Autoencoder-Driven Latent Representation Learning for Language-Agnostic Disordered Speech Classification using a Universal Feature set

Puneet Bawa*, Virender Kadyan†, Shareef Babu Kalluri†
* Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab, India
Email: erpuneetbawa@gmail.com
†University of Petroleum and Energy Studies, Dehradun, India
Email: {vkadyan,shareef}@ddn.upes.ac.in

***Abstract*—The disordered speech affects communication in various ways and presents overlapping acoustic characteristics, which makes automated classification more challenging. The traditional classification models are often disorder-specific and rely on isolated feature sets, which limits their applicability in real-world multilingual contexts. This study proposes a language-agnostic and disorder-generalized classification framework using a hybrid autoencoder-based latent space fusion approach, combining Vanilla, Variational, and SMOTE-Augmented Autoencoders. The proposed methodology establishes a universal set of common features across languages, making it adaptable to diverse linguistic backgrounds. It enables the classification of multiple disorders - including Autism Spectrum Disorder (ASD), Dysarthria, Hyperkinetic Dysarthria, Parkinson's Disease (PD), and Vox Senilis, based on shared acoustic features extracted from multilingual and imbalanced datasets. The proposed framework is evaluated on a combination of a self-collected Code-Mixed Children – Autism Spectrum Disorder (CoMiC-ASD) dataset and publicly available datasets. The CoMiC-ASD dataset comprises 55 speakers, including Hindi-English code-mixed speech from 28 children with ASD (23 male, 5 female) and 27 Typically Developing (TD) children (13 male, 14 female). The proposed model achieves 96.96%–99.49% accuracy across machine learning models, except for decision trees (88.34%).**

I. Introduction

Speech disorders, impacting a significant portion of the global population, pose considerable challenges to effective communication and overall well-being. An accurate and timely diagnosis is crucial for effective interventions and personalized treatment [1]. Over the past decade, researchers have increasingly recognized the diverse nature of speech-sound impairments [2]. It is now widely accepted that children with speech sound disorders do not constitute a uniform group [3]–[5]. These children vary in terms of severity, underlying causes, patterns of speech error, involvement of other linguistic components, response to treatment, and factors that influence long-term maintenance [6]. However, traditional diagnostic methods rely heavily on subjective clinical assessments, which can be time-consuming, resource-intensive, and prone to inter-rater variability [7]. The different disorders like Dysarthria [8], [9], Hyperkinetic Dysarthria [10], Vocal Fold Functionality disorders [11]–[13], Parkinson Disease [14]–[16], ASD [17], require separate evaluation methods to identify and classify them in a unified setting. Moreover, for evaluating these disorders, the dataset is limited and proprietary. This highlights the need for automated, objective, and scalable diagnostic tools. The previous works on the Dysarthria used support vector machines (SVM) with Mel-Frequency Cepstral Coefficients (MFCC), Constant-Q Transform (CQT), Chromagrams, Tonnetz coefficients, Zero Crossing Rate (ZCR), Spectral Roll-off, Spectral Contrast, and Tempo-related Features [18]. [10] have used MFCC features for Hyperkinetic Dysarthria classification, [12], [13] MFCC, Mel frequency energy line features and ZCR, mean-square energy have used for SVM classification to classify Vox Senilis, Larynzocele and pathalogical voices. Many attempts have been made on PD with different classification techniques such as random forest, linear regression, SVM using the characteristics of MFCC, Wavelet Transform (WT), tunable-Q WT [15], [16].

For each study on speech disorders, an isolated classification technique is used for individual task with a different set features on different languages. In this work, we proposed a universal set of common features across languages making it adaptable to multiple linguistic backgrounds, and different age groups, that can classify multiple speech disorders such as ASD, dysarthria, hyperkinetic dysarthria, PD, and Vox Senilis using the latent space fusion of different auto encoders first of its kind.

The key contributions in this paper are:

- We present a new collected ***Co***de ***Mi***xed ***C***hildren – Autism Spectrum Disorder (CoMiC-ASD) dataset with 55 speakers comprising of 28 ASD children (23 males and 5 females) and 27 TD children (13 males and 14 females).
- Unlike the traditional Synthetic Minority Over-sampling Technique (SMOTE) applied at the input level, this approach synthesizes speech data directly in the latent space, ensuring minority class disorders are well-represented and preventing model bias.
- Demonstrates significant improvements over traditional speech disorder classification models having the potential

Corresponding author: Shareef Babu Kalluri.

979-8-3315-7206-8/25/$31.00

to be used for early-stage screening of speech disorders across languages and demographics.

II. Data-Driven Latent Representation

This section introduces the three considered data-driven autoencoders (AE) for the process of generating synthetic data that captures the underlying patterns of the original dataset. In like manner, the working of the autoencoders on the basis of learning a compressed representation (latent space) has been explained.

A. Vanilla Autoencoder

A vanilla autoencoder consists of two main components: an encoder that compresses the input data into a lower-dimensional representation, and a decoder that reconstructs the input from this compressed representation [19]. The encoder ($f_{enc}(x)$) as detailed in equation (1) maps the input data $x \in R^n$ to a latent space $z \in R^m$, where $m < n$.

$$z_V = f_{enc}(x) = \sigma(W_{enc}x + b_{enc}) \tag{1}$$

where $W_{enc} \in R^{(m \times n)}$ corresponds to the weight matrix of the encoder with $b_{enc} \in R^{(m)}$ being the bias vector and σ being a non-linear activation ReLU function. Further, the decoder ($f_{dec}(z)$) maps the latent space z back to the original space, aiming to reconstruct x. Overall, the objective of the loss function or otherwise referred to as reconstruction error ($\mathcal{L}_{recon}$) is to minimize the difference between the input x and the reconstructed output $\hat{x}$ typically measured by Mean Squared Error as shown in equation (2).

$$\mathcal{L}_{recon} = \frac{1}{N}\sum_{i=1}^{N} \parallel x_i - \hat{x}_i \parallel^2 \tag{2}$$

Once the autoencoder has been trained, new synthetic data $\hat{x}_{new}$ is being generated by feeding random variations in the latent space z introducing small perturbations as evaluated in equation (3).

$$\hat{x}_{new} = f_{dec}(z_V + \epsilon) \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ corresponds a small random noise vector.

B. SMOTE-Augmented Autoencoder

In SMOTE-augmented autoencoders, synthetic samples are generated by interpolating between the existing data points in the minority class before passing them through an autoencoder [20]. For each data point x_i in the minority class, a synthetic sample has been generated using equation (4) by interpolating between x_i and one of its nearest neighbors x_{nn}

$$x_{syn} = x_i + \lambda(x_{nn} - x_i) \tag{4}$$

such that $\lambda \in [0, 1]$ corresponds to a random variable controlling the degree of interpolation. Therefore, for the process of augmentation, the sample has been generated using SMOTE such that the augmented data x_{syn} has been fed into vanilla autoencoder as evaluated in equation (5).

$$\hat{x}_{syn} = f_{dec}(f_{enc}(x_{syn})) \tag{5}$$

Similar to the vanilla AE, the SMOTE-augmented AE produces a latent representation as defined in equation (6), from the synthetic input.

$$z_{\text{SMOTE}} = f_{\text{enc}}(x_{\text{syn}}) \tag{6}$$

C. Variational Autoencoder

Variational Autoencoder (VAE) is a generative model that learns the probability distribution of the input data, allowing for better diversity in the generated samples [21]. Unlike vanilla autoencoder, the encoder does not output a single latent vector z but rather parameters of the distribution $q_\phi(z|x)$ which is typically a Gaussian distribution being evaluated in equation (7).

$$q_\phi(z_{VAE}|x) = \mathcal{N}(\mu(x), diag(\sigma(x)^2)) \tag{7}$$

where $\mu(x)$ refer to the mean vector and ϕ denotes the encoder network parameters such that the latent variable z is sampled from distribution as evaluated in equation (8).

$$z_{VAE} \sim \mathcal{N}(\mu(x), \sigma(x)^2) \tag{8}$$

The reconstruction loss consists of two parts as reconstruction loss and KL Divergence(D_{KL}) [22]. The reconstruction loss as evaluated in equation (9) measures how well the generated data matches the input. On the other hand, the KL Divergence (loss) as evaluated in equation (10) encourages the latent space to follow a prior distribution $p(z_{VAE})$, usually $\mathcal{N}(0, I)$.

$$\mathcal{L}_{recon} = E_{q\phi(z_{VAE}|x)}[\parallel x - \hat{x} \parallel^2] \tag{9}$$

$$\mathcal{L}_{KL} = D_{KL}[q\phi(z_{VAE}|x) \parallel p(z_{VAE})] \tag{10}$$

Thus, the reconstruction loss function is being evaluated as in equation (11) where δ corresponds to the hyperparameter that balances reconstruction and regularization.

$$\mathcal{L}_{VAE} = \mathcal{L}_{recon} + \delta\mathcal{L}_{KL} \tag{11}$$

D. Latent Fusion

The fusion process combines the latent representations from the three AEs to enhance the representational power and capture complementary features. The latent representations from these autoencoders are concatenated to form a unified representation as detailed in equation (12).

$$z_{fusion} = [z_V, z_{VAE}, z_{SMOTE}] \tag{12}$$

where z_V, z_{VAE}, z_{SMOTE} corresponds to the latent representation of Vanilla, Variational and SMOTE AEs. Furthermore, the reduction in the dimensionality of the fused latent

space z_{fusion} using principal component analysis (PCA) has been applied by first computing the covariance matrix. This makes fusion more compact and discriminative for downstream tasks where k eigenvectors have been selected to form the projection matrix W_{PCA} with reduced representation being computed as in equation (13). This ensures a compact representation while retaining significant variance in the data.

$$z_{\text{PCA}} = W_{\text{PCA}}^{\top} \left(z_{\text{fusion}} - \bar{z}_{\text{fusion}} \right) \tag{13}$$

Finally, the new dataset using the decoder as evaluated in equation (14) is being generated by sampling from the latent space.

$$\hat{x}_{\text{new}} = f_{\text{dec}}\big(z_{\text{fusion}} + \epsilon\big), \quad \epsilon \sim \mathcal{N}(0, I) \tag{14}$$

Overall, the latent vectors from Vanilla, SMOTE, and Variational AEs are concatenated and reduced via PCA, producing a compact fused representation by optimizing the latent space using a combined loss function ($\mathcal{L}_{fusion}$) as detailed in equation (15).

$$\mathcal{L}_{fusion} = \alpha \mathcal{L}_V + \beta \mathcal{L}_{VAE} + \gamma \mathcal{L}_{SMOTE} \tag{15}$$

where:

- $\mathcal{L}_V$: Reconstruction loss for Vanilla AE.
- $\mathcal{L}_{VAE}$: Reconstruction loss for VAE.
- $\mathcal{L}_{SMOTE}$: Reconstruction loss for SMOTE-Enhanced AE.
- α, β, γ: Weighting factors.

III. Methodology

The proposed methodology introduces a data-driven language-agnostic model for disordered speech classification by leveraging a hybrid autoencoder-based latent space fusion approach. The system integrates three types of autoencoders to capture universal speech disorder features across multiple languages.

A. Dataset Details and Pre-processing

The datasets used in this research has been designed to classify disordered speech from normal speech, comprising a mix of self-collected and publicly available datasets. This work covers six distinct speech disorders datasets, which include our self-collected ***Co***de ***Mi***xed ***C***hildren – Autism Spectrum Disorder (CoMiC-ASD) dataset, Dysarthria [8], Hyperkinetic (HK) Dysarthria [9], Parkinson's Disease (PD) [23], [24], Vox Senilis [11] and Normal speech samples. The CoMiC-ASD dataset was collected from Indian educational settings during structured speech tasks to ensure controlled conditions for capturing high-quality audio. The CoMiC-ASD dataset collection adhered to strict ethical guidelines to ensure participant safety and privacy[1]. The dataset comprises speech recordings collected from children diagnosed with mild/moderate ASD with a unique focus on code-mixed (Hindi and English) language usage.

The participants were engaged in one of these three specific speech tasks: 1) repeating predefined phrases, 2) engaging in spontaneous conversations, and 3) describing pictures. The CoMiC-ASD dataset comprises 23 male speakers and 5 female speakers, resulting in 589 male utterances and 302 female utterances. The ASD speakers falls within the age group of 8-16 years, such that the average age for male and female is 12.83 ± 2.46 (8-16 years) and 13.80 ± 1.64 (11-15 years), respectively. Besides this, we also collected normal children (typical development (TD)) speech data comprising of 27 children (13 Male, 14 female), such that the average age for male and female is 12.77 ± 1.69 (10-16 years) and 12.86 ± 2.88 (9-17 years), respectively. CoMiC-ASD dataset includes children's speech in code-mixed Hindi-English, representing bilingual,low-resource settings such that it has been included to assess model robustness across age and linguistic variability.

The datasets, other than CoMiC-ASD, include adult/elderly speakers: English (Dysarthria, Parkinson disease), Russian (HK Dysarthria), and German (Vox Senilis). Furthermore, only acoustic recordings from the TORGO dataset [8] were used (EMA excluded), and only dysarthric speakers were considered. Overall, the dataset contained 5803 uttered utterances comprising of 3700 male utterances and 2103 female utterances, and 4337 synthesized utterances have also added to the dataset. For normal speech, we considered 40% (241 utterances) of the total typically developing children's speech utterances from the CoMiC-ASD dataset, which consists of code-mixed English-Hindi language samples. The remaining 669 normal speech utterances were proportionally and equally selected from above considered speech datasets in different languages (English, German, and Russian) to ensure linguistic balance and diversity in the analysis. The selection of utterances was carefully made to maintain a balanced representation across languages and to prevent the model from being biased toward specific normal speech patterns.

TABLE I
UTTERANCE DETAILS OF COLLECTED COMIC-ASD DATASET, PUBLIC DATASETS, AND SYNTHESISED SAMPLES

Disorder	Male	Female	Total	Synthesised
CoMiC-ASD	589	302	891	768
Dysarthria	500	492	992	696
HK Dysarthria	2000	–	2000	0
Parkinson	168	450	618	986
Vox Senilis	224	168	392	1110
Normal	219	691	910	777

Details of the dataset is given in Table I. Likewise, speech from different age groups was included to test model robustness and generalizability, with all datasets processed through a standardized optimal-feature pipeline for consistency. As part of the data pre-processing, speech segments corresponding to attendants or facilitators were filtered out from the dataset, retaining only the utterances spoken by the participants themselves to ensure that model training was based solely on dis-

[1]The data collection protocols were approved by an institutional ethics review board with Reference No."IHEC/DHR/CU/PB/24/259"; Biomedical and Health Research (NECRBHR), Department of Health Research (DHR), Government of India

ordered speech characteristics. Additionally, the segmentation and annotation for the datasets have been performed using Praat Toolkit [25] to remove background noise, clicks and coughs manually. All audio files have been resampled to 16 kHz and normalized.

B. Feature Extraction

The various acoustic features have been extracted, focused on parameters such as frequency domain features and time domain features from both disordered speech and normal speech samples. Different set of features like MFCCs (13), Spectral Centroid (1) and Spectral Bandwidth (1) from librosa [26], Formant Frequencies (3), Pitch (1), Jitter (1) and Shimmer (1) from praat [25], Prosodic Features (2) from OpenSMILE [27], and LPC features (13) from Kaldi [28] has been enabled for precise computation of 36-D features ensuring robust representation of both temporal and spectral properties which are critical for classification of disordered speech.

C. Feature Importance

The analysis of feature importance across the disordered speech relies on acoustic features where each feature is represented as x_i with i indicating the specific feature index. Likewise, the importance of each feature vector (x_i) for the particular disorder is quantified upon the feature importance score (η_{ij}). Thus, the feature importance vector (η_j) over n number of features are being evaluated in equation (16) corresponding to each disorder (j).

$$\boldsymbol{\eta}_j = [\,\eta_{1j}, \eta_{2j}, \ldots, \eta_{nj}\,] \tag{16}$$

A radar plot, as illustrated in Figure 1, has been constructed using the normalized mean values for each feature vector, stratified by disorder class such that the angle corresponds to the feature index and the radius to the importance score (ηij). The feature importance vector η is balanced across features in the case of ASD (j_1). The notable contributions for the feature set, including SpeechRate, certain LPCs, and MFCCs have been found to be optimal, as they reflect reduced articulation speed and temporal irregularities commonly observed in autistic speech. Further, in case of Dysarthria (j_2), the significant variation in certain MFCCs and formant features (mainly $Formant\_F2$), which indicates impaired articulatory control and resonance. Conversely, in case of Hyper-Kinetic Dysarthria (j_3), the Jitter, Shimmer, and certain LPCs have been found to be optimal. The Pitch, certain MFCCs, jitter and SpeechRate has shown significant importance in case of Parkinson's Disease (j_4), which is consistent with voice tremors and instability associated. For Vox Senilis (j_5), MFCCs, Formants (mainly $Formant\_F2$) and Shimmer has indicated the relevance by representing glottal instability and articulatory decline associated with aging speech. For Normal speech (j_6), the notable contributions for the feature set including MFCCs, LPCs, and MeanIntensity have been found to be optimal, reflecting acoustic consistency and a well-regulated vocal tract system. Additionally, pearson correlation analysis was performed to examine linear dependencies among features. The feature sets including MFCCs, LPCs, Formants, Pitch, and Jitter were highly correlated, highlighting redundancy across certain acoustic measures. Thus, the feature selection process uses Euclidean distance between disorder centroids. This establishes the significance of the feature and the feature importance score I_j is calculated on the basis of the mean of Euclidean distance, as evaluated in equation (17).

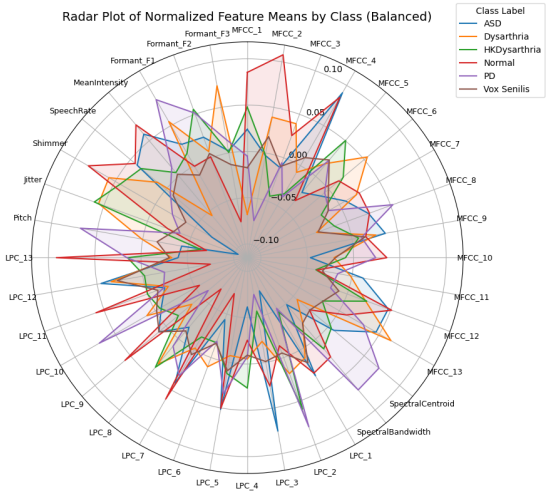

Fig. 1. Radar Plot of Feature Importance Across Disordered and Normal Speech

$$I_j = \frac{1}{|K|} \sum_{(a,b)\in K} \left|\, C_{a,j} - C_{b,j} \,\right| \tag{17}$$

where $C_{a,j}$ and $C_{b,j}$ denote the mean value of feature j for disorder centroids a and b, respectively, and K is the set of all unordered disorder pairs. The score I_j thus reflects the discriminative ability of feature j to separate disorders. Features are ranked according to I_j, and the top 20 features are selected to form the universal feature set.

D. Proposed System Architecture

Figure 2 illustrates the proposed language-agnostic and disorder-generalized architecture. In the autoencoder stage we have used Vanilla, SMOTE and Variational AEs for generating the synthetic data, more details are given in Section 2. The

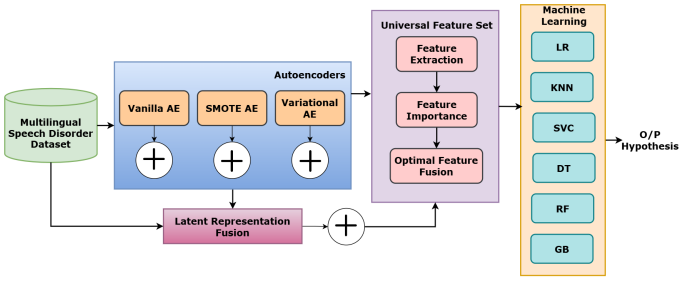

Fig. 2. Block Diagram of proposed language-agnostic and disorder-generalized architecture

TABLE II
PERFORMANCE MEASURE ACROSS ALL THE MACHINE LEARNING TECHNIQUES VS AUTO-ENCODER METHODS AND PROPOSED ARCHITECTURE

Model Type	Baseline		Vanilla AE		SMOTE		Variational AE		Proposed	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
LR	0.7605	0.7623	0.7105	0.7421	0.8074	0.8075	0.9215	0.9123	**0.9874**	**0.9812**
KNN	0.7823	0.7854	0.6832	0.7256	0.8142	0.8147	0.9123	0.9024	**0.9696**	**0.9628**
SVC	0.7734	0.7727	0.6927	0.7389	0.8648	0.8611	0.9234	0.9125	**0.9949**	**0.9922**
DT	0.7589	0.7589	0.5846	0.6258	0.8291	0.8262	0.8712	0.8605	**0.8834**	**0.8859**
RF	0.8432	0.8423	0.7049	0.7521	0.8416	0.8408	0.9021	0.8904	**0.9785**	**0.9716**
GB	0.8527	0.8528	0.7023	0.7645	0.8532	0.8491	0.9218	0.9128	**0.9718**	**0.9795**

synthesized data is combined with the original dataset to extract feature vectors and features of importance. These selected features, specific to each Speech-Language Disorder, are then utilized by machine learning models including Linear Regression (LR), K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB) for accurate classification of disordered speech. The evaluation metrics for each machine learning approach to classifying the SLD is shown in Table II. The proposed architecture is robust to language, age group and speech disorders.

IV. RESULTS AND DISCUSSIONS

For the baseline approach, we have used original data and all the extracted features (ref Sec III-B) without performing the feature selection for the classification of disordered speech data. This approach results into the classification accuracy of disordered speech ranging from 75.89% to 85.27% shown in Table II. The lower performance has multiple factors like class imbalance, disorder-specific features, cross-language feature variations (e.g., German (Vox Senilis) vs. Russian (Hyperkinetic Dysarthria)) across different machine-learning pipelines.

An autoencoder-based latent representation learning for improved feature compression & generalization technique has been investigated, and proposed a language-agnostic and disorder-generalized architecture. With the proposed architecture, the classification accuracy across multiple disorders and normal speech has significantly improved, ranging from 96.96% to 99.49% across all machine learning models, except for decision trees (88.34%). The corresponding F1-score and accuracy metrics are presented in Table II, while the confusion matrix for each specific disorder is illustrated in Figure 3.

The improvement in disorder classification is achieved by generalizing speech features across all disorders rather than focusing on a specific one. Additionally, the latent representation technique enhances disorder classification by reducing inter-disorder confusion. The figure 4 illustrates the high AUC values achieved by the SVC classifier, indicating strong class separability for each speech disorder. The compact clustering of ROC curves near the top-left corner reflects minimal inter-class overlap in the learned feature space. These results validate the effectiveness of latent feature fusion in capturing disorder-specific acoustic alterations, thereby enhancing the classifier's discriminative performance.

V. CONCLUSION

This paper addresses the challenges in diagnosing multiple speech disorders, emphasizing the need for automated

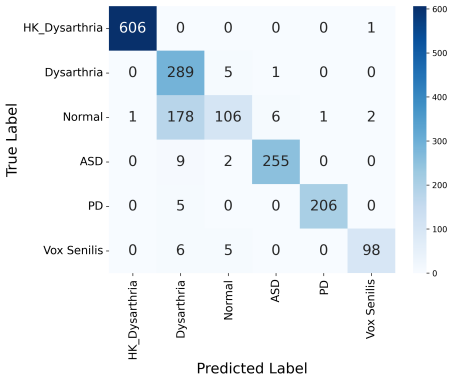

Fig. 3. Confusion matrix across disorders along with normal speech using proposed architecture and universal feature set

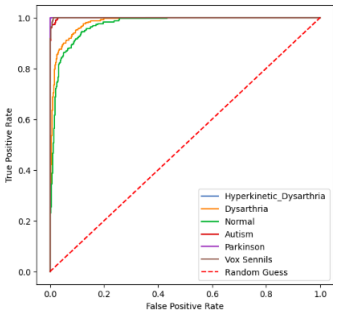

Fig. 4. ROC curve for SVC classifier for all the speech disorders

and scalable tools to overcome the limitations of traditional subjective methods. We introduce a new code-mix children – Autism Spectrum Disorder (CoMiC-ASD) dataset, collected from children aged 8–16 years. We propose a new approach for classifying multiple speech disorders by fusing latent space representations from Vanilla, Variational, and SMOTE Autoencoders. This method, the first of its kind, enhances classification performance for disorders such as Autism Spectrum Disorder, Dysarthria, Hyperkinetic Dysarthria, Parkinson's Disease, and Vox Senilis. Additionally, we establish a universal feature set including MFCCs, LPCs, formants (F1, F2, F3), Jitter, and Pitch to ensure adaptability across multiple languages and disordered speech. This approach showed a improvement in language-agnostic performance. Overall, the language-agnostic and disorder-generalized architecture using latent representation fusion has shown its ability to be an effective scaling strategy by tackling the problem of underrepresented disorders for use in real-world clinical applications.

References

[1] B. A. Lewis, L. A. Freebairn, and H. G. Taylor, "Follow-up of children with early expressive phonology disorders," *Journal of learning Disabilities*, vol. 33, no. 5, pp. 433–444, 2000.

[2] A. Tyler, "Subgroups, comorbidity, and treatment implications," *Speech sound disorders in children: In honor of Lawrence D. Shriberg*, pp. 71–92, 2010.

[3] E. Baker, "Management of speech impairment in children: The journey so far and the road ahead," *Advances in Speech Language Pathology*, vol. 8, no. 3, pp. 156–163, 2006.

[4] C. Bowen, *Children's speech sound disorders*. John Wiley & Sons, 2023.

[5] B. Dodd, *Differential diagnosis and treatment of children with speech disorder*. John Wiley & Sons, 2013.

[6] D. Barbara, "Differentiating speech delay from disorder: Does it matter?" *Topics in Language Disorders*, vol. 31, no. 2, pp. 96–111, 2011.

[7] G. P. Usha and J. S. R. Alex, "Speech assessment tool methods for speech impaired children: A systematic literature review on the state-of-the-art in speech impairment analysis," *Multimedia Tools and Applications*, vol. 82, no. 22, pp. 35 021–35 058, 2023.

[8] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language resources and evaluation*, vol. 46, pp. 523–541, 2012.

[9] A. M. Hashan, C. R. Dmitrievich, M. A. Valerievich, D. D. Vasilyevich, K. N. Alexandrovich, and B. B. Andreevich, "Deep learning based speech recognition for hyperkinetic dysarthria disorder," in *2024 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, IEEE, 2024, pp. 012–015.

[10] M. H. Antor, N. A. Khlebnikov, and B. A. Bredikhin, "Developing a hyperkinetic dysarthria speech classification system using residual learning," in *2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE)*, IEEE, 2024, pp. 1020–1023.

[11] B. Woldert-Jokisz, "Saarbruecken voice database," 2007.

[12] M. Zakariah, M. Al-Razgan, and T. Alfakih, "Pathological voice classification using meel features and svm-tabnet model," *Speech Communication*, vol. 162, p. 103 100, 2024.

[13] E. Özbay, F. A. Özbay, N. Khodadadi, F. S. Gharehchopogh, and S. Mirjalili, "Multifeature fusion method with metaheuristic optimization for automated voice pathology detection," *Journal of Voice*, 2024.

[14] D. Kumar, K. Peter, R. Sanjay, V. Rekha, Z. Poonam, and A. Sridhar, "Screening parkinson's diseases using sustained phonemes," *RMIT University, https://doi.org/10.25439/rmt*, vol. 12618755, p. v1, 2020.

[15] G. C. Oliveira, N. D. Pah, Q. C. Ngo, *et al.*, "A pilot study for speech assessment to detect the severity of parkinson's disease: An ensemble approach," *Computers in Biology and Medicine*, vol. 185, p. 109 565, 2025.

[16] H. Gunduz, "An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on parkinson's disease classification," *Biomedical Signal Processing and Control*, vol. 66, p. 102 452, 2021.

[17] Y. Hus and O. Segal, "Challenges surrounding the diagnosis of autism in children," *Neuropsychiatric disease and treatment*, pp. 3509–3529, 2021.

[18] A. S. Al-Ali, R. M. Haris, Y. Akbari, M. Saleh, S. Al-Maadeed, and M. Rajesh Kumar, "Integrating binary classification and clustering for multi-class dysarthria severity level classification: A two-stage approach," *Cluster Computing*, vol. 28, no. 2, p. 136, 2025.

[19] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[21] D. P. Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[22] S. Chatterjee, S. Maity, M. Bhattacharjee, S. Banerjee, A. K. Das, and W. Ding, "Variational autoencoder based imbalanced covid-19 detection using chest x-ray images," *New Generation Computing*, vol. 41, no. 1, pp. 25–60, 2023.

[23] Q. Yu, Y. Ma, and Y. Li, "Enhancing speech recognition for parkinson's disease patient using transfer learning technique," *Journal of Shanghai Jiaotong University (Science)*, vol. 27, no. 1, pp. 90–98, 2022.

[24] F. Prior, T. Virmani, A. Iyer, *et al.*, "Voice Samples for Patients with Parkinson's Disease and Healthy Controls," Aug. 2023. DOI: 10.6084/m9.figshare.23849127.v1. [Online]. Available: https://figshare.com/articles/dataset/Voice_Samples_for_Patients_with_Parkinson_s_Disease_and_Healthy_Controls/23849127.

[25] P. Boersma and D. Weenink, *Praat: Doing phonetics by computer*, Last accessed: 10 February 2025, 2025. [Online]. Available: https://www.fon.hum.uva.nl/praat/.

[26] B. M. et al., *Librosa: Audio and music signal processing in python - feature documentation*, Last accessed: 10 February 2025, 2025. [Online]. Available: https://librosa.org/doc/main/feature.html.

[27] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[28] D. P. et al., *Kaldi: A toolkit for speech recognition*, Last accessed: 10 February 2025, 2025. [Online]. Available: https://kaldi-asr.org/.