Journal of Education for Pure Science
Vol.16, No.1 (2026)
DOI: https://doi.org/10.32792/jeps.v16i1.790
Website: jceps.utq.edu.iq Email: jceps@eps.utq.edu.iq

Speech Intelligibility Score Detection Based on Light Weight Deep Learning

Osama Ali Mohammed*, [1], Noor D. AL-Shakarchy [2]

[1]Dept. of Computer Science, Faculty of Computer Science and Information Technology, Kerbala University, Kerbala, Iraq.
[2]Dept. of Computer Science, Faculty of Computer Science and Information Technology, Kerbala University, Kerbala, Iraq.
* Corresponding email: osama.ali@s.uokerbala.edu.iq; noor.d@uokerbala.edu.iq

Received 10/ 9/2025, Accepted 29/ 10 /2025, Published 1 / 3 /2026

Abstract:

Speech intelligibility evaluation is essential for assessing speech clarity in individuals with speech impairments, as it aids in clinical diagnosis and immediate treatment planning. Since traditional evaluation techniques are frequently time-consuming, subjective, and unsuitable for regular or extensive use, trustworthy automated systems are needed. In this study, we present a lightweight deep learning system for automating the classification of audio intelligibility using two-dimensional Convolutional Neural Networks (2D CNNs). Mel-Frequency Cepstral Coefficients (MFCC) and log-mel spectrogram features are extracted from segmented speech samples and used in the model. The UA-Speech and TORGO benchmark datasets were used for the experiments. The model was trained independently on male and female speech in UA-Speech to investigate adaptability. A female-trained model was then fine-tuned on male data to apply transfer learning. To assess robustness, additional tests were conducted using mixed-gender speech from the TORGO dataset. The suggested model achieves good classification accuracy under all training settings, according to the results, with spectrogram-based models exhibiting strong discriminative capacity and MFCC-based models converging more quickly. The suggested strategy offers a good compromise between accuracy and computing economy, according to comparisons with current methods. All things considered, this work presents a feasible and scalable method for automated speech intelligibility evaluation, with potential advantages for both clinical and research applications. In experiments, the proposed lightweight 2D CNN demonstrated strong reliability and robustness across various levels of intelligibility, achieving an overall accuracy of 98.9%, macro precision of 98.85%, recall of 98.82%, and F1-score of 98.83%.

Keywords: Classification, CNN, Deep learning, Dysarthria, Speech disorders, Speech intelligibility, UA-Speech

1-Introduction

The phonatory, articulatory, and respiratory speech subsystems must be coordinated by dozens of muscles for speech to be produced. Although neurological diseases such as Parkinson's disease (PD) and amyotrophic lateral sclerosis (ALS) are the main causes of speech disorders like dysarthria, Stuttering, cleft palate and lip, hearing loss, and many other medical conditions can all impair one's ability to speak clearly. As a result, clinicians treating speech

disorders must conduct speech intelligibility assessments to assess the efficacy of any speech therapy or medication[1].

Speech intelligibility, or the ability to understand spoken language, is essential for effective communication. Accurately measuring and assessing speech intelligibility is crucial across a variety of fields, including forensic investigation, education, healthcare, and telecommunications. Speech intelligibility score detection is the process of evaluating a listener's speech comprehension skills; it is sometimes represented by a score indicating how intelligible the audio stream appears[2].

Determining the severity level, another term for speech intelligibility, is the first step in planning speech therapy or other interventions for a person with dysarthria. The bias in subjective, costly, and time-consuming perceptual tests can be mitigated by using an objective evaluation of dysarthria severity. Additionally, dysarthric speech ASR (Automatic Speech Recognition) performance has improved as a result of a better comprehension of severity[3].

Basic intelligibility tests are commonly used by speech-language pathologists in voice clinics to assess speech intelligibility. However, because pathologists have a thorough understanding of patients' speech problems, subjective intelligibility tests can be expensive, time-consuming, and occasionally prejudiced. The development of an objective method for evaluating dysarthric speech is required. Excellent speech intelligibility is more crucial than ever, given the growing reliance on voice-based technologies and communication systems. Among the elements that could have a major influence on intelligibility are background noise, speech difficulties, and transmission quality[4].

The categorization of dysarthric speech becomes an essential diagnostic tool to differentiate between people with dysarthria and those with normal speech patterns. Current deep learning algorithms can distinguish between normal and dysarthric speech [5]. Recent years have witnessed a dramatic shift in the assessment of speech intelligibility, largely driven by advances in artificial intelligence and deep learning. Advanced neural network models can automatically and accurately assess speech intelligibility. These advanced systems analyze the audio input based on a number of factors, including frequency, pitch, and articulation, to determine intelligibility scores [6].

When factors such as cost-effectiveness, system dependability, energy efficiency, response speed, and ease of maintenance are considered, lightweight architecture is a strong alternative for many real-time applications. In this sense, it functions particularly well in situations that call for quick decisions and constrained resources. Because lightweight architecture-based solutions consume less memory, CPU, and system resources, they are employed in situations with strict resource limits.

The main innovations and contributions of this study are summarized as follows:

1- For automatic speech intelligibility score recognition, a lightweight 2D Convolutional Neural Network (2D CNN) is recommended, offering high performance while remaining computationally affordable compared to traditional CNN architectures.

2- After a thorough comparison of MFCC and log-mel spectrogram features, the study concludes that spectrogram-based inputs exhibit stronger discriminative power, even though MFCC features enable faster convergence.

3- The adaptability and robustness of the model across male and female speech are evaluated using a cross-gender transfer learning approach, which has not been explored in most CNN-based intelligibility studies.

4- To confirm the suggested model's generalization across various speakers and speech disorders, it is assessed on two benchmark datasets (UA-Speech and TORGO).

5- According to the experimental results, the model outperforms traditional CNN baselines with 98.9% accuracy, 98.85% precision, 98.82% recall, and 98.83% F1-score, while maintaining a small, scalable design suitable for real-time or clinical applications.

2- Related Work

The methods reported in the current literature regarding speech intelligibility evaluation may be categorized into the following major groups[7]:

- Physiological measurements based on advanced equipment,
- Perceptual judgment by a skilled clinician,
- Machine-based speech intelligibility evaluation based on machine learning and signal processing.

The application of automated techniques for evaluating dysarthric speech intelligibility is the primary focus of the present study. Various approaches have been proposed by experimenting with different combinations of model architecture and feature selection in the literature on automated intelligibility assessments. Large-scale research utilizing over 550,000 dysarthric speech samples trained multiple deep learning models and evaluated their generalization across datasets, including TORGO and UA-Speech. The results showed high classification accuracy and robustness, reinforcing the effectiveness of CNN-based models for speech intelligibility classification[8].

The current state of the art offers some hope, but experimental approaches can be costly to adopt and difficult to scale. Furthermore, they commonly use the same speaker's data for both training and other purposes, which introduces biases such as assessor familiarity and differences in rating scales. Automated intelligibility testing could help with distant diagnosis, personalize patient treatment, and increase the accuracy and accessibility of dysarthria diagnoses. Generally, automated intelligibility assessment approaches are allocated into two classes: Aiming to construct a classification model devoid of any background information on normal speech, reference-free methods of evaluating intelligibility concentrate on extracting acoustic characteristics thought to be strongly associated with intelligibility, and reference-based methods that evaluate intelligibility using normal speech signals as a benchmark. These methods base their predictions of intelligibility on healthy speech data that the models use to learn what makes speech comprehensible[9]. Furthermore, cross-language intelligibility assessment still presents an overwhelming challenge due to linguistic variations in phonetics, prosody, and speech patterns. Recent studies propose AI-driven frameworks that incorporate both language-universal and language-specific speech characteristics to improve intelligibility classification across different languages[10].

For intelligibility classification tasks, Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs) have been commonly utilized. Consuming generalized linear models (GMMs) and support vector machines (SVMs) Kadi et al. [11] examined several methods for classifying dysarthria intelligibility. Their studies used the TORGO and Nemours databases, and the input characteristics were MFCCs and seven aural cues from an earlier 'ear-based' model. Although no cross-database validation was done, the study concentrated on a three-class intelligibility severity classification using sentence-level data from both datasets for training and testing. Based on the results, the GMM-based approach outperformed SVMs in the dysarthric speech intelligibility classification. Specifically, the GMM achieved a high accuracy of 93.2% with both MFCCs and aural signals, but the SVM only achieved 76.6% accuracy using MFCCs as inputs. However, the models' performance on unseen speakers and datasets was limited because no cross-database validation was performed. Furthermore, the study used only sentence-level data, thereby missing the range of continuous speech. Even though GMMs performed better than SVMs in this particular scenario, the results point to the need for more reliable techniques that can handle a variety of datasets, speakers, and speech conditions, a requirement that the current work seeks to address.

In[12], a comprehensive comparison of deep learning approaches against an SVM baseline classifier was conducted. The study evaluated dense neural networks, Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks for three-class intelligibility classification on the TORGO dataset and four-class classification on the UA-Speech dataset, using MFCCs as input features. Notably, the two datasets were treated independently, with separate models trained and tested for each, and no cross-database validation was performed. This restricts our ability to comprehend how well these models generalize to novel datasets or speakers who are not visible. On both datasets, the results showed that CNN and dense neural networks outperformed the SVM baseline regarding the precision of the categorization. Curiously, the dense network outperformed with four layers, but accuracy dropped as the number of layers increased. Compared with more conventional shallow models, deep learning models perform much better on intelligibility classification tasks, as shown in this study. Even while deep learning models perform noticeably better than shallow classifiers, the current study fills a gap in the literature by highlighting the need for more reliable methods that can generalize across a variety of datasets and speech situations.

The intelligibility classification tasks have also noticed a rise in the utilization of Convolutional Neural Networks (CNNs), which are famous for their success in picture classification. More recent research has investigated their

potential for application to direct video data and to image-based representations of speech signals (e.g., spectrograms). For example, the authors[13] proposed a method for intelligibility classification that utilizes both audio and video data as input features. A pre-trained face detector was used to detect facial features and landmarks in the video, and MFCCs were extracted from the audio. The outputs of several CNN networks that had already processed these features were used to train a fully connected classifier. We found that this method achieved a 99% success rate on the UA-Speech dataset. However, the dependence on video input restricts its use in situations with only audio or poor video quality. Concerns regarding generalizability to unknown speakers or recording situations are further raised by the study's failure to report cross-dataset validation. Scalability may also be hampered by multi-modal CNN systems' computational complexity. The current study fills the need for reliable, audio-only intelligibility classification models, as indicated by these restrictions.

In a recent study[14], the effectiveness of CNNs and a suggested ResNet model for classifying dysarthria severity into four classes was compared. The study focused on short-duration voice segments (less than a second) to show the value of spectrograms as input features. When trained on spectrograms, the CNN and ResNet models produced results that were similar; on the UA-Speech dataset, one-second speech segments yielded an accuracy rate of about 86.6%. However, the ResNet model outperformed the CNN, achieving 91.8% accuracy when tested on entire spoken utterances. The ResNet model reached an amazing accuracy of 98.9%. Furthermore, the segment-length-based performance variation indicates the need for models that can consistently handle full-length speech across a range of scenarios; our study closes that gap.

ASR-based methods are a key focus within reference-based approaches to assessing intelligibility. These techniques use the observation that dysarthric speech is difficult for automatic speech recognition (ASR) systems trained solely on healthy speech to process effectively, and that the more severe the dysarthria, the worse the system performs. Automated speech recognition (ASR) techniques use ASR outputs to determine the accuracy rates of words or alphabets to determine the intelligibility of speech. For instance, [7] evaluated the intelligibility of dysarthric speech using the Mozilla DeepSpeech voice-to-alphabet system. The study employed two metrics: the Levenstein distance metric, which measures the "cost" of translating the ASR output string to the ground truth, and the sequence matcher metric, which determines the longest matching subsequence. Additionally, by testing and comparing different subsets of utterances from the UA-Speech dataset, the study investigated the relationship between predicted intelligibility scores and word-level perceptual ratings. However, their evaluation on a single dataset and their reliance on ASR systems trained solely on healthy speech limit the generalizability of their conclusions. Furthermore, intelligibility at the sentence level could not be adequately captured by word-level metrics, underscoring the need for models that can consistently assess dysarthric speech across speakers, datasets, and speech scenarios.

3- Proposed System

The proposed model uses a lightweight 2D Convolutional Neural Network (CNN) to classify speech intelligibility. The ability of 2D CNNs to capture spatial correlations in time-frequency representations of speech, including spectrograms and MFCC features, ultimately contributed to their appeal. Because convolutional neural networks (CNNs) can recognize local patterns in speech inputs, they do remarkably well on tests of intelligibility categorization.

The research used speech data processed with log-mel spectrograms and MFCC features, two techniques that provide detailed representations of speech characteristics. These representations enable the 2D CNN model to learn spatial hierarchies, enabling accurate classification of intelligibility levels. Unlike conventional machine learning techniques, CNNs use convolutional layers to automatically extract relevant features, eliminating the need for handcrafted data. The research used speech data processed with log-mel spectrograms and MFCC features, two techniques that provide detailed representations of speech characteristics. These representations enable the 2D CNN model to learn spatial hierarchies, enabling accurate classification of intelligibility levels. Unlike conventional

machine learning techniques, CNNs use convolutional layers to automatically extract relevant features, eliminating the need for handcrafted data.

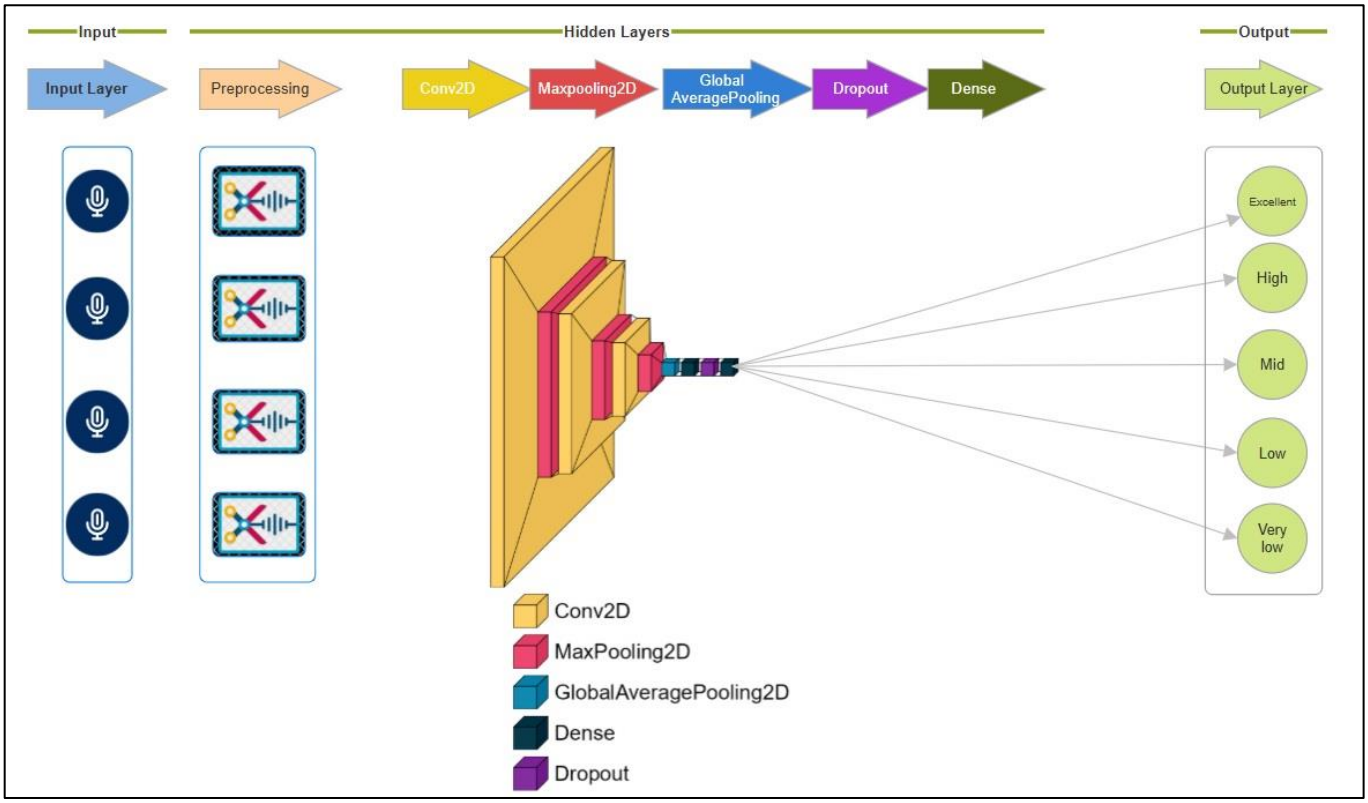

Fig. 1 - Model block diagram.

3.1 Dataset

UA-Speech and TORGO are two well-known datasets we used in our research on speech intelligibility. These datasets are ideal for evaluating our model's classification accuracy for intelligibility level because they contain speech recordings from people with varying degrees of speech impairment. The datasets were divided into 20% for testing and 80% for training. In UA-Speech, we performed classification into five classes based on intelligibility levels, while in TORGO, we categorised speech into four classes corresponding to different severity levels of speech impairment.

- UA-SPEECH: Included in this database are audio files and video clips of 19 individuals with cerebral palsy who have been diagnosed with dysarthric speech. Reading comprehension was assessed by having each participant complete three sets of 255 words. Each speaker contributes a total of 765 isolated words, with 455 of those being completely original. Ten digits, twenty-six letters of the radio alphabet, nineteen computer commands, and the one hundred widely used words in the Brown corpus of English (such as "it" and "is") make up the 155 common words blocks. On the other hand, to make the most of rare biphones, like "merchandise" and "naturalization," 100 unusual words were chosen for each chunk. Having five native English speakers in the United States try to transcribe a dysarthric speaker's speech was one way to gauge speech intelligibility. Perceptual intelligibility was measured as the percentage of correct answers. There are five groups that UA-Speech has identified for these as well: Very low (between 0% and 25%), low (between 26 and 50%), medium (between 51 and 75%), high (between 76 and 90%) and normal (between 90 and 100%) intelligibility [15].

- TORGO: Critical to studies examining dysarthric speech is the TORGO database, which is a product of a partnership between the University of Toronto, Ontario Federation for Cerebral Palsy, and Holland-Bloorview Kids, Rehab Hospital. This data set contains the housing information of 7 dysarthric speakers (F01, F03, F04, M01, M02, M03, M04, M05) as well as matched controls (FC01, FC02, FC03, MC01, MC02, MC03, MC04), where 'F' denotes a female and 'M' denotes a male. 'C' indicates control cases; We exclusively utilize the recordings made by the microphone array, which constitute a portion of the TORGO database. Using the intelligibility labels

in [a, b, c, d], we were able to generate 1200 utterances from 7 dysarthric speakers and 7 similarly situated controls. with "a" being the most understandable and "d" the least)[16].

Lastly, Table 1 displays the class-wise patient descriptions of speech intelligibility for the two datasets used in the model.

Table 1 - Class-wise speech Intelligibility patient description.

Intelligibility	UA-SPEECH	TORGO
Normal	The control speech used	-
High	F03, M02, M04, M12	The control speech used
Mid	F02, M07, M16	M02, M02, M04
Low	F04, M05, M11	F01, M05
Very low	F05, M08, M09, M10, M14	F03, F04, M03

3.2 Preprocessing

The segmentation allows consistent feature extraction and improves model performance by levelling input lengths. In addition to segmentation, normalization was employed to ensure uniform amplitude levels across all audio samples. Normalization makes the model more robust to variations in speech input by reducing the effect of adjustments to recording parameters and speaker volume. Normalizing each data set to a mean of 0 with 1 standard deviation was the first step in the feature extraction procedure. Quiet cutting was also used to eliminate the start and end gaps in each audio sample, ensuring that the recovered features are exclusive to the audio data. Each audio segment was compressed or padded to a preset duration of two seconds after equalization and quiet reduction. To ensure that all samples had the same feature dimensions, 13-coefficient MFCC (Mel-Frequency Cepstral Coefficients) features were also acquired. The recovered MFCCs were either zero-padded or truncated to a specified length to ensure the model's input size remained consistent. Additionally, log-mel spectrograms were computed, which capture the fundamental spectrum characteristics of speech sounds and offer a time-frequency representation. The Short-Time Fourier Transform (STFT) was used to create these spectrograms.

3.3 Feature extraction

The proposed model explored two types of features to represent speech intelligibility effectively: MFCC (Mel-Frequency Cepstral Coefficients) and Spectrograms.

- Spectrograms: Spectrograms based on the Mel scale's adjustment of the frequency axis in relation to human hearing are called Mel spectrograms, Mel-frequency spectrograms, or Mel-scaled spectrograms. A higher sensitivity of the human auditory system is associated with frequencies that are mapped on the Mel scale. Mel spectrogram construction uses the Fast Fourier Transform (FFT), along with other methods of calculating the power spectrum, starting with segmentation of the sound signal into frames. After that, the power spectrum is converted to the mel scale using triangular filter banks. Speech analysis tasks, such as dysarthria classification, benefit greatly from Mel spectrograms. They are in harmony with human perception and capture crucial spectral information, such as formant frequencies and spectral patterns. Mel spectrograms can be used as image-based features in dysarthria analysis or further processed to extract specific features. The dynamic range of an audio signal can be reduced while making its mel-frequency content easier to see and examine via log Mel spectrograms, which are produced by taking the logarithm of the Mel spectrogram's values [17].

- MFCC: For ASR systems, Mel Frequency Cepstral Coefficients (MFCCs) work well. Given their capacity to capture the spectral properties of speech signals, this could be seen as proof that they capture a large amount of the relevant information needed for speech understanding [18]. Consequently, they could be utilized in speech processing. We extracted 13 MFCC coefficients from each segmented audio sample and converted them into a 2D matrix representation suitable for CNN-based processing. Alterations to the vocal spectrum caused by dysarthria may explain why MFCC representations vary across severity levels. Speech therapists would greatly benefit from the development of an AI model capable of accurately distinguishing among these levels based on predetermined characteristics [19]. Compared to other features, MFCCs provide classifiers with the lowest

computational complexity [20]. The MFCC features were computed using a frame size of 1024 samples and a hop length of 256 samples, ensuring sufficient resolution for speech representation.

3.4 Methodology

Three convolutional stages and fully connected layers make up the suggested lightweight 2D CNN architecture. A max-pooling layer with a 2×2 pool size follows the first convolutional layer, which uses 32 filters with a 3×3 kernel. In the second level, there are 64 filters with another max-pooling operation, and in the third step, there are 128 filters with the same kernel and pooling setup. A Global Average Pooling (GAP) layer is used to maintain global information and lower the dimensionality of the feature maps. After that, the extracted features are regularized using an L2 penalty to avoid overfitting and run through a thick layer with 128 units and ReLU activation. A dropout layer with a rate of 0.5 comes next.

The classification probabilities are then provided by a SoftMax output layer with the same number of units as the target classes. It is appropriate for speech intelligibility evaluation tasks because of its small size, which guarantees a balance between representational capacity and computing efficiency. Using Adam optimizer and the categorical cross-entropy loss, the model was assembled. The architecture summary can be shown in Table 2. A lightweight model that balances classification performance and computational efficiency was a top priority in the architecture's design. The suggested network keeps a shallow depth and fewer parameters than deeper CNNs, which are frequently employed in computer vision tasks. This makes it better suited for real-time or resource-constrained settings.

The model can capture both low-level and high-level acoustic representations of speech by using gradually increasing convolutional filters (from 32 to 128). Additionally, the Global Average Pooling layer reduces overfitting and model complexity by eliminating the need for large fully connected layers. Additionally, the combination of L2 regularization and dropout offers increased resilience against variations in recording settings and speaker characteristics.

Table 2 - Summary of the proposed model architecture.

Layer (type)	Output shape	Criteria
Conv2d (32 filters, 3x3 kernel, ReLU activation)	(None, 126, 126, 32)	896
Max_pooling2d (2x2)	(None, 63, 63, 32)	0
Conv2d_1 (64 filters, 3x3 kernel, ReLU activation)	(None, 61, 61, 64)	18496
Max_pooling2d (2x2)	(None, 30, 30, 64)	0
Conv2d_2 (128 filters, 3x3 kernel, ReLU activation)	(None, 28, 28, 128)	73856
Max_pooling2d (2x2)	(None, 14, 14, 128)	0
Global_average_pooling2d(GlobalAveragePooling2D)	(None, 128)	0
Dense (Fully Connected Layer (128 neurons, ReLU, L2 regularization))	(None, 128)	16512
Dropout (0.5 for regularization)	(None, 128)	0
dense_1 (Dense)	(None, 5)	645
Total params: 110,405 Trainable params: 110,405 Non-trainable params: 0		

3.4 Training

The UA-Speech and TORGO benchmark datasets, split into 80% (80% training,20% validation) and 20% testing, were used for the experiments. The model was trained independently on male and female speech in UA-Speech to investigate adaptability. A female-trained model was then fine-tuned on male data to apply transfer learning. To assess robustness, additional tests were conducted using mixed-gender speech from the TORGO dataset**.** The dataset, including 46,418 spectrogram samples for training, 11,603 samples for validation, and 14,511 samples for testing spanning intelligibility classes was used to train the suggested lightweight 2D CNN model. The training procedure used up to 120 epochs, a batch size of 32, the Adam optimizer (with a starting learning rate of 0.001) was used to construct the model, and categorical cross-entropy loss was used for training. The best-performing model was maintained throughout training, and early halting with a patience of 10 was used to avoid overfitting. Additionally, a learning rate scheduler (ReduceLROnPlateau) was included, which decreased the learning rate by a factor of 0.5 when the validation loss plateaued for 5 consecutive epochs, with a minimum threshold of 1e-6. To address the

imbalance in intelligibility levels and prevent minority classes from being overshadowed during training, class weights were also implemented.

The NVIDIA GeForce GTX 1650 GPU (4 GB VRAM), Intel Core i9 processor,32 GB system RAM, TensorFlow with cuDNN acceleration, and a local workstation running Windows were used for all the experiments. TensorFlow/Keras, the primary deep learning framework, and Python 3.10 were installed in the software environment. Depending on the dataset size and feature type (spectrograms vs. MFCC), the average training time per trial ranged from 60 to 200 minutes.

The learning curves in Figure 2, which depict the model's performance over time during the training phase, provide additional evidence of the predicted model's effectiveness. By showing whether the model is underfitting, overfitting, or converging optimally, these curves aid in diagnosing training dynamics. As shown in Fig. 2, the loss and accuracy curves stabilize as training progresses, confirming that the model effectively minimizes the loss function. This behavior translates into strong classification performance.

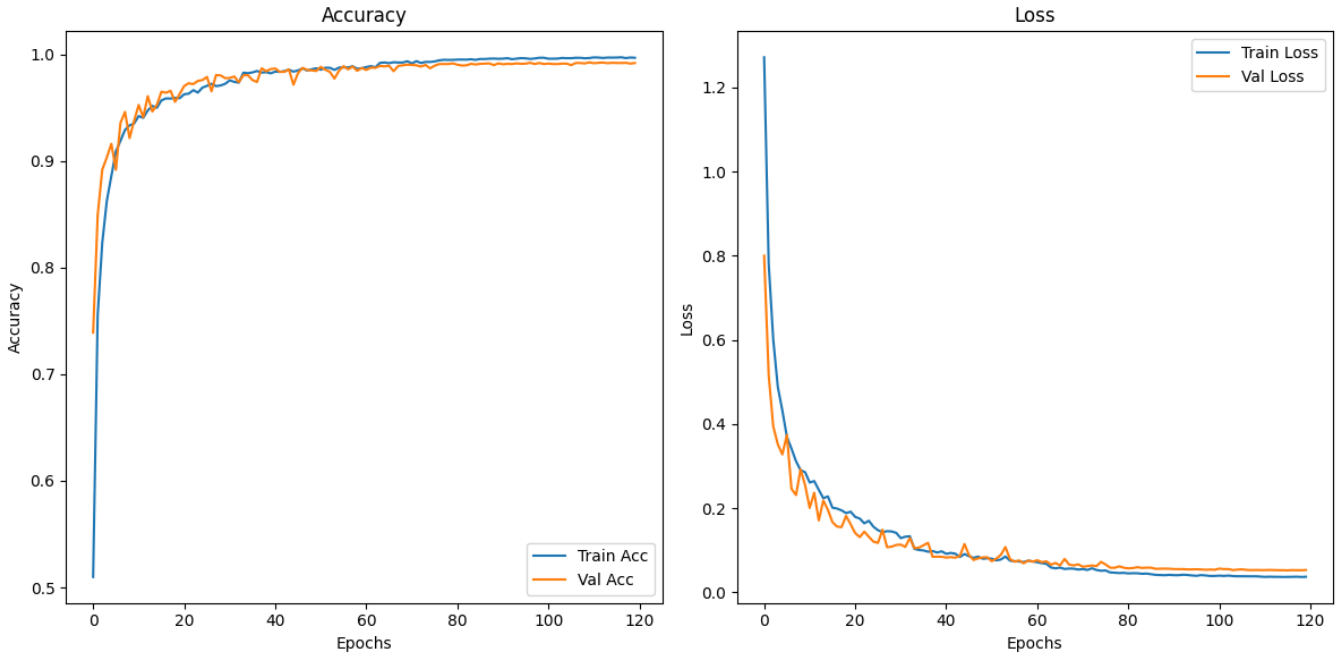

Fig. 2 - The model training behavior.

4- Results

The experimental results for the proposed lightweight CNN model are presented in Table 3, which also shows the classification accuracy across different datasets, speaker genders, and acoustic feature categories. The model's greatest performance on the UA-Speech dataset was 99.1% when trained on female speech using spectrogram features; the equivalent accuracies for male and mixed-gender spectrogram inputs were 98.8% and 98.6%, respectively. When MFCC characteristics were included, the model maintained comparable high accuracies of 98.9% for female speech, 98.5% for male speech, and 98.7% for mixed-gender speech. On the TORGO dataset, the model likewise demonstrated strong generalization, with 98.1% accuracy with spectrogram features and 98.4% accuracy with MFCC features. These results demonstrate that both spectrogram and MFCC characteristics provide reliable representations for intelligibility classification, with spectrograms providing UA-Speech with a slight advantage for female speech. Overall, the findings demonstrate that the proposed CNN design performs well across a variety of datasets and experimental setups.

Table 3 - The model's performance metrics.

Dataset	Speaker gender	Features	Classification accuracy
UA-SPEECH	Female	spectrogram	99.1%
	Male		98.8%
	Mix		98.9%
	Female	MFCC	98.6%
	Male		98.5%
	Mix		98.7%
TORGO	Mix	spectrogram	98.1%
		MFCC	98.4%

The classification performance of the suggested CNN model across the five intelligibility levels, excellent, high, low, mid, and very low, is depicted in the confusion matrix in Figure 3 for the UA-Speech dataset utilizing spectrogram features (female speaker). Similar patterns were seen in other experimental situations, which is why this experiment was selected as representative. With most samples properly classified along the diagonal, the model has great discriminative power. For example, precise identification was achieved for 4,164 samples in the excellent class, 1,864 samples in the high class, 2,712 samples in the low class, 2,634 samples in the mid class, and 2,985 samples in the very low class. There were extremely few misclassifications, mostly between nearby intelligibility levels (for example, high being incorrectly classed as low or very low, and vice versa). This tendency reflects the practical difficulty of differentiating borderline cases in speech and is consistent with the natural perceptual similarities between adjacent groups.

This tendency reflects the practical difficulty of differentiating borderline situations in speech intelligibility testing and is consistent with the intrinsic perceptual similarities between adjacent classes. The model's resilience in multi-class classification tasks is highlighted by the confusion matrix, which shows that it performs extremely reliably with little confusion among non-adjacent categories.

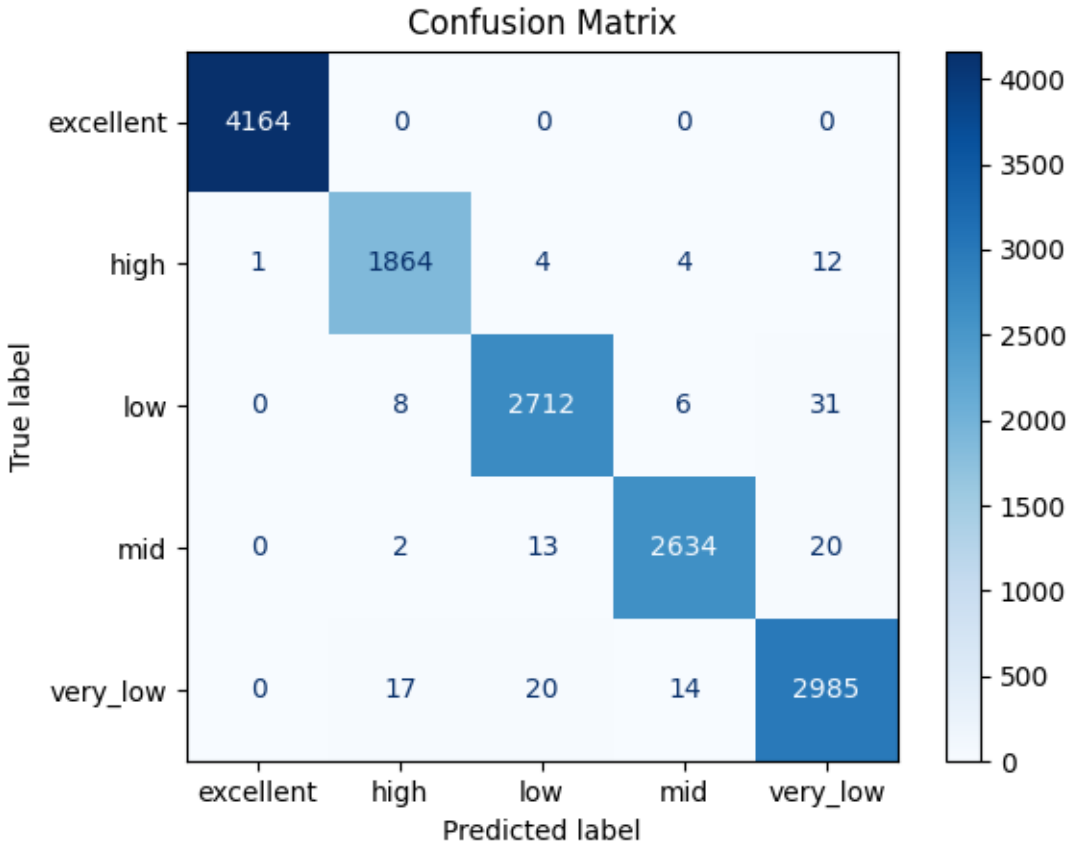

Fig. 3 - Five Level Speech Intelligibility Classification Confusion Matrix.

In order to determine how effective the studied model is, its performance was compared with related works that utilized the same datasets, UA-Speech and TORGO, for speech intelligibility classification. Both deep learning models, such as those based on Transformer architectures and Long Short-Term Memory (LSTM) networks, and more conventional machine learning techniques, such as Support Vector Machines (SVM) and Random Forest, have been used in a number of earlier investigations. Although the success of various techniques varied, several of them were computationally costly or required a large amount of handmade feature extraction. In addition to overall accuracy, the model's performance was evaluated using other metrics. Precision, recall, and F1-score were computed for each intelligibility class to provide a more thorough assessment of the model's performance across majority and minority classes. The method showed that it could correctly identify fewer common classes while still doing well overall, removing any potential biases brought on by unbalanced datasets. It had a 98.83% F1 score, a 98.85% precision, and a 98.82% recall.

While maintaining computing efficiency, the proposed lightweight 2D CNN model achieved performance comparable to or better than the competition using spectrograms and MFCC features. Unlike feature-engineering-dependent models, our CNN-based approach automatically derives spatial features from time-frequency representations, leading to robust classification with less training time. Furthermore, the incorporation of transfer learning in UA-Speech and the mixed-gender training methodology in TORGO improved generalization.

The performance of our suggested model and related research is thoroughly compared in Table 4, which also illustrates the benefits of our approach in terms of accuracy, feature extraction method, and computational efficiency.

Table 4 - Proposed model and some related works comparison.

Ref.	Dataset	Model	Features	Classification accuracy
[11]	Nemours and TORGO	GMM SVM GMM and SVM	MFCC and additional auditory modelling	93.2 % 75.3 % 78.8%
[12]	UA-Speech	SVM DNN CNN LSTM	MFCC	82.9% 93.6% 93.2% 75.1%
	TORGO	SVM DNN CNN LSTM	MFCC	82.7% 95.1% 96.2% 85.9%
[13]	UA-Speech	CNN	MFCC and recorded a video of the speaker	99.6%
[14]	UA-Speech	ResNet	Spectrogram	98.9%
[7]	UA-Speech	DeepSpeech ASR	Audio into a pretrained ASR model	98.0%
Proposed model	UA-Speech	CNN	MFCC Spectrogram	98.7% 98.9%
	TORGO		MFCC Spectrogram	98.4% 98.1%

The results from the mixed-gender subsets of the UA-Speech and TORGO datasets were selected for comparison with relevant studies for performance benchmarking since this scenario provides a more thorough evaluation across speaker variations. While our proposed model's accuracy was competitive, Ref. [13]'s results were slightly better. The utilization of multi-modal inputs, namely the combination of recorded speaker video and MFCC acoustic features, which offer complementary visual signals for intelligibility assessment, can help to explain this disparity. Nevertheless, adding video streams is computationally expensive and requires significantly more processing power, storage, and synchronization between modalities. However, the proposed method is lightweight and suitable for real-time or resource-constrained applications since it just employs acoustic information. Furthermore, the observed classification accuracy difference between our model and Ref. [13] is insignificant, demonstrating the effectiveness of our approach considering its lower computational cost.

To determine the importance of design components in the model's exceptional performance, ablation study was carried out. Network depth, preprocessing methods (such normalization and silent removal), and input feature selection (spectrograms vs. MFCCs) were among the components that were altered or removed in order to carry out the investigations. Even while MFCCs provided faster convergence but marginally worse accuracy, the results demonstrated that spectrogram-based inputs significantly enhanced performance. The UA-Speech and TORGO datasets demonstrated the need of careful preprocessing and architecture decisions in attaining reliable performance. Preprocessing approaches, in particular silence removal and normalization, enhanced model stability and generalization. These results suggest that the high accuracy attained was a result of the model design and preprocessing method working together rather than just the dataset's properties.

5- Discussion

The proposed model is lightweight and maintains competitive accuracy compared to conventional deep learning frameworks because of its straightforward architectural design and fewer parameters. This architecture choice ensures decreased memory utilization, quicker training and inference, and suitable for deployment in real-time or resource-constrained applications. The simplified model achieves comparable accuracy to traditional CNN and ResNet architectures with a fraction of the necessary parameters, demonstrating successful feature extraction without

sacrificing efficiency. Oversimplification may somewhat hinder performance on very complex speech patterns, but this design choice likely reduces overfitting, particularly in datasets with limited examples of abnormal speech.

Regarding feature representations, two distinct approaches spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs) were investigated independently. Spectrophotograms provide a thorough two-dimensional time-frequency description of audio signals. Each utterance typically contains thousands of characteristics, and both fine-grained temporal and spectral patterns are recorded. In contrast, MFCCs are compact descriptors (usually 13–40 coefficients per frame) that highlight perceptually relevant spectrum features while drastically reducing dimensionality to mimic human aural perception. The study demonstrated the benefits of spectrograms' high spectral richness as well as the effectiveness and resilience of MFCCs by assessing the model on each representation independently.

Additionally, research showed that spectrogram-based models had somewhat greater peak performance by utilizing their wider feature space, although models trained using MFCC features often converged faster during training because to their decreased dimensionality and compact representation. Because the wider feature space captures fine-grained spectral and temporal variations that are essential for abnormal speech classification, models trained on spectrograms performed better than MFCC-based models. In scenarios where rapid model retraining or deployment on devices with limited computational resources is necessary, MFCC-based models converged more quickly, but at the cost of somewhat lower peak accuracy.

Gender-related differences were also seen across experiments. On average, female voices were slightly more accurate than male ones. Due to physiological and auditory factors, such as the greater fundamental frequency (F0) and unique formant spacing of female speech, this may enhance feature separability. However, both male and female voices were included to ensure the model's robustness to speaker heterogeneity. Because mixed-gender subgroup results are commonly employed in related studies and suggest larger generality, they were highlighted for fair comparability with prior work. Because to differences in fundamental frequency and formant spacing, female voices achieved 99.1% accuracy while male voices achieved 98.8%. This is in line with past studies showing gender-based feature separability.

By using weighted loss function strategies to address class imbalance, minority classes were successfully represented throughout training. Dropout layers, early stopping based on validation results, and regularization techniques were used to prevent overfitting, and the depth and capacity of the model were carefully adjusted to match the size of the dataset to prevent underfitting. Even with weighted loss functions techniques, the rarest intelligibility classes sometimes displayed noticeably worse accuracy, suggesting that future research should look into oversampling or synthetic data augmentation to improve performance even more. Regularization techniques like dropout and early ending stabilized training and reduced overfitting, as seen by the minimal difference between training and validation accuracies.

Future studies will examine self-supervised pre-trained speech models for more thorough feature extraction and use visual cues such lip movements to broaden the existing approach for assessing multi-modal intelligibility. Tests must also be extended to incorporate a greater variety of clinical datasets and real-world scenarios in order to validate the model's robustness and applicability for usage in actual speech rehabilitation systems. The categorization of intelligibility may be enhanced by visual cues like lip movements, which record extra articulatory information not present in auditory signals alone. Self-supervised pre-trained models could provide richer, more thorough feature representations, making them more robust to a variety of speakers and datasets. The requirement for huge multi-modal datasets and guaranteeing real-time implementation in clinical contexts are two obstacles that still exist, nevertheless. The current analysis focused on audio-only features, which may not capture all articulatory cues relevant in multi-modal settings; a significant class imbalance, though reduced, may still affect performance on rare intelligibility levels; and the study was primarily evaluated on a small number of datasets, which may limit generalizability to other dysarthric populations. Comprehending these limits aids in contextualizing the findings and guiding following investigations.

6- Conclusion

This study introduces a two-dimensional CNN to build a deep learning system that classifies speech clarity at multiple levels, balancing computational efficiency with a lightweight architecture. In addition to being appropriate for real-world applications, this lightweight design achieved high classification accuracy while maintaining low computational cost, using spectrograms and MFCCs as network inputs. Two datasets, UASPEECH and TORGO, were used for the experiments, and speakers of both sexes at varying intelligibility levels demonstrated the model's resilience. Because spectrograms have a richer deep spectrum, they showed a little bit more accuracy in experiments. However, because of its compact feature representation, training with MFCC inputs to the network showed faster convergence, demonstrating its potential for real-time applications. By eliminating the need for manually created features and adopting a deep learning solution that balances accuracy, efficiency, and minimal resource consumption, the suggested approach outperforms many conventional techniques. To enhance performance in clinical diagnostics, assistive communication technology, and therapeutic tools, future research will focus on expanding the quantity and variety of speech datasets.

7- Data availability

Two databases were used in this research:

- **UA-Speech** is only available to researchers employed by government and academic institutions. Applications for commercial use are prohibited. Data must be protected from theft and cannot be redistributed. For a copy of the dataset, contact **uaspeech-request@lists.illinois.edu** via email.
- **The TORGO** dataset can be accessed at:

  https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html

Conflicts Of Interest

The authors claim that there is no conflict of interest.

References

[1] **S. Venugopalan , Joel Shor, Manoj Plakal, Jimmy Tobin, Katrin Tomanek, Jordan R. Green, Michael P. Brenner**, "Comparing Supervised Models And Learned Speech Representations For Classifying Intelligibility Of Disordered Speech On Selected Phrases," July 08, 2021, *arXiv*: arXiv:2107.03985. doi: 10.48550/arXiv.2107.03985.

[2] **J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan**, "Automatic intelligibility classification of sentence-level pathological speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 132–144, Jan. 2015, doi: 10.1016/j.csl.2014.02.001.

[3] **C. Bhat and H. Strik**, "Speech Technology for Automatic Recognition and Assessment of Dysarthric Speech: An Overview," *J Speech Lang Hear Res*, vol. 68, no. 2, pp. 547–577, Feb. 2025, doi: 10.1044/2024_JSLHR-23-00740.

[4] **F. Javanmardi, S. R. Kadiri, and P. Alku**, "Pre-trained models for detection and severity level classification of dysarthria from speech," *Speech Communication*, vol. 158, p. 103047, Mar. 2024, doi: 10.1016/j.specom.2024.103047.

[5] **S. M. Shabber and E. P. Sumesh**, "AFM signal model for dysarthric speech classification using speech biomarkers," *Front. Hum. Neurosci.*, vol. 18, p. 1346297, Feb. 2024, doi: 10.3389/fnhum.2024.1346297.

[6] **C. Spille, S. D. Ewert, B. Kollmeier, and B. T. Meyer**, "Predicting speech intelligibility with deep neural networks," *Computer Speech & Language*, vol. 48, pp. 51–66, Mar. 2018, doi: 10.1016/j.csl.2017.10.004.

[7] **A. Tripathi, S. Bhosale, and S. K. Kopparapu**, "Automatic Speaker Independent Dysarthric Speech Intelligibility Assessment System," *Computer Speech & Language*, vol. 69, p. 101213, Sept. 2021, doi: 10.1016/j.csl.2021.101213.

[8] **S. Venugopalan, Jimmy Tobin, Samuel J. Yang, Katie Seaver, Richard J.N. Cave, Pan-Pan Jiang, Neil Zeghidour, Rus Heywood, Jordan Green, Michael P. Brenner**., "Speech Intelligibility Classifiers from 550k Disordered Speech Samples," Mar. 15, 2023, *arXiv*: arXiv:2303.07533. doi: 10.48550/arXiv.2303.07533.

[9] **A. Huang, K. Hall, C. Watson, and S. R. Shahamiri,** "A Review of Automated Intelligibility Assessment for Dysarthric Speakers," in *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Bucharest, Romania: IEEE, Oct. 2021, pp. 19–24. doi: 10.1109/SpeD53181.2021.9587400.

[10] **E. Yeo, J. M. Liss, V. Berisha, and D. R. Mortensen,** "Potential Applications of Artificial Intelligence for Cross-language Intelligibility Assessment of Dysarthric Speech". doi.org/10.48550/arXiv.2501.15858.

[11] **K. L. Kadi, S. A. Selouani, B. Boudraa, and M. Boudraa**, "Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 1, pp. 233–247, 2016, doi: 10.1016/j.bbe.2015.11.004.

[12] **A. A. Joshy and R. Rajan**, "Automated Dysarthria Severity Classification: A Study on Acoustic Features and Deep Learning Techniques," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1147–1157, 2022, doi: 10.1109/TNSRE.2022.3169814.

[13] **H. Tong, H. Sharifzadeh, and I. McLoughlin**, "Automatic Assessment of Dysarthric Severity Level Using Audio-Video Cross-Modal Approach in Deep Learning," in *Interspeech 2020*, ISCA, Oct. 2020, pp. 4786–4790. doi: 10.21437/Interspeech.2020-1997.

[14] **S. Gupta, Ankur T. Patil, Mirali Purohit, Mihir Parmar, Maitreya Patel, Hemant A. Patil, Rodrigo Capobianco Guido**, "Residual Neural Network precisely quantifies dysarthria severity-level based on short-duration speech segments," *Neural Networks*, vol. 139, pp. 105–117, July 2021, doi: 10.1016/j.neunet.2021.02.008.

[15] **H. Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas Huang, Kenneth Watkin, Simone Frame.,** "Dysarthric speech database for universal access research," in *Interspeech 2008*, ISCA: ISCA, Sept. 2008. doi: 10.21437/interspeech.2008-480.

[16] **F. Rudzicz, A. K. Namasivayam, and T. Wolff**, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Lang Resources & Evaluation*, vol. 46, no. 4, pp. 523–541, Dec. 2012, doi: 10.1007/s10579-011-9145-0.

[17] **A. Al-Ali, Somaya Al-Maadeed, Moutaz Saleh, Rani Chinnappa Naidu, Zachariah C Alex, Prakash Ramachandran, Rajeev Khoodeeram and Rajesh Kumar**., "Classification of Dysarthria based on the Levels of Severity. A Systematic Review," Oct. 11, 2023, *arXiv*: arXiv:2310.07264. doi: 10.48550/arXiv.2310.07264.

[18] **A. H. Andersen, E. Schoenmaker, and S. Van De Par,** "Speech intelligibility prediction as a classification problem," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, Vietri sul Mare, Salerno, Italy: IEEE, Sept. 2016, pp. 1–6. doi: 10.1109/MLSP.2016.7738814.

[19] **A. S. Al-Ali, R. M. Haris, Y. Akbari, M. Saleh, S. Al-Maadeed, and M. Rajesh Kumar,** "Integrating binary classification and clustering for multi-class dysarthria severity level classification: a two-stage approach," *Cluster Comput*, vol. 28, no. 2, p. 136, Apr. 2025, doi: 10.1007/s10586-024-04748-1.

[20] **P. N. Chowdary, V. S. Aravind, G. V. N. S. L. V. Vardhan, M. S. Akshay, M. S. Aashish, and J. L. G**, "A Few-Shot Approach to Dysarthric Speech Intelligibility Level Classification Using Transformers," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, July 2023, pp. 1–6. doi: 10.1109/ICCCNT56998.2023.10308067.