

Gated Multi-Graph Fusion via Graph Attention Networks for Alzheimer’s Disease Detection

Jinyu Li¹, Xiao Wei^{1,2}, Bin Wen^{1,2}, Kai Li², Yuqin Lin³, Xiaobao Wang¹,
Longbiao Wang^{1,4,*}, Jianwu Dang^{2,*}

¹ School of Future Technology, Tianjin University, Tianjin, China

² Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

³ College of Computer and Data Science, Fuzhou University, Fuzhou, China

⁴ Huiyan Technology (Tianjin) Co., Ltd, Tianjin, China

lijinyu536@tju.edu.cn, longbiao-wang@tju.edu.cn

Abstract

Spontaneous speech is a vital non-invasive biomarker for Alzheimer’s Disease (AD), yet many systems overlook non-linear structural disruptions and clinical heterogeneity in pathological language. We propose a Multi-View Gated Graph Attention Network that transcribes audio via Automatic Speech Recognition (ASR) to construct semantic, dependency, and co-occurrence graphs, characterizing speech through a “content–structure–flow” framework. Notably, the co-occurrence graph leverages Pointwise Mutual Information (PMI) from a normative corpus to quantify narrative logic and linguistic deviation. To address symptomatic diversity, an adaptive gated fusion mechanism dynamically integrates these views. Evaluated on the ADReSSo dataset, our model achieves 90.00% accuracy. Ablation results confirm that the PMI-based graph and heterogeneity-aware gating are essential for robust classification across diverse clinical populations. Our source code is publicly available at <https://github.com/opecac/AD>.

Index Terms: Alzheimer’s Disease Detection, Graph Neural Networks, Spontaneous Speech Analysis, Gated Fusion Mechanism

1. Introduction

Dementia, primarily AD, represents an escalating global health crisis hallmarked by progressive cognitive decline that manifests early in spontaneous speech [1, 2]. The “Cookie Theft” picture description task is extensively recognized for its clinical utility in eliciting these critical linguistic markers in a controlled environment [3]. Initial efforts in speech-based dementia detection predominantly relied on traditional machine learning paradigms [4, 5]. These methodologies utilized manual feature engineering to extract handcrafted acoustic parameters—such as pitch and jitter—alongside lexical diversity metrics, which were subsequently classified using algorithms like Support Vector Machines (SVM) or Random Forests [6, 7]. While these models offered high interpretability [8], they were constrained by their dependence on expert-defined features and their inherent inability to capture the latent, non-linear dependencies found in spontaneous discourse. The subsequent evolution of deep learning, specifically the application of Large Language Models (LLMs) such as BERT, has substantially improved diagnostic performance [9, 10]. Nonetheless, these models primarily learn semantic representations implicitly [11, 12, 13] and often fail to sufficiently characterize the complex, multi-

dimensional structural degradation characteristic of pathological speech. Recently, researchers have explored the use of Graph Neural Networks (GNNs) to model the intricate topological structure of language, yielding promising results [14, 15]. However, current research in this domain remains insufficient in its modeling of linguistic features. Furthermore, existing frameworks typically rely on simplistic fusion strategies (e.g., concatenation or fixed-weight summation), limiting the model’s adaptability to diverse symptomatic manifestations across the clinical spectrum.

In this paper, we propose a multi-view graph learning framework that addresses two critical gaps in current research through the following innovations: First, we introduce a “content–structure–flow” trinity for comprehensive linguistic modeling. Traditional methods often overlook the “flow” of discourse—the logical progression of speech. While semantic graphs capture “content” and dependency graphs capture “structure,” we specifically design a Co-occurrence Graph based on PMI derived from healthy normative data. This graph uniquely reflects a subject’s ability to describe local events. In patients with AD, speech often exhibits disordered sequences, logical jumps, and repetitive loops [16]. By integrating this co-occurrence perspective, our model provides a full-spectrum characterization of linguistic patterns, capturing not just what is said, but how the narrative logic unfolds. Second, we propose a Gated Fusion Mechanism to address the clinical heterogeneity of AD. Clinical observations demonstrate that dementia symptoms are not monolithic: some patients exhibit “syntactic collapse” (simplified grammar with preserved vocabulary), while others suffer from “semantic void” (fluent but meaningless speech) [17]. Our gated network adaptively assigns weights to the semantic, dependency, and co-occurrence representations on a per-sample basis. This allows the framework to capture sample-specific features, dynamically focusing on the most discriminative linguistic markers for each individual.

In summary, our contributions are as follows:

1. Discourse Flow Analysis: We utilize a PMI-based co-occurrence graph to quantify the deviation of event description logic from healthy norms.
2. A Holistic Multi-Graph Framework: We integrate semantic, dependency, and co-occurrence graphs to model the “content–structure–flow” of spontaneous speech.
3. Heterogeneity-aware Fusion: We implemented a gated fusion mechanism that accounts for the diversity of AD symptoms, enhancing the model’s ability to adapt to varying symptomatic presentations.

*indicates the corresponding authors.

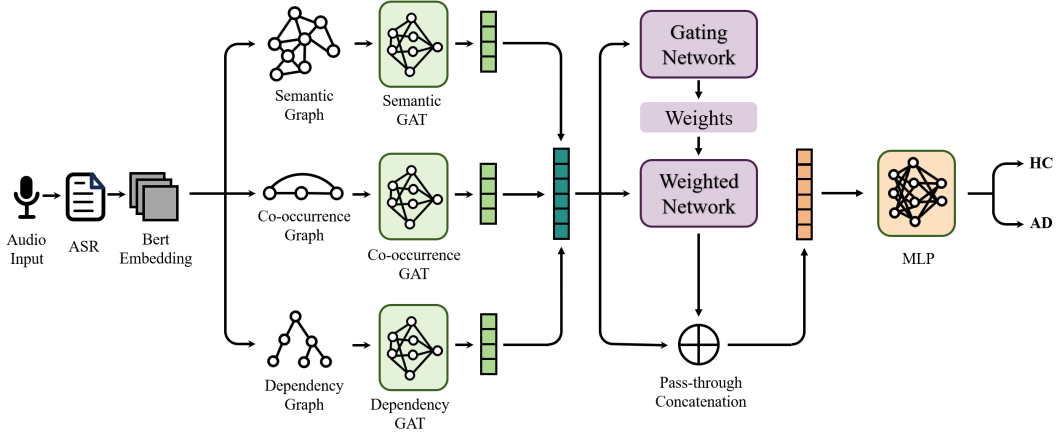


Figure 1: **Overview of the proposed Multi-View Gated Graph Attention Network framework for dementia detection.** (1) Audio is transcribed via ASR and converted into word-level embeddings. (2) Semantic, co-occurrence, and dependency graphs are constructed to capture multi-scale linguistic patterns. (3) Topological features are extracted by GAT layers and aggregated via global pooling. (4) These multi-view features are adaptively fused by a gating network. (5) The fused features are fed into an MLP for final classification.

2. Multi-View Gated GAT

We propose a multi-dimensional graph learning framework, as illustrated in Figure 1, designed to capture the “content–structure–flow” of spontaneous speech. This architecture consists of five main modules as detailed below.

2.1. Speech Transcription and Node Embedding

Automatic Speech Recognition (ASR): Given the raw audio signal of a subject describing the “Cookie Theft” picture, we utilize Whisper [18] for transcription. Whisper’s large-scale weak supervision training makes it exceptionally robust to the disfluencies common in dementia-impacted speech. The output is a sequence of N words: $S = \{w_1, w_2, \dots, w_N\}$.

Node Feature Initialization: To represent each word in a high-dimensional semantic space, we use a pre-trained BERT-base [19] model. Since BERT employs a WordPiece tokenizer, a word w_i may be split into k sub-word tokens $\{t_{i,1}, t_{i,2}, \dots, t_{i,k}\}$. To maintain a 1-to-1 mapping between the transcript and graph nodes, we define the node representation $\mathbf{x}_i \in \mathbb{R}^{768}$ as the mean of its constituent token embeddings:

$$\mathbf{x}_i = \text{MeanPooling}(\text{BERT}(t_{i,1}), \dots, \text{BERT}(t_{i,k})). \quad (1)$$

The final node feature matrix for a transcript is $\mathbf{X} \in \mathbb{R}^{N \times 768}$.

2.2. Multi-View Graph Construction

We define three graphs $\mathcal{G}_{sem}$, $\mathcal{G}_{syn}$, and $\mathcal{G}_{co}$ sharing the same vertex set $\mathcal{V}$, where $|\mathcal{V}| = N$.

2.2.1. Semantic Graph: Content Representation

This graph models the conceptual density and semantic relationships. We compute the cosine similarity between every pair of word embeddings $\mathbf{x}_i$ and $\mathbf{x}_j$ [20]. An edge $e_{ij} \in \mathcal{E}_{sem}$ is created if the similarity exceeds a predefined threshold τ_s :

$$\text{Sim}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|} > \tau_s. \quad (2)$$

This captures the global semantic structural characteristics of the subject.

2.2.2. Dependency Graph: Structural Integrity

To capture grammatical decay (e.g., loss of complex clauses), we use spaCy [21] to perform dependency parsing. An edge $e_{ij} \in \mathcal{E}_{syn}$ is established if there is a direct syntactic dependency (e.g., nominal subject, direct object, modifier) between w_i and w_j . This graph reflects the structural complexity of the subject’s language [22, 23].

2.2.3. Co-occurrence Graph: Discourse Flow via Normative PMI

To capture the “flow” dimension of the “content–structure–flow” trinity, we design a co-occurrence graph that quantifies the logical progression and temporal coherence of speech. Traditional sequential models often fail to detect the subtle “logical jumps” or repetitive loops characteristic of AD-impacted discourse. Our approach addresses this by measuring how much a subject’s word associations deviate from established healthy linguistic norms.

We first construct a Normative Corpus $\mathcal{D}_{hc}$ using only transcripts from healthy controls. This corpus serves as a linguistic reference baseline for characterizing natural word associations within the context of the ‘Cookie Theft’ task. We calculate the Pointwise Mutual Information (PMI) [24, 25] for all word pairs (w_i, w_j) that appear within a sliding window of size n across $\mathcal{D}_{hc}$:

$$\text{PMI}(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}, \quad (3)$$

where $P(w_i)$ and $P(w_j)$ are the individual unigram probabilities, and $P(w_i, w_j)$ is the joint probability of co-occurrence within the window. A high PMI indicates a strong, statistically significant association between words (e.g., “sink” and “overflowing”) that reflects healthy narrative logic.

For a specific subject’s transcript S , we construct the co-occurrence graph $\mathcal{G}_{co}$ by mapping the pre-calculated normative weights onto the subject’s actual word sequence. An edge $e_{ij} \in \mathcal{E}_{co}$ is established between word w_i and any subsequent word w_{i+k} (where $1 \leq k \leq n$) if and only if their normative PMI exceeds a predefined threshold τ_c .

The structural density inherent in the co-occurrence graph $\mathcal{G}_{co}$ functions as a direct proxy for discourse coherence, effectively capturing the nuanced temporal "flow" of spontaneous speech. Within a healthy flow scenario—where a subject adheres to a logical and conventional narrative trajectory—the resulting $\mathcal{G}_{co}$ is defined by a high density of edges paired with significant PMI weights. Together, these elements construct a robustly well-connected topological "skeleton" composed of normative word associations. Conversely, the pathological disruption observed in patients with AD—which manifests as labored word retrieval or disjointed logical transitions—produces word pairs that are statistically rare or nonsensical compared to the normative baseline. This ultimately yields a fragmented graph topology marked by sparse or anomalous connectivity.

2.3. View-Specific Encoding via GAT

To learn the importance of different neighbors within each graph, we employ a Graph Attention Network (GAT) [26]. For a node i in graph $k \in \{sem, syn, co\}$, the attention coefficient α_{ij} is computed as:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^\top [\mathbf{W}\mathbf{x}_i \parallel \mathbf{W}\mathbf{x}_j]))}{\sum_{l \in \mathcal{N}_i} \exp(\text{LeakyReLU}(\mathbf{a}^\top [\mathbf{W}\mathbf{x}_i \parallel \mathbf{W}\mathbf{x}_l]))}, \quad (4)$$

where $\mathbf{a}$ and $\mathbf{W}$ are learnable parameters and $\mathcal{N}_i$ is the neighborhood of node i . The updated node features $\mathbf{h}'_i$ are aggregated to form a global graph representation $\mathbf{z}_k$ using Global Mean Pooling:

$$\mathbf{z}_k = \frac{1}{N} \sum_{i=1}^N \mathbf{h}'_i, \quad \mathbf{z}_k \in \mathbb{R}^{d_{gat}}. \quad (5)$$

2.4. Heterogeneity-Aware Gated Fusion

Given that AD symptoms vary (e.g., some lose syntax, others lose semantic logic), we use a gated mechanism to dynamically weight the three views [27]. First, we concatenate the three graph vectors: $\mathbf{Z}_{cat} = [\mathbf{z}_{sem} \parallel \mathbf{z}_{syn} \parallel \mathbf{z}_{co}]$. The gating network computes a weight vector $\mathbf{g}$:

$$\mathbf{g} = \text{Softmax}(\mathbf{W}_g \mathbf{Z}_{cat} + \mathbf{b}_g), \quad \mathbf{g} = [\beta_{sem}, \beta_{syn}, \beta_{co}]. \quad (6)$$

The fused representation $\mathbf{z}_{fused}$ is a weighted sum:

$$\mathbf{z}_{fused} = \beta_{sem} \mathbf{z}_{sem} + \beta_{syn} \mathbf{z}_{syn} + \beta_{co} \mathbf{z}_{co}. \quad (7)$$

To prevent the loss of individual view details during fusion, we apply a pass-through concatenation:

$$\mathbf{z}_{final} = [\mathbf{z}_{fused} \parallel \mathbf{z}_{sem} \parallel \mathbf{z}_{syn} \parallel \mathbf{z}_{co}]. \quad (8)$$

This ensures that the classifier has access to both the "optimized mixture" and the raw specific features of each linguistic dimension.

2.5. Classification and Objective Function

The final vector $\mathbf{z}_{final}$ is passed through a Multi-Layer Perceptron (MLP). To classify the subject as either AD or Healthy Control (HC), the model outputs a predicted probability $\hat{y} = \text{Sigmoid}(\text{MLP}(\mathbf{z}_{final}))$. To improve the model's generalization and prevent over-confident predictions, we employ Label Smoothing during training. Instead of using hard binary labels $y \in \{0, 1\}$, we transform them into soft targets y_{ls} :

$$y_{ls} = y(1 - \alpha) + \frac{\alpha}{K}, \quad (9)$$

where $\alpha = 0.2$ is the smoothing parameter and $K = 2$ is the number of classes. The model is then optimized using the Smoothed Binary Cross-Entropy (BCE) Loss:

$$\mathcal{L} = -\frac{1}{M} \sum_{m=1}^M [y_{ls,m} \log(\hat{y}_m) + (1 - y_{ls,m}) \log(1 - \hat{y}_m)]. \quad (10)$$

This approach encourages the model to learn more robust latent representations by penalizing extreme logit values, thereby enhancing both calibration and classification performance on the clinical dataset.

3. Experiments

3.1. Experimental settings

3.1.1. Dataset

Our experiments were conducted using the standardized ADReSSo 2021 Challenge dataset [28], which is a curated subset of the Pitt Corpus within the DementiaBank database. Samples were collected via the "Cookie Theft" picture description task—a clinical gold standard for assessing narrative speech and cognitive status. The corpus comprises audio recordings and transcripts from 237 English speakers, including 122 individuals diagnosed with AD and 115 HC. Following the ADReSSo 2021 protocol, the data is partitioned into a training set of 166 participants (87 AD, 79 HC) and a test set of 71 participants (35 AD, 36 HC). We cropped the raw audio files based on provided timestamps to retain only segments containing subject speech. Since our approach constructs graph nodes directly from word-level tokens in subject narratives, samples lacking valid subject-specific segments—specifically five from the training set and one from the test set—were excluded as they could not support graph construction. To ensure experimental fairness and rigorous comparison, we re-implemented all baseline methods and evaluated them on this filtered version of the dataset.

3.1.2. Implementation details

All experiments were conducted on an NVIDIA GeForce RTX 4090D GPU. For the model architecture, the input dimension was set to 768 to match the BERT-base embeddings. The GAT component consisted of a single layer with 2 attention heads and a hidden dimension of 128. The subsequent MLP utilized a hidden layer of 256 units. To prevent overfitting, we applied a dropout rate of 0.5 and a weight decay of 0.003. Regarding the hyperparameter configuration, the model was trained for a maximum of 100 epochs with a batch size of 8. We employed the Adam optimizer with an initial learning rate of 1×10^{-3} , coupled with a ReduceLROnPlateau scheduler. Furthermore, we implemented label smoothing of 0.2 to improve the model's calibration and robustness. Through extensive experimentation, we determined the optimal hyperparameters for graph construction: a sliding window size of $n = 3$ and an edge addition threshold of $\tau_c = 0.3$ for the co-occurrence graph, and a threshold of $\tau_s = 0.8$ for the semantic graph, which collectively yielded the best performance in our final results.

3.2. Main results and discussions

As illustrated in Table 1, our proposed framework demonstrates superior performance, consistently outperforming all baseline methods across all metrics. Specifically, it achieves a 5-fold cross-validation accuracy of 88.81% on the training set and an accuracy of 90.00% on the test set.

Table 1: Performance comparison of different methods on 5-fold cross-validation and test set (%).

Method	5-Fold Cross-Validation Results				Test Set Results			
	Accuracy	F1-Score	Recall	Precision	Accuracy	F1-Score	Recall	Precision
Luz et al. [28]	78.22	77.96	73.40	84.17	79.71	78.12	73.53	83.33
Balogopalan et al. [11]	80.78	82.24	83.79	81.28	82.86	83.33	85.71	81.08
Ajrouti et al. [12]	85.08	85.58	83.86	88.43	81.43	81.16	80.00	82.35
Cai et al. [14]	84.38	85.54	86.14	85.10	84.29	83.58	80.00	87.50
Ortiz-Perez et al. [29]	86.88	86.71	87.31	86.79	81.43	80.60	77.14	84.38
Ours	88.81	88.81	88.81	89.43	90.00	89.86	88.57	91.18

Compared to the baseline established by Luz et al. , our model demonstrates a substantial performance leap. In contrast to Ajrouti et al., who utilized OpenAI’s text embeddings as high-dimensional features for direct classification, our approach employs BERT to establish a rigorous word-to-node mapping via token averaging. This strategy leverages deep semantic representations while simultaneously preserving the discrete symbolic structure of the text. Unlike the approach by Balagopalan et al., which utilizes BERT hidden features for direct classification, our multi-view gated graph attention network framework enables structural linguistic modeling by characterizing speech patterns through complementary perspectives.

While Cai et al. focused on dependency or dynamic graphs, our integration of a PMI-based co-occurrence graph effectively captures temporal anomalies in speech flow. Pathological markers common in AD patients—such as disjointed logical transitions or labored word retrieval—are explicitly manifested through abnormal fluctuations in PMI weights. Simultaneously, the dependency graph models the integrity of the linguistic “skeleton,” while the semantic graph utilizes cosine similarity to detect subtle semantic drift.

The method proposed by Ortiz-Perez et al. achieved state-of-the-art performance under the cross-validation setting. However, it was observed that this configuration leads to a significant performance degradation when the model is evaluated on the test set. This phenomenon may be attributed to the model overfitting to its respective validation split in each fold, yielding inflated overall metrics. For a more stringent assessment of performance, we evaluated the proposed approach against the test set, yielding an accuracy rate of 81.43%.

3.3. Ablation Study

To evaluate the individual contribution of each architectural component to the diagnostic performance of our proposed method, we conducted a series of ablation experiments on the test set. By systematically removing specific graph views and the fusion mechanism, we generated five ablation variants for comparison. The results are summarized in Table 2.

The most significant performance degradation occurred in the w/o Graph Structure variant, which relies solely on the average pooling of BERT embeddings without any topological modeling. This variant achieved the lowest scores across all metrics (Acc: 81.43%, F1: 80.60%), underscoring the critical importance of capturing non-linear structural dependencies in pathological speech.

Among the specific linguistic views, the removal of the Co-occurrence Graph led to a substantial decline in the F1-score to 85.29%. This confirms that the temporal “flow” and logical progression captured via PMI are indispensable biomarkers

Table 2: Ablation study results on test set (%).

Model Variant	Accuracy	F1-Score
w/o Semantic Graph	85.71	86.11
w/o Dependency Graph	87.14	86.57
w/o Co-occurrence Graph	85.71	85.29
w/o Gated Fusion	87.14	87.32
w/o Graph Structure	81.43	80.60
Ours	90.00	89.86

for Alzheimer’s Disease. Similarly, the removal of the Semantic Graph and Dependency Graph resulted in accuracy drops to 85.71% and 87.14%, respectively. These results underscore that while semantic content provides the “substance” of the diagnosis, the syntactic “skeleton” provides the necessary linguistic constraints for robust classification.

The w/o Gated Fusion variant, which utilizes a static averaging of the three graph outputs instead of the dynamic weighting network, saw its accuracy fall to 87.14%. This confirms that a “one-size-fits-all” approach to linguistic feature integration is insufficient for AD detection. The gated network’s ability to adaptively prioritize different linguistic views based on individual patient speech patterns—such as prioritizing syntactic errors in one patient and semantic incoherence in another—is a critical factor in achieving the full model’s accuracy of 90.00%.

4. Conclusion

Experimental evaluations on the ADReSSo 2021 dataset validate that the Multi-View Gated Graph Attention Network provides a robust framework for identifying AD by holistically characterizing the “content-structure-flow” features of spontaneous speech. One of the primary contributions of this study is the pivotal role played by the co-occurrence graph; by utilizing normative PMI to quantify discourse logic, it successfully captures the incoherent logical transitions and repetitive patterns characteristic of cognitive decline that traditional sequential models are typically unable to detect. Ablation results indicate that removing this “flow” dimension significantly impairs diagnostic performance, confirming that narrative progression is a crucial feature. Furthermore, unlike conventional static fusion strategies, our gated network adaptively assigns weights to linguistic features on a per-sample basis, effectively mitigating the challenges posed by the inherent clinical heterogeneity of AD. Future research will prioritize evolving the framework into a comprehensive multimodal system by integrating acoustic prosody as additional graph-based perspectives.

5. Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant U23B2053 and by the National Talent Program under Grants E4G008, E55304, E43301, and E476.

6. Generative AI Use Disclosure

During the preparation of this manuscript, we used generative AI tools to polish the English language and improve readability. These tools were not used to generate any scientific claims, experimental results, or significant parts of the manuscript.

7. References

- [1] H. Lindsay, J. Tröger, and A. König, “Language Impairment in Alzheimer’s Disease—Robust and Explainable Evidence for AD-Related Deterioration of Spontaneous Speech Through Multilingual Machine Learning,” *Frontiers in Aging Neuroscience*, vol. Volume 13 - 2021, 2021.
- [2] M. Gumus, M. Koo, C. M. Studzinski, A. Bhan, J. Robin, and S. E. Black, “Linguistic changes in neurodegenerative diseases relate to clinical symptoms,” *Frontiers in Neurology*, vol. 15, p. 1373341, Mar. 2024.
- [3] X. Qi, Q. Zhou, J. Dong, and W. Bao, “Noninvasive automatic detection of Alzheimer’s disease from spontaneous speech: a review,” *Frontiers in Aging Neuroscience*, vol. 15, p. 1224723, Aug. 2023.
- [4] Y. Qiao, X. Yin, D. Wiechmann, and E. Kerz, “Alzheimer’s Disease Detection from Spontaneous Speech Through Combining Linguistic Complexity and (Dis)Fluency Features with Pretrained Language Models,” in *Interspeech 2021*. ISCA, Aug. 2021, pp. 3805–3809.
- [5] M. Kurdi, “Automatic diagnosis of alzheimer’s disease using lexical features extracted from language samples,” *Journal of Medical Artificial Intelligence*, vol. 7, pp. 13–13, Jun. 2024.
- [6] Z. S. Syed, M. S. S. Syed, M. Lech, and E. Pirogova, “Tackling the ADRESSO Challenge 2021: The MUET-RMIT System for Alzheimer’s Dementia Recognition from Spontaneous Speech,” in *Interspeech 2021*. ISCA, Aug. 2021, pp. 3815–3819.
- [7] R. Shankar, Z. Goh, F. Devi, and Q. Xu, “A systematic review of explainable artificial intelligence methods for speech-based cognitive decline detection,” *npj Digital Medicine*, vol. 8, no. 1, p. 724, Nov. 2025.
- [8] L. Calzà, G. Gagliardi, R. Rossini Favretti, and F. Tamburini, “Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia,” *Computer Speech & Language*, vol. 65, p. 101113, Jan. 2021.
- [9] A. Balagopalan and J. Novikova, “Comparing Acoustic-Based Approaches for Alzheimer’s Disease Detection,” in *Interspeech 2021*. ISCA, Aug. 2021, pp. 3800–3804.
- [10] X. Wei, B. Wen, Y. Lin, K. Li, M. Gu, X. Wang, L. Wang, and J. Dang, “Breaking Data Efficiency Dilemma: A Federated and Augmented Learning Framework For Alzheimer’s Disease Detection via Speech,” Feb. 2026, arXiv:2602.14655 [cs].
- [11] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, “To BERT or not to BERT: Comparing Speech and Language-Based Approaches for Alzheimer’s Disease Detection,” in *Interspeech 2020*. ISCA, Oct. 2020, pp. 2167–2171. [Online]. Available: https://www.isca-archive.org/interspeech_2020/balagopalan20.interspeech.html
- [12] K. Ajroudi, M. I. Khedher, O. Jemai, and M. A. El-Yacoubi, “Exploring the Efficacy of Text Embeddings in Early Dementia Diagnosis from Speech,” in *2024 16th International Conference on Human System Interaction (HSI)*, Jul. 2024, pp. 1–6, iSSN: 2158-2254.
- [13] K. Chlasta, P. Struzik, and G. M. Wójcik, “Enhancing dementia and cognitive decline detection with large language models and speech representation learning,” *Frontiers in Neuroinformatics*, vol. 19, Dec. 2025.
- [14] H. Cai, X. Huang, Z. Liu, W. Liao, H. Dai, Z. Wu, D. Zhu, H. Ren, Q. Li, T. Liu, and X. Li, “Exploring Multimodal Approaches for Alzheimer’s Disease Detection Using Patient Speech Transcript and Audio Data,” Jul. 2023, arXiv:2307.02514 [eess].
- [15] A. E. Hallani, A. Chakhtouna, and A. Adib, “Graph Attention Networks with Dual-Edge Connectivity for Alzheimer’s Disease Detection from Speech,” *IEEE Transactions on Artificial Intelligence*, pp. 1–11, 2025.
- [16] E. Burke, J. Gunstad, and P. Hamrick, “Comparing global and local semantic coherence of spontaneous speech in persons with Alzheimer’s disease and healthy controls,” *Applied Corpus Linguistics*, vol. 3, no. 3, p. 100064, Dec. 2023.
- [17] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, “Linguistic Features Identify Alzheimer’s Disease in Narrative Speech,” *Journal of Alzheimer’s Disease*, vol. 49, no. 2, pp. 407–422, Nov. 2015.
- [18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 28492–28518.
- [19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.
- [20] Y. Chen, L. Wu, and M. J. Zaki, “Iterative Deep Graph Learning for Graph Neural Networks: Better and Robust Node Embeddings,” Oct. 2020, arXiv:2006.13009 [cs].
- [21] M. Honnibal and I. Montani, *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*, 2017, software available from <https://spacy.io/>.
- [22] O. Ivanova, I. Martínez-Nicolás, E. García-Piñuela, and J. J. G. Meilán, “Defying syntactic preservation in Alzheimer’s disease: what type of impairment predicts syntactic change in dementia (if it does) and why?” *Frontiers in Language Sciences*, vol. 2, p. 1199107, Aug. 2023.
- [23] Z. Lian and Z. Wang, “Dependency Grammar Approach to the Syntactic Complexity in the Discourse of Alzheimer Patients,” *Behavioral Sciences*, vol. 15, no. 10, Sep. 2025.
- [24] K. W. Church and P. Hanks, “Word Association Norms, Mutual Information, and Lexicography,” *Computational Linguistics*, vol. 16, no. 1, pp. 22–29, 1990.
- [25] L. Yao, C. Mao, and Y. Luo, “Graph Convolutional Networks for Text Classification,” Nov. 2018, arXiv:1809.05679 [cs].
- [26] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph Attention Networks,” Feb. 2018, arXiv:1710.10903 [stat].
- [27] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, “Gated Multimodal Units for Information Fusion,” Feb. 2017, arXiv:1702.01992 [stat].
- [28] S. Luz, F. Haider, S. D. L. Fuente, D. Fromm, and B. MacWhinney, “Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge,” in *Interspeech 2021*. ISCA, Aug. 2021, pp. 3780–3784.
- [29] D. Ortiz-Perez, M. Benavent-Lledo, J. Rodríguez-Juan, J. García-Rodríguez, and D. Tomás, “CogniAlign: Word-level multimodal speech alignment with gated cross-attention for Alzheimer’s detection,” *Knowledge-Based Systems*, vol. 329, p. 114264, Nov. 2025.