
Cognitive Echo: Detecting Neurodegenerative Disorders from Speech

Ibrahim Aldarmaki

Carnegie Mellon University
Pittsburgh, PA 15213
ialdarma@andrew.cmu.edu

Hajer Dahmani

Carnegie Mellon University
Pittsburgh, PA 15213
hdahmani@andrew.cmu.edu

Nikitha Srikanth

Carnegie Mellon University
Pittsburgh, PA 15213
nsrikant@andrew.cmu.edu

Dominic Dimambro

Carnegie Mellon University
Pittsburgh, PA 15213
ddimambr@andrew.cmu.edu

Abstract

Accurate and early detection of neurodegenerative diseases like Amyotrophic Lateral Sclerosis (ALS) is critical, yet current diagnostic methods are often subjective and may miss subtle early symptoms. Speech provides a non-invasive, information-rich biomarker that reflects both cognitive and motor changes. In this work, we investigate automated classification of ALS and dysarthria severity from speech using deep learning. We implement a Vision Transformer (ViT) baseline and propose a novel architecture that outperforms this baseline through extensive experimentation and ablation studies. The final architecture, comprising a pretrained encoder, attention pooling and weighted cross-entropy loss achieved a Macro F1 score of 0.824. These results highlight the potential of speech-based biomarkers for objective monitoring and early detection of neurodegenerative disorders, paving the way for more accessible and precise clinical assessment.

1 Motivation

Early and accurate detection of neurodegenerative diseases is essential, as timely intervention can slow disease progression and significantly improve quality of life. However, current diagnostic practices rely heavily on subjective clinical assessments, which often miss subtle early symptoms and lead to delayed treatment. This limitation motivates the need for more objective, scalable, and sensitive diagnostic tools.

Speech offers a promising solution: it is non-invasive, easy to collect, and rich in cognitive and motor information. Our objective is to develop an automated speech-based biomarker that (1) qualitatively improves diagnostic objectivity by reducing reliance on subjective clinical judgment, and (2) quantitatively enhances early-stage detection accuracy by identifying subtle speech changes associated with neurodegeneration. By achieving these goals, this approach can reduce diagnostic uncertainty, enable earlier intervention, and support better long-term outcomes for patients, caregivers, and healthcare providers.

2 Objective

We aim to develop deep learning and signal-processing methods that classify disease status and predict clinically relevant severity levels directly from speech. The classification task consists of

five target categories: ALS with severe dysarthria (Class 1), ALS with moderate dysarthria (Class 2), ALS with mild dysarthria (Class 3), ALS with no dysarthria (Class 4), and Healthy (Class 5). Our objective is to improve upon the current benchmark model, a Vision Transformer (ViT), which achieves an average F1-score of 0.437.

To address this challenge, we propose a novel speech-based classification architecture that leverages pre-trained self-supervised audio models and attention-driven representation learning. At a high level, our approach extracts informative embeddings from multiple recordings per subject using Wav2Vec [2], aggregates them into a unified subject-level representation, and performs classification using a lightweight MLP classifier. This design aims to enhance sensitivity to subtle acoustic biomarkers while improving robustness across subjects, ultimately enabling more accurate disease severity classification compared to existing baselines.

3 Related work and background

3.1 Background

Speech has been widely studied as a signal for a wide range of biological and demographic indicators. For example, age and gender [6], and other cultural markers [5] can be predicted from acoustic cues. Neurodegenerative disorders such as Parkinson’s Disease (PD) , Amyotrophic Lateral Sclerosis (ALS), and Alzheimer’s Disease (AD), affect motor, cognitive, affective, and feedback pathways, that have a direct effect on speech [8]. This includes changes in voice, prosody, articulation, and language. These correlations have opened up opportunities to perform non-invasive and low-cost disease monitoring and diagnosis.

There are a wide range of features that are typically studied to diagnose speech impairments, often with low inter-rater agreement among clinicians [9]. This problem is exacerbated when some features used for diagnosis work for a certain population may not work for another. These factors highlight the need for robust and generalizable approaches.

In addition, datasets such as VOC-ALS have been developed to support research in this area, offering voice recordings from ALS patients and healthy controls across a variety of speech tasks [3]. The ICASSP 2026 Speech Analysis for Neurodegenerative Diseases (SAND) Challenge [1] builds upon this dataset, providing a standardized benchmark for model evaluation. The current baseline employs a Vision Transformer (ViT) architecture and achieves an average F1-score of 0.606, serving as a reference point for ALS detection and dysarthria severity classification.

3.2 Literature review

Recent work has focused on detecting early signs of neurodegenerative impairment, such as Mild Cognitive Impairment (MCI) or preclinical Alzheimers disease, directly from spontaneous speech [7]. The PROCESS Grand Challenge at ICASSP 2025 formalized benchmark tasks for dementia detection from speech[10]. Likewise, the SAND challenge emphasizes speech as a key modality for diagnosing and tracking neurodegenerative diseases [1].

Several existing approaches to speech-based disease detection rely on machine learning models built around large sets of handcrafted features. One line of work, for example, employs thousands of acoustic, articulatory, and sensory features to train ALS detection models [11]. Other work combines predetermined speech markers of cognitive decline with acoustic embeddings derived from models such as Whisper to detect Alzheimer’s disease [4]. While incorporating deep embeddings reduces interpretability, it enables the discovery of discriminative patterns that hand-engineered features may overlook. Moreover, acoustic embeddings preserve temporal detail and facilitate multimodal integration. With continued advances in deep learning, there remains substantial opportunity to explore richer multimodal fusion strategies informed by domain expertise.

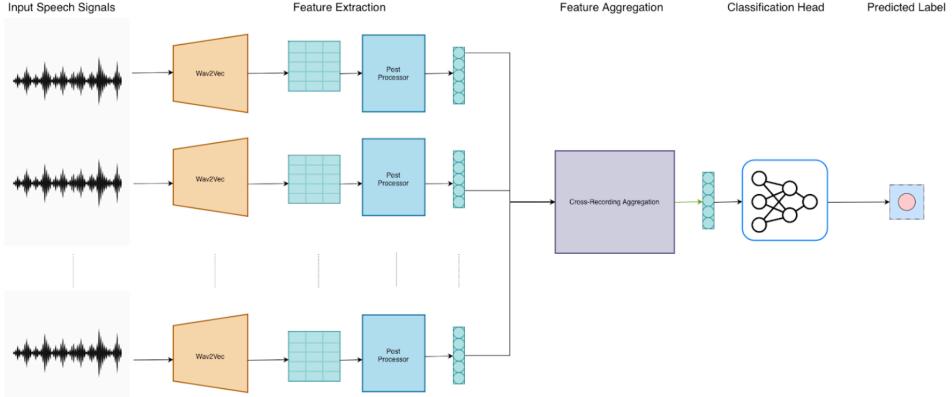


Figure 1: Our proposed ALS detection model.

Table 1: Model architecture and main hyper-parameters for ALS detection.

Component	Description	Hyper-parameters / Notes
Input	Multi-phonation audio, shape $[B, R, T]$, $R=8$ phonations.	Sample rate: 16 kHz; max length 5s
Feature extractor	Wav2Vec2-base + LoRA, input $[B, R, T] \rightarrow [B, R, T_{feat}, D]$.	LoRA: $r = 8$, $\alpha = 16$, modules: {q,k,v_proj}; $D = 768$
Feature / padding mask	Downsamples attention mask $[B, R, T] \rightarrow [B, R, T_{feat}]$.	Handles time padding
Sequence post-processor	Attention-based pooling over T_{feat} per recording.	$[B, R, T_{feat}, D] \rightarrow [B, R, D]$
Multi-recording fusion	Fuses recordings by mean: $[B, R, D] \rightarrow [B, D]$.	Fusion = mean
Classifier	2-layer MLP: Linear(D→256) → ReLU → Linear(256→num_classes).	Hidden dim = 768; num classes = 5
Loss	Weighted cross-entropy + label smoothing 0.05.	Class weights = $1/\sqrt{\text{freq}}$, normalized
Optimizer & scheduler	AdamW; ReduceLROnPlateau (monitor val loss).	lr=3e-4, weight_decay=1e-4; factor=0.5, patience=6, min_lr=1e-6
Training setup	FP16 training, WeightedRandomSampler for class balance.	Batch size = 3; epochs = 50

4 Methodology

4.1 Model description

Figure 1 is an overview of the architecture of our ALS detection model. As shown in Table 1, the input consists of multi-phonation audio recordings resampled at 16 kHz. Features are extracted using a pretrained Wav2Vec2-base model with LoRA adapters. Attention-based pooling is applied over the temporal dimension of each recording, followed by mean fusion across recordings. Classification is performed via a 2-layer MLP with 5 output classes. The model is trained with weighted cross-entropy with label smoothing (0.05) and class weights $1/\sqrt{\text{freq}}$. Optimization uses AdamW with a ReduceLROnPlateau scheduler. FP16 mixed-precision training and a WeightedRandomSampler handle class imbalance.

4.2 Dataset

We use the SAND Dataset [1, 3], which contains voice recordings from 339 Italian speakers, including 205 ALS patients (121 males, 84 females) and 134 healthy controls (72 males, 62 females), aged 18–90. Each participant performed two types of speech tasks:

1. Sustained vowels (/a/, /e/, /i/, /o/, /u/) held for at least 5 seconds.
2. Rapid syllable repetitions (/pa/, /ta/, /ka/) spoken in a single breath.

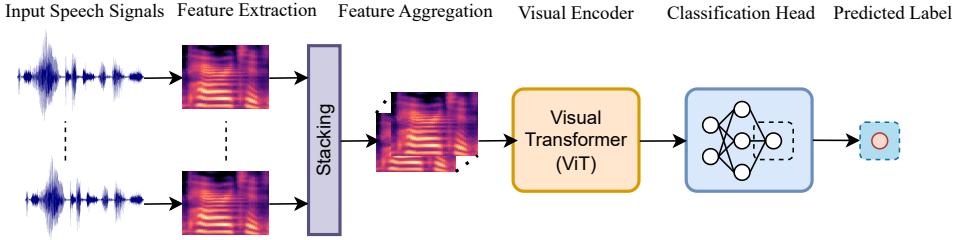


Figure 2: Overview of the speech classification pipeline based on Mel-spectrogram features and a Vision Transformer (ViT) encoder. Each input speech signal is converted into a Mel-spectrogram, resulting in eight feature maps that are stacked to form a combined acoustic representation. The concatenated spectrogram is processed by the ViT to learn contextual time-frequency representations, followed by an MLP classification head that outputs the final predicted class label.

ALS severity is labeled using ALSFRS-R scores, an integer value from 1 to 4 for ALS patients, or 5 for healthy subjects, defining five classes:

- Class 1: ALS with Severe dysarthria
- Class 2: ALS with Moderate dysarthria
- Class 3: ALS with Mild dysarthria
- Class 4: ALS with no dysarthria
- Class 5: Healthy controls

The distribution of classes in the given training set is:

- Class 1: 2.2%
- Class 2: 9.55%
- Class 3: 20.95%
- Class 4: 27.94%
- Class 5: 39.33%

Metadata for each participant includes ID, age, sex, and, for longitudinal assessments, the number of months between recordings and ALSFRS-R scores at each assessment. Figures 3 and 4 present the class distribution by sex and the age distribution across the dataset, respectively.

Each subject provides eight phonations (/a/, /e/, /i/, /o/, /u/, /pa/, /ta/, /ka/), sampled at 16 kHz with a maximum duration of 5 seconds.

Waveforms are augmented during training with Gaussian noise, random gain, and speed perturbation. Each recording is padded per batch to the maximum length, producing attention masks to ignore padding during feature extraction. Class imbalance is addressed using a *WeightedRandomSampler* based on the inverse square root of class frequencies.

Features are extracted using a pretrained Wav2Vec2 model with LoRA adapters, producing hidden states per recording. Attention-based temporal pooling summarizes each recording, followed by mean fusion across recordings. A 2-layer MLP performs classification.

During evaluation, the best model is selected based on validation macro-F1.

4.3 Evaluation Metric

We use the macro-F1 score to evaluate the performance of our classification models. This metric provides a balanced measure of performance across all classes, making it suitable for assessing ALS/dysarthria severity without allowing majority classes to dominate.

For each class $c \in \{0, 1, 2, 3, 4\}$ corresponding to the five severity levels, we compute precision, recall, and F1-score in a one-vs-rest manner. Let:

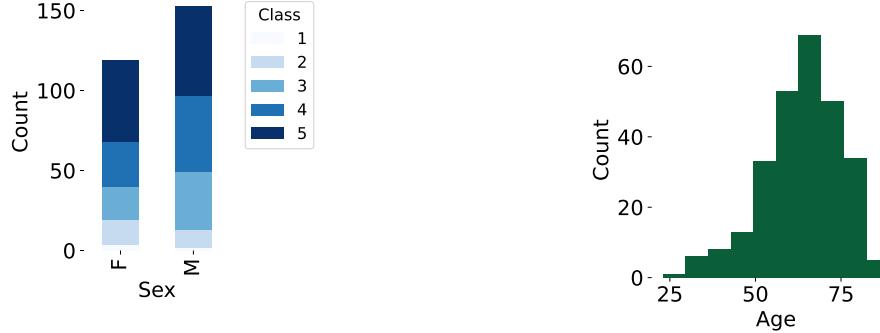


Figure 3: Stratified distribution of class labels across males and females.

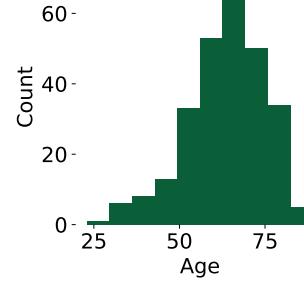


Figure 4: Histogram of age distribution.

TP_c : number of true positive for class c
 FP_c : number of false positives for class c
 FN_c : number of false negatives for class c

Then:

$$Precision_c = \frac{TP_c}{TP_c + FP_c}, Recall_c = \frac{TP_c}{TP_c + FN_c}.$$

Then the F1-score for class c is computed directly as:

$$F1_c = \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c}$$

The **macro-F1 score** is the average of the class-wise F1-scores:

$$Macro-F1 = \frac{1}{C} \sum_{c=0}^{C-1} F1_c, \quad C = 5$$

This formulation ensures that each class contributes equally to the overall score, making it robust to class imbalance.

4.4 Loss function

We use the weighted Cross-Entropy (CE) loss as the loss function for our classification model because it provides the best trade-off: it allows rare dysarthria classes to contribute more to the gradient while avoiding instability caused by overcorrecting.

Let $count_c$ denote the number of training samples for class c , and let $freq_c$ denote the normalized frequency:

$$freq_c = \frac{count_c}{\sum_{i=1}^C count_i}, \quad C = 5$$

The weight for class c is then:

$$w_c = \frac{1}{\sqrt{freq_c}}$$

Finally, the weights are normalized by their mean to ensure numerical stability:

$$w_c^{norm} = \frac{w_c}{\frac{1}{C} \sum_{i=1}^C w_i}$$

The Weighted CE Loss is then computed as:

$$L = - \sum_{c=1}^C w_c^{norm} \cdot y_c \log(\hat{y}_c)$$

where y_c is the ground-truth label and \hat{y}_c is the predicted probability for class c . This weighting scheme ensures that underrepresented classes contribute more to the learning process without destabilizing training, resulting in improved performance on imbalanced ALS severity datasets.

The model parameters θ are optimized using the Adam optimizer with weight decay:

$$\theta \leftarrow \theta - \eta \frac{\hat{m}}{\sqrt{\hat{v}} + \epsilon} - \lambda \theta$$

where:

- $\eta = 3 \times 10^{-4}$ is the learning rate,
- \hat{m} and \hat{v} are the bias-corrected first and second moment estimates,
- $\epsilon = 10^{-8}$ is a small constant for numerical stability,
- $\lambda = 10^{-4}$ is the weight decay coefficient.

A learning rate scheduler ('ReduceLROnPlateau') reduces η when the validation loss plateaus. Mixed-precision (FP16) training is used on compatible GPUs to improve efficiency.

4.5 Experimental depth and iteration

We conducted a series of experiments to determine the optimal configuration for our ALS detection model. The experimentation proceeded in two main stages: (1) evaluating different loss functions to handle class imbalance, and (2) testing different feature extractors and temporal pooling strategies.

Loss Function Ablations

Table 2 summarizes the impact of various loss functions given the class imbalance in our dataset.

Loss Function	Macro F1
Baseline CE	0.427
Weighted CE ($w = 1/\sqrt{freq_c}$)	0.454
Class-Balanced Focal Loss ($\gamma = 1.5$)	0.351
Class-Balanced Focal Loss ($\gamma = 1.0$)	0.272

Table 2: Comparison of different loss functions. The highlighted weighted cross-entropy loss performs the best, indicating better handling of underrepresented ALS severity classes.

Feature Extractor and Pooling Ablations

After selecting the weighted cross-entropy loss, we evaluated different feature extractors and temporal pooling methods using Wav2Vec-LoRA embeddings. Results are summarized in Table 3.

Method	Val Loss	Macro F1	Comments
ViT Baseline	1.42	0.437	Input features: MFCC-Spectrograms
Wav2Vec-LoRA + Oversampling	1.21	0.635	Wav2Vec embeddings, inverse-sqrt class sampling
Feature Extractor: Attention	0.943	0.824	Learnable temporal pooling over hidden states
Feature Extractor: Convolution	1.01	0.793	1D temporal convolution over feature sequence
Feature Extractor: Multi-Head Attention	1.25	0.637	Multiple attention heads capture diverse patterns

Table 3: Impact of different feature extractors and pooling strategies on validation loss and Macro F1. The highlighted Attention-based temporal pooling provided the best performance.

5 Baseline and extensions

5.1 Baseline selection and evaluation

For our baseline, we attempted to replicate the Vision Transformer (ViT) setup described in the SAND Challenge [1], which reports an average F1-score of 0.606 for ALS detection and dysarthria severity classification. The original challenge provides no details regarding the model setup, architecture, or data preprocessing. Therefore, we implemented the baseline ourselves, experimenting with different configurations and conducting ablation studies to evaluate the impact of various design choices.

We implemented a ViT-based model adapted for multi-phonation audio input as follows:

- **ViT encoder:** We use a ViT model (`vit_base_patch16_224`) from the `timm` library without pretraining. The model takes Mel-spectrograms as input and outputs embedding features rather than classification logits.
- **Phonation aggregation:** Embeddings from multiple phonations are averaged to produce a single participant-level representation.
- **Classification head:** A dropout layer followed by a fully connected layer maps the aggregated features to the five ALS/dysarthria classes.

Vision Transformers were chosen as the baseline due to their strong capability to model global dependencies through self-attention, which is particularly beneficial for capturing complex temporal-spectral relationships in speech spectrograms. Unlike convolutional architectures that focus on local patterns, ViTs capture a broader view of the input, helping the model identify subtle differences in articulation and voice patterns related to ALS and dysarthria. Figure 2 illustrates the overall architecture of the baseline we chose to implement.

5.2 Implemented extensions/experiments

Our baseline ViT model achieved a macro-F1 score of **0.437** as shown in Table 4. While this established a functional foundation, several limitations motivated the design of our model extensions: MFCCs discard important fine-grained acoustic detail, the dataset is highly imbalanced across severity levels, and temporal pooling fails to prioritize segments of speech that are informative for dysarthria. The following extensions were implemented to address these issues.

Loss function variants

We first explored ways to mitigate class imbalance. Standard cross-entropy underweights rare-dysarthria classes, leading the model to favor predictions on majority classes. Two alternatives were evaluated:

- **Weighted cross-entropy loss:** We applied an inverse square-root frequency weighting to amplify the contribution of minority classes while preserving training stability.
- **Class-balanced focal loss:** This loss function reweights gradients using the "effective number" of samples and adds a modulating factor to emphasize difficult examples. This loss tended to place excessive emphasis on rare classes, and this failure helped guide us toward weighted cross-entropy as a more balanced alternative. These comparisons can be found in Table 2.

Sampling strategy

To further address class imbalance, we incorporated a `WeightedRandomSampler`, ensuring that each batch contained a more uniform distribution of severity levels.

Transition from ViT to Wav2Vec

We replaced MFCC inputs with raw-audio embeddings extracted using Wav2Vec [2]. Pretrained representations have been shown to encode articulatory information, which is highly relevant for the detection of dysarthria.

Parameter-efficient fine-tuning with LoRA

To fit Wav2Vec to our task without overfitting the small dataset, we applied Low-Rank Adaptation (LoRA) to preserve the pretrained backbone while introducing a small number of trainable parameters.

5.2.1 Sequence post-processing modules

Because Wav2Vec produces variable-length embeddings, we implemented several pooling strategies:

- last-dimension pooling
- mean and max pooling
- temporal convolution pooling
- multi-head self-attention pooling
- learned scalar attention pooling

These modules were designed to aggregate temporal information within each recording and control how dysarthria-relevant frames are weighted. Among these, the learned scalar attention pooling mechanism proved to be the most effective, as shown in Table 3.

5.3 Baseline reproduction evidence

The systematic exploration of hyperparameters and their effects on F1-score, as shown in Table 4, demonstrates that our ViT baseline was faithfully reproduced and validated.

ID	Learning Rate	Dropout	LR Scheduler	Patience / Restarts	Augment (T/F Mask)	Batch Size	F1 Score
1	1.00E-04	0	ROP	3, mode = max (f1 score), fact=0.5	✗	3	0.397
2	1.00E-05	0	ROP	3, mode = max (f1 score), fact=0.5	✗	16	0.394
3	1.00E-05	0	ROP	3, mode = max (f1 score), fact=0.5	✗	64	0.369
4	1.00E-05	0.3	ROP	3, mode = max (f1 score), fact=0.5	✗	16	0.401
5	1.00E-05	0.3	ROP	8, mode = max (f1 score), fact=0.5	✗	16	0.387
6	1.00E-06	0.5	ROP	5, mode = max (f1 score), fact=0.5	✓	16	0.336
7	1.00E-05	0.5	ROP	8, mode = max (f1 score), fact=0.5	✓	16	0.354
8	1.00E-05	0.5	CA	tmax = num_epochs	✓	16	0.385
9	1.00E-05	0.5	CAWR	t0 = 5, min = 1e-7	✓	16	0.394
10	1.00E-05	0.5	CAWR	t0 = 10, min = 1e-7	✓	16	0.409
11	1.00E-05	0.5	CAWR	t0 = 5, t_mult = 2, min = 1e-7	✓	16	0.368
12	1.00E-06	0.5	CAWR	t0 = 5, min = 1e-7	✓	16	0.339
13	5.00E-06	0.5	CAWR	t0 = 5, min = 1e-7	✓	16	0.411
14	5.00E-06	0.5	CAWR	t0 = 5, min = 1e-7	✓	64	0.338
15	1.00E-05	0.5	ROP	3, mode = max (f1 score), fact=0.5	✗	64	0.304
16	1.00E-05	0.5	ROP	3, mode = max (f1 score), fact=0.5	✗	16	0.317
17	1.00E-05	0.3	CAWR	t0 = 5, min = 1e-7	✗	16	0.400
18	5.00E-06	0.3	CAWR	t0 = 5, min = 1e-7	✗	16	0.420
19	5.00E-06	0.3	ROP	3, mode = max (f1 score), fact=0.5	✗	16	0.437
20	5.00E-06	0.3	ROP	3, mode = max (f1 score), fact=0.5	✗	16	0.432

Table 4: Ablation study of Vision Transformer (ViT) training configurations. All experiments use a ViT model with random weight initialization, trained for 50 epochs using the AdamW optimizer. The table summarizes the effects of learning rate, dropout, learning rate scheduler (ROP = ReduceLROn-Plateau, CA = Cosine Annealing, CAWR = Cosine Annealing with Warm Restarts), patience/restart settings, batch size, and the use of time-frequency masking (if ✓, $t = 20$, $f = 8$). Ablation **19** achieves the highest validation F1-score, highlighting the benefit of its configuration for stable and effective optimization.

6 Results and analysis

6.1 Results

From Table 3, it is evident that using an attention-based temporal pooling mechanism over Wav2Vec embeddings significantly improves performance. When combined with the weighted cross-entropy loss, as shown in Table 2, these design choices yielded our best-performing ALS detection model. The final architecture comprising Wav2Vec, LoRA fine-tuning, attention pooling, weighted cross-entropy loss, and weighted sampling achieved a Macro F1 score of 0.824, representing an improvement of approximately 1.9x over the baseline of 0.437.

To evaluate robustness, we repeated training with different random seeds over 50 epochs. Table 5 reports the Macro F1 scores across runs, demonstrating consistent performance with a mean of 0.776 and standard deviation of 0.039.

Table 5: Robustness of the final architecture across multiple seeds.

Seed	Macro F1
99	0.706
77	0.763
101	0.764
55	0.765
66	0.800
88	0.808
111	0.824
Mean	0.776
Std	0.039

The per-class performance for the best model (seed = 111) is presented in Table 6 and Table 7. Classes 0 and 1 are perfectly classified, while classes 3 and 4 show higher misclassification rates, reflecting class imbalance challenges.

Table 6: Per-class performance of the best model (seed = 111).

Class	Precision	Sensitivity	F1-score	Support
0	1.000	1.000	1.000	1
1	1.000	1.000	1.000	2
2	0.875	0.875	0.875	8
3	0.500	0.750	0.600	12
4	0.818	0.529	0.643	17

Table 7: False negative rates per class for the best model (seed = 111).

Class	False Negative Rate
0	0.0
1	0.0
2	0.125
3	0.25
4	0.471

6.2 Error failure case analysis

To assess the robustness of the proposed model, we examined the class-wise error patterns using the confusion matrix shown in Figure 5. The matrix reveals that while the model achieves strong diagonal dominance indicating high accuracy across most classes, several systematic failure modes emerge.

First, the largest source of error arises from confusions between adjacent or acoustically similar classes. These off-diagonal concentrations suggest that the model struggles when the underlying speech characteristics share overlapping feature distributions. This behavior is consistent with our feature extraction pipeline, where subtle variations may be attenuated during temporal pooling, causing ambiguity in the final embeddings.

Second, certain classes exhibit asymmetric misclassification patterns. For example, class 4 may be frequently misidentified as class 3, while the inverse occurs rarely. This asymmetry indicates that some classes have broader intra-class variability, making them harder to model, or that the learned representation captures dominant features that more strongly resemble neighboring categories.

Additionally, a small set of errors occurs in low-support classes, which aligns with the data imbalance present in the training set as shown in Figure 6. These cases highlight a limitation where the classifier tends to overfit classes with more abundant samples and underrepresent minority classes during decision-making.

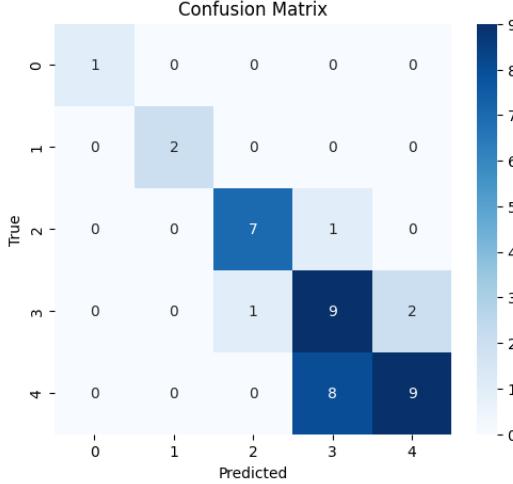


Figure 5: Confusion matrix illustrating class-wise prediction performance and error patterns.

6.3 Sensitivity/Ablation analysis

After selecting the weighted cross-entropy loss, we performed a sensitivity analysis to evaluate the impact of different feature extractors and temporal pooling strategies on model performance using Wav2Vec-LoRA embeddings. Table 3 summarizes the results.

From the table, it is clear that the choice of feature extractor and pooling method has a substantial effect on both validation loss and Macro F1 score. Notably, the attention-based temporal pooling mechanism achieved the best performance, with a validation loss of 0.943 and a Macro F1 of 0.824, outperforming both convolutional and multi-head attention alternatives.

The superior performance of the attention-based pooling can be attributed to its ability to dynamically weigh temporal information across the entire sequence of embeddings, allowing the model to focus on the most informative parts of the speech signal. In contrast, convolutional pooling applies fixed local filters, which may fail to capture long-range dependencies, and multi-head attention, while theoretically expressive, did not yield the same improvements, possibly due to overparameterization or insufficient training data for effective head specialization. These findings indicate that learnable attention pooling effectively leverages the rich representations of Wav2Vec embeddings, resulting in a substantial performance gain over simpler or more rigid pooling strategies. Thus, this demonstrates that careful selection of temporal aggregation mechanisms is critical, and that attention-based pooling provides a robust method for extracting informative features from pretrained speech embeddings.

7 Discussion

Our experiments show that replacing spectral features and transformer-based pooling with a pre-trained Wav2Vec encoder, paired with an attention-based aggregation mechanism, significantly improves dysarthria severity classification performance. The improvement from our ViT baseline (Macro-F1 of 0.437) to our best Wav2Vec-LoRA model (Macro-F1 of 0.824) demonstrates the value of self-supervised speech representations for this type of clinical task.

Even with these gains there are several limitations that remain. The model still struggles with the mild and no-dysarthria classes, where acoustic differences are extremely subtle. This leads to confusion between adjacent severity levels, which is reflected in the inherent difficulty of distinguishing these classes.

Additionally, our limited dataset size made our evaluation metrics volatile and limited in their granularity. Although we do see consistent good performance in the rare classes across runs, further validation of the model in these classes would be necessary before any clinical application were to be considered.

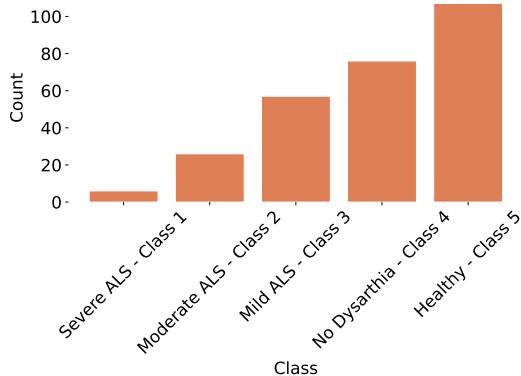


Figure 6: Class distribution in the dataset. This chart highlights the severe class imbalance, with certain classes significantly outweighing others in terms of frequency.

8 Future directions

There are several promising directions for extending this work. First, expanding the dataset by performing high-quality dysarthria specific data-augmentations could help address our scarce-data bottleneck. Augmentations that mimic dysarthria-related symptoms, like slowed articulation, irregular phonation, and increased pausing, may help the model learn early-stage markers more effectively.

Second, since dysarthria severity is inherently ordered, future models could incorporate ordinal classification objectives, or regression-style loss functions. Approaches such as ordinal cross-entropy may reduce confusion between adjacent severity levels and make model predictions more aligned with clinical scales.

Another future direction is to explicitly model the structure of the multi-task recordings. An architecture that processes each speech task separately and aggregates across tasks could better capture content-dependent cues, and provide insight into what tasks are best for dysarthria severity classification.

Finally, interpretability techniques such as attention visualization could be used to identify which acoustic regions most influence model predictions. This could increase clinical trust as well as help validate whether the model focuses on meaningful speech characteristics.

9 Conclusion

The objective of this project was to develop a model capable of predicting ALS dysarthria severity from short audio recordings. Our initial ViT baseline using MFCC features showed limited performance, motivating a shift toward more expressive speech representations.

By adopting a Wav2Vec encoder with LoRA fine-tuning and an attention-based pooling mechanism, we substantially improved classification performance. Our best model achieved a Macro-F1 of 0.824, nearly doubling our baseline performance. Our ablation studies demonstrated that attention pooling and weighted cross-entropy were important contributors to performance, whereas alternatives like class-balanced focal loss or mean-pooling were less effective in our architecture. Weighted sampling further helped stabilize our model’s performance for rare classes.

Our findings highlight the potential of pretrained speech models for clinical audio analysis, as well as emphasize the importance of addressing data imbalance and subtle class distinctions in the detection of neurodegenerative disorders.

10 Administrative details

10.1 Team contributions

Our work was highly collaborative, with each team member taking on essential responsibilities while maintaining continuous coordination. Hajar led the documentation and proposed the idea of incorporating a pretrained Wav2Vec model, designing the overall architecture. Ibrahim focused on implementing data augmentation techniques, integrating Wav2Vec, and developing the MLP classifier. Nikitha was responsible for implementing the feature extractor and conducting experiments to identify the most effective configuration. Dominic handled running experiments, selecting appropriate loss functions, performing ablations, and analyzing hyperparameter performance. Each contribution was critical to the projects progress, reflecting a truly balanced and collaborative effort.

10.2 GitHub repository

All project code, preprocessing pipelines, model implementations, and documentation will be maintained in a GitHub repository to ensure version control and reproducibility. The repository is available at:

[https://github.com/dimadome7/
Cognitive-Echo-Detecting-Neurodegenerative-Disorders-from-Speech](https://github.com/dimadome7/Cognitive-Echo-Detecting-Neurodegenerative-Disorders-from-Speech)

References

- [1] Speech analysis for neurodegenerative diseases (sand) challenge. <https://www.sand.icar.cnr.it>, 2025. Accessed: September 26, 2025.
- [2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.
- [3] Raffaele Dubbioso, Myriam Spisto, Laura Verde, Valentina Virginia Iuzzolino, Gianmaria Senerchia, Elena Salvatore, Giuseppe De Pietro, Ivano De Falco, and Giovanna Sannino. Voice signals database of als patients with different dysarthria severity and healthy controls. *Scientific Data*, 11(1):800, 2024.
- [4] Yifan Gao, Long Guo, and Hong Liu. Leveraging multimodal methods and spontaneous speech for alzheimers disease identification. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE, 2025.
- [5] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.
- [6] Damian Kwasny and Daria Hemmerling. Gender and age estimation methods based on speech using deep neural networks. *Sensors*, 21(14):4785, 2021.
- [7] Saturnino Luz, Fasih Haider, Sofia De la Fuente, Davida Fromm, and Brian MacWhinney. Detecting cognitive decline using speech only: The adreso challenge. In *Proc. of Interspeech 2021*, pages 3780–3784, 2021.
- [8] Cathy J. Price. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, 62(2):816–847, 2012.
- [9] Shakeel A. Sheikh, Md. Sahidullah, and Ina Kodrasi. Overview of automatic speech analysis and technologies for neurodegenerative disorders: Diagnosis and assistive applications. *IEEE Journal of Selected Topics in Signal Processing*, , 2025.
- [10] Fuxiang Tao, Bahman Mirheidari, Madhurananda Pahar, Sophie Young, Yao Xiao, Hend Elghazaly, Fritz Peters, Caitlin Illingworth, Dorota Braun, Ronan OMalley, et al. Early dementia

- detection using multiple spontaneous speech prompts: The process challenge. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE, 2025.
- [11] Jun Wang, Prasanna V. Kothalkar, Beiming Cao, and Daragh Heitzman. Towards automatic detection of amyotrophic lateral sclerosis from speech acoustic and articulatory samples. In *Interspeech 2016*, pages 1195–1199, 2016.