Diagnosis of Dysarthria Severity and Explanation Generation Using XAI-Enhanced CLINIC-GENIE on Diadochokinetic Tasks

Anonymous ACL submission

Abstract

Deep neural network classifiers for dysarthria severity face limitations regarding interpretability and treatment guidance. To overcome these, we introduce **CLINIC-GENIE**, an explainable two-stage framework consisting of: (1) **CL**assification model using **IN**tegrated **I**nformation from **C**linically explainable acoustic features and speech representations **(CLINIC)**, a dysarthria severity classification model combining acoustic and speech embeddings with Clinically Explainable Acoustic Features (CEAFs) for enhanced interpretability and performance; and (2) **G**eneration of **E**xplanations from **N**umerical features using **I**nterpretability and patient **E**xamples **(GENIE)**, a module translating numerical data, such as CEAFs and their Shapley values, into intuitive natural language explanations via a large language model. In the severity classification experiments on the DDK dataset, CLINIC achieved a balanced accuracy of **0.952**, a 17.3% improvement over using CEAFs alone. In evaluation of the generated diagnosis, certified speech-language pathologists rated explanations from CLINIC-GENIE highly, with an average fidelity score of **4.94**, confirming enhanced clinical utility through intuitive, human-like interpretations. These results demonstrate that CLINIC-GENIE enhances clinical utility by improving classification accuracy and providing intuitive, human-like explanations. The code will be made publicly available on GitHub.

1 Introduction

Dysarthria is a motor speech disorder characterized by impaired articulation, phonation, and resonance resulting from neurological damage (Duffy et al., 2012). Clinical assessment of its severity is essential for formulating appropriate treatment plans and monitoring disease progression (Joshy and Rajan, 2021). However, continuous monitoring of treatment and disease progression by clinical pathologists is resource-intensive. For this reason, research on automatic dysarthria severity diagnosis using deep learning has been actively pursued.

Recent deep learning approaches have primarily focused on severity prediction alone, utilizing only one or two types of features such as mel-spectrograms (Suhas et al., 2020; Joshy and Rajan, 2023; Rathod et al., 2023), Wav2Vec 2.0 representations (Baevski et al., 2020), or Mel-Frequency Cepstral Coefficients (MFCC) (Hernandez et al., 2020; Bhattacharjee et al., 2023; Yeo et al., 2022).

While these methods enable accurate prediction of dysarthria presence and severity directly from speech, relying on such limited feature sets can overlook the complexity of speech disorders, and these black-box models lack the interpretability essential for clinical applications. These models have limited clinical applicability because they fail to explain the specific factors influencing their predictions. Therefore, explainable AI (XAI) is regarded as a prerequisite for safely integrating AI-based decision-support systems into clinical practice (Mancini et al., 2024; Shen et al., 2025).

To address this lack of explainability, we focused on the diadochokinetic (DDK) task among dysarthria assessment tools. The DDK task—rapid repetition of syllables like 'pa-ta-ka'—remains clinically valuable due to its ability to measure oral motor control without requiring linguistic competence, making it suitable even for severely impaired speakers (Wang et al., 2009; Segal et al., 2022). Furthermore, this task allows us to leverage features that speech-language pathologists (SLPs) actually use when providing explanations to patients.

In this study, we term these features Clinically Explainable Acoustic Features (CEAFs) and propose a **CL**assification model using **IN**tegrated **I**nformation from **C**EA and speech representations (CLINIC) that utilizes them. CLINIC incorporates 12 CEAFs along with speaker gender as inputs, in addition to mel-spectrograms and Wav2Vec 2.0 embeddings, to accurately diagnose dysarthria severity. Simultaneously, through Shapley values of the CEAFs, our model can provide information about which acoustic characteristics influence the prediction and how these features relate to clinically validated pathophysiological mechanisms (Shapley, 1953; Lundberg and Lee, 2017).

In clinical practice, it is crucial that patients with dysarthria recognize any issues in their own speech, as this awareness is a key factor in determining appropriate rehabilitation strategies and establishing effective treatment plans. However, the severity of dysarthric speech cannot be determined by any single acoustic feature, and these features do not vary in a strictly linear fashion, which limits the utility of simple interpretive approaches such as rule-based methods. For example, a high value on a particular acoustic feature cannot be immediately deemed abnormal; instead, it is more important to consider whether that value is relatively abnormal in light of interactions with factors such as the patient's gender or other characteristics. Interpretive validity is enhanced not by absolute values alone, but by determining whether a feature represents a relative outlier within a population sharing similar demographic attributes. Accordingly, a natural language explanation module called **G**eneration of **E**xplanations from **N**umerical features using **I**nterpretability and patient **E**xamples (GENIE) is proposed in this study. GENIE combines attributiion (Shapley value) analysis with RAG-based (Lewis et al., 2020) case retrieval to translate numeric prediction contributions into patient-friendly explanations. In GENIE, similar cases are retrieved based on CLINIC outputs and an evaluation metric termed CEAFs. Medical prompts are then utilized to generate large language model (LLM)-based natural language explanations, thereby providing patients with intuitive and clinically meaningful narratives. The generated explanations were validated through automated evaluation using G-EVAL(Liu et al., 2023) as well as expert assessments by SLPs. By converting complex numeric information into clinically interpretable explanations, this approach was found to enhance the transparency and trustworthiness of the AI model.

The primary contributions of our work are as follows:

- **CLINIC:** Clinically Explainable Acoustic Features (CEAFs), derived from the assessment criteria used by speech-language pathologists in real-world clinical setting, were employed to enhance the interpretability of the model. Additionally, integrating CEAFs with mel-spectrogram and Wav2Vec 2.0 embeddings led to improved severity classification performance.

- **GENIE:** enables effective interpretation of the patient's complex and nonlinear speech characteristics by quantitatively assessing the contribution of each feature and generating precise, persuasive explanations through comparison with similar patient cases.

- **Integrated Medical Speech Analysis Framework (CLINIC-GENIE).** To the best of our knowledge, this study presents the first implementation of a medical speech analysis framework that integrates classification (CLINIC), attribution of CEAFs (Shapley values), and natural language explanation (GENIE) into a single pipeline. By unifying analytical components that were previously addressed separately in explainable AI (XAI) research, this framework introduces a novel XAI approach that simultaneously satisfies both interpretability and clinical applicability.

2 Related Work

2.1 Deep Learning for Dysarthria and Other Speech-Based Disease Classification

Various studies have explored automatic methods for analyzing speech with dysarthria samples. Traditionally, MFCC (Hernandez et al., 2020; Bhattacharjee et al., 2023; Yeo et al., 2022), mel-spectrograms (Suhas et al., 2020; Joshy and Rajan, 2023; Rathod et al., 2023), or self-supervised representations (e.g., Wav2Vec 2.0, HuBERT (Hsu et al., 2021)) (Sanjay et al., 2024; Samptur et al., 2024) have typically been employed as input features, while some researchers have used additional speech features (e.g., F0) or combined them with MFCC (Hernandez et al., 2020; Yeo et al., 2022; van Bemmel et al.). From a model architecture perspective, these features are commonly fed into DNN-based classifiers (Hernandez et al., 2020; Bhattacharjee et al., 2023; Yeo et al., 2022; Suhas et al., 2020; Joshy and Rajan, 2023; Rathod et al., 2023; Sanjay et al., 2024), which leverage either acoustic representations (e.g., MFCC or mel-spectrogram) or self-supervised representations to predict dysarthria severity.

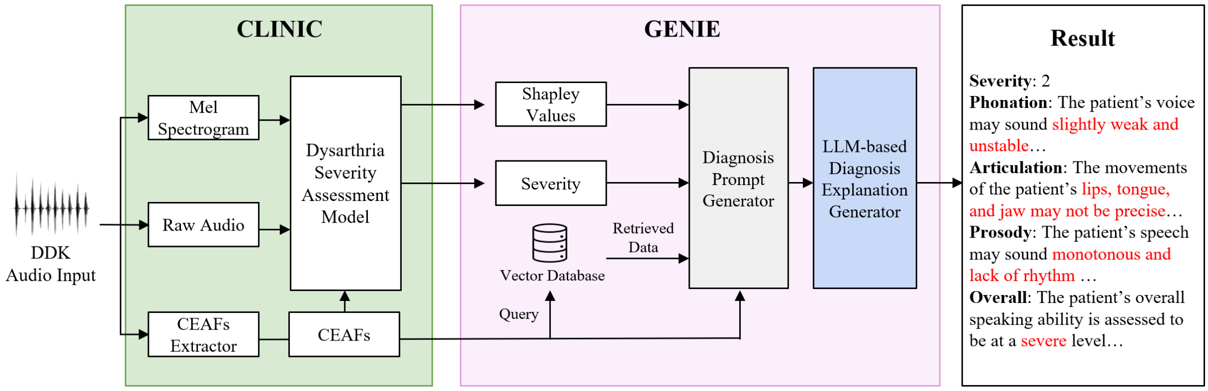


Figure 1: The overall architecture of CLINIC-GENIE. **CLINIC:** From the DDK audio input, three representations are derived: a mel-spectrogram, the raw audio waveform, and CEAFs extracted by the CEAFs Extractor. These features are integrated within the Dysarthria Severity Assessment Model to predict dysarthria severity. **GENIE:** The resulting CEAF vector is employed as a query to a vector database so that acoustically similar patient cases can be retrieved. The retrieved cases, together with the Shapley values, severity, and CEAFs are assembled into a Diagnosis Prompt Generator and provided to a LLM-based Diagnosis Explanation Generator.

These approaches achieved plausible performance, lacking explainability for their predictions.

2.2 Explainability in Deep Learning Models

To enhance interpretability, Shapley values (Shapley, 1953; Lundberg and Lee, 2017) were introduced, assigning fair and transparent contributions to features based on cooperative game theory. Similarly, Integrated Gradients (Sundararajan et al., 2017) and DeepLIFT (Shrikumar et al., 2017) compare inputs against a baseline capturing relative feature importance or activation differences to clarify how each feature influences predictions of the model.

However, in medical and healthcare settings, simply using these XAI method to identify "which factors influenced the outcome" may not be sufficient. Models in these contexts must utilize data in ways that closely align with real clinical evidence and be easily understood by patients, as these aspects directly impact treatment decisions (Markus et al., 2021; Amann et al., 2020).

2.3 Translating Numerical Data into Natural Language Explanations

Some studies leverage LLMs to convert numerical data into natural language explanations. For example, iPrompt (Singh et al., 2022) proposes an algorithm that automatically generates explanations using LLMs to clarify patterns in data. In addition, there has been research on converting Shapley values into more accessible natural language explanations, thereby making the prediction process clearer to a broader audience (Zeng, 2024).

2.4 Large Language Models in Healthcare

LLMs have recently emerged as powerful tools in healthcare applications, offering new capabilities for generating clinical explanations, interpreting medical data, and supporting healthcare professionals in decision-making processes (Thirunavukarasu et al., 2023; Nazi and Peng, 2024).

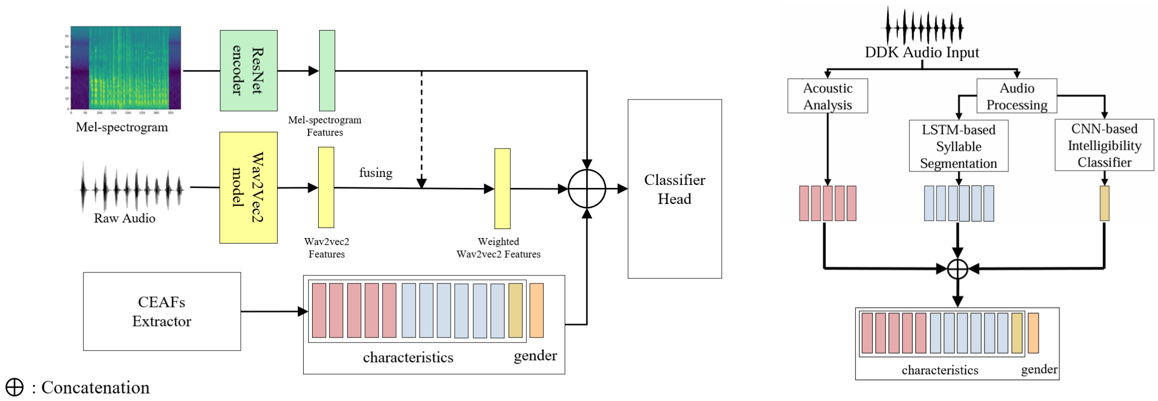
RAG (Lewis et al., 2020) combines LLM with retrieval systems to provide more accurate and reliable explanations. Before generating explanations, RAG retrieves relevant clinical data to ensure that the explanations are factual and precise (Xiong et al., 2024). This approach is crucial for providing personalized dysarthria diagnoses and treatment plans, where limited clinical data are available.

3 Interpretable Dysarthria Diagnosis System

Figure 2 provides an overview of the CLINIC-GENIE, which consists of two main components: (1) **CLINIC**, a severity classification model that incorporates CEAFs and mel-spectrogram and Wav2Vec 2.0 representations extracted from dysarthric speech, and (2) **GENIE**, a natural-language explanation generator that using a large language model.

3.1 CLINIC: A Seveity Classification Model

The CLINIC integrates CEAFs with mel-spectrogram and Wav2Vec 2.0 representations



(a) Dysarthria Severity Assessment Model (b) CEAFs Extractor

Figure 2: Overall structure of **CLINIC**. (a) Dysarthria Severity Assessment Model integrates CEAFs with representations derived from mel-spectrograms and raw audio (via Wav2Vec 2.0). Each representation is encoded separately, and their embeddings are concatenated into a single vector, subsequently fed into the Classifier Head for classification (b) The architecture of the CEAFs Extractor, which derives speaker characteristics information by analyzing DDK audio input through acoustic analysis, LSTM-based syllable segmentation, and CNN-based intelligibility classification.

to effectively capture complementary aspects of dysarthric speech. Specifically, mel-spectrograms encode detailed local acoustic characteristics (Hershey et al., 2017), whereas Wav2Vec 2.0 embeddings provide global contextual information by modeling broader temporal dependencies within speech signals (Baevski et al., 2020). Additionally, CEAFs enable clinically meaningful interpretations of acoustic features, facilitating a rational assessment process and enhancing the model's explainability.

3.1.1 CEAFs: Clinically Explainable Acoustic Features

Two types of DDK tasks, Alternating Motion Rate (AMR) and Sequential Motion Rate (SMR), are used to extract key characteristics relevant to dysarthria evaluation (Darley et al., 1969; Duffy et al., 2012). AMR assesses articulatory speed and consistency by repeating the same syllable(e.g., /pa/, /ta/, or /ka/), SMR evaluates the ability to rapidly transition between different articulatory positions using syllable sequences, such as /pataka/(Darley et al., 1969; Duffy et al., 2012). These tasks provide insights into the coordination, speed, and consistency of articulatory movements, which are essential for accurately assessing dysarthria severity (Darley et al., 1969; Duffy et al., 2012).

Table 1 presents 12 CEAFs along with their definitions and the corresponding extraction methods which are illustrated in Figure 2 (b). CEAFs were derived from two primary sources: the Mayo Clinic rating system (Darley et al., 1969), a widely recognized framework for dysarthria evaluation, which provided the basis for feature extraction, and NeuroSpeech (Orozco-Arroyave et al., 2018), a software tool for automated DDK analysis, which was used to derive the CEAFs. Together, these features enable a comprehensive evaluation of phonatory, prosodic, and articulatory aspects within the DDK task. CEAFs were extracted using acoustic analysis, an LSTM-based syllable segmentation model, and a CNN-based intelligibility classifier (Oh et al., 2023), as detailed in Appendix B.

3.1.2 Dysarthria Severity Assessment

Figure 2 (a) illustrates the architecture of the proposed dysarthria severity assessment model, which integrates multiple acoustic features using a joint representation learning approach (Huang et al., 2020).

CEAFs measured from the DDK task, along with gender information, were normalized using min-max scaling to mitigate scale discrepancies. The normalized features were subsequently processed through a fully connected layer to generate embedding vectors.

Features are extracted from the mel-spectrogram using a ResNet (He et al., 2016) model, capturing averaged characteristics across the frequency and time axes. These features are utilized as embedding vectors. The Wav2Vec 2.0 (Baevski et al., 2020) model processes raw audio signals to gener-

Characteristic	Definition	Extraction Method
F0 variability (st)	Variance of the fundamental frequency (semitones)	Acoustic Analysis
F0 variability (Hz)	Variance of the fundamental frequency (Hz)	
Avg. energy (dB)	Mean signal energy	
Energy variability (dB)	Standard deviation of energy	
Max. energy (dB)	Maximum signal energy	
DDK rate (syll/s)	Number of syllables per second	LSTM-based Syllable segmentation
DDK mean duration (ms)	Average syllable duration	
DDK regularity	Standard deviation of syllable durations	
Pause rate (pauses/s)	Number of pauses per second	
Pause mean dur. (ms)	Average pause duration	
Pause regularity	Standard deviation of pause durations	
Intelligibility score	Listener's understanding of the spoken content at the syllable level	CNN-based Intelligibility Classifier

Table 1: Clinically Explainable Acoustic Features (CEAFs) automatically extracted from the DDK task.

ate frame-level representation vectors, which are subsequently used as embedding vectors for raw audio. Specifically we utilized the publicly released wav2vec2-large-xlsr-53 pre-trained model (Conneau et al., 2021). Pre-training on roughly 56k h of speech across 53 languages enable to capturing the complex acoustic cues of dysarthria. Previous studies have demonstrated that combining mel-spectrogram and Wav2Vec 2.0 features enhances the ability to capture both local and global information. In this study, Wav2Vec 2.0 embeddings derived from raw audio are fused with mel-spectrogram feature vectors extracted from the ResNet model through an attention based mechanism. The fused representations are utilized as input features for the dysarthria severity assessment model.

Three embedding vectors are concatenated into a single vector: a CEAF vector, a mel-spectrogram vector, and a fused vector that combines the mel-spectrogram and Wav2Vec 2.0 representations. This concatenated vector is forwarded to the final classifier head, which predicts the probabilities of dysarthria severity. A weighted categorical cross entropy loss function (Cui et al., 2019) is employed during training to mitigate data imbalance.

3.1.3 Extraction Shapley Values

Shapley values quantify how much each feature contributes to the model output by contrasting the prediction obtained with the feature at its actual value against the prediction when that feature is fixed at a baseline (typically its expected value). The original formulation of Shapley values is detailed in Appendix A

If we can determine how strongly each CEAF influences the predicted severity, we can capture valuable cues for diagnosing dysarthria. Using the CLINIC, we first obtain the predicted severity and then compute Shapley values to extract the numerical contribution of each CEAF. The following section explains how these numerical scores are transformed into natural-language explanations.

3.2 GENIE: A Natural Language Explanation Generator

GENIE is a LLM module that combines previously predicted severity with Shapley values to produce patient-specific diagnostic narratives. Using RAG, the module retrieves prior cases with comparable assessment profiles and contrasts their CEAFs, thereby generating fine-grained, clinician-oriented explanations that highlight each patient's salient deficits and recommended focal points.

3.2.1 Retrieval of Analogous Patient Cases

The retrieval component operates entirely at inference time, with no additional training required. For each test instance, we construct a structured feature-based query from the CLINIC output. Specifically, we form a dictionary mapping each CEAFs name to its numeric value, and we include the predicted severity under the key `finalprediction`. The vector database $\mathcal{D}_{\text{DB}} = \{d_1, \ldots, d_n\}$ is constructed from the training set, with each patient represented as a document d_i containing that patient's CEAFs vector and ground-truth severity. Each document d_i is embedded as a 3,072-dimensional vector representation e_{d_i} using (`text-embedding-3-large` model) model, and all such document embeddings are stored and indexed using ChromaDB (Contributors, 2023) with a Hierarchical Navigable Small World index (Malkov and Yashunin, 2018). The test-time query d_Q is similarly embedded as e_{d_Q}, and cosine similarity is computed between the

Severity	CEAFs only	CEAFs + Mel	CEAFs + Wav2Vec 2.0	CLINIC (ours)
0 (Healthy)	0.750	**1.000**	**1.000**	**1.000**
1 (Mild to Moderate)	0.837	**0.980**	0.898	0.857
2 (Severe)	0.750	0.500	0.500	**1.000**
Balanced Acc.	0.779	0.827	0.799	**0.952**

Table 2: Accuracy by severity and input configuration.

query and each document embedding:

$$sim(e_{d_Q}, e_{d_i}) = \frac{e_{d_Q} \cdot e_{d_i}}{|e_{d_Q}||e_{d_i}|} \quad (1)$$

Finally, the top-k most similar documents are selected as relevant patient cases, which are then used to construct prompts for input to a Diagnosis Prompt Generator.

3.2.2 Diagnosis Prompt Generator

The diagnosis prompt generator uses four types of input as conditions: (1) the severity prediction from the CLINIC, (2) the numeric values of the CEAFs, (3) the Shapley values corresponding to each CEAF, and (4) relevant patient cases. The prompt generator is conditioned on the four task-specific DDK severity scores and is additionally provided with the patient's final severity, which is obtained through majority voting over those scores. For each patient, severity for the four DDK tasks is predicted by the CLINIC module, and a final severity label is assigned based on majority voting among these predictions. In clinical practice, it is essential for patients to understand which aspects of their condition require improvement. While CEAFs sufficiently describe the patient's acoustic profile, they are insufficient to identify the most influential features affecting the patient. Therefore, Shapley values are incorporated to explicitly highlight the features that contribute most significantly. The core prompt components are in Appendix H and a full example prompt is provided in Appendix I.1.

3.2.3 LLM-based Diagnosis Explanation Generator

Using the prompt generated by the diagnosis prompt generator, the LLM-based Diagnosis Explanation Generator employs OpenAI GPT-4o (OpenAI, 2025) to synthesize a diagnostic explanation spanning four clinical dimensions of dysarthria—phonation, articulation, prosody, and overall severity. The generator analyzes the CEAF values in conjunction with their corresponding Shapley attributions, thereby smoothing the underlying information and identifying which acoustic features exert the greatest influence and which remain deficient for the patient. The system prompt instructs the LLM to (i) interpret the provided inputs, (ii) discuss each CEAF in proportion to its Shapley value, and (iii) produce a patient-friendly diagnostic report in Korean. By explicitly decomposing the reasoning process into these sequential steps, the module is operated in a chain-of-thought (Wei et al., 2022) paradigm. To ensure consistency and mitigate hallucinations in the generated text, a fixed prompt template is employed and the generation temperature is set to 0.1.

4 Experiment

4.1 Dataset

The dataset consists of 59 healthy controls (HCs) and 321 patients, totaling 380 participants aged between 20 and 84 years. The healthy controls and patients were recruited in collaboration with [anonymized for review]. The data collection process, including recordings of the DDK task and clinical assessments, was approved by the Institutional Review Board (IRB) of the participating institutions, and informed consent was obtained from all participants. To collect corpus for dysarthria assessment, recordings of the DDK task were gathered from the speakers. DDK utterances consist of repeated syllables such as 'pa', 'ta', 'ka', and 'pataka'. Clinical data, including dysarthria severity and gender, were also collected. The severity of dysarthria is categorized into three levels: Healthy (0), Mild to Moderate (1), and Severe (2). A neurosurgeon assessed and labeled the severity using the National Institute of Health Stroke Scale (NIHSS) criteria (Kwah and Diong, 2014). The HCs were classified as severity 0, while patients were categorized as severity 1 or 2. In total, the dataset comprises 1,536 utterances, collected from 59 individuals with a severity 0, 290 with a severity 1, and 31 with a severity 2, including 239 male and 141 female speakers. Because obtaining data from patients with severity 2 is challenging in real-world clinical settings, the number of participants in this category is relatively small. For severity classification evaluation, we conducted testing using 244 utterances from 61 speakers (31 males and 30 females) who were not included in the training and validation process of the CLINIC. Among them, 8 speakers were labeled with severity 0, 49 with severity 1, and 4 with severity 2. Throughout the entire dataset, patient IDs were used instead of names to ensure anonymity.

4.2 Dysarthria Severity Assessment

To examine how best to exploit CEAFs information, we ran four ablation experiments under a unified classifier head. First every speech-derived representation (mel-spectrogram, Wav2Vec 2.0 features, or both) is routed through the embedding procedure described in 3.2. Then these embedding vectors are passed through a fully connected (FC) layer, producing a 128-dimensional vector. The 13-dimensional CEAFs vector follows a parallel two-layer FC path that also produces a 128-dimensional embedding. Detailed experiment method and model configurations for all variants are provided in Appendix C and Appendix D, separately. Test set accuracies for each severity are summarized in Table 2.

4.2.1 Results

As Table 2 shows, leveraging CEAFs with additional speech representations generally improved performance over using CEAFs alone. Providing patients with accurate and timely diagnoses is critically important (Ball et al., 2015), especially for those with severe severity. Therefore, when selecting our model, we considered not only the overall performance but also how accurately it predicted the severity 2. Although models using CEAFs and mel-spectrograms performed well for patients with severity 1, they accurately predicted only half of the severity 2 cases. In contrast, the CLINIC successfully identified all severity 2 patients. Therefore, we selected the CLINIC as our final system. More detailed results, including the confusion matrix, can be found in Appendix E.

4.3 Effectiveness of GENIE in Generalization

The experiment was designed to determine whether each component of the GENIE is indispensable. To verify explanatory effectiveness, two evaluation protocols, automated and human expert evaluation, were applied, and the text for each pipeline was generated with GPT-4o using $k = 3$.

4.3.1 Medical Explanation Evaluator

Medical Explanation Evaluator framework was developed to automatically assess the generated texts by GENIE. The framework, instantiated with GPT-4o, applies the g-eval (Liu et al., 2023) methodology on the full test set and evaluates each method's outputs across five metrics on a 0–100 scale. We set the generation temperature to 0.1. To ensure ethical integrity and fairness, the reported results represent the average values obtained from five repeated runs. The selection criteria and descriptions of the metrics are presented in Appendix G. Among the evaluation metrics, Semantic Equivalence and Fidelity were selected as the major criteria because they indicate how faithfully the generated explanations reflect clinical reasoning. Consistency, Relevance, and Patient-friendliness were designated as minor criteria. The detailed prompt used for the Evaluator can be found in Appendix I.2.

Reference data were compiled by three SLPs after they listened to the patient recordings in the test set. Textual descriptions were produced for five aspects: severity, phonation, prosody, articulation, and overall assessment. Examples of the reference data can be found in Appendix J.

According to the table 3, the vanilla baseline provides only CEAFs to the LLM in the first row. Semantic Equivalence was observed at 62.95 and Fidelity at 61.43, the bottom values for each metric.
CLINIC integration: CLINIC directly contributed to improvements in nearly every metric, yielding markedly closer alignment with clinical judgments than the baselines.
RAG-based contextualisation: When RAG was added to the CLINIC-only system, increases were observed in both Semantic Equivalence and Fidelity, indicating that contextual information supplied by similar patient cases endowed the explanations with richer content and stronger semantic coherence.
Exposure of Shapley attributions: When Shapley values were introduced, every metric increases by a further two to three points, and near-maximal values were achieved across the board. Notably, Semantic Equivalence reached 83.93 and Fidelity 79.38, confirming that an explicit disclosure of the model's reasoning maximised the perceived trustworthiness of the generated explanations.

Configuration				Mean Scores				
CEAFs	Model for Pred. Severity	RAG	Shap	Semantic Eq.	Fidelity	Consistency	Relevance	Patient–friendliness
O	X	X	X	62.95	61.43	78.75	71.95	79.93
O	CEAFs + Mel	O	O	82.23	77.23	89.37	85.00	**92.43**
O	CEAFs + Wav2Vec 2.0	O	O	82.68	77.77	88.93	84.36	91.57
O	CLINIC	X	X	77.95	73.57	86.07	81.32	89.66
O	CLINIC	O	X	81.34	76.43	87.41	83.57	90.23
O	CLINIC	O	O	**83.93**	**79.38**	**89.38**	**85.39**	91.30

Table 3: Component-wise ablation results for GENIE in the dysarthria-specific automatic evaluation. Presence (O) or absence (X) indicates whether each module is included. "Model for Pred. Severity" denotes the model that produced the predicted severity. An "X" indicates that no predicted-severity component is included at all. RAG shows whether similar-patient inputs are provided, and Shap shows whether Shapley values are included. Scores are reported on a 0–100 scale for five quality metrics generated by the LLM. Detailed descriptions of the prompt are provided in Appendix I.2.

Configuration				Mean Ratings				
CEAFs	Model for Pred.Severity	RAG	Shap	Semantic Eq.	Fidelity	Consistency	Relevance	Patient–friendliness
O	X	X	X	-	3.81	4.47	4.94	4.64
O	CEAFs + Mel	O	O	-	4.50	4.31	**5.00**	4.56
O	CEAFs + Wav2Vec 2.0	O	O	-	4.83	**4.94**	**5.00**	4.50
O	CLINIC	X	X	-	4.86	4.56	4.89	4.67
O	CLINIC	O	X	-	4.89	4.75	**5.00**	4.67
O	CLINIC	O	O	-	**4.94**	4.92	**5.00**	**4.69**

Table 4: Component-wise ablation results for the CLINIC-GENIE based on human expert evaluation(1–5 Likert).

4.3.2 Human Expert Evaluation

An expert evaluation was conducted in which 12 patients, randomly selected at a rate of four per severity, were assessed. Three SLPs first listened to each patient's DDK voice recording and then rated the explanations generated by each method on four metrics, using a 1–5 Likert scale. Because the explanations were evaluated directly by clinical pathologists, the Semantic Equivalence metric used in automatic evaluation was omitted. As shown in Table 4, the GENIE configuration that integrates CLINIC prediction, RAG retrieval, and Shpley values attribution achieves the best performance under expert review. Its Fidelity score rises from 3.81 in the baseline to 4.94, an improvement of almost 30 percent. The high agreement between expert evaluation scores and the automatic evaluation results in Table 3 supports the reliability of the evaluation metrics. Both Table 3 and Table 4 show that the lowest scores were obtained by the vanilla baseline model using only CEAFs, while the highest scores were achieved by the pipeline proposed in this paper. Additionally, the relative score distributions between the two evaluations are largely similar. Although differences in evaluation methods cause some variance in absolute scores, the relative rankings and score trends remain consistent, demonstrating the reliability of the proposed automatic evaluation system.

5 Conclusion

An integrated framework, **CLINIC-GENIE**, is proposed for the simultaneous classification and explanation of DDK speech. By combing CEAFs, mel-spectrograms, and Wav2Vec 2.0 representations, the CLINIC module attains a balanced accuracy of 0.952 and correctly identifies all severe cases. The GENIE module combines Shapley attributions with RAG-retrieved analogous cases to generate patient-oriented explanations covering four clinical dimensions: phonation, articulation, prosody, and overall severity and achieves top scores on nearly every automatic and expert metric. These results suggest that the framework can help clinicians and patients intuitively understand the rationale behind AI decisions, thereby accelerating early diagnosis and personalized rehabilitation planning while mitigating the wider societal burden of dysarthria care.

6 Limitations

The clinical corpus used in this study is imbalanced across severities, with markedly fewer speakers in the severity 2. This scarcity can constrain the model capacity. Future work will focus on enlarging and rebalancing the dataset—particularly by recruiting more severe speakers or exploring data-augmentation strategies to mitigate this limitation. Additionally, our framework is trained and evaluated solely on DDK speech. Its ability to generalize to more natural speech has not yet been verified.

References

Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20:1–9.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

John R Ball, Bryan T Miller, and Erin P Balogh. 2015. Improving diagnosis in health care.

Tanuka Bhattacharjee, Anjali Jayakumar, Yamini Belur, Atchayaram Nalini, Ravi Yadav, and Prasanta Kumar Ghosh. 2023. Transfer learning to aid dysarthria severity classification for patients with amyotrophic lateral sclerosis. In *Proc. INTERSPEECH*, volume 2023, pages 1543–1547.

Paul Boersma and David Weenink. Praat: Doing Phonetics by Computer [Computer Program]. Version 6.4.27, retrieved 27 January 2025.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2426–2430. Introduces the *wav2vec 2.0 XLSR-53* multilingual pre-trained model.

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

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *Proc. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9260–9269.

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.

Joseph R Duffy and 1 others. 2012. *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Health Sciences.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Abner Hernandez, Sunhee Kim, and Minhwa Chung. 2020. Prosody-based measures for automatic severity assessment of dysarthric speech. *Applied Sciences*, 10(19):6999.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, and 1 others. 2017. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

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. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. 2020. Fusion of Medical Imaging and Electronic Health Records Using Deep Learning: A Systematic Review and Implementation Guidelines. *npj Digital Medicine*, 3(1):1–9.

Amlu Anna Joshy and Rajeev Rajan. 2021. Automated dysarthria severity classification using deep learning frameworks. In *2020 28th European signal processing conference (EUSIPCO)*, pages 116–120. IEEE.

Amlu Anna Joshy and Rajeev Rajan. 2023. Dysarthria severity classification using multi-head attention and multi-task learning. *Speech Communication*, 147:1–11.

Li Khim Kwah and Joanna Diong. 2014. National institutes of health stroke scale (nihss). *Journal of physiotherapy*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *Proc. International Conference on Learning Representations*.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30.

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.

Eleonora Mancini, Francesco Paissan, Paolo Torroni, Mirco Ravanelli, and Cem Subakan. 2024. Investigating the effectiveness of explainability methods in parkinson's detection from speech. *arXiv preprint arXiv:2411.08013*.

Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. 2021. The Role of Explainability in Creating Trustworthy Artificial Intelligence for Health Care: a Comprehensive Survey of The Terminology, Design Choices, And Evaluation Strategies. *Journal of Biomedical Informatics*, 113:103655.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI.

J. Oh, H. Park, and J. Kim. 2023. Speech Intelligibility Prediction of Dysarthri Using Deep Convolutional Networks. In *Proc. 18th Asia Pacific International Conference on Information Science and Technology*.

OpenAI. 2025. Gpt–4o model. `https://platform.openai.com/docs/models/gpt-4o`. Accessed: 2025-05-20.

Juan Rafael Orozco-Arroyave, Juan Camilo Vásquez-Correa, Jesús Francisco Vargas-Bonilla, R. Arora, N. Dehak, P. S. Nidadavolu, H. Christensen, F. Rudzicz, M. Yancheva, H. Chinaei, A. Vann, N. Vogler, T. Bocklet, M. Cernak, J. Hannink, and Elmar Nöth. 2018. NeuroSpeech: An Open-Source Software for Parkinson's Speech Analysis. *Digital Signal Processing*, 77:207–221.

Siddharth Rathod, Monil Charola, Akshat Vora, Yash Jogi, and Hemant A Patil. 2023. Whisper features for dysarthric severity-level classification. *Small*, 12(768):12.

Neelesh Samptur, Tanuka Bhattacharjee, Anirudh Chakravarty K, Seena Vengalil, Yamini Belur, Atchayaram Nalini, and Prasanta Kumar Ghosh. 2024. Exploring syllable discriminability during diadochokinetic task with increasing dysarthria severity for patients with amyotrophic lateral sclerosis. In *Proc. Interspeech 2024*, pages 4114–4118.

B Sanjay, Priyadharshini Mk, P Vijayalakshmi, and T Nagarajan. 2024. Severity classification and dysarthric speech detection using self-supervised representations. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 621–628.

Kaitlin Schuessler. 2010. Performance of Alternating Motion Rate (AMR) in Individuals With Parkinson's Disease Under External And Internal Cueing Conditions. Master's thesis, University of Colorado at Boulder.

Yael Segal, Kasia Hitczenko, Matthew Goldrick, Adam Buchwald, Angela Roberts, and Joseph Keshet. 2022. Ddktor: Automatic diadochokinetic speech analysis. *arXiv preprint arXiv:2206.14639*.

Lloyd S Shapley. 1953. A value for n-person games. *Contribution to the Theory of Games*, 2.

Matthew Shen, Pouria Mortezaagha, and Arya Rahgozar. 2025. Explainable artificial intelligence to diagnose early Parkinson's disease via voice analysis. *Scientific Reports*, 15:11687.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR.

Chandan Singh, John X Morris, Jyoti Aneja, Alexander M Rush, and Jianfeng Gao. 2022. iprompt: Explaining data patterns in natural language via interpretable autoprompting. *ArXiv preprint*, 2210.

BN Suhas, Jhansi Mallela, Aravind Illa, BK Yamini, Nalini Atchayaram, Ravi Yadav, Dipanjan Gope, and Prasanta Kumar Ghosh. 2020. Speech task based automatic classification of als and parkinson's disease and their severity using log mel spectrograms. In *2020 international conference on signal processing and communications (SPCOM)*, pages 1–5. IEEE.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

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.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Loes van Bemmel, Chiara Pesenti, Xue Wei, and Helmer Strik. Automatic assessments of dysarthric speech: the usability of acoustic-phonetic features.

Yu-Tsai Wang, Ray D Kent, Joseph R Duffy, and Jack E Thomas. 2009. Analysis of diadochokinesis in ataxic dysarthria using the motor speech profile program™. *Folia Phoniatrica et Logopaedica*, 61(1):1–11.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Vincent Y. Zhao, Kelvin Chen, J. Zhou, and et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, New Orleans, LA.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Eun Jung Yeo, Kwanghee Choi, Sunhee Kim, and Minhwa Chung. 2022. Cross-lingual dysarthria severity classification for english, korean, and tamil. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 566–574. IEEE.

Xianlong Zeng. 2024. Enhancing the interpretability of shap values using large language models. *arXiv preprint arXiv:2409.00079*.

A Shpely Values Formulation

The contribution for feature i is defined as :

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \cdot (f(S \cup \{i\}) - f(S)) \tag{2}$$

following the original Shapley value formulation (Shapley, 1953) and its adaptation for model explanations (Lundberg and Lee, 2017), where F denotes the full set of features, S is a subset not containing i, and $f(\cdot)$ represents the expected model output when only the features in the given set are known (with the others marginalized).

B CEAFs Extraction Method

This appendix provides detailed descriptions of the methods used to extract CEAFs, including acoustic analysis, an LSTM-based syllable segmentation model, and a CNN-based intelligibility classifier.

Acoustic analysis was performed using the Praat software(Boersma and Weenink). The LSTM-based model quantified the rate, duration, and regularity of pronunciation and respiration by segmenting audio into speech and non-speech frames. The model consists of 16 LSTM layers and a fully connected (FC) layer. Raw audio signals were converted into spectrograms and fed into the model, which classified each frame as speech or non-speech. Frame-level predictions were aggregated into segment-level results by grouping consecutive frames with identical classifications. Speech segments shorter than 0.07 seconds were classified as silence, and silence segments longer than 0.14 seconds were used to calculate the pause rate. These threshold values (0.07 and 0.14 seconds) were determined based on the best performance observed on the training set. The silence threshold of 0.14 seconds was determined based on previous AMR task research, which found that healthy adults produce syllables at an average rate of approximately 0.143 seconds per syllable(Schuessler, 2010). The intelligibility classifier employs a ResNeXt-based CNN model(Oh et al., 2023) to classify speech samples into one of five ordinal intelligibility levels, ranging from 1 (least intelligible) to 5 (most intelligible).

C Experiments Details

For the dysarthria severity assessment model, the dataset was divided into training, validation, and testing sets following an 8:1:1 ratio, stratified by severity levels. The model was trained using the AdamW(Loshchilov and Hutter, 2018) optimizer with a learning rate of 0.00003. Model selection was performed on the validation set using macro-F1. For each model, we predicted a severity for every utterance and then applied majority voting across all utterances produced by a given patient to derive that patient's final dysarthria severity.

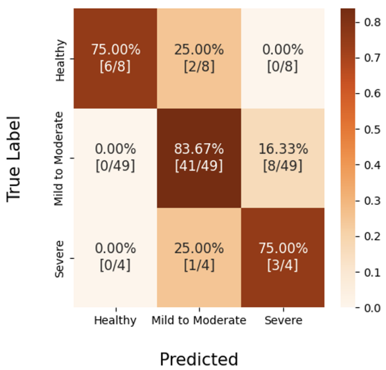
D Classification Model Configuration

The detailed information about the models used in the severity assessment experiments is provided in Table 5. All models share an identical CEAFs layer structure, takes as input the 12 CEAFs along with the speaker's gender. The classifier heads adapt to the dimensionality of the combined features (128-dim for model with only CEAFs, 256-dim for others). This design allows us to systematically assess how different speech representations contribute to dysarthria severity classification performance.

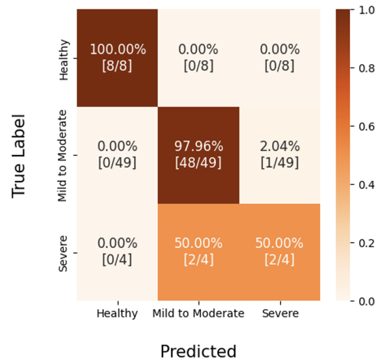
In Table 5, the "Mel-Path" and "W2V Path" columns indicate the processing pipelines for mel-spectrogram and Wav2Vec 2.0 representations, respectively, showing how these inputs are integrated into the overall model architecture.

E Detailed Result

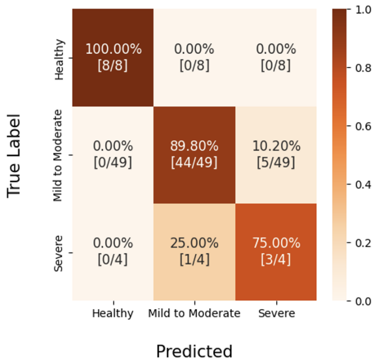
The confusion matrices of the severity classification results for 4.2 are illustrated in Fig. 3.



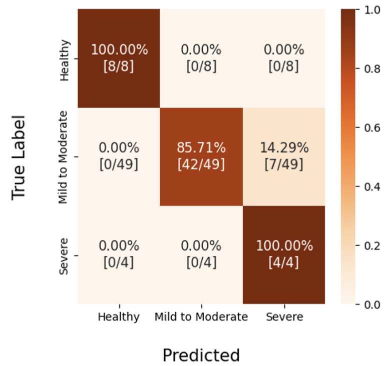
(a) CEAFs Only



(b) CEAFs + mel-spectrogram



(c) CEAFs + Wav2Vec 2.0



(d) CLINIC (Ours, CEAFs + mel-spectrogram + Wav2Vec 2.0)

Figure 3: Confusion matrices of severity classification results for four different feature combinations: (a) CEAFs Only, (b) CEAFs + mel-spectrogram, (c) CEAFs + Wav2Vec 2.0, and (d) CLINIC.

Table 5: Detailed architecture of the proposed models for dysarthria severity classification. (BN: Batch Normalization)

Model	Architecture
1. Only CEAFs	**CEAFs Layer:** 13 → 128 → BN, Dropout → 128 → BN, Dropout **Classifier:** 128 → 128 → 128 → 3 (with BN, ReLU, Dropout 0.3)
2. Mel + CEAFs	**CEAFs Layer:** Same as Model 1 **Mel Path:** Mel-Spectrogram → ResNet-50 → 2048-dim → Linear → 512 → 128 **Classifier:** Concat[CEAFs(128), Mel(128)] = 256 → 128 → 128 → 3 (with BN, ReLU, Dropout 0.3)
3. W2V + CEAFs	**CEAFs Layer:** Same as Model 1 **W2V Path:** Wav2Vec 2.0 (frozen) → 1024-dim → Linear → 128 → BN → ReLU → Dropout → Attention Pool → 128-dim **Classifier:** Concat[CEAFs(128), W2V(128)] = 256 → 128 → 128 → 3 (with BN, ReLU, Dropout 0.3)
4. CLINIC (Ours)	**CEAFs Layer:** Same as Model 1 **Mel Path:** Mel-Spectrogram → ResNet-50 → 2048-dim → Linear → 749 → BN → Dropout **W2V Path:** Wav2Vec2 → Cross-attention with 749-dim ResNet feature → Linear 1024 → 768 → BN → Dropout → Concat[ResNet(749), W2V(768)] = 1517 → 128 → BN → Dropout **Classifier:** Concat[Audio(128), CEAFs(128)] = 256 → 128 → 128 → 3 (with BN, ReLU, Dropout 0.3)

F K-Shot Experiment

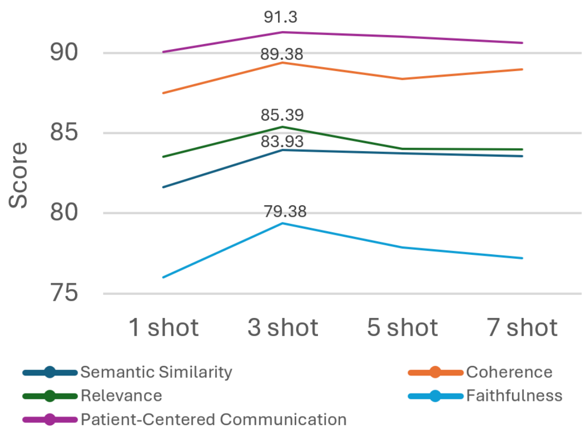


Figure 4: Shot Experiment

Figure 4 presents an ablation study on the number of similar patient cases provided during generation. The 3-shot setting yielded the best overall performance, achieving the highest or near-highest scores across most evaluation metrics. In contrast, the 7-shot setting exhibited a slight decline in performance, indicating that increasing the number of retrieved cases does not necessarily lead to better generation quality. Notably,*Patient-Centered Communication* remained consistently high across all settings, suggesting that the model reliably generates patient-friendly explanations regardless of the number of reference cases. On the other hand, *Faithfulness* showed a downward trend as the number of shots increased, highlighting the need for caution when incorporating a larger number of external cases.

G Medical Explanation Evaluator Metrics

Semantic Equivalence: The semantic overlap between the system output and the reference report written by SLPs is quantified. The evaluation considers not only surface level lexical matches but also whether the patient's condition is captured comprehensively and accurately.
Fidelity: Agreement between the generated explanation and the clinicians' own assessment. Omitting a clinically observed feature deficit, for instance, is scored low.
Consistency: Logical agreement between the predicted severity and the accompanying narrative.
Relevance: Topical adequacy of the text to the task of dysarthric speech analysis. Irrelevant digressions are penalised.
Patient-friendliness: Clarity and accessibility of the explanation for lay readers. Narratives that avoid technical jargon receive higher scores.

H Prompt Structure of GENIE

- **Sys**$_m sg : Thesys_msgservesasthecomponentdefiningthe$ $friendlyKoreanreport.$
-
- **Explanation for Each Feature:** A concise reference text summarises the clinical meaning of each CEAF, allowing the LLM to

ground its narrative in domain-appropriate semantics.

- **Relevant Patient information:** The top-k analogous cases retrieved in the previous stage provide concrete clinical comparators, thereby increasing the specificity and credibility of the generated explanation.

- **Final Output Template:** All outputs adhere to a fixed JSON schema containing the fields Severity, Phonation, Articulation, Prosody, Overall, ensuring consistent formatting across patients.

I Prompt Templates

I.1 Example of GENIE

This is a detailed example of the diagnosis prompt generator section of GENIE.

Example of GENIE Prompt

<s>[INST]«SYS»
Role : Please take on the role of a doctor and explain the information in a way that is clear and reassuring to the patient.

Data Sensitivity and Analysis Guide: I am responsible for analyzing raw patient data to evaluate key factors related to health status. By quantifying various data categories, such as test results, gender, age, and individual feature data, I comprehensively assess the patient's health.

Contextual Understanding and Interpretation Guide: I focus on understanding and evaluating the context of medical data. This approach ensures that I do not evaluate the data on a superficial level but instead gain a deep understanding of the context to accurately interpret the analysis results.

Adaptive and Feedback-Oriented Guide: I continuously improve the method of medical data evaluation over time. By incorporating feedback from various sources, I refine the analysis criteria regularly. For example, I gather feedback from healthcare professionals, patients, and the latest research findings, and use this to continuously modify and improve the data evaluation standards. **«/SYS»**

Instruction : Analyze the given information to describe the characteristics of the patient. Pred final severity refers to the severity level of the patient as finally predicted by the DNN.

SHAP value represents the impact of each feature on the classification of severity (a higher value indicates a greater influence, while a lower value indicates less influence).

For each category, provide explanations focusing on the features that had the greatest impact according to the SHAP values.

Based on the predicted severity, write the patient explanation in Korean, using simple and intuitive words that are easy for general patients to understand. Express severity as a numerical value. Rephrase the explanation using simple, everyday words instead of technical terms. For the articulation section, please describe the patient's performance separately for the syllables <puh>, <tuh>, and <kuh>.

Ensure the output follows the Output Template format in JSON file with four keys: (severity, Phonation, Articulation, Prosody, Overall)

Explanation for Each Feature :

<Patient Information>

- speaker : "name of the speaker"
- severity : "severity of dysarthria of the patient (0 - simmilar to normal person, 1 - mild, 2 - severe)"
- age : "age of the patient"
- gender : "gender of the patient (0 - male, 1 - female)"

<ddk low-level features>

- intelligibility : "How clearly a person speaks so that speech is comprehensible to a listener"
- var F0 semitones : "Variance of the fundamental frequency in semitones"
- var F0 Hz : "Variance of the fundamental frequency in Hz"
- avg Energy : "Average of vocal energy"

- var Energy : "Standard deviation of vocal energy"
- max Energy : "Maximum value of vocal energy"
- ddk rate : "The number of syllables pronounced per second"
- ddk average : "Average time of each syllables pronounced"
- ddk std : "Standard deviation of the time of each syllables pronounced"
- ddk pause rate : "The number of pause per second"
- ddk pause average : "Average time of each pause"
- ddk pause std : "Standard deviation of the time of each pause"
- task : 2 - repeating "puh", 3 - repeating "tuh", 4 - repeating "kuh", 5 - repeating "puh tuh kuh"

Severity:

0: Normal

1: Mild to moderate

2: Severe

<Phonation>

This refers to how strong and stable the voice sounds when speaking. For example, if the voice is too weak, shaky, or sounds breathy, it may indicate a problem with phonation.

<Articulation>

This describes how accurately the lips, tongue, and jaw move to form speech sounds. Imprecise articulation can cause speech to sound slurred or unclear.

<Prosody>

This includes the rhythm, pitch, and speed of speech, which help convey emotion and naturalness. When prosody is impaired, speech may sound flat, monotone, or emotionally unexpressive.

Reference Data(information of other patients) :

1 reference data :

Severity : 1 , speaker: nia HS0027 severity: 1, gender: 1 task id: 2, intelligibility: 4, var f0 semitones: 73.433, var f0 hz: 29.183, avg energy: 69.307, var energy: 26.122, max energy: 80.175, ddk rate: 2.092, ddk average: 230.313, ddk std: 43.954, ddk pause rate: 0.131, ddk pause average: 216.875, ddk pause std: 437.475

task id: 3, intelligibility: 4, var f0 semitones: 72.326, var f0 hz: 28.265, ...

task id: 4, intelligibility: 4, var f0 semitones: 93.234, var f0 hz: 48.798, ...

task id: 5, intelligibility: 4, var f0 semitones: 83.135, var f0 hz: 38.258, ...

2 reference data :

Severity: 1, speaker: nia HS0159, severity: 1, gender: 1, ddk feature info: ...

3 reference data :

Severity: 1, speaker: nia HS0109, severity: 1, gender: 0, ddk feature info: ...

Input Data :

speaker: nia HS0079, data info: gender: 1,

ddk feature info:

task id: 2, gender: 1, intelligibility: 4, var f0 semitones: 44.034, var f0 hz: 13.227, avg energy: 62.455, var energy: 24.292, max energy: 73.25, ddk rate: 1.213, ddk average: 287.5, ddk std: 110.701, ddk pause rate: 0.152, ddk pause average: 396.563, ddk pause std: 748.029

task id: 3, gender: 1, intelligibility: 4, var f0 semitones: 88.085, var f0 hz: 43.288...

task id: 4, gender: 1, intelligibility: 4, var f0 semitones: 62.411, var f0 hz: 21.849...

task id: 5, gender: 1, intelligibility: 4, var f0 semitones: 97.624, var f0 hz: ...

SHAP Value : 'id': 'nia HS0079',

shap class: 0, gender: 0.571, intelligibility: 0.471, var f0 semitones: 0.457, var f0 hz: 0.505, avg energy: 0.436, var energy: 0.631, max energy: 0.42, ddk rate: 0.544, ddk average: 0.581, ddk std: 0.515, ddk pause rate: 0.508, ddk pause average: 0.600, ddk pause std: 0.591

shap class: 1, gender: 0.793, intelligibility: 0.618, var f0 semitones: 0.499, ...

shap class: 2, gender: 0.182, intelligibility: 0.332, var f0 semitones: 0.382, var f0 hz: 0.437, ...

Each Task Pred Severity : ['task': 2, 'ddk pred severity': 1, 'task': 3, 'ddk pred severity': 1, 'task': 4, 'ddk pred severity': 1, 'task': 5, 'ddk pred severity': 1]

Final Pred Severity : 1

Output Template :

Severity :

Phonation :

Articulation :

Prosody :

Overall :

I.2 Example of Medical Explanation Evaluator

This is a detailed example of the Medical Explanation Evaluator prompt.

Example of Medical Explanation Evaluator Prompt

<s>[INST]«SYS»
Role : Please take on the role of a doctor and explain the information in a way that is clear and reassuring to the patient.

Data Sensitivity and Analysis Guide: I am responsible for analyzing raw patient data to evaluate key factors related to health status. By quantifying various data categories, such as test results, gender, age, and individual feature data, I comprehensively assess the patient's health.

Contextual Understanding and Interpretation Guide: I focus on understanding and evaluating the context of medical data. This approach ensures that I do not evaluate the data on a superficial level but instead gain a deep understanding of the context to accurately interpret the analysis results.

Adaptive and Feedback-Oriented Guide: I continuously improve the method of medical data evaluation over time. By incorporating feedback from various sources, I refine the analysis criteria regularly. For example, I gather feedback from healthcare professionals, patients, and the latest research findings, and use this to continuously modify and improve the data evaluation standards.

Instruction :

Evaluation: Provide a score (1-100) for each criterion, followed by a brief explanation of why you assigned that score. Please evaluate whether each feature has been accurately extracted.

Explanation for Each Feature :

<Patient Information>

- speaker : "name of the speaker"
- severity : "severity of dysarthria of the patient (0 - simmilar to normal person, 1 - mild, 2 - severe)"
- age : "age of the patient"
- gender : "gender of the patient (0 - male, 1 - female)"

<ddk low-level features>

- intelligibility : "How clearly a person speaks so that speech is comprehensible to a listener"
- var F0 semitones : "Variance of the fundamental frequency in semitones"
- var F0 Hz : "Variance of the fundamental frequency in Hz"
- avg Energy : "Average of vocal energy"
- var Energy : "Standard deviation of vocal energy"
- max Energy : "Maximum value of vocal energy"
- ddk rate : "The number of syllables pronounced per second"
- ddk average : "Average time of each syllables pronounced"
- ddk std : "Standard deviation of the time of each syllables pronounced"
- ddk pause rate : "The number of pause per second"
- ddk pause average : "Average time of each pause"
- ddk pause std : "Standard deviation of the time of each pause"

- task : 2 - repeating "puh", 3 - repeating "tuh", 4 - repeating "kuh", 5 - repeating "puh tuh kuh"

Severity:

0: Normal

1: Mild to moderate

2: Severe

<Phonation>

This refers to how strong and stable the voice sounds when speaking. For example, if the voice is too weak, shaky, or sounds breathy, it may indicate a problem with phonation.

<Articulation>

This describes how accurately the lips, tongue, and jaw move to form speech sounds. Imprecise articulation can cause speech to sound slurred or unclear.

<Prosody>

This includes the rhythm, pitch, and speed of speech, which help convey emotion and naturalness. When prosody is impaired, speech may sound flat, monotone, or emotionally unexpressive.

Evalution:

Evaluate the generated response using the following criteria:

Semantic Similarity (1-100): Does the generated response convey the same meaning as the reference answer, even if the wording differs? Focus on whether the key ideas and intent are preserved. Please evaluate whether the severity level determined by the clinical pathologist matches the severity level predicted by the model. If the severity levels differ between the reference data and the generated data, assign a lower score.

Coherence (1-100): Evaluate whether the response is well-structured and logically organized. Check if it reads naturally without grammatical or syntactical errors. Evaluate whether an appropriate explanation has been generated based on the corresponding severity level.

Relevance (1-100): Evaluate whether the response stays focused on the topic and directly addresses the prompt. Confirm that patient-related features are appropriately explained without including unnecessary information.

Faithfulness (1-100): Exaggerations or inaccuracies regarding these features are grounds for point deductions. Please evaluate whether the severity level determined by the clinical pathologist matches the severity level predicted by the model. In addition, if key features mentioned in the reference data are missing from the generated explanation, a penalty should be applied

Patient-friendliness (1-100): Evaluate whether the response minimizes the use of technical jargon and explains things in simple, easy-to-understand terms. Also, check if any difficult medical terminology is used without explanation.

Evaluation Form (scores ONLY):

- Semantic Similarity :
- Coherence:
- Relevance:
- Faithfulness:
- Patient-Centered Communication:

Please derive it as a json file according to the output format

Reference Answer :

Severity : 1

Phonation : The voice gradually becomes quieter, accompanied by noticeable tremors.

Articulation : The movements of the tongue, lips, and jaw are slow, and the syllables /puh/, /tuh/, and /kuh/ were all pronounced slowly. Consonants were pronounced weakly, and there was difficulty in producing rapid transitions between sounds.

Prosody : The speech rate was consistent, and no significant issues with prosody were observed.

Overall : Due to short phonation, the voice volume decreased during speech, and vocal tremors were observed. Consonants were pronounced weakly. Therefore, vocal training and speech therapy may be necessary to achieve more stable phonation.

Generated Response :

Severity : 1

Phonation : Your voice is generally stable, but there may be occasional slight tremors. This may indicate a minor issue with phonation.

Articulation : The movements of the lips, tongue, and jaw are mostly accurate, but pronunciation may occasionally sound unclear. This may suggest slight difficulties with articulation.

Prosody : The rhythm and speed of your speech are generally natural.

Overall : Overall, your speaking ability may have slight difficulties, but is generally well maintained. This is considered a mild issue and is unlikely to significantly interfere with daily life.

J Reference Data

The following is the reference data written by SLPs for each severity level. It was actually written in Korean.

Reference data for patients with severity level 0

Severity : 0

Phonation : The voice quality, pitch, and volume are all normal with no particular abnormalities.

Articulation : The speech organs such as the tongue, lips, and jaw moved at a normal speed for the sounds /p/, /t/, /k/, and articulation in connected speech tasks was at an average level.

Prosody : The speech rate was regular, and no particular issues were observed in prosody.

Overall : There appear to be no special problems with speech overall. The voice and articulation are appropriate for the age group.

Reference data for patients with severity level 1

Severity : 1

Phonation : The voice gradually becomes quieter, accompanied by noticeable tremors.

Articulation : The movements of the tongue, lips, and jaw are slow, and the syllables /puh/, /tuh/, and /kuh/ were all pronounced slowly. Consonants were pronounced weakly, and there was difficulty in producing rapid transitions between sounds.

Prosody : The speech rate was consistent, and no significant issues with prosody were observed.

Overall : Due to short phonation, the voice volume decreased during speech, and vocal tremors were observed. Consonants were pronounced weakly. Therefore, vocal training and speech therapy may be necessary to achieve more stable phonation.

Generated text for patients with severity level 2

Severity : 2

Phonation : Overall, the articulatory organs moved slowly and with limited motion during speech.

Articulation : It was difficult to move the articulatory organs quickly to produce distinct sounds accurately, resulting in distorted or substituted pronunciation

Prosody : There were many prolonged pauses during speech, causing slow speech rate and disrupted flow.

Overall : Low speech intelligibility is affected by articulation, phonation, and prosody. Due to the muscles' lack of agility and accuracy, the speech flow is disrupted, leading to reduced clarity.

K Result of GENIE

These are examples of texts generated by GENIE for each severity level. It was actually written in Korean.

Generated text for patients with severity level 0

Severity : 0

Phonation : The patient's voice is strong and steady. The voice is neither weak, trembling, nor breathy.

Articulation : The patient's movements of the lips, tongue, and jaw are precise, resulting in clear and distinct pronunciation. Sounds like <puh>, <tuh>, and <kuh> were pronounced accurately.

Prosody :The patient's speaking rate, rhythm, and pitch are natural, with good emotional expression.

Overall : The patient's speaking ability is within the normal range. Pronunciation, voice, and speech rhythm are all natural and clear. There is no need for concern.

Generated text for patients with severity level 1

Severity : 1

Phonation :Your phonation is generally stable but may have slight variations. Your voice might be a bit weak or shaky, but it is not a major concern.

Articulation : Movements of the lips, tongue, and jaw are mostly accurate, though sometimes your pronunciation may sound unclear. You might experience some difficulty repeating the sounds 'puh', 'tuh', and 'kuh'.

Prosody : The rhythm and speed of your speech are mostly natural, but it may occasionally sound monotonous. Emotional expression might be somewhat limited.

Overall : Overall, your speaking ability shows mild issues. It should not significantly affect everyday conversations. Paying a bit more attention to pronunciation and speaking slowly can make your speech clearer.

Reference data for patients with severity level 2

Severity : 2

Phonation :The patient's voice may sound somewhat weak and unstable, indicating possible phonation issues.

Articulation : The movements of the lips, tongue, and jaw are inaccurate, which can cause unclear pronunciation. There may be difficulty producing the sounds 'puh', 'tuh', and 'kuh'.

Prosody : The patient's speech may sound monotonous and lack emotional expression, suggesting problems with rhythm and speech rate.

Overall : The patient's overall speaking ability is assessed as severe. This means difficulties in phonation, articulation, and prosody. However, do not worry. Improvement is possible with appropriate therapy and practice.