IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 January 2024, accepted 16 January 2024, date of publication 19 January 2024, date of current version 20 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3356350

RESEARCH ARTICLE

Disfluency Assessment Using Deep Super Learners

SHEENA CHRISTABEL PRAVIN[1], SUSAN ELIAS[1], VISHAL BALAJI SIVARAMAN[2], G. ROHITH[1], AND Y. ASNATH VICTY PHAMILA[3]

[1]School of Electronics Engineering, Vellore Institute of Technology, Chennai 600127, India
[2]Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA
[3]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai 600127, India

Corresponding author: Sheena Christabel Pravin (sheenachristabel.p@vit.ac.in)

This work was supported by the Vellore Institute of Technology, Chennai Campus, Chennai, India.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT The use of machine learning algorithms for the assessment of speech fluency is increasingly becoming recognized globally due to their ability to quickly identify speech impairments. This approach is preferred over manual diagnosis, as it reduces the likelihood of human error and minimizes the delay in commencing the therapy. A pipelined deep learner-dual classifier (PDL-DC) is proposed for the automated detection of speech impairment. The assessment of individuals' speech fluency consisted of two distinct phases: the classification of speech disfluencies and the categorization of fluency disorders. Speech disfluencies, including revisions, prolongations, whole-word repetitions, word-medial repetitions, and filled pauses, were categorized into distinct groupings. The second aspect of classification pertains to the assessment of fluency levels, wherein speakers are classified into three categories: healthy individuals, individuals with stuttering, and individuals with Specific Language Impairment (SLI). The proposed model's implementation of a pipelined design enables the dual validation of a subject's fluency. The proposed model demonstrates an average classification accuracy, precision, and recall of 97%.

INDEX TERMS Fluency assessment, speech impairment, pipelined deep learner-dual classifier, healthy, stuttering, specific language impairment.

I. INTRODUCTION

Speech is essential for communication because it allows us to express ourselves and use platforms that are speech-based. Disfluency is any interruption in speaking and can hurt a person's quality of life. Speech impairments make it difficult for the speaker to render speech fluently. The nature of speech impairments varies depending on the type of disability [1]. Disfluency patterns are often indicative of speech disorders, such as stuttering. This research is motivated by the potential impact on clinical speech therapy, where a deeper understanding of disfluencies can lead to improved diagnostic tools and personalized therapeutic interventions. Filler disfluency identification, which detects and counts any spoken interjections (such as "okay," "right," etc.), is a component of several disfluency detection algorithms. Further investigation indicates that these applications essentially seek a list of interjections from the user before using the Speech-to-Text (STT) technology to match any interjections in the list with the spoken phrase. This is effective for interjections like "um" and "uh", as long as the STT tool being used contains the required embeddings, but it can result in substantial categorization errors for the majority of other utterances that are meaningful words, such as "like," which is frequently used as a filler word in English. Since they come in a range of shapes and sizes, stuttering and other disfluencies are challenging to characterize. This significantly complicates the problem space because variables like gender, age, accent, and language may have a sure impact on each stutter's content.

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

It is difficult to distinguish between all types of stuttering using a single model because there are multiple classes of stuttering, each with distinct sub-classes and significantly different structures. Even a specific sort of stutter applied to a single phrase can be performed in several ways. People are exceptional at spotting stutters because of their familiarity with them, but machine-learning models have traditionally struggled with this. As a result, it is critical to properly discern the disfluencies associated with each of the speech impairments to provide an accurate diagnosis of the specific speech impediment. This would allow for more prompt treatment for the individuals. Moreover, disfluencies caused by multilingualism limit the accuracy of the manual diagnosis of speech impediments. Manual evaluation tools, such as the Iowa Scale [2], rate the severity of stuttering on a range of 1 to 8. The Stuttering Severity Instrument (SSI-4) was also used to differentiate between mild and severe stuttering severity levels based on the syllabic count, which revealed substantial disparities across Speech Language Pathologists (SLPs). Consequently, distinguishing a speech-language impediment from stuttering is extremely difficult [3], [4]. As a result, a machine learning-enabled evaluation approach would assist SLPs in detecting the appropriate speech impediment and eventually recommending the appropriate therapy.

An autoencoder-based super learning architecture named Pipelined Deep Learner—Dual Classifier (PDL—DC) is proposed to rapidly and accurately measure fluency and its disorders. Autoencoders find latent features, find anomalies, classify images, and process natural language. An autoencoder has an input encoder, a hidden layer, and an output decoder. After training on the input features, the encoder expects test data class labels after reconstructing the input. The decoder restructures the bottleneck representation to create a rebuilt feature vector.

The significant contributions of this research work are:

- A unique deep learning framework for the Pipelined Deep Learner-Dual Classifier (PDL-DC) with the Speech Disfluency Categorization and the Fluency Disorder Classification is proposed. Testing of the proposed model on the state-of-the-art UCLASS dataset for validating the efficacy of the proposed model.
- The core goal of deciphering the type of speech fluency impairment of an individual is targeted in two phases for accurate prediction of the fluency disorder. Phase I includes speech disfluency classification, while Phase II incorporates the categorization of speech fluency disorders into three classes: stuttering, specific language impairment, and healthy. The aforementioned phases of classification are two separate entities, and to the best of the author's knowledge, the combination or connection of one system to another was not attempted earlier.

II. BACK GROUND

A speech disfluency is any aberration or generally unusual component of one's speaking patterns. There are hundreds of distinct types of speech disfluencies, which are sometimes lumped along with language and swallowing difficulties. Stuttering, often known as stammering, is a disease characterized by problems with the consistency of the flow and fluency of speech. This frequently entails unintentional additions of sounds and phrases, as well as a delay or difficulty in continuing steadily through a sentence. Despite being described as a condition, stuttering can occur in anyone's speech and is frequently triggered by stress or anxiousness. The speech disfluency classification models, which incorporate a unique hybrid deep ensemble [5], [6], [7] for categorizing varied speech disfluencies in the UCLASS [8] and Fluency Bank datasets [9], achieve an accuracy range of 97 to 98.1 %. FluentNet is a discrete model that comprises a residual CNN [10] that learns frame-level representations of the speech spectrum. In the UCLASS dataset, the residual CNN was followed by cascading Bi-LSTM, which displayed an average accuracy of 91.75 %. Additionally, a disfluency recognition system using a deep residual network and Bi-LSTM [11] obtained a miss rate of 10.03 % while categorizing distinct stuttering disfluencies. An innovative strategy that uses a decision tree model based on prosodic characteristics collected from a voice data corpus [12] achieved 75.5% accuracy on average. The best recall accuracy of the speech disfluency system is between 84.8 and 89.7% for fluent vs. neutral-disfluent classes and between 87.7 and 96.9% for fluent-neutral vs. disfluent classes. This is done by running prosodic features from a Japanese speech corpus through several support vector machines [13]. Using Mel Frequency Cepstral Coefficients (MFCCs) and phoneme probabilities from the datasets of UCLASS, Fluency Bank, and SEP-28K, the new method displayed a constant accuracy of 94% when training a neural network [14] to classify stuttering into four disfluency classes. According to experimental data, the extraction of acoustic features [15] from the UCLASS stuttered speech corpus presented a 98 % recognition rate when identifying two types of pauses. Using a dedicated Gaussian mixture model, glottal features and other conventional speech features like MFCCs, intensity, cepstra, and pitch were extracted from speech signals to detect three distinct classifications of disfluencies [16]. When trained on the Cognitive Stroop test database, the baseline accuracy was 79–84% [17]. A k-NN classifier may detect stuttering by extracting speech features, mainly MFCC, from a voice corpus [18].

A novel approach that begins with the extraction of MFCC features from the UCLASS database and then uses a combination of a k-NN classifier and support vector machines [19] yields accuracy in the range of 86 % to 93 % in distinguishing fluent speech from disfluent speech. After extracting MFCCs from the speech dataset, a Stuttering Speech Recognition (SSR) system [20] could distinguish class labels at a 94 % accuracy rate using an adaptive optimization-based artificial neural network. Another SSR used support vector machines, Gaussian mixture models, and vector quantization to analyze prosodic and source characteristics from the UCLASS dataset achieving 90.51 %, 84.33 %, and 86.71 % accuracy [8] respectively. An SSR system [21], [22] that used MFCC

TABLE 1. Characteristics of the disfluency types.

Disfluency Type	Description
Filled Pause	Fillers used to conceal pauses in speech
Prolongation	Extended vowels during speech rendition
Revision	Introducing amendments to the original speech rendition with an altered word or phrase
Word Repetition	Reiteration of the whole word
Word-medial Repetition	Reiteration of part words during speech rendition

TABLE 2. Number of disfluencies in the dataset.

Disfluency Type	Count
Filled Pause	37
Prolongation	923
Revision	290
Word Repetition	997
Word-medial Repetitions	513

features extracted from the UCLASS database showed an average accuracy between 93.10% and 95.88% when using the DMFCC-Vector Quantization (VQ) framework for analysis. Notably, a few studies in the recent literature [5], [10], [14], [15], [18], [22], have attempted to measure the degree of disfluency using traditional machine-learning approaches. While machine learning approaches are less sensitive in phonation deviation evaluation, a deep learner can provide accurate categorization. Identifying Specific Language Impairment (SLI) and stuttering is another important but sometimes overlooked part of a system that can be used in real-time technology-assisted treatment. This motivated the authors to propose a hybrid deep learning model with the creation of a new dataset and also compared it with the state-of-the-art UCLASS dataset [8].

III. RESEARCH FINDINGS

A. DATASET DESCRIPTION

The disfluent speech data corpus used in this study is made up of spontaneous English speech samples from 27 bilingual children, 14 male and 13 female, whose first language is Tamil and their second language is English. They are between the ages of four and seven years. The speech samples were recorded after informed consent was obtained from a parent or legal guardian. All methods were carried out under relevant guidelines and regulations as in [23]. The sampling frequency maintained all through the recording was 16 kHz. Bilingual children demonstrated several disfluencies during their speech rendition in their second language of communication other than their mother tongue, according to careful observation. Their disfluent speech utterances were evaluated to identify glottal feature abnormalities caused by speech disfluencies. In the children's utterances, five disfluencies were observed: filled pauses, prolongations, revisions, word repetitions and word-medial repetitions. Table 1 provides a summary of the disfluencies as well as a description of the same.

TABLE 3. Count of disfluencies as per fluency disorder.

Disfluency Type	Count
Healthy disfluencies	290
SLI disfluencies	1033
Stuttering disfluencies	1437

From an extensive literature review and practical interaction with children aged four to seven, the authors concluded that those with stuttering disorder exhibited more prolongation and word-medial repetitions, whereas those with Specific Language Impairment (SLI) exhibited filled pauses and whole-word repetitions. The healthy patients were seen to be fluent but exhibited revision disfluencies at times.

Table 2 and Table 3 elaborate on the total number of healthy disfluencies such as revisions, SLI disfluencies such as whole-word repetitions and filled gaps, and stuttering disfluencies such as prolongations and word-medial repetitions.

PDL-DC was rigorously tested on the UCLASS dataset using Leave-One-Subject-Out (LOSO) cross-validation to confirm the effectiveness of the proposed model. The output of models evaluated against this dataset is presented as the mean of 25 experiments. A test set from one participant and training data from 24 people are combined to create each experiment. The UCLASS dataset was subjected to a 10-fold cross-validation procedure, with a randomly chosen selection of 90% of the samples from each stutter and 90% of the clean samples being utilized for training. Tests were conducted on the remaining 10% of both clean and stuttering samples. Following training over 30 epochs for all experiments, the loss exhibited no variation.

B. FEATURE EXTRACTION

Over voiced segments, a total of 9 glottal parameters and their four statistical functionals were calculated. This gave us 36 glottal characteristics. To learn more about the glottal features, the difference between different kinds of stuttering was also brought up. The glottal characteristics from the speech frames as shown in Table. 4 was used to train the proposed hybrid deep learning model. The mean open quotient, which represents the open and closure peaks of the signals at the glottis, was calculated as the proportion of the opening phase to the whole duration of the glottal signal cycle, as shown in equation (1). The disfluencies had a longer open quotient, indicating that the stuttering participants' voice start time was prolonged.

$$\text{Mean Open Quotient} = \frac{\textit{Glottal opening phase period}}{\textit{Duration of glottal cycle}} \quad (1)$$

The Open Quotient (OQ) variability was measured as the rate of opening phase duration of the glottal flow signal, measured using equation (2).

$$\text{OQ (Variability)} = \frac{\textit{Opening phase duration}}{\textit{Duration of glottal cycle}} \quad (2)$$

TABLE 4. Glottal features.

Glottal Features	Feature Description
GCI	Glottal Closure Instants
Mean OQ	Average Open Quotient for consecutive glottal pulses
OQ Variability	Inconsistency of Open Quotient for successive glottal pulses
Mean NAQ	Averaged and Normalized Amplitude Quotient for successive glottal pulses
NAQ Variability	Inconsistency in Normalized Amplitude Quotient for successive glottal pulses
Average H1H2	Variation in the first two harmonics of the glottal function
H1H2 Variability	Inconsistency in the difference between the first two harmonics of the glottal pulses
Mean HRF	Average Harmonic Richness Factor
HRF Variability	Inconsistency in Harmonic Richness Factor

The Normalized Amplitude Quotient viz. NAQ parametrizes the closing phase of the glottis and was measured to be the ratio of the quotient of amplitude to the time taken for one glottal cycle to complete as shown in equation (3).

$$\text{NAQ} = \frac{Amplitude\ Quotient}{Duration\ of\ glottal\ cycle} \tag{3}$$

The parameter H1H2 is the difference in the amplitude of the first two glottal formants H1 and H2 which is a measure of vocal quality. It was computed as per equation (4). The change in the open quotient reflected on the H1H2 parameter is shown in Figure 1(a).

$$\text{H1H2} = 10log_{10}\left(\frac{|A_s(f_0)|}{|A_s(2*f_0)|}\right) \tag{4}$$

where, $A_s(f_0)$ is the spectral amplitude at the fundamental frequency (f_0) is the spectral amplitude at twice the f_0.

The Harmonic Richness Factor is measured as the fraction of the spectral amplitude of harmonics in total w.r.t. the amplitude of fundamental frequency as shown in equation (5).

$$\text{Harmonic Richness Factor} = 10\ log_{10}\left(\frac{\sum |A_s(f_{12})|}{|A_s(f_0)|}\right) \tag{5}$$

where, $A_s(f_n)$ is the spectral amplitude of the n^{th} harmonic.

As shown in Figure 1, the mean Glottal Closure Instants (GCI) for individuals with specific language impairment and healthy instances were found to be lower due to a considerably longer glottal closure time required in the successive glottal cycles. Those who stutter had a lower normalized amplitude quotient during subsequent glottal cycles. Because of the disturbed voice onset, the stuttering individuals had the lowest average open quotient for successive glottal cycles. Also, the variance in the first harmonics of the glottal pulse increased significantly for the stuttering people. The density distributions of the average OQ, NAQ, and H1H2 were different for each disfluency, as shown in Figure 2. Furthermore, the mean harmonic richness factor separated the filled pauses from the other disfluencies. Since glottal characteristics distinguished disfluencies distinctively, they were used to train the hybrid deep learning model for disfluency classification.

C. PROPOSED MODEL

The proposed hybrid model was made as a precise screening tool to find out the speech-related issues in a person, as a result of an intrinsic categorization of speech disfluencies. The suggested hybrid model was hailed as the Pipelined Deep Learner-Dual Classifier. Figure 3 displays the proposed PDL-DC architecture, which includes a hybrid deep learner for dual categorization. The suggested PDL-DC model, as shown in the figure, combines a convolutional autoencoder [24] with a super learner model [25] for fluency evaluation. The procedure is divided into two stages: the disfluency categorization phase and the fluency disorder classification phase. The Speech Disfluency Categorization phase aims to decipher the individual's speech disfluency uttered using the proposed hybrid deep learner model. The Fluency Disorder Classification phase deciphers the type of speech fluency disorder suffered by the individual using the same hybrid deep learner model. Furthermore, the suggested stages are diagnosed independently of one another. In this research work, the convolutional autoencoder was incorporated for feature engineering, to narrow down the best features collated with the proposed PDL-DC model for executing accurate diagnosis at each phase. The convolutional autoencoder model is composed of the encoder-decoder section each made of three strata of convolutional layers. Each of the layers has been constructed with a different number of neurons viz. first layer was built with 36 neurons identical to the last layer of the decoder. Further, the kernel size in each layer of the encoder and decoder was tweaked to (2,2).

The model summary of the CAE with its encoder and decoder is presented in Figure 5. The proposed PDL-DC model, as previously stated, includes a super learner model that intends to do analysis and prediction with high accuracy and precision. The super learner is a composite of machine learners that use various categorization algorithms. The super learner model's process seeks to maintain an ideal weight average of all the models. As a consequence, the super learner model has been shown to provide accurate predictions. The super-learner model in this study was built using eight of the best-performing base models: logistic regression, passive aggressor classifier, k-nearest neighbor's classifier, random forest classifier, bagging classifier, extra trees classifier, extreme gradient boosting (XGB) classifier, and multi-layered perceptron classifier. Furthermore, the suggested super learner model has been combined with a meta-learner, the XGB Classifier, for the prediction of class labels. The architecture of the super learner is seen in detail in Figure 4. The entire dataset utilized for model training, validation and testing the Pipelined Deep Learner - Dual Classifier as a whole was partitioned into a 60:20:20 ratio,

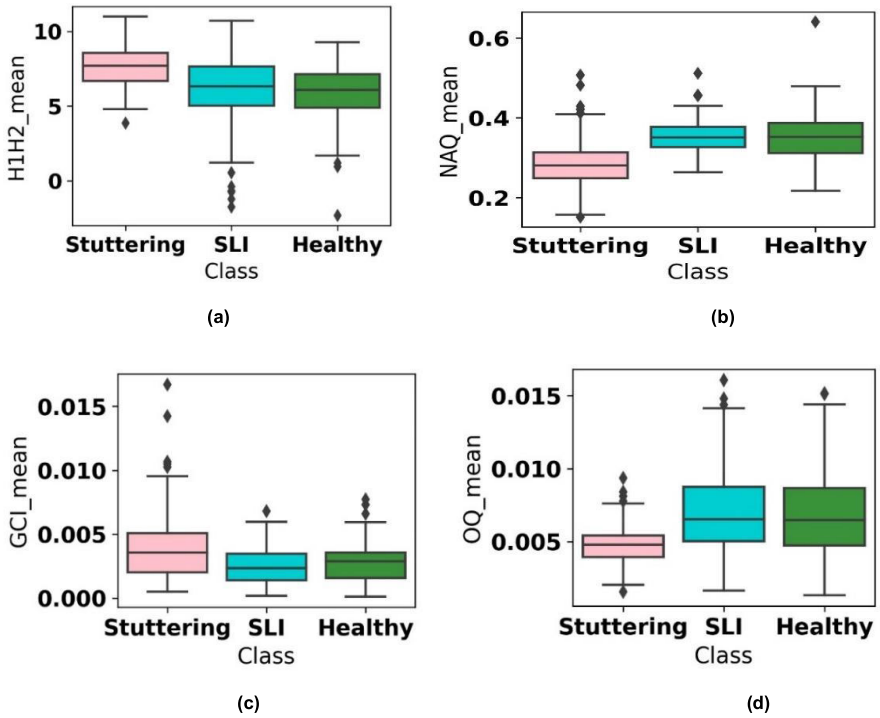


FIGURE 1. Glottal feature comparison of Stuttering, SLI, and Healthy speech.

with 60% of the data utilized for the model's learning phase while the remaining 40% was dedicated for model validation and testing [26].

This is conventionally followed to train and test the proposed model. Finally, the activation functions of each layer of the encoder and decoder were modified to Rectified Linear Unit (ReLU) with the same padding.

IV. RESULTS AND DISCUSSION

The hypothesized phases of the hybrid model were assessed using conventional classification performance measures such as Cohen's Kappa coefficient, classification accuracy, F1-score along with its weighted, macro and micro measurements, balanced accuracy, Jaccard score. Further, the error metrics such as Hamming Loss were also investigated. Also, the Confusion matrix was plotted to evaluate the effectiveness of the base-models and the proposed Pipelined Deep Learner-Dual Classifier (PDL-DC) model in distinguishing one class label from other labels.

A. CLASSIFICATION

An evaluation metric that describes the model's ability to identify the output labels uniquely. It is the fraction of the right predictions to the total number of classifications made by the model. For 20 trials, the mean categorization accuracy was calculated using equation (6).

$$\text{Classification Accuracy} = \frac{Number\ of\ Correct\ Forecasts}{Total\ Number\ of\ Predictions} \tag{6}$$

From Table.5, it is interpreted that there is an improvement in the model testing during the *Fluency Disorder Classification phase* by 0.820 as compared to the *Speech Disfluency Categorization phase*. Also, while validating the model, the *Fluency Disorder Classification phase* scores 0.70 % points better than the *Speech Disfluency Categorization phase.* This proves the efficacy of the model in identifying the correct disfluency disorders and categorizing them properly.

B. KAPPA COEFFICIENT

A statistical metric that is used to measure the level of correlation between the predicted and the true labels in a model's prediction is *Cohen's kappa coefficient*, presented as *kappa* in the equation (7). The range of *kappa* is between 0 and 1; a score close to 1 indicates a competent classifier.

$$kappa = \frac{TL - PL}{1 - PL} \tag{7}$$

where TL: True label and PL: Predicted label

Table.5 demonstrates that when the proposed model is validated and tested, it differs just by 1 % point in *kappa* value between the *Speech Disfluency Categorization phase* and the *Fluency Disorder Classification phase*. This displays the model's accuracy in correlating the true and predicted labels.

C. F1 SCORE

The proposed phases of the hybrid model were rigorously evaluated by calculating the F1 score. It is the measure of the

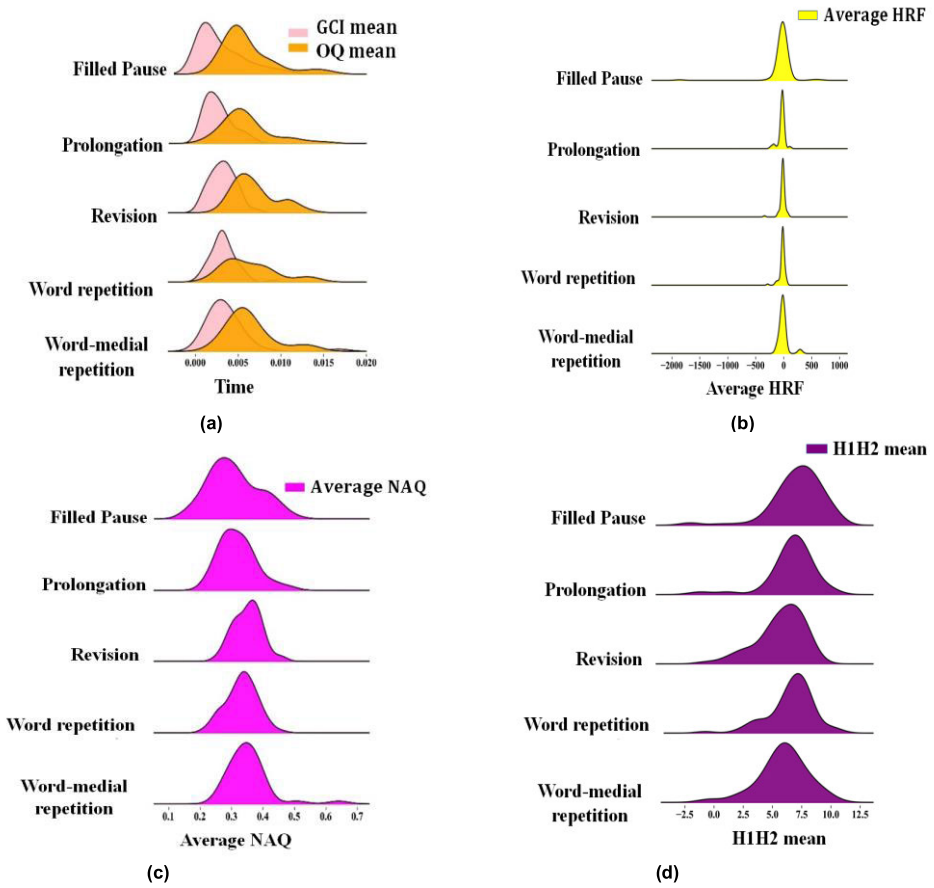


FIGURE 2. Comparison of disfluency classes' w.r.t. the Glottal features.

mean taken between the recall and precision scores. It was calculated using equation (8).

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{8}$$

After the model was verified and evaluated, the proposed model during the *Fluency Disorder Classification phase* surpassed in performance over the *Speech Disfluency Categorization phase* by one percentage point, as shown in Table 5. This improvement is seen in all three cases viz. macro, micro and weighted average. This shows how well the proposed model detects and categorizes fluency disorders.

D. JACCARD SIMILARITY DISORDER

The predicted labels and the true labels were compared by measuring the Jaccard Similarity Index (JSI). It gives the score of the similitude between the classification model's prediction and the true class labels. It was measured utilizing equation (9). The JSI disorder ranges from 0 to 1. A higher JSI rating emphasizes accurate categorization.

$$JSI\left(y_{true}, y_{pred}\right) = \frac{\left|y_{true} \bigcap y_{pred}\right|}{\left|y_{true} \bigcup y_{pred}\right|} \tag{9}$$

The proposed model during the *Fluency Disorder Classification phase* performs better by 2 % of JSI than during the *Speech Disfluency Categorization phase*. This improvement is seen in all three instances, namely macro, micro, and weighted average. This demonstrates how accurately the proposed model discerns the similarities between the predicted label and the actual label which indicates how likely the specific disfluencies have occurred.

E. HAMMING LOSS

The Hamming Loss is given by the proportion of true classification labels to the detected labels, as depicted in equation (10).

$$Hamming\ Loss\left(y_{true}, y_{pred}\right) = \frac{\left|No.\ of\ y_{true}\right|}{\left|No.\ of\ y_{pred}\right|} \tag{10}$$

The results obtained during the evaluation of the postulated phases of the proposed hybrid model are illustrated in Table 5. It is inferred that during the *Speech Disfluency Categorization phase* and the *Fluency Disorder Classification phase,* the proposed model maintains an average of 0.02 loss. This indicates the loss is almost nil.

The efficacy of the proposed Pipelined Deep Learner-Dual Classifier (PDL-DC) model in both phases of classification was measured using confusion matrices. Along the diagonal elements of the confusion matrix, the true classifications are

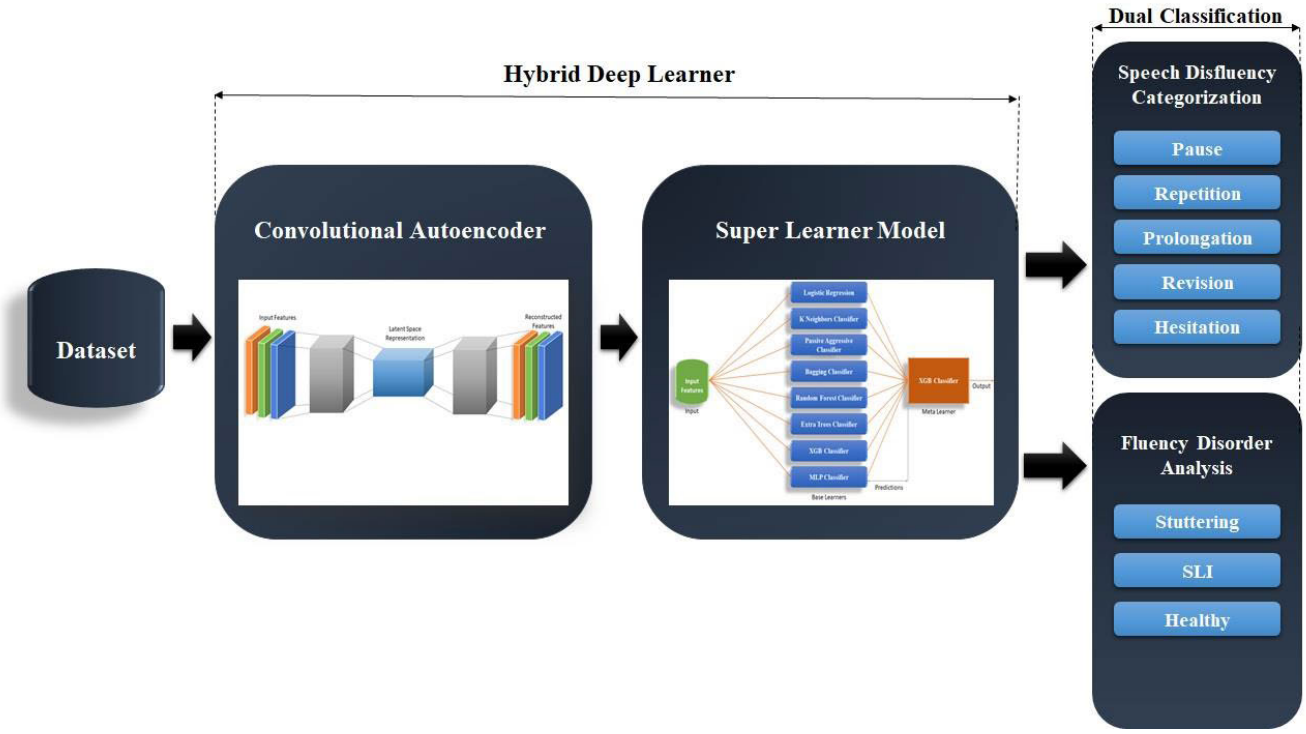


FIGURE 3. Proposed pipelined deep learner deep learning framework.

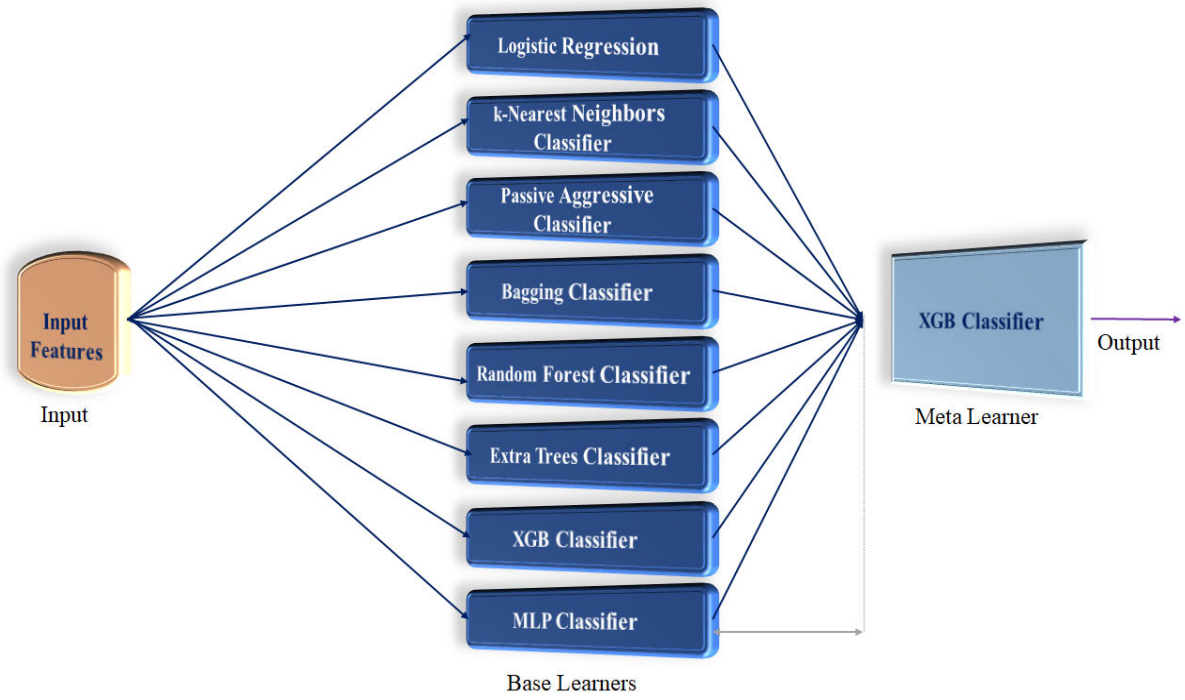


FIGURE 4. Super learner model architecture.

presented. The proposed PDL-DC model was trained over 70% of the available dataset while the remaining 30% of the instances in the dataset were used as validation instances. Models were trained with a batch size of 32 and a learning rate of 0.001 using the Adam optimizer. With an early stopping requirement of 15, early stopping was employed based on

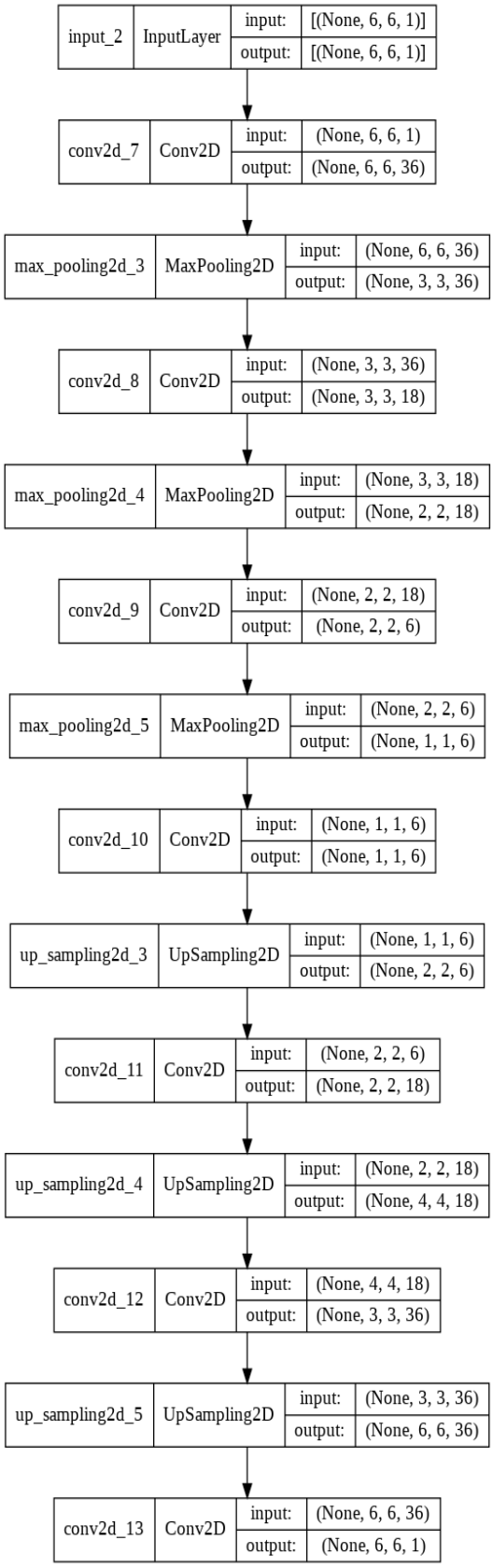


FIGURE 5. Encoder layout of the convolutional autoencoder.

Algorithm 1 Pseudocode for the CAE-based Feature Engineering

```
begin
    set vector size of input feature as 'd', the number of hidden layers as 'l', and the number of classes as 'C' initialize the epochs, number of neurons in the hidden layers and learning rate
    build a CAE with 'd' input nodes, 'l' layers of hidden nodes with 64 units;
    set number of units at the bottleneck layer as 10;
    set number of neurons at the output layer of the CAE as 'd' to match with the input layer;
    set CAE weights W_t and Biases V^(1), V^(2) to an initial zero;
    for each training epoch
    for databatch
        compute feature resynthesis:
        x̃ = φ(W_t.φ(W_t.x + V^(1) + V^(2))
        Evaluate the loss of reconstruction:
        D_KL (P(ỹ) | P(y)) D_KL (P(ỹ) | P(y)) = − Σ_{ỹ∈y} P (y) log (P(y) | P(ỹ))
        set an update to modify the CAE weights W_t and V^(1), V^(2)
        end
    end
    discard output layer;
    extricate feature vector from the bottleneck layer;
end
```

loss. Also, it was found that the proposed hybrid model's postulated phases yielded superior results in both phases of classification.

F. CONFUSION MATRIX

The deployment of stutter detection using the proposed model does not just apply to individuals with long-term stuttering difficulties, but it may also appeal to the rest of the globe since it can aid with communication skill enhancement. The contemporary models in the literature were compared with the proposed Pipelined Deep Learner - Dual Classifier model's performance which is presented in Table 6. The literature in [18, 22, 15, 10, 5, and 14] is compared against the proposed method. In literature [18, 22, 15, 10, 5 and 14] and the proposed work, UCLASS dataset is commonly used for testing the conventional disfluency classification models. UCLASS dataset comprises recordings from 128 stuttering children and adults. In this research work, the UCLASS annotations of 25 subjects released by [10] were utilized. While testing the proposed model with UCLASS dataset for *Speech Disfluency Categorization phase*, the results of testing and validation accuracy are almost the same. This proves the efficacy of the model for the state-of-the-art dataset. The difference

Algorithm 2 Pseudocode for the Super-Learner-based Prediction

```
begin
    Extricate bottleneck features from the CAE;
    for 1 to total number of epochs:
        fragment data into k folds;
        for the respective fold:
            for every base − model in the Super Learner :
                fit the base-model on the bottleneck features in the present fold;
                validate each base-model by computing class probabilities;
                build classification probabilities for the base-model;
            end
        end
        update weights of the base-models;
        compute mean likelihood of all the base-models;
        if the computed loss is lesser thanthe preceding epochloss then
            conjoin mean likelihood of the class labels to the feature set;
        else
            store mean likelihood of the present epoch; break for;
        end
    end
    for
        every sample in the test dataset do
        for
            every base-model in the Super Learner do
            get individual class likelihood from the base-model on the test dataset;
            compute class-label likelihood for the base-model;
        end
    end
    get mean likelihood of all base-models;
    classify the speech disfluency labels;
    classify the fluency disorder labels;
    formulate the confusion matrix;
end
```

in validation accuracy in percentage between the proposed technique and the literature [18, 22, 15, 10, 5 and 14] are 10.92, 4.31, 11.56, -0.30, and 3.39 respectively. This shows the proposed method outperforms other methods relatively in validating the samples. The difference in testing accuracy in percentage between the proposed technique and the literature [18, 22, 15, 10, 5, and 14] are 4.94, 2.36, 4.75, -0.10, and 4.27 respectively. This shows that the proposed method outperforms other methods relatively by effective testing of the samples.

TABLE 5. Evaluation of the proposed hybrid model.

Metrics		Phase-I (Speech Disfluency Categorization)		Phase-II (Fluency Disorder Classification)	
		Model Validation	Model Testing	Model Validation	Model Testing
Classification Accuracy		98.4	97.5	99.1	98.3
Cohen's Kappa Coefficient		0.96	0.95	0.97	0.96
F1 Score	Macro	0.98	0.97	0.99	0.98
	Micro	0.98	0.97	0.99	0.98
	Weighted	0.98	0.97	0.99	0.98
Jaccard Similarity Index	Macro	0.96	0.95	0.98	0.97
	Micro	0.96	0.95	0.98	0.97
	Weighted	0.96	0.95	0.98	0.97
Hamming Loss		0.02	0.03	0.02	0.02

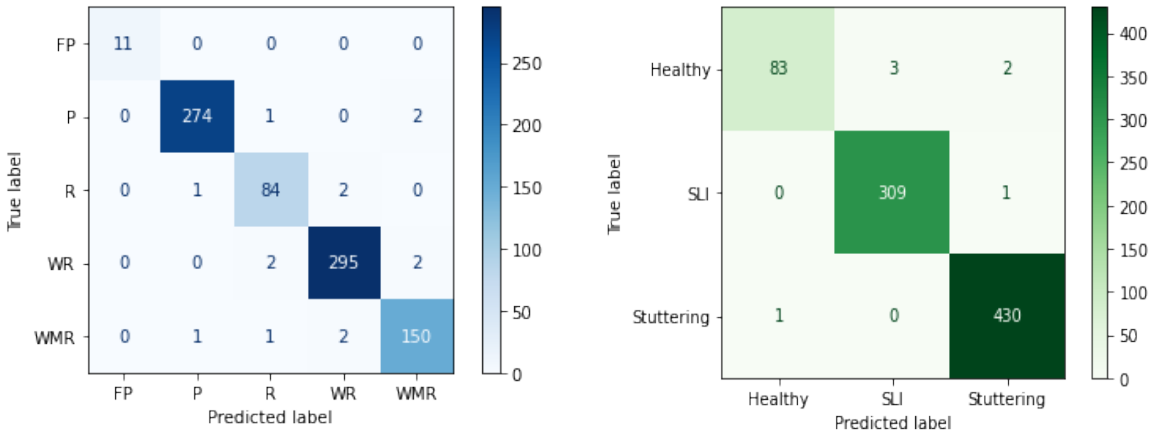
During the *Speech Disfluency Categorization phase,* the hybrid model proved to classify well the classes of filled pause, prolongation, word repetition and word-medial repetition as depicted in Figure 6 (a). There is 100 % accuracy in identifying the filled pause. In identifying the prolongation, almost 97 % of identification accuracy was seen. Since there is a thin line of demarcation between the whole word repetition and word-medial repetition, error rate differed just by a percent in distinguishing them. Also, in 96% of instances, the suggested model was able to identify the revision disfluencies. On the whole, the proposed model was able to identify speech disfluencies in 97.5% of the cases, which reflects the efficacy of the model during the *Speech Disfluency Categorization phase*. Further, during the *Fluency Disorder Analysis phase* of the hybrid model, the classes of Stuttering, SLI and Healthy were categorized as depicted in Figure 6 (b). From Fig. 6 (b), the proposed model was able to identify the *Specific Language Impairment* (SLI) and *stuttering* by 99%. Stuttering is characterized by glitches observed in the flow and fluency of speech. This frequently entails unintentional additions of sounds and phrases, as well as a delay or difficulty continuing steadily through a sentence, and sometimes resembles the SLI. Although labeled as a condition, stuttering can be observed occasionally in healthy subjects' speech as well and is frequently caused by stress or anxiety. In a nutshell, it has been established that the proposed hybrid model outperforms any conventional model in terms of performance and resource utilization in the task of distinguishing disfluencies and demarcating fluency disorders.

V. CONCLUSION

The proposed Pipelined Deep Learner-Dual Classifier model has been built to render a holistic assessment of the fluency of a person. Stuttering and SLI remain the most difficult speech

TABLE 6. Comparison with contemporary research on disfluency classification.

Authors	Model		Dataset	Task	Validation Accuracy (%)	Testing Accuracy (%)
P., Mahesha et al. [18]	K-Nearest Neighbor and Support Vector Machine		UCLASS Dataset	Stuttered Speech Analysis	86.67	93.34
P., Mahesha et al. [22]	DMFCC-Vector Quantization		UCLASS Dataset	Stuttered Speech Analysis	93.10	95.88
F. Afroz et al. [15]	Custom Classification Model		UCLASS stuttered speech corpus	Stuttered Speech Analysis	98	98
Kourkounakis, Tedd, et al. [10]	Custom Residual Convolutional Neural Network		UCLASS Dataset	Speech Disfluency Categorization	91.75	91.75
Pravin, S.C. and Palanivelan, M. [5] [2]	Hybrid Deep Ensemble		UCLASS and Fluency Bank Dataset	Speech Disfluency Categorization	97	98.1
M. Jouaiti and K. Dautenhahn [14]	Custom Neural Network		UCLASS, Fluency Bank, and SEP-28K Dataset [27]	Speech Disfluency Categorization	94	94
This work	Pipelined Deep Learner - Dual Classifier (PDL-DC)	Phase I	Custom Glottal Features Dataset	Speech Disfluency Categorization	97.5	98.4
		Phase II		Fluency Disorder Classification	98.3	99.1



(a) (b)

FIGURE 6. (a) Confusion matrix heatmap of phase i classification: speech disfluency categorization and (b) heatmap of phase ii model: fluency disorder classification.

disorders to distinguish because of a thin line of demarcation between them, which makes it tricky to model using simple algorithms. Thus, the PDL-DC model was designed to precisely classify the stuttering disfluencies from SLI disfluencies. PDL-DC could recognize five forms of speech disfluencies: filled pauses, prolongations, revisions, word repeats and word-medial repetitions. This model employs a hybrid combination of convolutional autoencoder and super learning model for learning efficient temporal correlations and spectral frame-level speech representations from disfluent speech.

To correctly detect the needed influences and significant sections of speech, the proposed model provides state-of-the-art disfluency classification results compared to previous work in the field. As there is no dataset available with instances of disfluencies from subjects exhibiting stuttering and SLI, the authors intended to create a disfluent speech dataset to permit an in-depth study on disfluency identification. Experimentally, the suggested hybrid model demonstrated greater classification accuracy and reduced computational complexity than other traditional ensembles described in the current literature. As a consequence, the proposed model may be used as an accurate fluency assessment tool to distinguish speech disfluencies pronounced by subjects with varied fluency disorders and to analyse the presence of stuttering disorder or a particular language impairment in patients based on the real-time data provided by Clinicians. Further, fluency assessment tests may be carried out rapidly for students which would reveal their proficiency in the language.

In future, the building robust models to assess the intensity of various speech-related disorders will be focused. This would also involve deeper analysis into the physiological bases behind the origin of speech disorders in subjects. Further, an automated speech therapy suggestive aid and AI-enabled speech therapy are also thought-out to be a viable future work.

REFERENCES

[1] S. C. Pravin, V. B. Sivaraman, and J. Saranya, "Deep ensemble models for speech emotion classification," *Microprocessors Microsyst.*, vol. 98, Apr. 2023, Art. no. 104790.

[2] D. Sherman, "Clinical and experimental use of the Iowa scale of severity of stuttering," *J. Speech Hearing Disorders*, vol. 17, no. 3, pp. 316–320, Sep. 1952.

[3] B. Walsh, F. Tian, J. A. Tourville, M. A. Yücel, T. Kuczek, and A. J. Bostianó, "Hemodynamics of speech production: An fNIRS investigation of children who stutter," *Sci. Rep.*, vol. 7, no. 1, p. 4034, Jun. 2017.

[4] A. G. Sares, M. L. D. Deroche, D. M. Shiller, and V. L. Gracco, "Timing variability of sensorimotor integration during vocalization in individuals who stutter," *Sci. Rep.*, vol. 8, no. 1, p. 16340, Nov. 2018.

[5] S. C. Pravin and M. Palanivelan, "A hybrid deep ensemble for speech disfluency classification," *Circuits, Syst., Signal Process.*, vol. 40, no. 8, pp. 3968–3995, Aug. 2021.

[6] S. C. Pravin and M. Palanivelan, "WDSAE-DNDT based speech fluency disorder classification," *Malaysian J. Comput. Sci.*, vol. 35, no. 3, pp. 222–242, Jul. 2022.

[7] S. C. Pravin and M. Palanivelan, "Regularized deep LSTM autoencoder for phonological deviation assessment," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 35, no. 4, Mar. 2021, Art. no. 2152002.

[8] P. Howell, S. Davis, and J. Bartrip, "The University College London archive of stuttered speech (UCLASS)," *J. Speech, Lang., Hearing Res.*, vol. 52, no. 2, pp. 556–569, Apr. 2009.

[9] N. Bernstein Ratner and B. MacWhinney, "Fluency bank: A new resource for fluency research and practice," *J. Fluency Disorders*, vol. 56, pp. 69–80, Jun. 2018.

[10] T. Kourkounakis, A. Hajavi, and A. Etemad, "FluentNet: End-to-end detection of speech disfluency with deep learning," 2020, *arXiv:2009.11394*.

[11] T. Kourkounakis, A. Hajavi, and A. Etemad, "Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6089–6093.

[12] E. Shriberg, R. Bates, and A. Stolcke, "A prosody-only decision-tree model for disfluency detection," in *Proc. Eurospeech*, 1997, pp. 2383–2386.

[13] H. Deng, Y. Lin, T. Utsuro, A. Kobayashi, H. Nishizaki, and J. Hoshino, "Automatic fluency evaluation of spontaneous speech using disfluency-based features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 9239–9243.

[14] M. Jouaiti and K. Dautenhahn, "Dysfluency classification in stuttered speech using deep learning for real-time applications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6482–6486.

[15] F. Afroz and S. G. Koolagudi, "Recognition and classification of pauses in stuttered speech using acoustic features," in *Proc. 6th Int. Conf. Signal Process. Integr. Netw. (SPIN)*, Mar. 2019, pp. 921–926.

[16] T. F. Yap, J. Epps, E. H. C. Choi, and E. Ambikairajah, "Glottal features for speech-based cognitive load classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 5234–5237.

[17] B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah, "Speech-based cognitive load monitoring system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 2041–2044.

[18] P. Mahesha and D. S. Vinod, "Feature based classification of dysfluent and normal speech," in *Proc. 2nd Int. Conf. Comput. Sci., Eng. Inf. Technol.*, New York, NY, USA, Oct. 2012, pp. 594–597.

[19] P. Mahesha, "An approach for classification of dysfluent and fluent speech using K-NN and SVM," *Int. J. Comput. Sci., Eng. Appl.*, vol. 2, no. 6, pp. 23–32, Dec. 2012.

[20] G. Manjula, M. Shivakumar, and Y. V. Geetha, "Adaptive optimization based neural network for classification of stuttered speech," in *Proc. 3rd Int. Conf. Cryptogr., Secur. Privacy*, New York, NY, USA, Jan. 2019, pp. 93–98.

[21] P. Mahesha and D. S. Vinod, "Characterization of stuttering dysfluencies using distinctive prosodic and source features," in *Advances in Ubiquitous Sensing Applications for Healthcare, Classification Techniques for Medical Image Analysis and Computer Aided Diagnosis*, N. Dey, Ed. New York, NY, USA: Academic, 2019, pp. 89–107.

[22] P. Mahesha and D. S. Vinod, "Delta cepstral analysis for classification of repetition and prolongation stuttering dysfluencies," in *Proc. ICISP*, 2014, pp. 408–414.

[23] L. H. Finestack, B. Payesteh, J. R. Disher, and H. M. Julien, "Reporting child language sampling procedures," *J. Speech, Lang., Hearing Res.*, vol. 57, no. 6, pp. 2274–2279, Dec. 2014.

[24] M. S. Seyfioglu, A. M. Özbayoglu, and S. Z. Gürbüz, "Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 4, pp. 1709–1723, Aug. 2018.

[25] S. Goyal and P. K. Bhatia, "Heterogeneous stacked ensemble classifier for software defect prediction," *Multimedia Tools Appl.*, vol. 81, no. 26, pp. 37033–37055, Nov. 2022.

[26] S. Saha, J. Roy, B. Pradhan, and T. K. Hembram, "Hybrid ensemble machine learning approaches for landslide susceptibility mapping using different sampling ratios at east Sikkim himalayan, India," *Adv. Space Res.*, vol. 68, no. 7, pp. 2819–2840, Oct. 2021.

[27] C. Lea, V. Mitra, A. Joshi, S. Kajarekar, and J. P. Bigham, "SEP-28k: A dataset for stuttering event detection from podcasts with people who stutter," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6798–6802.