NeuroXVocal: Detection and Explanation of Alzheimer's Disease through Non-invasive Analysis of Picture-prompted Speech

Nikolaos Ntampakis[1,2][0009−0001−7546−4104], Konstantinos Diamantaras[1][0000−0003−1373−4022], Ioanna Chouvarda[3][0000−0001−8915−6658], Magda Tsolaki[4][0000−0002−2072−8010], Panagiotis Sarigiannidis[5,2][0000−0001−6042−0355], and Vasileios Argyriou[6][0000−0003−4679−8049]

[1] International Hellenic University, Sindos, Greece
nikontam1@iee.ihu.gr
[2] MetaMind Innovations, Kozani, Greece
nntampakis@metamind.gr
[3] Aristotle University of Thessaloniki, Thessaloniki, Greece
[4] Greek Association of Alzheimer's Disease & Related Disorders, Thessaloniki, Greece
[5] University of Western Macedonia, Kozani, Greece
[6] Kingston University London, London, UK

Abstract. The early diagnosis of Alzheimer's Disease (AD) through non invasive methods remains a significant healthcare challenge. We present NeuroXVocal, the first end-to-end explainable AD classification system that achieves state-of-the-art performance while providing clinically interpretable explanations. Our novel dual-component architecture consists of: (1) Neuro, a multimodal classifier implementing a unique transformer based fusion strategy that projects acoustic, textual, and speech embeddings into a common dimensional space for complex cross-modal interactions; and (2) XVocal, a specialized RAG-based explainer that retrieves relevant clinical literature to generate evidence-based explanations. Unlike previous approaches using late fusion or simple concatenation, our architecture enables both robust classification and meaningful clinical insights. Using the IS2021 ADReSSo Challenge benchmark dataset, NeuroXVocal achieved 95.77% accuracy, significantly outperforming previous state-of-the-art. Medical professionals validated the clinical relevance of XVocal's explanations through structured evaluation. This work advances beyond pure classification to bridge the gap between machine learning predictions and clinical decision-making. Code available at:
https://github.com/NNtamp/NeuroXVocal.

Keywords: Alzheimer · Multimodal · Explainable Healthcare AI.

1 Introduction

Alzheimer's Disease (AD) has emerged as a critical global health concern, affecting over 55 million people worldwide with nearly 10 million new cases annually [1]. Early detection through non-invasive methods remains crucial for effective intervention and treatment planning. While traditional diagnostic approaches rely on neuroimaging or invasive procedures, recent advances in artificial intelligence have opened new possibilities for early detection through speech analysis [2, 3]. This paper presents NeuroXVocal, a novel dual-component system that not only classifies but also explains its diagnostic predictions through speech analysis of patients describing images, whether they are identified as having Alzheimer's disease or being cognitively healthy. The relationship between cognitive decline and speech patterns has been extensively studied using the ADReSSo benchmark dataset [4]. Syed et al. achieved significant results using functionals of deep textual embeddings, reporting 84.51% accuracy in AD detection [5]. Shah et al. further investigated language-agnostic speech representations, demonstrating the effectiveness of speech intelligibility features with 79.6% accuracy [6]. More recently, Fu et al. proposed a multimodal fusion method combining acoustic and semantic information using ImageBind audio encoder and ELMo, achieving 90.3% accuracy [7]. Li et al. demonstrated promising results using Whisper-based transfer learning, achieving 84.51% accuracy and 84.50% F1-score by innovatively using full transcripts as prompts during fine-tuning [8]. The latest advancement by Lee et al. introduced a graph neural network leveraging image-text similarity from vision language models, achieving 88.73% accuracy [9]. While these approaches have shown promising results in classification, the field has seen limited progress in explaining the reasoning behind diagnostic predictions. Recent work by Iqbal et al. employed Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) to provide insights into linguistic markers of cognitive decline [10]. Similarly, Bang et al. explored the use of LLMs for generating evidence-based explanations of speech patterns, though their approach was limited by the interpretability of the underlying language model [11]. However, these studies still face challenges in providing comprehensive, clinically-actionable explanations that bridge the gap between machine learning predictions and medical decision-making. Building upon these foundations, we present NeuroXVocal, which addresses these limitations through the following key contributions:

1. Novel Architecture: First end-to-end framework seamlessly integrating AD classification with clinically interpretable explanations, where multimodal features contribute to both diagnosis and explanation generation.
2. Advanced Fusion Strategy: A transformer-based architecture that projects acoustic features, textual features, and speech embeddings into a common dimensional space before fusion, enabling complex cross-modal interactions superior to existing late-fusion approaches.
3. Clinical Explainability: Introduction of XVocal, a specialized RAG component that retrieves relevant AD research to generate evidence-based explanations. Medical professionals validated its clinical relevance, confirming its potential as a diagnostic support tool.

4. State-of-the-art Performance: Achievement of 95.77% accuracy on the ADReSSo benchmark, significantly outperforming existing methods while maintaining interpretability.

2 Methodology

Our proposed novel NeuroXVocal system consists of two primary components, as in Fig. 1: (1) the Neuro classifier for AD detection through multimodal analysis of speech data, and (2) the XVocal explainer for generating clinically-interpretable justifications. The system processes input audio samples through multiple parallel streams to extract complementary features before fusion and classification.

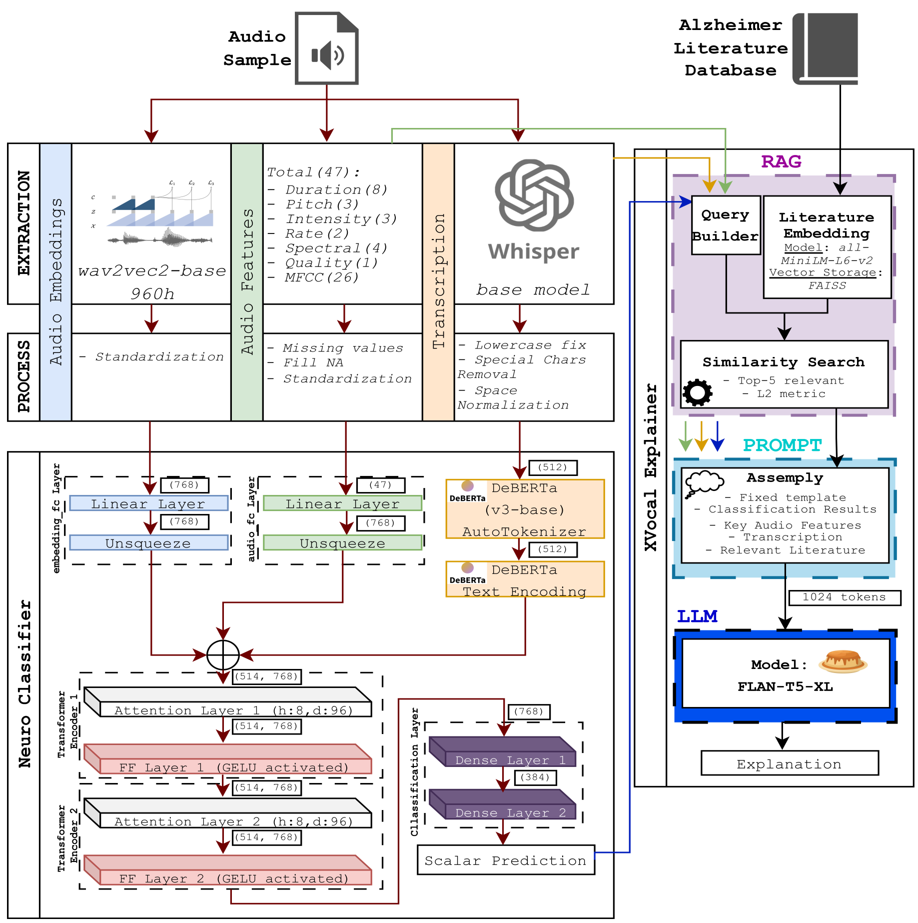

Fig. 1. NeuroXVocal Architecture

2.1 Feature Extraction and Processing

Let x be an input audio sample. From this input, we extract three distinct feature representations. The acoustic features $f_a(x) = \phi_a(x) \in \mathbb{R}^{47}$ comprise temporal characteristics (speech/pause ratios), prosodic features (pitch, intensity), articulation metrics, spectral properties, voice quality indicators (jitter, shimmer, harmonics-to-noise ratio), and 13 Mel Frequency Cepstral Coefficients(MFCC) coefficients with their standard deviations. These features (47 in total) undergo standardization and missing value imputation. For speech embeddings, we employ Wav2Vec2-base-960h [12] after converting audio to mono and resampling to 16kHz:

$$f_e(x) = \text{Mean}(\text{Wav2Vec2}(\text{Preprocess}(x))) \in \mathbb{R}^{768} \tag{1}$$

where the embeddings are standardized before further processing. The textual features are obtained using Whisper ASR [13] for transcription followed by DeBERTa-v3-base [14] encoding:

$$f_t(x) = \text{DeBERTa}(\text{Preprocess}(\text{Whisper}(x))) \in \mathbb{R}^{768} \tag{2}$$

where preprocess includes lowercase conversion, special character removal, and space normalization.

2.2 Neuro Classifier

The classification component implements a novel fusion architecture. We first project the acoustic and speech embedding features to a common dimensional space:

$$h_a = \text{Linear}(f_a(x)) \in \mathbb{R}^{768}, \quad h_e = \text{Linear}(f_e(x)) \in \mathbb{R}^{768} \tag{3}$$

The fusion process concatenates these projections with the text embeddings in the projected dimensions of 514×768:

$$H = [h_a; h_e; f_t(x)] \in \mathbb{R}^{514 \times 768} \tag{4}$$

A two-layer transformer encoder processes this representation. Each layer implements multi-head attention with 8 heads, where the input H is projected to queries (Q), keys (K), and values (V):

$$\text{Attention}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_8)W^O \tag{5}$$

where W^O is the output projection matrix, and each attention head is computed as:

$$\text{head}_i = \text{softmax}(\frac{QW_i^Q(KW_i^K)^T}{\sqrt{d_k}})VW_i^V \tag{6}$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{768 \times 96}$ are learned parameter matrices, and $d_k = 96$ is the head dimension. This is followed by a feed-forward network with GELU activation:

$$Z = \text{FFN}(\text{Attention}(H)) \in \mathbb{R}^{514 \times 768} \tag{7}$$

The final classification uses a two-layer classifier:

$$p(y|x) = \sigma(\text{Dense}(z_0)) \tag{8}$$

where σ is the sigmoid activation function for binary classification.

2.3 XVocal Explainer

The novel explainability component implements a RAG approach that processes the extracted features along with the Neuro classifier's prediction. The explanation generation begins by constructing a structured prompt query q (available in project's GitHub repository) through a template:

$$q = \{\text{class}(p(y|x)) \oplus \text{features}(f_a(x)) \oplus \text{speech}(f_e(x)) \oplus \text{transcript}(f_t(x))\} \tag{9}$$

The relevant literature corpus $\mathcal{L}$ is preprocessed into semantic chunks by splitting each document into paragraphs and then into individual sentences to create a fine-grained context pool $\{c_1, ..., c_n\}$. Using all-MiniLM-L6-v2 [16], we construct a dense vector index:

$$E_c = \{\text{MiniLM}(c_i) \in \mathbb{R}^{384} | c_i \in \mathcal{L}\} \tag{10}$$

where each chunk is encoded into a 384-dimensional embedding space. These embeddings are indexed using FAISS [15] L2 distance metric:

$$\mathcal{I} = \text{FAISS}_{\text{L2}}(E_c) \tag{11}$$

For retrieval, the query q is encoded in the same embedding space and the top 5 most relevant chunks are retrieved using nearest neighbor search:

$$\mathcal{L}_r = \{\mathcal{I}.\text{search}(\text{MiniLM}(q), k = 5)\} \tag{12}$$

The final explanation is generated using FLAN-T5-XL [17]:

$$E = \text{FLAN-T5}(q \oplus \mathcal{L}_r; \tau, p) \tag{13}$$

where τ and p are the temperature and top-p sampling parameters respectively, controlling the generation coherence.

3 Experiments and Results

3.1 Implementation Details

All experiments were conducted on Ubuntu using 8xNVIDIA A16 GPUs with 126GB system RAM. The Neuro classifier trained for a maximum of 200 epochs. For the XVocal component, we used FAISS (CPU) for retrieval and deployed using 4-bit quantization. Each training round was completed on an average of 9 hours in our setup.

3.2 Dataset

We utilised the ADReSSo Challenge dataset [4] for the probable AD prediction task. The data is organized in the diagnosis folder, with 166 patients in the training set (79 cognitively normal [cn], 87 probable Alzheimer's disease [ad]) and 71 patients in the test set. The test set is kept independent for transparent evaluation. The dataset is accessible through DementiaBank membership, requiring registration and administrator approval. The complete dataset documentation is available through our project repository.

3.3 Results

Table 1. Performance comparison on ADReSSo dataset. A: Acoustic, T: Text, S: Speech embeddings

Methodology	Modalities	5-fold Accuracy(± std%)	Acc[%]	F1-score[%]
Syed et al.(2021) [5]	T		84.51%	84.45%
Shah et al.(2023) [6]	A+T		79.60%	
Fu et al.(2024) [7]	A+T		90.3%	91.4%
Li et al.(2024) [8]	T+S		84.51%	84.5%
Lee et al.(2025) [9]	T+S		88.73%	88.23%
(Neuro)XVocal	**A+T+S**	**96.24% ± 2.47%**	**95.77%**	**95.76%**

Regarding the results of the Neuro Classifier incorporated in our NeuroX-Vocal methodology, we compared with prominent and recent state-of-the-art methodologies as shown in Table 1. To evaluate the performance, we utilized the widely adopted accuracy and F1-score metrics. As demonstrated in Table 1, our Neuro classifier achieved robust performance across multiple evaluation scenarios. In the 5-fold cross-validation setting, we obtained an average accuracy of 96.24% with a standard deviation of 2.47%. When trained on the full training set and evaluated on the independent test set, our method achieved 95.77% accuracy and 95.76% F1-score, substantially outperforming all previous approaches. To assess the clinical relevance and utility of XVocal's explanations, we conducted a comprehensive qualitative evaluation with medical experts. Each expert evaluated explanations for 20 patient cases (10 AD, 10 CN) using a structured questionnaire with 10 criteria[1], rated on a 5-point Likert scale. For the knowledge base of the RAG component, we have incorporated a curated corpus of 10 seminal publications [18–27] covering linguistic markers, spontaneous speech analysis, and LLM applications in AD detection.

The evaluation results (Table 2) demonstrate strong performance across multiple dimensions of clinical utility. XVocal achieved notably high scores in AD marker identification (3.98) and explanation clarity (3.96), indicating its effectiveness in highlighting relevant diagnostic features. The system also performed

[1] Questionnaire available at: https://forms.gle/rAFuC6ediUYrqQzf8

Table 2. Criteria and expert evaluation results for XVocal's explanations

Assessment Focus	Scale	Mean Score
Clear justification of diagnosis	1-Not clear, 5-Very clear	3.96
Pertinence of identified markers	1-Not relevant, 5-Highly relevant	3.85
Consistency with medical knowledge	1-No alignment, 5-High alignment	3.63
Explanation-based confidence	1-Not confident, 5-Highly confident	3.63
Recognition of disease indicators	1-No markers identified, 5-Highly appropriate markers identified	3.98
Utility for diagnosis	1-Not useful, 5-Highly useful	3.70
Coherence and plausibility	1-Not sound, 5-Very sound	3.74
Expected consensus	1-Very unlikely, 5-Very likely	3.56
Robustness of reasoning	1-Not at all plausible, 5-Highly plausible	3.77
Potential for misinterpretation	1-Not misleading, 5-Highly misleading	2.38

well in identifying relevant linguistic features (3.85) and maintaining logical soundness (3.74), suggesting reliable diagnostic reasoning. Particularly noteworthy is the low score for potentially misleading aspects (2.38), indicating that experts found minimal risk of misinterpretation in XVocal's explanations. This is crucial for clinical applications where accuracy and reliability are paramount. The system also demonstrated good alignment with clinical understanding (3.63) and strong utility for supporting diagnostic decisions (3.70). XVocal successfully identified key speech markers such as increased pause durations and reduced semantic fluency, connecting these features to established AD literature.

4 Ablation Study

To assess the contribution of each modality, we conducted systematic experiments by removing components and adapting the network architecture accordingly. For each combination, we modified the dimensions of the input layers to match the sizes of the feature vector. The fusion layer and attention mechanisms were adjusted proportionally while maintaining the core architecture design.

Results, as shown in Table 3, demonstrate the synergistic effect of multimodal fusion, with transcription features providing the strongest individual contribution when combined with audio embeddings (91.30%). The transcription features prove crucial, as configurations lacking this component show reduced performance (84.78%). Acoustic features seems to be the weaker modality, suggesting they capture complementary speech characteristics. The optimal performance (95.77%) achieved with all three modalities indicates each component contributes unique discriminative information essential for robust AD detection.

Table 3. Ablation study results showing modality combinations

Audio Embed.	Audio Feat.	Text Trans.	Accuracy[%]	F1-score[%]
✓	✓	✓	95.77	95.76
	✓	✓	89.86	89.86
✓		✓	91.30	91.29
✓	✓		84.78	84.70

5 Conclusion

We presented NeuroXVocal, a novel dual-component system that advances the state-of-the-art in both AD detection accuracy and clinical interpretability. Our key contributions include: (1) the first end-to-end framework seamlessly integrating high-accuracy classification (95.77%) with evidence-based explanations, (2) a transformer-based architecture enabling superior cross-modal fusion through common dimensional space projection, and (3) a specialized RAG-based explainer validated by medical professionals for clinical relevance. Unlike previous approaches focusing solely on classification, NeuroXVocal bridges the critical gap between machine learning predictions and clinical decision-making. Future work will focus on developing a real-time inference pipeline and implementing streaming audio processing for immediate feature extraction. We plan to extend the system with an interface for clinical deployment, incorporating incremental learning capabilities to adapt to new data patterns. Additionally, we aim to expand the knowledge base with continuous literature updates and enhance the RAG component with domain-specific prompt engineering for more targeted explanations. Further validation through large-scale clinical trials will help establish NeuroXVocal's efficacy as a practical diagnostic support tool.

Acknowledgments. This project has received funding from the European Union's Horizon Europe research and innovation programme (G.A. No. 101135800 - RAIDO) and by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (G.N. 10099264). The authors extend their gratitude to the medical experts team of Prof. Magda Tsolaki for their invaluable contribution to the evaluation process and validation of the XVocal component.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. World Health Organization.: Global status report on the public health response to dementia. WHO Press, Geneva (2021)
2. Wong, K.K., et al.: Enhance early diagnosis accuracy of Alzheimer's disease by elucidating interactions between amyloid cascade and tau propagation. In: MICCAI 2023, pp. 1–10. Springer, Vancouver (2023).

3. Luz, S., Haider, F., de la Fuente, S., et al.: Noninvasive automatic detection of Alzheimer's disease through speech analysis. In: Frontiers in Aging Neuroscience, vol. 15, Article 1224723 (2023)
4. Luz, S., Haider, F., de la Fuente, S., Fromm, D., MacWhinney, B. (2021).: ADReSSo Challenge Dataset [Data set]. DementiaBank. Retrieved January 9, 2025, from https://dementia.talkbank.org/ADReSSo-2021/
5. Syed, Z.S., Shah, M.S., Lech, M., et al.: Tackling the ADRESSO Challenge 2021: The MUET-RMIT System for Alzheimer's Dementia Recognition from Spontaneous Speech. In: Interspeech 2021, pp. 3780–3784. ISCA, Brno (2021)
6. Shah, Z., Sawalha, J., Tasnim, M., et al.: Exploring language-agnostic speech representations using domain knowledge for detecting Alzheimer's dementia. In: ICASSP 2023, pp. 1–5. IEEE Press, Rhodes (2023)
7. Fu, Y., Xu, L., Zhang, Y., et al.: Classification and diagnosis model for Alzheimer's disease based on multimodal data fusion. Medicine 103(52), (2024)
8. Li, J., Zhang, W.Q.: Whisper-Based Transfer Learning for Alzheimer Disease Classification: Leveraging Speech Segments with Full Transcripts as Prompts. In: ICASSP 2024, pp. 11211–11215. IEEE Press, Seoul (2024)
9. Lee, B., Bang, J.U., Song, H.J., et al.: Alzheimer's disease recognition using graph neural network by leveraging image-text similarity from vision language model. Scientific Reports 15, 997 (2025)
10. Iqbal, F., Syed, Z.S., Syed, M.S.S., Syed, A.S.: An Explainable AI Approach to Speech-Based Alzheimer's Detection Using Linguistic Features. In: ISCA SMM 2024, pp. 1–6. ISCA Press (2024)
11. Bang, J.-U., Han, S.-H., Kang, B.-O.: Alzheimer's Disease Recognition from Spontaneous Speech Using Large Language Models. In: ETRI Journal 46(1), pp. 1–10 (2024)
12. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems 33, 12449–12460 (2020)
13. Radford, A., Kim, J. W., Hallacy, C., et al.: Robust Speech Recognition via Large-Scale Weak Supervision. OpenAI (2022)
14. He, P., Gao, J., Chen, W.: DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. International Conference on Learning Representations (ICLR) (2023)
15. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., et al.: The Faiss Library. IEEE Transactions on Big Data 7(3), 535–547 (2024)
16. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. Advances in Neural Information Processing Systems 33, 5776–5788 (2020)
17. Chung, H. W., Hou, L., Longpre, S., et al.: Scaling Instruction-Finetuned Language Models. In: Proceedings of the Association for Computational Linguistics (ACL), pp. 1–12 (2023)
18. Fraser, K.C., Meltzer, J.A., Rudzicz, F.: Linguistic features identify Alzheimer's disease in narrative speech. Journal of Alzheimer's Disease 49(2), 407–422 (2016)
19. Tóth, L., et al.: Speech recognition-based classification of mild cognitive impairment and dementia. LNCS 11096, 89–98. Springer, Heidelberg (2018)
20. Ahmed, S., Haigh, A.M., de Jager, C.A., Garrard, P.: Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. Brain 136(12), 3727–3737 (2013)

21. Forbes-McKay, K.E., Shanks, M.F., Venneri, A.: Charting the decline in spontaneous writing in Alzheimer's disease: a longitudinal study. Acta Neuropsychiatrica 25(6), 320–327 (2013)
22. Taler, V., Phillips, N.A.: Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. Journal of Clinical and Experimental Neuropsychology 30(5), 501–556 (2008)
23. König, A., et al.: Automatic Speech Analysis for the Assessment of Alzheimer's Disease: A Review. Journal of Alzheimer's Disease Reports 2(1), 1–15 (2018)
24. Ding, K., Chetty, M., Noori Hoshyar, A., et al.: Speech based detection of Alzheimer's disease: a survey of AI techniques, datasets and challenges. Artificial Intelligence Review 57, 325 (2024)
25. Qi, X., Zhou, Q., Dong, J., Bao, W.: Noninvasive automatic detection of Alzheimer's disease from spontaneous speech: a review. Frontiers in Aging Neuroscience 15, Article 1224723 (2023)
26. Chen, J., Ye, J., Tang, F., Zhou, J.: Automatic Detection of Alzheimer's Disease Using Spontaneous Speech Only. In: Interspeech 2021, pp. 3830–3834 (2021)
27. Agbavor, F., Liang, H.: Predicting dementia from spontaneous speech using large language models. PLOS Digital Health 1(12), e0000168 (2022)