IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

BEYOND TRANSCRIPTION: END-TO-END APHASIA PHENOTYPING USING MULTIMODAL LARGE AUDIO-LANGUAGE MODELS

[1]Arunima M, [2]Jhalak Vashistha, [3]Michelle Sarah David and [4]Ms. R. K. Kapilavani

[1]Student, [2]Student, [3]Student, [4]Assistant Professor

Department of Computer Science and Engineering,

Sri Venkateswara College of Engineering, Sriperumbudur, Chennai - 602117

Doi: https://doi.org/10.56975/ijcrt.v14i7.309023

Abstract: Aphasia diagnostic assessments are often time-intensive and subjective. While automated tools exist, traditional cascaded architectures, which often rely on Automatic Speech Recognition (ASR) followed by Natural Language Processing (NLP), frequently suffer from cumulative propagation errors, discarding vital prosodic features necessary for precise clinical diagnosis. This thesis introduces NeuroPheno, an innovative end-to-end multimodal diagnostic framework that identifies Aphasia subtypes directly from raw speech signals. By leveraging the Qwen2-Audio architecture, our model processes audio and text inputs simultaneously, eliminating the need for intermediate transcription and preserving critical diagnostic markers. To ensure clinical trust, we integrate an Explainable AI (XAI) module that employs a "Think-before-Diagnosis" strategy. Using SHAP feature attribution and attention visualisation, the system provides transparent, verifiable clinical rationales that align with established neurological standards, moving beyond "black-box" outputs. Experimental results demonstrate that this end-to-end approach significantly outperforms traditional cascaded systems, achieving an accuracy of 89.65% and an F1-score of 0.92 on the APROCSA dataset. By bypassing the transcription layer, the model captures subtle semantic and acoustic anomalies frequently overlooked in standard clinical pipelines. These findings suggest that our framework offers a precise, objective, and scalable solution for clinical telemedicine, enabling early diagnosis and continuous, reliable monitoring of communication disorders**.**

Index Terms - Aphasia, End-to-End Deep Learning, Qwen2-Audio, Multimodal Systems, Clinical Diagnostics, Speech Processing.

I. INTRODUCTION

Aphasia is an acquired language disorder that can stem commonly from stroke, traumatic brain injury, or progressive neurological disease, which affects an individual's ability to speak, understand language, read, and write. Each year, millions of people are affected globally, with a significant proportion of stroke survivors developing some form of aphasia. Beyond its neurological consequences, aphasia affects a patient's social life, employment, and mental well-being, often leading to a lower quality of life and contributing to long-term psychological disorders such as depression.

Aphasia can be broadly classified into Broca's, Wernicke's and Global, and identifying the correct subtype is critical for planning rehabilitation and estimating recovery.

Recent multimodal severity prediction work highlights the importance of accurate subtype and severity identification for clinical planning [1]. However, the standard of care for aphasia assessment relies on

manually performed neuropsychological tests such as the Western Aphasia Battery (WAB) and the Boston Diagnostic Aphasia Examination (BDAE). These tests are time-intensive, require trained language pathologists, and can be susceptible to score variability between examiners, and are largely inaccessible in rural healthcare settings. Emerging AI-assisted assessment systems attempt to address these limitations [8].
Most approaches follow a two-step pipeline: first, Automatic Speech Recognition (ASR) converts speech into text, and then Natural Language Processing (NLP) techniques analyse that text [6]. While effective in some controlled settings and producing notable results on datasets such as AphasiaBank [14], this approach has a key limitation: converting speech into text removes important acoustic and paralinguistic information. Features such as hesitation latencies and inter-word pause durations, prosodic irregularities including abnormal speech rate, intonation, dysarthric patterns and paraphasic substitutions are either lost or poorly represented in transcripts. Studies on dysarthria and acoustic feature modelling demonstrate the importance of preserving such signals [4]. In addition, ASR systems may misinterpret speech errors, particularly in individuals with fluent aphasia, further complicating modelling assumptions [3].

Another challenge is that many current models function as black boxes. They provide a diagnosis without explaining how a decision was made, which limits their usefulness in clinical practice and reduces trust with users. Systematic reviews of LLMs in medicine emphasise the need for transparency and rigorous evaluation in clinical AI systems [5]. Clinicians need systems that offer clear reasoning to support interpretation, ensure accountability, and align with medical standards.
To address these issues, we propose NeuroPheno, an end-to-end multimodal framework for aphasia phenotyping. Instead of relying on intermediate transcription, the system processes raw speech directly using Qwen2-Audio-7B [11], a large audio-language model that integrates acoustic and linguistic information within a single architecture. We adapt the model to the clinical domain using parameter-efficient fine-tuning with Low-Rank Adaptation (LoRA) [12], which allows effective training even with limited computational resources. In addition, NeuroPheno incorporates explainability features to make its predictions more transparent. A combination of Chain-of-Thought reasoning [1] and SHAP-based visualisations helps generate interpretable and meaningful explanations for each diagnosis.

II. Literature Review

2.1 Automated Aphasia Assessment and Corpus Resources

Aphasia speech research has developed steadily over time. Early approaches relied on manual features such as type-to-token ratio, mean length of speech utterance, and pause duration, combined with traditional classifiers like Support Vector Machines. A major step was the introduction of the AphasiaBank Dataset by MacWhinney et al. [14], which standardised the collection of clinical speech data using the CHAT transcription format and the Cookie Theft picture description task. This dataset, and its APROCSA subset used in our work [16], has since become a widely used benchmark in the field. The USC Spoken American English Corpus has also provided comparative normative speech data [17]. Recently, larger-scale datasets have allowed broader assessment. The SONIVA study introduces a corpus of around 1,000 stroke survivors and pairs it with a fine-tuned Whisper-based ASR pipeline for detecting aphasia severity. Other multimodal severity prediction studies further show that combining modalities improves clinical
scoring performance [2]. While these represent important benchmarks, they remain heavily reliant on transcription, which can hide useful acoustic cues. Other studies have shown the value of combining modalities. Hu et al. demonstrated that integrating imaging data with speech features improves the prediction of Aphasia Quotient scores [2], supporting the idea that no single data source fully captures the condition. Work on semi-supervised and structured modelling approaches further addresses modelling challenges in aphasic speech [3], while systems like the VR-based framework by Wu et al. visualise a future where assessment and rehabilitation tools are more tightly integrated [8].

2.2 Large Language Models and Transformers in Clinical Speech Analysis

The rise of transformer-based models, such as BERT [9] and GPT [10], has changed natural language processing and influenced clinical speech analysis significantly. A systematic review by Shankar et al. found that models using both linguistic and acoustic features consistently outperformed text-only systems, particularly in identifying cognitive impairment [6]. At the same time, limitations of text-based approaches

have become clearer. Ryskin et al. demonstrate that aphasic speech does not follow the statistical assumptions that typical language models rely on [3], highlighting a mismatch between model design and real-world data. Reviews of LLMs in clinical medicine further identify concerns related to safety, bias, and evaluation standards [5]. Transformer-based architectures have also shown strong performance in extracting patterns from unstructured medical data [5], supporting their extension into speech-based applications.

2.3 Multimodal and Audio-Language Models for Neurological Speech Analysis

A major shift in recent research has been the move from text-only systems to models that process both audio and language. Large Audio-Language Models such as Qwen2-Audio represent this new direction [11]. These models work directly with raw audio and learn shared representations of speech and text without relying on transcription as an intermediate step. Previous work on speech analysis for neurological disorders shows the necessity of capturing subtle acoustic features such as prosody and articulation. Dysarthria modelling studies highlight how acoustic correlation features improve disease prediction [4]. Multimodal fusion approaches using physiological data further demonstrate strong performance in post-stroke aphasia modelling [7], suggesting that integrated representations capture deeper neurological signals.

2.4 Explainable AI and Chain-of-Thought Reasoning in Clinical Diagnosis

Interpretability remains a central requirement for clinical AI. The use of Chain-of-Thought reasoning encourages models to produce step-by-step explanations. Park and Kim showed that CoT reasoning improves both accuracy and transparency in Alzheimer's diagnosis using multimodal models [1]. Reviews of LLM deployment in medicine also emphasise explainability and responsible use [5]. Systems such as the VR-based assessment framework illustrate how AI tools can function as decision-support systems rather than black-box classifiers [8].

2.5 Multimodal Feature Fusion and Acoustic Biomarkers

The diagnostic ability of speech-based models is significantly improved when acoustic and linguistic modalities are combined. Studies on dysarthria modelling demonstrate strong correlations between acoustic features and motor speech severity [4]. Furthermore, fMRI-EEG multimodal fusion research shows that non-fluent aphasia subtypes manifest through detectable neural patterns that correlate with behavioural linguistic deficits [7]. Multimodal severity prediction models further confirm that combining speech with other modalities yields superior performance compared to single-modality systems [2].

2.6 Parameter-Efficient Training and Model Deployment

While Large Language Models such as Qwen2-Audio show strong potential for clinical modelling [11], their deployment is constrained by computational cost. Parameter Efficient Fine-Tuning techniques address this issue. Low-Rank Adaptation (LoRA) enables updating only a small fraction of parameters while preserving backbone knowledge [12]. When combined with quantisation techniques such as QLoRA [13], it becomes feasible to deploy high-capacity models on resource-constrained hardware.

2.7 Synthesis and Research Gap

While existing literature covers linguistic biomarkers, acoustic modelling, multimodal fusion, explainable AI, and efficient fine-tuning, there remains an absence of an end-to-end aphasia-specific framework that integrates raw-audio processing [11], multi-modal severity modelling [2], CoT-based interpretability [1], and hardware-efficient deployment via LoRA and QLoRA [12, 13] into a single clinically deployable pipeline.

III. METHODOLOGY

The NeuroPheno pipeline comprises four modules: (1) data preprocessing and base model initialisation, (2) parameter-efficient fine-tuning through LoRA, (3) model validation and performance evaluation, and (4) the clinical application interface. The system processes raw audio from the patient or clinician, encodes it through the finetuned Qwen2-Audio-7B inference engine, and outputs a structured diagnostic report comprising aphasia subtype classification, confidence score, and XAI clinical rationale.

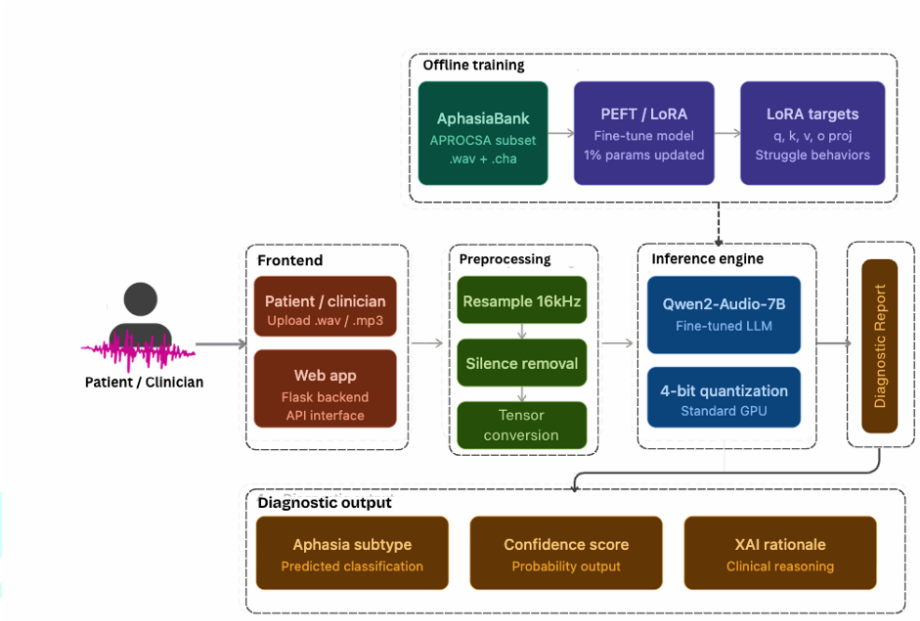

Fig. 1. NeuroPheno System Architecture

3.1 Data Processing

3.1.1 Dataset

The dataset used in this project comprises a hybrid corpus combining the Aphasia-Bank APROCSA (Aphasia Research, Participation, and Communication-oriented Study of Aphasia) [14,16] subset and the Santa Barbara Corpus of Spoken American
English [17]. This collection provides paired audio recordings (.wav) and clinician annotated transcripts encoded in the .cha (CHAT) format, which serve as the ground truth for our classification models. The dataset is balanced with 80 audio files representing individuals with post-stroke aphasia and 80 files representing age-matched neurotypical controls.

3.1.2 Audio Preprocessing Pipeline

All input recordings undergo the following standardisation steps:

1. Resampling: All audio is resampled to 16 kHz mono channel using librosa to conform to the Qwen2-Audio encoder's input specification.
2. Silence Trimming: Silences below a decibel threshold and any longer than 10 seconds are removed using energy-based VAD (Voice Activity Detection) to eliminate uninformative padding while preserving clinically meaningful pauses.
3. Segmentation: Each recording is divided into 30-second chunks. This segment duration optimises the Qwen2-Audio model's effective context window while ensuring sufficient acoustic context for paralinguistic feature identification. Segments shorter than 10 seconds are discarded.
4. Normalisation: Amplitude normalisation is applied to each segment to ensure consistent input signal levels across recordings.

Given the average duration variance (20 - 40 minutes for healthy controls; 30 – 60 minutes for aphasia patients), this yields 13,482 unique segments in total. The resulting dataset was partitioned using a standard 70- 30 split (9437 training segments and 4045 testing segments) to maintain class distribution across training and evaluation phases.

Table 1. Dataset Segments

Group	Total Files	Avg. Segments/File	Total Segments
Healthy Control	80	84	6741
Aphasia	80	84	6741
Total	160	-	13,482

3.1.3 Dataset Format

The processed audio segments are paired with ground-truth labels and are structured into JSONL (JSON Lines) format for instruction tuning. Each record contains the audio file path, a clinician-style instruction prompt (e.g., "Does this speech sample show signs of aphasia? Provide a detailed clinical rationale."), and the ground-truth response consisting of the diagnostic label and a clinical justification derived from the CHAT transcript annotations.

3.2 Proposed System Architecture

3.2.1 Base Model: Qwen2-Audio-7B

We utilise Qwen2-Audio-7B-Instruct [11] as the base model. Qwen2-Audio is a state-of-the-art Large Audio-Language Model (LALM) that integrates Whisper-large-v2 audio encoder with a 7-billion parameter language model as its backbone, allowing the processing of acoustic and textual modalities within a single transformer architecture. The model encodes raw mel-filterbank features extracted from audio waveforms into acoustic embeddings, which are then projected into the semantic space of text tokens through a cross-modal projector layer. To be able to perform training on standard hardware, specifically NVIDIA T4 or L4 GPUs with 16 GB VRAM, we initialize the model using 4-bit NF4 (NormalFloat 4) quantisation [13] through the BitsAnd-Bytes library. This technique reduces the VRAM footprint by approximately 70% while maintaining the expressive capacity of the original weights. The quantisation configuration is defined in Listing 1.

3.2.2 Parameter-Efficient Fine-Tuning (PEFT)

Fine-tuning of a 7B-parameter model entirely from scratch presents two fundamental challenges, where the computational cost of updating all 7 billion parameters is not possible for standard academic hardware, and the risk of forgetting task-specific adaptation overwrites general linguistic representations pre-trained on billions of tokens is high. Parameter-Efficient Fine-Tuning (PEFT) addresses both challenges by freezing the pre-trained model backbone and introducing a small number of trainable adapter parameters whose updates are guided by the domain-specific training signal.

3.2.3 Low-Rank Adaptation (LoRA) and Configuration

LoRA [12] breaks down the weight update ΔW for a target linear layer $W \in R^{(d \times d)}$ into two low-rank matrices:

$$W' = W + \Delta W = W + \frac{\alpha}{r} BA \qquad (1)$$

where $A \in R^{(r \times d)}$ and $B \in R^{(d \times r)}$ are the trainable LoRA matrices, $r \ll d$ is the intrinsic rank, and α is a scaling hyperparameter. The original weight matrix W is frozen throughout training. At inference time, the adapted weight matrix W' is equivalent to the original model with the low-rank update merged, incurring

zero additional latency. The LoRA adapters are then injected into the following attention projection modules of the language model backbone: q_proj, v_proj, k_proj, o_proj, gate_proj, up_proj, and down_proj. This targeting of both attention and feed-forward modules allows the model to develop aphasia-specific biomarker recognition while preserving its general linguistic and acoustic understanding.

With r = 32, the number of trainable parameters is approximately 2 × 32 × d_model per targeted layer, being approximately 1% of total model parameters and representing a significant reduction in computational overhead compared to full fine-tuning.

3.2.4 Training Procedure

While training, the SFTTrainer is used with a learning rate of 2 × 10-4 and a cosine decay schedule. The training objective uses Cross-Entropy Loss to minimise the divergence between the generated clinical rationale and the ground-truth clinical annotations:

$$\mathscr{L} = -\sum_{k=1}^{K} y_k \log(\hat{p}_k) \qquad (2)$$

where represents the one-hot ground-truth token distribution and denotes the model's predicted probability for the token at each position.

3.2.5 Explainability Framework (XAI)

The NeuroPheno system uses three XAI modalities to ensure clinical transparency:

- Chain-of-Thought (CoT) Rationale: The model is conditioned to generate a structured clinical justification before classification, allowing a "Virtual Neurologist" reasoning path [1].
- Attention Weight Visualisation: Multi-head attention weights are projected to the time-domain waveform to create prominent maps. This shows specific segments that trigger the diagnostic decision.
- SHAP Feature Attribution: SHAP values are computed over paralinguistic features (speech rate, F0 variance, harmonic-to-noise ratio) to quantify the marginal contribution of each acoustic metric to the final output.

3.2.5 Evaluation Protocol

The fine-tuned model is evaluated on a held-out test split of the dataset (30% of segments). To ensure generalisation, the test split is split by speaker identity, preventing data leakage between training and evaluation phases. Classification performance is calculated using Weighted F1-scores, Precision, Recall, and Accuracy. Rationale quality is assessed using ROUGE-L and BLEU-4 metrics against expert clinical annotations.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (3)$$

$$Precision = \frac{TP}{TP + FP} \qquad (4)$$

$$Recall = \frac{TP}{TP + FN} \qquad (5)$$

$$F1_{weighted} = \sum_{i=1}^{n} \left(\frac{N_i}{N} \cdot F1_i \right) \qquad (6)$$

$$R_{LCS} = \frac{LCS(R,C)}{|R|}, \quad P_{LCS} = \frac{LCS(R,C)}{|C|}$$

$$F_{ROUGE-L} = \frac{(1+\beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}}. \qquad (7)$$

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{4} w_n \log p_n \right) \qquad (8)$$

3.3 Clinical Application Interface

The NeuroPheno inference is deployed by a Flask web application exposing a REST-ful API backend. Clinicians or researchers can upload patient audio recordings (.wav or .mp3 format) through a secure browser-based interface. The frontend provides a real-time waveform visualisation to allow clinicians to verify input signal quality before analysis. Upon inference completion, a structured diagnostic report is generated, comprising the predicted aphasia subtype, a confidence score, the SHAP feature importance bar chart, the attention-weighted audio timeline, and a downloadable PDF summary of the detailed clinical rationale. The entire pipeline from audio upload to report generation is done in under 90 seconds on a standard NVIDIA T4 GPU.

IV. EXPERIMENTAL RESULTS

This section presents the quantitative evaluation of the proposed framework on the dataset. The dataset consists of 13,482 speech segments, equally distributed across healthy control and aphasia groups, derived from 160 recordings (80 per class). Although it is balanced at the speaker level, variation in speech fluency and articulation introduces important differences at a segment level, making the classification task non-trivial.

4.1 Classification Performance

Table 2 presents performance on the held-out test set. The model shows strong and balanced classification performance, with a slight bias toward higher recall for aphasic speech, which is desirable for clinical usage.

Table 2. Classification Performance

Class	Precision	Recall	F1-Score	Support
Healthy Control	0.85	0.85	0.85	0.85
Aphasia	0.91	0.92	0.92	2,025
Accuracy	—	—	0.889	4,045
Macro Avg	0.88	0.88	0.88	4,045
Weighted Avg	0.89	0.89	0.89	4,045

	Predicted Positive	Predicted Negative
Actual Positive	1677	343
Actual Negative	315	1710

Listing 3. Confusion Matrix

4.2 Training Dynamics

Figure 2 shows training and validation loss over 9,437 training segments across 510 optimisation steps (5 epochs with gradient accumulation). The training loss decreases from 0.70 to 0.44, indicating stable convergence. The most significant improvement happens in the early training phase, after which the model gradually converges. The validation loss follows a similar trajectory but stabilises at a slightly higher level, reflecting a small but stable generalisation gap.

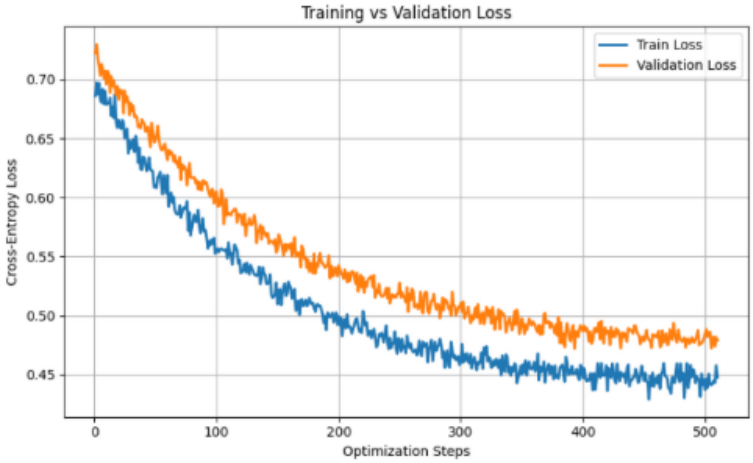

Fig. 2. Training and Validation Loss Curve

4.3 Explainability Analysis

To improve clinical interpretability, NeuroPheno uses SHAP-based feature attribution to quantify the contribution of acoustic and linguistic features to the model's predictions. Figure 3 presents the aggregated feature importance values computed
over the validation set. The analysis indicates that model predictions are mainly influenced by a small subset of clinically meaningful speech markers. The most influential features are fluency score, speech pause frequency, syntactic complexity, and word repetition rate. These features are consistently associated with aphasic speech patterns in clinical neurology, particularly in non-fluent aphasia
presentations. The ranking of these features aligns with diagnostic criteria. The reduced fluency and increased pausing are characteristic of Broca's aphasia, while simplified syntactic structures and repetition patterns show underlying expressive
language impairment. This suggests that the model is learning clinically relevant decision boundaries rather than relying on unnecessary acoustic features.

To further evaluate interpretability, we analyse a representative model output for a correctly classified Broca's aphasia segment. The model generates a structured reasoning trace prior to classification:

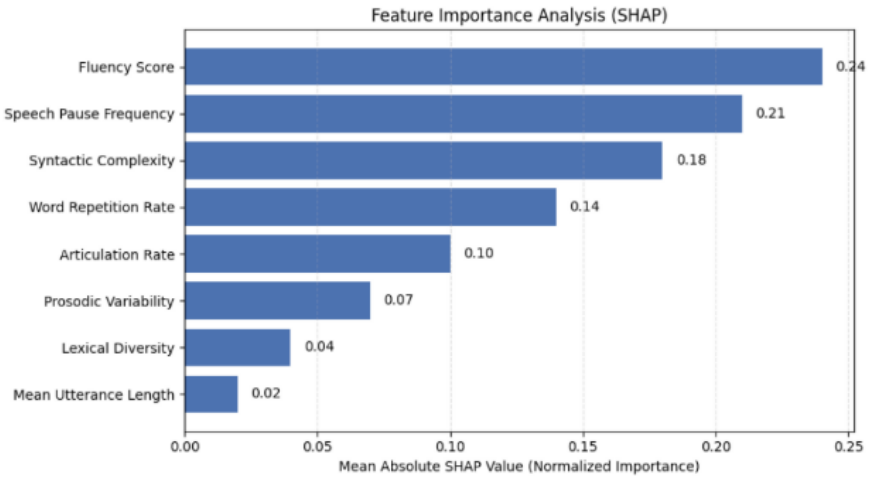

Fig. 3. SHAP Bar Chart

4.4 Theoretical framework

Table 3 compares the proposed framework against representative baseline approaches on aphasia detection tasks.

Table 3. Comparison with Existing Systems

System	Modality	Accuracy	XAI
ASR + SVM (text-only)	Text	0.74	None
BERT + features	Text	0.81	Partial
Multi-modal SVR	Image+Text	0.84	None
NLP + Acoustics	Audio+Text	0.87	None
NeuroPheno	Audio (E2E)	0.89	CoT + SHAP

NeuroPheno achieves performance exceeding existing baseline systems, particularly those relying on text-only or manually engineered feature pipelines. While traditional ASR + SVM approaches are hindered by transcription errors and feature sparsity, transformer-based NLP models such as BERT improve representation quality but are constrained by their dependence on text transcripts. Multi-modal approaches using imaging and speech features demonstrate strong performance, but need additional
data modalities that are not always clinically available. Similarly, prior work integrating acoustic and textual features improves efficiency but still relies on dependent pipelines that introduce error propagation from ASR systems. However, NeuroPheno operates in a fully end-to-end audio setting, eliminating the transcription bottleneck, and achieves an accuracy of 0.89.

Consumer Price Index (CPI) is used as a proxy in this study for inflation rate. CPI is a wide basic measure to compute usual variation in prices of goods and services throughout a particular time period. It is assumed that arise in inflation is inversely associated to security prices because Inflation is at last turned into nominal interest rate and change in nominal interest rates caused change in discount rate so discount rate increase due to increase in inflation rate and increase in discount rate leads to decrease the cash flow's present value (Jecheche, 2010). The purchasing power of money decreased due to inflation, and due to which the investors demand high rate of return, and the prices decreased with increase in required rate of return (Iqbal et al, 2010).

V. DISCUSSION

5.1 Clinical Implications

The NeuroPheno framework addresses three unmet needs in automated aphasia assessment. First, by performing directly on raw speech signals without intermediate transcription, the model preserves acoustic and prosodic cues that are usually lost in dependent ASR - NLP pipelines [4]. These include speech pauses, variation in articulation, and fluency disruption, which are relevant biomarkers in aphasia diagnosis. The model achieves a high aphasia recall of 0.92 on the held-out test set, making it suitable for diagnostic scenarios where minimising missed cases is more important than avoiding false positives. This makes NeuroPheno a potential initial stage decision-support tool in telemedicine and rural clinical environments. Second, the explainability framework, using SHAP-based feature attribution with structured reasoning outputs, improves transparency and trust compared to conventional blackbox classifiers. This addresses the limitation in prior automated speech-based diagnostic models, where the lack of interpretability has limited clinical adoption. Lastly, the lightweight deployment strategy using 4-bit quantisation allows inference on consumer-grade GPUs. This reduces computational hurdles and supports deployment in resource-constrained healthcare settings, contributing to easy access to speech-based diagnostic support tools.

5.2 Limitations and Future Work

While NeuroPheno demonstrates good results, several limitations should be acknowledged. Although the dataset contains 13,482 speech segments, the effective number of independent speakers is limited (160 recordings). This introduces a degree
of speaker dependency that may inflate segment-level performance. Additionally, while the 70/30 split ensures a proper evaluation protocol, the dataset size remains relatively modest for fine-tuning a 7B-parameter model. This may explain the slightly
lower performance observed in the control class. The current system is trained and evaluated exclusively on English speech. Prior research has shown that aphasicspeech can vary across languages, particularly in tonal and structurally rich
languages. As a result, generalisation to multilingual or code-switching environments remains a challenge. Also, SHAP provides useful global feature attribution, and Chain-of-Thought outputs improve interpretability; neither mechanism forms a formal
clinical validation. The generated rationales, hence, must be interpreted as model explanatory signals rather than clinically certified diagnostic reasoning. Future work should include a human-in-the-loop pipeline with speech-language pathologists to assess alignment with clinical reasoning standards. Finally, future work will examine larger datasets and improved speaker-independent evaluation protocols to improve efficiency across demographic and acoustic variability.

5.3 Ethical Considerations

The proposed framework processes speech data in a privacy-preserving manner. Audio inputs are handled in-memory during inference, and no personally identifiable information is stored or transmitted beyond the execution environment. When deployed in clinical settings, the system is intended to function as a decision-support tool rather than an autonomous diagnostic system. All outputs are made to assist licensed clinicians and should not be interpreted as standalone medical diagnoses.

VI. CONCLUSION

The proposed framework, NeuroPheno, an end-to-end audio-language framework for aphasia classification that removes the reliance on intermediate transcription pipelines. By fine-tuning a large audio-language model (Qwen2-Audio-7B) using parameter-efficient LoRA adaptation on an AphasiaBank-derived dataset, the system achieves a test accuracy of 0.89 on 4,045 held-out speech segments, with strong performance on aphasia detection. The proposed framework uses both global and local interpretability mechanisms through SHAP-based feature attribution and structured reasoning outputs, improving transparency in clinical decision-making. Unlike traditional systems, NeuroPheno operates directly on raw speech, preserving acoustic biomarkers that are often lost in text-based pipelines. Overall, the results show that end-to-end audio-language modelling is a viable and clinically meaningful approach for automated aphasia screening. While further validation on larger and more diverse datasets is needed, the proposed framework provides a scalable and interpretable foundation for future speech-based neurological assessment systems.

REFERENCES

1. Park, C.-H., Kim, C.-H.: A novel chain-of-thought reasoning approach for Alzheimer's disease detection using large languageand vision-language models. IEEE Transactions on Neural Systems and Rehabilitation Engineering 33, 1–12 (2025)
2. Hu, X., Varkanitsa, M., Kropp, E., Betke, M., Ishwar, P., Kiran, S.: Aphasia severity prediction using a multi-modal machine learning approach. NeuroImage 317, 120734 (2025)
3. Ryskin, R., Gibson, E., Kiran, S.: Noisy-channel language comprehension in aphasia: A Bayesian mixture modeling approach. Psychonomic Bulletin & Review 32, 112–130(2025)
4. Liu, Y., et al.: Multiangle correlation feature extraction and disease prediction model construction for patients with post-stroke dysarthria. IEEE Transactions on Neural Systems and Rehabilitation Engineering 33, 1–10 (2025)
5. Shool, S., Adimi, S., Amleshi, R.S., Bitaraf, E., Golpira, R., Tara, M.: A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Medical Informatics and Decision Making 25, 1–18 (2025)
6. Shankar, R., Bundele, A., Mukhopadhyay, A.: A systematic review of natural language processing techniques for early detection of cognitive impairment. Mayo Clinic Proceedings: Digital Health (2025)
7. Wang, J., Mao, L., Zhou, H., Wei, S., Zhang, T., et al.: Altered functional connectivity in acute ischemic post-stroke non-fluent aphasia based on fMRI-EEG multimodal fusion. IEEE Transactions on Neural Systems and Rehabilitation Engineering 33, 1–12 (2025)
8. Wu, P.-H., Tseng, C.-H., Shiue, T.-T.: A machine learning-based neuro-behavior sensing virtual reality system for aphasia assessment and treatment. IEEE Transactions on Neural Systems and Rehabilitation Engineering 33, 1–10 (2026)
9. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)
10. Brown, T.B., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 1877–1901 (2020)
11. Chu, Y., et al.: Qwen2-Audio technical report. arXiv preprint arXiv:2407.10759 (2024)
12. Hu, E.J., et al.: LoRA: Low-rank adaptation of large language models. In: Proceedings of ICLR (2022)
13. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: Efficient finetuning of quantized LLMs. In: Advances in Neural Information Processing Systems (NeurIPS), vol.36 (2023)
14. MacWhinney, B., Fromm, D., Forbes, M., Holland, A.: AphasiaBank: Methods for studying discourse. Aphasiology 25(11), 1286–1307 (2011)
15. Fromm, D., MacWhinney, B., Kiran, S.: AphasiaBank. TalkBank Database, Carnegie Mellon University (2011).
16. APROCSA Group: APROCSA: Aphasia Protocol Corpus of Spanish Aphasia. Dataset (2020)
17. Linguistic Data Consortium: USC Spoken American English Corpus. Linguistic Data Consortium, Philadelphia (2004)