Dyxplain: A SHAP-Guided LLM Framework for Interpretable Dysarthria Assessment

Anonymous ACL submission

Abstract

Effective dysarthria assessment requires accurate severity classification and interpretable feedback. While SHapley Additive exPlanations (SHAP) enhance model interpretability, their numerical nature limits accessibility for non-experts. To address this, we propose *Dyxplain*, a novel framework that improves accessibility and personalization in AI-driven speech disorder assessment by integrating acoustic feature analysis, Machine Learning (ML)-based dysarthria classification, SHAP-driven feature importance evaluation, and LLM-based explanation generation. Dyxplain processes paragraph reading speech data, identifies key acoustic features using SHAP, and generates intuitive explanations via an LLM. It achieves a speaker-level F1 score of 88.6%, improving severity classification performance by 11.7% over ML-only models, while also enhancing interpretability with a 37.5% increase in semantic similarity to clinical pathologist annotations compared to LLM-only methods.

1 Introduction

Dysarthria is a motor speech disorder caused by damage to the nervous system, leading to paralysis or weakness of the muscles involved in speech production. This condition affects various speech subsystems, including phonation, articulation, and prosody, resulting in severe communication impairments, emotional distress, and social isolation (Duffy, 2019). An automatic dysarthria diagnosis system that closely aligns with clinical judgments can provide valuable support to clinicians in diagnosis and treatment planning. Therefore, there is a need for automated and objective tools that can consistently and efficiently assess dysarthria, contributing to the accuracy and effectiveness of clinical diagnoses and treatments.

Recent studies have been aiming to automatically assess dysarthria while ensuring interpretability. Some studies have applied SHapley Additive exPlanations (SHAP) (Lundberg, 2017) to models (Zhang et al., 2023), while others have incorporated intermediate layers that learn perceptual dimensions to enhance interpretability (Xu et al., 2023). Other studies have also explored interpretability by analyzing the activation patterns of hidden neurons linked to phonetic features (Abderrazek et al., 2023; Tu et al., 2017). However, these approaches have limitations, as they do not directly provide clinically useful information, often leading to results that are non-intuitive in practical applications.

In the medical domain, approaches leveraging large language models (LLMs) have been proposed to provide clearer and more intuitive explanations. (Sharma et al., 2023) evaluated ChatGPT's performance on the USMLE, confirming the potential of LLM as tools for medical education and clinical decision support. Study explored the practical applicability of LLM in clinical settings. MedPaLM (Reddy et al., 2023) was trained on extensive medical datasets to generate reliable answers to clinical questions. Similarly, A study was conducted that examined the current state of LLM applications in the medical field-including diagnostic support, patient communication, and clinical note summarization (Peng et al., 2023). There was also a study on Medagent (Tang et al., 2023) that utilized an AI agent methodology, in which multiple agents were constructed to perform specific roles in complex information analysis, thereby enabling appropriate decision-making. However, most existing studies rely on pre-processed textual data or require large-scale datasets, posing challenges in terms of time and cost. In particular, in the medical field, automated analysis systems that utilize real-time patient data are still in the early stages of research, with most existing studies focusing on models trained on structured medical documents or records.

To overcome these limitations, we propose *Dyxplain*, a novel framework that automatically classifies and analyzes a patient's condition using audio

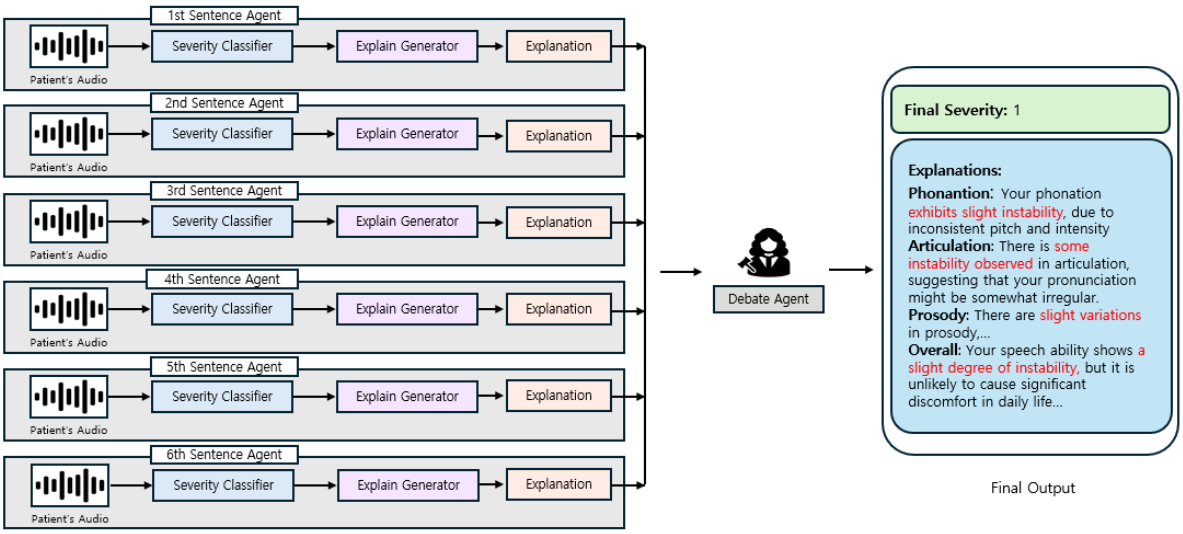

Figure 1: Overview of **Dyxplain**. The Dyxplain is composed of six Sentence Agents and one Debate Agent. Each Sentence Agent is provided with one of the six uttered sentences from the patient and is responsible for performing severity classification and analysis. The Debate Agent is subsequently given the outputs generated by the Sentence Agents and produces the patient's final severity label along with an integrated explanation.

data as input. The proposed system extracts acoustic features from the patient's speech, classifies dysarthria severity using machine learning (ML), and generates intuitive natural language explanations through an LLM. In the medical domain, interpreting data requires more than numerical analysis; it demands complex reasoning that incorporates clinical standards, patient metadata, and interactions between multiple features. To address this, we incorporate a Chain-of-Thought (CoT) (Wei et al., 2022) prompting strategy that enables the model to perform step-by-step reasoning, allowing for the generation of clinically meaningful and logically structured explanations. Additionally, a Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) mechanism is employed to retrieve similar patient cases, which are then integrated into the prompt to help the LLM produce more accurate and personalized outputs. This approach moves beyond reliance on static, pre-structured documents and establishes a real-time, interpretable medical analysis pipeline based on actual patient speech data. The main contributions of this work are as follows:

- To our knowledge, this is the first approach that integrates ML and LLMs using speech data collected from a paragraph reading task.
- A medical pipeline specialized for dysarthria using RAG (Lewis et al., 2020) and CoT (Wei et al., 2022) was established to improve patient severity analysis and classification performance.
- Effective analysis and severity classification were challenging for a dataset of long reading passages, our study demonstrates that adopting AI agent techniques significantly improves severity classification performance and enables the generation of interpretable explanations.

Table 1: Speech acoustic features extracted for paragraph reading task

Speech subsystem	Feature
Phonation	Pitch(std, min, max)
	Intensity (std, max)
	Jitter (DDP, PPQ)
	Shimmer (APQ)
	Harmonics-to-Noise Ratio
Articulation	F1
	F2
	Avg CER
Prosody	Speaking rate
	Articulation rate
	Avg syllable duration
	Phonation ratio
	Pause frequency ratio
	Pause duration ratio
	Pause ratio

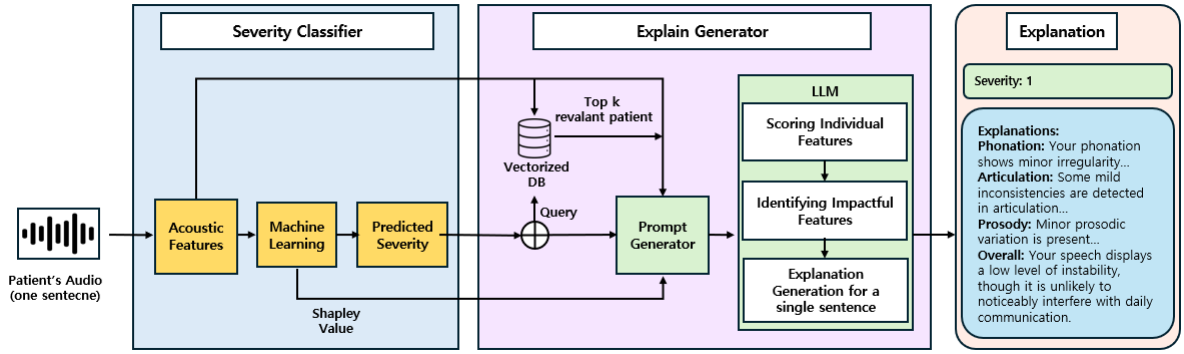

Figure 2: This figure illustrates the n-th Sentence Agent, which analyzes one of the six sentences spoken by the patient. In the **Severity Classifier**, acoustic features are extracted from the patient's audio and used to predict severity levels via a ML model. To explain the basis of the model's prediction, SHAP values are also computed. In the **Explain Generator**, the predicted severity and acoustic features are used to retrieve Top-k similar cases from a vectorized patient database. Based on the retrieved information, the Prompt Generator constructs a prompt, which is used as input to the LLM. Using the CoT, the model performs step-by-step analysis of the patient's condition and generates a natural language explanation.

2 Dataset

The "autumn" paragraph (Kim, 2005) reading task is a standardized Korean test designed to assess various aspects of speech, such as vocal quality, respiration rate, and pauses. This task involves reading a paragraph consisting of 369 syllables across 6 sentences, which includes all Korean consonants and vowels, reflecting their frequency of occurrence in the language. Audio data were recorded at 44.1kHz sampling rate, and downsampled to 16kHz for model input.

Data collection posed significant challenges, resulting in a relatively small dataset. In particular, individuals with severe dysarthria were less likely to participate, leading to a limited number of high-severity samples, a common issue in speech disorder research. The dataset consists of 326 patients (212M, 114F) and 60 healthy control (HC) speakers (34M, 26F) aged between 20 and 84 years. Each speaker read six sentences, with some participating in multiple recording sessions, totaling 2,364 sentences across all speakers. Clinical data, including gender and dysarthria severity, were also collected. Severity was categorized as normal (0), mild to moderate (1), or severe (2), with labels assigned by a neurosurgeon using National Institute of Health Stroke Scale (NIHSS) (Kwah and Diong, 2014) criteria. The HCs are classified as severity 0, while the patients are categorized as severity 1 or 2. Recruitment of healthy controls and patients was conducted in collaboration with [anonymized for review]. The study protocol, encompassing recordings and clinical assessments, was reviewed and approved by the participating institutions' IRBs, and all individuals provided informed consent.

For model training and evaluation, the dataset was split into training, validation, and test sets. The distribution of speakers was as follows: 260 in training (0:40, 1:196, 2:24), 59 in validation (0:11, 1:45, 2:3), and 67 in the test set (0:9, 1:49, 2:9), ensuring a balanced representation of severity levels. Each speaker was assigned to a single subset to prevent data leakage between training and evaluation.

3 Method

Dyxplain is composed of six Sentence Agents, each responsible for performing classification and analysis on an individual sentence, and a single Debate Agent, which serves as the final decision maker by aggregating the results. Each Sentence Agent is provided with one of the six sentences read by the patient and performs severity classification and explanation generation independently. The Debate Agent is subsequently given the sentence-level predictions and explanations, which are integrated to produce the final severity assessment at the patient level. The overall pipeline is illustrated in Fig. 1. Moreover, the operational flow for each sentence agent is shown in Fig. 2.

3.1 Sentence Agent

3.1.1 Severity Classifier

Guided by the Mayo Clinic classification system (Darley et al., 1969) for dysarthria, we selected a set

Table 2: Context and Prompt Information

Context	Prompt
Explanation for Each features	"Vocalization-related characteristic values." - Pitch - pitch_max: The frequency that determines the pitch or tone of the voice. It is primarily used to measure the highness or lowness of a voice. ...
Scoring Instruction	Analyze the given information to assess the patient's speech condition. SHAP value indicates the impact of each acoustic feature on classifying the patient into a specific severity level-a higher value suggests a greater influence, while a lower value suggests less. Based on the predicted severity, write a description of the patient's condition in Korean using simple and easy-to-understand words so that the general public (i.e., patients) can comprehend it. The description should simulate a real-world scenario where a doctor explains the condition to a patient in person.
Sentence Explain Generation Instruction	**Input Data Description:** `acoustic_feature` contains features used to assess dysarthria severity, along with values derived from a deep learning model. `pred_severity` represents the predicted severity level of the patient's speech condition based on the machine learning model. **Analysis Process:** Evaluate speech features using `acoustic_feature`. Score each feature on a scale from 0 to 100 based on `pred_severity`
Debate Agent Instruction	The paragraph reading task consists of a total of six sentences, and `pred_severity` refers to the severity of the patient's condition predicted by each sentence agent. As the **debate agent**, please determine the final severity and provide an appropriate explanation. The severity is generally calculated based on the frequency of occurrence across features. Follow the tagging criteria below: 0: Normal to Mild 1: Mild to moderate 2: Severe
Relevant Patient Information	Severity: 1 {'data_info': [{'speaker': '024', 'info': {'severity': 0.0, 'gender': 0, 'acoustic_feature': [{'total_feature': [{'severity': 0.0, 'f1': 747.29, 'f2': 1689.414, 'pitch': 175.563, 'pitch_std': 68.112 ...}
Input Data	{'speaker': '049', 'data_info': {'gender': 0, 'paragraph_characteristic_info': {'total_characteristic': {'severity': 0, 'gender': 0, 'f1': 565.53, 'f2': 1493.16, 'pitch': 106.60, 'pitch_std': 24.21 ...}
Feautres Score	{"acoustic features ": {"Pitch": 20, "Steady Pitch": 25, ... "Articulation rate" : 80 }
Impact Features	{'pos': ('Pitch', 20), 'neg': ("Articulation rate": 80), ...}
Shap Value	"F1": -0.21, "F2": -0.10, "Pitch": 0.12, "Steady Pitch": -0.05, "Max Intensity: 0.03, ... "Articulation_rate": 0.67,
Scoring Output Template	Paragraph features : {"Pitch" : , "Steady Pitch" : , "Steady Intensity" : , "Max Intensity" : ...
Explanation Output Template	Severity: Phonation: Articulation: Prosody: Overall:

of representative acoustic features covering phonation, articulation, and prosody. These features were extracted from sentence-level segmented speech using the Praat software (Boersma and Weenink, 2009), as summarized in Table 1. Specifically, 13 features were related to phonation, 3 to articulation, and 7 to prosody.

The Average Character Error Rate (Avg CER) for each speaker was calculated using the Whisper model (Radford et al., 2023), which computes the CER for each sentence. Although all features were extracted from sentence-level segments, the CER was averaged across the six sentences to reduce variability at the utterance level. This averaging provides a more robust estimate of overall intelligibility and aligns with clinical assessment practices.

Additionally, each speaker's age and gender were incorporated to account for potential variation in feature distribution, resulting in a total of 25 features. To assess dysarthria severity, the LightGBM (LGBM) model (Ke et al., 2017) was trained using sentence-level feature vectors as input. During

inference, SHAP values were computed for each sentence to analyze the contribution of individual features to severity classification across classes 0, 1, and 2.

SHAP values quantify the contribution of each feature to the model's prediction by computing its marginal effect across all possible feature subsets. In this study, TreeSHAP (Lundberg et al., 2020) was employed to efficiently calculate SHAP values for the LGBM model.

3.1.2 Explain Generator

The Explain Generator is designed to retrieve relevant patient cases using RAG and to construct prompts based on the retrieved information. These prompts are subsequently processed by an LLM through CoT reasoning, whereby the influence of individual acoustic feature is assessed, sentence-level severity is classified, and personalized explanations are generated using SHAP-derived scores and rule-based identification of impactful features.

Generating Prompt using RAG

To analyze the influence of each acoustic feature, relevant information is retrieved from similar patients using RAG, and prompts are generated based on this information. Each sentence of a patient is represented as a document containing the corresponding acoustic features and severity label. Each document is embedded into a 3,072-dimensional vector using OpenAI's `text-embedding-3-large` model (`text-embedding-3-large model`). All sentence-level embeddings from the training set are stored in a vector database using ChromaDB (Contributors, 2023) and indexed using the Hierarchical Navigable Small World (Malkov and Yashunin, 2018) algorithm.

For each test instance, a query is created based on the input sentence, and its embedding is compared to all entries in the database using cosine similarity. The top-k most similar patient cases are then selected based on this similarity, and their information is used to generate the final prompt, which is subsequently processed by the LLM through Cot reasoning.

Generating Explanation of Sentence using CoT

- **Scoring individual acoustic features:**
  In this study, OpenAI's GPT-4o (Achiam et al., 2023) model was employed as LLM. To ensure consistent output, the temperature was fixed at 0.1. The scoring process involves analyzing the correlation between each acoustic features and the severity, and expressing this relationship as a score. The input configuration of the LLM first includes the "Explanation for Each Acoustic Feature" presented in Appendix A. This document provides the definition, clinical relevance, and interpretation notes for features such as Pitch, Intensity, Jitter (DDP/PPQ), Shimmer (APQ), HNR, F1/F2, and Speaking/Articulation rate, including factors like noise or microphone distance that may affect interpretation. It helps the model reason based on meaning rather than terminology, ensures consistent interpretation of feature directionality, and standardizes feature scales to improve comparability across sentences and patients. By serving as a semantic prior, it enhances the stability of scoring and explanation generation. Next, Relevant Patient Information obtained through the RAG method is incorporated to provide a personalized reference baseline that reflects physiological differences such as gender- and age-related variations in F0 or formants. This reduces distribution shift and improves the fairness and validity of model outputs. The Scoring Instruction defines the quantitative goal of linking features to severity, the evaluation scale (0-100), and output consistency constraints (e.g., fixed temperature and format). It minimizes output variability, ensures comparability across patients, and maintains clarity and accountability in clinical communication. The standardized output schema also supports seamless post-processing, visualization, and data storage. Finally, the Input Data include patient metadata (e.g., ID and gender), predicted severity, and individual acoustic feature values.

  This Module examines how specific features have impacted the sentence query and identifies which features have had a stronger influence on the sentence query. Each characteristic is then scored based on its influence, and this score is used to determine its importance, allowing for the generation of more personalized explanations. Refer to Table 2 for examples of each input.

- **Identifying impactful features**
  In this part, the scored results from the

previous step are used as input to determine whether each feature is classified as a positive acoustic or a negative acoustic. pos (positive acoustic), indicates that this acoustic feature is more similar to that of the HC group. On the other hand, neg (neagtive acoustic) indicates that the acoustic is more similar to that of the patient group. The pos and neg features are classified using a rule-based approach: if the score is 25 or below, it is classified as pos, and if the score is 80 or above, it is classified as neg. Scores falling between 25 and 80 are considered non-impactful and excluded from further interpretation. These thresholds were heuristically determined. The pos and neg features are identified as impactful acoustics, and this identification process enables a clear distinction between normal speech patterns and those that require improvement.

- **Explanation Generation for a single sentence with Shapley Values**
  The LLM receives as input the *Explanation for Each Characteristic* presented in Table 1, Sentence Explain Generation Instruction, Relevant Patient Information, Input Data, and the *Final Output Template*, along with the features scores and impactful features extracted in previous steps. Negative features, where the patient showed poor performance, are highlighted in the speech subsystem explanation, and the model provides appropriate suggestions for treatment. Conversely, if positive features are present, the corresponding subsystems include encouraging text that acknowledges the patient's strong performance. However, explanations based solely on the impactful features generated by the LLM may lack accuracy and interpretability. To address this limitation, SHAP values were additionally incorporated to enhance the explainability of the output. Since SHAP values quantitatively indicate the degree to which each feature contributes to the severity classification, they were considered well-suited for this task. By jointly utilizing the LLM-generated impactful features and the SHAP values, the overall interpretability and reliability of the explanation were improved.

The Explain Generator produces the evaluation of the patient's sentence-level speech. The evaluation consists of five components: severity, phonation, articulation, prosody, and overall. These components are structured in JSON format to facilitate consistent interpretation and support future clinical applications.

3.2 Debate Agent

The Debate Agent generates the final severity and evaluation of the patient by comparing and analyzing the results produced by six sentence agents. This analysis process employs the GPT-4o model, and to ensure consistent text generation, the temperature parameter is set to 0. The inputs used include the sentence-level evaluations generated by the sentence agents and the Debate Agent Instruction described in Table 2. Even for sentences spoken by the same patient, the predicted severity values may vary depending on the order in which the sentences are read. This effect is particularly pronounced in dysarthric patients, where respiratory fatigue often leads to increased variability in breathing and phonation during the latter part of the paragraph. As a result, while earlier sentences may resemble those of healthy individuals, later sentences can exhibit noticeable deviations. In other words, agents may produce different severity predictions for the same patient. The Debate Agent takes the outputs generated by the sentence-level agents as input and reanalyzes them to derive the patient's final severity and overall evaluation. The resulting output provides not only a detailed description of the patient's current speech impairment but also suggestions for treatment direction and areas that may require further diagnosis.

4 Experiment

4.1 Patient severity classification by sentence and speaker

To evaluate the effectiveness of the proposed agent-based framework, we conducted comparative experiments using three different methods: (1) **LGBM**, which performs severity classification on each sentence using a machine learning model and aggregates the results via majority voting; (2) **LGBM + PatientOne LLM**, which feeds the acoustic features of all six sentences into a single LLM for direct prediction without agent-based decomposition; and (3) **Dyxplain(Proposed)**, in which six Sentence Agents individually analyze each utterance, and a Debate Agent integrates the sentence-level results to produce the final severity prediction. The

Table 3: Sentence-level and Speaker-level Severity Classification

Method	Sentence-level F1 score (%)		Speaker-level F1 score (%)	
	Macro	Micro	Macro	Micro
LGBM	76.66	81.11	76.96	82.67
LGBM + PatientOne LLM	-	-	80.27	83.58
Dyxplain(Proposed)	**84.87**	**85.99**	**88.66**	**91.04**

Table 4: Performance Evaluation on Generating Assessment of Explainable Features

Method	Components	Semantic Similarity	Coherence	Relevance	Faithfulness
Vanilla	Acoustic Feature	47.05	66.25	52.41	41.16
Self-discover	Acoustic Feature	72.05	81.16	67.32	60.09
Dyxplain	Acoustic Feature	76.43	85.63	78.41	70.18
Dyxplain	SHAP	79.11	87.59	80.14	72.23
Dyxplain	Acoustic Feature + SHAP	**84.55**	**88.30**	**86.45**	**78.39**

LLM model used was GPT-4o.

As shown in Table 3, the proposed method significantly outperforms all baselines in both sentence-level and speaker-level classification. At the speaker level, our model achieved 88.66% macro-F1 and 91.04% micro-F1, demonstrating its robustness in assessing long-form speech inputs such as paragraph reading tasks. Notably, while the PatientOne setting leverages a single LLM to process aggregated input, it lacks the fine-grained reasoning enabled by agent decomposition. In contrast, the proposed framework benefits from sentence-level granularity and Debate Agent integration, allowing for precise modeling of intra-sentence variation-especially critical in dysarthric speech, where later sentences tend to degrade due to respiratory fatigue. These results confirm that our agent-based approach enhances both classification performance and interpretability in dysarthria assessment.

4.2 Comparison of patient explanation generation using G-eval

Table 4 presents the results of an experiment designed to evaluate the effectiveness of the medically specialized prompt incorporating CoT and RAG in generating interpretable explanations. This experiment also investigates whether the core components introduced in the Dyxplain framework-namely, acoustic features and SHAP values-are essential for improving the quality of explanations.

The experiment was conducted based on the G-eval (Liu et al., 2023) methodology, using GPT-4o as the backbone model. The reference data consisted of textual evaluations produced by three clinical pathologists who listened to patient speech and analyzed it in terms of phonation, articulation, prosody, and overall explanation. These reference texts were compared against the generated outputs from each method. G-eval was used to evaluate whether the generated explanations could produce results comparable to those of clinical pathologists, using four evaluation metrics.

4.2.1 Baseline

- **Vanilla:** A simple prompting strategy where only acoustic features are provided as input, with a basic instruction such as "analyze the values and diagnose the patient's condition." Neither CoT nor RAG is utilized in this setting.

- **Self-discover (Zhou et al., 2024):** A generic CoT-based prompting method adapted from prior work on complex reasoning tasks. Acoustic features are used as input, and the model is expected to reason autonomously. However, it does not incorporate SHAP values or any medical-specific prompt structure.

4.2.2 Metrics

The generated results are evaluated using four metrics, each with a score ranging from 0 to 100.

- **Semantic Similarity:** Evaluates how semantically similar the answer is to the reference answer.

- **Coherence:** Assesses whether the response is structurally complete and logically connected.

- **Relevance:** Measures how closely the response is related to the core subject of the problem.

- **Faithfulness:** It is a metric that measures the extent to which distortion or misunderstanding is avoided.

The results show that Dyxplain (Acoustic Feature + SHAP) achieves the highest performance across all evaluation metrics: Semantic Similarity (84.55), Coherence (88.30), Relevance (86.45), and Faithfulness (78.39). While using only SHAP or acoustic features individually also resulted in notable performance gains, the combination of both yielded the most interpretable and clinically aligned explanations. In particular, improvements in Faithfulness and Relevance suggest that the generated explanations accurately reflect the underlying model behavior-a crucial requirement for medical applications. These findings demonstrate that effective generation of interpretable explanations requires more than raw feature inputs. The integration of model interpretability signals (SHAP), structured reasoning mechanisms (CoT), and domain-specific knowledge (acoustic features) is critical for producing high-quality explanations that are suitable for real-world clinical settings.

5 Conclusion and Future Works

5.1 Conclusion

This study presents *Dyxplain*, an integrated pipeline for dysarthria evaluation, combining acoustic feature extraction, SHAP-based feature importance evaluation, and explanation generation using an LLM, resulting in significantly enhanced severity classification performance and interpretability compared to traditional ML models. In particular, by leveraging a dysarthria-specific medical pipeline, a speaker-level macro F1 score of 88.66% in severity classification and a semantic similarity score of 84.55 with clinical experts were achieved, demonstrating the effective analysis of patients' acoustic features and the factors influencing their condition. These results indicate that the proposed pipeline can provide valuable assessments and treatment guidance in real-world medical settings for both clinicians and patients.

5.2 Future Works

We will broaden evaluation beyond the current single-site Korean paragraph-reading setting by testing on public and multi-center dysarthria corpora and in cross-language scenarios. In particular, we plan cross-dataset studies on TORGO, UASpeech, ALS-TIMIT, and PC-GITA, alongside additional institutional cohorts to quantify domain and device shift. Beyond paragraph reading, we will include spontaneous monologues, sustained vowels, and diadochokinetic (DDK) tasks to study cross-task generalization and to develop multi-task training that learns shared and task-specific representations across phonation, articulation, and prosody.

To strengthen baselines, we will benchmark end-to-end and self-supervised encoders such as wav2vec2.0, HuBERT, and WavLM, as well as ECAPA-TDNN-style time-delay networks and Transformer/LSTM sequence models (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022; Desplanques et al., 2020). We will explore ordinal regression (e.g., CORAL) and continuous severity regression, and extend outputs to subsystem-level scoring with calibrated uncertainty. Robustness studies will target noise, reverberation, and device/channel shift, incorporating data augmentation (e.g., SpecAugment, time-stretch/pitch-shift), domain adaptation and test-time adaptation, and reporting of calibration and uncertainty (ECE/ACE, reliability diagrams, conformal prediction sets). We will also evaluate retrieval and aggregation design, including alternative embeddings and similarity metrics, top-k selection, and sentence-to-patient fusion strategies, and compare SHAP with Integrated Gradients (Sundararajan et al., 2017), DeepSHAP (Lundberg and Lee, 2017), LIME (Ribeiro et al., 2016), counterfactual exemplars, and perturbation-based faithfulness tests.

Human-centered evaluation will include blinded clinician ratings to assess correctness, usefulness, and actionability, as well as workflow metrics such as decision time and inter-rater agreement. We will study longitudinal sensitivity to therapeutic change and alignment with standardized clinical scales, and investigate personalization via patient-conditioned priors and adaptive weighting across sentences. To improve reproducibility and access, we will add open-source variants using widely available embeddings and LLMs (e.g., Llama/Mistral/Qwen) (Touvron et al., 2023; Jiang et al., 2023; Yang et al., 2024), release code, prompts, and ablation-ready scripts with fixed seeds, and document privacy-preserving deployment options (de-identification, access control, federated/edge inference).

6 Limitations

The present study is confined to a single language, a single standardized elicitation (paragraph reading), and a single-institution recording environment, with relatively few severe cases. This scope limits external validity and may bias decision boundaries, particularly under domain shift across microphones, rooms, noise conditions, or dialectal/prosodic patterns. Labels were derived from a coarse three-level severity scale, and inter-rater agreement was not measured; potential label noise and construct mismatch between NIHSS-derived severity and subsystem-specific impairment could affect both training and evaluation. The pipeline also includes an ASR-derived intelligibility proxy (Avg CER) computed with a specific recognizer; ASR errors and language/domain biases can be conflated with clinical signal, leading to over- or underestimation of impairment when recognition struggles for non-clinical reasons.

Methodologically, interpretability relies on TreeSHAP for a gradient-boosted decision tree and on heuristic impact thresholds that were not tuned with comprehensive sensitivity analyses. Alternative attribution and explanation strategies (Integrated Gradients, DeepSHAP, LIME, counterfactuals) were not compared, and the aggregation of sentence-level evidence into patient-level rationales was not formally validated for stability. Retrieval and Debate Agent behavior may be sensitive to embedding choice, similarity metric, top-k selection, prompt design, and temperature; we did not provide systematic ablations. Reported results emphasize point estimates without confidence intervals, statistical significance tests, or calibration and uncertainty analyses, which constrains interpretability of small observed gains.

Evaluation of explanation quality relied heavily on an LLM-based metric (G-eval) rather than blinded clinician judgments, prospective user studies, or downstream decision-support impact in clinical workflows. We did not report group-wise performance or calibration differences across demographics or devices, nor conduct out-of-distribution (OOD; out-of-distribution) robustness analyses, despite including age and gender as covariates. Finally, the use of closed-source embeddings and LLMs constrains exact reproducibility, auditability, and long-term accessibility; and although RAG uses de-identified summaries, retrieval over patient-like cases raises residual privacy and governance considerations that would warrant stricter safeguards and documentation in clinical deployments.

References

Sondes Abderrazek, Corinne Fredouille, Alain Ghio, Muriel Lalain, Christine Meunier, and Virginie Woisard. 2023. Interpreting deep representations of phonetic features via neuro-based concept detector: application to speech disorders due to head and neck cancer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:200–214. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Paul Boersma and David Weenink. 2009. Praat: doing phonetics by computer (version 5.1.13).

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Chroma Contributors. 2023. Chroma: Ai-native open-source embedding (vector) database. https://github.com/chroma-core/chroma. Accessed: 2025-05-20.

Frederic L. Darley, Arnold E. Aronson, and Joe R. Brown. 1969. Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research*, 12(2):246–269.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, interspeech$_2$020.$ISCA$.

Joseph R Duffy. 2019. *Motor speech disorders*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: a highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

H. Kim. 2005. Neurogenic speech and language disorders. In H. Sim, Y. Kim, J. Kim, H. Kim, S. Bae, M. Shin, S. Lee, and J. Han, editors, *Introduction to communicative disorders*, pages 149–178. Hakjisa, Seoul, Korea.

Li Khim Kwah and Joanna Diong. 2014. National institutes of health stroke scale (nihss). *Journal of Physiotherapy*, 60(1):61.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Scott Lundberg. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Preprint*, arXiv:1705.07874.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, and 1 others. 2023. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Soumya Reddy, Gabrielle Schwartzman, and R Hal Flowers. 2023. Chatgpt in dermatology clinical practice: potential uses and pitfalls. *Cutis*, 112(2).

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Preprint*, arXiv:1602.04938.

Prabin Sharma, Kisan Thapa, Dikshya Thapa, Prastab Dhakal, Mala Deep Upadhaya, Santosh Adhikari, and Salik Ram Khanal. 2023. Performance of chatgpt on usmle: unlocking the potential of large language models for ai-assisted medical education. *arXiv preprint arXiv:2307.00112*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *Preprint*, arXiv:1703.01365.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. Medagents: large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

OpenAI `text-embedding-3-large` model. OpenAI `text-embedding-3-large` model. `https://platform.openai.com/docs/models/text-embedding-3-large`. Accessed: 2025-05-20.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Ming Tu, Visar Berisha, and Julie Liss. 2017. Interpretable objective assessment of dysarthric speech based on deep neural networks. In *Interspeech*, pages 1849–1853.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Lingfeng Xu, Julie Liss, and Visar Berisha. 2023. Dysarthria detection based on a deep learning model with a clinically-interpretable layer. *JASA Express Letters*, 3(1):015201.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Zhenglin Zhang, Xiaolong Shang, Li-Zhuang Yang, Wenlong Ai, Jiawei Wang, Hongzhi Wang, Stephen T.C. Wong, Xun Wang, and Hai Li. 2023. Artificial intelligence-powered acoustic analysis system for dysarthria severity assessment. *Advanced Intelligent Systems*, 5(10):2300097. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aisy.202300097.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V Le, Ed Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024. Self-discover: Large language models self-compose reasoning structures. *Advances in Neural Information Processing Systems*, 37:126032–126058.

A Details of Features

Speech Acoustic Features Extracted for Paragraph Reading Task

- **Phonation**
  - **Pitch (std, min, max)**: Fundamental frequency (F0) and its standard deviation, minimum, and maximum value.
  - **Intensity (std, max)**: Speech signal's intensity and its standard deviation and maximum value.
  - **Jitter (DDP, PPQ)**: Measures of pitch variation; DDP (Difference of Differences of Periods) and PPQ (Pitch Period Perturbation Quotient).
  - **Shimmer (APQ)**: Measures of amplitude variation; APQ (Amplitude Perturbation Quotient).
  - **Harmonics-to-Noise Ratio**: Ratio of harmonic sound to noise in the voice.
- **Articulation**
  - **F1**: Frequency of the first formant.
  - **F2**: Frequency of the second formant.
  - **Avg CER**: Average Character Error Rate (CER) calculated across sentences in a paragraph.
- **Prosody**
  - **Speaking rate**: Rate of spoken syllables per second, including pauses.
  - **Articulation rate**: Rate of spoken syllables per second, excluding pauses.
  - **Avg syllable duration**: Average duration of syllables in the speech signal.
  - **Phonation ratio**: Ratio of speaking time to total time.
  - **Pause frequency ratio**: Ratio of the number of pauses to the total speech duration.
  - **Pause duration ratio**: Ratio of the number of pauses to the total pause duration.
  - **Pause ratio**: Ratio of total pause duration to the total speech duration.