RESEARCH PAPER

COGNICARE: Early Alzheimer's Disease Detection Using Multimodal Speech and Text Analysis

Dr. C. Padmaja[1*], Dr. Sushitha Susan Joseph[2], Dr. Selvi C[3]

[1*]Associate Professor, Dept. of ECE, G. Narayanamma Institute of Technology & Science, Hyderabad, India.
Email: c.padmaja@gnits.ac.in; ORCID: 0000-0003-0521-916X

[2]Assistant Professor, Dept. of CSE, Indian Institute of Information Technology Kottayam, Kottayam, Kerala, India.
Email: sushithasj@iiitkottayam.ac.in

[3]Assistant Professor, Dept. of CSE, Indian Institute of Information Technology Kottayam, Kottayam, Kerala, India.
Email: selvic@iiitkottayam.ac.in

ABSTRACT

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by memory loss, cognitive decline, and impaired communication abilities. Early diagnosis is critical for timely intervention and improved patient management; however, conventional diagnostic approaches such as Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and neuropsychological assessments are often expensive, time-consuming, and inaccessible for large-scale screening. This paper presents COGNICARE, a machine learning-based multimodal framework for early Alzheimer's disease risk prediction using speech and text data. The proposed system utilizes acoustic features extracted from speech recordings and linguistic features derived from textual descriptions. An XGBoost classifier is employed for speech-based analysis, while Logistic Regression is used for text-based classification. Predictions from both modalities are combined using an Adaptive Fusion Algorithm to obtain a comprehensive cognitive risk score. The system is developed as a user-friendly web application using Gradio to facilitate preliminary screening and monitoring. Experimental evaluation on the DementiaBank Pitt Corpus demonstrates promising performance, achieving 88.00% accuracy for speech-based detection and 94.12% accuracy for text-based detection. The proposed multimodal framework offers a cost-effective, non-invasive, and scalable solution for early Alzheimer's disease assessment.

Keywords: Alzheimer's Disease, Dementia Detection, Speech Analysis, Natural Language Processing, XGBoost, Logistic Regression, Multimodal Learning, Machine Learning, Healthcare Analytics.

How to cite this article: Padmaja C, Joseph SS, Selvi C. COGNICARE: Early Alzheimer's Disease Detection Using Multimodal Speech and Text Analysis. Int J Drug Deliv Technol. 2026;16(61s):1416-1422. DOI: 10.25258/ijddt.16.61s.159

Source of support: Nil

Conflict of interest: None

INTRODUCTION

Alzheimer's disease (AD) is one of the most prevalent neurodegenerative disorders affecting millions of individuals worldwide. The disease causes a gradual deterioration of cognitive functions, including memory, reasoning, language, and communication abilities. Clinical manifestations generally progress through stages such as Subjective Memory Loss (SML), Mild Cognitive Impairment (MCI), and Alzheimer's Dementia (AD).

Traditional diagnostic procedures involve neuroimaging techniques such as MRI and PET scans, alongside comprehensive clinical evaluations. Although effective, these approaches are costly, resource-intensive, and often detect the disease only after substantial neurological damage has occurred. There is therefore a need for an accurate, reliable, and accessible system capable of leveraging both speech and textual information to facilitate early risk prediction and support healthcare professionals in preliminary screening and monitoring.

Recent advances in Artificial Intelligence (AI), Machine Learning (ML), speech processing, and Natural Language Processing (NLP) have enabled the development of automated diagnostic support systems. Since speech and language impairments are among the earliest indicators of cognitive decline, speech recordings and textual transcripts provide valuable biomarkers for Alzheimer's detection.

This research proposes COGNICARE, a multimodal machine learning system that integrates speech and text analysis to provide an accessible and reliable mechanism

for early Alzheimer's risk prediction. The objectives of this work include:

1. Developing a machine learning-based framework for Alzheimer's risk prediction using speech and text.

2. Identifying cognitive decline through speech characteristics such as: Pauses, Pitch variation, Speaking rate and Fluency

3. Detecting linguistic abnormalities through text analysis.

4. Combining speech and language assessments using multimodal fusion.

5. Designing a user-friendly web application for practical deployment.

The remainder of this paper is organized as follows. Section II presents the related work on speech- and text-based Alzheimer's disease detection techniques. Section III describes the proposed methodology, including data collection, preprocessing, feature extraction, and multimodal fusion. Section IV presents the experimental setup and performance evaluation results. Finally, Section V concludes the paper and outlines future research directions.

LITERATURE SURVEY

Recent advances in Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP), and speech analysis have significantly contributed to the development of automated systems for Alzheimer's disease (AD) detection. Researchers have explored various speech and linguistic biomarkers to identify cognitive impairment at an early stage.

Yang et al. [1] presented a comprehensive review of learning-based speech analysis techniques for Alzheimer's disease detection. The study highlighted the effectiveness of acoustic and linguistic features in identifying cognitive decline and discussed the potential of machine learning algorithms for non-invasive diagnosis. However, the review primarily focused on existing methodologies and did not propose a deployable diagnostic framework.

Ding et al. [2] conducted an extensive survey of AI techniques, datasets, and challenges associated with speech-based Alzheimer's disease detection. Their work provided valuable insights into publicly available datasets, feature extraction methods, and classification techniques. The authors emphasized the need for robust multimodal systems capable of integrating diverse cognitive biomarkers. Nevertheless, the study remained a survey and lacked experimental validation.

Bang et al. [3] investigated the use of spontaneous speech as a biomarker for Alzheimer's disease recognition. The study demonstrated that speech characteristics such as fluency, pauses, and lexical richness can effectively differentiate cognitively healthy individuals from patients with Alzheimer's disease. The findings confirmed the potential of speech analysis for early-stage cognitive assessment.

Cai et al. [4] proposed a multimodal framework that combined speech transcripts and acoustic information for Alzheimer's disease detection. Their results indicated that integrating linguistic and audio features improved classification performance compared with single-modality approaches. The study highlighted the importance of multimodal learning; however, the proposed framework involved relatively complex feature extraction and modeling procedures.

Ahn et al. [5] explored deep learning techniques for speech-based Alzheimer's disease detection. By employing neural network architectures, the researchers demonstrated improved performance in capturing complex speech patterns associated with cognitive decline. Although the results were promising, deep learning models required substantial computational resources and large amounts of training data, limiting their applicability in lightweight clinical screening systems.

Runde et al. [6] focused on enhancing Alzheimer's disease detection using Natural Language Processing techniques. Their study showed that linguistic features extracted from patient speech transcripts could effectively identify cognitive impairment. The authors reported improved diagnostic performance through advanced text-processing methods, reinforcing the significance of language abnormalities as early indicators of dementia.

Oiza-Zapata et al. [7] proposed an explainable artificial intelligence framework for Alzheimer's detection from speech using Shapley Additive Explanations (SHAP) for feature selection and model interpretation. Their findings demonstrated that explainable speech biomarkers can improve clinical trust while maintaining high classification performance.

Lee et al. [8] developed a multimodal framework that integrates image, text, and audio information obtained from picture-description tasks. Their study showed that combining multiple modalities significantly improves Alzheimer's disease recognition compared with single-modality approaches, highlighting the importance of multimodal learning for cognitive assessment.

Azadmaleki et al. [9] introduced SpeechCARE, a dynamic multimodal framework that combines acoustic, linguistic, and demographic information using transformer-based architectures. The proposed model

achieved promising performance in detecting cognitive impairment and Alzheimer's disease across diverse speech tasks and linguistic contexts.

To further enhance diagnostic performance, Favaro et al. [10] investigated the fusion of speech and eye-movement biomarkers for Alzheimer's disease detection. Experimental results demonstrated that multimodal fusion improves predictive accuracy by leveraging complementary cognitive indicators obtained from different behavioral modalities.

Ksibi et al. [11] proposed a multimodal Siamese neural network architecture for dementia detection from speech recordings. Their approach improved feature representation learning and classification accuracy by effectively modeling similarities and differences between cognitively healthy and impaired individuals.

Flick and Ostrand [12] explored context-sensitive linguistic features extracted from connected speech for cognitive impairment prediction. Their study demonstrated that automatically generated contextual language features provide better discrimination than conventional lexical and syntactic measures.

With the emergence of Large Language Models (LLMs), Llaca-Sánchez et al. [13] investigated the use of transformer-based embeddings derived from speech transcripts. Their findings indicated that LLM-generated representations capture subtle linguistic biomarkers associated with cognitive decline and outperform traditional text representations in several classification tasks.

Ortiz-Perez et al. [14] proposed CogniAlign, a multimodal framework that aligns speech and text information at the word level using gated cross-attention mechanisms. The model achieved state-of-the-art performance by effectively integrating prosodic and linguistic information for Alzheimer's disease detection.

Pérez-Toro et al. [15] evaluated automated speech biomarkers across multiple languages to study their cross-linguistic generalizability. Their results demonstrated that several speech-based indicators remain effective across different languages, improving the applicability of automated screening systems in multilingual settings.

Wang et al. [16] combined speech-derived digital biomarkers with biological fluid biomarkers using machine learning techniques. The multimodal approach significantly improved prediction of cognitive impairment, demonstrating the potential of integrating behavioral and biological indicators for clinical decision support.

Despite these advances, most existing studies primarily focus on either speech analysis or text analysis independently. While some multimodal approaches have been proposed, many involve computationally intensive deep learning architectures that may not be suitable for practical deployment in resource-constrained healthcare environments. Furthermore, limited research has investigated lightweight machine learning frameworks capable of integrating speech and text information while maintaining high predictive performance.

To address these limitations, the proposed COGNICARE framework employs a multimodal approach that combines acoustic and linguistic features using XGBoost-based speech analysis, Logistic Regression-based text analysis, and an Adaptive Fusion Algorithm. This integration aims to provide a cost-effective, non-invasive, and scalable solution for early Alzheimer’s disease risk prediction.

METHODOLOGY

The proposed COGNICARE framework consists of three major components:

- Audio Analysis Module
- Text Analysis Module
- Adaptive Fusion Module

The system operates as a sequential, 4-stage pipeline that processes user data from initial login to final diagnostic feedback shown in figure 1. In the first stage, an authentication module manages secure user access through dedicated patient and doctor portals while ensuring privacy and role-based authorization. The second stage performs data processing, where audio recordings and textual descriptions are collected and converted into structured feature vectors through acoustic and linguistic feature extraction. The third stage utilizes machine learning models to analyze the extracted features and generate a cognitive risk score, which is subsequently categorized into low, moderate, or high-risk levels. Finally, the output stage stores assessment records in a centralized database and presents results through patient and doctor dashboards, enabling continuous monitoring, historical analysis, and timely clinical intervention.

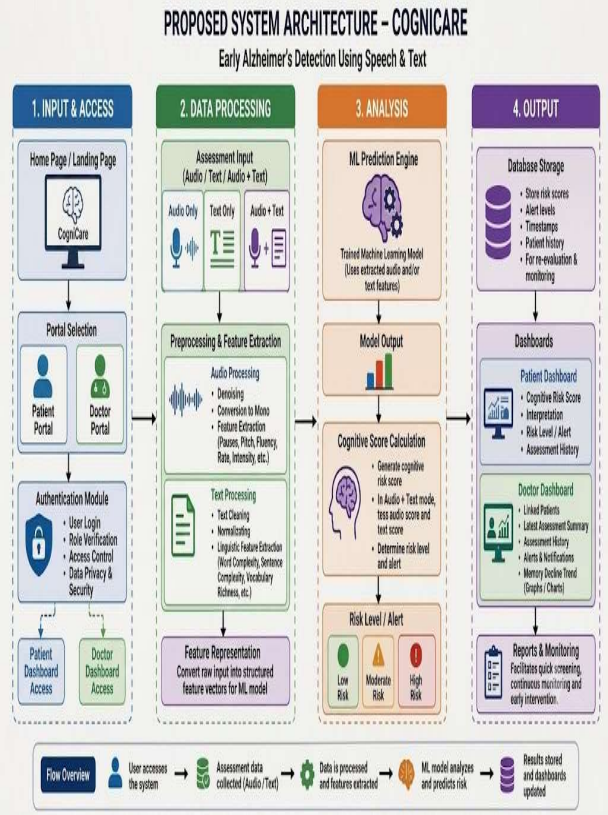

Fig. 1 Proposed methodology overview

The proposed study utilizes the DementiaBank Pitt Corpus, one of the most widely used benchmark datasets for Alzheimer's disease detection research. The dataset contains speech recordings and corresponding transcripts collected from healthy control subjects and patients diagnosed with Alzheimer's disease.

The corpus supports multimodal analysis by providing both acoustic and linguistic information. The dataset characteristics are summarized in Table I.

Table 1. Dataset Characteristics

Parameter	**Value**
Dataset Name	DementiaBank Pitt Corpus
Dataset Type	Multimodal (Audio + Text)
Original Audio Samples	530
Augmented Audio Samples	1590
Audio Sampling Rate	16 kHz
Transcript Files	552
Linguistic Features	10
Final Text Feature Dimension	12,010

Speech recordings and textual descriptions are collected from participants. Speech signals undergo acoustic feature extraction to capture biomarkers associated with cognitive decline, while textual transcripts are processed to derive linguistic and semantic features. The extracted features are then analyzed independently using machine learning classifiers. The speech modality is classified using XGBoost, whereas the textual modality is analyzed using Logistic Regression. Finally, predictions from both modalities are integrated through an Adaptive Fusion Algorithm to generate a comprehensive cognitive risk score.

The proposed COGNICARE framework employs two machine learning models to analyze different modalities of patient data. Since speech and text contain distinct cognitive biomarkers, separate classifiers are trained for each modality. The audio classification module utilizes Extreme Gradient Boosting (XGBoost) to analyze acoustic features extracted from speech recordings, while the text classification module employs Logistic Regression to evaluate linguistic characteristics derived from textual descriptions. The outputs of both classifiers are subsequently combined through an Adaptive Fusion Algorithm to generate a comprehensive cognitive risk assessment.

A. Speech-Based Classification Using XGBoost

Speech impairment is one of the earliest indicators of Alzheimer's disease. Patients often exhibit slower speech, increased hesitation, longer pauses, reduced pitch variation, and decreased fluency. To effectively model these complex relationships, XGBoost is employed as the speech classifier.

XGBoost (Extreme Gradient Boosting) is an ensemble learning algorithm based on Gradient Boosting Decision Trees (GBDT). Instead of relying on a single decision tree, XGBoost constructs multiple trees sequentially, where each new tree learns from the errors made by the previous trees. This iterative learning process improves prediction accuracy while reducing bias and variance.

For each speech recording, 146 acoustic features are extracted, including: Speech rate, Pause frequency, Average pause duration, Pitch variation, Fluency measures, Rhythm characteristics and Acoustic descriptors. The final prediction of XGBoost is obtained by summing the outputs of all decision trees:

$$\hat{y} = \sum_{k=1}^{K} f_k(x) \qquad (1)$$

where:

$\hat{y}$ = predicted output

$f_k(x)$ = output of the kth decision tree

K = total number of trees

After training, the model outputs class probabilities representing the likelihood of Alzheimer's disease. Thus, the speech-based cognitive risk score is computed, which indicate that speech characteristics provide valuable biomarkers for detecting cognitive decline.

B. Text-Based Classification Using Logistic Regression

Language impairment is another prominent symptom of Alzheimer's disease. Patients frequently exhibit reduced vocabulary richness, repetitive word usage, simplified sentence structures, and decreased linguistic complexity.

To analyze these patterns, Logistic Regression is employed as the text classifier. Logistic Regression is a supervised machine learning algorithm widely used for binary classification problems due to its simplicity, interpretability, and computational efficiency.

The text preprocessing module extracts: Word-level TF-IDF features, Character-level TF-IDF features, Vocabulary richness, Sentence length, Repetition rate, Word complexity measures. The resulting feature vector contains 12,010 dimensions.

Logistic Regression estimates the probability of Alzheimer's disease using the sigmoid function:

$$P(y = 1) = \frac{1}{1 + e^{-z}} \quad (2)$$

Where

$$z = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b \quad (3)$$

where z is the weighted combination of input features. The decision rule is defined as follows:

Probability > 0.5 -> Alzheimer's Disease

Probability $\leqslant$ 0.5 -> Healthy Control

C. Adaptive Fusion Algorithm

Although each modality provides valuable information independently, relying on a single modality may lead to incomplete assessments. For example, poor audio quality may reduce speech classification accuracy, while transcription errors or limited text input may affect linguistic analysis. Therefore, integrating information from both modalities can provide a more reliable and comprehensive evaluation of cognitive health.

To address this challenge, the proposed COGNICARE framework employs an Adaptive Fusion Algorithm that combines predictions generated by the speech-based XGBoost classifier and the text-based Logistic Regression classifier. The fusion process operates at the decision level. Instead of combining raw features, each modality independently generates a probability score indicating the likelihood of Alzheimer's disease.

The outputs are: Audio Score (generated by XGBoost) and Text Score (generated by Logistic Regression). These scores are subsequently combined using weighted averaging to obtain a Final Cognitive Risk Score. The final risk score is computed as:

$$Final\ Score = w_a\ (Audio\ Score) + w_t\ (Text\ Score) \quad (4)$$

Where w_a and w_t weights assigned to the audio and text modalities, respectively. The weights determine the relative contribution of each modality to the final prediction

- $\mathbf{w_a}$ = Audio weight
- $\mathbf{w_t}$ = Text weight

$$w_a + w_t = 1 \quad (5)$$

The final cognitive risk score is mapped to predefined risk categories to facilitate clinical interpretation shown in table 2. These categories allow healthcare professionals to quickly identify patients requiring further clinical evaluation.

Table 2. Cognitive risk score parameters

Score Range	Risk Level
0 – 39	Low Risk
40 – 69	Moderate Risk
70 – 100	High Risk

These scores are subsequently combined using the Adaptive Fusion Algorithm to generate the Final Cognitive Risk Score, which is used to categorize the patient into Low, Moderate, or High-Risk groups.

This multimodal assessment improves the reliability of early Alzheimer's disease screening and supports healthcare professionals in identifying individuals who may benefit from further neuropsychological evaluation and clinical intervention.

RESULTS AND DISCUSSION

The proposed COGNICARE framework functions as a clinical decision-support system rather than a standalone diagnostic tool. Following multimodal analysis, the generated Cognitive Risk Score is displayed on the Doctor Dashboard together with the patient's speech transcript, extracted biomarkers, historical assessments, and risk trends.

The physician evaluates both the AI-generated results and the content of the patient's picture description. Particular attention is given to the completeness of the description, vocabulary richness, sentence structure,

fluency, repetition patterns, and the patient's ability to identify key events in the Cookie Theft Picture.

To demonstrate the practical operation of the proposed system, a patient is presented with the Cookie Theft Picture and asked to describe the events occurring in the scene shown in figure 2.

Speech analysis captures abnormalities such as reduced fluency, prolonged pauses, and decreased pitch variation. Text analysis identifies linguistic deficits including repetitive language, reduced vocabulary richness, and simplified sentence structures.

The multimodal design therefore offers a practical solution for scalable cognitive screening.

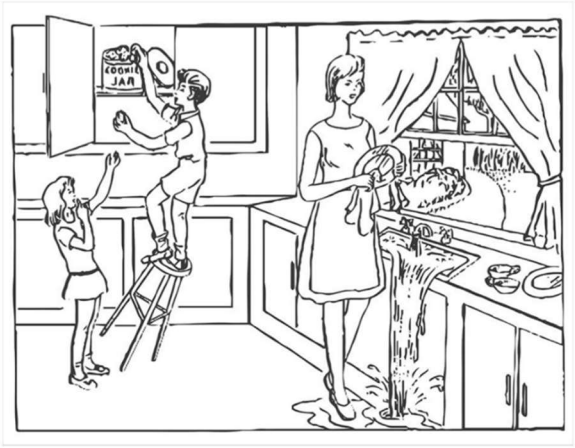

Fig. 2 Cognitive Assessment using the Cookie Theft Picture [17]

A cognitively healthy individual typically provides a detailed description such as: "A boy is standing on a stool taking cookies from a jar while his sister watches. Their mother is washing dishes and the sink is overflowing." In contrast, an individual with cognitive impairment may produce fragmented responses containing frequent pauses, hesitations, and missing details.

Table 3. Audio/text Processing Example

Audio Processing		**Text Processing**	
Feature	**Value**	**Feature**	**Value**
Speech Rate	92 words/min	Vocabulary Richness	Low
Pause Count	15	Sentence Length	3.2
Avg. Pause Duration	1.8 sec	Repetition Rate	High
Pitch Variation	Low	Word Complexity	Low
Fluency Score	78%	Fluency Score	90%

Final Cognitive Risk Score can be

$$Final\ Score = 0.4\ (78) + 0.6\ (90) = 85.2$$

The patient is therefore classified as **High Risk**.

The proposed multimodal framework demonstrates the effectiveness of combining speech and text modalities for Alzheimer's disease detection.

Experimental observations indicate that text-based analysis achieved higher standalone accuracy than speech analysis. However, combining both modalities through adaptive fusion increases reliability and robustness by compensating for weaknesses in individual modalities.

CONCLUSION

This paper presented COGNICARE, a machine learning-based multimodal framework for early Alzheimer's disease detection using speech and text data. The system employs XGBoost for acoustic analysis, Logistic Regression for linguistic analysis, and an Adaptive Fusion Algorithm for multimodal integration. Experimental evaluation on the DementiaBank Pitt Corpus demonstrated promising performance with accuracies of 88.00% and 94.12% for audio and text modalities respectively. The proposed framework provides a cost-effective, non-invasive, and user-friendly alternative to conventional clinical screening methods.

Future work will focus on integrating advanced deep learning architectures including CNNs, LSTMs, Transformers, and Large Language Models to further improve predictive performance and capture complex cognitive biomarkers.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the All India Council for Technical Education (AICTE) for providing the opportunity to pursue the **AICTE sponsored Quality Improvement Programme (QIP) Postgraduate Program** in "Artificial Intelligence and Data Science" at the Indian Institute of Information Technology (IIIT) Kottayam. The academic exposure, research-oriented environment, and institutional support provided through this program have been instrumental in the successful completion of this work.

The author also extend their heartfelt appreciation to the **Indian Institute of Information Technology (IIIT) Kottayam** for providing excellent infrastructure, resources, and a stimulating academic atmosphere that facilitated the execution of this research project.

The author also extends deepest gratitude to Dr. Sushitha Susan Joseph and Dr. Selvi for their invaluable guidance, constant encouragement, insightful suggestions, and expert supervision

throughout the course of this research. The mentorship, constructive feedback, and unwavering support significantly contributed to the successful completion of this project and the preparation of this manuscript.

The author further acknowledges all faculty associated with the AICTE QIP PG Program at IIIT Kottayam for their support, cooperation, and encouragement. Their contributions have been greatly appreciated and have played an important role in the completion of QIP PG research work.

REFERENCES

1. Q. Yang *et al.*, "Learning-based speech analysis for Alzheimer's disease detection: A review," *Alzheimer's Research & Therapy*, vol. 14, no. 186, 2022.
2. K. Ding, S. Chetty, and D. Burns, "Detecting Alzheimer's disease from speech: A survey of AI techniques, datasets and challenges," *Artificial Intelligence Review*, vol. 57, 2024.
3. J. U. Bang, S. Kim, and Y. M. Ro, "Spontaneous speech can help recognize Alzheimer's disease," *ETRI Journal*, vol. 46, no. 4, pp. 811–822, 2024.
4. H. Cai *et al.*, "Using speech transcripts and audio to detect Alzheimer's disease," *arXiv preprint arXiv:2307.02514*, 2023.
5. K. Ahn, J. Kim, and J. Kim, "Deep learning helps detect Alzheimer's disease through speech data," *Bioengineering*, vol. 10, no. 9, p. 1093, 2023.
6. B. S. Runde, S. Alapati, and R. S. Desarkar, "Improving Alzheimer's disease detection from speech with natural language processing," *Brain Sciences*, vol. 14, no. 3, p. 211, 2024.
7. M. Oiza-Zapata and J. Gallardo-Antolín, "Alzheimer's disease detection from speech using Shapley additive explanations for feature selection and enhanced interpretability," *Electronics*, vol. 14, no. 11, 2025.
8. J. Lee, Y. Kim, and H. Park, "Multimodal Alzheimer's disease recognition from image, text and audio," *Scientific Reports*, vol. 15, 2025.
9. A. Azadmaleki, M. Shahin, and D. Blackburn, "SpeechCARE: Dynamic multimodal modeling for cognitive screening in diverse linguistic and speech task contexts," *npj Digital Medicine*, vol. 8, 2025.
10. M. Favaro, A. Costa, and G. Ricci, "Advancing Alzheimer's disease detection via multimodal fusion of speech and eye movement data," *Journal of Alzheimer's Disease*, vol. 103, no. 1, pp. 1–15, 2025.
11. M. Ksibi, S. Ben Salem, and H. Amiri, "Multimodal Siamese networks for dementia detection from speech in women," *Scientific Reports*, vol. 15, 2025.
12. S. Flick and R. Ostrand, "Automatically calculated context-sensitive features of connected speech improve prediction of impairment in Alzheimer's disease," *Journal of Speech, Language, and Hearing Research*, vol. 68, no. 2, 2025.
13. A. Llaca-Sánchez, J. Martínez, and P. Gómez, "Exploring LLM embedding potential for dementia detection using audio transcripts," *Applied Sciences*, vol. 15, no. 4, 2025.
14. J. Ortiz-Perez, L. Hernandez, and M. Cruz, "CogniAlign: Word-level multimodal speech alignment with gated cross-attention for Alzheimer's detection," *Knowledge-Based Systems*, vol. 312, 2025.
15. D. Pérez-Toro, M. Hernández, and A. López, "Automated speech markers of Alzheimer dementia: Test of cross-linguistic generalizability," *Alzheimer's & Dementia*, vol. 21, no. 1, 2025.
16. Y. Wang, X. Li, and H. Zhang, "Speech digital biomarker combined with fluid biomarkers predict cognitive impairment through machine learning," *Alzheimer's Research & Therapy*, vol. 17, no. 1, 2025.
17. H. Goodglass, E. Kaplan, and B. Barresi, Boston Diagnostic Aphasia Examination, 3rd ed. Philadelphia, PA, USA: Lippincott Williams & Wilkins, 2001. The Cookie Theft Picture Description Task is one of the standard elicitation tasks within the BDAE.