WDSAE-DNDT BASED SPEECH FLUENCY DISORDER CLASSIFICATION

Sheena Christabel Pravin [1*], ***M. Palanivelan*** [2]

[1]School of Electronics Engineering (SENSE), Vellore Institute of Technology, Chennai, India

[2]Professor, Department of ECE, Rajalakshmi Engineering College, Chennai, India

Email: sheenachristabel.p@vit.ac.in[1*](corresponding author), palanivelan.m@rajalakshmi.edu.in[2]

DOI: https://doi.org/10.22452/mjcs.vol35no3.3

ABSTRACT

In this paper, Weight Decorrelated Stacked Autoencoder-Deep Neural Decision Trees (WDSAE-DNDT), a novel hybrid model is proposed for automating the assessment of children's speech fluency disorders by discerning their disfluencies. In fluency disorder classification, it is imperative to know how each feature contributes to the disorder classification rather than the diagnosis itself and so the depth modified DNDT acts as the best discriminator since it is interpretable by its very nature. The WDSAE presents DNDT with a high-level latent representation of the disfluent speech. A fusion feature vector was built by combining the prosodic cues from disfluent speech segments combined with the WDSAE-based Bottleneck features. The proposed hybrid model was compared with the performance of the experimented baseline models. Further analysis was carried out to check the impact of tree cut points for each feature and epochs on the accuracy of prediction of the hybrid model. The proposed hybrid model when trained on the fusion feature set has shown appreciable improvement in the area under the Receiver Operating Characteristics (ROC) curve, classification accuracy, Kappa statistical value, and Jaccard similarity index. The WDSAE-DNDT demonstrates high precision than the baseline models in setting clinical benchmark to distinguish subjects with dysphemia from those with Specific Language Impairment.

Keywords: WDSAE-DNDT, Speech Fluency Disorder, Bottle Neck Features, Dysphemia, Specific Language Impairment

1.0 INTRODUCTION

Fluency disorder is prevalent among children during their language-learning phase, often undermined and neglected by parents and care-givers, which may lead to persistent speech disorders as the child grows into an adolescent. Also, demarcating natural childhood disfluencies due to bilingualism from stuttering disfluencies has for decades posed a major challenge in the diagnosis of speech impairment to the Speech and Language Pathologists (SLPs) [1]. Automatic assessment of fluency disorders is a difficult task as the subjects have different speaking rate, accents and family background. Disfluencies in speech highlight the inherent fluency disorder in the child. Dysphemia is one such fluency disorder characterized by stuttering where repetitions, pauses, or drawn out syllables, words, and phrases disrupt the smooth flow of speech [2]. It is due to acute irregularity co-ordination between the nervous system and the speech production system. They are symbolized by irregular breathing patterns and reflected as perceivable interruptions in speech utterances. There are both physiological and psychological elucidations for the occurrence of disfluencies [3-5]. Disfluencies in children with dysphemia are accompanied frequently by anxiety during communication, significant variation in breathing, hyper-tension, vocalization, articulation, rhythm, speaking rate and accent of speech [6] with physiological struggle in production of certain phonemes. It is also characterized by excessive repetition of words and a sharp hesitation to communicate. On the other hand, Specific Language Impairment (SLI), dominantly observed in multi-lingual children, is yet another fluency disorder that occurs during the (second) language-development process of a child [7]. SLI, characterized by a language delay, affects the reading, writing, speaking and listening skills of children during their language acquisition phase. SLI, when untreated, leads to learning disability in children. Children with SLI do exhibit disruptions of speech sounds which are manifested as repetitions, prolongations, blocks and dysrhythmic phonation patterns but the nature of disfluencies in dysphemic speakers subtly vary from those with SLI. It is indeed essential to demarcate SLI from dysphemia for early therapeutic intervention. The course of demarcation being a tedious manual process for the SLPs can be automized through machine learning and deep learning techniques. Estimation of speech parameters can assess the health of one's voice but the precision of these parameters in discerning a specific speech disorder is related to competent machine learning algorithms with sharp discrimination accuracy. Fluency disorder classification thus involves the study of speech parameters governing the disorders and is also oriented to the application of efficient classification techniques. Fluency disorder predictions, though weighed to be complex, aid in

the prognosis of specific speech disorders [8]. Early detection of fluency disorders is inevitable in distinguishing dysphemia from Specific Language Impairment (SLI). In this work, both temporal and spectral prosodic features were estimated from the disfluent speech. Through regress experiments, it was observed that the energy, duration and fundamental frequency-related speech parameters are highly sensitive in distinguishing speech disorders.

The conventional machine learning classifiers like the k-means clustering, Gaussian Mixture Model (GMM) [9], Support Vector Machines (SVM) [10], Decision Trees, Bayesian Classifier [11], XGBoost classifier [12], Artificial Neural Networks (ANNs) [13-15] have been widely used for speech disorder classification. With machine learning based disease predictions becoming pervasive and impacting many aspects of our day-to-day lives [16-17], the focal point of research has moved beyond model performance to vital factors such as interpretability [18]. The deep learning models with hierarchical layers, including the Long Short-Term Memory (LSTMs) were applied for speech disorder classification [19]. In recent times, hybrid machine learning algorithms have attracted research interests in speech disorder classification as they are powerful enough to surpass the conventional models in terms of accuracy of classification. Hierarchical training models have shown to improve performance in terms of accuracy and precision, for instance the Convolutional Deep Belief Networks connected with the Convolutional Neural Networks (CNNs) was proposed for automatic detection of voice disorders [20] while the SVM model was combined with Sequential Halving And Classification (SHAC) framework which apparently performed better than the Long Short-Term Memory-Fully Convolutional Network (LSTM-FCN), a hybrid classifier in distinguishing one speech disorder from another [12]. The main contributions of this paper are as follows:

- A prosody-based fluency evaluation of children was carried out to distinguish subjects with dysphemia from those with specific language impairment. Profound speech analysis revealed that a few prosody features including the measure of mean squared error of the estimated energy contour with a 1-degree polynomial which sharply indicated the rise and fall of energy levels in the speech utterances. Also, a measure of the regression coefficient between F0 contour and a linear regression line helped deciphering pitch deviations in children with dysphemia.

- Weight decorrelation was introduced in the Stacked Autoencoder to facilitate better feature learning and dimensionality reduction. The agglomeration of the weight decorrelated stacked autoencoder with the DNDT model, which was depth-modified yielded the best accuracy of classification even with the available sparse data than the competing models in the literature and the other baseline models.

Though the key advantage of deep learning models is automatic higher-level feature extraction, they demand hours of speech training data for accurate classification. Due to the unavailability of such huge on-demand data, alternatively, the extraction of prosodic features was given precedence as they were found significant in distinguishing fluency disorders.

From this point forward, the paper is organized into five sections. In section 2, related work in the existing literature is highlighted. In section 3, the disfluent speech corpus and the prosodic feature extraction from disfluent speech are explained along with the architecture of the proposed WDSAE model for feature-dimensionality reduction. It also includes a comprehensive description of the proposed novel hybrid WDSAE-DNDT model with the fusion feature vector. Experimental results are presented in Section 4 while section 5 concludes the observations elucidating the prospective future work.

2.0 RELATED WORK

The prevailing literature holds information on the study of disfluent speech features [21-24] but very less content on assessment of fluency disorders using machine learning models. An early work on distinguishing childhood disfluencies from fluency disorders using the Artificial Neural Networks, trained on the Back-Propagation algorithm was carried out by [25] and hierarchical Self-Organising Map (SOMs) were introduced by [15] to classify fluent utterances from non-fluent utterances to diagnose speech fluency disorders. The first SOM facilitates dimensionality reduction of the perceptual features derived from the 1/3-octave filters, which follows classification by the second SOM. Being an unsupervised model, the SOM yields low classification accuracy of 76.67%. The Radial Basis Function (RBF) and the MLP networks [14] trained on the speech parameters such as the centre frequencies between 100 Hz and 10 kHz presented a competitive accuracy of 88.1% and 94.9% respectively. Yet they report that the RBF classifier exhibited less generalization ability in classifying unknown data. Deep Neural Network-Hidden Markov Model (DNN-HMM) based acoustic models [26] when trained on data as large as 175 hours of spoken speech yielded a reduced Word Error Rate (WER) of 21.2 % for adult fluency assessment and 12% WER for child fluency assessment. Bi-directional LSTM-based fluency assessment and scoring model [27] with feed forward

attention layer could produce a Pearson correlation of 0.602 between the machine-rated and human-rated fluency score to conclude on a suitable therapy for subjects detected with fluency disorders. However, the results of the fluency score is neither cross-validated nor a thorough statistical evaluation of the model was carried out except the Pearson's correlation to prove the accuracy of the model. Recently, a deep learning framework with feed-forward layer and multi-head attention layer appended with audio encoding was proposed by [28] for fast-screening of language-based fluency disorder in children yielding an overall accuracy of 76.1%. Both Acoustic features, along with the transcriptional features from the Automatic Speech Recognizer (ASR) have been embedded into the training set of the fast-screening deep learning framework. However, the word error rate of the ASR has not been disclosed, which is a significant metric to understand the precision of the transcriptional features. The classifiers used in the existing literature, along with the language, dataset size and the performance metrics are presented in Table 1.

Table 1: Classifiers in the Existing Literature

Authors	Year	Classifiers	Speech Features	Language	Size of the Dataset	Evaluation Metrics
Zhang, X., et. al. [28]	2020	Deep learning framework with feed-forward layer and multi-head attention layer	Initial consonant, Tone, Vowel Syllable count, Speech speed, Pronunciation duration, Content restatement/replication, Redundant articles, Pause count Pause duration, The wrong usage of grammar, Keywords missing, Information organization	English	2200 audio samples	Accuracy:76.1%
Chen, L., et. al. [27]	2018	Bidirectional Long-Short Term Memory	Fluency, Rhythm, Intonation, Stress, Pronunciation, Grammar, Vocabulary Use	English	5488 spoken responses	Pearson's Coefficient:0.602
Cheng, J. et. al. [26]	2015	DNN-HMM	Mel Frequency Cepstral Coefficients, Mel Scale Filter bank	Adult English dataset	175.2 hours data	WER:21.2%
				Child English dataset	227.2 hours data	WER:14.5%
				Adult Chinese dataset	168.7 hours data	WER:12.0% t
Świetlicka et al [14]	2009	RBF	1/3-octave filter bank features	Polish speech dataset	118 data samples	Accuracy:88.1%
		MLP				Accuracy:94.9%
Szczurowska et. al. [15]	2009	Kohonen networks	Sound intensity levels from twenty one 1/3 octave digital filters	Polish speech dataset	110 speech utterances	Accuracy:76.67%
Geetha et al. [25]	2000	Artificial Neural Networks	Disfluency types, frequency and duration of disfluencies, secondary behaviours, Speech rate, Stuttering Severity Index scores	English	Speech samples from 51 children	Accuracy: 92%

Grounded on the existing literature, the machine learning and deep learning models do well in classifying fluency disorders. However, the lack of natural interpretability of each speech feature and its significance in distinguishing one fluency disorder from the rest is a serious drawback in these models, which mask the decision flow process at the outset of fluency disorder classification, where it is often more important to know how each feature contributes to the prediction rather than the conclusion itself. The Deep Neural Decision trees (DNDT) [29] with modified depth have a clear advantage in this aspect, as one can easily follow the progression of the classification course being a tree and know exactly how a prediction is being made.

Conventional speech datasets for Specific Language Impairment and Dysphemia in the literature include the University College London's. Archive of Stuttered Speech (UCLASS) dataset [41] and the Laboratory of Artificial Neural Network Applications (LANNA) dataset [42]. The LANNA dataset contains two subgroups of recordings of children's speech from different types of speakers. The first subgroup (healthy) consists of recordings of children without speech disorders; the second subgroup (patients) consists of recordings of children with SLI. The UCLASS dataset is a collection of spontaneous stuttered or dysphemic speech database in English language. The UCLASS Release 1 dataset consists of conversational speech from 18 female subjects and 120 male subjects. In this research work, the proposed model was trained on the speech features extracted from disfluent speech samples from bilingual children.

3.0 METHOD

This research work aims to automate the detection of speech fluency disorders in children by proposing a Weight Decorrelated Stacked Autoencoder (WDSAE) conglomerated with the depth-modified Deep Neural Decision Trees. The WDSAE-based high-level feature extraction generates deep latent representations of the speech disfluency features in combination with the prosodic features from the disfluent speech corpus. The features are then presented to the DNDT model which in turn classifies fluency disorders with high precision and accuracy. Since the Deep Neural Decision Trees are interpretable, the contribution of each feature towards the classification of fluency disorders is intricately tracked.

3.1 Speech Corpus

The disfluent speech corpus for fluency disorder classification was built by collecting speech samples from 27 bilingual children (14 boys and 13 girls), ranging in the age between 5 and 7 years with a mean and standard deviation of the subjects' age being 5.7 and 0.98 respectively. Tamil-English speaking bilingual children (subjects) were asked to produce spontaneous speech on a minimum of 5 topics they chose from the visual stimuli passed down to them. Fluent children were categorized as healthy subjects based on the recommendation of the Speech Language Pathologist. Based on the manual investigation and disfluency analysis on the spontaneous speech of the subjects by the Speech Language Pathologist, 7 subjects were diagnosed with SLI and 5 speakers were pronounced with dysphemia with an average age of 6 and the rest were declared healthy. Spontaneous speech samples were recorded for a total of 125 minutes; around 4.5 minutes per subject with one recording per subject. The sampling rate was set to 16 KHz. Voice samples were digitalized into a tablet for perceptual and spectral analyses of selected parameters. The disfluent speech segments were manually annotated using PRAAT tool [30] by following the annotation style of [31]. The disfluencies exhibited by the healthy, SLI and dysphemic children were annotated under their respective labels. A total of 944 disfluencies were annotated at the disfluency boundaries followed by prosodic speech feature extraction [32].

3.2 Feature Extraction and Analysis

Prosody-based speech features viz. energy, duration and pitch related 38 speech features as displayed in Table 2 were extracted from the speech disfluencies to evaluate the subjects, as the prosodic features play a dominant role in producing fluent speech. The manual investigation and disfluency analysis on the spontaneous speech of the subjects were conducted using the Stuttering Severity Index–4 tool at the clinic of the Speech Language Pathologist.

Table 2: Prosodic Features Extracted from Disfluent Speech

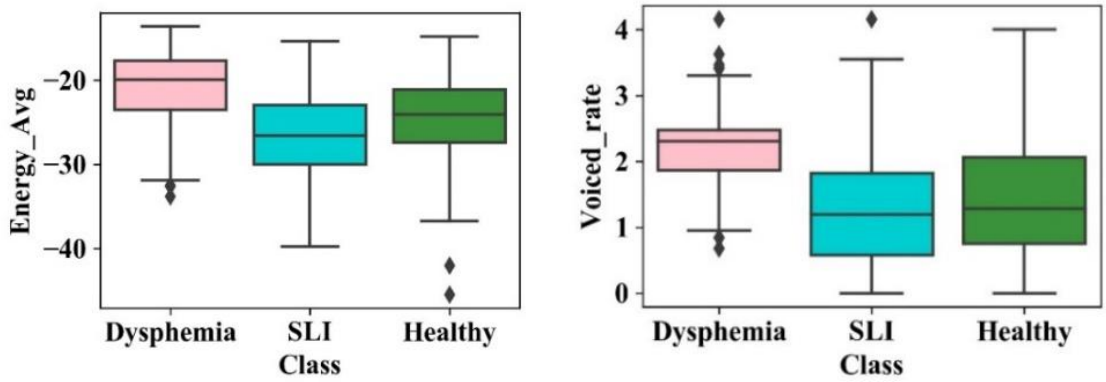
Prosodic Features		Description
Energy-related Features	**Voiced_Energy_Avg**	Average Energy of the voiced speech segments
	Voiced_Energy_Std	Standard Deviation of Energy of the voiced segments
	Voiced_Energy_Max	Maximum of Energy of the voiced segments
	Voiced Rate	Number of voiced segment count per second
	UnVoiced_energy_Reg	Unvoiced energy Regularity
	Voiced_energy_Reg	Regularity of Voiced energy
	Tilt Energy Contour	Average tilt of energy contour
	Reg. Coeff.	Regression coefficient between the energy contour and a linear regression
	Delta Energy Avg	Average Delta energy within voiced segments
	Delta Energy Std	Standard deviation of Delta energy within voiced segments
	MSE_Energy	Mean square error of the reconstructed energy contour
Duration-related Features	**Voiced_duration_Avg**	Average duration of voiced speech segments
	Voiced_duration_std	Standard deviation of voiced speech segments
	Pause Rate	Number of pauses counts per second
	Pause_duration_Avg	Average duration of pause
	Pause_duration_Std	Standard deviation of the duration of pauses
	Silence Duration	(Silence duration)/(Voiced duration + Unvoiced durations)
	Voiced_to_Unvoiced Ratio	(Voiced duration)/(Unvoiced durations)
	UnVoiced duration	(Unvoiced duration)/(Voiced + Unvoiced durations)
	Voiced duration	(Voiced duration)/(Voiced + Unvoiced durations)
	Voiced_sil_dur Ratio	(Voiced duration)/(Silence durations)
	UnVoiced_sil_dur Ratio	(Unvoiced duration)/(Silence durations)
	UnVoiced_dur_reg	Unvoiced duration Regularity
	Voiced_dur_reg	Regularity of Voiced duration
	PauseDur_reg	Regularity of Pause duration
	Duration_Voiced_max	Maximum duration of voiced segments
	Duration_unVoiced_max	Maximum duration of unvoiced segments
	Duration_Voiced_min	Minimum duration of voiced segments
	Duration_unVoiced_min	Minimum duration of unvoiced segments
	VUV Rate	rate (# of voiced segments) / (# of unvoiced segments)
Fundamental Frequency-related Features	**F_0_avg**	Average fundamental frequency in voiced segments
	F_0_std	Standard deviation of fundamental frequency in Hz
	F_0_var	Variability of F_0 in Hz
	F_0_max	Maximum of the fundamental frequency in Hz
	F_0_Avg_tilt	Average tilt of fundamental frequency
	F_0_tilt_regularity	Tilt regularity of fundamental frequency
	MSE_F_0	Mean square error of the reconstructed F_0
	Reg_Coeff_F_0_LR	Regression coefficient between the F_0 contour and a linear regression

The average energy estimates the intensity of glottal excitation. The disfluent segments from the dysphemic subjects score high on average energy as revealed in Fig. 1(a). A total of 11 energy-related prosodic features were extracted

including the average energy, standard deviation and maximum value of the voiced segments. The regularity of the voiced/unvoiced energy was measured by computing the standard deviation of energy, related to the fluctuations of energy from the mean energy level of the speech segments. The mean squared error of the estimated energy contour with a 1-degree polynomial was scored to capture the rise and fall of energy levels.

The durational prosodic features give the physical interpretation of the speech signal conventionally measured from the pause rate, pause duration, average duration of the voiced components, rate of unvoiced components, maximum and minimum duration of unvoiced components and the like. The pause rate was calculated over the recurrences of silence intervals measured with a threshold upon the logarithmic value of energy of the consecutive speech frames. The rate of unvoiced segments and speech rate are vital measures that distinguish fluency disorders. The ratio of Voiced to Unvoiced (VUV) components was measured using the equation (1).

$$VUV = \frac{No.of\ voiced\ speech\ segments}{No.of\ unvoiced\ speech\ segments} \tag{1}$$



(a) (b)

Fig.. 1: (a) Average Energy, (b) Voiced Speech rate of Dysphemic subjects compared with the SLI and Healthy subjects

The increase in the speech rate is an undermining factor of the dysphemic speech production system since the speech parameters at higher speaking rate demarcates the subjects with SLI and those with dysphemia as shown in **Fig. 1(b).** A total of 19 durational features were extracted for analysing the durational anomalies in different fluency disorders.

The spectral features related to pitch or fundamental frequency (F_0) describes the perceptual properties of the speech like the maximum frequency each child can reach. They also represent the intonational precedents of the subjects. The mean fundamental frequency was measured along with its maximum value and other functionals namely the standard deviation and variance. Additionally, the F_0 tilt was measured, which is a measure of the tilt of pitch accent by perceiving the rise and fall of spectral energy in the fundamental frequency contour as given in equation (2). A total of 8 pitch-related features were extracted from the speech segments.

$$F_0\ \text{tilt} = \frac{|E_{rise}| - |E_{fall}|}{|E_{rise}| + |E_{fall}|} \tag{2}$$

where, E_{rise} and E_{fall} indicate the rise and fall in spectral energy. Also, a measure of regression coefficient between the F0 contour and a linear regression line helped deciphering pitch deviations in the subjects.

3.3 WEIGHT-DECORRELATED STACKED AUTOENCODER

The Weight-Decorrelated Stacked Autoencoder (WDSAE) proposed as a feature compressor network, institute a deep learning algorithm in which a number of autoencoders are stacked one upon another with hierarchical weight decorrelation to realise intricate features. Autoencoder is an exceptional type of neural network architecture used to reconstruct the input.

3.3.1 Training the WDSAE

Autoencoders when stacked in stages learn higher level features from the input data [33]. These feature representations are then decoded and reconstructed to display the actual data. The Autoencoder (AE) has a bottleneck layer with meagre neurons as the intermediate layer, forcing them to create compressed representations of the input that can be used by the decoder to reproduce the original input. In this work, the WDSAE architecture was built with three autoencoders stacked one upon another as shown in **Fig. 2**. The autoencoders were regularised with sequential 20% dropout on the neural units which introduces weight decorrelation and avoids over-fitting. Thus, the WDSAE learns deep features from the tempo-spectral prosodic speech parameters.

Each AE in the WDSAE architecture was built with one input layer, one hidden layer and one output layer. The BottleNeck layer of the AEs was built with fewer neurons than the input layer to facilitate feature compression. The first AE is trained with the input prosodic feature vector and the succeeding autoencoders were trained from the output of the previous AE. The first autoencoder in the WDSAE has 'n' input units and maps the input feature vector x $\in R^n$. The reconstructed input was fed to the stacked autoencoders in tandem with sequentially reduced number of units facilitated by dropout. The dropout on the unit 'm' in the hidden layer 'H' is described as given in equation (3).

$$V_m^H = \theta(\sum_{l<n} \sum_k \nabla_k^l \left(W_t^{(1)}\right)_{mk} I_k^l) \quad (3)$$

∇_k^l is the selector variable for dropout following Bernoulli's distribution; I_k^l being the input to the final autoencoder in the WDSAE architecture. The hidden latent vector 'h' is extracted from the BottleNeck layer 'H' of the last autoencoder with the trained encoder whose weight matrix is $Wt^{(1)}$ and bias vector $B^{(1)}$as given in equation (4).

$$h = \theta \left(\left(W_t^{(1)}\right)_{mk} I_k + B^{(1)}\right) \quad (4)$$

$'\theta'$ is the non-linear hyperbolic tangent activation function as defined in equation (5).

$$\theta(x) = \frac{1-e^{-2x}}{1+e^{-2x}} \quad (5)$$

Then, the input vector was decoded back from the hidden vector to produce a reconstructed vector $\tilde{x}$ using the Decoder's learned weight matrix $Wt^{(2)}$ and bias vector $B^{(2)}$ as given in equation (6).

$$\theta \left(Wt^{(2)}h + B^{(2)}\right) \quad (6)$$

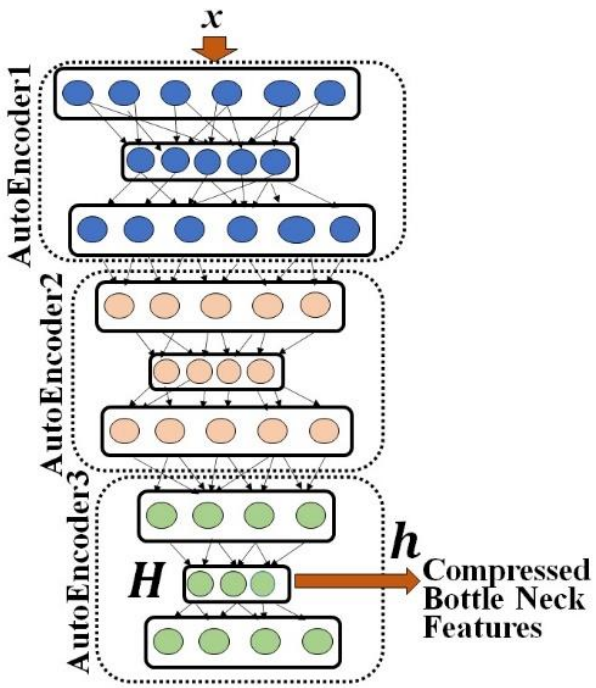


Fig 2: Weight-Decorrelated Stacked Autoencoder

The learning parameters of the AE now comprises of ($Wt^{(1)}$, $Wt^{(2)}$ $B^{(1)}$, $B^{(2)}$).

3.3.2 WDSAE-based Feature Reconstruction

The goal of training the WDSAE is to keep the loss function minimum. It is measured between the input and the reconstructed feature vector as cited in equation (7).

$$arg\ \min_{W_t^1, W_t^2, B^{(1)}, B^{(2)}} [L(x, \tilde{x})] \tag{7}$$

The 'Categorical Cross Entropy' [40] defined in equation (8) is the loss function chosen for this task as it involves multi-class classification.

$$L(x, \tilde{x}) = \frac{-1}{N} \sum_{i=1}^{N} \sum_{k=1}^{n} x_{ik} \log(\tilde{x}_{ik}) + (1 - x_{ik}) \log(1 - \tilde{x}_{ik}) \tag{8}$$

where $\tilde{x}$ is the reconstructed feature vector, 'N' is the size of the mini-batch input vector since loss is measured on a mini-batch of inputs and 'n' is the size of the input feature vector. The weights of the autoencoders are updated by following the updating formulae specified in equations (9)-(12).

$$\left(W_t^{(1)}\right)_{new} = \left(W_t^{(1)}\right)_{old} - \propto \frac{\partial\ L(x,\tilde{x})}{\partial\ W} \tag{9}$$

$$\left(W_t^{(2)}\right)_{new} = \left(W_t^{(2)}\right)_{old} - \propto \frac{\partial\ L(x,\tilde{x})}{\partial\ W} \tag{10}$$

$$B^{(1)}{}_{new} = B^{(1)}{}_{old} - \propto \frac{\partial\ L(x,\tilde{x})}{\partial\ B^{(1)}} \tag{11}$$

$$B^{(2)}{}_{new} = B^{(2)}{}_{old} - \propto \frac{\partial\ L(x,\tilde{x})}{\partial\ B^{(1)}} \tag{12}$$

where, $\propto$ is the learning rate

The partial derivatives of the loss function over the learning parameters ($Wt^{(1)}$, $Wt^{(2)}$ $B^{(1)}$, $B^{(2)}$) are detailed in the equations (13)-(16).

$$\frac{\partial\ L(x,\tilde{x})}{\partial\left(W_t^{(1)}\right)_{jl}} = \frac{-1}{N} \sum_{i=1}^{N} \left\{ \sum_{k=1}^{n} \left[\frac{x_{ik} - \tilde{x}_{ik}}{\tilde{x}_{ik}(1-\tilde{x}_{ik})} \theta'\left(\delta_{ik}^{\tilde{x}}\right) W_{kl} \theta'(\delta_{il}^{h})\, x_{ij} \right] + \theta'(\delta_{ik}^{\tilde{x}}) \theta(\delta_{il}^{h}) \right\} \tag{13}$$

$$\frac{\partial\ L(x,\tilde{x})}{\partial B_j^{(1)}} = \frac{-1}{N} \sum_{i=1}^{N} \sum_{k=1}^{n} \left[\frac{x_{ik} - \tilde{x}_{ik}}{\tilde{x}_{ik}(1-\tilde{x}_{ik})} \theta'\left(\delta_{ik}^{\tilde{x}}\right) W_{jl} \theta'(\delta_{ij}^{h}) \right] \tag{14}$$

$$\frac{\partial\ L(x,\tilde{x})}{\partial\left(W_t^{(2)}\right)_{jl}} = \frac{-1}{N} \sum_{i=1}^{N} \left\{ \sum_{k=1}^{n} \theta'(\delta_{ik}^{\tilde{x}}) \theta(\delta_{il}^{h}) \right\} \tag{15}$$

$$\frac{\partial\ L(x,\tilde{x})}{\partial B_j^{(2)}} = \frac{-1}{N} \sum_{i=1}^{N} \sum_{k=1}^{n} \left[\frac{x_{ik} - \tilde{x}_{ik}}{\tilde{x}_{ik}(1-\tilde{x}_{ik})}\ \theta'(\delta_{ik}^{\tilde{x}}) \right] \tag{16}$$

where, W_{jl} is the weight relating the j^{th} input and the l^{th} hidden unit; $B_j^{(1)}$ being the bias of the j^{th} unit in the hidden layer. When the WDSAE model was able to reconstruct the actual input impeccably from the hidden vector 'h', subsequent to the training phase, the reconstruction layer of the final AE was discarded and the Bottle Neck feature vector 'h' was extracted. The feature vector, 'h' holds adequate knowledge on the input and is a deeper latent representation of the input 'x'. Stacking the autoencoders with dropout has an agile and better minimization of the loss function and hierarchical training progressively extracts deeper and higher-level features from the input. The Adadelta optimizer [34] was used to update the weight parameters for quantitative minimization of the loss function. Notably, the Adadelta optimizer adopts a global dynamic learning rate with minimal computation when applied over the Stochastic Gradient Descent algorithm. Before training the model, the 'max absolute scaling' was introduced on the features. WDSAE was constructed using the TensorFlow library [35] using the Python 3 programming language that supports the distributed architecture to extract compressed Bottle Neck Features (BNFs) in an unsupervised fashion The BNFs can be looked upon as dimensionally reduced input features with nonlinear transformation.

3.4 Proposed WDSAE-DNDT Model for Fluency Disorder Classification

In the proposed deep-learning framework, the Weight Decorrelated Stacked Autoencoder (WDSAE) is integrated with the DNDT model to form the hybrid WDSAE-DNDT model. The whole process flow is depicted in **Fig. 3**. In this schema, the prosodic features extracted from the disfluent speech segments were primarily taken into

consideration, since they held the most essential cues for differentiating disfluent segments of the dysphemic subjects from the healthy and SLI subjects. These features were then concatenated to the latent features given by the WDSAE based BNFs, forming a fusion feature set comprising of both prosodic and latent information which serves to yield high classification accuracy as verified in section 4.0. Following the learning phase of the WDSAE-DNDT model, the fluency disorder class labels were eventually acquired across each test sample.

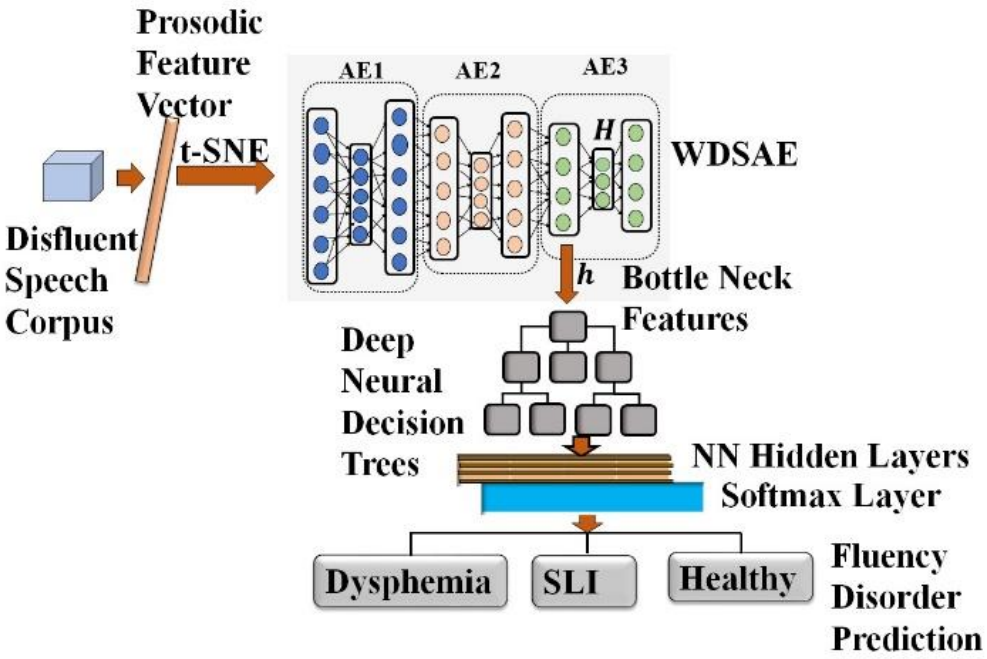


Fig. 3: Proposed WDSAE-DNDT Model

3.4.1 Learning and Classification by the WDSAE-DNDT Model

Parameter learning happens in DNDT by back-propagation using the Stochastic Gradient Descent (SGD) algorithm which allows synchronised parameter learning as against the conventional Greedy algorithm [36] used by the decision trees which involves complicated greedy splitting and was found to be sub-optimal. The DNDT additionally performs an inherent feature selection which was observed by running the model a number of times to observe the features it ignored in the trailing trials when the cut points for the specific features were removed. The DNDT was thus trained on mini-batches using the SGD algorithm with 2000 training epochs, measuring the cross-entropy loss at each iteration. The Neural Net (NN) was built with four hidden layers with 50 neurons in each layer and a softmax activation function was arrayed at the output layer of the NN as shown in equation (17).

$$f(x) = \text{softmax}((\text{v } x + \text{b})/\tau) \quad (17)$$

where, v is a constant given by [1,2,3,…..k+1] , τ is a temperature parameter that governs the randomness of classification and approximates the one-hot vector; the parameter 'b' is defined in terms of tree cut points as shown in equation (18)

$$\text{b} = [0, -\gamma_1, -\gamma_1 - \gamma_2, \ldots, -\gamma_1 - \gamma_2 - \ldots - \gamma_k] \quad (18)$$

The temperature hyper-parameter was set to 0.1 to ensure the model's sensitivity to low probable features. The differentiable soft binning function is thus leveraged to split the tree nodes into multiple leaves and not restricted to binary splits. This soft binning function takes in the feature 'x' and gives out the index of the bins ranging from 0 to 'k', to which the feature 'x' belongs to. Therefore, it is essential to have 'k' cut points. Monotonically increasing cut points $[\gamma_1, \gamma_2, \gamma_{3,\ldots}\gamma_k]$ were used as denoted in equation (19).

$$[\gamma_1 < \gamma_2 < \gamma_{3<\ldots}\gamma_k] \quad (19)$$

The features were binned, each to a Neural Net to find the index of the leaf node from where the feature 'x' originated. Given the binning function, the decision tree was built through a Kronecker product of each feature, binned to its NN to discover the final nodes given by equation (20). Assuming an input feature vector, $x \epsilon R^u$, where 'u' is the dimension of the new fusion feature vector formed by combining the prosodic features and BNFs.

$$y_i = \text{f}_1(\text{x}_1) \otimes \text{f}_2(\text{x}_2) \otimes \ldots \text{f}_\text{L}(\text{x}_\text{u}) \quad (20)$$

Each leaf at the final node is assumed to behave as a linear classifier that classifies the test instances into any of the three classes by majority voting.

3.4.2 Interpretability of DNDT

The Deep Neural Decision Trees (DNDT) model is interpretable by its very nature, as it is a tree. **Fig. 4** explains the intermediate branching decisions in a tree with the Gini index computed at each node. The split at each node of the tree is optimized by the speech dataset fed into it. The 'value' vector in individual nodes of the tree expresses the total number of observations that were categorized into that node from each of the three classification labels. With many features contributing in decision-making, it is imperative to resolve the significance and relevance of each of the features. Thus, the best relevant feature is placed at the root node which is traversed down by splitting the nodes.

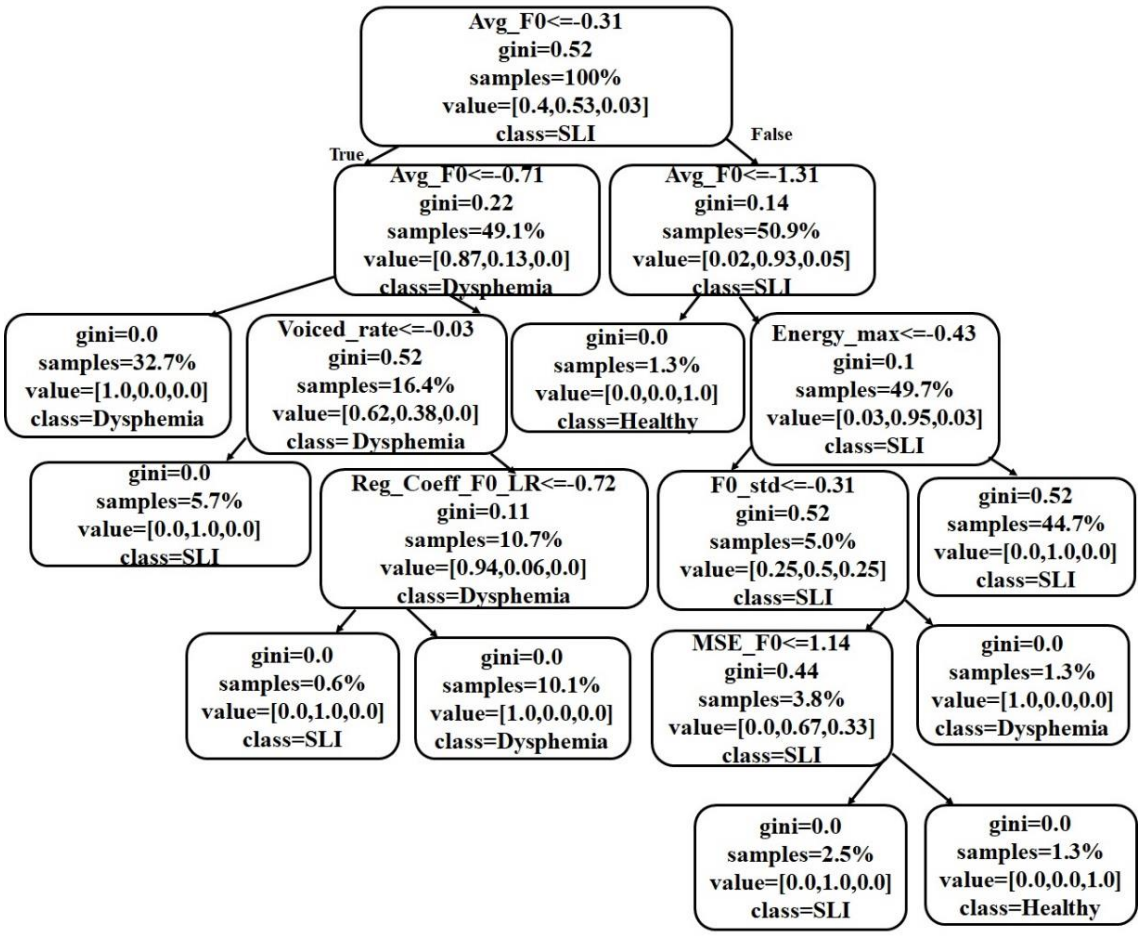


Fig.4: Visualization of a DNDT implemented as a conventional Decision Tree, trained on Prosodic features

Down the tree, the value of uncertainty is measured which is the level of impurity stated by the Gini index given in equation (21).

$$Gini\ index = 1 - \sum_{j=1}^{M} (p_j)^2 \tag{21}$$

It is observed that the impurity level at the branching nodes decreases and tends to a null value at the final nodes leading to an improved classification which implies the presence of best split at each node. The feature 'Avg_F0', which is the mean value of the pitch, was able to distinguish healthy subjects from those with Specific Language Impairment (SLI). A value of voiced rate less than -0.03 points to Dysphemic subjects while variability in F_0 less than 1.14 indicates SLI. The proposed hybrid WDSAE-DNDT model for fluency disorder classification is illustrated by a pseudocode given under Algorithm 1.

Algorithm 1: Pseudocode for the proposed WDSAE-DNDT Model

start
  initialize mini-batch size N, training epochs, learning rate, number of Autoencoders 'L', number of neurons in the hidden layers of AE m[A], input feature vector size n, the number of classes C
for each layer in the WDSAE given as **A** (1<A<L):
  Build an Autoencoder with D input units and d hidden units
  if A is the first layer of WDSAE
    D=n
  else
    D=m[A-1]
  end
  Set the number of output units of the AE = **D**
  initialize AE weights **W** and Biases $\mathbf{B^{(1)}}$, $\mathbf{B^{(2)}}$ to zero
  for each training epoch
    for each mini-batch of data
      Compute Feature reconstruction:
      $\tilde{\mathbf{x}}=\boldsymbol{\theta}(\mathbf{W}.\ \boldsymbol{\theta}(\mathbf{W}.\mathbf{x}+\mathbf{B}^{(1)})+\mathbf{B}^{(2)})$
      Compute the Loss function:
      $\boldsymbol{L}(\mathbf{x},\tilde{\mathbf{x}})=\frac{-1}{\mathbf{N}}\sum_{\mathbf{i=1}}^{\mathbf{N}}\sum_{\mathbf{k=1}}^{\mathbf{n}}\mathbf{x_{ik}}\,\mathbf{log}(\tilde{\mathbf{x}}_{\mathbf{ik}})+\mathbf{1}-\mathbf{x_{ik}})\,\mathbf{log}(\mathbf{1}-\tilde{\mathbf{x}}_{\mathbf{ik}})$
      Update AE Weights
    end
  end
  Remove the output layer
end
append prosodic cues with Bottleneck features
initialise the Neural Net (NN) in DNDT model with 'K' input neurons, 'M' hidden layers
initialise Decision Tree with 'gini' criterion and 'best' splitter
for each epoch:
  Build the NN with the softmax activation function:
  $\boldsymbol{f}(\boldsymbol{x})$= **softmax((v** $\boldsymbol{x}$ **+ b)/**$\boldsymbol{\tau}$**)**
  Furnish the NN with mini-batch of 'u' dimensional fusion features
  Find the final leaf nodes '$\boldsymbol{y_i}$' through a Kronecker product of each feature binned to its NN:
    $\boldsymbol{y_i}=\mathbf{f_1}(\mathbf{x_1})\otimes\mathbf{f_2}(\mathbf{x_2})\otimes\ldots.\,\mathbf{f_L}(\mathbf{x}_u)$
end
Classify the test data into any of the C classes at the final nodes $\boldsymbol{y_i}$ by majority voting
end

3.5 Evaluation Method

The proposed WDSAE-DNDT model was evaluated over various metrics including precision, recall, F-score, weighted average precision, macro-average precision, mean validation accuracy and mean test accuracy. Also, a 5-fold cross-validation report on the performance of the proposed WDSAE-DNDT model was generated over five parallel sets of data to assess the best part of the data on which the proposed model performed well.

4.0 RESULTS

For comparison with other standard classification models, the baseline experimentation was carried out first with the DNDT model trained on disfluent speech prosodic features, then with the DNN and MLP.

4.1 Comparison with DNDT model Trained on Prosodic Cues

The stand-alone Deep Neural Decision Trees (DNDT) model [27] was trained on the prosodic feature set comprising of duration, energy and fundamental frequency-based speech features. The prosodic feature vector was fed directly to the DNDT model trained with 2000 epochs with the same feature settings, enabling the relevant feature cut points to observe the performance of DNDT at a learning rate of 0.1 with 256 batch size and the cross entropy being the loss function. The classification accuracy of the stand-alone DNDT model on the test instances was observed to be 83% with a precision of 0.88 in detecting dysphemia and was found less accurate than the proposed model. The number of training epochs were varied from 1000 to 2000 in steps of 500 and maximum accuracy was obtained at 2000 epochs.

4.2 Comparison with DNN and MLP

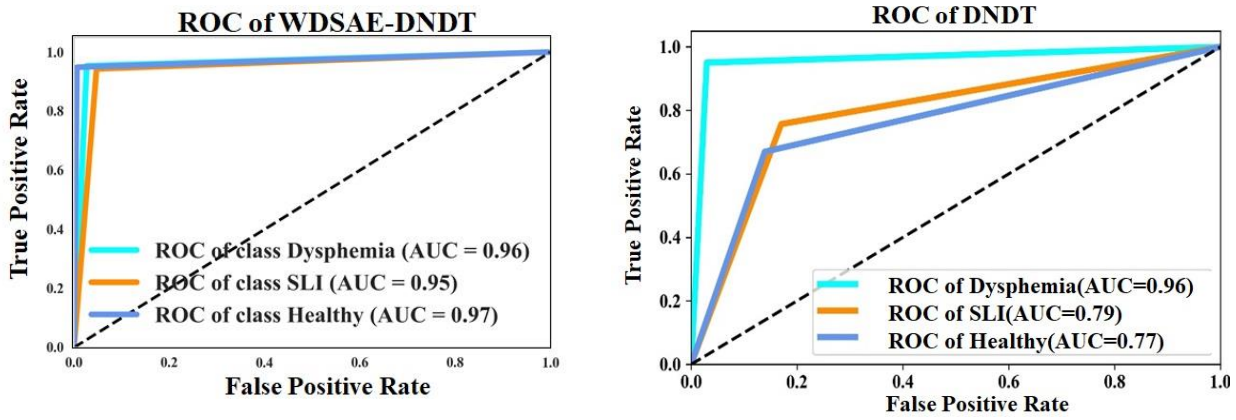
Baseline experiments were further conducted with Deep Neural Network (DNN) model using the scikit flow library in Python [35] and Multi-Layer Perceptron (MLP) classifier using scikit-learn library [37].

A 3-class DNN classification model was built with 10 hidden layers with 20 neurons on each layer, trained on the prosodic speech features. It was also run on 2000 epochs with a batch size of 256 samples and a test data size of 0.20 at a learning rate of 0.1. The cross-entropy loss function and Adam optimizer were chosen as in the previous models. The Rectified Linear Units (ReLU) activation function adopted for DNN showed faster convergence. At the output layer, the softmax activation function ensured that the summation of the activation of each unit at the output layer was unity, such that the output was estimated as conditional probabilities, followed by classification into definite classes.

The MLP model with feed-forward neural network architecture was also built with 2 hidden layers with 20 neurons each. The MLP model was trained with prosodic features, batched with 256 samples. The optimizer, activation functions, train/test split ratio and epochs used in training the DNN model were emulated in the MLP model as well but the learning rate was altered to 0.01 as the MLP model performed well at this learning rate. Though the 'Recall' measure of the DNN and MLP models in predicting the class 'Dysphemia' was considerably good, they showed fairly low 'Precision' in predicting the fluency disorders which dropped their F scores compared to the proposed model.

4.3 ROC - based Evaluation

The Receiver Operating Characteristics (ROC) graph describes the relative trade-off between the rate of true positives and the false positives. The ROC chart is inadequate in envisaging and choosing classifiers on the basis of their performance. To have a comparative study of the performance of the implemented models, the ROC performance curve was reduced to Area Under the Curve (AUC), a single scalar measure of the mean true positive rate of the model over the possible false positive rates.



(a) (b)

Fig. 5: ROC and AUC for (a) Proposed WDSAE-DNDT (b) DNDT Model

The WDSAE-DNDT classifier with the fusion feature set (Prosodic features+ Bottle Neck Features) has the highest AUC of 0.96, 0.95, 0.97 for Dysphemia, SLI and Healthy classes respectively and perform well than the other experimental baselines. From **Fig. 5(a)and 5(b)**, it is observed that AUC of the proposed WDSAE-DNDT model with an average AUC of 0.96 for all classes proves the model's high diagnostic ability in discriminating fluency disorders from healthy speech. The average AUC for DNDT, DNN and MLP amounts to 0.84, 0.84, and 0.76 respectively. Though the other models show an appreciable AUC in classifying dysphemia, they exhibit very low AUC for the other two classes and fail in distinguishing healthy subjects from those with SLI.

4.4 Precision, Recall and F-score of the Models

For classification tasks, the terms true positive or the number of precise classifications, true negatives or the number of substitutions and deletions, false positives or incorrect deletions, the number of substitutions and insertions, and false negatives or incorrect insertions compare the results of the classifier models. Now, comparing the classifiers based on standard evaluation metrics, presented in **Table 3.**

Let,
Sub=Substitutions; Del= Deletions; Ins= Insertions; Cor=Correct Classification; R=Recall, P=Precision
N=Total Number of Classification
Then, the evaluation metrics are as given in equation (22).

$$\text{Precision} = \frac{\text{Cor}}{\text{Cor+Sub+Ins}} \quad \text{Recall} = \frac{\text{Cor}}{\text{Cor+Sub+Del}}$$
$$\text{F-score} = \frac{2*(\text{R}*\text{P})}{(\text{R+P})} \qquad (22)$$

High precision corresponds to low false positive estimate whereas Recall is an estimate of the fraction of positive events that are rightly predicted. It is the correlation between the accurately predicted positive events to all the observed outcomes in the true class. F-score represents the harmonic mean between the recall and precision values. F-score is highly appropriate than classification accuracy, especially in the case of uneven class distribution. F-score ranges between 0 and 1. A better F-score value is closer to 1. It is apparent from **Table 3** that the proposed WDSAE-DNDT model outperforms all the baseline models in predicting true positives (precision) well. Among the baselines, the stand-alone DNDT has a precision and recall higher compared to the DNN and MLP models while the latter achieve a nearly equal F-score in predicting dysphemia. In the classification of SLI and healthy samples, DNN exceeds in the F-score than the MLP but rests always lower than the F-score measure of the proposed model. These evaluation scores are presented graphically in **Fig. 6 (a)** and **(b)**.

Table 3: Classification Accuracy

	Proposed WDSAE-DNDT			**DNDT**			**DNN**			**MLP**		
Class	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Dysphemia	**0.99**	**0.98**	**0.99**	0.88	0.98	0.93	0.85	0.97	0.91	0.85	0.95	0.90
SLI	**0.96**	**0.93**	**0.95**	0.80	0.76	0.78	0.75	0.73	0.74	0.68	0.61	0.65
Healthy	**0.92**	**0.95**	**0.94**	0.75	0.70	0.72	0.70	0.68	0.69	0.61	0.63	0.62

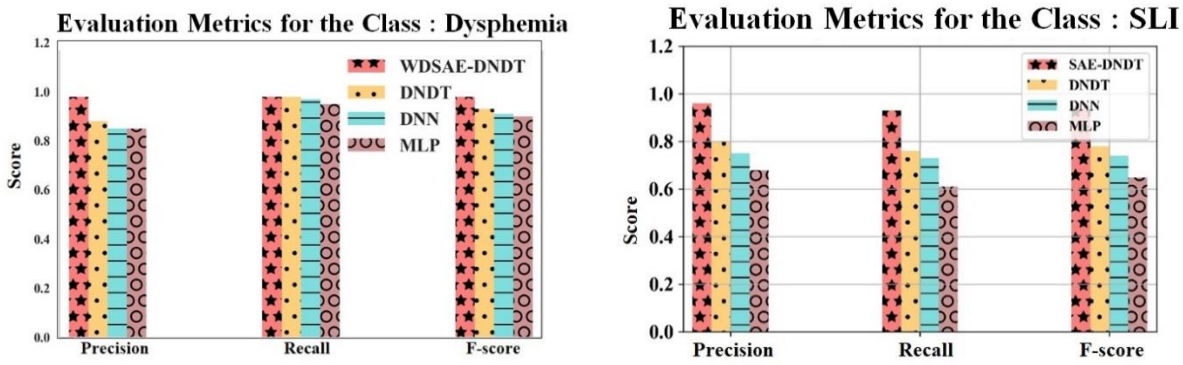


Fig.6: *Precision, Recall and F-score for the class (a) Dysphemia (b) SLI*

4.5 Cross Validation of the Proposed Model

Cross-validation strategy was used to assess the performance of the proposed WDSAE-DNDT model. The training feature-set was split into '5' parallel sets to complete the five-fold cross-validation which is presented in **Fig.7**. The proposed WDSAE-DNDT model yields better validation accuracy sans over-fitting which is evident from the graph. DNN model has better validation accuracy than the rest of the baselines.

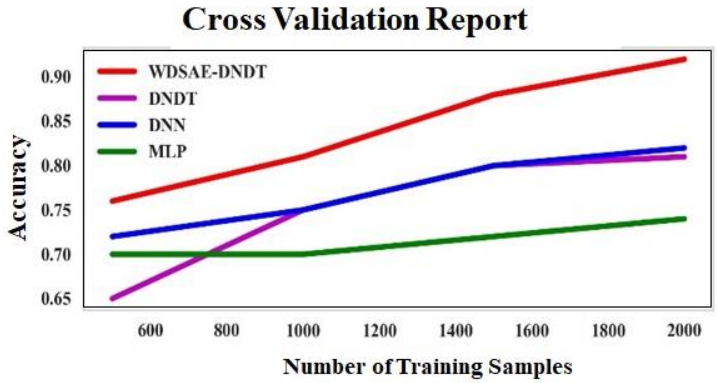


Fig.7: Cross Validation of the Models

4.6 Classification Accuracy and Average Precision- based Evaluation

Evaluation of the machine learning models based on accuracy is an essential part of choosing the classifier. The classification accuracy is the correlation between the number of precise predictions to the total number of input samples. If Cor=Correct Classification and N=Total Number of classifications, classification accuracy is defined as given in equation (23).

$$\text{Classification Accuracy} = \frac{\text{Cor}}{\text{N}} \tag{23}$$

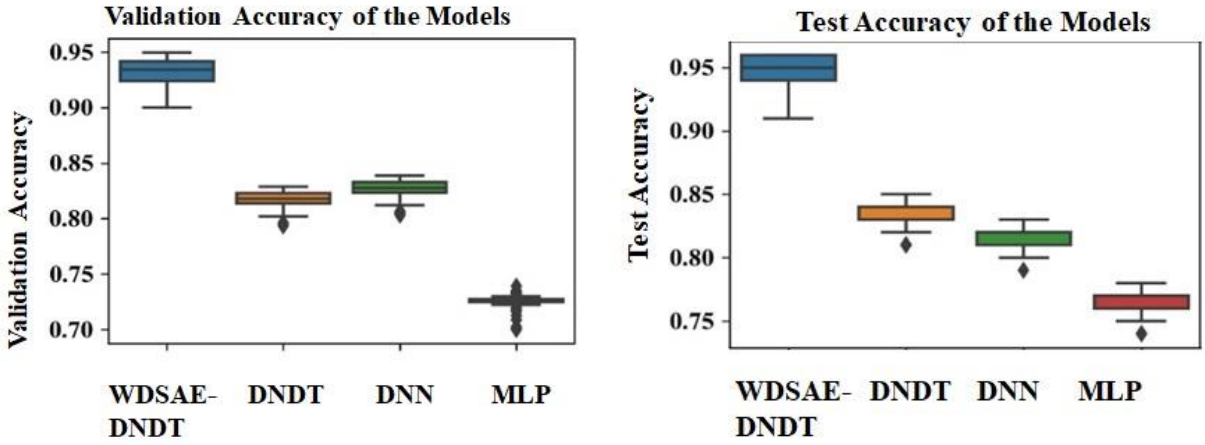
The proposed WDSAE-DNDT model was validated with 20% of the training set and has superior training and test accuracy than the other models. Notably, the DNN shows a slight increase in validation accuracy than the DNDT but proves to be less accurate on the test set when compared to the DNDT model. Additionally, the weighted

average precision was measured as the average of the precision per class label. The macro-average precision amounts to the unweighted average of precisions for each class label. Macro-average which is exhaustively used for multi-class classification was measured for all the experimented models in this work. Since the values of weighted average and macro-average precisions pertain to individual classes, the scores taken for the evaluation of dysphemia class alone is presented in **Table 4** along with the mean validation accuracy and mean test-accuracy of the models measured over the values taken across 75 trials of each model.

Table 4: Evaluation Metrics

	Proposed WDSAE-DNDT	**DNDT**	**DNN**	**MLP**
Weighted-Average Precision	0.96	0.80	0.79	0.71
Macro-Average Precision	0.96	0.81	0.79	0.72
Mean Validation-Accuracy	**0.9312**	0.81	0.82	0.74
Mean Test-Accuracy	**0.9525**	0.83	0.81	0.77

The Validation accuracy and Test accuracy scores of the models are graphically described in **Fig. 8(a) and (b)**. The proposed hybrid model outweighs the rest of the baselines in predicting dysphemia, demonstrated by its high macro-average and weighted-average precision scores.



(a) (b)

Fig.8: (a) Validation Accuracy and (b) Test Accuracy of the models

4.7 Cut-points based WDSAE-DNDT model performance

During the learning phase of the depth-modified Deep Neural Decision Tree, it was found possible to ignore certain features by de-activating its cut points. This corresponds to disabling the feature, so that it does not impact prediction. By increasing the number of cut points for each feature, a larger model was obtained which significantly improved the classification accuracy as shown in Fig. .9. A minimum of 1 cut point to a maximum of 10 cut points were experimented in our trials.

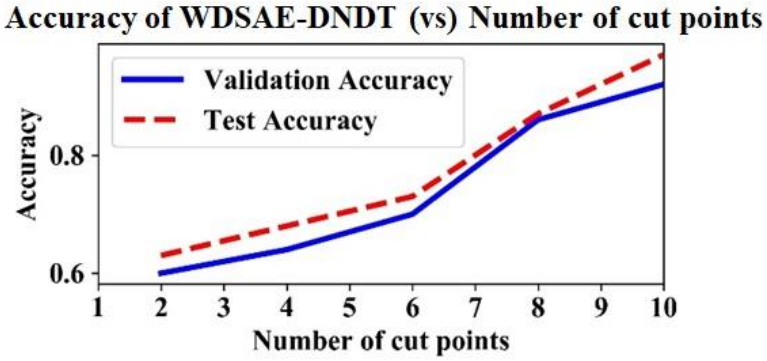


Fig. .9 : Accuracy of WDSAE-DNDT with increasing number of Tree cut points

4.8 Epochs-based WDSAE-DNDT model performance

It was observed during the validation phase that the classification accuracy of the proposed deep learning framework improved with the increase in number of epochs; one epoch being a complete iteration of the feature set through the model. As the number of epochs increased, the number of times the weights were modified in the model increased and the curve rose from underfitting bend to an optimal fit. As presented in **Fig. 10,** the hybrid WDSAE-DNDT model has the highest classification test accuracy at 2000 epochs, which indicates the model's abstinence from over-fitting.

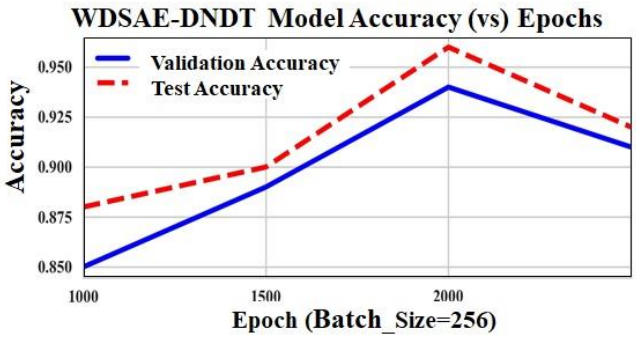


Fig.10: *Accuracy of WDSAE-DNDT model for increasing number of epochs*

4.9 Statistical Evaluation of the Models

The statistical evaluation of the models was carried out to uncover the percentage of test instances precisely classified by each of the models. Statistical metrics namely the Cohen's Kappa Coefficient (κ) [38] and Jaccard Similarity Index (JSI) score [39] were measured for all the models. κ is essentially a correlation coefficient that conveys the closeness between predicted and the true labels as defined in equation (24).

$$\kappa=\frac{AL-PL}{(1-PL)};\ 0<\kappa<1 \tag{24}$$

where, AL: Actual Label
PL: Predicted Label

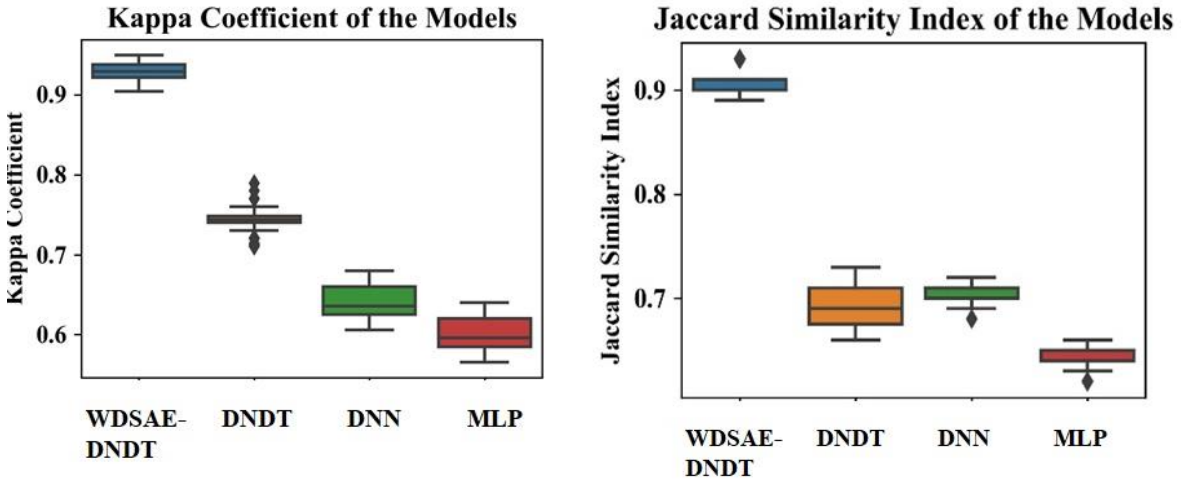
Table 5: Statistical Evaluation of the Models

	Proposed WDSAE-DNDT	**DNDT**	**DNN**	**MLP**
κ	**0.92**	0.74	0.66	0.62
JSI	**0.90**	0.71	0.70	0.65

Since the Kappa Coefficient is influenced by only the instances of data which may have been rightly classified by the model, another statistical metric, JSI is measured for validation. The Jaccard Similarity Index score compares the symmetric difference to the union of the predicted and actual labels of the dataset as given in equation (25)

$$JSI(PL,AL)=\frac{|PL\cap AL|}{|PL\ U\ AL|};\ 0<JSI(PL,AL)<1 \tag{25}$$

The proposed WDSAE-DNDT model achieves the highest mean Kappa Coefficient and mean Jaccard Similarity Index score taken across the 75 trials as shown in **Table 5**. The κ and JSI scores of the models are graphically portrayed in **Fig. 11 (a)-(b).**



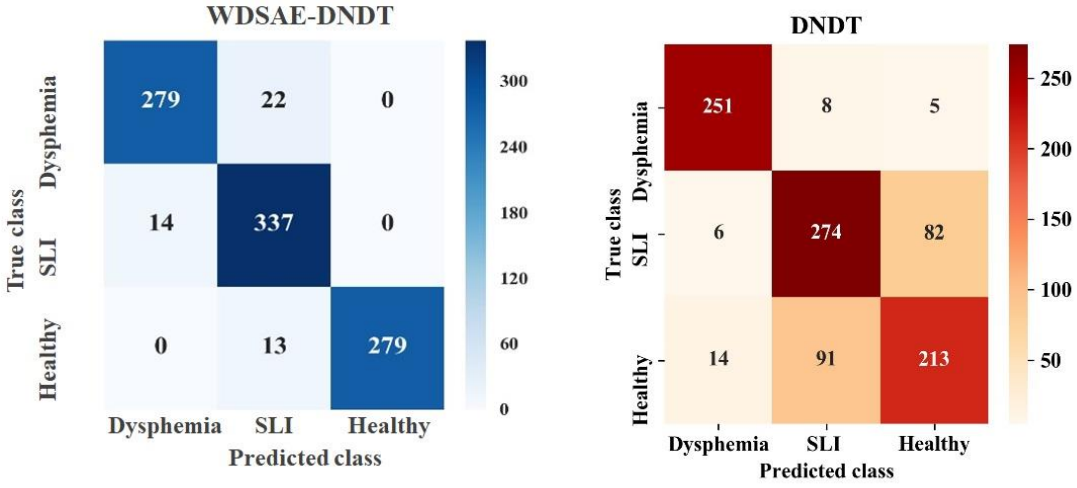
(a) (b)

Fig. 11: (a) Kappa Coefficients (b) Jaccard Similarity Index values of the Proposed and Baseline models

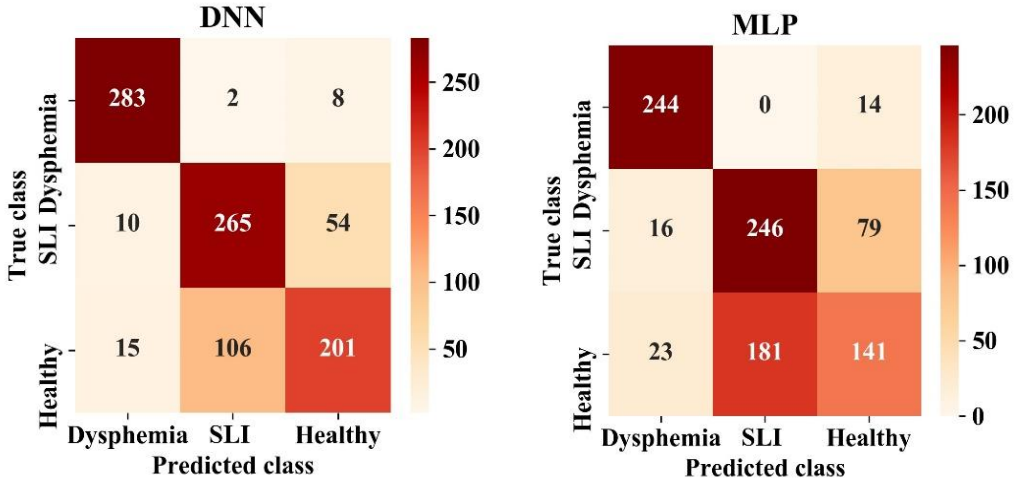
In the above figures, the centre line in the box plots indicates the median value of the Kappa Coefficient; the edges indicating the 25th and 75th percentiles of the observation with the whiskers extending to mark the minimum and maximum values of the readings taken during the 75 trials. The baseline models comparatively display κ values lesser than 0.75 which mirrors the low accuracy of the models in the task of prediction. In the same way, the JSI score comparison between models reveal lower JSI scores among the baselines which indicate the height of dissimilarity between actual labels and predicted labels by the classifiers; the DNN shows an almost equal Jaccard score as the DNDT though the Kappa value of the DNN is lesser which apparently makes the DNDT a better stand-alone classification model than DNN for the task of fluency disorder classification.

4.10 Confusion Matrices of the Models

The confusion matrix of the proposed WDSAE-DNDT model along with the other baselines are given in **Figures 12 (a)-(d).** In the SAE-DNDT model, with a total of 944 test samples, 301 samples fit to the dysphemic class, 351 samples to the SLI class and 292 Healthy samples.



(a) (b)



(c) (d)

Figure 12: Confusion Matrix of (a) WDSAE-DNDT (b) DNDT (c) DNN (d) MLP

The proposed hybrid model could classify 279 dysphemic samples, 337 SLI samples and 279 healthy samples yielding an overall average accuracy of 95%. Owing to the common fixation of the hyper-parameter 'random_state' to be equal to 1 in all the models for a random split in samples from different classes, the number of samples under each class in the train/validation/test set subtly varies in the confusion matrices of all the evaluated models but in every model, the total number of test samples amount to 944 as cited above.

The proposed hybrid model, Weight Decorrelated Stacked Autoencoder-DNDT (WDSAE-DNDT), yielded the best accuracy of classification even with the available sparse data than the competing models in the literature and the other baseline models. Also, the proposed WDSAE-DNDT exhibited higher Precision, Jaccard score, Kappa coefficient, Recall, F-score and the better confusion matrix than the conventional models. Weight decorrelation introduced in the proposed model facilitated better feature learning and dimensionality reduction. Also, the WDSAE-DNDT performed well when trained on the available sparse data by automatic higher-level feature extraction from the input to distinguish one fluency disorder from the other for accurate classification. Thus, WDSAE-DNDT is proposed for fluency disorder classification.

5.0 CONCLUSION

In this paper, Weight Decorrelated Stacked Autoencoder-Deep Neural Decision Tree, a novel hybrid deep learning framework was proposed for fluency disorder classification to distinguish Specific Language Impairment from Dysphemia and to present clinical benchmarks that can help Clinicians evaluate their therapy decisions. The weight decorrelation introduced in the Stacked Autoencoder and the enhanced hidden layer depth of the modified DNDT significantly contributed in improving the prediction accuracy of the proposed WDSAE-DNDT model which yields the best average prediction accuracy of 95% when trained on the available sparse data and is shown to perform well with the increase in the number of epochs and tree cut points. The key findings of this work by profound speech analysis are the inference on a few prosodic features viz. the mean squared error of the estimated energy contour with one-degree polynomial, the regression coefficient between the F0 contour and a linear regression along with the fundamental frequency and its variant measures such the average, variance and standard deviation of pitch, the average, maximum and standard deviation of spectral energy, the voiced rate that greatly contributed to distinguish speech fluency disorders.

In future, the authors wish to take forward the fluency disorder classification task to the early diagnosis of the onset of cognitive diseases such as the Alzheimer's disease and Parkinson's disease in adults based on the disfluent speech patterns of the patients using robust and accurate deep learning algorithms.

DATA AVAILABILITY STATEMENT

The disfluent speech dataset generated and analysed during the current study is available from the corresponding author on reasonable request.

ACKNOWLEDGMENT

This project is partially funded by AICTE, India under the Research Progress Scheme (Grant Reference No. 8-40/RIFD/RPS/Policy-1/2017-18) dated 15th March 2019. The Author is the Co-Investigator of the project and the Co-author is the Principal Investigator.

REFERENCES

[1] K. Govindarajan, J. Paradis, "Narrative abilities of bilingual children with and without Developmental Language Disorder (SLI): Differentiation and the role of age and input factors", *Journal of Communication Disorder*, Vol.77, 2019, pp.1-16.

[2] Robert West, Edward Nusbaum, "A Motor Test for Dysphemia (stuttering)", Quarterly Journal of Speech, Vol. 15, No.4, 1929, pp. 469-479.

[3] M. Corley, L.J. MacGregor, D.I. Donaldson, "It's the Way That You, Er, Say It: Hesitations In Speech Affect Language Comprehension", Cognition, Vol. 105, No.3, 2007, pp. 658–668.

[4] S.H. Fraundorf, D.G. Watson, "The disfluent discourse: Effects of Filled Pauses on Recall", Journal of Memory Language, Vol. 65, No. 2, 2011, pp. 161–175.

[5] K. McDougall, M. Duckworth, "Profiling fluency: An Analysis Of Individual Variation in Disfluencies in Adult Males. Speech Communication, Vol. 95, 2017, pp. 16–27.

[6] O. Bloodstein: '*A Handbook on Stuttering'*, Singular Publishing Group, Inc., San Diego, 1995, CA.178-181.

[7] R.B. Gillam, E. D. Peña, L.M. Bedore, et al., "Identification of specific language impairment in bilingual children: I. Assessment in English", *Journal of Speech, Language, and Hearing Research*, Vol. 56, No. 6, 2013, pp. 1813–1823.

[8] A. Czy˙zewski, A. Kaczmarek, B. Kostek, "Intelligent processing of stuttered speech", *Journal of Intelligent Information Systems*, Vol. 21, No.2, 2003, pp. 143–171.

[9] J.D. Arias-Londono, J.I. Godino-Llorente, N. Saenz-Lechon, et al., "Detection of Pathological Voices using Complexity Measures, Noise Parameters, and Mel-Cepstral Coefficient", *IEEE Transaction, on Biomedical Engineering*, 2011, 58, 370–379.

[10] I. Hammami, L. Salhi, and S. Labidi, "Pathological voices detection using Support Vector Machine", *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, Monastir, 2016, pp. 662-666.

[11] L. Verde, G. DePietro, G. Sannino, "Voice Disorder Identification by Using Machine Learning Techniques", *IEEE Access*, Vol.6, 2018, pp.16246–16255

[12] M. Pishgar, F. Karim, S. Majumdar, H. Darabi, "Pathological Voice Classification Using Mel-Cepstrum Vectors and Support Vector Machine", 2018 IEEE International Conference on Big Data (Big Data), pp. 5267-5271.

[13] J. Nayak, P.S. Bhat, R. Acharya, U.V. Aithal, "Classification and Analysis of Speech Abnormalities", ITBM-RBM, Vol. 26, No. 5, 2005, pp. 319–327.

[14] I. Swietlicka, W. Kuniszyk-Jó´zkowiak, E. Smołka, "Artificial Neural Networks in the Disabled Speech Analysis", Computer Recognition System, Vol. 57, 2009, pp. 347–354.

[15] I. Szczurowska, W. Kuniszyk-Jozkowiak, E. Smolka, "Speech nonfluency detection using Kohonen networks". Neural Computing and Applications, Vol. 18, No.7, 2009, pp. 677–687.

[16] S.O. Orimaye Jojo S-M. Wong, KJ Golden, et al., "Predicting probable Alzheimer's disease using linguistic deficits and biomarkers", BMC Bioinformatics, Vol. 18, No.34, 2017, pp.1-13.

[17] K.C. Fraser, J.A. Meltzer, F. Rudzicz, "Linguistic Features Identify Alzheimer's Disease in Narrative Speech", Journal of Alzheimers Disease, Vol. 49, No.2, 2016, pp.407-22.

[18] Weller, Adrian, "Challenges for Transparency", Proceedings of the ICML Workshop on Human Interpretability in Machine Learning, 2017, pp. 55–62.

[19] V. Gupta, "Voice Disorder Detection Using Long Short-Term Memory (LSTM) Model *Arxiv.org*, 2018, [online] Available: https://arxiv.org/pdf/1812.01779.

[20] H. Wu, J. Soraghan, A. Lowit, G. Di-Caterina, "A Deep Learning Method for Pathological Voice Detection Using Convolutional Deep Belief Networks", Proceedings of Interspeech 2018, pp. 446-450.

[21] Watson, B. Jennifer, Byrd, T. Courtney and Edna J. Carlo, "Chapter 7: Disfluent Speech Characteristics of Monolingual Spanish-Speaking Children". Multilingual Aspects of Fluency Disorders, edited by Peter Howell and John Van Borsel, Bristol, Blue Ridge Summit: Multilingual Matters, 2011, pp. 169-191.

[22] A.L. Leclercq, P. Suaire and A. Moyse, "Beyond stuttering: Speech disfluencies in normally fluent French-speaking children at age 4". Clinical Linguistics & Phonetics, Vol. 32, no. 2, 2018, pp. 166–179.

[23] Sheena Christabel Pravin and Palanivelan, M 2021, 'Regularized Deep LSTM Autoencoder for Phonological Deviation Assessment', International Journal of Pattern Recognition and Artificial Intelligence, vol. 35, no.4, 2021, p. 2152002.

[24] Sheena Christabel Pravin and Palanivelan, M 2021, 'A Hybrid Deep Ensemble for Speech Disfluency Classification', Circuits, Systems, and Signal Processing, Springer, vol. 40, no.8, 2021, pp. 3968-3995.

[25] Y.V. Geetha, K. Pratibha, R. Ashok, et al., "Classification of childhood disfluencies using neural networks", Journal of Fluency Disorders, Vol. 25, 2000, pp. 99-117.

[26] J. Cheng, X. A. Chen, Metallinou, "Deep neural network acoustic models for spoken assessment applications", Speech Communication, Vol. 73, 2015, pp.1427.

[27] L. Chen L, J. Tao, S. Ghaffarzadegan, et al. "End-to-end neural network based automated speech scoring", in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE 2018, pp. 6234-8.

[28] X. Zhang, F. Qin, Z. Chen, L. Gao, G. Qiu, S. Lu, "Fast screening for children's developmental language disorders via comprehensive speech ability evaluation—using a novel deep learning framework", Ann Transl Med, Vol. 8, No. 11, 2020 pp. 1-14.

[29] Y.Yang, I.G. Morillo, T.M. Hospedales, "Deep Neural Decision Trees", ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden, 2018.

[30] P. Boersma, D. Weenink, D., 2013. Praat: Doing phonetics by computer [Computer program]. Version 5.3.51, retrieved 2 February 2020 from http://www.praat.org/

[31] J. Moore, M. Kronenthal, M. and S. Ashby, 'Guidelines for AMI speech transcriptions', AMI Deliverable, 2005.

[32] J. R. Orozco-Arroyave, J. C. Vásquez-Correa et al., "NeuroSpeech: An open-source software for Parkinson's speech analysis". Digital Signal Processing, Vol.77, 2018, pp.207-221.

[33] S. Liu, C. Zhang, J. Ma, "Stacked auto-encoders for feature extraction with neural networks", International Conference on Bio-Inspired Computing- Theories and 442 Applications, Springer, Singapore, 2016, pp. 377–384.

[34] M.D. Zeiler, "ADADELTA: An adaptive learning rate method", arXiv:1212.5701[cs.LG].

[35] Martín Abadi, Ashish Agarwal, Paul Barham, et al., "TensorFlow: Large-scale machine learning on heterogeneous systems", 2016, arXiv:1603.04467.

[36] Norouzi, Mohammad, Collins, et al., "Efficient non-greedy optimization of decision trees", In NIPS, 2015.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python", Journal of Machine Learning, 2011, Res.12, 2825.

[38] W. D. Thompson, S. D. Walter, "A Reappraisal of the kappa coefficient", Journal of Clinical Epidemiology, Vol. 41, No. 10, 1988, pp. 949–958.

[39] Tan, Pang-Ning; Steinbach, et al., "*Introduction to Data Mining*", 2005, ISBN 0-321-32136-7.

[40] Shie Mannor, Dori Peleg and Reuven Rubinstein, "The cross-entropy method for classification", In Proceedings of the 22nd international conference on Machine learning, 2005.

[41] Howell, P, Davis, S & Bartrip, J 2009, The University College London Archive of Stuttered Speech (UCLASS).

[42] LANNA—Laboratory of Artificial Neural Network Applications, Department of Circuit Theory at the FEECTU in Prague. http://ajatubar.feld.cvut.cz/lanna/