

SAND CHALLENGE: FOUR APPROACHES FOR DYSARTHRIA SEVERITY CLASSIFICATION

Gauri Deshpande, Harish Battula, Ashish Panda, Sunil Kumar Kopparapu

TCS Research, Tata Consultancy Services Limited, India

ABSTRACT

This paper presents a unified study of four distinct modeling approaches for classifying dysarthria severity in the Speech Analysis for Neurodegenerative Diseases (SAND) challenge. All models tackle the same five-class classification task (ALS patients rated 1–4 in severity, and 5 for healthy controls) using a common dataset of ALS patient speech recordings. We investigate: (1) a ViT-AVE method leveraging a Vision Transformer on spectrogram images with an averaged-loss training strategy, (2) a 1D-CNN approach using eight 1-D convolutional neural networks (CNNs) with majority-vote fusion, (3) a BiLSTM-OF approach using nine BiLSTM models with majority-vote fusion, and (4) a Hierarchical XGBoost ensemble that combines glottal and formant features through a two-stage learning framework. Each method is described, and their performances on a validation set of 53 speakers are compared. Results show that while the feature-engineered XGBoost ensemble achieves the highest macro-F1 (0.86), the deep learning models (ViT, CNN, BiLSTM) attain competitive F1-scores (0.64–0.70) and offer complementary insights into the problem. We discuss the trade-offs of each approach and highlight how they can be seen as complementary strategies addressing different facets of dysarthria classification. In conclusion, combining domain-specific features with data-driven methods appears promising for robust dysarthria severity prediction.

Index Terms— BiLSTM, CNN, Glottal Features, Phase Features, Late Fusion, Hierarchical Modeling, ViT

1. INTRODUCTION

Dysarthria is a motor speech disorder common in neurodegenerative diseases such as Amyotrophic Lateral Sclerosis (ALS). The Speech Analysis for Neuro-degenerative Diseases (SAND) challenge [1] at ICASSP 2026 focuses on automatic classification of dysarthria severity into five levels. Task #1 of this challenge asks participants to predict the severity class (“ALSFRS-R” score category) for each speaker’s voice, given a fixed set of short utterances (spoken vowels and syllables). Class labels range from 1 (most severe dysarthric speech) to 4 (milder dysarthria in ALS patients), and 5 for healthy control speakers.

All approaches in this study use the same dataset provided in SAND challenge. The dataset contains recorded utterances from 219 ALS patients for training and 53 speakers for validation. Each speaker provides 8 specific utterances: five sustained phonations (vowels A, E, I, O, U) and three repetitive rhythmic syllables (KA, PA, TA). These utterances capture different aspects of speech production (vowel phonation versus articulatory rhythm). The classification task is challenging due to severe class imbalance – for example, only 4 speakers are labeled Class 1 (most severe) while 86 are Class 5 (healthy) in the training set. This imbalance necessitates strategies like data augmentation and weighted loss to avoid biasing toward the majority class. Additionally, there is a gender imbalance, with male-to-female ratios of 1.28 in training and 1.30 in validation sets (see Table 1).

In this paper, we consolidate four complementary models developed for SAND Task #1, integrating our findings into a single cohesive report. Despite differing methodologies, all four aim to maximize classification accuracy on the same task and dataset. By unifying their perspectives, we provide a comprehensive view of how diverse techniques, ranging from deep learning on raw spectrograms to machine learning on engineered features, can contribute to the dysarthria severity classification problem. The following sections describe each approach’s methodology, followed by a comparison of their performance and a discussion on their complementary strengths.

Table 1: Gender distribution across classes, SAND Task #1 dataset.

| Class | Training Baseline | | | Validation Baseline | | |
|--------------------------|-------------------|-----|--------------------------|---------------------|----|----|
| | F | M | Total | Class | F | M |
| 1 | 3 | 1 | 4 | 1 | 1 | 1 |
| 2 | 12 | 10 | 22 | 2 | 3 | 1 |
| 3 | 16 | 29 | 45 | 3 | 5 | 7 |
| 4 | 24 | 38 | 62 | 4 | 4 | 10 |
| 5 | 41 | 45 | 86 | 5 | 10 | 11 |
| Total | 96 | 123 | 219 | Total | 23 | 30 |
| Male/Female ratio = 1.28 | | | Male/Female ratio = 1.30 | | | |

2. METHODOLOGY

We developed four different models to address the five-class dysarthria classification. Each approach leverages a unique modeling technique and fusion strategy for the multiple utterances per speaker. In the following, we detail each approach: ViT-AVE, Hierarchical XGBoost, 1D-CNN, and BiLSTM-OF.

2.1. ViT Model with Averaged Loss (ViT-AVE)

The ViT-AVE approach uses a vision transformer model to classify dysarthria severity from spectrogram images. We started with a pre-trained Vision Transformer (ViT-B16) model (originally developed for image recognition) and fine-tuned it on the speech spectrogram data. Each audio utterance (vowel or syllable) was converted to a 2D spectrogram image (see Fig. 1) using `librosa`; images were standardized to 512×256 pixels representing the time-frequency content. Basic data augmentation was applied to these spectrogram images during training (random horizontal flips, rotations, and color jitter for brightness/contrast/saturation) to increase variability.

A key novelty of ViT-AVE is how it handles multiple utterances per speaker. During training, we computed the loss for each of a speaker’s 8 (5 phonations, 3 rhythm) spectrograms and then averaged the loss across all utterances of that speaker before back-propagation. This averaged loss approach aligns with the assumption that a speaker’s overall severity label should be reflected consistently across all their utterances. It effectively treats the 8 utterances as a set, stabilizing training by not over-weighting any single utterance. In the inference (validation) phase, we feed all 8 spectrograms of a speaker through the ViT model and then average the predicted class probabilities of all utterances to make the final decision. This probability averaging (late fusion) is analogous to an ensemble vote, but weighted by confidence, and was found to improve robustness.

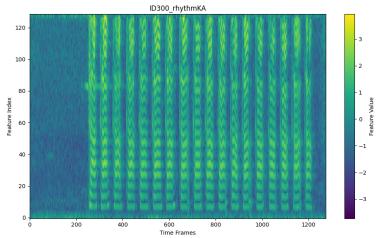


Fig. 1: Spectrogram of ID300 for the rhythm KA.

2.1.1. Architecture

The ViT-AVE model is a PyTorch pretrained model ViT-B16 [2] transformer (12 layer, 768 dim hidden) with its final classification head replaced by a new fully connected layer tuned

for 5 classes. By leveraging the ViT’s strength in image analysis, the model can capture fine-grained spectral patterns in the voice recordings.

2.1.2. Result

On the validation set (53 speakers), the ViT-AVE achieved a macro-averaged F1 score of 0.68 (68.09%). Notably, this approach demonstrated that even with relatively few training samples, a pre-trained transformer can be re-purposed for audio classification when combined with appropriate augmentation and a fusion strategy. We observed that data scarcity and imbalance caused a noisy validation curve (high variance in F1 across epochs), indicating the model was somewhat sensitive to training fluctuations. Nonetheless, ViT-AVE provided a solid baseline, outperforming a trivial majority-class predictor by leveraging spectral image-based features and multi-utterance averaging.

2.2. Hierarchical XGBoost with Glottal and Formant Features

In contrast to end-to-end neural approaches, we explored a two-stage hierarchical (see Figure 2) model using gradient boosted trees (XGBoost) enriched with expert-crafted speech features. This approach explicitly incorporates domain knowledge about dysarthric speech by extracting glottal pulse, extracted using the SEDREAMS algorithm [3], and formant frequency features, alongside patient demographics (age group and gender), as inputs to the classifier.

2.2.1. Feature Extraction

From each audio recording, we computed 12 acoustic features that have known relevance in characterizing speech impairment: 5 vowel formant frequencies (F1–F5), and 7 glottal pulse parameters (see Fig. 3). The glottal features were derived using the SEDREAMS algorithm to detect glottal closure instants (GCIs) – essentially markers of vocal fold vibration – from which measures such as pitch period statistics and amplitude are obtained. These features capture voice quality and articulation characteristics that may correlate with dysarthria severity (e.g., unstable pitch or reduced articulation clarity). In addition, two demographic features (speaker’s age group and gender) were included, since dysarthria manifestations can vary with age and differ between male and female speakers.

2.2.2. Architecture

The classification is performed in two stages. In Stage 1, we train eight binary XGBoost classifiers (see Table 2), each specializing in a particular decision sub-problem or subgroup of data. These were designed based on observed confusions and the metadata: for example, Model 1 distinguishes Class 3 vs

Table 2: Configuration of eight XGBoost models for hierarchical approach.

| Model # | Group (M-F) | Classification | SC | N_{est} | F_l (ms) | Audio _d |
|---------|---------------|----------------|----|-----------|------------|--------------------|
| 1 | F | 3 vs 4 & 5 | A | 200 | 100 | Full |
| 2 | F | 4 vs 5 | KA | 100 | 100 | Initial 20 s |
| 3 | < 60- M | 3 vs 4 & 5 | U | 100 | 50 | Later 10 s |
| 4 | < 60- M | 4 vs 5 | O | 100 | 500 | Full |
| 5 | ≥ 60 - M | 3 vs 4 & 5 | E | 200 | 100 | Initial 20 s |
| 6 | ≥ 60 - M | 4 vs 5 | I | 100 | 100 | Initial 20 s |
| 7 | M & F | 1 vs 2 | I | 100 | 50 | Full |
| 8 | M & F | 1 vs 2 | U | 100 | 50 | Full |

Class 4 & 5 for female speakers, Model 2 separates Class 4 vs Class 5 for female speakers, Models 3–6 handle similar binary splits for male speakers stratified by age (mid-age (< 60 vs old-age ≥ 60), and Models 7 and 8 focus on differentiating lower severity classes (Class 1 vs Class 2) across all speakers. Each XGBoost model uses the 12 features as input, and all are trained with a small learning rate (0.01) to ensure fine-grained learning. The binary outputs from these 8 models (essentially predictions or confidence scores for each sub-task) are then collected. In Stage 2, we feed the collection of eight Stage-1 predictions plus the encoded age (normalized to [0, 1]) and gender (encoded as 0.9 for female and 0.1 for male) into a higher-level classifier. Stage 2 is implemented as a simple decision tree (depth=5, 100 trees) that learns to map the pattern of binary outcomes to the final five-class decision. Figure 2 illustrates this hierarchy: the first layer of models focuses on specific binary distinctions, and the second layer integrates them for the overall classification.

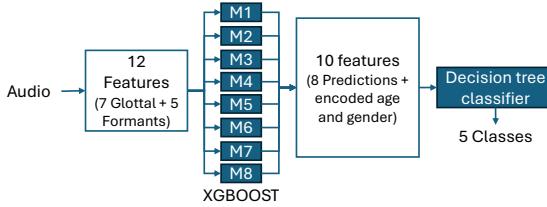


Fig. 2: Block diagram of hierarchical XGBoost approach.

This hierarchical ensemble effectively performs a form of late fusion and decision fusion: it uses specialized classifiers to capture subtle differences (for example, differences in how females vs males present Class 4 vs 5), then combines their outputs. By incorporating demographics and dividing the task, it tackles the heterogeneous nature of the data (gender and age effects on speech) in a structured way.

2.2.3. Result

The hierarchical XGBoost approach yielded the highest performance among our methods. On the validation set of 53 speakers, it achieved a macro F1 score of 0.8644 (86.44%), with an overall accuracy of 0.88741. The confusion ma-

trix indicated only slight confusion between adjacent classes (e.g., Class 3 vs 5, Class 4 vs 5). This high accuracy demonstrates the effectiveness of combining expert features with a tailored classification strategy. The model’s success suggests that in this limited-data scenario, incorporating prior knowledge (via features like formants and glottal pulses) can significantly boost performance. It’s worth noting that certain utterance types were excluded from this approach (the unvoiced PA and TA were found less useful for glottal feature extraction), indicating the approach made informed decisions about which data aspects to leverage. Overall, this method provides a strong benchmark, showing that a carefully designed feature-based model can excel in dysarthria classification.

Remark 1 Separate models are used for male and female speakers because pitch and glottal characteristics vary significantly across gender and age groups. Experiments showed that phonation categories A, E, I, O, U and rhythm KA were most effective, while voiceless rhythms (PA, TA) were excluded due to poor glottal pulse extraction.

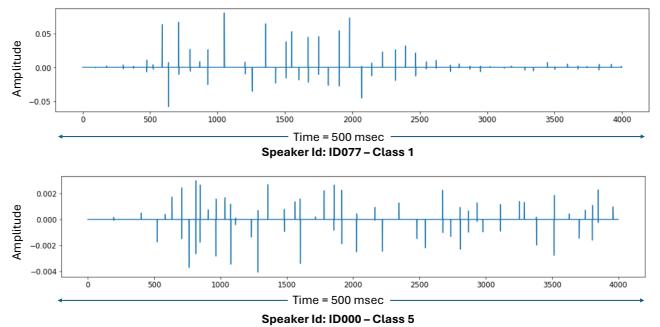


Fig. 3: Glottal parameters extracted from speech signals for speakers 'ID077' (Class 1) and 'ID000' (Class 5).

2.3. 1D-CNN : 1D-CNN with Output Fusion

The 1D-CNN approach (1-Dimensional Convolutional Neural Network with Output Fusion) addresses the classification

by training separate CNN models for each type of utterance and then fusing their outputs via majority voting (see Fig. 4). The rationale is that each utterance (vowel or rhythm) might carry complementary information about the speaker’s speech deficit, so specialized models can learn features particular to each phonation, and a fusion decision then aggregates these insights.

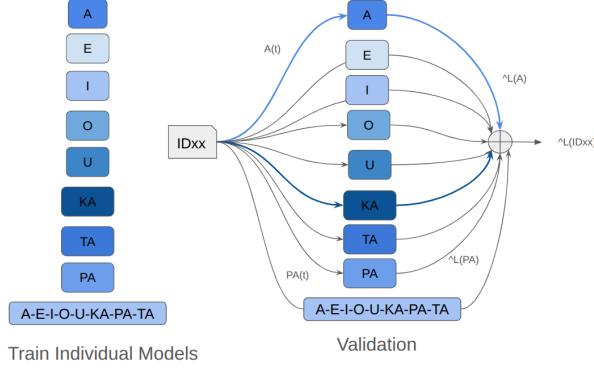


Fig. 4: Architecture for late fusion model.

2.3.1. Architecture & Features

We built eight parallel 1-D CNN models, one for each of the 5 vowel phonations (A, E, I, O, U) and 3 rhythmic syllables (KA, PA, TA). Each CNN takes as input the raw audio of its designated utterance type, transformed into a sequence of acoustic feature frames. Unlike ViT-AVE which used spectrogram images, 1D-CNN operates on frame-based features.

We extracted a rich set of phase-based acoustic features per frame: specifically, 54-dimensional vectors comprising Phase Cepstral Coefficients (13), Group Delay Cepstral Coefficients (13), Modified Group Delay (13), Instantaneous Frequency (13), and other phase statistics like coherence and entropy. These features collectively capture fine details of the speech signal’s phase and frequency content (e.g., the shape of the spectrum, source-filter characteristics, and frequency modulation cues), which are informative for characterizing slurred or irregular speech in dysarthria. Each audio utterance was windowed (20 ms frames with 10 ms hop, 8 kHz sampling) and truncated or zero-padded to a fixed length of $T = 500$ frames (≈ 5.01 s) so that all inputs had equal length. The CNN architecture for each task is relatively small, namely,

1. 3 Conv1d layers with BatchNorm, ReLU, Dropout
2. Global Average Pooling (across T frames)
3. Layer Normalization
4. Fully Connected Layers: $256 \rightarrow 128 \rightarrow 5$

The fully connected layer maps to the 5 output classes. This yields a probability distribution over the 5 severity levels for

that single utterance. Each of the 8 models has about 0.31 million parameters. We trained each CNN with a cross-entropy loss, using class-weighting inversely proportional to class frequency to handle the imbalance (so errors on rare classes counted more) and used the AdamW optimizer (learning rate 0.001).

2.3.2. Fusion Strategy

At inference time, a given speaker provides 8 utterances (A, E, I, O, U, KA, PA, TA). Each utterance is fed to its corresponding CNN model, producing 8 independent predicted labels. The final classification for the speaker is then obtained by majority voting across these 8 predicted labels. In other words, the class that appears most frequently among the eight CNN outputs is chosen as the overall diagnosis for that speaker. This late fusion by majority vote leverages the *wisdom of the ensemble*. This fusion strategy reduces the impact of any single utterance or model mis-prediction, as long as the majority are correct. This method assumes that each utterance contributes equally to the final decision and treats all output votes uniformly.

2.3.3. Result

In initial experiments using all 8 utterance models, 1D-CNN achieved a validation accuracy of 0.6226 and a macro F1 of 0.5656. Analysis of feature importance using SHAP [4] indicated that certain feature types and certain utterances were more informative. By selecting the top 20 most important features and focusing on a subset of the best-performing utterance models (specifically using only the models for E, O, KA, PA, TA) while dropping weaker ones, the performance improved significantly. The refined 1D-CNN system recorded a macro F1 score of 0.6398 (63.98%) on the validation set, with a weighted F1 of 0.63. This improvement highlights that not all utterances contributed equally – the chosen five utterances had provided the most discriminatory power – and that feature selection helped remove noisy or redundant inputs. Even though 1D-CNN’s accuracy ($\approx 64\%$) was lower than the transformer and XGBoost models, it demonstrated a viable approach using a straightforward architecture. The ensemble of simple CNNs was able to capture complementary information from different sounds, and the majority vote fusion yielded a stable combined decision. The approach underlined the importance of task-specific modeling (separating vowels and syllables) and feature engineering (phase features) for this problem, bridging a gap between end-to-end learning and expert features: the models learned internal representations, but on carefully chosen input features for each phonation.

2.4. BiLSTM-OF : BiLSTM Ensemble with Output Fusion

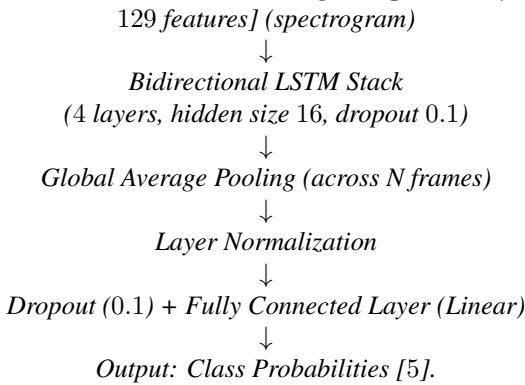
The BiLSTM-OF approach extends the idea of utterance-wise modeling to sequence models, using Bidirectional LSTM (BiLSTM) networks for each utterance type and fusing their outputs with majority voting (see Fig. 4). Recurrent neural networks like LSTMs can capture temporal dynamics in speech, which might be beneficial for the rhythmic syllable utterances or any temporal patterns in phonation not captured by static features.

2.4.1. Architecture & Preprocessing

We prepared the input for BiLSTMs similarly to 1D-CNN , but focusing on time-domain spectrogram features. All utterances were first trimmed to remove leading and trailing silences to ensure the model focuses only on the active speech portions. We then converted each utterance into a spectrogram using Short-Time Fourier Transform (STFT) with 20 ms window, 10 ms hop, and 256-point FFT (see Fig. 1). The resulting magnitude spectrogram can be viewed as a sequence of frames (time steps) with 129 frequency-bin features each (given the 256 FFT, we have 129 unique frequencies). Unlike VIT-AVE 's 2D image approach, here the spectrogram is fed as a time-series matrix to the LSTM. We constructed nine BiLSTM models: one for each of the same 8 utterances (A, E, I, O, U, KA, PA, TA) as in 1D-CNN , plus one extra model for a combined *phonation-rhythm* (C_+ ; concatenation of A-E-I-O-U-KA-TA-PA) input.

Remark 2 This combined input, denoted C_+ , is created by concatenating all 8 utterances of a speaker in sequence, forming a single longer waveform representing that speaker's full set of sounds. The C_+ model is intended to capture cross-utterance cues and overall speech characteristics when all content is heard together, complementing the individual utterance models.

Each BiLSTM model is a lightweight recurrent network with an architecture as follows: *Input Sequence [N frames, 129 features] (spectrogram)*



This yields about 31.6k trainable parameters per model. The design ensures the model can capture temporal patterns

within each utterance (via the BiLSTM's recurrent connections) while eventually producing an utterance-level class probability. We used cross-entropy loss with class weighting (inversely proportional to class frequency, as in 1D-CNN) to address the imbalance. Training used the Adam optimizer ($\text{lr}=0.001$) and early stopping (patience 20 epochs) to prevent overfitting. Because the dataset is small, we used a batch size of 1 to maximize the training samples usage per epoch. Figure 5 shows a sample train and the validation plot for MODEL(C_+) and MODEL(TA). As can be observed the F1-score is 0.47(0.43) on the validation set achieved after 14(17) epochs respectively for MODEL(C_+) and MODEL(TA).

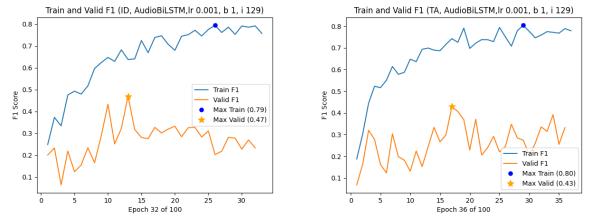


Fig. 5: Sample train and validation F1-score versus epoch for C_+ and rhythm TA.

2.4.2. Fusion Strategy

During inference, a speaker's 8 trimmed utterances are run through the corresponding 9 BiLSTM models (for A, E, I, O, U, KA, PA, TA, C_+) to produce 9 predicted labels. The final classification is determined by majority vote among these 9 outputs like in case of 1D-CNN (with one extra vote coming from the C_+ model).

2.4.3. Result

Initially, using all 9 BiLSTM models, the system achieved about 62.26% macro F1 (very similar to the baseline CNN ensemble). We observed that certain models (particularly those for the sustained vowels) were less accurate, whereas the models for the rhythmic syllables (KA, PA, TA) and the combined utterance C_+ were more reliable. Therefore, in a refined experiment, we pruned the ensemble down to 4 models – using only the three rhythm utterance models (KA, PA, TA) and the C_+ model, which were the strongest contributors. With the majority voting among just these four outputs, the performance improved markedly. The best BiLSTM-OF configuration achieved a macro F1 score of 0.7042 (70.42%) on the validation set, with a weighted F1 of 0.6450 and accuracy of ≈ 0.642 . This was a substantial gain over using all utterances indiscriminately, indicating that for LSTM models the rhythmic utterances carried more consistent signals of severity (possibly because these fast repetitive syllables exaggerate speech impairments like slurring or timing irregularities). The result of 70% F1 also surpassed the 1D-CNN

's 64%, suggesting that capturing temporal dynamics via BiLSTM (and including the combined utterance) provided an edge. However, BiLSTM-OF did not reach the performance of the feature-based XGBoost model, underlining that purely data-driven approaches faced challenges with the limited data. Nonetheless, BiLSTM-OF demonstrates an effective strategy: by tailoring models to different utterance types and wisely combining them, one can boost classification performance.

2.5. Results and Comparison

All four approaches were evaluated under the same conditions on the SAND validation set (53 speakers, true labels unknown to models during training).

Table 3 shows the confusion matrices for the four approaches while Table 4 provides a comparison of the key characteristics and results of all the four approaches. We compare the model architecture, input features used, the fusion strategy to handle multiple utterances, and the achieved macro F1 score (the primary challenge metric) on the validation data.

2.5.1. Observations

The feature-based Hierarchical XGBoost approach outperformed in terms of raw macro-F1 (0.86 vs 0.68–0.70 for the best neural models) which emphasizes the value of carefully chosen features and tailored sub-tasks in this challenge. By using formant and glottal features grounded in speech science, it appears to capture the relevant cues of dysarthria more directly, and the hierarchical structure effectively handled different speaker demographics and class confusions. On the other hand, the ViT-AVE and BiLSTM-OF approaches demonstrate that purely data-driven, end-to-end learning can also reach reasonable performance (0.68 and 0.70 F1 respectively) even with limited data, by exploiting augmentation, pre-training, or ensemble techniques. ViT-AVE's use of a pretrained transformer allows it to generalize from a small dataset by transferring knowledge from images, and its averaging strategy smooths out per-utterance variability. BiLSTM-OF and 1D-CNN both emphasize the benefit of treating each utterance type separately.

It is interesting to note that both the fusion strategies employed, namely, averaging probabilities and majority voting have merit. ViT-AVE's probability averaging is a soft fusion, potentially leveraging the confidence of predictions, whereas 1D-CNN and BiLSTM-OF's majority vote is a hard decision fusion that can be more robust when individual model confidence is not well-calibrated. In our experiments, the majority vote (with selected strong models) in BiLSTM-OF gave a higher F1 than ViT-AVE's averaging, but this may also be due to differences in feature learning capability (spectrogram image vs. sequential modeling).

All models had to confront the class imbalance issue. The deep learning models did so through data augmentation and

weighted losses, while the XGBoost model implicitly handled it by splitting tasks (ensuring minor classes got focused binary classifiers). As a result, each approach was able to identify Class 1 speakers (most scarce) to some extent.

Remark 3 *Automatic Speech Recognition techniques were unsuccessful on this dataset as the transcription (ASR output) provided are insufficient. For example, the rhythm sounds PA, TA and KA are repeated but there is no output transcription with information about how many times these sounds were repeated. If these rhythm sounds were annotated to indicate how many times each speaker uttered these, then ASR techniques could have been used to better advantage. Also, information about how the manual grading of the speech signals were done could have provided better background for designing systems. For example, we have assumed in all our experiments that the classification was arrived at by averaging over all the utterances. However exact methodology of manual classification would have provided a better background for the design of the loss function.*

3. CONCLUSION

We explored four complementary approaches for Task #1 of the SAND Challenge, which involves five-class classification of dysarthria severity using speech recordings. All the four approaches (ViT-AVE, 1D-CNN, BiLSTM-OF, and Hierarchical XGBoost) offer a unique perspective on the same classification task, from end-to-end neural modeling to hierarchical feature-driven classification. Comparative analysis shows that incorporating expert knowledge via features (formants, glottal pulses) can yield superior accuracy (macro F1 0.86), but deep learning methods with the right design (pre-training, data augmentation, ensemble fusion) can also achieve good performance (F1 0.70). The approaches are largely complementary: for instance, the transformer and LSTM models automatically learn abstract representations of dysarthric speech, while the XGBoost approach confirms the relevance of specific known biomarkers of dysarthria. All models face common challenges of limited and imbalanced data, which they address through augmentation, specialized modeling per utterance, and task decomposition.

Through this analysis, one can appreciate how different strategies effect the goal of classifying ALS speech severity. This collective insight suggests that future work could explore hybrid models that blend these approaches, such as feeding engineered features into neural networks or using neural outputs as features in boosting, to further improve robustness. Moreover, cross-validation with larger datasets or fusion of the complementary models could push performance beyond what each achieved.

In summary, the SAND Task 1 challenge has been approached from multiple angles by us: from spectrogram-based deep learning to interpretable feature-based classi-

Table 3: Confusion matrices for different approaches: (a) ViT-AVE , (b) 1D-CNN , (c) BiLSTM-OF , and (d) XGBoost.

| (a) ViT-AVE | | | | | | (b) 1D-CNN | | | | | (c) BiLSTM-OF | | | | | (d) XGBoost | | | | | | | |
|-------------|---|---|---|---|---|------------|---|---|---|---|---------------|-----------|----|---|---|-------------|---|-----------|----|---|---|---|----|
| \hat{L} | 1 | 2 | 3 | 4 | 5 | \hat{L} | 1 | 2 | 3 | 4 | 5 | \hat{L} | 1 | 2 | 3 | 4 | 5 | \hat{L} | 1 | 2 | 3 | 4 | 5 |
| L | 1 | 2 | - | - | - | L | 1 | 1 | - | - | - | L | 1 | 2 | - | - | - | L | 1 | 1 | - | - | 1 |
| | 2 | - | 2 | 2 | - | | 2 | 1 | 3 | - | - | | 2 | - | 2 | - | 2 | | 2 | - | 3 | - | 1 |
| | 3 | - | - | 8 | 1 | 2 | | 3 | - | 1 | 7 | - | 4 | 3 | - | - | 7 | 3 | 2 | 3 | - | 8 | 3 |
| | 4 | - | - | 4 | 6 | 4 | | 4 | - | 3 | 1 | 5 | 5 | 4 | - | - | 1 | 9 | 4 | 4 | - | - | 12 |
| | 5 | - | - | 5 | 2 | 14 | | 5 | - | 2 | - | 2 | 17 | 5 | - | - | 3 | 4 | 14 | 5 | - | - | 3 |

| Approach | Architecture | Input Features | Fusion Strategy | Macro-F1 |
|-----------|----------------------------|-----------------------------------------------|------------------------------------------------------|-------------|
| ViT-AVE | ViT-B16 Transformer | Spectrogram images (512x256) | Average probability across 8 utterances | 0.68 |
| 1D-CNN | 1-D CNN (8-model ensemble) | Phase-based spectral features (54 dims/frame) | Majority vote across 8 model outputs | 0.64 |
| BiLSTM-OF | BiLSTM (9-model ensemble) | STFT magnitude spectrogram (129 dims/frame) | Majority vote across 9 model outputs | 0.70 |
| XGBoost | 8 XGBoost + Decision Tree | 5 formants + 7 glottal features + age/gender | Hierarchical ensemble (binary models → 5-class tree) | 0.86 |

Table 4: Comparison of four approaches for dysarthria severity classification in the SAND challenge.

fiers. The highest performing model (hierarchical XGBoost) demonstrates the effectiveness of domain-inspired features, while the deep learning models highlight the promise of end-to-end learning even in data-scarce settings. This consolidated study serves as a comprehensive reference for researchers interested in dysarthria classification, illustrating that there is no one-size-fits-all solution – instead, a combination of diverse techniques may be the key to robust and generalizable performance in this field.

4. REFERENCES

- [1] “SAND Challenge – Speech Analysis for Neurodegenerative Diseases: Task One,” <https://www.sand.icar.cnr.it/#task-one>, 2026, Accessed: 2025-10-01.
- [2] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai, “An image is worth 16x16 words: Transformers for image recognition at scale,” 2021.
- [3] Thomas Drugman and Thierry Dutoit, “Glottal closure and opening instant detection from speech signals,” in *Proc. Interspeech 2009*, 2009, pp. 2891–2894.
- [4] Scott M Lundberg and Su-In Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.