



Hierarchical Convolutional Spectral Feature Learning for Multi-Class Disfluency Classification in the SEP-28k Stuttering Speech Corpus

Afrin Alam*^{lb} and Amritanjali Amritanjali¹^{lb}

* Birla Institute of Technology Mesra, Ranchi, India.

¹ Birla Institute of Technology Mesra, Ranchi, India.

phdcs10055.23@bitmesra.ac.in*

amritanjali@bitmesra.ac.in¹

Abstract: This paper presents a deep learning framework that is used to perform automatic multi-classification of stuttering on the Sep28k dataset. This approach extracts Mel-Frequency Cepstral Coefficients (MFCCs), which help represent the spectral information of the stuttered speech audio signal, whereas a convolutional neural network (CNN) is employed for multi-class classification. The audio data are partitioned into fixed 5-second segments and converted into 40-dimensional MFCC representations to ensure uniform inputs. The proposed features are then encoded, and supervised multi-class classifiers are trained with an 80:20 train-test split. The proposed model is composed of sequential 2D convolutional layers, which are supported by max-pooling and batch normalization. The Dropout Layer within a fully connected Layer to improve generalization and minimize overfitting. The Model undergoes 30 epochs of training using Adam and multi-class log loss. This experiment achieved a test accuracy of 87.9% with good precision, recall, and F1-scores. From the confusion matrix, it is demonstrated that most stuttering subtypes are correctly classified, with only small errors between similar-sounding classes. The result shows that CNN using MFCC features is proved effective in detecting the pattern in stuttering dysfluencies. Although the model is planned as a CNN-BiLSTM system, it currently uses only the CNN layer. Although BiLSTM in the future may enhance time-based learning and it will definitely improve accuracy.

Keywords: Audio feature analysis, Convolutional Neural Network (CNN), Deep Neural Network, Mel-Frequency Cepstral Coefficients, Multiple-class categorization, SEP-28k dataset, Stuttering.

1. Introduction

Stuttering or stammering is a speech disorder that mainly occurs when there is a disruption in the normal flow of speech. This type of disorder takes various forms,

such as dysarthria, apraxia, stuttering, cluttering, lisping, etc. [1] General symptoms of stuttering are phrases or repetition of words, stretching of words, silence or pauses before, after, or within a certain word, and adding extra words such as "um" or "oo" when facing difficulty in moving to the next word [1,2]. The survey study found that approximately 1 million people suffered from these dysfunctions[3]. Whereas in India, nearly 11-12 million people suffer from this disorder. [4] It is observed that this type of dysfluency is more common among males than females. [2] Some common causes of stuttering include stress, delayed developmental progress in childhood, and irregularities in speech-motor control. There are other factors that cause stuttering, and the most common is genetic [2,3,5]. Now, based on the characteristics of speech, stutters are classified into various types, namely block, prolongation, word repetition, sound repetition, and interjection[1, 2]. Prolongation happens when a person, while speaking, holds out a sound for too long. For example, she [eeeeeeee] is going for dinner [rrrrrrr]. This makes the sound abnormal by stretching the word. Block stuttering occurs when a gap or complete stoppage of airflow occurs, resulting in a pause when a person speaks. For example, I [long pause for 5 seconds] went to school today. Interjection type of stuttering happens when a person, while speaking, inserts extra words or sounds such as "um", "uh", "ahh" while speaking something. For example, I have "um" some work 'uh' today. Sound repetition is a common stuttering in which parts of words are repeated multiple times. For example. I [ma- ma-ma-ma] made lunch. Next comes word repetition, which occurs when a person, while communicating with someone, repeats a word/phase multiple times. For example, I [love- love – love] you. [2, 5,6] Manual detection of stuttering is always time-consuming and tedious work. Identification of stuttering becomes a challenging task due to many major factors, namely a person's age, gender, language, speaker task, and the context in which speech is produced. [2,7] Some known stuttering treatments include speech therapy, cognitive-behavioral therapy, advanced assistive devices, support groups or organizations, and medication. In speech therapy, techniques such as slow speech, breathing control, and programs for children like the Lidcombe Program are used to improve speech quality.[1,5] Cognitive Behavioral Therapy (CBT) aids in reducing the anxiety and negative thoughts related to dysfluency. Assistive devices also improve speech fluency, and support groups connect with people who stutter for shared experiences and encouragement. Medication is rarely used but may help in specific cases (e.g., to reduce anxiety)[2, 5,6].

2. Related Work

Over the past decade, the rapid advancement of automatic stuttering detection systems has witnessed remarkable improvement, primarily benefiting from breakthroughs in deep neural network models and speech signal processing technologies. Initial computational studies on a stuttering detection system rely totally on hand-

crafted acoustic features and traditional classification models, whereas modern investigations predominantly employ deep neural network architectures that are capable of learning automatic discriminative speech silent representations from raw processed speech signals. This results in superior and effective detection and classification.

Deep Learning Architectures for Stuttering Detection:

Deep learning architecture provides benefits for the identification of stuttering dysfluencies because it automatically learns discriminative features from high-dimensional input, decreasing its dependency on manual feature engineering. This ShutterNet is represented as a pioneering advancement in the field of stuttering detection, leveraging a Time Delay Neural Network (TDNN) framework, eliminating the dependence on an automatic speech recognition (ASR)-dependent component.

Sheik, S.A., Sahidullah [7] demonstrated ShutterNet, which is a deep learning model used for the detection of different speech dysfluency types. Unlike any other ASR-based model, it directly uses the acoustic signal without any ASR module. The model is built using a time delay neural network (TDNN) to extract the contextual speech pattern and is tested using the UCLASS dataset with over 100 speakers. This framework achieves better performance than other state-of-the-art residual neural network methods. Furthermore, due to TDNN's parameter-sharing mechanism, the model requires fewer trainable parameters, making it more computationally efficient. Sheik, S.A., and Sahidullah [8] introduced a multi-branch (MB) framework with a class-weighted loss applied to the SEP-28k dataset, achieving better results than the baseline StutterNet. The use of data augmentation further boosts the macro F1-score by 4.18%. Additionally, a multi-context (MC) variant of StutterNet is introduced, providing an extra 4.48% improvement in F1-score. In cross-corpus evaluations, incorporating data augmentation results in a 13.23% performance increase compared to models trained without augmentation.

Transformer and Pre-trained Representation Models

Modern deep learning approaches use transformers and self-supervised learning techniques for detecting stuttering dysfluencies, rather than focusing on small segments. Basak, K., Mishra [9] demonstrated TranStutter, a convolution-free Transformer architecture for stuttering type classification. It employs multi-head self-attention and positional encoding to effectively capture temporal speech patterns, outperforming traditional approaches. The model is evaluated on the SEP-28k and FluencyBank datasets, achieving accuracies of 88.1% and 80.6%, respectively. These results demonstrate its ability to detect speech disfluencies and its potential to enhance diagnostic performance in speech pathology.

Researchers have explored self-supervised learning architecture, namely the Wav2Vec2.0 and Whisper encoders, as they can learn useful speech embeddings

that are suitable and effective across multilingual and multi-speaker contexts. Ameer, H., Latif,[10] proposed Whisper for detecting and classifying disfluency types in stuttered speech and contributed in three main areas: improving the SEP-28k dataset, applying Whisper for classification, and proposing an efficient encoder layer freezing strategy. The optimized model achieves an average F1-score of 0.81, reflecting strong performance. The findings also show that deeper encoder layers play a more important role in identifying disfluencies than earlier layers. Overall, the study presents a more efficient and reliable approach, supporting continued progress in this domain.

Deep Learning for Multi-class Stuttering Event Recognition

Recent studies have shifted beyond differentiating multi-class event dysfluency classification, rather than binary classification; its main purpose is to differentiate between individual disfluency events, namely sound repetition, word repetition, prolongation, and blocking. It enhances the more detailed, clinical usefulness of meaningful information from automated systems. An initial study was done by Kour et al.[11] In which a framework was demonstrated that combines a deep residual network with a bidirectional long short-term memory layer for stutter classification, achieving an average miss rate of 10.03% and exceeding state-of-the-art performance by approximately 27%. Unlike existing work that relies on a language model, this proposed methodology uses only acoustic features, allowing it to detect multiple stutter dysfluency patterns without the need for a speech recognition system. Kourko, Kourkounakis, T., and Hajavi [12] proposed a deep learning framework named FluentNet for the identification of multiple types of speech dysfluencies. It uses a Squeeze-and-Excitation Residual CNN to extract rich spectral frame-level features, followed by bidirectional LSTM layers to model temporal dependencies. To concentrate or focus on the most relevant speech segment, an attention mechanism is applied, hence enhancing the overall accuracy of this model. This experiment is also conducted on the Libristutter dataset, achieving strong results compared to other existing benchmark methods.

The contributions of this work are:

1. A computationally lightweight hierarchical CNN framework was proposed for performing eight-class speech disfluency classification on the SEP-28k corpus.
2. The model utilized 40-dimensional MFCC-based acoustic representations to effectively capture speech characteristics and facilitate accurate disfluency recognition.
3. The network was designed with a minimal parameter configuration of 73,416 trainable parameters, reducing computational complexity while preserving classification effectiveness.
4. From the experimental result, it is demonstrated that this proposed framework gains a satisfactory performance result, achieving an accuracy of

87.90%, a weighted average precision of 88%, a recall of 88%, and an F-1 score of 88%,

5. The developed framework provides an adaptable approach for automatic stuttering assessment and can be applied in computer-assisted speech monitoring and therapeutic support systems.

3. Dataset

3.1 .SEP-28k (Stuttering Events in Podcasts):-

This dataset comprises 28,177 clips. (Each audio clip is 3 seconds long. And each clip is labeled into five event types, namely blocks, prolongations, sound repetitions, word repetitions, and interjections. [13] . This dataset was created to address the issue that public datasets face in building a well-generalizable ASIS model. This dataset was the first publicly available, annotated dataset with stuttering labels. Audio was recorded from public podcasts, mainly featuring people who stutter and interviewing others who stutter[8,16]In 2021, Colin Lea [13]and his research team assembled the SEP-28k dataset to fill the gap in large, publicly available stuttered speech resources. The dataset contains approximately 28,000 three-second audio excerpts grouped into five dysfluency classes. This dataset is collected from publicly available podcasts, mainly consisting of interviews in which people who stutter interact with other individuals who stutter. It is observed that, by evaluating several acoustic models on both SEP-28k and the public FluencyBank dataset, the F-1 score improves by 28% and 24%, respectively, simply by increasing the amount of training data. Working with the SEP-28k corpus, Nouf Alhakbani [14]and her colleagues (2025) proposed an automated detection system for the detection of speech dysfluencies . This methodology is based on deep learning models such as convolutional neural networks (CNN) and convolutional long short-term memory (ConvLSTM). The experiment is conducted on two benchmark datasets, FluencyBank and SEP-28K, where it demonstrates strong performance with detection accuracies of 0.97 for CNN and 0.96 for ConvLSTM. This approach achieves satisfactory performance compared to previous approaches and confirms the effectiveness of the proposed method in enhancing stuttering assessment. In 2022, Sebastian P. Bayerl [15] and his co-researchers examined how data partitioning methods affect or influence the dysfluency detection system. wav2vec 2.0 models with a classification head are used, along with support vector machines (SVM) that rely on features extracted from wav2vec 2.0 representations for dysfluency detection.To analyze the performance of the model both speaker-dependent and speaker-independent splits of the SEP-28k dataset are utilized . Results show that the dataset is dominated by a small number of speakers, affecting evaluation reliability. This suggests that random splits can inflate reported accuracy because the same speakers may appear in both sets, whereas speaker-exclusive splits offer a more realistic view of model generalization. The authors called for standardized split configurations to ensure fair and reproducible benchmarking.

Research on SEP-28k has advanced from dataset creation to the use of sophisticated deep learning models. CNN and ConvLSTM achieve the highest accuracy, while transformer-based and augmentation methods enhance temporal modeling and address class imbalance. Remaining challenges include minority class detection, speaker generalization, and integrating linguistic features. Future work will focus on hybrid models, standardized evaluation, and multi-modal integration to improve reliability and practical use.

4. Experimental Workflow
4.1Data Preprocessing

The SEP-28k speech dataset was prepared using Librosa, where all audio clips were standardized to a duration of 5 seconds and processed at a sampling rate of 22.05 kHz. Forty-dimensional MFCC features were extracted from each speech sample using an FFT size of 2048 and a hop length of 512. The obtained frame-level feature values were averaged to produce fixed-length acoustic representations. The corresponding class labels were encoded using LabelEncoder and converted into one-hot vectors for multi-class classification. The generated feature vectors were reshaped into $40 \times 1 \times 1$ input dimensions, and the dataset was randomly divided into training and testing sets with an 80:20 split using `random_state = 42`.

Parameter	Value
Dataset	Sep28k Shuttering Speech corpus(28,177 clips)
Audio Duration	5 s
Sampling Rate	22.05 kHz
Feature Extraction	40 MFCCs
Input Shape	$40 \times 1 \times 1$
Data Split	80% Training / 20% Testing

Table 1. Data Preprocessing Parameters of the proposed CNN model

4.2 Implementation Work

This experimental framework is mainly for performing multitask classification. The first step is acoustic time–frequency feature, namely MFCCs or spectrogram, are scaled and reshaped into a two–dimensional form so as to perform a convolutional processing. MFCCs are used to capture meaningful spectral characteristics of speech signals that are associated with human auditory perception. Their compact feature representation enables the CNN model to recognize distinctive acoustic patterns from different stuttering and disfluency classes, resulting in improved multi-class classification performance.

The architecture begins with a Conv2D layer using 32 filters to learn an elementary spatial pattern from the input feature map. Then a MaxPooling2D layer is assiduous for reducing the spatial size of the representation and minimizing the computational cost. Batch normalization is introduced to ensure stable learning, accelerate convergence, and minimize the variation in the training process. In the next step, a second Conv2D layer consisting of 64 filters is implemented. This extracts the relevant feature characteristics. This stage is again followed by max pooling and batch normalization to enhance the model's overall generalization ability and regularization. After completing the feature extraction task, the resulting multidimensional maps are flattened into a one-dimensional vector, which is then forwarded to a fully connected dense layer containing 128 neurons. This layer performs higher-level abstraction and non-linear integration of the extracted information. Dropout layers are incorporated to randomly deactivate generalization. The final output stage is a dense layer with 8 neurons, each representing a class, and it employs a softmax activation function to generate probability values for multi-class classification. Overall, the network consists of 73,416 trainable parameters, so as to consistently maintain the well-balanced trade-off between representational capacity and computational efficiency. In the training phases, the model's weights are refined via backpropagation with an Adam, which is an effective optimizer, while categorical cross-entropy serves as the loss function. The performance of a model is estimated by classification accuracy. The overall process follows a systematic sequence: extraction of input features → learning of representations through convolutional layers → spatial dimensionality reduction → application of normalization → feature combination using dense layers → final classification with regularization. This structured hierarchical architecture enables effective learning of distinctive patterns from the input image, resulting in reliable multi-class prediction accuracy.

4.3 Hyperparameter Settings

The proposed CNN-based framework was developed in TensorFlow/Keras and consisted of two convolutional blocks with 32 and 64 filters. Each block was followed by batch normalization and 2×1 max-pooling layers to improve feature extraction and representation learning. The extracted features were flattened and passed through a fully connected layer containing 128 ReLU neurons with a dropout rate of 0.3, followed by an eight-class Softmax output layer. The model was trained using the Adam optimizer with a learning rate of 0.001 and categorical cross-entropy loss for 30 epochs with a batch size of 32. The network contains 73,416 parameters, and its performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrix analysis.

Hyperparameter	Value
Optimizer	Adam
Learning Rate	0.001
Batch Size	32
Epochs	30
Convolution Layers	2 (32, 64 filters)
Kernel Size	3×1
Activation Function	ReLU
Dropout	0.30
Output Activation	Softmax (8 classes)
Loss Function	Categorical Cross-Entropy

Table 2. Hyperparameter Configuration of The Proposed CNN Model

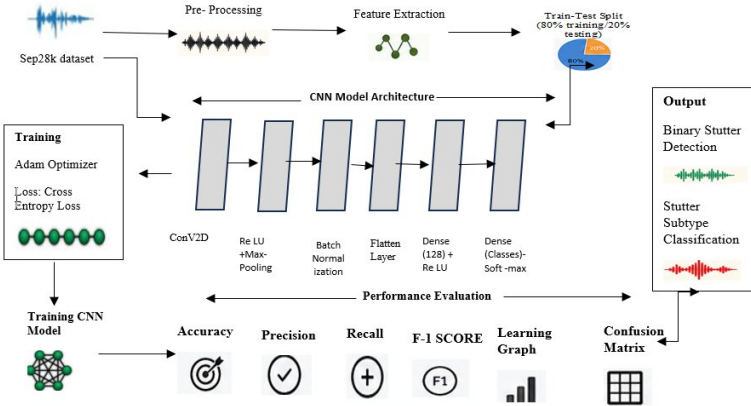


Fig. 1. Work Flow Diagram of Proposed CNN Model

Fig 1 illustrates the workflow diagram of the implemented method. The proposed architecture employs a hierarchical convolutional spectral feature learning strategy, in which successive convolutional layers progressively learn and encode spectral representations from MFCC-based inputs. This multi-level representation learning improves the model's ability to differentiate among various speech disfluency categories while maintaining a lightweight, computationally efficient network design.

5. Result and Discussion

The proposed framework achieved an overall classification accuracy of 87.90% for multi-class speech disfluency recognition, with a weighted precision of 88.00%,

recall (sensitivity) of 88.00%, an F1-score of 88.00%, and an overall specificity of 96.80%. Figure 3 illustrates the confusion matrix obtained from the evaluation of the proposed framework on the SEP-28k dataset. The clear dominance of correct predictions along the diagonal demonstrates that the model effectively identified the majority of samples belonging to all eight speech disfluency categories. The WomenWhoStutter class obtained the highest number of correctly classified samples (1,727), followed by StutterTalk (912), HeStutters (623), and StutteringIsCool (620). The few classification errors were mainly observed among acoustically similar categories, particularly between HeStutters and WomenWhoStutter, and between StrongVoices and StutterTalk. These results highlight the capability of the proposed framework to accurately separate the eight speech disfluency classes while maintaining an overall accuracy of 87.90%. Figure 2 presents the training and validation performance of the proposed framework through accuracy and loss curves obtained over 30 epochs. Both accuracy curves exhibit a progressive improvement during training, reaching approximately 91.8% for the training data and 87.9% for the validation data, indicating effective learning of discriminative speech characteristics and satisfactory generalization. At the same time, the loss values for both training and validation datasets decrease gradually and stabilize during the final epochs, confirming successful convergence of the model. A slight separation between the training and validation curves appears after nearly 15 epochs, indicating only a limited overfitting tendency. Overall, the observed learning patterns confirm the stability, robustness, and dependable performance of the proposed framework. Overall, the obtained results demonstrate that the proposed MFCC-based convolutional feature learning framework is an efficient approach for automated multi-class stuttering recognition. Future work involving temporal sequence modeling techniques may further strengthen the representation of dynamic speech patterns and improve the classification of complex disfluency events.

5.1 Clinical Relevance

The proposed hierarchical convolutional spectral feature learning framework serves as an automated platform for recognizing and classifying multiple speech disfluency types, enabling objective and early stuttering evaluation. By incorporating a lightweight CNN architecture, the framework effectively captures discriminative speech characteristics to distinguish different disfluency patterns. This capability supports speech-language pathologists in clinical screening, diagnostic assessment, and the evaluation of therapy progress while minimizing reliance on subjective clinical interpretation. Furthermore, its computationally efficient design allows seamless integration into real-time speech therapy systems and telehealth services, improving the reliability, consistency, and accessibility of automated stuttering assessment.

LAYER(TYPE)	OUTPUT SHAPE	PARAM#
Conv2d		
Max_pooling2d	(None, 38,1, 32) (None, 19,1, 32) (None, 17, 1, 64)	128
Batch_normalization		0
Conv_2D_1		128
Max_pooling2d_1	(None 8,1, 64)	6,208
Batch_normalization_1	(None 8,1, 64)	0 256
Flatten	(None, 512)	0
Dense	(None, 128)	65,664
Dropout	(None, 128)	0
Dense_1	(None, 8)	1032

Table 3. CNN architecture used for stuttering detection .

Performance Metric	Value (Weighted Average %)
Accuracy	87.90
Precision	88.00
Recall (Sensitivity)	88.00
F1-score	88.00
Specificity	96.80

Table 4. Performance Matrix of the Proposed Model

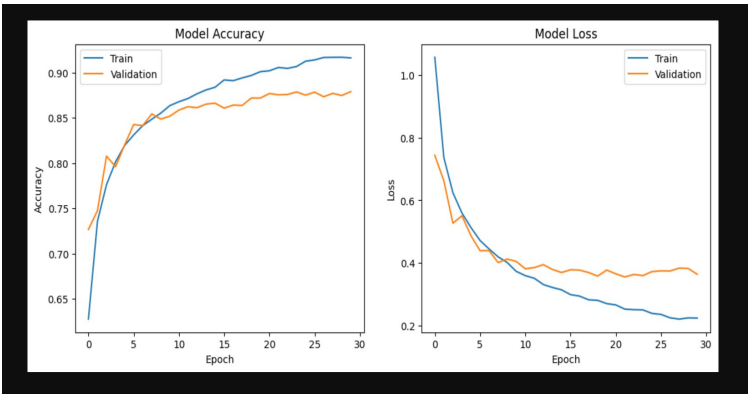


Fig. 2. The training accuracy and loss plot obtained during training the model.

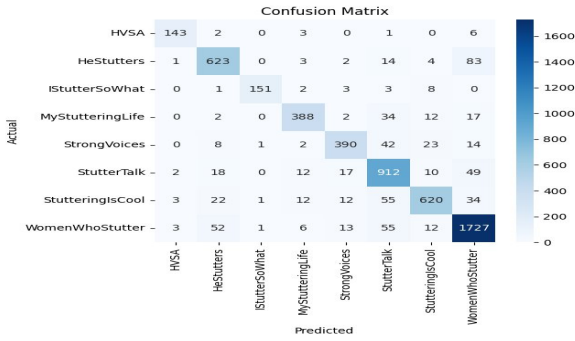


Fig. 3. Confusion Matrix obtained during training the model.

RESEARCH WORK	DATASET USED	FEATURE EXTRACTION	TECHNIQUE USED	ACCURACY ACHIEVED
SHEIKH, S. A. ET AL [7]	UCLASS DATASET	MFCC	TDNN (STUTTERNET)	~83–85%
basak k. mishra[9]	sep28k dataset and fluency bank dataset	2D MEL SPECTROGRAM	TRANSHUTTER(TCONVOLUTION FREE TRANSFORMER BASED DL MODEL)	80.6% on fluencybank dataset and 88.1% on sep28k-dataset
sheikh, s.a., sahidullah, [16]	sep 28k dataset	deep speech embeddings & wav2vec2.0 embeddings	ml classifier	80- 85%
sharma, n.m., mahapatra[17]	uclass dataset	mfcc	knn	87.1%
sheikh, s.a., sahidullah. [18]	sep28k dataset	mfcc	multi-task (mtl) and adversarial learning (adv)	60.3%
PROPOSED WORK	SEP-28K	40-D MFCC (STATISTICAL MEAN REPRESENTATION)	DEEP CNN (2 CONV LAYERS + DENSE + DROPOUT)	87.9 %

Table 5. Comparative Analysis of Existing Stuttering Detection Work.

6. Conclusion and Future Work

This research proposes a hierarchical convolutional spectral feature learning framework for automated multi-class speech disfluency recognition using the SEP-28k

corpus. The framework integrates 40-dimensional Mel-Frequency Cepstral Coefficient (MFCC) features with a lightweight convolutional neural network (CNN) to learn representative spectral patterns from stuttered speech signals. Experimental results indicate that the proposed model achieves an overall classification accuracy of 87.90%, together with weighted precision, recall, and F1-score of 88% across the eight speech disfluency categories. The micro-averaged evaluation further reports 90% precision, 87% recall, and an F1 Score of 88%, demonstrating balanced performance across all classes. The confusion matrix and the corresponding training–validation curves also verify the model’s stable convergence and dependable classification capability. With its compact architecture and low computational complexity, the proposed framework provides an effective solution for automatic speech disfluency recognition and is suitable for integration into computer-assisted stuttering assessment systems. Future studies will focus on incorporating self-supervised speech representation models, including HuBERT and Wav2Vec 2.0, alongside attention-based neural architectures to capture richer temporal dependencies and contextual information in speech. Furthermore, validating the proposed framework on multilingual and cross-corpus stuttering datasets while employing advanced augmentation techniques and explainable AI methods may further improve its robustness, generalization performance, and applicability in practical speech therapy and clinical assessment.

Acknowledgement

The authors would like to express their sincere appreciation to their institution for providing the computational facilities and support required to carry out this research.

References

- [1]Hegde, S., Shetty, S., Rai, S., Dodderi, T.: A survey on machine learning approaches for automatic detection of voice disorders. *Journal of Voice* 33(6), 947–e11 (2019).
- [2]Barrett, L., Hu, J., Howell, P.: Systematic review of machine learning approaches for detecting develop mental stuttering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30, 1160–1172 (2022).
- [3]Yairi, E., Ambrose, N.: Epidemiology of stuttering: 21st century advances. *Journal of Fluency Disorders* 38(2), 66–87 (2013).
- [4]The Indian Stammering Association: Stammering FAQ. <http://stammer.in/stammering-faq/>, last accessed 2026.
- [5]Guitar, B.: Stuttering: An integrated approach to its nature and treatment. 2nd edn. Lippincott Williams & Wilkins, Philadelphia (2013).

- [6] Sheikh, S.A., Sahidullah, M., Hirsch, F., Ouni, S.: Machine learning for stuttering identification: Review, challenges and future directions. *Neurocomputing* 514, 385–402 (2022).
- [7] Sheikh, S.A., Sahidullah, M., Hirsch, F., Ouni, S.: StutterNet: Stuttering detection using time delay neural network. In: *Proceedings of the 29th European Signal Processing Conference (EUSIPCO 2021)*, pp. 426–430. IEEE (2021)
- [8] Sheikh, S.A., Sahidullah, M., Hirsch, F., Ouni, S.: Advancing stuttering detection via data augmentation, class-balanced loss, and multi-contextual deep learning. *IEEE Journal of Biomedical and Health Informatics* 27(5), 2553–2564 (2023).
- [9] Basak, K., Mishra, N., Chang, H.T.: TranStutter: A convolution-free transformer-based deep learning method to classify stuttered speech using 2D mel-spectrogram visualization and attention-based feature representation. *Sensors* 23(19), 8033 (2023).
- [10] Ameer, H., Latif, S., Latif, R., Mukhtar, S.: Whisper in focus: Enhancing stuttered speech classification with encoder layer optimization. *arXiv preprint arXiv:2311.05203* (2023).
- [11] Kourkounakis, T., Hajavi, A. and Etemad, A., 2020, May. Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6089–6093). IEEE.
- [12] Kourkounakis, T., Hajavi, A. and Etemad, A., 2020. FluentNet: end-to-end detection of speech disfluency with deep learning. *arXiv preprint arXiv:2009.11394*.
- [13] Lea, C., Mitra, V., Joshi, A., Kajarekar, S., Bigham, J.P.: Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter. In: *ICASSP 2021*, pp. 6798–6802. IEEE (2021).
- [14] Alhakbani, N., Alnashwan, R., Al-Nafjan, A. and Almudhi, A., 2025. Automated stuttering detection using deep learning techniques. *Journal of Clinical Medicine*, 14(10), p.3552.
- [15] Bayerl, S.P., Wagner, D., Nöth, E., Bocklet, T. and Riedhammer, K., 2022, September. The influence of dataset partitioning on dysfluency detection systems. In *International Conference on Text, Speech, and Dialogue* (pp. 423–436). Cham: Springer International Publishing.
- [16] Sheikh, S.A., Sahidullah, M., Hirsch, F. and Ouni, S., 2022. Introducing ECAPA-TDNN and Wav2Vec2. 0 embeddings to stuttering detection. *arXiv preprint arXiv:2204.01564*.
- [17] Sharma, N.M., Mahapatra, P.K. and Gandhi, V., 2024, February. Analysis of Machine Learning Methods for Speech Disfluencies' Classification. In *International Congress on Information and Communication Technology* (pp. 13–22). Singapore: Springer Nature Singapore.
- [18] Sheikh, S.A., Sahidullah, M., Hirsch, F. and Ouni, S., 2022 August. Robust stuttering detection via multi-task and adversarial learning. In *2022 30th European signal processing conference (EUSIPCO)* (pp. 190–194). IEEE.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

