SpeechHGT: A Multimodal Hypergraph Transformer for Speech-Based Early Alzheimer's Disease Detection

Shagufta Abid[1], **Dongyu Zhang**[1,*], **Ahsan Shehzad**[1], **Jing Ren**[2], **Shuo Yu**[1,3], **Hongfei Lin**[1], **Feng Xia**[2]
[1]Dalian University of Technology, China
[2]RMIT University, Australia
[3]Key Laboratory of Social Computing and Cognitive Intelligence, Ministry of Education, China
{shagufta.abid,ahsan.shehzad}@outlook.com, {zhangdongyu,hflin}@dlut.edu.cn, {shuo.yu, jing.ren, f.xia}@ieee.org

Abstract

Early detection of Alzheimer's disease (AD) through spontaneous speech analysis represents a promising, non-invasive diagnostic approach. Existing methods predominantly rely on fusion-based multimodal deep learning, effectively integrating linguistic and acoustic features. However, these methods inadequately model higher-order interactions between modalities, reducing diagnostic accuracy. To address this, we introduce SpeechHGT, a multimodal hypergraph transformer designed to capture and learn higher-order interactions in spontaneous speech features. SpeechHGT encodes multimodal features as hypergraphs, where nodes represent individual features and hyperedges represent grouped interactions. A novel hypergraph attention mechanism enables robust modeling of both pairwise and higher-order interactions. Experimental evaluations on the DementiaBank datasets reveal that SpeechHGT achieves state-of-the-art performance, surpassing baseline models in accuracy and F1 score. These results highlight the potential of hypergraph-based models to improve AI-driven diagnostic tools for early AD detection.

1 Introduction

Early diagnosis of Alzheimer's disease (AD) is crucial for timely intervention and improved patient outcomes [Alberdi *et al.*, 2016; Shehzad *et al.*, 2025]. AD is a neurodegenerative disorder characterized by memory loss, cognitive decline, and behavioral changes [Marvi *et al.*, 2024; Zhang *et al.*, 2024]. While neuroimaging techniques like MRI and PET can detect brain alterations, their utility is limited by high costs, restricted accessibility, and radiation exposure from repeated PET scans [Ahmed *et al.*, 2019; Yu *et al.*, 2024; Yang *et al.*, 2022]. Consequently, there is a growing need for cost-effective, non-invasive diagnostic methods [Petti *et al.*, 2020; Ding *et al.*, 2024]. Speech analysis shows promise for early AD detection, leveraging both linguistic (e.g., word choice, syntactic complexity) and acoustic (e.g., speech rate, pitch) features [Pulido *et al.*, 2020; Pacheco-Lorenzo *et al.*, 2024]. However, effectively integrating these diverse speech features to capture the intricate patterns of cognitive decline necessitates further research and advanced modeling approaches.

Previous methods in speech-based AD detection can be categorized into unimodal and multimodal approaches, each employing distinct computational methodologies [Latif *et al.*, 2021; Shehzad *et al.*, 2024]. Acoustic methods utilize prosodic features, such as pitch, formant frequencies, and temporal variations, to identify vocal anomalies linked to AD-related neurodegeneration [Luz *et al.*, 2024; Zhang *et al.*, 2021]. Linguistic approaches focus on lexical, syntactic, and semantic features, examining word frequency, sentence structure, and narrative coherence to detect cognitive impairments. However, unimodal methods often neglect cross-modal interactions, leading to incomplete assessments and reduced diagnostic accuracy. Therefore, multimodal architectures integrating information from multiple sources are adopted [Vrindha *et al.*, 2023; Venugopalan *et al.*, 2021]. These systems integrate acoustic and linguistic representations through hierarchical fusion strategies [Turrisi *et al.*, 2024]. Recent advances employ graph transformers to model complex intermodal relationships in multimodal data, enhancing diagnostic performance [Ektefaie *et al.*, 2023; Peng *et al.*, 2024]. By leveraging attention mechanisms, these models effectively capture intricate feature dependencies, demonstrating significant potential for improving AD classification from spontaneous speech [Bessadok *et al.*, 2022].

Despite advancements, current multimodal speech analysis techniques often miss crucial, complex interactions between linguistic and acoustic features—interactions vital for early Alzheimer's disease (AD) detection [Ying *et al.*, 2023; Priyadarshinee *et al.*, 2023]. These interactions, like the interplay of pitch, tempo, and prosody, reflect cognitive impairments in AD. Traditional methods, such as linear aggregation, typically treat these features independently, assuming simple additive relationships [Ilias and Askounis, 2022a]. This overlooks the inherent non-linear dependencies in speech data [Pérez-Toro *et al.*, 2021], potentially leading to delayed or inaccurate diagnoses. We hypothesize that incorporating

*Corresponding author.

these higher-order interactions will significantly improve diagnostic accuracy, aligning with neurolinguistic theories of cognitive decline [Dell, 1986]. This study aims to validate this hypothesis and enhance non-invasive diagnostic tools for early AD detection.

To address these challenges, we propose SpeechHGT[1], a multimodal hypergraph transformer, to model higher-order interactions between linguistic and acoustic speech features for improved AD detection. SpeechHGT extracts discriminative features from preprocessed audio, representing AD-related speech characteristics. We construct a multimodal hypergraph where nodes denote individual features, and hyperedges capture grouped interactions. This hypergraph integrates both simple edges for pairwise relations and hyperedges for higher-order dependencies. To process this hypergraph-structured data, we design a novel hyperedge attention-based transformer model, which captures both pairwise and higher-order interactions. Transformed node features are aggregated for binary classification, distinguishing AD from speech samples. Experimental results demonstrate that SpeechHGT outperforms baseline models in accuracy and F1-score, offering an effective approach for early AD detection and improved diagnostic reliability.

Our contributions are as follows.

1. We propose SpeechHGT, a novel multimodal hypergraph transformer that captures higher-order interactions between linguistic and acoustic speech features, overcoming the limitations of existing fusion-based approaches for AD detection in capturing complex dependencies.
2. We design a dual-layer hypergraph attention mechanism that effectively models both pairwise and higher-order dependencies, which can improve the integration of multimodal speech features for robust classification.
3. Extensive experiments on multiple real-world datasets show that SpeechHGT outperforms state-of-the-art methods in speech-based AD classification. It achieves higher accuracy, and F1-score on all benchmark datasets, demonstrating its effectiveness in improving early AD diagnosis from spontaneous speech.

2 Related Work

2.1 Speech Analysis for Brain Disease Diagnosis

The diagnosis of speech-based neurodegenerative diseases traditionally relies on acoustic or linguistic representations [Luz *et al.*, 2024]. Acoustic methods analyze prosodic and voice quality characteristics, including pitch contours, speaking rate, and jitter, to identify early markers of AD. [Luz *et al.*, 2020] proposes a standardized acoustic preprocessing framework, demonstrating that prosodic indices alone reveal measurable cognitive impairment. Linguistic methods, in contrast, focus on transcribed speech, examining lexical diversity, syntactic complexity, and semantic coherence. [Searle *et al.*, 2020] shows that advanced language embeddings, such as DistilBERT [Sanh, 2019], can enhance detection accuracy in machine learning models using textual transcripts. Despite their utility, unimodal approaches fail to integrate prosodic and linguistic features, limiting the exploration of holistic speech characteristics that are essential for comprehensive diagnostic assessments.

Current multimodal speech analysis techniques aim to enhance diagnostic accuracy by integrating acoustic and linguistic features using early or late fusion strategies [Ilias and Askounis, 2022b]. [Martinc and Pollak, 2020] shows that optimized combinations of text and audio outperform unimodal approaches in AD detection. Multimodal deep learning models, such as BiLSTM or Transformer architectures, improve feature integration by combining acoustic waveforms with textual transcripts. [Rohanian *et al.*, 2021] demonstrates that gating mechanisms in sequence models align prosodic and lexical-semantic features to enhance predictions. [Zhu *et al.*, 2021] refines semantic embeddings using non-semantic features, like pause duration, via Wav2vec. However, existing methods often fail to model complex intermodal relationships and higher-order dependencies, neglecting critical biomarkers such as semantic confusion and speech disfluencies in AD.

2.2 Graph Transformers

Graph transformers integrate the representational power of graph neural networks (GNNs) with attention-based Transformer mechanisms to model relational data [Liu *et al.*, 2021]. These architectures propagate node-level information across structured connections while using attention coefficients to weight node or edge importance. Recent advancements in protein folding and language modeling demonstrate their ability to address complex relational data domains [Ying *et al.*, 2021]. In clinical research, graph-based methods model disease progression, predict pathological links, and identify biomarkers [Luo *et al.*, 2024]. Their strength lies in capturing long-range dependencies while preserving structural information. However, applying Graph Transformers to multimodal data presents challenges, such as heterogeneous feature spaces and sparse cross-modal relationships [Li *et al.*, 2024].

3 Design of SpeechHGT

3.1 Problem Formulation

This study proposes SpeechHGT, a multimodal hypergraph transformer, for early AD detection using DementiaBank speech data. Speech features include linguistic (F_L) and acoustic (F_A), forming a combined set $F = F_L \cup F_A$. A hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ models higher-order feature interactions, where nodes ($\mathcal{V}$) represent features and hyperedges ($\mathcal{E}$) encode relationships. SpeechHGT learns node (h_v) and hyperedge (h_e) embeddings via hypergraph attention, outputting a binary classification ($y \in \{0, 1\}$) for AD presence. The model optimizes accuracy by minimizing the loss function $\mathcal{L}$. This approach enhances AD detection and provides insights into cognitive decline. Figure 1 illustrates the framework.

[1]The source codes are available at: https://github.com/Ahsan-Shehzad/SpeechHGT.

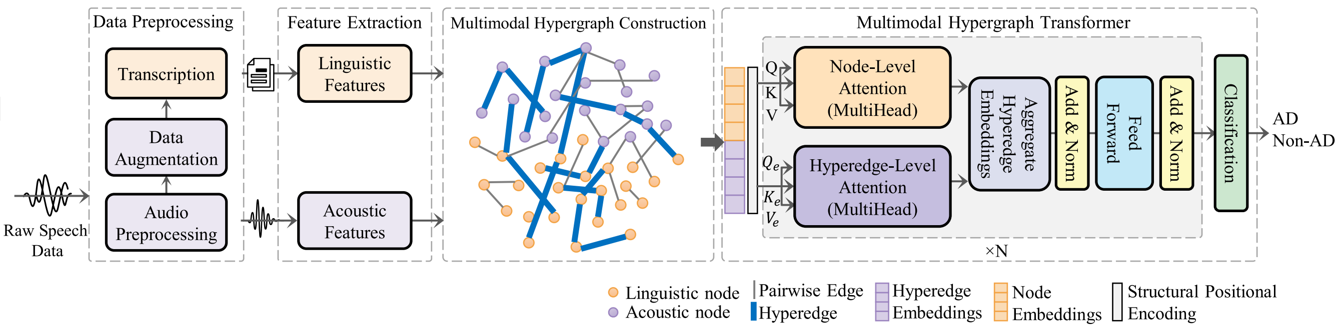

Figure 1: Illustration of SpeechHGT framework.

3.2 Audio Preprocessing

We preprocess raw audio to standardize data quality for reliable feature extraction and classification. This includes noise reduction, volume normalization, segmentation, augmentation, and transcription. All recordings are converted to WAV, resampled to 16 kHz with 16-bit depth, and set to mono using SoX[2]. NoisePy[3] applies spectral subtraction-based noise reduction using STFT-estimated noise spectrum $N(f)$. RMS normalization ensures uniform amplitude. Voice activity detection (VAD) via py-webrtcvad[4] segments recordings. These steps improve analytical robustness.

3.3 Data Augmentation

We apply data augmentation to enhance diversity and model robustness. Speed perturbation adjusts playback speed ($\alpha \in \{0.9, 1.0, 1.1\}$) while preserving pitch. Pitch shifting modifies frequency components via the semitone factor β ($\beta \in \{-2, -1, 1, 2\}$). Gaussian noise $n(t) \sim \mathcal{N}(0, \sigma^2)$ is added, maintaining a 20 dB SNR. Augmentations are implemented using librosa[5].

3.4 Transcription

We use Whisper[6], an ASR model, to transcribe audio $x(t)$ into text T [Liu *et al.*, 2024]. It captures filler words, disfluencies, and unintelligible segments: $T = f_{\text{ASR}}(x(t); \Theta)$, where f_{ASR} is the model and Θ its parameters. This transcription supports linguistic analysis and feature extraction.

3.5 Feature Extraction

Linguistic Feature

We analyze speech transcripts for cognitive and communicative disruptions linked to Alzheimer's Disease (AD). Utilizing advanced Natural Language Processing (NLP) tools like spaCy[7], NLTK[8], and Transformers[9], we extract a wide range of linguistic features. Lexical analysis includes word count, Type-Token Ratio (TTR), Part-of-Speech (POS) tag distributions, Brunet's Index, and Honore's Statistic. Syntactic features encompass sentence complexity, grammatical correctness, parsing tree depth, and clause-to-sentence ratios. Semantic features, derived from contextual embeddings (e.g., BERT), evaluate coherence, semantic similarity, and named entity detection. We use Latent Dirichlet Allocation (LDA) for topic modeling. Discourse analysis involves pausing patterns, pronoun usage, narrative coherence, and topic maintenance. All features are consolidated into a structured vector $\mathbf{L}_i$ for each audio sample.

Acoustic Feature

We analyze audio signals to capture prosodic, articulatory, and spectral properties of speech. Using LibROSA and OpenSMILE[10], we extract phonation features, including jitter, shimmer, Harmonics-to-Noise Ratio (HNR), and Cepstral Peak Prominence (CPP), which reflect vocal stability and clarity. Temporal features, such as speaking rate, silent and filled pause durations, and turn-taking timing, characterize fluency and rhythm. Spectral features, including formant frequencies (F1, F2), Mel-Frequency Cepstral Coefficients (MFCCs), spectral slope, flux, centroid, and bandwidth, describe spectral energy distribution and dynamics. Energy-based features, such as Zero-Crossing Rate (ZCR), intensity contours, sub-band energy distribution, and loudness profiles, quantify energy variations. All acoustic features are encapsulated into a feature vector $\mathbf{A}_i$, providing a comprehensive representation of the audio signal.

3.6 Multimodal Hypergraph Construction

Defining Nodes

Each node in the hypergraph represents a unique linguistic or acoustic feature extracted from the DementiaBank dataset. We define the set of nodes as $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$, where each node v_i corresponds to a specific feature F_i. Each feature is assigned a unique identifier, such as F_1 for Vocabulary Richness or F_2 for Pitch Variability. The attribute vector $\mathbf{a}_i$ for each node v_i includes the feature type (linguistic or acoustic) and its statistical properties, specifically the mean (μ_i) and variance (σ_i^2). Accordingly, the feature vector for node v_i is expressed as:

$$\mathbf{x}_i = [ID(F_i), \text{Type}_i, \mu_i, \sigma_i^2]. \tag{1}$$

[2]http://sox.sourceforge.net/
[3]https://github.com/noisepy/NoisePy
[4]https://github.com/wiseman/py-webrtcvad
[5]https://librosa.org
[6]https://github.com/openai/whisper
[7]https://spacy.io/
[8]https://www.nltk.org/
[9]https://huggingface.co/docs/transformers
[10]https://audeering.com/opensmile/

Identifying Hyperedges

We identify hyperedges to capture higher-order interactions influencing AD detection through statistical correlation analysis and clustering. First, we compute Pearson correlation coefficients (ρ_{ij}) for all feature pairs (F_i, F_j):

$$\rho_{ij} = \frac{\text{cov}(F_i, F_j)}{\sigma_i \sigma_j}. \tag{2}$$

Next, we apply the spectral clustering algorithm to the correlation matrix to cluster features with high inter-correlations. Let $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$ denote the resulting set of clusters. Each cluster C_k corresponds to a hyperedge e_k defined as:

$$e_k = \{v_i \in \mathcal{V} \mid F_i \in C_k\}. \tag{3}$$

This method ensures that hyperedges represent synergistic feature groups with collective relevance to AD detection.

Incorporating Pairwise Edges

We incorporate pairwise edges to capture direct interactions between individual feature pairs, complementing hyperedges. For a given feature pair (F_i, F_j), we compute the mutual information $I(F_i; F_j)$ to quantify feature dependency:

$$I(F_i; F_j) = \sum_{f_i \in F_i} \sum_{f_j \in F_j} p(f_i, f_j) \log \left(\frac{p(f_i, f_j)}{p(f_i)p(f_j)} \right). \tag{4}$$

A pairwise edge is established between nodes v_i and v_j if $I(F_i; F_j)$ exceeds a predefined threshold θ. These pairwise edges capture direct feature dependencies absent in higher-order groupings, enhancing the hypergraph's structural complexity and representational depth.

Representing the Hypergraph

We represent the hypergraph using an incidence matrix $\mathbf{H} \in \{0,1\}^{n \times m}$, where n is the number of nodes and m is the number of hyperedges. Each matrix element $H_{i,k}$ is defined as:

$$H_{i,k} = \begin{cases} 1 & \text{if node } v_i \text{ belongs to hyperedge } e_k \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

We encode nodes and hyperedges with feature vectors to facilitate learning within the hypergraph framework. Each node v_i is represented by a normalized feature vector $\mathbf{x}_i$, defined as:

$$\mathbf{x}_i = \left[\frac{\mu_i - \mu_{\text{min}}}{\mu_{\text{max}} - \mu_{\text{min}}}, \frac{\sigma_i^2 - \sigma_{\text{min}}^2}{\sigma_{\text{max}}^2 - \sigma_{\text{min}}^2}, \text{Type}_i \right]. \tag{6}$$

Hyperedge features are aggregated from constituent node vectors. The aggregated feature vector $\mathbf{y}_k$ for hyperedge e_k is computed as:

$$\mathbf{y}_k = \frac{1}{|e_k|} \sum_{v_i \in e_k} \mathbf{x}_i. \tag{7}$$

This encoding captures collective node information, enabling the multimodal hypergraph transformer (SpeechHGT) to learn complex feature interactions. The hypergraph construction module outputs the incidence matrix $\mathbf{H}$, node features $\{\mathbf{x}_i\}_{i=1}^n$, and hyperedge features $\{\mathbf{y}_k\}_{k=1}^m$.

3.7 Multimodal Hypergraph Transformer

The Hypergraph Transformer constitutes the core of our SpeechHGT architecture, designed to leverage hypergraph-structured data for capturing pairwise and higher-order feature interactions in spontaneous speech.

Hypergraph Attention Mechanism

We design the hypergraph attention mechanism to extend standard self-attention for modeling higher-order interactions. Our approach employs dual attention layers: Node-Level Attention and Hyperedge-Level Attention. The Node-Level Attention captures pairwise node interactions, formulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} \right) V, \tag{8}$$

where Q, K, and V denote the query, key, and value matrices, and d_k represents the key vector dimensionality. To complement this, we implement Hyperedge-Level Attention, which aggregates features across hyperedges to capture higher-order dependencies:

$$\text{Hyperedge-Attention}(Q_e, K_e, V_e) = \text{softmax}\left(\frac{Q_e K_e^\top}{\sqrt{d_e}} \right) V_e. \tag{9}$$

Here, Q_e, K_e, and V_e correspond to hyperedge-specific matrices, with d_e as the dimensionality of hyperedge key vectors. This dual-layer design allows us to capture both direct and collective feature interactions effectively.

Aggregate Hyperedge Embeddings

We represent the collective influence of hyperedges on connected nodes using an iterative embedding update mechanism. Each hyperedge $e \in E$ starts with an initial embedding $\mathbf{h}_e^{(0)}$, which we iteratively refine based on connected node features. The embedding update rule is defined as:

$$\mathbf{h}_e^{(l+1)} = \sigma \left(W_e \cdot \text{Mean} \left(\{ \mathbf{h}_v^{(l)} \mid v \in e \} \right) + b_e \right). \tag{10}$$

Here, W_e and b_e are trainable parameters, σ represents the activation function, and $\mathbf{h}_v^{(l)}$ denotes the node embedding at layer l. This mechanism ensures that hyperedge embeddings effectively capture aggregated information from their associated nodes.

Structural Positional Encodings

We incorporate structural positional encodings to capture nodes' structural relationships and positional contexts within the hypergraph. These encodings preserve structural integrity and enable the learning of positional dependencies. Each node's degree, defined by the number of hyperedges it participates in, is encoded as: $\mathbf{p}_v^{\text{degree}} = \text{Linear}\left(\log(1 + \text{deg}(v))\right)$, where $\text{deg}(v)$ represents the degree of node v, and Linear denotes a linear transformation. We also encode centrality measures, including betweenness ($\beta(v)$) and closeness ($\gamma(v)$) centrality, as: $\mathbf{p}_v^{\text{centrality}} = \text{Linear}\left(\beta(v), \gamma(v)\right)$. The final positional encoding combines degree and centrality encodings:

$$\mathbf{p}_v = \mathbf{p}_v^{\text{degree}} \parallel \mathbf{p}_v^{\text{centrality}}. \tag{11}$$

We integrate these encodings into node representations, enhancing the Transformer's ability to leverage structural information. This approach improves the model's ability to capture nuanced feature interactions essential for Alzheimer's disease detection.

Classification

The Classification module utilizes transformed node features from the Hypergraph Transformer to perform binary classification of AD in speech samples. It converts node embeddings, which capture linguistic and acoustic interactions, into a unified graph-level representation $\mathbf{z}$ through an attention-based readout function:

$$\mathbf{z} = \sum_{v \in V} \alpha_v \mathbf{h}_v, \tag{12}$$

where α_v is the attention weight for node v, calculated as:

$$\alpha_v = \frac{\exp(\mathbf{w}^\top \mathbf{h}_v)}{\sum_{u \in V} \exp(\mathbf{w}^\top \mathbf{h}_u)}. \tag{13}$$

The attention mechanism ensures that significant nodes have a larger influence on $\mathbf{z}$. The aggregated representation $\mathbf{z}$ is then processed through a fully connected layer followed by a sigmoid function to produce the probability $\hat{y}$ of AD:

$$\hat{y} = \sigma\left(\mathbf{W}\mathbf{z} + b\right). \tag{14}$$

The model is optimized using binary cross-entropy loss. The pseudocode of SpeechHGT is given in Algorithm 1.

Algorithm 1 SpeechHGT Algorithm

Input: Raw speech audio $\mathbf{X}$
Output: AD prediction $\hat{y}$

procedure SPEECHHGT($\mathbf{X}$)
  $\mathbf{X}_P \leftarrow$ Preprocess($\mathbf{X}$)
  $\mathbf{X}_A \leftarrow$ Augment($\mathbf{X}_P$)
  $T \leftarrow$ Transcribe($\mathbf{X}_A$)
  $\mathbf{L} \leftarrow$ ExtractLinguisticFeatures(T)
  $\mathbf{A} \leftarrow$ ExtractAcousticFeatures($\mathbf{X}_A$)
  $\mathcal{V} \leftarrow \mathbf{L} \cup \mathbf{A}$
  $\mathcal{E} \leftarrow$ HypergraphConstruction($\mathcal{V}$)
  $\mathbf{H}_0 \leftarrow$ InitializeEmbeddings($\mathcal{V}, \mathcal{E}$)
  $\mathbf{H}_S \leftarrow$ ApplyStructuralEncodings($\mathbf{H}_0, \mathcal{E}$)
  Initialize Transformer weights θ
  $E \leftarrow$ number of training epochs
  for $e \leftarrow 1$ to E **do**
    $\mathbf{H}_N \leftarrow$ NodeLevelAttention($\mathbf{H}_S, \mathcal{E}$)
    $\mathbf{H}_E \leftarrow$ HyperedgeLevelAttention($\mathbf{H}_N, \mathcal{E}$)
  end for
  $\mathbf{Z} \leftarrow$ AggregateEmbeddings($\mathbf{H}_E$)
  $\hat{y} \leftarrow \sigma(W_n \ldots \sigma(W_1 \mathbf{Z} + b_1) \cdots + b_n)$
  Compute loss $\mathcal{L}_{\text{cls}}$
  Update θ using backpropagation
  return $\hat{y}$
end procedure

Dataset	Number of Samples	Average Age (Years)	Gender (M,F))
ADReSS	156 (78 AD + 78 CN)	66.8 AD, 66.8 CN	44.9% M, 55.1% F
ADReSSo	237 (122 AD + 115 CN)	69.38 AD, 66.06 CN	34.9% M, 65.1% F
ADReSS-M	271 (132 AD + 139 CN)	69.9 AD, 66.2 CN	33.6% M, 66.4% F

Table 1: Summary of key features and characteristics of datasets.

4 Experiments

4.1 Experimental Settings

Datasets

We evaluate SpeechHGT using three benchmark datasets from the DementiaBank repository: ADReSS, ADReSSo, and ADReSS-M. These datasets, derived from the Cookie Theft Picture Description Task, are designed for AD detection based on spontaneous speech data. Table 1 summarizes their key characteristics, with details provided below.

- The ADReSS dataset [Luz *et al.*, 2020] includes 156 samples (78 AD, 78 cognitively normal [CN]) with balanced gender representation (44.9% male, 55.1% female) and a mean age of 66.8 years. A 70/30 train-test split is employed, but its small size limits robust training and increases overfitting risk.
- The ADReSSo dataset [Luz *et al.*, 2021] contains 237 samples (122 AD, 115 CN), with a higher female representation (34.9% male, 65.1% female) and greater age variability. The mean ages are 69.38 years (AD) and 66.06 years (CN). Its moderate size and demographic diversity enable model evaluation under variable conditions.
- The ADReSS-M dataset [Luz *et al.*, 2024], the largest, comprises 271 samples (132 AD, 139 CN), with a gender distribution of 33.6% male and 66.4% female. The dataset uses an 80/20 train-test split, offering stability for model training while presenting demographic imbalance challenges.

4.2 Baselines

Challenge Baselines

These correspond to the methodologies established in the ADReSS, ADReSSo, and ADReSS-M challenges, which serve as standardized benchmarks for AD classification and cognitive score prediction.

Conventional Machine Learning Models

We implement Random Forest (RF), Support Vector Machines (SVM), and AdaBoost, focusing on linguistic features extracted from speech data. These models benchmark classical techniques against deep learning methods.

Unimodal Speech Models

These models utilize either linguistic or acoustic features for Alzheimer's detection. For instance, [Searle *et al.*, 2020] employ TF-IDF and DistilBERT embeddings, while [Pérez-Toro *et al.*, 2021] use X-vectors, prosody, and emotional embeddings. We replicate their feature extraction pipelines and training procedures.

Methods		ADReSS Dataset				ADReSSo Dataset				ADReSS-M Dataset			
Type	**Models**	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
Challenge Baselines	Baselines	62.50	63.50	62.50	62.00	78.87	79.00	79.00	77.78	73.91	75.00	68.20	71.40
Conventional ML models	RF	61.25	65.66	58.18	57.05	63.86	72.13	50.57	59.46	60.84	67.74	48.28	56.38
	SVM	59.64	61.90	59.77	60.82	6539	5874	5301	55.84	58.43	60.71	58.62	60.82
	AdaBoost	60.42	63.16	53.33	55.81	64.84	59.30	44.25	50.39	61.23	55.65	32.86	41.15
Unimodal Speech Methods	Searle et al.	81.00	80.50	83.00	85.00	82.16	84.12	79.84	81.32	77.31	81.82	73.56	76.24
	Pérez-Toro et al.	76.20	77.00	75.32	76.15	78.00	88.89	71.43	80.00	75.92	76.94	75.92	75.80
Multimodal Speech Methods	Martinc et al.	77.08	76.50	76.50	77.00	81.67	81.69	81.94	81.69	72.73	73.51	72.73	72.50
	Rohanian et al.	79.17	79.37	79.17	79.13	84.00	83.30	84.16	81.43	70.42	71.72	70.42	69.88
	Zhu et al.	77.08	80.95	70.83	75.56	83.10	83.55	83.02	70.91	75.08	77.27	70.83	73.91
	Chen et al.	80.42	81.72	80.42	79.88	80.42	81.72	80.42	79.88	79.57	72.73	66.67	69.57
	Tamm et al.	74.65	80.56	80.56	76.32	80.06	80.69	78.39	78.72	78.30	75.00	73.42	74.30
	Lin et al.	83.15	82.12	76.70	79.27	84.51	83.64	79.92	81.66	**83.44**	82.67	77.21	79.70
Ours	SpeechHGT	**86.32**	**86.14**	**85.28**	**86.69**	**88.18**	**89.27**	**88.54**	**87.86**	82.82	**83.17**	**82.41**	**81.37**

Table 2: Performance comparison with different baselines (%).

Multimodal Speech Models

These models integrate linguistic and acoustic features to address AD-related speech complexities. The baselines include [Martinc and Pollak, 2020], [Rohanian *et al.*, 2021], [Zhu *et al.*, 2021], [Chen *et al.*, 2023], [Tamm *et al.*, 2023], [Lin and Washington, 2024]. We adhere to the architectures and parameter configurations described in their experiments.

4.3 Implementation Details

We implement the SpeechHGT framework using the PyTorch Geometric library to model multimodal hypergraphs efficiently. The architecture includes two key modules: multimodal hypergraph construction and the multimodal hypergraph transformer. We evaluate the framework on ADReSS, ADReSSo, and ADReSS-M datasets, training and testing separately. The training uses the Adam optimizer (learning rate: 0.001, batch size: 32) with early stopping based on validation loss (patience: 10 epochs). Hyperparameters, including attention heads, hidden dimensions, and dropout rates, are optimized via grid search. All experiments utilize an NVIDIA RTX 4090 GPU, Intel i9 13th Gen CPU, and 64GB RAM, ensuring scalability and computational efficiency. This implementation achieves robust higher-order interaction modeling, validating our framework's effectiveness.

4.4 Performance of SpeechHGT

We evaluate SpeechHGT on ADReSS, ADReSSo, and ADReSS-M datasets using standard binary classification metrics. Results (Table 2) confirm high precision and reliability in AD detection. On ADReSS, the model achieves an accuracy of 86.32%, precision of 86.14%, recall of 85.28%, and F1-score of 86.69%. For ADReSSo, it attains an accuracy of 88.18%, with precision of 89.27% and recall of 88.54%, leveraging hypergraph attention for linguistic-acoustic dependencies. Despite greater heterogeneity in ADReSS-M, it maintains an accuracy of 82.82%, demonstrating robustness while identifying areas for improved adaptation to outlier speech patterns.

4.5 Comparison with Baseline Methods

The proposed SpeechHGT model demonstrates consistent superiority across the ADReSS, ADReSSo, and ADReSS-M datasets compared to four baseline categories: challenge baselines, conventional machine learning models (e.g., RF, SVM, AdaBoost), unimodal speech models, and state-of-the-art multimodal approaches (Table 2). On average, SpeechHGT achieves 85.77% accuracy, 86.33% precision, 85.74% recall, and 85.97% F1-score, outperforming the best-performing model in each baseline group. Compared to challenge baselines (62.50% accuracy), SpeechHGT improves performance by 23.27% on average. Against conventional ML models (best: 65.39% accuracy), it achieves gains of 20.38–24.98%. Unimodal models like [Pérez-Toro *et al.*, 2021] are outperformed accuracy by 10.18% and F1-score by 7.86%. Finally, SpeechHGT surpasses state-of-the-art multimodal baselines, improving F1-score by 4.26% on average. While SpeechHGT slightly underperforms [Lin and Washington, 2024] in accuracy on ADReSS-M (-0.62%), it outperforms them in F1-score (+1.67%) and demonstrates superior consistency across all datasets, highlighting its robustness and generalizability.

4.6 Ablation Study

We conducted an ablation study to evaluate SpeechHGT framework components by disabling specific elements. First, we removed hyperedge-level attention, limiting interactions to pairwise features. Next, hyperedges were eliminated, reducing the graph to pairwise connections. We also assessed modality-specific contributions by excluding linguistic and acoustic features individually. The impact of structural positional encodings (e.g., node degrees, centrality) was also evaluated. Comparisons with a multimodal baseline highlighted their significance. As shown in Table 3, ADReSS accuracy dropped from 86.32% to 78.55% without hyperedge-level attention, and further to 76.82% without hyperedges. Removing linguistic and acoustic features resulted in 81.14% and 80.28% accuracies, respectively, while eliminating structural positional encoding led to an 82.87% accuracy.

4.7 Analysis of Higher-Order Interactions

We employ the SpeechHGT model to systematically identify and quantify higher-order interactions indicative of Alzheimer's Disease (AD)-related cognitive decline. This model utilizes a hypergraph attention mechanism, which effectively prioritizes clinically significant speech features while concurrently minimizing extraneous noise. Key results, illustrated in Figure 4, underscore the pivotal role

Figure 2: The accuracy and F1-score of SpeechHGT w.r.t. different k values on three datasets.

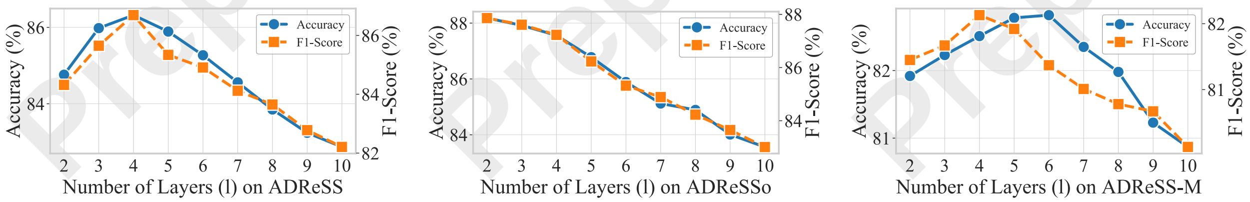

Figure 3: The accuracy and F1-score of SpeechHGT w.r.t. different l values on three datasets.

Model Variant	**ADReSS Accuracy**	**ADReSSo Accuracy**	**ADReSS-M Accuracy**
SpeechHGT (Full Model)	**86.32**	**88.18**	**82.82**
w/o Hyperedge Attention	78.55	79.36	76.19
w/o Hyperedges	76.82	77.60	74.54
w/o Linguistic Features	81.14	82.01	78.68
w/o Acoustic Features	80.28	81.13	77.85
w/o Structural Positional Encodings	82.87	83.77	79.51

Table 3: Ablation study on different components of SpeechHGT on three datasets.

of these interactions, such as semantic coherence and narrative structuring, in AD detection. Identified disruptions in logical flow and prosodic control align with neurolinguistic theories linking conceptual organization and motor-speech processes to cognitive decline [Rumelhart *et al.*, 1986; Dell, 1986].

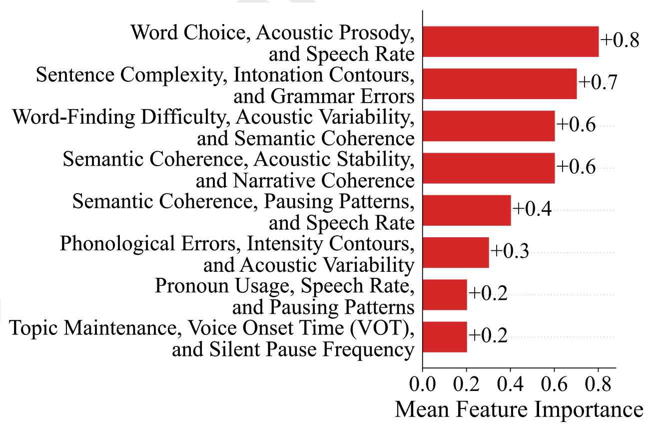

Figure 4: Importance ranking of different higher-order interactions in AD detection.

4.8 Parameter Analysis

The selection of hyperparameters significantly impacts SpeechHGT's ability to model higher-order interactions in spontaneous speech. We analyze the number of clusters (k) for hyperedge construction and transformer depth (l) for hierarchical representation learning. A grid search across $k \in [2, 10]$ and $l \in [2, 10]$ on ADReSS, ADReSSo, and ADReSS-M datasets identifies dataset-specific optima. Fixed hyperparameters (learning rate: 10^{-4}, batch size: 16, attention heads: 8, dropout: 0.3) isolate the effects of k and l. As illustrated in Figure 2 and Figure 3. Optimal cluster counts (k^*) are $k = 3$ for ADReSS and ADReSS-M and $k = 4$ for ADReSSo, with higher k reducing accuracy due to hyperedge fragmentation. Transformer depth (l^*) varies, favoring $l = 4$ for ADReSS, $l = 2$ for ADReSSo, and $l = 6$ for ADReSS-M, reflecting dataset-size dependencies. These results confirm SpeechHGT's sensitivity to parameter tuning across heterogeneous datasets.

5 Conclusion

This study presents SpeechHGT, a novel multimodal hypergraph transformer designed to address the limitations of existing fusion-based models in AD detection through spontaneous speech analysis. By introducing a hypergraph-based approach to represent and learn higher-order interactions between linguistic and acoustic features, SpeechHGT achieved significant improvements in diagnostic accuracy, F1-score on the benchmark datasets, outperforming state-of-the-art methods. Future research will explore the application of SpeechHGT to other neurodegenerative diseases and datasets, alongside architectural enhancements to further refine its diagnostic capabilities. These findings underscore the potential of hypergraph-based learning frameworks to advance non-invasive, speech-based diagnostic tools, providing new insights into the cognitive decline associated with Alzheimer's disease.

Acknowledgments

This work is supported by the National Language Commission Key Program (No. ZDI145-80) and the Natural Science Foundation of Liaoning (Grant No. 2024-MSLH-073).

References

[Ahmed *et al.*, 2019] Md Rishad Ahmed, Yuan Zhang, Zhiquan Feng, Benny Lo, Omer T. Inan, and Hongen Liao. Neuroimaging and machine learning for dementia diagnosis: Recent advancements and future prospects. *IEEE Reviews in Biomedical Engineering*, 12:19–33, 2019.

[Alberdi *et al.*, 2016] Ane Alberdi, Asier Aztiria, and Adrian Basarab. On the early diagnosis of alzheimer's disease from multimodal signals: A survey. *Artificial intelligence in medicine*, 71:1–29, 2016.

[Bessadok *et al.*, 2022] Alaa Bessadok, Mohamed Ali Mahjoub, and Islem Rekik. Graph neural networks in network neuroscience. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5833–5848, 2022.

[Chen *et al.*, 2023] Xuchu Chen, Yu Pu, Jinpeng Li, and Wei-Qiang Zhang. Cross-lingual alzheimer's disease detection based on paralinguistic and pre-trained features. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 1–2. IEEE, June 2023.

[Dell, 1986] Gary S. Dell. A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93(3):283–321, 1986.

[Ding *et al.*, 2024] Kewen Ding, Madhu Chetty, Azadeh Noori Hoshyar, Tanusri Bhattacharya, and Britt Klein. Speech based detection of alzheimer's disease: a survey of ai techniques, datasets and challenges. *Artificial Intelligence Review*, 57(12):1–43, 2024.

[Ektefaie *et al.*, 2023] Yasha Ektefaie, George Dasoulas, Ayush Noori, Maha Farhat, and Marinka Zitnik. Multimodal learning with graphs. *Nature Machine Intelligence*, 5(4):340–350, April 2023.

[Ilias and Askounis, 2022a] Loukas Ilias and Dimitris Askounis. Multimodal deep learning models for detecting dementia from speech and transcripts. *Frontiers in Aging Neuroscience*, 14:830943, 2022.

[Ilias and Askounis, 2022b] Loukas Ilias and Dimitris Askounis. Multimodal deep learning models for detecting dementia from speech and transcripts. *Frontiers in Aging Neuroscience*, 14, March 2022.

[Latif *et al.*, 2021] Siddique Latif, Junaid Qadir, Adnan Qayyum, Muhammad Usama, and Shahzad Younis. Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, 14:342–356, 2021.

[Li *et al.*, 2024] Fan Li, Xiaoyang Wang, Dawei Cheng, Wenjie Zhang, Ying Zhang, and Xuemin Lin. Hypergraph self-supervised learning with sampling-efficient signals. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, IJCAI-2024. International Joint Conferences on Artificial Intelligence Organization, August 2024.

[Lin and Washington, 2024] Kaiying Lin and Peter Y. Washington. Multimodal deep learning for dementia classification using text and audio. *Scientific Reports*, 14(1), June 2024.

[Liu *et al.*, 2021] Rui Liu, Berrak Sisman, and Haizhou Li. Graphspeech: Syntax-aware graph attention network for neural speech synthesis. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 6059–6063. IEEE, June 2021.

[Liu *et al.*, 2024] Xiaoqian Liu, Guoqiang Hu, Yangfan Du, Erfeng He, YingFeng Luo, Chen Xu, Tong Xiao, and Jingbo Zhu. Recent advances in end-to-end simultaneous speech translation. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, IJCAI-2024, page 8142–8150. International Joint Conferences on Artificial Intelligence Organization, August 2024.

[Luo *et al.*, 2024] Xuexiong Luo, Jia Wu, Jian Yang, Shan Xue, Amin Beheshti, Quan Z. Sheng, David McAlpine, Paul Sowman, Alexis Giral, and Philip S. Yu. Graph neural networks for brain graph learning: A survey. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, IJCAI-2024, page 8170–8178. International Joint Conferences on Artificial Intelligence Organization, August 2024.

[Luz *et al.*, 2020] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. In *Interspeech 2020*, interspeech_2020. ISCA, October 2020.

[Luz *et al.*, 2021] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Detecting cognitive decline using speech only: The adresso challenge. In *Interspeech 2021*, interspeech_2021. ISCA, August 2021.

[Luz *et al.*, 2024] Saturnino Luz, Fasih Haider, Davida Fromm, Ioulietta Lazarou, Ioannis Kompatsiaris, and Brian MacWhinney. An overview of the adress-m signal processing grand challenge on multilingual alzheimer's dementia recognition through spontaneous speech. *IEEE Open Journal of Signal Processing*, 5:738–749, 2024.

[Martinc and Pollak, 2020] Matej Martinc and Senja Pollak. Tackling the adress challenge: A multimodal approach to the automated recognition of alzheimer's dementia. In *Interspeech 2020*, interspeech_2020, page 2157–2161. ISCA, October 2020.

[Marvi *et al.*, 2024] Fahimeh Marvi, Yun-Hsuan Chen, and Mohamad Sawan. Alzheimer's disease diagnosis in the preclinical stage: Normal aging or dementia. *IEEE Reviews in Biomedical Engineering*, page 1–18, 2024.

[Pacheco-Lorenzo *et al.*, 2024] Moisés R. Pacheco-Lorenzo, Heidi Christensen, Luis E. Anido-Rifón, Manuel J. Fernández-Iglesias, and Sonia M. Valladares-Rodríguez. Analysis of voice biomarkers for the detection of cognitive impairment. *IEEE Access*, 12:122840–122851, 2024.

[Peng *et al.*, 2024] Ciyuan Peng, Jiayuan He, and Feng Xia. Learning on multimodal graphs: A survey. *ArXiv Preprint ArXiv:2402.05322*, 2024.

[Petti *et al.*, 2020] Ulla Petti, Simon Baker, and Anna Korhonen. A systematic literature review of automatic alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association*, 27(11):1784–1797, 2020.

[Priyadarshinee *et al.*, 2023] Prachee Priyadarshinee, Christopher Johann Clarke, Jan Melechovsky, Cindy Ming Ying Lin, Balamurali BT, and Jer-Ming Chen. Alzheimer's dementia speech (audio vs. text): Multi-modal machine learning at high vs. low resolution. *Applied Sciences*, 13(7):4244, 2023.

[Pulido *et al.*, 2020] María Luisa Barragán Pulido, Jesús Bernardino Alonso Hernández, Miguel Ángel Ferrer Ballester, Carlos Manuel Travieso González, Jiří Mekyska, and Zdeněk Smékal. Alzheimer's disease and automatic speech analysis: a review. *Expert systems with applications*, 150:113213, 2020.

[Pérez-Toro *et al.*, 2021] P.A. Pérez-Toro, S.P. Bayerl, T. Arias-Vergara, J.C. Vásquez-Correa, P. Klumpp, M. Schuster, Elmar Nöth, J.R. Orozco-Arroyave, and K. Riedhammer. Influence of the interviewer on the automatic assessment of alzheimer's disease in the context of the adresso challenge. In *Interspeech 2021*, interspeech_2021, page 3785–3789. ISCA, August 2021.

[Rohanian *et al.*, 2021] Morteza Rohanian, Julian Hough, and Matthew Purver. Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. In *Interspeech 2021*, interspeech_2021, page 3820–3824. ISCA, August 2021.

[Rumelhart *et al.*, 1986] David E Rumelhart, James L McClelland, PDP Research Group, et al. *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press, 1986.

[Sanh, 2019] V Sanh. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[Searle *et al.*, 2020] Thomas Searle, Zina Ibrahim, and Richard Dobson. Comparing natural language processing techniques for alzheimer's dementia prediction in spontaneous speech. In *Interspeech 2020*, interspeech_2020, page 2192–2196. ISCA, October 2020.

[Shehzad *et al.*, 2024] Ahsan Shehzad, Feng Xia, Shagufta Abid, Ciyuan Peng, Shuo Yu, Dongyu Zhang, and Karin Verspoor. Graph transformers: A survey, 2024.

[Shehzad *et al.*, 2025] Ahsan Shehzad, Dongyu Zhang, Shuo Yu, Shagufta Abid, and Feng Xia. Dynamic graph transformer for brain disorder diagnosis. *IEEE Journal of Biomedical and Health Informatics*, page 1–14, 2025.

[Tamm *et al.*, 2023] Bastiaan Tamm, Rik Vandenberghe, and Hugo Van Hamme. Cross-lingual transfer learning for alzheimer's detection from spontaneous speech. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 1–2. IEEE, June 2023.

[Turrisi *et al.*, 2024] Rosanna Turrisi, Margherita Squillario, Giulia Abate, Daniela Uberti, and Annalisa Barla. An overview of data integration in neuroscience with focus on alzheimer's disease. *IEEE Journal of Biomedical and Health Informatics*, 28(4):1824–1835, April 2024.

[Venugopalan *et al.*, 2021] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D Wang. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific reports*, 11(1):3254, 2021.

[Vrindha *et al.*, 2023] M. K. Vrindha, V. Geethu, P. R. Anurenjan, S. Deepak, and K. G. Sreeni. A review of alzheimer's disease detection from spontaneous speech and text. In *2023 International Conference on Control, Communication and Computing (ICCC)*, page 1–5. IEEE, May 2023.

[Yang *et al.*, 2022] Qin Yang, Xin Li, Xinyun Ding, Feiyang Xu, and Zhenhua Ling. Deep learning-based speech analysis for alzheimer's disease detection: a literature review. *Alzheimer's Research & Therapy*, 14(1):186, 2022.

[Ying *et al.*, 2021] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 28877–28888. Curran Associates, Inc., 2021.

[Ying *et al.*, 2023] Yangwei Ying, Tao Yang, and Hong Zhou. Multimodal fusion for alzheimer's disease recognition. *Applied Intelligence*, 53(12):16029–16040, 2023.

[Yu *et al.*, 2024] Shuo Yu, Shan Jin, Ming Li, Tabinda Sarwar, and Feng Xia. Long-range brain graph transformer. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Zhang *et al.*, 2021] Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. Multimet: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, page 3214–3225. Association for Computational Linguistics, 2021.

[Zhang *et al.*, 2024] Lu Zhang, Junqi Qu, Haotian Ma, Tong Chen, Tianming Liu, and Dajiang Zhu. Exploring alzheimer's disease: a comprehensive brain connectome-based survey. *Psychoradiology*, 4:kkad033, 2024.

[Zhu *et al.*, 2021] Youxiang Zhu, Abdelrahman Obyat, Xiaohui Liang, John A. Batsis, and Robert M. Roth. Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection. In *Interspeech 2021*, interspeech_2021, page 3790–3794. ISCA, August 2021.