion on this work.

II. RELATED WORKS

During the past decade, numerous machine learning and deep learning techniques have been simulated to address various challenge of dysarthric speech classification. Early studies primarily focused on classical machine learning approaches that utilized hand-crafted features. Celin et al. in [1] introduced a two-level data augmentation method—Virtual Microphone Array synthesis coupled with Multi-Resolution Feature Extraction (VM-MRFE) is recognised with isolated-word dysarthric speech recognition; this method significantly lowered word

error rates (WER), especially for speakers with low and very low intelligibility.

Chandrashekar et al. in [2] explored various spectro-temporal representations (e.g., STFT, Mel spectrograms, time-frequency CNNs) for comprehensibility assessment of dysarthric speech, demonstrating that time-frequency CNN configurations outperform those focusing on time or frequency alone. Naeini et al. in [3] evaluated two augmentation strategies recordings of healthy individuals varying rate/loudness and synthetic speech with accented variations for fine-tuning ASR models; their approach matched or exceeded models fine-tuned on real dysarthric data, achieving a WER of ~5.7% and strong segmentation accuracy (IoU ≈ 89.2%).

Radha & Bansal et al. in [4] proposed a machine-driven detection and severity assessment system for dysarthria using raw speech input; although the abstract isn't directly accessible, the study likely focuses on end-to-end modeling for severity classification. Takashima et al. in [5] investigated cross-linguistic knowledge transferability for dysarthric speech recognition, showing that models trained on one language could effectively adapt to dysarthric speech in another language via transfer learning techniques.

Hsieh et al. in [6] presented a hierarchical curriculum learning framework utilizing multi-level knowledge distillation aimed at improving dysarthric speech recognition, though specific abstract details aren't available. He, Seng et al. In [7] has introduced a cooperative AI dysarthric speech acknowledgment system using data augmentation powered by a Generative Adversarial Network, presumably aiming to bolster recognition performance through synthetic data generation.

Chandrashekar et al. in [8] assessed perceptually enhanced STFT (e.g., mel-warped) and Constant-Q transform spectrograms using CNNs for both word- and sentence-level intelligibility assessment; Constant-Q and enhanced mel-STFT representations showed superior classification performance.

Vinotha et al. in [9] introduces a hybrid Transformer-CTC model trained on the UA-Speech corpus, achieving a robust Word Recognition Accuracy (WRA) of 89%, effectively aligning speech features with character sequences without requiring precise frame-level alignment. The approach demonstrates

strong transcription performance across varying dysarthria severity levels, offering a significant advancement for assistive speech recognition in dysarthric speakers.

Al-Ali et al. in [10] provides a systematic review exploring current AI techniques for classifying dysarthria severity, emphasizing the need for objective, consistent assessment tools beyond subjective clinical evaluations. The study analyzes effective acoustic features and machine learning methods, aiming to guide future developments in diagnostic and therapeutic tools. Bhat et al. in [11] likely proposes a BLSTM (Bidirectional Long Short-Term Memory) based method for assessing sentence-level intelligibility in dysarthric speech, implying that sequential modeling via BLSTM enhances the automation of intelligibility evaluation.

Yang et al. in [12] has used speech recognition methods tailored to dysarthric speech, focusing on the digit recognition task possibly exploring how specialized ASR systems can better recognize simple, constrained vocabulary commonly used in dysarthria research. Lee et al. in [13] has discussed integration of AI-based speech monitoring into consumer electronics to detect dysarthria, suggesting potential for ubiquitous, real-time dysarthria detection in everyday devices, thus enabling early intervention and increased accessibility.

Mani Sekhar et al. in [14] applied transfer learning alongside CNN models to detect dysarthric speech, leveraging pre-trained architectures adapted to the dysarthric domain. The approach likely improves detection accuracy while reducing training complexity, by capitalizing on learned representations from broader datasets.

This paper provides research gap in the existing models by providing compact CNN model to achieve better performance metrics compared to other state of the art techniques that are applied on the dysarthria data.

III. METHODOLOGY

The proposed methodology follows a basic structured format for dysarthria detection involving data collection, preprocessing, extracting features with Mel-frequency Cepstral Coefficients and converting to scalograms, model selection, training, performance

evaluation and result analysis as depicted in the Fig. 1. below. The figure clearly shows the different stages involved while detecting the dysarthric signal from the healthy one. The performance of the model is evaluated based on the 4:1 split of trained and test dataset.

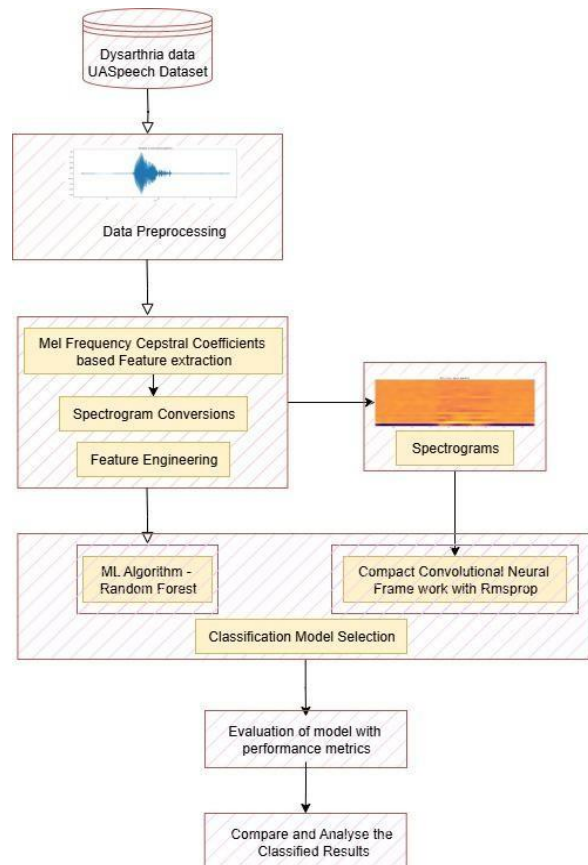


Fig. 1. Block diagram of the proposed methodology

3.1. Dataset Description

The UA Speech dataset [1] is a widely used resource in the field of auditory communication and processing, especially for research involving dysarthric speech recognition.

It was developed by the University of Illinois at Urbana-Champaign (UIUC) and is specifically modelled to assist the evolution of robust automatic speech recognition (ASR) systems for speakers with speech vitiate.

The dataset includes recordings from 15 dysarthric speakers and 13 control (healthy) speakers. The data is audio files of 16-bit, 16KHz .WAV format. These audio files are visualized as shown in the Fig. 2. given below that clearly plots the audio wave for both dysarthric and healthy signal.

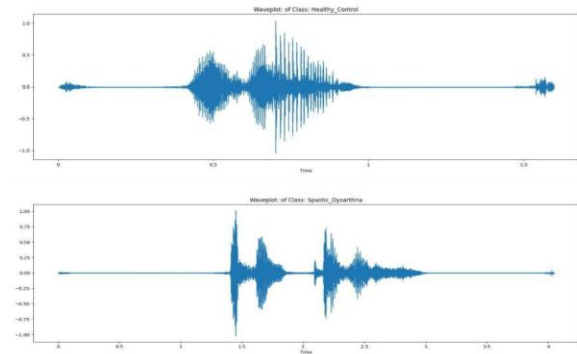


Fig. 2. Instance Waveplot of the Healthy and Dysarthric sample

3.2. Data Preprocessing

The critical step of Data pre-processing is to modify the UA Speech dataset suitable for training machine learning models and to assess these models. The primary goal is to improve the choice of the input data and increase the number samples for each class and make them balanced by sampling the existing data into small frames and ensure consistency across all samples, particularly when dealing with speech from dysarthric speakers, which often exhibits high variability and noise.

3.2.1. Data Sampling

The UA Speech dataset contains 15 samples of dysarthria (cerebral palsy) and around 13 samples of Healthy control which is not sufficient to train any ML or DL based model. To address this scenario a data sampling technique can be applied on the existing data that generates required number of subset of data samples for each class. Around 4000 subset samples were created for each class making the dataset balanced. Hence the entire dataset is of size 8000 samples of audio signals.

3.2.2. MFCC based Feature Extraction

Mel Frequency based feature extraction is a foundational step in processing speech signals for machine learning and classification tasks. The technique transforms the raw time-domain signal into a frequency-domain representation that captures the perceptual characteristics of human hearing. This is particularly useful in analyzing dysarthric speech, where subtle spectral distortions can affect intelligibility.

Using this feature extraction technique around 128

features has been extracted and are used to train the machine learning model. The dataset comprises of 8000 samples each with 128 features, hence the dataset is of size 8000 x 128. This dataset can be applied data cleaning and later on to machine learning models.

3.2.3. Conversion to Spectrograms

After Mel frequency-based feature extraction the obtained features are shown on 2D spectrograms to train the compact CNN frame work. These figures are saved individually for each sample of data, likewise around 8000 spectrograms are generated as shown in the Fig. 3. below.

The size of these spectrograms can be resized according to the network implemented.

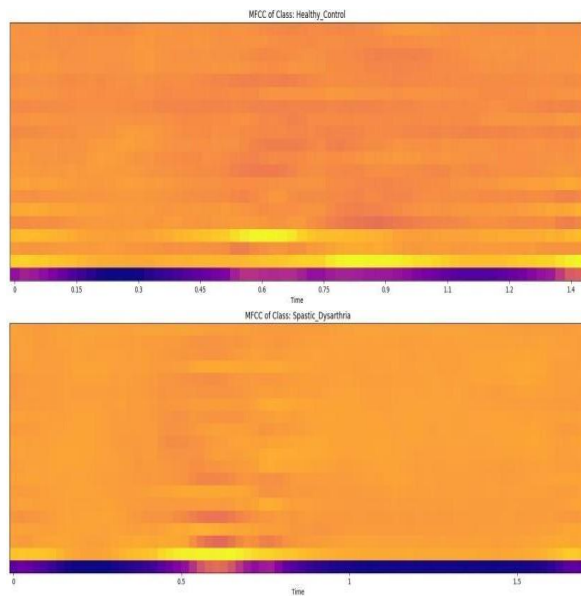


Fig. 3. Instance Spectrograms of the Healthy and Dysarthric samples

3.3. Compact CNN Architecture with Rmsprop Optimizer.

Convolutional Neural Networks (CNNs) are widely used in image classification, segmentation and object detection. The architecture is configured to adaptively learn spatial hierarchies of characteristics from input images. The sequential based convolutional neural network is implemented using the basic layers as shown below in the Fig. 4. While compiling the model rmsprop optimizer is used and the data is applied to this model to classify the samples between dysarthria and healthy.

Model: "sequential_1"

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 14, 6, 32)	320
batch_normalization_1 (BatchNormalization)	(None, 14, 6, 32)	128
max_pooling2d_1 (MaxPooling2D)	(None, 7, 3, 32)	0
dropout_2 (Dropout)	(None, 7, 3, 32)	0
flatten_1 (Flatten)	(None, 672)	0
dense_2 (Dense)	(None, 128)	86,144
dropout_3 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 1)	129

Total params: 86,721 (338.75 KB)

Trainable params: 86,657 (338.50 KB)

Non-trainable params: 64 (256.00 B)

Fig. 4. Layers used to implement the Compact-Convolutional Neural Network

3.4. Training and Validation

The entire data consists of 8000 samples of which 4000 were healthy control samples and 4000 were dysarthric samples. Of these 80 percent is applied to training as well validation together to the model and train_val curve of the model is given in the Fig. 5. below. Clearly the dataset is apt to train the model as loss is in exponential decay function.

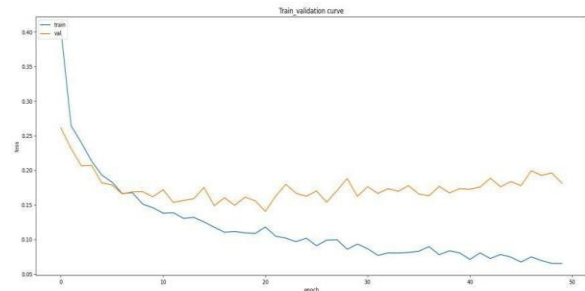


Fig. 5. Train and Validation curve when applied to the model

IV. EXPERIMENTAL RESULTS

In this study, the classification of dysarthric speech signals was carried out using MFCC spectrograms as input features. Two machine learning models were implemented and compared: a Random Forest (RF) classifier and a Compact Convolutional Neural Network (CNN) model. Their performance was quantitatively evaluated using various classification metrics in addition to confusion matrix analysis.

The Random Forest classifier was trained on spectrogram-based feature vectors extracted from dysarthric and control speech samples. The confusion

matrix given in Fig. 6 revealed that the model performed reasonably well in identifying healthy controls but exhibited misclassifications among dysarthria severity levels, particularly in borderline cases.

The overall accuracy and F1 score were 93%, with precision and recall values of 93.5% and 93.5%, respectively as shown in the Fig. 7. These results indicate that the RF model, while robust to feature variability, struggled to capture complex non-linear relationships inherent in dysarthric speech signals.

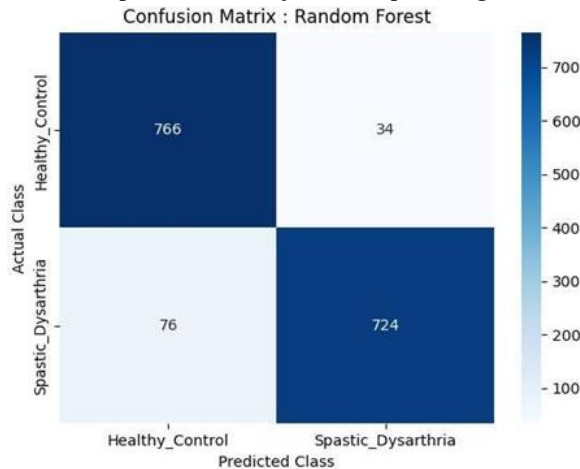


Fig. 6. Confusion Matrix of Random Forest

	precision	recall	f1-score	support
Healthy_Control	0.91	0.96	0.93	800
Spastic_Dysarthria	0.96	0.91	0.93	800
accuracy			0.93	1600
macro avg	0.93	0.93	0.93	1600
weighted avg	0.93	0.93	0.93	1600

Fig. 7. Classification Report obtained using Random Forest

To exploit the spatial and temporal dependencies within the MFCC spectrograms, a Compact CNN architecture was developed. Fig. 8. give a more balanced distribution of correct predictions across all classes with CNN when compared to the Random Forest model.

The CNN achieved a higher overall accuracy of 96%, with precision, recall, and F1-score values of 96% each respectively as given in the classification report in the Fig. 9. below. This demonstrates that deep feature extraction from spectrogram representations improve the model's ability to distinguish between subtle variations in dysarthric and control speech samples.

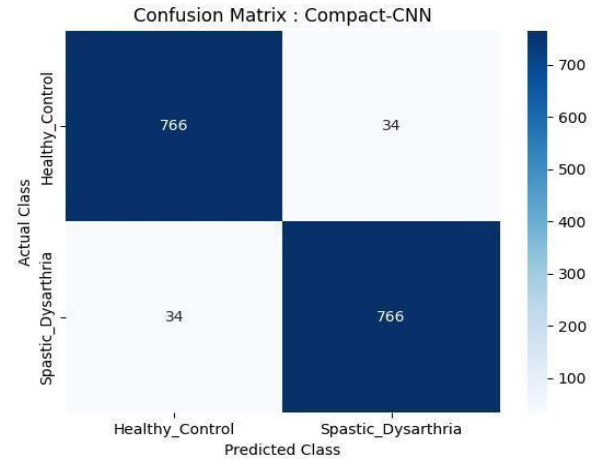


Fig. 8. Confusion Matrix using Compact-CNN

	precision	recall	f1-score	support
Healthy_Control	0.96	0.96	0.96	800
Spastic_Dysarthria	0.96	0.96	0.96	800
accuracy			0.96	1600
macro avg	0.96	0.96	0.96	1600
weighted avg	0.96	0.96	0.96	1600

Fig. 9. Classification Report obtained using Compact-CNN

Compact CNN model achieved a consistently higher ROC curve across all classes, yielding an average AUC of 0.9902. This suggests that the CNN effectively captures discriminative spectro-temporal features in MFCC spectrograms, enabling more reliable differentiation between healthy and dysarthric speech.

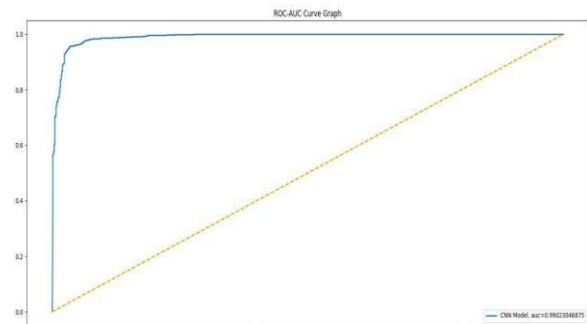


Fig. 10. AUC-ROC curve obtained using Compact-CNN

Comparative Analysis

Table. 1 summarizes the comparative performance of both models. The CNN outperformed the RF across all metrics, particularly in F1-score and AUC, highlighting its ability to exploit spectro-temporal features embedded in MFCC spectrograms. Figure Y

presents the ROC curves for both models, showing that the CNN curve is consistently above that of the RF model, indicating superior true-positive and false-positive trade-offs.

While the RF model remains computationally efficient, the Compact CNN provides a more reliable solution for dysarthric speech classification.

Table. 1. Comparison of the Performance metrics of two models

Performance Metrics	Random Forest	Compact-CNN
Accuracy	93%	96%
F1 score	93%	96%
Precision	93.5%	96%
Recall	93.5%	96%

V. CONCLUSIONS

In this work, an investigation on the classification of dysarthric speech using MFCC spectrogram representations with both a Random Forest classifier and a Compact CNN model has been carried out. The experimental results demonstrated that while the Random Forest achieved reasonable accuracy and computational efficiency, it struggled to reliably differentiate between closely related dysarthria severity levels.

In contrast, the Compact CNN consistently outperformed the Random Forest across multiple evaluation metrics, including accuracy, F1-score, and AUC-ROC, owing to its ability to automatically extract discriminative spectro-temporal features from spectrograms.

The ROC analysis further confirmed that the CNN model maintained superior trade-offs between sensitivity and specificity, highlighting its robustness in dysarthric speech recognition. These findings suggest that deep learning approaches, particularly lightweight CNN architectures, are better suited for practical deployment in dysarthria classification tasks. Future work could explore the integration of data augmentation, transfer learning from large-scale speech corpora, and attention-based architectures to further enhance recognition accuracy and generalization across different dysarthria severities and languages

REFERENCES

- [1] T. A. M. Celin, T. Nagarajan, and P. Vijayalakshmi, "Data augmentation using virtual microphone array synthesis and multi-resolution feature extraction for isolated word dysarthric speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 346–354, Feb. 2020, doi: 10.1109/JSTSP.2020.2972161.
- [2] H. M. Chandrashekar, V. Karjigi, and N. Sreedevi, "Spectro-temporal representation of speech for intelligibility assessment of dysarthria," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 390–399, Feb. 2020, doi: 10.1109/JSTSP.2019.2949912.
- [3] S. A. Naeini, L. Simmatis, D. Jafari, Y. Yunusova, and B. Taati, "Improving dysarthric speech segmentation with emulated and synthetic augmentation," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 12, pp. 382–389, 2024, doi: 10.1109/JTEHM.2024.3375323.
- [4] K. Radha and M. Bansal, "Automated detection and severity assessment of dysarthria using raw speech," in *Proc. 14th Int. Conf. Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India, 2023, pp. 1–7, doi: 10.1109/ICCCNT56998.2023.10307923.
- [5] Y. Takashima, R. Takashima, T. Takiguchi, and Y. Ariki, "Knowledge transferability between the speech data of persons with dysarthria speaking different languages for dysarthric speech recognition," *IEEE Access*, vol. 7, pp. 164320–164326, 2019, doi: 10.1109/ACCESS.2019.2951856.
- [6] I.-T. Hsieh and C.-H. Wu, "Hierarchical curriculum learning for dysarthric speech recognition via multi-level knowledge distillation," *IEEE Transactions on Audio, Speech, and Language Processing*, 2025, doi: 10.1109/TASLPRO.2025.3597438.
- [7] Y. He, K. P. Seng, and L.-M. Ang, "Collaborative AI dysarthric speech recognition system with data augmentation using generative adversarial neural network," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 33, pp. 2097–2111, 2025, doi: 10.1109/TNSRE.2025.3570383.

- [8] C. H. M., V. Karjigi, and N. Sreedevi, "Investigation of different time-frequency representations for intelligibility assessment of dysarthric speech," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 12, pp. 2880–2889, Dec. 2020, doi: 10.1109/TNSRE.2020.3035392.
- [9] R. Vinotha, D. Hepsiba, L. D. V. Anand, P. M. Bruntha, L. Dinh, and H. Dang, "Empowering dysarthric communication: Hybrid Transformer-CTC-based speech recognition system," *IEEE Access*, vol. 13, pp. 82479–82491, 2025, doi: 10.1109/ACCESS.2025.3568342.
- [10] A. Al-Ali et al., "The detection of dysarthria severity levels using AI models: A review," *IEEE Access*, vol. 12, pp. 48223–48238, 2024, doi: 10.1109/ACCESS.2024.3382574.
- [11] C. Bhat and H. Strik, "Automatic assessment of sentence-level dysarthria intelligibility using BLSTM," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 322–330, Feb. 2020, doi: 10.1109/JSTSP.2020.2967652.
- [12] J. Yang, H. Guo, X. Xu, and H. Bu, "A study of speech recognition techniques for dysarthria speeches based on digit recognition," in *Proc. 5th Int. Conf. Electronic Communication and Artificial Intelligence (ICECAI)*, Shenzhen, China, 2024, pp. 382–386, doi: 10.1109/ICECAI62591.2024.10675073.
- [13] S. Lee et al., "AI-enabled speech monitoring for dysarthria detection in consumer electronics," in *Proc. IEEE Int. Conf. Consumer Electronics (ICCE)*, Las Vegas, NV, USA, 2025, pp. 1–4, doi: 10.1109/ICCE63647.2025.10929844.
- [14] S. R. Mani Sekhar, G. Kashyap, A. Bhansali, A. Abishek A., and K. Singh, "Dysarthric-speech detection using transfer learning with convolutional neural networks," *ICT Express*, vol. 8, no. 1, pp. 61–64, 2022, doi: 10.1016/j.icte.2021.07.004.