

An Explainable Machine Learning Framework for Dysarthria Detection Using Deep Acoustic Embeddings and Prosodic Cues

Md Tanjil Islam Aronno¹, Md. Sabbir Hossain Molla², Md Shovon Rahman³, Md Tanvir Rahman Sahed⁴

^{1,3}Department of Electrical and Electronic Engineering, Bangladesh University of Engineering & Technology
Dhaka, Bangladesh

^{2,4}Department of Electrical, Electronic & Communication Engineering, Military Institute of Science & Technology
Dhaka, Bangladesh

¹aronnondc@gmail.com, ²sabbir.eece@gmail.com, ³shovonrahman600@gmail.com, ⁴tanvir.sahed00@gmail.com

Abstract—This paper introduces a machine learning framework for binary dysarthria classification using acoustic and prosodic cues collected from speech recordings of dysarthric and non-dysarthric patients. Deep CNN14 embeddings are used to learn the audio signals’ spectral and temporal patterns and to improve the classification accuracy, other prosodic descriptors such as pitch variability, energy dynamics, and spectral flatness are added. We train and combine several machine learning classifiers, such as KNN, Random Forest, XGBoost and CatBoost, and a soft-voting ensemble model. Adding prosody to the experiment made the overall accuracy go up from 97% to 98.1%, with big drops in both false positives and false negatives. SHAP-based explainable AI analysis shows that pitch-related and energy based parameters are two of the most important factors in detecting dysarthria. The proposed method presents a balanced framework that integrates deep acoustic representations with comprehensible features to attain reliable and clear dysarthric speech assessment.

Index Terms—dysarthria, UASpeech, prosody, ensemble, machine learning.

I. INTRODUCTION

Dysarthria and other speech problems can make it very hard for a person to communicate. These problems are often caused by neurological disorders like Parkinson’s disease, cerebral palsy, or amyotrophic lateral sclerosis (ALS). Dysarthric speech usually has problems with articulation, prosody, phonation, and resonance. Accurately and swiftly identifying these issues can aid in early diagnosis, improved treatment planning, and the development of speech based interfaces that are user friendly for those who have these issues.

In recent years, the use of machine learning (ML) to identify dysarthria has increased. Compared to conventional clinical testing, which is frequently subjective, time consuming, and require expert assistance, this is a possible substitute. By examining speech patterns and acoustic characteristics, automated systems provide an impartial means of assessing and monitoring individuals in both clinical and home environments. Further they offer an objective way to evaluate and keep an eye on people in both clinical and home settings by looking at speech patterns and acoustic features. However, creating strong detection systems is still hard because speakers are different, there are limited annotated datasets available, and

models need to be able to predict well while also being easy to understand.

Previous research investigated handcrafted acoustic features and statistical classifiers, while modern studies have utilized deep learning architectures trained on spectrograms or raw audio. Pretrained audio encoders have made things even better in situations where there isn’t much to work with. Even with these improvements, many current methods only use deep embeddings or handcrafted features. They don’t look at whether combining them could make the model more reliable and clear.

This paper presents a dysarthria detection framework that combines learned audio embeddings from the CNN14 model with spectral and prosodic features. A soft voting ensemble combines the outputs of multiple base classifiers, such as tree-based, linear, and distance-based models, that are trained on these features. We use SHAP (SHapley Additive Explanations) to measure how much each feature contributes to the model in order to make it easier to understand. The UASpeech dataset is used to test the system because it gives a controlled setting for finding speech problems in a wide range of speakers and severity levels.

The goal of this work is to create a detection pipeline that not only improves classification performance but also gives clear information about the acoustic features that contributes mostly to detect dysarthric speech. The proposed framework can tell the difference between dysarthric speech and healthy speech as seen from the higher overall accuracy and macro-averaged evaluation matrices. In situations where automatic speech recognition (ASR) is used, dysarthric speakers may need a stronger and different set up of ASR module. In that cases, our framework can work as one of the good classifier of normal and dysarthric speaker.

II. LITERATURE REVIEW

The use of machine learning and deep learning for dysarthria detection has progressed over the past decade, driven by advances in data augmentation, transfer learning, and model interpretability. Despite these developments, exist-

ing approaches involve trade-offs related to generalizability, computational efficiency, and clinical applicability.

Balaji et al. presented a deep learning model aimed at improving dysarthric speech comprehensibility for therapeutic communication using speech reconstruction techniques [1]. Although effective, the model depends on high-quality data from specific disorders, which limits its transferability. Badge et al. examined the use of AI-based speech analysis for early stroke detection and identified speech symptoms such as sudden-onset dysarthria [2]. While the results were promising, the study was constrained by a narrow dataset and the absence of large-scale clinical validation.

Tan et al. proposed CBA-Whisper, a fine-tuned Whisper model employing curriculum learning and AdaLoRA for low-resource dysarthric speech recognition [3]. The method reduced training instability but required high memory resources and relied on Whisper’s architecture, limiting deployment on edge devices. Prabhala et al. introduced an ensemble deep learning framework using a stacking strategy to detect dysarthric speech [4]. Although the model achieved high classification performance, it lacked explainability, which is important for medical decision support.

Jones et al. conducted a pilot study using convolutional neural networks (CNNs) to detect early Parkinson’s disease through dysarthric speech patterns [5]. The findings indicated diagnostic potential, but the small sample size and lack of longitudinal data limited clinical applicability. Hassan et al. proposed a hybrid WaveNet and CNN-BiLSTM architecture for dysarthria detection in patients with ALS and cerebral palsy [6]. While the model outperformed individual networks, it required substantial computational resources and was sensitive to background noise. Prem et al. investigated lightweight CNN models for dysarthria classification, enabling real-time performance with limited accuracy loss [7]. However, the shallow architectures struggled with more severe speech degradation.

Midharshini et al. developed a machine learning model to classify dysarthria severity levels and reported encouraging results across impairment grades [8]. Despite strong accuracy, the approach did not provide feature-level explainability. Mahum et al. proposed a Swin Transformer-based framework that outperformed CNN baselines across several metrics [9]. The model required extensive preprocessing and was not evaluated under real-time or multilingual conditions. Sasikala et al. explored electroglottography signals combined with machine learning for identifying voice pathologies, including dysarthria [10]. Although physiological signals improved diagnosis, the requirement for specialized equipment limited practical use.

Yao et al. provided an analysis of artificial intelligence applications in neurological speech disorder identification, focusing on challenges in dysarthria modeling [11]. While informative, the study did not include implementation or comparative evaluation. Favaro et al. proposed a multimodal system combining speech, facial cues, and lexical features for detecting neurological disorders such as dysarthria [12]. The approach achieved good accuracy but was limited by

data synchronization challenges and the lack of open-access datasets.

Recent self-supervised speech encoders such as wav2vec 2.0 and HuBERT [13] have been shown to learn effective representations that improve a wide range of downstream speech tasks beyond automatic speech recognition, including speaker diarization and separation.

Overall, many existing systems prioritize performance over interpretability or rely on limited datasets, which affects generalization. To address these limitations, we propose a hybrid ensemble framework that combines CNN14-based audio embeddings with spectral and prosodic features. A soft-voting ensemble and SHAP analysis are used to identify features contributing to dysarthric speech detection. The framework is evaluated on the UASpeech dataset, providing a balance between accuracy, generalization, and explainability.

III. METHODOLOGY

The methodology section incorporates the description of the overall process, from the dataset collection to the model evaluation. The steps involved in the overall classification framework is illustrated in Fig 1.

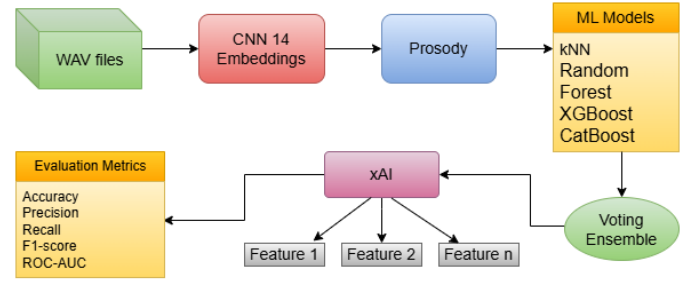


Fig. 1: Overall classification framework

A. DATASET DESCRIPTION

A popular benchmark dataset for studies on dysarthric speech recognition and intelligibility evaluation is the UA-Speech [14] corpus. It includes recordings of 13 control speakers who speak normally and 15 people who have dysarthria mainly due to cerebral palsy. A fixed vocabulary of roughly 765 isolated words, including letters, numbers, computer commands, and everyday words, is uttered by each participant, producing between 2,800 and 5,500 audio files per speaker. The recordings were made using high-quality setups at 44.1 kHz, 16-bit PCM resolution, in a quiet, microphone-equipped setting. Severity-aware modelling and assessment are made easier by the corpus’s metadata on speaker gender, age, and dysarthria severity level (very low, low, moderate, or high).

B. PREPROCESSING

In order to provide a consistent time–frequency footprint for CNN14 and prosody extraction, all waveforms are peak-normalized and forced to a fixed 6.4 s duration at 16 kHz via centre pad. Before calculating frame-level contours, we use

the (*librosa.effects.trim(top_db=20)*) function from Python’s Librosa library to apply silence trimming for prosodic analysis. Train-only fitting of the *SimpleImputer(strategy="mean")* and *StandardScaler()* functions is used to process the resultant tabular feature matrix (CNN14 embeddings $\pm$ prosody).

We use *GroupShuffleSplit(test_size=0.2, random_state=42)* to implement a speaker-partitioned split in order to stop speaker leakage. To make sure no speaker shows up in both the training and test sets, speaker identifiers are parsed from file paths (such as M01, F05) and passed as groups. In addition to providing a realistic generalization setting for clinical deployment, this design maintains subject independence. 80% of the data were used for training and the remaining 20% was used as test data. All scaling steps are fit on the training speakers only and then applied to rest of the speakers.

C. FEATURE EXTRACTION & FUSION

1) *CNN14 Embeddings*: We have used PANN’s (Pre-trained AudioSet CNN) CNN14, which is a strong, task-agnostic acoustic embedding. These embeddings capture time-frequency patterns such as modulation, timbre, and onsets that transfer well to downstream speech tasks, particularly when labeled data is scarce. We get 2048-D acoustic embeddings by setting up the model for 16 kHz audio with $f_{min}=50$ Hz and $f_{max}=8$ kHz, and then loading it with pretrained weights. Each normalized waveform of 6.4 seconds is sent once, and the embedding vector from the second-to-last layer is kept (emb_0 ... emb_2047).

2) *Prosody Features*: Prosody shows how stable phonation is, how loud it is, and how well the spectrum is balanced. These are all things that usually change in dysarthric speech. In our research, we integrate prosodic features with CNN14 embeddings and note enhanced classification performance. Using a 1024-sample analysis window, we get all of the prosodic and spectral-shape descriptors at 16 kHz. Using the *librosa.feature* functions, we find the RMS energy, zero-crossing rate, spectral centroid, spectral bandwidth, spectral rolloff (95%), and spectral flatness. Using *librosa.yin(fmin=50, fmax=500)*, we estimate pitch (F0). Unvoiced frames are given a value of NaN, and a voiced-ratio metric (the fraction of frames with finite F0) is used to summarize the results. We calculate seven robust statistics for each temporal contour: mean, standard deviation, median, 25th percentile, 75th percentile, maximum, and minimum. This gives us a compact 50-dimensional feature vector. Furthermore, prosodic features employ certain clinical intuitions that are depicted in Table I.

3) *Fusion and Leakage Handling*: Feature fusion is the process of combining the 2048-D CNN14 vector and the 50-D prosody vector early on to make a 2098-D representation for each clip. The same imputer and scaler that were used to train the model are used on the fused matrix. Data leakage is managed through (i) speaker-grouped data partitioning and (ii) limiting all preprocessing and hyperparameters to the training split before evaluation.

TABLE I: Prosodic and Spectral-Shape Features with Clinical Intuition

Feature	Clinical Intuition
Pitch (F0)	Intonation and phonation stability; low variance indicates monotone speech.
Voiced Ratio	Voicing consistency; low values suggest breathiness or dropout.
RMS Energy	Loudness control; reduced range indicates weak respiratory effort.
ZCR	Signal noisiness; high values imply imprecise articulation.
Spectral Centroid	Spectral brightness; low centroid reflects muffled speech.
Spectral Bandwidth	Energy spread; wider bands denote articulation instability.
Spectral Rolloff (95%)	High-frequency energy proportion; higher rolloff shows frication.
Spectral Flatness	Tonality vs. noise; high values mark rough or breathy phonation.

D. IMPLEMENTATION MACHINE LEARNING MODELS & ENSEMBLE

1) *Machine Learning Models*: We used four classifiers that worked well together and were trained on the combined on both feature set, CNN14 only and fusion of CNN14 with prosody. The training process adhered to the previously outlined speaker-held-out protocol, which involved fitting pre-processing exclusively on the training split and evaluating on the held-out test split. Below are short descriptions of the models:

- **K-Nearest Neighbours (KNN)**: A classifier that doesn’t use parameters and gives a label based on the majority of the k nearest training samples. It captures the local shape of the feature space and gives a simple starting point for combining acoustic and prosodic representations.
- **Random Forest (RF)**: A group of decision trees that were trained on bootstrap copies and feature subsampling. RF lowers variance by bagging and gives well-calibrated probabilities by averaging over different trees. This is helpful when there is a moderate class imbalance.
- **Extreme Gradient Boosting (XGBoost)**: A normalized gradient-boosted tree ensemble that fits residuals stage by stage using a logistic loss. We use tree-level sparsity, column/row subsampling, and depth control to find a good balance between bias and variance and make it easier to generalize on high-dimensional fused features.
- **Categorical Boosting (CatBoost)**: A gradient-boosted tree model with strong regularisation defaults and ordered boosting. Even though our features are numbers, CatBoost’s ordered boosting and class-weighting make it work well even when there is an imbalance, and it still trains quickly on GPUs.

2) *Ensembling the Models*: We used a “soft-voting” scheme (equal weights) to combine the base learners to make them more robust and able to generalize. Let $\hat{p}_m(y=1 | \mathbf{x})$ be the predicted probability from model $m \in \{1, \dots, M\}$. The

ensemble probability is

$$\hat{p}_{\text{ens}}(y=1 | \mathbf{x}) = \frac{1}{M} \sum_{m=1}^M \hat{p}_m(y=1 | \mathbf{x}), \quad (1)$$

and the final decision is obtained by thresholding

$$\hat{y} = \begin{cases} 1, & \text{if } \hat{p}_{\text{ens}}(y=1 | \mathbf{x}) \geq \tau, \tau=0.5 \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

In our implementation, the ensemble is a VotingClassifier that uses KNN, RandomForest, XGBoost, and CatBoost. Each model is trained on both feature space and adds the same amount to $\hat{p}_{\text{ens}}$.

E. Feature Importance Analysis Using SHAP

SHAP is used to do model attribution on the fitted tree models and their ensemble. We use *shap.TreeExplainer* for XGBoost and Random Forest and the built-in *get_feature_importance()* for CatBoost. When we combine SHAP across tree models, we work in the raw space to make sure that the score domain stays the same. Features are clearly labeled: CNN14 dimensions are labeled as *emb_0* through *emb_2047*, and prosodic descriptors are labeled with their statistical names, such as *f0_std*, *rms* enabling both global feature importance analysis and the examination of their effects in specific prediction scenarios. In practice, prosodic measures associated with F0 variability, RMS dynamics, spectral flatness, and voiced ratio serve as interpretable determinants of the dysarthria decision, whereas selected (*emb_digits*) coordinates encapsulate supplementary latent acoustic patterns.

F. EVALUATION matrices

Evaluation employs a single speaker-held-out protocol utilizing *GroupShuffleSplit*, with 80% of the data allocated for training and 20% for testing. We check how well the models work by looking at Accuracy, Precision, Recall, F1-macro, and ROC-AUC.

Accuracy: The accuracy of a sample is the number of samples that were correctly classified:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}. \quad (3)$$

Precision: Precision is the percentage of predicted positives that are actually positives:

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (4)$$

Recall: Recall (sensitivity) is the percentage of real positives that are correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (5)$$

F1-Score: The F1-score is the harmonic mean of Precision (*P*) and Recall (*R*):

$$F1 = \frac{2PR}{P+R}, \text{ where } P = \text{Precision}, R = \text{Recall}. \quad (6)$$

ROC-AUC: The Receiver Operating Characteristic Area Under the Curve (ROC-AUC) shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for all thresholds. It is the same as the chance that a randomly chosen positive gets a higher score than a randomly chosen negative.

IV. RESULT ANALYSIS

The overall evaluation of the ML models over the UASPEECH dataset includes the comparison with and without embeddings along with the inclusion of ROC-AUC curves, confusion matrices. Finally, the most contributing features to the detection of a dysarthric case have been conducted with the XAI analysis.

A. PERFORMANCE EVALUATION OF THE PROPOSED MODELS

We only used CNN14 to test how well the ML classifiers worked, and then we added prosody to see how well they worked together. Adding prosodic descriptors has consistently made every model work better. Table II shows that XGBoost and the ensemble classifier had the highest overall accuracy and macro-F1 of about 0.98. The Random Forest, on the other hand, had the lowest accuracy. The average increase in F1 after adding prosody features was about 1.1% (from 0.97 to 0.981), which shows that prosodic cues add to the information that CNN14 latent representations already have. 9

TABLE II: Performance with and without prosody (macro-averaged matrices).

Features	Models	Acc.	Pre.	Rec.	F1-Score
CNN14	KNN	0.97	0.97	0.97	0.97
	Random Forest	0.93	0.93	0.93	0.93
	XGBoost	0.97	0.97	0.97	0.97
	CatBoost	0.96	0.96	0.96	0.96
	Ensemble	0.97	0.97	0.97	0.97
CNN14+Prosody	KNN	0.98	0.98	0.98	0.98
	Random Forest	0.94	0.94	0.94	0.94
	XGBoost	0.98	0.98	0.98	0.98
	CatBoost	0.97	0.97	0.97	0.97
	Ensemble	0.98	0.98	0.98	0.98

Acc.=Accuracy, Pre.=Precision, Rec.=Recall, F1-Score=F1-Score

1) *Confusion Matrix Analysis:* As can be seen from the confusion matrices in Fig 2 and Fig 3, using prosody with CNN14 decreases both error types: FP drops from 329 to 200 and FN drops from 470 to 318. Prosody, as opposed to a simple threshold shift, indicates a sharper decision boundary by improving dysarthric speech detection while avoiding over-flagging controls.

2) *Model Robustness and Discussion:* To avoid identity leakage between train and test data, a speaker-partitioned *GroupShuffleSplit* was used for both training and evaluation of the models. On the held-out set, all classifiers obtained nearly identical accuracy and macro-F1 values, indicating little overfitting. Generalization to unseen speakers is further supported by the robust and steady ROC-AUC displayed in Fig 4.

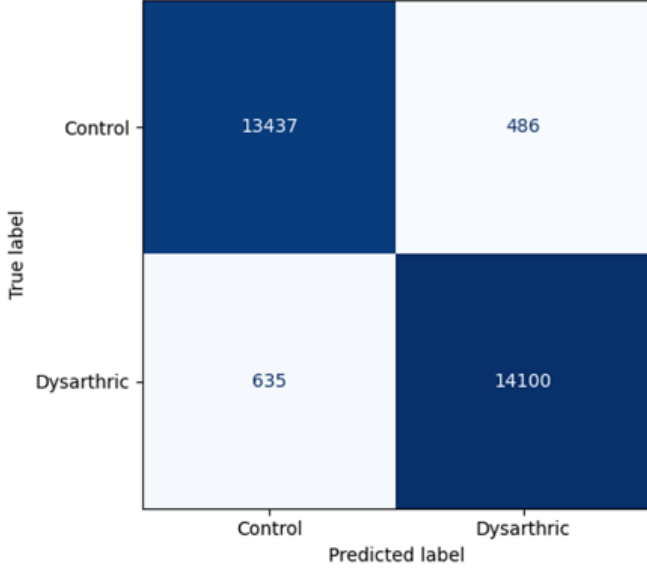


Fig. 2: Global SHAP feature-importance ranking for the CNN14+prosody model

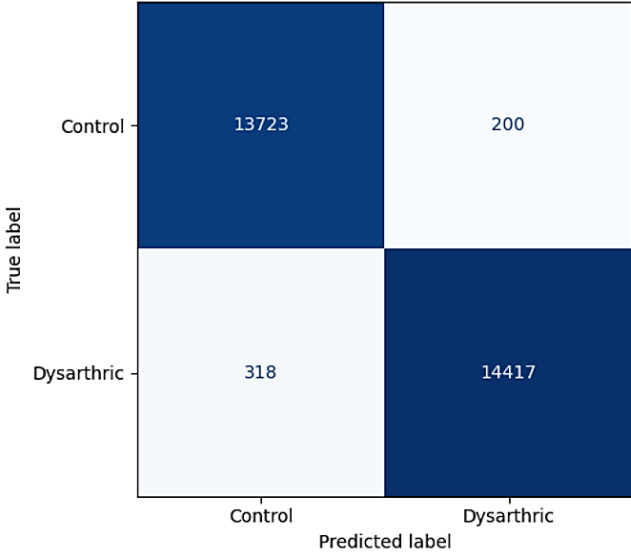


Fig. 3: Confusion matrices of the ensemble classifier with CNN14 only

B. Explainable AI (XAI) Analysis

While assessing the best ensemble classifier, XAI analysis was used to interpret the ways in which individual features influenced the classification of dysarthria. Key CNN14 latent embeddings are shown alongside prosodic descriptors like pitch variability ($f0_std$), energy dynamics (rms_std), spectral flatness, spectral centroid, and voiced ratio. Reduced F0 variations are often reported in spastic presentations, whereas larger F0 and intensity fluctuations are commonly associated with irregular prosody seen in ataxic-like patterns. Since UASpeech labels are binary (dysarthric vs. control), prosodic classifica-

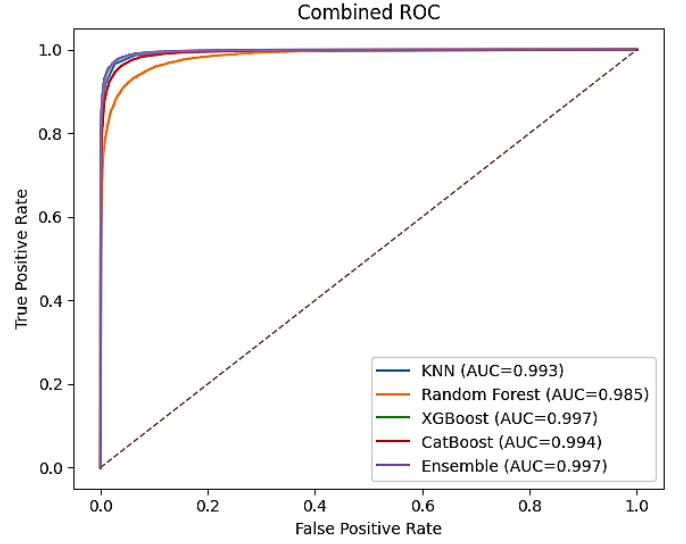


Fig. 4: Confusion matrices of the ensemble classifier with prosody features

tions are not feasible here. The feature-importance plot (Fig. 5) ranks variables according to their average contribution. This demonstrates that the fused model uses both deep acoustic representations and interpretable prosodic cues to differentiate dysarthric speech from control speech.

The impact of feature values on model predictions is visualised by the beeswarm plot of Fig. 6. With colour indicating feature magnitude (blue for low, red for high) and horizontal position indicating SHAP influence, each point represents a sample. While points on the left support the control class, those on the right raise the likelihood of dysarthria. While higher voiced ratios and more pitch variability are linked to control speech, high spectral flatness and large RMS variability push predictions towards dysarthria, reflecting unstable phonation. Additionally, speaker-specific variability is reflected in the spread of points, confirming that dysarthric cues appear differently in different people.

V. CONCLUSION

In summary, the fusion of CNN14 and prosodic features yields a significant and consistent improvement across all models, particularly enhancing recall and F1 scores. The SHAP-based interpretability analysis further indicates that the improvements originate from physiologically meaningful acoustic cues. Hence, the proposed approach demonstrates both high predictive performance and clinical plausibility, making it a reliable framework for automatic dysarthria detection. UASpeech recordings are collected under controlled conditions; performance may degrade under background noise or mismatched microphones.

In future, we intend to use openAI's Whisper or META'S Wav2Vec2.0 as feature extractors as they have become popular as well as used widely for feature extraction purposes. Furthermore, the development of deep learning based models

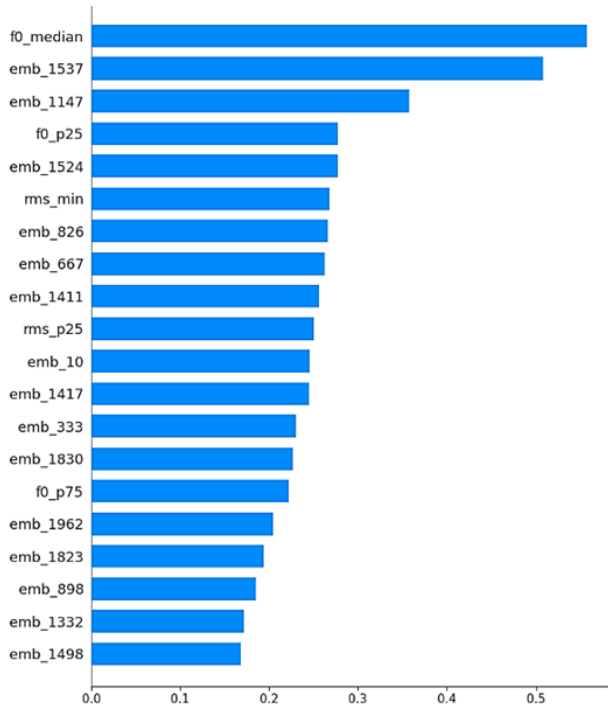


Fig. 5: Global SHAP feature-importance ranking for the CNN14+prosody model

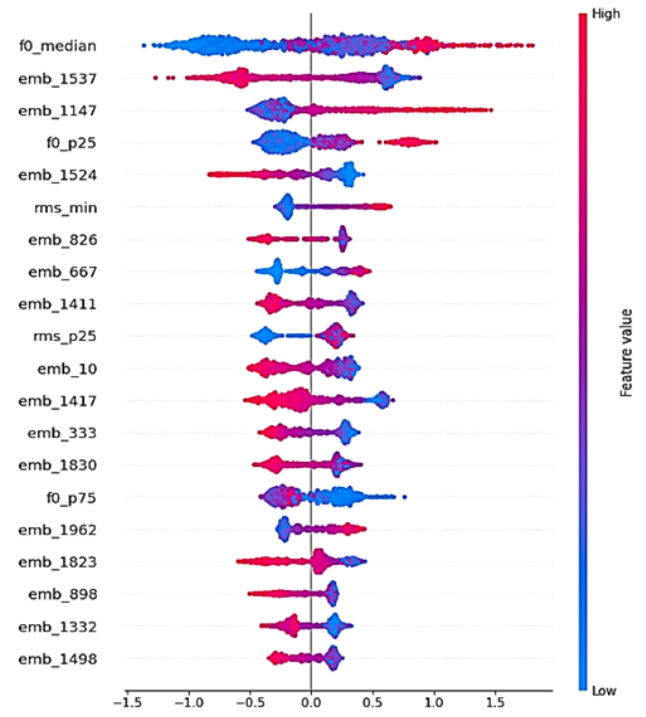


Fig. 6: SHAP beeswarm plot showing feature-level impact on dysarthria prediction

to classify properly in noisy environments will certainly work as a significant contribution to the research community.

REFERENCES

- [1] AP Yeshwanth Balaji, TR Eshwanth Karti, K Nithish Ariyha, J Vikash, and G Jyothish Lal. Enhancing dysarthric speech for improved clinical communication: A deep learning approach. In *Machine Learning and Deep Learning Modeling and Algorithms with Applications in Medical and Health Care*, pages 1–22. Springer, 2025.
- [2] Helen Badge, Pui Lok Annette Fung, Christopher Blair, Dennis Cordato, Megan Trebilcock, Suraj Narayanan Sasikumar, Rumbi Teramayi, Lauren Christie, Paul Middleton, and Mark Parsons. The feasibility of an automated artificial intelligence system to use voice and speech in detecting large vessel occlusion stroke. *Digital Health Week*, 4(1), 2025.
- [3] Tianyi Tan, Xinan Chen, Xiaohuai Le, Wenzhi Fan, Xianjun Xia, Chuanzeng Huang, and Jing Lu. Cba-whisper: Curriculum learning-based adalora fine-tuning on whisper for low-resource dysarthric speech recognition. In *Proc. Interspeech 2025*, pages 3309–3313, 2025.
- [4] Jagat Chaitanya Prabhala, Ravi Ragoju, Venkatanaresbhabu Kupilli, and Christophe Chesneau. Enhanced early detection of dysarthric speech disabilities using stacking ensemble deep learning model. *Machine Learning with Applications*, page 100721, 2025.
- [5] Sara A Jones, Jeremy Cosgrove, He Wang, and Ryan K Mathew. Automated detection of speech disorders in parkinson’s disease using deep convolutional neural networks: A pilot study. *medRxiv*, pages 2025–07, 2025.
- [6] Esraa Hassan, Abeer Saber, Tarek Abd El-Hafeez, T Medhat, and Mahmoud Y Shams. Enhanced dysarthria detection in cerebral palsy and als patients using wavenet and cnn-bilstm models: A comparative study with model interpretability. *Biomedical Signal Processing and Control*, 110:108128, 2025.
- [7] Daniel Prem, Kumudha Raimond, and Babita Prem. A comparative study of lightweight cnns for dysarthria classification. In *2025 5th International Conference on Pervasive Computing and Social Networking (ICPCSN)*, pages 172–177. IEEE, 2025.
- [8] B Midharshini, P Shanmugapriya, and V Mohan. Detection of dysarthria severity levels using ml models. In *2025 2nd International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)*, pages 1–15. IEEE, 2025.
- [9] Rabbia Mahum, Ismaila Ganiyu, Lotfi Hidri, Ahmed M El-Sherbeeney, and Haseeb Hassan. A novel swin transformer based framework for speech recognition for dysarthria. *Scientific Reports*, 15(1):20070, 2025.
- [10] S Sasikala, Arun Kumar, T Inbavel, S Darshini, et al. Voice pathology detection using machine learning and electroglottography signals. In *2025 3rd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, pages 1–4. IEEE, 2025.
- [11] Dongren Yao, Aki Koivu, and Kristina Simonyan. Applications of artificial intelligence in neurological voice disorders. *World Journal of Otorhinolaryngology-Head and Neck Surgery*, 2025.
- [12] Anna Favaro, Chelsie Motley, Tianyu Cao, Miguel Iglesias, Ankur Butala, Esther Oh, Robert Stevens, Jesús Villalba, Najim Dehak, and Laureano Moro-Velázquez. A multi-modal array of interpretable features to evaluate language and speech patterns in different neurological disorders. pages 532–539, 01 2023.
- [13] Séverin Baroudi, Hervé Bredin, Joseph Razik, and Ricard Marxer. On the use of self-supervised representation learning for speaker diarization and separation. *arXiv preprint arXiv:2512.15224*, 2025.
- [14] Heejin Kim, Mark Hasegawa Johnson, Jonathan Gunderson, Adrienne Perlman, Thomas Huang, Kenneth Watkin, Simone Frame, Harsh Vardhan Sharma, and Xi Zhou. Uaspeech. 2023.