PLOS Digital Health

Speaker-independent dysarthria severity classification using self-supervised transformers and multi-task learning

--Manuscript Draft--

Manuscript Number:	PDIG-D-24-00243
Article Type:	Research Article
Full Title:	Speaker-independent dysarthria severity classification using self-supervised transformers and multi-task learning
Short Title:	Speaker-independent dysarthria severity classification using self-supervised transformers
Corresponding Author:	Balasundaram Kadirvelu Imperial College London London, UNITED KINGDOM
Order of Authors:	Balasundaram Kadirvelu
	Lauren Stumpf
	Sigourney Waibel, PhD
	A. Aldo Faisal, PhD
Keywords:	Dysarthria; speech intelligibility assessment; Self-supervised models; Transformers; Multi-task learning; Contrastive learning; Deep learning
Abstract:	Dysarthria, characterised by slurred speech, is a hallmark of many neurological disorders and brain trauma. Clinical assessment requires an audio-visual investigation by a trained healthcare expert, who evaluates criteria such as respiration, phonation, articulation, resonance, and prosody during speech. Quantitative assessment of dysarthria is challenging due to its complexity, variability, and the subjective nature of human-observation-based scoring methods. We present a novel machine-learning framework using transformers for stratifying and monitoring patient speech. Our framework integrates a wav2vec 2.0 model, pre-trained on raw speech data from healthy individuals. To reduce reliance on speaker-specific characteristics and effectively manage the intrinsic intra-class variability of dysarthric speech, we employ a contrastive learning strategy with a multi-task objective: cross-entropy loss for classifying dysarthria severity, and triplet margin loss to ensure latent embeddings are grouped by severity rather than by speaker. This Speaker-Agnostic Latent Regularisation (SALR) framework provides an objective, accessible, and cost-effective alternative to traditional assessments. Evaluated on the Universal Access Speech dataset with leave-one-speaker-out cross-validation, our SALR framework achieved an accuracy of 70.5% and an F1 score of 59.2%, surpassing the previous benchmark of 54%. This represents a 16.5% increase in accuracy or a relative improvement of over 30%. Explainability analysis indicates that our multi-task objective enhances the ordinal structure of the latent space, reducing dependence on speaker-specific cues and demonstrating robustness and generalisability. In conclusion, the SALR framework sets a new benchmark in speaker-independent dysarthria severity classification, with potential implications for broader clinical applications in automated verbal assessments.
Additional Information:	
Question	**Response**
Government Employee Are you or any of the contributing authors an employee of the United States government?	No - No authors are employees of the U.S. government.

Manuscripts authored by one or more US Government employees are not copyrighted, but are licensed under a CC0 Public Domain Dedication, which allows unlimited distribution and reuse of the article for any lawful purpose. This is a legal requirement for US Government employees. This will be typeset if the manuscript is accepted for publication.	
Financial Disclosure Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review the submission guidelines for detailed requirements. View published research articles from *PLOS Digital Health* for specific examples. This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate. **Funded studies** Enter a statement with the following details: • Initials of the authors who received each award • Grant numbers awarded to each author • The full name of each funder • URL of each funder website • Did the sponsors or funders play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript? Did you receive funding for this work?	The author(s) received no specific funding for this work.
Competing Interests On behalf of all authors, disclose any competing interests that could be perceived to bias this work.	No competing interests

This statement will be typeset if the manuscript is accepted for publication. Review the instructions link below and PLOS Digital Health's competing interests policy to determine what information must be disclosed at submission.	
Data Availability Provide a **Data Availability Statement** in the box below. This statement should detail where the data used in this submission can be accessed. This statement will be typeset if the manuscript is accepted for publication. Before publication, authors are required to make all data underlying their findings fully available, without restriction. Review our PLOS Data Policy page for detailed information on this policy. Instructions for writing your Data Availability statement can be accessed via the Instructions link below.	The UA speech dataset employed in this study was previously published [Kim, Heejin, et al. "Dysarthric Speech Database for Universal Access Research." Interspeech, vol. 2008, 2008] and is accessible upon request, subject to ethical considerations outlined by the original authors.

Speaker-independent dysarthria severity classification using self-supervised transformers and multi-task learning

Balasundaram Kadirvelu[1◕], Lauren Stumpf[1◕], Sigourney Waibel[1], A Aldo Faisal[1,2*],

1 Brain & Behaviour Lab, Department of Computing and Department of Bioengineering, Imperial College London, London, United Kingdom

2 Chair in Digital Health, Faculty of Life Sciences, University of Bayreuth, Bayreuth, Germany

◕These authors contributed equally to this work.

* a.faisal@imperial.ac.uk

Abstract

Dysarthria, characterised by slurred speech, is a hallmark of many neurological disorders and brain trauma. Clinical assessment requires an audio-visual investigation by a trained healthcare expert, who evaluates criteria such as respiration, phonation, articulation, resonance, and prosody during speech. Quantitative assessment of dysarthria is challenging due to its complexity, variability, and the subjective nature of human-observation-based scoring methods. We present a novel machine-learning framework using transformers for stratifying and monitoring patient speech. Our framework integrates a wav2vec 2.0 model, pre-trained on raw speech data from healthy individuals. To reduce reliance on speaker-specific characteristics and effectively manage the intrinsic intra-class variability of dysarthric speech, we employ a contrastive learning strategy with a multi-task objective: cross-entropy loss for classifying dysarthria severity, and triplet margin loss to ensure latent embeddings are grouped by severity rather than by speaker. This Speaker-Agnostic Latent Regularisation (SALR) framework provides an objective, accessible, and cost-effective alternative to traditional assessments. Evaluated on the Universal Access Speech dataset with leave-one-speaker-out cross-validation, our SALR framework achieved an accuracy of 70.5% and an F1 score of 59.2%, surpassing the previous benchmark of 54%. This represents a 16.5% increase in accuracy or a relative improvement of over 30%. Explainability analysis indicates that our multi-task objective enhances the ordinal structure of the latent space, reducing dependence on speaker-specific cues and demonstrating robustness and generalisability. In conclusion, the SALR framework sets a new benchmark in speaker-independent dysarthria severity classification, with potential implications for broader clinical applications in automated verbal assessments.

Author Summary

Dysarthria, a speech impairment caused by neurological conditions, is a common symptom of several disorders, including stroke, head trauma, brain tumours, Parkinson's disease, multiple sclerosis, motor neuron disease, and cerebral palsy. Accurate assessment of dysarthria is challenging due to the complex nature of speech disorders, the variability among patients, and the biases inherent in human observation. Traditional methods for evaluating dysarthria are often subjective and rely heavily on expert opinions. There is a clear need for more standardised, efficient and accessible tools to assess dysarthria. We have developed a novel deep-learning framework to classify dysarthria severity levels directly from speech recordings without needing expert input. Our framework, tested using the Universal Access Speech dataset, achieved a classification accuracy of 70.5%, surpassing the previous benchmark by a 16.5% increase in accuracy. The results indicate that our framework provides a more consistent and objective way to classify dysarthria severity compared to traditional assessments. This advancement could lead to more reliable dysarthria evaluations in clinical environments, potentially impacting treatment approaches and improving patient care.

Introduction

Dysarthria, characterised by impaired control over speech muscles due to neurological conditions, has a profound impact on communication and quality of life [1]. Various neurological disorders, including stroke, head trauma, brain tumours, Parkinson's disease, multiple sclerosis, motor neuron disease, and cerebral palsy manifest dysarthria, leading to a spectrum of speech abnormalities [2, 3]. The complex nature of dysarthria, influenced by underlying pathology and individual patient characteristics, presents significant challenges in both assessment and management [4]. Effective assessment of dysarthria is crucial not only for understanding its severity but also for monitoring disease progression and tailoring therapeutic interventions [5].

The traditional approach to dysarthria assessment involves auditory-perceptual evaluations by experienced speech-language pathologists. However, this method is subjective and may lack consistency, underlining the need for more objective and standardised assessment tools [6]. With advancements in technology, automated, machine learning-based tools have emerged as promising alternatives, offering the potential for more objective, efficient and accessible dysarthria assessments which can be especially advantageous for individuals facing mobility challenges due to co-occurring physical disabilities [7].

Recent studies [8–13] have explored a variety of machine-learning techniques for automating the assessment of dysarthria, highlighting their potential to revolutionise diagnostics in this field. Gupta et al. [8] employed short-duration speech segments analysed via Residual Neural Networks (ResNet) to classify dysarthria severity levels. Shih et al. [9] developed an integrated model combining convolutional neural networks and gated recurrent units to detect dysarthria. Joshy et al. [10] utilised deep neural networks to classify dysarthria through low-dimensional feature representations derived from subspace modelling. Tripathi et al. [11] processed outputs from the DeepSpeech end-to-end speech-to-text engine to extract features for their analysis. Tong et al. [12] proposed a cross-modal deep learning framework that integrates both audio and video data to classify dysarthria severity levels. Lastly, in a recent study, Joshy et al. [13] examined the effectiveness of multi-head attention mechanisms and multi-task learning in the automated classification of dysarthria severity levels.

Despite these advancements, developing accurate and reliable automated tools remains a significant challenge [14]. The variability in speech patterns among individuals with dysarthria, which is influenced by the type and severity of the underlying neurological condition, complicates the development of effective diagnostic models. Additionally, the scarcity of extensive dysarthric speech datasets, exacerbated by the difficulties in collecting prolonged speech samples from individuals with severe dysarthria, hampers the training of advanced machine learning models that require large amounts of data [15, 16].

Recent advancements in deep learning, particularly transformer models, have shown potential in various speech processing tasks [17, 18]. Their ability to capture contextual information across entire input sequences makes them well-suited for modelling the nuanced effects of dysarthria on speech [13]. In this study, we propose a novel framework that leverages a transformer model trained on healthy speech to assess the severity of dysarthria. Our methodology exploits the wav2vec 2.0 [19] model, a state-of-the-art self-supervised transformer model, to extract meaningful speech representations. Through self-supervised pre-training on healthy speech, the wav2vec 2.0 model acquires an understanding of speech's fundamental structure, a characteristic we leverage in our framework to overcome data scarcity constraints. Furthermore, our framework incorporates a multi-task learning strategy to prevent over-fitting and to accommodate the inherent intra-class variability observed in dysarthric speech. Through rigorous evaluation and validation, we demonstrate the effectiveness of our proposed framework, thereby contributing to the advancement of more accurate and accessible assessments for dysarthria.

Materials and methods

Dataset

We used the Universal Access dysarthric speech corpus (UA-Speech) [20], a comprehensive and commonly used English language dataset for dysarthric speech research. The dataset comprises recordings of spoken words from 15 subjects with dysarthria and 13 age-matched healthy controls. Each participant read three blocks of 255 words each. Each block contained 155 words that were repeated across blocks, and 100 uncommon words that were unique to each block. The 155 repeated words included 10 digits, 26 radio alphabet letters, 19 computer commands, and 100 common words from the Brown Corpus. The unique uncommon words in each block were selected from novels in Project Gutenberg to maximise phone-sequence diversity. This resulted in a total of 765 isolated words per subject, with 455 distinct words. These recordings were captured using a seven-channel microphone array and five native American English speakers transcribed the recordings. Each subject's speech intelligibility was calculated based on the average percentage of words correctly transcribed. Subjects were categorised into four levels of dysarthric severity based on their intelligibility ratings: very low severity(76-100% intelligible), low severity (51-75% intelligible), medium severity (26-50% intelligible), and high severity (0-25% intelligible). For a detailed overview of UA-Speech, readers are referred to [20].

Finetuning the wav2vec 2.0 model

We chose to use wav2vec 2.0 [19] for our pretrained transformer over other models like Audio Spectrogram Transformer [18] and HuBERT [21] based on empirical evidence from early experimentation. The wav2vec 2.0 model was pretrained on an expansive 960-hour dataset from diverse audio-book libraries and the entire pretraining process was distributed across 64 V100 GPUs and spanned 1.6 days. The underlying transformer architecture of this model consists of 12 transformer blocks. Each block has a model dimension of 768, an inner feed-forward network dimension of 3072, and 8 attention heads. We utilise the `facebook/wav2vec2-base` model available from the 4.33.1 version of the HuggingFace library [22]. To fine-tune our model for the specialised task of classifying dysarthria severity, we added a linear classification head comprising two linear layers with a ReLU activation. The fine-tuning training was conducted with a batch size of four, using the Adam optimiser set with a learning rate of 0.0005, betas configured to $(0.9, 0.98)$, and an epsilon value of 1×10^{-8}.

Speaker-Agnostic Latent Regularisation (SALR) Framework

Our initial experiments demonstrated a significant challenge with simply fine-tuning an off-the-shelf wav2vec 2.0 model: its tendency to overfit specific speakers. This could be attributed to the limited diversity within the UA-Speech dataset, which only includes 15 distinct dysarthric speakers. Instead of effectively learning the characteristics specific to dysarthria severity, the model appears to be leveraging speaker-specific cues. This approach can minimise the training loss, through recognising the speaker's identity and subsequently assigning a dysarthria severity label. But this approach struggles with new, previously unheard speakers, highlighting a gap in the model's ability to generalise. This issue extends to the latent space, potentially leading to the formation of speaker-centric clusters. Different words uttered by the same speaker are more closely embedded in the latent space compared to the same words spoken by different speakers, even if those speakers have the same level of dysarthria severity. This entangled representation of words is because the complexity of a word — defined by its syllables, phonetic structures, and the necessary motor control for pronunciation — directly impacts how prominently dysarthric symptoms manifest. Without a clear representation in the latent space that accounts for word complexity, the model faces challenges.

To address these issues in the latent space, we introduce a regularisation contrastive loss framework called Speaker-Agnostic Latent Regularisation (SALR) to disentangle the embeddings. Our framework (Fig 1) represents a specialised configuration that enhances the fine-tuned wav2vec 2.0 model with additional components tailored to accomplish an auxiliary task alongside the primary dysarthria classification. The auxiliary task in this framework is a contrastive learning task which aims to ensure that the separation between word embeddings within a shared severity classification becomes speaker-agnostic, thereby preventing the model from learning embeddings that embed speaker-specific characteristics. Specifically, the framework consists of an extra head designed for the auxiliary task, a weighted loss function crafted to balance the learning objectives of both the primary and auxiliary tasks, and a training regimen that specifies how the weighted loss function is applied.

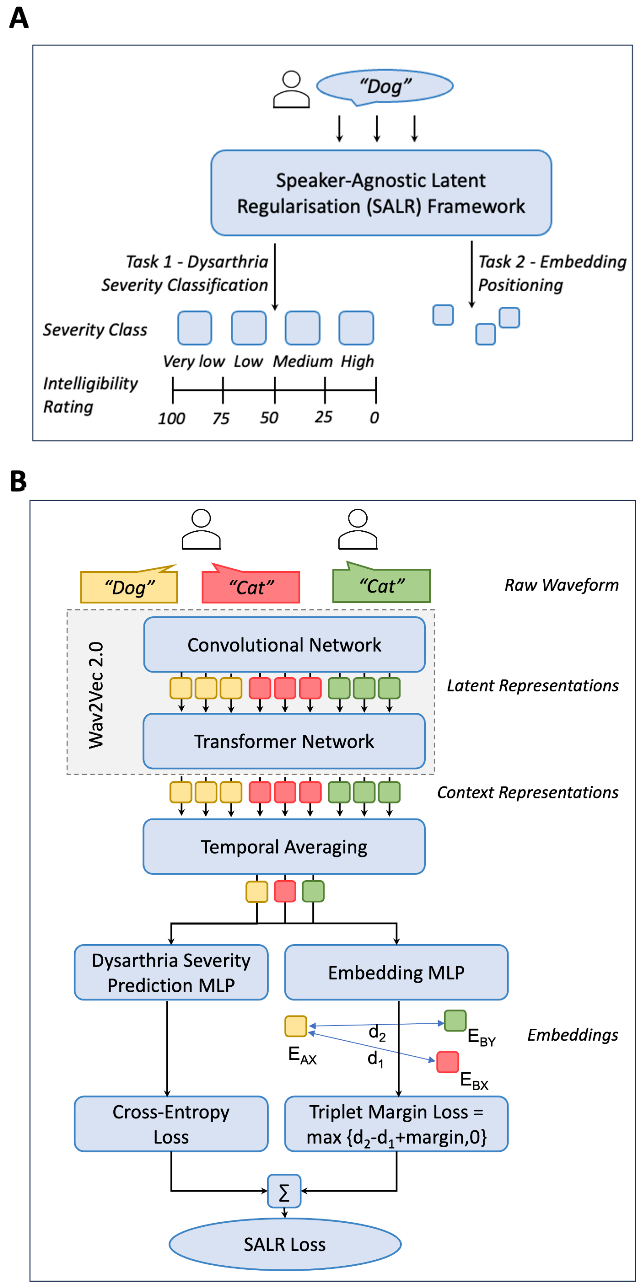


Fig 1. Speaker-Agnostic Latent Regularisation (SALR) framework. A. Conceptual overview of the SALR framework, illustrating its multi-task learning approach. The framework incorporates a primary task of dysarthria severity classification and an auxiliary task utilising contrastive learning to generate speaker-agnostic embeddings in the latent space. **B.** Detailed architecture of the SALR Framework, highlighting the computation pathways for the combined SALR loss, which includes both cross-entropy and triplet margin losses to optimise embedding separation and accuracy.

To illustrate this framework, let E_{AX} represent the embedding of a 'Word A' articulated by 'Patient X', E_{BX} for 'Word B' spoken by 'Patient X', and E_{BY} for 'Word B' spoken by 'Patient Y', where both 'Patient X' and 'Patient Y' share the same dysarthria severity label. We define the distance d_1 as the distance between E_{AX} and E_{BX} and d_2 as the distance between E_{AX} and E_{BY}.

The objective is to make d_1 approximately equal to d_2. This ensures that within a particular severity classification, variations in embeddings stem from the words themselves, not the speakers. Initially, we hypothesise that d_1 will be smaller than d_2. This is because the distance d_1 captures the distance between words spoken by the same speaker, and as previously discussed, our latent space tends to be influenced by speaker-specific traits.

To achieve our objective, we utilise triplet margin loss [23] with specific aims. First, we intend to push away E_{BX} from the anchor E_{AX} by considering E_{BX} as the negative sample and E_{AX} as the anchor. Given that these embeddings originate from the same speaker, we expect them to be closely located in the latent space. Thus to balance d_1 and d_2, the distance d_1 needs to be expanded. Simultaneously, we aim to pull E_{BY} closer to E_{AX} by designating E_{BY} as the positive sample and retaining E_{AX} as the anchor. Since these embeddings are from different speakers, we hypothesise that their distance in the latent space will be larger. Thus to equate d_1 and d_2, the distance d_2 should be contracted. We note that the triplet margin loss is designed to ensure the anchor embedding, E_{AX}, is nearer to the positive sample E_{BY} than to the negative sample E_{BX}, by a specific distance known as the margin, m. However, by intentionally keeping m minimal and taking into account the initial distances between the embeddings, we aim to make the distances between the anchor-positive and anchor-negative pairs approximately the same and get rid of the initial disparity.

We hypothesise that implementing this regularising loss will be beneficial because it shifts the model's focus from identifying speakers to distinguishing words. Specifically, the model should be able to differentiate between two words regardless of whether they are spoken by the same person or by different individuals with the same dysarthria severity. For example, by creating a greater distance between the anchor embedding E_{AX} and E_{BX}, the model is forced to learn an embedding that focuses on the differences between the two words rather than relying on the speaker's identity thereby making the embeddings speaker-independent given a severity class. This enhanced ability to discriminate between words allows for more accurate comparisons as it can disentangle the complexity of the word from the dysarthria.

The triplet margin loss (TML) is defined as:

$$\text{TML}(E_{AX}, E_{BY}, E_{BX}) = \max(0, \quad d(E_{AX}, E_{BX}) - d(E_{AX}, E_{BY}) + m) \tag{1}$$

In Eq.1, E_{AX} acts as the anchor, E_{BY} is the positive sample, and E_{BX} is the negative sample. The term m is a predefined margin set to 0.05. The distance function $d(x, y)$ is chosen to be the L_2 Euclidean distance.

The final loss L is expressed as $L = \epsilon L_{\text{reg}} + \gamma L_{\text{CE}}$, where ϵ serves as the weighting parameter, and is set at 0.01. The parameter γ starts at 0 for 3000 steps, allowing the model to focus on contrastive regularisation. It is then updated to 1, incorporating cross-entropy loss into the training regimen. These parameters were determined through experimentation on the 10 control patients, data we do not use in training or testing. This helps to show the robustness and generalisability of our approach, as the parameters were not tuned on data from the control group, thereby validating its potential for real-world clinical applications.

Baseline models

To contextualise our findings, we compare our results with three established baselines: XGBoost [24], Multi-layer Perceptron (MLP) [25], and CNN-LSTM [26]. XGBoost is a gradient-boosted decision tree algorithm designed for speed and performance, which excels in classification and regression tasks. It iteratively corrects the mistakes of the previous trees, and the final prediction is the sum of the predictions from all the trees. Multi-layer Perceptron (MLP) is a class of feedforward artificial neural networks, consisting of at least two layers of nodes. Each node is a neuron with a nonlinear activation function. CNN-LSTM is a hybrid neural network model that combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs). This model is capable of capturing both spatial and temporal dependencies in data, making it particularly suitable for sequence prediction problems.

For models requiring tabular data (XGBoost and MLP), we utilised the extended

Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [27], an extensive feature set of 88 acoustic features for speech analysis. For the LSTM-CNN model, we opted for an end-to-end learning approach with spectrograms as input. Spectrograms are visual representations of the spectrum of frequencies of a signal as they vary with time, and are especially useful for capturing the irregularities in dysarthric speech. To optimise the hyperparameters of our baseline models, we employed Bayesian optimisation using the Optuna library [28]. This method systematically explores the hyperparameter space to find the optimal set, balancing exploration and exploitation.

Ethics Statement

The UA speech dataset employed in this study was previously published [20] and is accessible upon request, subject to ethical considerations by the original authors. All participants in the dataset were adults (over 18 years of age) and provided explicit consent for the use and dissemination of their data. This study's use of the dataset aligns with the original consent and does not require further approval from our Institutional Review Board.

Results

Speaker-dependent vs speaker-independent splits

In our investigation, we examined both speaker-dependent and speaker-independent data splits for training and evaluating models for dysarthric speech severity classification. The speaker-dependent split facilitates the model's training and testing on data from the same individuals, albeit with different words during the testing phase. Although this method aids in model training due to the consistency of voice patterns, its applicability in a clinical environment is restricted, as it fails to validate the model against new patients—a fundamental requirement for an automated diagnostic tool. Consequently, we focused on the speaker-independent data split setup, in which the model is trained and assessed on data from distinct groups of speakers. This ensures the model's capacity to generalise across unfamiliar voices. Our study presents findings on the speaker-independent multi-class severity classification task, requiring the model to categorise the severity of dysarthric speech into four distinct levels: very low, low, medium, and high. This approach is vital as it aligns directly with clinical relevance and the model's efficacy across diverse patient conditions.

Speaker-independent multi-class severity results

In this study, we evaluated the performance of the models using a leave-one-subject-out cross-validation method. With a total of 15 speakers in our dataset, we conducted 15 iterations of training and testing, recording the average test results. In each iteration, data from 14 subjects (comprising 465 utterances for each subject, with three repetitions of 155 digits/alphabets/common words from the dataset) were used for training. The 300 uncommon words from the remaining subject were used for testing. This process was systematically repeated for all 15 subjects, ensuring each subject's data was exclusively utilised for testing once. This rigorous methodology guarantees the robustness and reliability of our findings for not only new dysarthric patients but also new vocabulary. The leave-one-subject-out cross-validation process was repeated five times, with mean and standard deviation values recorded.

Tables 1 and 2 present the performance metrics of our proposed frameworks compared with the baseline models: XGBoost, MLP, and LSTM-CNN. The baseline models demonstrated sub-optimal performance, each yielding an accuracy below 50%. Conversely, the fine-tuned wav2vec model achieved an accuracy of 64.81%. Notably, our innovative SALR framework surpassed all comparative models, including the fine-tuned wav2vec 2.0 model, achieving the highest accuracy of $70.48 \pm 1.11\%$ and the highest F1 score of $59.23 \pm 1.54\%$.

While aggregate metrics such as F1 score and accuracy provide substantial insights, further insight is obtained through the analysis of the confusion matrices. The confusion matrices (Fig 2) illustrate our models' proficiency in classifying extreme dysarthric severities but also highlight challenges in differentiating between low and medium severity classes. Specifically, the fine-tuned wav2vec 2.0 model (Fig 2A) frequently misclassified instances of low as medium severity and medium as low and high. In comparison, the SALR framework (Fig 2B) struggled with distinguishing medium instances, often mis-classifying medium as low severity.

Table 1. Table comparing the performance of various models, presenting mean ± standard deviation of accuracy scores for each of the 15 patients across five iterations of leave-one-subject-out cross-validation.

Patient Code	MLP	LSTM-CNN	XGBoost	Finetuned wav2vec 2.0	SALR
M04	51.78 ± 1.84	73.49 ± 2.34	62.83 ± 2.98	87.19 ± 0.95	79.62 ± 0.79
F03	57.39 ± 3.22	70.32 ± 4.33	61.91 ± 2.43	89.46 ± 0.84	77.00 ± 0.67
M12	61.89 ± 1.84	74.47 ± 2.33	71.89 ± 2.42	92.40 ± 0.74	88.64 ± 0.93
M01	56.74 ± 2.33	60.83 ± 3.42	63.84 ± 1.89	78.13 ± 1.09	81.18 ± 0.93
M07	15.98 ± 5.84	7.38 ± 4.84	10.18 ± 4.33	7.67 ± 1.89	20.00 ± 1.57
F02	9.39 ± 3.84	5.38 ± 3.33	7.85 ± 5.75	22.74 ± 1.89	21.20 ± 2.39
M16	13.73 ± 4.39	8.48 ± 4.72	11.43 ± 3.27	26.02 ± 2.00	19.12 ± 1.32
M11	8.48 ± 3.28	13.49 ± 5.47	13.43 ± 4.33	32.49 ± 1.04	58.10 ± 1.32
F04	6.48 ± 3.82	7.43 ± 4.37	11.85 ± 4.46	25.83 ± 2.38	61.60 ± 1.01
M05	9.04 ± 4.84	9.49 ± 5.79	16.89 ± 4.23	26.58 ± 1.00	60.53 ± 1.39
M09	73.38 ± 2.38	72.78 ± 2.80	81.89 ± 2.89	92.58 ± 0.89	98.43 ± 0.89
M08	72.37 ± 2.80	80.41 ± 3.24	78.94 ± 1.98	95.55 ± 0.79	97.20 ± 0.71
M10	76.47 ± 1.98	75.49 ± 1.32	81.89 ± 2.89	96.05 ± 1.84	97.70 ± 0.90
M14	77.90 ± 2.81	79.95 ± 1.39	79.80 ± 1.39	97.98 ± 0.67	98.10 ± 0.90
F05	74.24 ± 3.81	72.38 ± 1.43	80.84 ± 1.23	95.36 ± 0.84	98.80 ± 0.93
Average	44.35 ± 3.27	47.45 ± 3.34	49.03 ± 2.98	64.81 ± 1.26	**70.48** ± 1.11

Table 2. Table comparing the performance of various models, presenting mean ± standard deviation of F1 score across five runs of leave-one-subject-out cross-validation

MLP	LSTM-CNN	XGBoost	Finetuned wav2vec 2.0	SALR
27.44 ± 3.83	29.95 ± 4.15	37.75 ± 3.20	52.39 ± 2.13	**59.23** ± 1.54

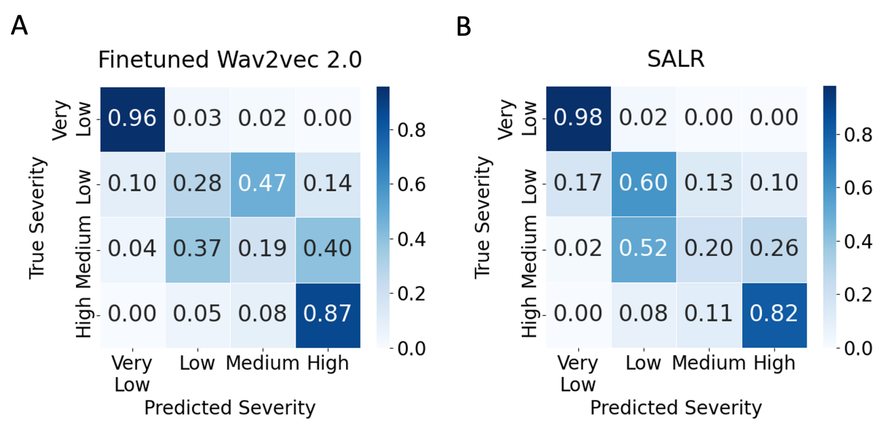


Fig 2. Normalised confusion matrices. A. Fine-tuned wav2vec 2.0 model **B.** SALR framework

To contextualise our findings within the broader scope of existing research, we compared our results with those from previous studies on speaker-independent dysarthria classification. Tripathi et al. [11] reported the highest accuracy of 53.90%

using features obtained from DeepSpeech—a deep learning-based speech-to-text engine—combined with an SVM classifier under a leave-one-subject-out cross-validation scheme. In contrast, our initial results using a fine-tuned wav2vec 2.0 model showed a better classification accuracy of 64.81%. We achieved further improvements using our SALR framework, which reached an accuracy of 70.48% (see Fig 3). This comparative analysis highlights the significant advancements made by our SALR framework over previous methods. Utilising the same test set and cross-validation scheme, our study ensures a rigorous and fair comparison, demonstrating notable enhancements in methodological approach and classification accuracy, crucial for effective implementation in various clinical settings.

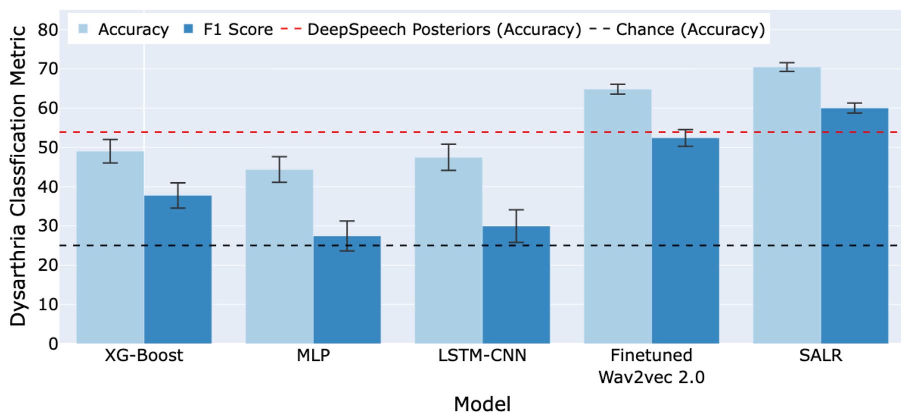


Fig 3. Comparative performance of various models for speaker-independent multi-class dysarthria severity classification on the UA-Speech dataset. Accuracy and F1 scores of our models compared to chance predictions and the existing benchmark set by Tripathi et al. [11] using DeepSpeech posteriors. Error bars represent the standard deviation across five repetitions, illustrating the consistency of model performance.

Interpretation of the latent space analysis

To further assess the impact of our frameworks on the model's representation of speech data, we conducted a t-SNE analysis of the latent space. Fig 4 provides visual insights into how the models organise the latent representations of both the fine-tuned wav2vec 2.0 model and the SALR framework with respect to dysarthria severity and speaker identity.

In the fine-tuned wav2vec 2.0 model, the latent space displays a lack of structured organisation with respect to ordinal severity levels. High-severity samples often cluster closely to both mid and low-severity samples (Fig 4A). Additionally, distinct clusters corresponding to different speakers are evident (Fig 4C). In contrast, the SALR multi-task framework introduces a clearer ordinal structure to the latent space (Fig 4B). Speaker clusters within this framework are also more dispersed (Fig 4D). These observations highlight the effectiveness of the contrastive loss in our SALR framework, which successfully disentangles speaker-specific cues from severity assessments in the latent space.

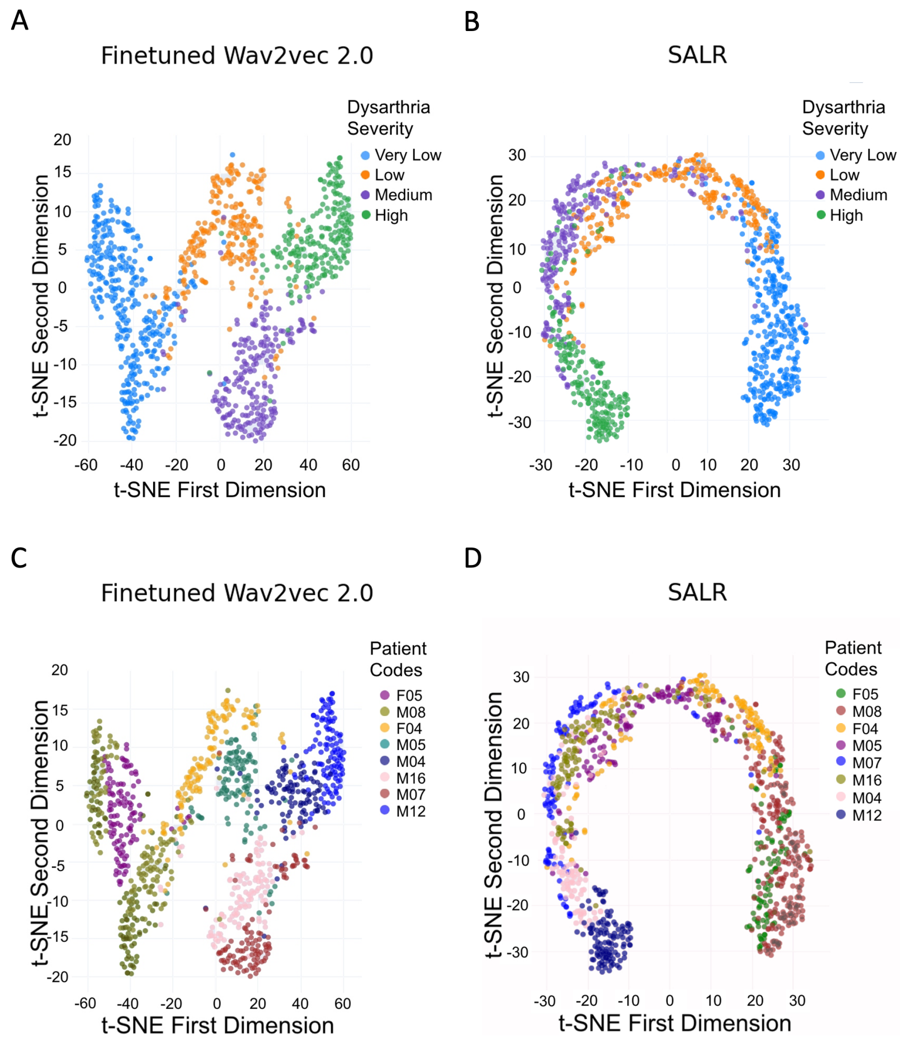


Fig 4. Visualisation of t-SNE embeddings. Data points are coloured according to patient severity (first row) and patient code (second row) for the fine-tuned wav2vec 2.0 model **(A, C)** and the SALR framework **(B, D)**. These visualisations support our hypothesis that the SALR framework organises the latent space in alignment with severity levels (first row) and disperses speaker clusters (second row).

Discussion

The primary focus of the study was developing an automated, machine learning-based tool for classifying dysarthria severity levels in a speaker-independent manner. We first fine-tuned a base wav2vec 2.0 model, a state-of-the-art self-supervised transformer model, trained on healthy speech for the task of dysarthria severity assessment. The fine-tuned model outperformed other traditional baseline models based on XGBoost, MLP, and LSTM-CNN. Our analysis indicated that the model could achieve more accurate results by focusing on dysarthria-specific speech features rather than on individual idiosyncrasies unrelated to the actual severity of the condition. To counteract the model's tendency to overfit to individual speakers and improve generalisation, we introduced the novel SALR multi-task framework. This framework significantly improved the model's performance, achieving an accuracy of 70.48% and an F1 score of 59.23%, marking a 16.58% increase over the published benchmark. Further analysis shows that the SALR multi-task objectives not only enhance numerical performance but also help organise the latent space in a manner that aligns with severity levels. This reduces the model's reliance on speaker-specific cues, thus boosting performance and validating our hypothesis about the effects of the multi-task framework.

Using confusion matrices for speaker-independent evaluation, we found that while the framework excels in categorising extreme severity classes, it faces challenges in distinguishing between low and medium severity levels. These challenges are primarily attributable to the limited number of patient samples available for these categories post-segmentation, with only two patients remaining in each of the low and medium categories, thereby limiting the model's learning efficacy. Compounding this issue is the lack of distinct boundaries between these classes. For instance, patient M16 categorised under medium severity with an intelligibility rating of 43%, stands in stark contrast to other individuals within the same category, such as M07 and F02, who have ratings of 28% and 29%, respectively. Conversely, the lowest-rated individual in the adjacent low category, M05, had a rating of 58%. This disparity in intelligibility ratings approximately equates the gap between M16 and either its own category or the neighbouring low category, thereby blurring the classification boundaries. Consequently, the model's ability to accurately classify ambiguous cases like M16 may be compromised due to the combined factors of data sparsity and ambiguous class distinctions.

While the use of transformer-based frameworks is effective, it introduces challenges related to interpretability. Although our latent space visualisation provides some insights into the model's functionality, it is beneficial to adopt additional methods, such as attention heatmaps or layer-wise relevance propagation, to gain a fuller understanding of the model's decision-making processes. This is particularly important as we move toward automated dysarthria severity assessments, where transparency and interpretability are crucial. Another promising direction for advancing the field might be exploring self-supervised pre-training tasks on dysarthric samples rather than on normal speech. Researchers should be aware of the computational requirements for this approach, as seen in the original training of the wav2vec 2.0 model, which utilised 64 GPUs [19].

The implications of our findings for clinical practice and research are substantial. The ability of the SALR framework to provide reliable and accurate assessments of dysarthria severity in a speaker-independent manner is particularly relevant for clinical settings. Integrating automated tools like our framework in clinical practice could significantly enhance diagnostic processes. This advancement could facilitate more objective and efficient assessments of dysarthria, contributing to improved patient care and management [29]. SALR offers the opportunity to reduce the unmet demand for speech and language assessments in terms of both quantity and quality. This is particularly important for healthcare systems strained by staff shortages, rapidly ageing populations, and increased healthcare service demand. Speech and language therapy is a profession with substantial training requirements, limiting the availability of experts in many countries. For example, the UK has a vacancy rate of around 25%and recognises it as a shortage profession [30]. AI-based speech assessment could support diagnostic assessments and assist in training professionals, particularly in the initial stages of their training, e.g. see [31, 32]. In daily operations, this technology could be used in conjunction with human raters or as an autonomous system for rapid initial assessments; or, with more training data, for systematic assessments.

The accessibility and cost-effectiveness of our AI-based approach could enhance the speed and precision of dysarthria assessments, particularly benefiting individuals with mobility challenges due to co-occurring physical disabilities [33] and thus potentially allow assessment via videoconferencing, as in dermatology [34]. Crucially, remote or home assessment would enable true patient-centric evaluation of patient capability, a rapidly growing domain of digital healthcare [35, 36]. Thus, speech rehabilitation of dysarthria could be potentially even entirely technologically guided as in other forms of motor rehabilitation in a multi-modal, multi-sensory AI-guided treatment in the real world [37, 38] for smart rehabilitation. However, any real-world deployment would require careful assessment and comparison of commercial and clinical grade speech recording methodologies, as was done in other domains of sensing [39].

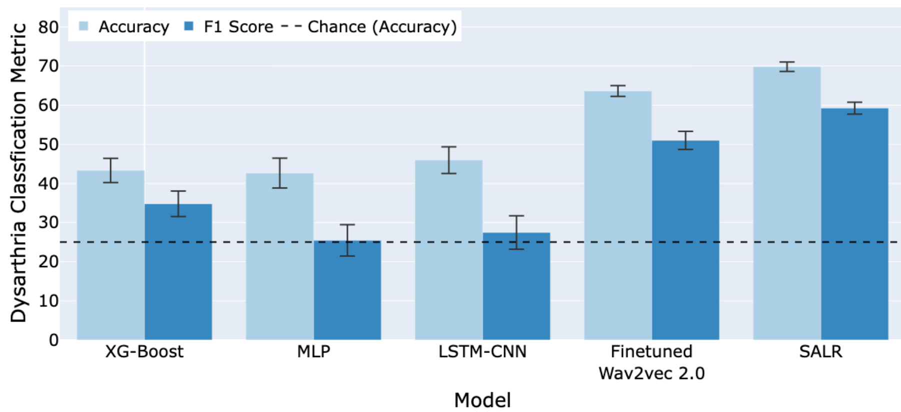
The SALR framework could serve as a valuable tool for monitoring disease progression and rehabilitation progression, offering insights into the efficacy of interventions over time – turning SALR into a digital biomarker. These digital insights can inform both the development of targeted therapies and the refinement of existing treatment protocols [40]. Our methodology lays the groundwork for further developments using self-supervised and semi-supervised transformer-based learning models in other areas of biomedical time series, where data scarcity, inter-patient variability and the need for high generalisability are common challenges [41, 42].

Conclusion

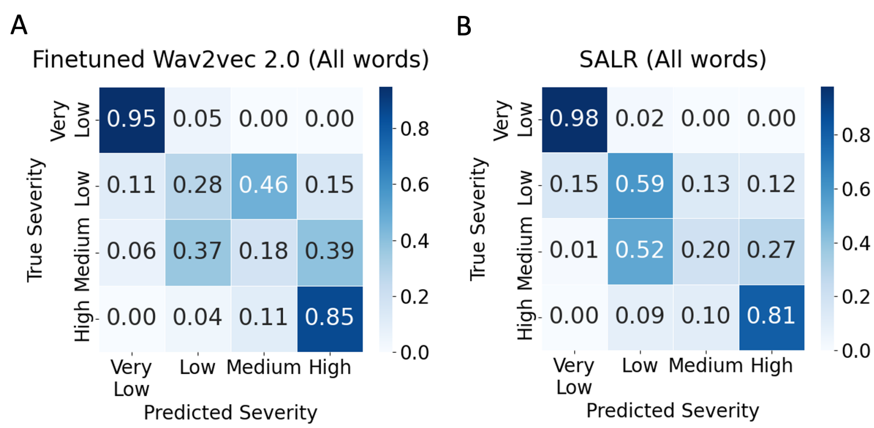
Our results demonstrate that we can address long-standing challenges in dysarthria severity assessment by introducing a novel multi-task deep-learning framework leveraging the wav2vec 2.0 transformer model. Our automated approach sets a new benchmark for speaker-independent, multi-class dysarthria severity classification on the Universal Access speech dataset, showing substantial improvements in accuracy. These findings highlight the potential of our method to offer more precise, efficient, and clinically relevant automated assessments of dysarthria severity.

Supporting information

S1 Fig. Comparative performance of various models for speaker-independent multi-class dysarthria severity classification on the UA-Speech dataset (tested on all words of the test subject using leave-one-subject-out cross-validation). We have plotted the accuracy and F1 scores of our models compared to chance predictions for the case when all 765 utterances from the test subject were used in the test set. This test case checks for system performance for new speakers (but not new vocabulary). Error bars represent the standard deviation across five repetitions.



S2 Fig. Normalised confusion matrices for the case when all 765 utterances from the test subject were used in the test set. **A.** fine-tuned wav2vec 2.0 model, **B.** SALR framework



Acknowledgments

AAF acknowledges the support from the UKRI Turing AI Fellowship (EP/V025449/1). BK acknowledges support from the NIHR Research Support Service.

Authors' Contributions

AAF conceptualised the study. LS performed the analysis with technical inputs from BK & AAF and clinical inputs from SW. BK reviewed the data analysis. LS and BK wrote the initial draft of the manuscript. BK wrote the revised draft of the manuscript, which AAF reviewed and edited. All authors reviewed the manuscript.

Conflicts of Interest

None declared.

References

1. Darley FL, Aronson AE, Brown JR. Differential diagnostic patterns of dysarthria. Journal of speech and hearing research. 1969;12(2):246–269.

2. Whitehill TL, Ciocca V. Speech errors in Cantonese speaking adults with cerebral palsy. Clinical linguistics & phonetics. 2000;14(2):111–130.

3. Scott S, Caird FI. Speech therapy for Parkinson's disease. Journal of Neurology, Neurosurgery & Psychiatry. 1983;46(2):140–144.

4. Joy NM, Umesh S. Improving acoustic models in TORGO dysarthric speech database. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2018;26(3):637–645.

5. Freed DB. Motor speech disorders: diagnosis and treatment. Plural Publishing; 2018.

6. Gavidia-Ceballos L, Hansen JHL. Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection. IEEE Transactions on Biomedical Engineering. 1996;43(4):373–383.

7. Baghai-Ravary L, Beet SW. Automatic speech signal analysis for clinical diagnosis and assessment of speech disorders. Springer Science & Business Media; 2012.

8. Gupta S, Patil AT, Purohit M, Parmar M, Patel M, Patil HA, et al. Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. Neural Networks. 2021;139:105–117.

9. Shih DH, Liao CH, Wu TW, Xu XY, Shih MH. Dysarthria Speech Detection Using Convolutional Neural Networks with Gated Recurrent Unit. Healthcare (Switzerland). 2022;10. doi:10.3390/healthcare10101956.

10. Joshy AA, Rajan R. Automated Dysarthria Severity Classification: A Study on Acoustic Features and Deep Learning Techniques. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2022;30:1147–1157. doi:10.1109/TNSRE.2022.3169814.

11. Tripathi A, Bhosale S, Kopparapu SK. Improved speaker independent dysarthria intelligibility classification using deepspeech posteriors. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020. p. 6114–6118.

12. Tong H, Sharifzadeh H, McLoughlin I. Automatic assessment of dysarthric severity level using audio-video cross-modal approach in deep learning. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. vol. 2020-October. International Speech Communication Association; 2020. p. 4786–4790.

13. Joshy AA, Rajan R. Dysarthria severity classification using multi-head attention and multi-task learning. Speech Communication. 2023;147:1–11. doi:10.1016/j.specom.2022.12.004.

14. Al-Ali A, Al-Maadeed S, Saleh M, Naidu RC, Alex ZC, Ramachandran P, et al. The Detection of Dysarthria Severity Levels Using AI Models: A Review. IEEE Access. 2024;.

15. Hawley MS, Enderby P, Green P, Cunningham S, Brownsell S, Carmichael J, et al. A speech-controlled environmental control system for people with severe dysarthria. Medical Engineering & Physics. 2007;29(5):586–593.

16. Christensen H, Cunningham SP, Fox C, Green PD, Hain T. A comparative study of adaptive, automatic recognition of disordered speech. In: Interspeech. Portland; 2012. p. 1776–1779.

17. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.

18. Gong YA, Chung YA, Glass J. AST: Audio spectrogram transformer. arXiv preprint arXiv:210401778. 2021;.

19. Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems. 2020;33:12449–12460.

20. Kim H, Hasegawa-Johnson M, Perlman A, Gunderson JR, Huang TS, Watkin KL, et al. Dysarthric speech database for universal access research. In: Interspeech. vol. 2008; 2008. p. 1741–1744.

21. Hsu WN, Bolte B, Tsai YHH, Lakhotia K, Salakhutdinov R, Mohamed A. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021;29:3451–3460.

22. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:191003771. 2019;.

23. Dong X, Shen J. Triplet loss in siamese network for object tracking. In: Proceedings of the European conference on computer vision (ECCV); 2018. p. 459–474.

24. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 785–794.

25. Murtagh F. Multilayer perceptrons for classification and regression. Neurocomputing. 1991;2(5-6):183–197.

26. Shih DHH, Liao CH, Wu TW, Xu XY, Shih MH. Dysarthria Speech Detection Using Convolutional Neural Networks with Gated Recurrent Unit. Healthcare. 2022;10:1956.

27. Eyben F, Scherer KR, Schuller BW, Sundberg J, André E, Busso C, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE transactions on affective computing. 2015;7(2):190–202.

28. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining; 2019. p. 2623–2631.

29. Chandrashekar H, Karjigi V, Sreedevi N. Breathiness indices for classification of dysarthria based on type and speech intelligibility. In: 2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET). IEEE; 2019. p. 266–270.

30. Royal College of Speech and Language Therapists. Workforce Planning in England. Royal College of Speech and Language Therapists; 2023. Available from: `https://www.rcslt.org/wp-content/uploads/2023/04/Workforce-planning-in-England.pdf`.

31. Milling M, Pokorny FB, Bartl-Pokorny KD, Schuller BW. Is speech the new blood? recent progress in ai-based disease detection from audio in a nutshell. Frontiers in digital health. 2022;4:886615.

32. Deka C, Shrivastava A, Abraham AK, Nautiyal S, Chauhan P. AI-based automated speech therapy tools for persons with speech sound disorder: a systematic literature review. Speech, Language and Hearing. 2024; p. 1–22.

33. Enderby P, Cantrell A, John A, Pickstone C, Fryer K, Palmer R. Guidance for providers of speech and language therapy services: dysarthria. Asia Pacific Journal of Speech, Language and Hearing. 2010;13(3):171–190.

34. Brinker TJ, Hekler A, Von Kalle C, Schadendorf D, Esser S, Berking C, et al. Teledermatology: comparison of store-and-forward versus live interactive video conferencing. Journal of medical Internet research. 2018;20(10):e11871.

35. Ricotti V, Kadirvelu B, Selby V, Festenstein R, Mercuri E, Voit T, et al. Wearable full-body motion tracking of activities of daily living predicts disease trajectory in Duchenne muscular dystrophy. Nature medicine. 2023;29(1):95–103.

36. Kadirvelu B, Gavriel C, Nageshwaran S, Chan JPK, Nethisinghe S, Athanasopoulos S, et al. A wearable motion capture suit and machine learning predict disease progression in Friedreich's ataxia. Nature Medicine. 2023;29(1):86–94.

37. Haar S, Faisal AA. Brain activity reveals multiple motor-learning mechanisms in a real-world task. Frontiers in Human Neuroscience. 2020;14:354.

38. Haar S, Sundar G, Faisal AA. Embodied virtual reality for the study of real-world motor learning. Plos one. 2021;16(1):e0245717.

39. Auepanwiriyakul C, Waibel S, Songa J, Bentley P, Faisal AA. Accuracy and acceptability of wearable motion tracking for inpatient monitoring using smartwatches. Sensors. 2020;20(24):7313.

40. Beijer L, Rietveld T. Asynchronous telemedicine applications in rehabilitation of acquired speech-language disorders in neurologic patients. Smart Homecare Technology and TeleHealth. 2015; p. 39–48.

41. Huang SC, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digital Medicine. 2023;6(1):74.

42. Wolf D, Payer T, Lisson CS, Lisson CG, Beer M, Götz M, et al. Self-supervised pre-training with contrastive and masked autoencoder methods for dealing with small datasets in deep learning for medical imaging. Scientific Reports. 2023;13(1):20260.