NeuroImage 327 (2026) 121739

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Neural-linguistic analysis for Alzheimer’s detection: A deep learning approach informed by cognitive neuroscience☆,☆☆

Jianhui Lv [a], Shalli Rani [b], Keqin Li [c], Ning Liu [d,*]

[a] Multi-modal Data Fusion and Precision Medicine Laboratory, The First Affiliated Hospital of Jinzhou Medical University, Jinzhou 121000, China
[b] Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab 140401, India
[c] College of Computer Science, State University of New York, New Paltz, NY 12561, USA
[d] Department of Imaging, The First Affiliated Hospital of Jinzhou Medical University, Jinzhou 121000, China

ARTICLE INFO

Keywords:
Alzheimer’s disease detection
Cognitive neuroscience
Neural-linguistic feature
Acoustic-symbolic transformation

ABSTRACT

Alzheimer’s disease (AD) is a progressive neurodegenerative disorder that disrupts cognitive function across multiple domains, particularly affecting language networks and speech production pathways in the brain. Patients demonstrate symptoms including aphasia, reduced syntactic complexity, and diminished verbal fluency that reflects underlying neural pathology in language-related cortical areas. Current detection methods rely on resource-intensive neuroimaging, invasive biomarker sampling, and extensive neuropsychological testing, creating substantial barriers to early diagnosis. While researchers have explored using acoustic features, paralinguistic markers, and text-based features for AD detection, existing approaches face fundamental limitations: traditional acoustic methods fail to capture semantic-cognitive content, text transcription is labor-intensive, and automatic speech recognition quality suffers due to pronunciation variations and cognitive impairments in elderly populations. This paper introduces cognitive acoustic symbolic transformation for ALzheimer’s (COASTAL), a neurobiologically-inspired framework that models hierarchical speech processing pathways. COASTAL transforms acoustic patterns into discrete symbolic elements through a specialized transformation module before applying contextual analysis that mirrors prefrontal-temporal language networks. Evaluated on the ADReSSo corpus, COASTAL achieved 70.42% accuracy, outperforming established baselines by 5.63%. Integration with complementary self-supervised approaches through hierarchical fusion improved performance to 77.46%. Analysis revealed that preserving fine-grained temporal features through shallower transformation architecture significantly enhanced diagnostic accuracy, aligning with neuropsychological evidence that subtle timing patterns in speech provide sensitive markers of cognitive decline.

1. Introduction

Alzheimer’s disease (AD) is a neurodegenerative disorder with worldwide impact, currently affecting approximately 50 million individuals, with projections indicating this number may triple by 2050 (Vogt et al., 2023; Klepl et al., 2022; Zhao et al., 2024). AD is categorized as a neurodegenerative disorder, and from a cognitive neuroscientific perspective, it represents the gradual dismantling of the complex systems of an organism’s brain along with the progressive loss of cognitive abilities (Liu et al., 2024; Niazi et al., 2024). Deficits in memory, executive skills, and language are hallmarks of this disease (Liampas et al., 2023). By-patient observation, AD is associated with a loss of neurons and synapse dysfunction, especially in the medial temporal lobe, posterior cingulate, and association cortices (Mangalmurti and Lukens, 2022; Griffiths et al., 2023). This explains the clinical features of impaired memory and language processing. Remarkably, the neural changes accompany behavioral alterations, particularly changes in speech patterns, long before a formal diagnosis can be made (Robin et al., 2023). This critical period enables AD to be diagnosed and treated far earlier than current methods that focus on advanced biomarkers, which are, in fact, the consequence of early neurodegenerative changes (Lardelli et al., 2025; Ginsberg M. J. Blaser, 2024).

Diagnosing AD traditionally involves neuropsychological tests, PET scans for amyloid plaques, and the tau and amyloid protein tests from

☆ This article is part of a Special issue entitled: ‘AI for Early Dementia Detection’ published in NeuroImage.

☆☆ This work is supported by Natural Science Foundation of Liaoning Province - General Project in 2026 by Ning Liu: The mechanism by which astaxanthin improves cell pyroptosis through the PI3K/Akt/NF-kB axis and the application research of cognitive acoustic framework analysis in Alzheimer’s disease.

* Corresponding authors.
E-mail addresses: lvjianhui2012@163.com (J. Lv), shalli.rani@chitkara.edu.in (S. Rani), lik@newpaltz.edu (K. Li), liun@jzmu.edu.cn (N. Liu).

https://doi.org/10.1016/j.neuroimage.2026.121739
Received 16 June 2025; Received in revised form 29 December 2025; Accepted 19 January 2026
Available online 20 January 2026

cerebrospinal fluid analysis (Zhou et al., 2025; Yang et al., 2024a). Such tests are expensive, and only because of their cost, invasiveness, and the time required to identify significant structural changes in the brain, they are usually reserved for the later stages of the disease (Aye et al., 2024). This creates a difficult hurdle in identifying AD at an earlier stage when interventions might be more beneficial (Qin et al., 2024). Studies in cognitive neuroscience have pointed out that prose and speech production and understanding are some of the first areas to be impacted in early AD, with hallmark changes emerging due to the progressive loss of the semantic memory network, phonological processing systems, and the frontal executive control systems that are responsible for the output of fluent speech (Lv et al., 2025). These alterations provide insight into AD progression, which aligns with understanding the cerebrum's affected areas during the onset of Alzheimer's, thus offering a non-invasive opportunity to study the disease process (Monfared et al., 2022; Dao et al., 2025).

The evolution of AI and its branches have changed the paradigm for processing and analyzing complex systematic patterns in speech and language data (Zhang, 2025; Luo et al., 2024). Such changes have opened up new avenues in developing non-invasive screening tools for AD. Previous studies have tried to solve the problem with approaches like acoustic feature analysis, paralinguistic marker analysis, and semantic modeling based on text analytics (Koenig et al., 2023). The Computational Paralinguistics Challenge and extended Geneva Minimalistic Acoustic Parameter Set datasets have been shown to perform well in predicting AD using acoustic features (Bayerl et al., 2023; Garcia-Gutierrez et al., 2023). It has been reported that certain prosodic features of speech, like rhythm and stress patterns, are disrupted during AD, and the degree to which these features are disrupted is proportional to the degree of AD, indicating disruption of frontal-subcortical circuits involved in speech motor control (Maiella et al., 2024). Paralinguistic studies have demonstrated that Alzheimer's disease progressively disrupts frontal-subcortical circuits responsible for coordinating speech timing, manifesting as altered speech rate and abnormal temporal organization of phrases and pauses. These disruptions reflect neurodegeneration affecting the neural networks that allocate attentional resources and maintain working memory during speech production. Analyzed data deriving metrics such as type-token ratio, syntactic complexity, and coherence measures demonstrate that AD profoundly affects a person's ability to access and structure grammatical elements using language due to the gradual decline of neural networks dedicated to language processing (Hsu et al., 2025).

They do not help much in accurately differentiating AD from other neurologic conditions that may have similar acoustic profiles due to the inability to model the speech's semantic content. Automated transcription techniques that analyze discourse to extract the linguistic features of text provide some semantic information, but the automated systems perform poorly on elderly speakers. Moreover, most existing approaches consider speech an acoustic signal or a linguistic unit. The paper introduces an innovative neural-linguistic framework that integrates acoustic and semantic processing pathways, thus addressing these gaps.

Accordingly, the main contributions of this paper are summarized as follows.

- We propose cognitive acoustic symbolic transformation for ALzheimer's (COASTAL), a novel neurobiologically-inspired framework for detecting AD from spontaneous speech.
- We address fundamental limitations in existing approaches by developing a hierarchical processing architecture that transforms acoustic signals into discrete symbolic representations before applying contextual sequence analysis, mirroring the organization of language networks in the human brain.
- We demonstrate that shallower transformation architectures (2-layer) more effectively preserve fine-grained temporal features critical for cognitive assessment, achieving 70.42% accuracy on the ADReSSo dataset and outperforming established baseline methods by 5.63%.

The rest of the paper is organized as follows: Section 2 introduces the COASTAL framework. Section 3 presents the experimental methodology and results. Finally, Section 4 concludes the paper.

2. Neurally-inspired computational framework for cognitive deficit detection

The architecture transforms acoustic representations into symbolic linguistic elements through a cascade of processing steps that parallel neural pathways in human speech understanding. This computational framework converts acoustic spectral patterns to quantized representational tokens and subsequently analyzes temporal dependencies within these discrete sequences to identify cognitive markers associated with neurodegeneration. The design draws from neuroimaging studies (Robertson et al., 2024) mapping the ventral speech processing stream that extends from primary auditory cortex through superior temporal regions to inferior frontal areas. Our system implements a dual-component architecture: an acoustic-symbolic transformation module ($\mathcal{T}$) and a contextual sequence analyzer ($\mathcal{S}$). Fig. 1 provides a schematic representation of the complete architectural framework (Becker et al., 1994).

2.1. Acoustic-symbolic transformation module

The acoustic-symbolic transformation component employs a modified variational inference framework fundamentally different from standard approaches. While conventional variational learning (Li et al., 2023) utilizes encoder–decoder architectures ($\mathcal{E}_\psi$ and $\mathcal{D}_\omega$), our implementation incorporates specialized constraints reflecting cognitive processing limitations.

The transformation process begins with acoustic input sequences $\boldsymbol{A} = \{a_1, a_2, \ldots, a_m\}$ that undergo non-linear projection into a representational manifold $\mathcal{M}$. The architecture assumes that cognitive representations occupy discrete regions within this manifold, corresponding to attractor states in neural dynamics. The inference process can be formalized through the mapping function $\mathcal{E}_\psi : \boldsymbol{A} \to \boldsymbol{v}$, which projects acoustic patterns onto latent variables $\boldsymbol{v}$ constrained by prior distribution $\mathcal{P}(\boldsymbol{v})$ modeling expected cognitive representations.

The generative component $\mathcal{D}_\omega(\boldsymbol{A}|\boldsymbol{v})$ reconstructs acoustic patterns from these latent representations. For temporal sequence $\boldsymbol{A}$ with M elements, the conditional generative process incorporates context-sensitivity through:

$$\mathcal{D}_\omega(\boldsymbol{A}|\boldsymbol{v}) = \prod_{j=1}^{M} \mathcal{D}_\omega(a_j|\{a_k\}_{k<j}, \boldsymbol{v}) \cdot \mathcal{R}(\boldsymbol{A}, \boldsymbol{v}, j) \tag{1}$$

where $\mathcal{R}(\boldsymbol{A}, \boldsymbol{v}, j)$ represents a recurrence function modeling working memory constraints that limit integration across distant sequence elements - a cognitive limitation particularly affected in AD.

The posterior distribution $\mathcal{P}(\boldsymbol{v}|\boldsymbol{A})$ cannot be directly computed due to the combinatorial complexity of possible attractor configurations. We therefore introduce approximation function $\mathcal{Q}_\psi(\boldsymbol{v}|\boldsymbol{A})$ optimized through divergence minimization. This leads to our evidence lower bound formulation:

$$\begin{aligned}\mathcal{L}_{ELBO} &= \mathbb{E}_{\mathcal{Q}_\psi(\boldsymbol{v}|\boldsymbol{A})}\left[\log \mathcal{D}_\omega(\boldsymbol{A}|\boldsymbol{v})\right] - \\ \gamma \cdot D_{KL}&\left[\mathcal{Q}_\psi(\boldsymbol{v}|\boldsymbol{A})||\mathcal{P}(\boldsymbol{v})\right] + \lambda \cdot \Omega(\psi, \omega)\end{aligned} \tag{2}$$

The first component represents reconstruction fidelity; the second implements regularization through divergence between approximate posterior and cognitive prior, and $\Omega(\psi, \omega)$ incorporates neurocognitive constraints on representational capacity with an importance weight λ.

Our implementation extends this framework through discretization operations that quantize continuous representations into symbolic elements $\boldsymbol{c}$. The transformation process incorporates neural biophysical constraints by modeling each representation as activation across

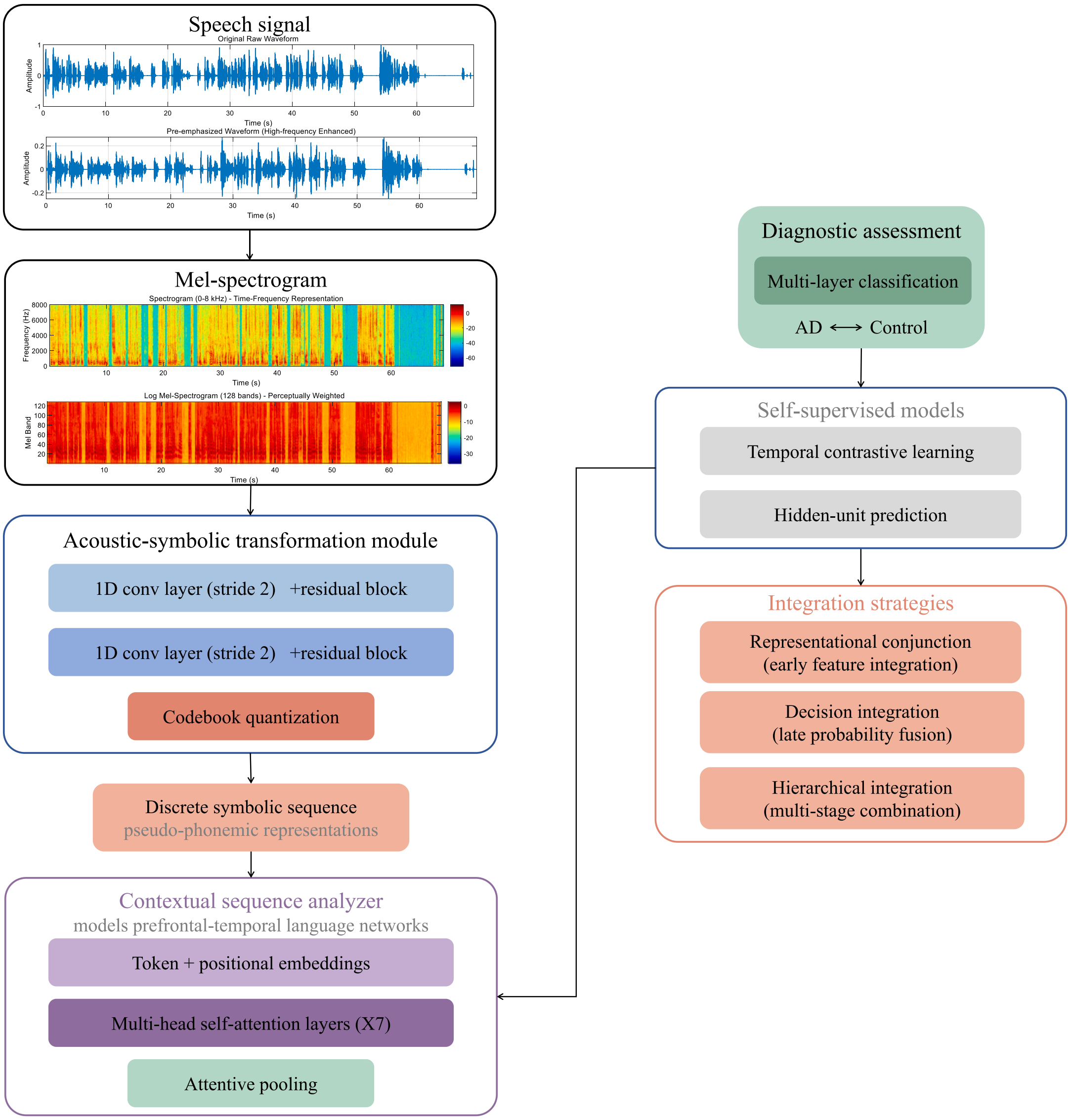
Speech signal
Original Raw Waveform
Amplitude
1
0
-1
0
10
20
30
40
50
60
Time (s)
Pre-emphasized Waveform (High-frequency Enhanced)
Amplitude
0.2
0
-0.2
Time (s)
Mel-spectrogram
Spectrogram (0-8 kHz) - Time-Frequency Representation
Frequency (Hz)
8000
6000
4000
2000
0
-20
-40
-60
Time (s)
Log Mel-Spectrogram (128 bands) - Perceptually Weighted
Mel Band
120
100
80
60
40
20
-10
-30
Time (s)
Acoustic-symbolic transformation module
1D conv layer (stride 2) +residual block
1D conv layer (stride 2) +residual block
Codebook quantization
Discrete symbolic sequence
pseudo-phonemic representations
Contextual sequence analyzer
models prefrontal-temporal language networks
Token + positional embeddings
Multi-head self-attention layers (X7)
Attentive pooling
Diagnostic assessment
Multi-layer classification
AD ⟷ Control
Self-supervised models
Temporal contrastive learning
Hidden-unit prediction
Integration strategies
Representational conjunction
(early feature integration)
Decision integration
(late probability fusion)
Hierarchical integration
(multi-stage combination)

Fig. 1. Cognitive-linguistic processing framework.

K distinct neuronal ensembles with competitive dynamics, formally expressed as:

$$\boldsymbol{c} = \mathrm{Quantize}(\mathcal{E}_\psi(\boldsymbol{A})) = \sum_{i=1}^{K} \delta_i \cdot \mathrm{OneHot}(\arg\max_i \mathcal{S}_i(\mathcal{E}_\psi(\boldsymbol{A}))) \tag{3}$$

where $\mathcal{S}_i$ represents the activation function for the ith neural ensemble and δ_i is its corresponding symbolic representation in a learned codebook. This competitive selection process parallels winner-take-all dynamics in cortical microcircuits, particularly in speech perception regions where categorical boundaries emerge from continuous acoustic input (Koever et al., 2013).

The discretization operation (Tian et al., 2021) introduces non-differentiability in computational workflows. To address this limitation while maintaining biological plausibility, we implement a temperature-controlled relaxation using softened categorical distributions:

$$\tilde{c} = \sum_{i=1}^{K} \delta_i \cdot \frac{\exp((\mathcal{S}_i(\mathcal{E}_\psi(\boldsymbol{A})) + g_i)/\tau)}{\sum_{j=1}^{K} \exp((\mathcal{S}_j(\mathcal{E}_\psi(\boldsymbol{A})) + g_j)/\tau)} \tag{4}$$

where g_i represents stochastic perturbations modeling neural noise, and τ controls selectivity, paralleling attentional modulation in auditory processing. This modified objective function becomes:

$$\mathcal{L}_{modified} = \mathbb{E}_{\mathcal{Q}_\psi(\boldsymbol{c}|\boldsymbol{A})}\left[\log \mathcal{D}_\omega(\boldsymbol{A}|\boldsymbol{c})\right] - \alpha \cdot D_{KL}\left[\mathcal{Q}_\psi(\boldsymbol{c}|\boldsymbol{A})||\mathcal{P}(\boldsymbol{c})\right] + \xi \cdot \boldsymbol{\Phi}(\boldsymbol{c}) \tag{5}$$

where $\boldsymbol{\Phi}(\boldsymbol{c})$ incorporates additional constraints on symbolic representations reflecting neurological limitations in phonological processing characteristic of neurodegenerative conditions.

Architecturally, the acoustic-symbolic transformation module incorporates hierarchical processing through cascaded convolutional operations (Song et al., 2022) with increasing temporal receptive fields, mirroring the progressive integration across longer timescales observed in ascending auditory pathways. Each processing layer incorporates skip connections implementing predictive coding principles:

$$\mathcal{F}(\boldsymbol{H}_l) = \mathrm{Conv1D}(\boldsymbol{H}_{l-1}) + \mathrm{SkipConnect}(\boldsymbol{H}_{l-1}, \boldsymbol{H}_{l-2}) \tag{6}$$

where $\boldsymbol{H}_l$ represents feature activations at layer l. This architecture creates representations capturing both phonemic content and suprasegmental features (rhythm, prosody, hesitations) that serve as critical diagnostic markers for cognitive decline.

2.2. Bidirectional contextual encoder model

The bidirectional contextual encoder (β) consists of multiple bidirectional Transformer encoder layers that model the sequential relationships within the pseudo-phoneme sequences generated by the η model. This component is inspired by neuroscientific evidence of bidirectional processing in language comprehension networks, where both preceding and following contexts influence word interpretation in temporal and frontal language areas.

The β model is trained using two self-supervised tasks that mirror aspects of human language processing: masked language modeling (MLM) (Wu and Chung, 2022) and next sentence prediction (NSP) (Liao et al., 2024). In the MLM task, we randomly mask 15% of the tokens in a training sequence. For each selected token at position i, we apply one of three transformations: replace with [MASK] token (80% probability), replace with a random token (10% probability), or leave unchanged (10% probability). The model then predicts the original token using the surrounding context, similar to how humans use contextual cues to resolve ambiguous or degraded speech signals.

For the NSP task, we select sequence pairs A and B from the dataset. Fifty percent of pairs have B as the actual continuation of A, while in the remaining 50%, B is a random sequence. A [SEP] token separates the sequences, and the model performs binary classification to determine if the sequences are continuous. This task helps the model learn discourse-level relationships, which are particularly disrupted in AD due to impairments in executive function and working memory.

The architecture of the β model consists of 6 Transformer encoder layers, each with 12 attention heads and 768-dimensional feature embeddings. This design allows the model to capture long-range dependencies in language that are typically impaired in AD patients due to disrupted connectivity between frontal and temporal language regions. While formal interpretability analysis of the attention mechanisms was beyond the scope of this initial study, the multi-head attention architecture theoretically enables the model to learn different aspects of linguistic relationships simultaneously, with each head potentially specializing in different dependency types such as syntactic, semantic, or discourse-level connections.

The embeddings from the β model provide rich linguistic representations that capture syntax, semantics, and discourse-level features relevant to detecting cognitive impairment. By processing discrete pseudo-phoneme sequences rather than text, the model preserves important paralinguistic information like hesitations, repetitions, and dysfluencies that are known markers of cognitive decline.

2.3. Neural-linguistic model training process

Training our neural-linguistic model occurs in three distinct phases, mirroring the brain's hierarchical organization of language processing from acoustic feature extraction to phonological encoding to semantic integration.

In the first phase, we train the η model to compress mel-spectrograms into sequences of pseudo-phonemes. The η encoder transforms the input spectrogram into a sequence of discrete codes drawn from a learned codebook. These discrete codes function as pseudo-phonemes, representing fundamental units of speech similar to how the brain's dorsal auditory stream transforms continuous acoustic signals into discrete phonological representations.

The encoder architecture employs multiple 1D convolutional layers with stride 2, each followed by a residual block that employs the following transformation:

$$\mathbf{F}(\mathbf{y}) = \mathbf{F}_c(\mathbf{y}) + \mathbf{y} \tag{7}$$

where $\mathbf{F}_c$ represents the convolutional transformation and $\mathbf{y}$ is the input to the residual block. This residual connection helps maintain gradient flow during training, analogous to the parallel processing pathways observed in cortical auditory processing.

The discrete nature of the output is achieved through vector quantization, where each continuous vector produced by the encoder is mapped to its nearest neighbor in a learned codebook $\mathcal{C} = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_K\}$ containing K prototype vectors:

$$q(\mathbf{d}|\mathbf{x}) = \delta(\mathbf{d} - \mathbf{e}_k) \quad \text{where} \quad k = \arg\min_j \|\mathbf{z}_e(\mathbf{x}) - \mathbf{e}_j\|_2 \tag{8}$$

where $\mathbf{z}_e(\mathbf{x})$ is the output of the encoder and δ is the Dirac delta function. Since this quantization operation is non-differentiable, we employ the straight-through estimator for backpropagation:

$$\nabla_{\mathbf{z}_e} \mathcal{L} \approx \nabla_{\mathbf{d}} \mathcal{L} \tag{9}$$

The training objective for the η model combines reconstruction loss with commitment loss:

$$\mathcal{L}_\eta = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \alpha \|\text{sg}[\mathbf{z}_e(\mathbf{x})] - \mathbf{d}\|_2^2 + \gamma \|\mathbf{z}_e(\mathbf{x}) - \text{sg}[\mathbf{d}]\|_2^2 \tag{10}$$

where sg[$\cdot$] denotes stop-gradient, α controls the commitment of the encoder to its outputs, and γ regulates the divergence between encoder outputs and codebook vectors.

In the second phase, we train the β model using the pseudo-phoneme sequences generated by the η encoder. The β model vocabulary consists of the η codebook indices plus special tokens [PAD], [CLS], [SEP], and [MASK]. The positional embeddings (**PE**) are added to the token embeddings (**TE**) to form the input representation:

$$\mathbf{H}_0 = \mathbf{TE} + \mathbf{PE} \tag{11}$$

The transformer encoder layers then process this input through self-attention mechanisms and feed-forward networks:

$$\mathbf{H}_l = \text{TransformerLayer}_l(\mathbf{H}_{l-1}) \tag{12}$$

where $l \in \{1, 2, \dots, L\}$ and $L = 6$ in our implementation. For the MLM task, the training objective is:

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{i \in \mathcal{M}} \log p_\beta(x_i | \mathbf{x}_{\backslash \mathcal{M}}) \tag{13}$$

where $\mathcal{M}$ is the set of masked token positions and $\mathbf{x}_{\backslash \mathcal{M}}$ represents the input with tokens at positions in $\mathcal{M}$ masked.

For the NSP task, the objective is:

$$\mathcal{L}_{\text{NSP}} = -\mathbb{E}_{(\mathbf{A},\mathbf{B},y)} \log p_\beta(y | [\text{CLS}], \mathbf{A}, [\text{SEP}], \mathbf{B}) \tag{14}$$

where $y \in \{0, 1\}$ indicates whether sequence **B** follows sequence **A** in the original data.

The combined training objective for the β model is:

$$\mathcal{L}_\beta = \mathcal{L}_{\text{MLM}} + \lambda \mathcal{L}_{\text{NSP}} \tag{15}$$

where λ balances the contribution of the two tasks.

In the third phase, for AD detection, we freeze the parameters of both η and β models and use them to extract features from audio samples. We process the audio through the η encoder to obtain pseudo-phoneme sequences, which are then fed into the β model. The final hidden state corresponding to the [CLS] token is average-pooled over the time dimension to produce a fixed-dimensional representation:

$$\mathbf{r} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{H}_L^{(t)} \tag{16}$$

This representation is then passed through a classification network consisting of two fully connected layers with a rectified linear unit (ReLU) activation function between them.

2.4. Neurobiologically-inspired self-supervised learning frameworks

Recent advances in computational neuroscience have produced self-supervised representation learning (Oh et al., 2023) approaches that capture hierarchical structure in speech signals. We explore integrating our cognitive processing framework with these neurobiologically-inspired architectures to enhance detection sensitivity for neurodegenerative markers.

Temporal contrastive learning ($\mathcal{W}$) implements a predictive processing framework that captures acoustic-phonetic and sequential speech signals' dependencies. The architecture consists of three specialized components: acoustic feature extractor, predictive context network, and contrastive learning mechanism. This organization parallels hierarchical predictive processing observed in auditory cortical networks, where each processing level predicts upcoming input patterns.

During forward computation, raw waveform signals transform multiple convolutional layers with increasing temporal receptive fields:

$$\zeta_t = \mathrm{ConvFeatureExtractor}(\mathrm{Waveform}, t) \tag{17}$$

These representations are then processed through a contextualization network implementing multiple self-attending layers with causal masking:

$$\chi_t = \mathrm{ContextNetwork}(\zeta_{1:t}) \tag{18}$$

The architecture implements a contrastive learning objective that discriminates between genuine future representations and distractor patterns:

$$\mathcal{L}_{\mathcal{W}} = -\log \frac{\exp(\mathrm{sim}(\chi_t, \zeta_{t+\delta})/\kappa)}{\sum_{\tilde{\zeta}\in\mathcal{Z}_t} \exp(\mathrm{sim}(\chi_t, \tilde{\zeta})/\kappa)} \tag{19}$$

where $\mathrm{sim}(\cdot,\cdot)$ represents cosine similarity between vectors, $\mathcal{Z}_t$ contains the target future representation $\zeta_{t+\delta}$ and N distractor representations, and κ controls solution sharpness. This objective encourages the model to develop predictive representations encoding temporal dependencies across multiple timescales - a capability specifically degraded in neurodegenerative conditions affecting prefrontal function.

The hidden-unit prediction framework ($\mathcal{H}$) implements a different biologically-inspired approach. This architecture first clusters acoustic features using unsupervised learning to establish perceptual categories similar to phonetic feature detectors in the superior temporal cortex. The model then predicts these categories from masked contextual windows, implementing a computational version of the predictive coding theory of cortical function.

The objective function of this framework becomes:

$$\mathcal{L}_{\mathcal{H}} = -\sum_{i\in\mathcal{M}} \log p_\theta(v_i|\chi_{\setminus\mathcal{M}}) \tag{20}$$

where v_i represents the cluster assignment for frame i, and $\chi_{\setminus\mathcal{M}}$ represents the input features with frames in mask set $\mathcal{M}$ masked out. The masking strategy implements an irregular pattern with varying mask lengths modeling attention fluctuations:

$$P_{\mathrm{mask}}(i) = f(\Phi_i, \Psi_{i-k:i+k}) \tag{21}$$

where Φ_i represents local acoustic properties and $\Psi_{i-k:i+k}$ represents contextual influence from surrounding frames. This masking procedure specifically targets acoustically salient regions, forcing the model to develop robust contextual prediction capabilities that rely on intact working memory function.

These neurobiologically-inspired frameworks capture complementary aspects of speech processing: $\mathcal{W}$ excels at learning predictive acoustic-sequential representations, while $\mathcal{H}$ develops abstract categorical representations paralleling phonological perception. By integrating these frameworks with our cognitive assessment system, we aim to leverage their complementary strengths for enhanced detection of neurocognitive decline markers.

2.5. Multimodal integration strategies for neuropathological assessment

To combine neurocognitive markers extracted from different processing pathways, we implement integration strategies inspired by multimodal association areas in the brain. These strategies model how distributed information processing streams converge in tertiary association cortices to support complex cognitive functions.

A representational conjunction strategy implements early information integration in a manner similar to multisensory convergence in association cortices. This approach combines feature representations from different processing streams before diagnostic assessment:

$$\boldsymbol{r}_{\mathrm{conjunctive}} = \Psi(\boldsymbol{r}_{\mathcal{TS}}, \boldsymbol{r}_{\mathcal{W}}, \boldsymbol{r}_{\mathcal{H}}) \tag{22}$$

where $\boldsymbol{r}_{\mathcal{TS}}$, $\boldsymbol{r}_{\mathcal{W}}$, and $\boldsymbol{r}_{\mathcal{H}}$ are representations from our cognitive framework, temporal contrastive learning, and hidden-unit prediction frameworks, respectively. The conjunction function Ψ implements a sophisticated multimodal integration mechanism:

$$\Psi(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n) = \sum_{i=1}^{n} \mathbf{W}_i \cdot g_i(\boldsymbol{r}_i) + \sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathbf{B}_{ij} \cdot h_{ij}(\boldsymbol{r}_i, \boldsymbol{r}_j) \tag{23}$$

where g_i represents individual transformation functions, h_{ij} represents pairwise interaction functions, and $\mathbf{W}_i$ and $\mathbf{B}_{ij}$ are learned weight matrices. This formulation explicitly captures both individual feature contributions and cross-stream interactions that may reveal subtle patterns of neurocognitive dysfunction. Here, 'individual' refers to separate processing streams rather than patient-specific cognitive profiles. The model learns to weight different neural pathways based on their diagnostic relevance across the population, though future work could incorporate personalized cognitive assessments.

Decision integration strategy implements a late fusion approach inspired by convergent decision processes in frontal executive networks. This method combines diagnostic assessments made by individual processing streams:

$$\hat{y}_{\mathrm{integrated}} = \Omega(\hat{y}_{\mathcal{TS}}, \hat{y}_{\mathcal{W}}, \hat{y}_{\mathcal{H}}) \tag{24}$$

where $\hat{y}_{\mathcal{TS}}$, $\hat{y}_{\mathcal{W}}$, and $\hat{y}_{\mathcal{H}}$ represent diagnostic outputs from different processing streams. The decision integration function Ω implements a dynamic weighting mechanism sensitive to certainty indicators:

$$\Omega(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n) = \sum_{i=1}^{n} \omega_i \cdot \hat{y}_i \tag{25}$$

$$\omega_i = \frac{\exp(c_i \cdot \phi_i)}{\sum_{j=1}^{n} \exp(c_j \cdot \phi_j)} \tag{26}$$

where c_i represents confidence indicators for each processing stream, and ϕ_i represents reliability weights optimized during system development. This confidence-weighted approach mimics clinical decision-making processes where different diagnostic indicators are weighted according to their reliability and specificity.

The hierarchical integration strategy implements a more sophisticated approach inspired by the nested processing hierarchy of the prefrontal cortex. For example, consider processing a speech segment containing a hesitation pattern. The first stage might transform individual stream outputs (COASTAL detecting the temporal gap, temporal contrastive learning identifying disrupted acoustic continuity, hidden-unit prediction recognizing unusual categorical transitions) into normalized representations. The second stage would then combine pairs of these representations to identify interaction patterns (such as hesitations coinciding with semantic access difficulties), while the final stage integrates all information to determine overall cognitive status. This method applies a staged integration process with increasing abstraction levels:

$$\boldsymbol{r}_i^{(1)} = \mathrm{TransformStream}_i(\boldsymbol{r}_i) \tag{27}$$

$$\boldsymbol{r}_{ij}^{(2)} = \mathrm{IntegratePair}_{ij}(\boldsymbol{r}_i^{(1)}, \boldsymbol{r}_j^{(1)}) \tag{28}$$

$$\boldsymbol{r}^{(3)} = \mathrm{FinalIntegration}(\{\boldsymbol{r}_{ij}^{(2)}\}) \tag{29}$$

$$\hat{y}_{\mathrm{hierarchical}} = \mathrm{DiagnosticAssessment}(\boldsymbol{r}^{(3)}) \tag{30}$$

This hierarchical approach enables the system to identify complex interaction patterns across processing streams that may provide sensitive markers for subtle neurocognitive changes preceding clinical manifestation of AD.

3. Experimental protocol and neurocognitive analysis

3.1. Dementia speech corpus

The INTERSPEECH 2021 ADReSSo Challenge corpus (Syed et al., 2021) was used to assess the COASTAL framework. This corpus is a benchmark for the comparative assessment of automatic speech AD detection systems. The ADReSSo corpus is a remarkable development in the computational meritocratic resource landscape for cognitive evaluation because of its controlled demographic balances and acoustic preprocessing to reduce confounding factors.

The creation of the ADReSSo corpus seeks to build upon the most evident gaps in existing spontaneous speech datasets tailored for dementia-focused research. First, the dataset upholds acoustic integrity by employing targeted preprocessing to reduce the recording environment's variance while maintaining the diagnostics of relevant speech elements. Second, it enforces stringent demographic control by balancing age and gender to eliminate the potentiality of confounding effects on the classifiers. These design decisions resolve methodological issues concerning ecological validity and demographic fairness raised by cognitive neuroscientists in the context of computational approaches to neurological evaluation.

The entire corpus contains 237 speech recordings organized within diagnostic and experimental categories, while the development corpus has 166 recordings (87 from clinically diagnosed Alzheimer's patients and 79 from age-matched healthy controls). The evaluation corpus also has 71 recordings with the same categorical distribution. All speech samples were collected using the standardized "Cookie Theft" picture description task from the Boston Diagnostic Aphasia Examination, which is a neuro-linguistic test widely used in clinical neuropsychology to assess communicative impairments due to neurodegenerative diseases. This task was chosen because cognitive neuroscience supports its ability to prompt spontaneous language use. It requires the coordination of several cognitive systems affected by AD, such as semantic memory, executive control, and visuospatial processing.

The dataset spans 5.05 h, with recordings between 22 and 268 s. Participants were digitized at 44.1 kHz and 16 bits and underwent preprocessing, which fixed noise profile removal and amplitude normalization to reduce variability from the recording environment. This method supports the reduction of variability that could obscure computational analysis while retaining the speech attributes important for cognitive assessment.

Participants were screened and grouped based on age (53 to 84 years) and English proficiency. All participants were native English speakers. Gender proportions were balanced across experimental conditions to avoid influence from gender-related idiosyncratic pronunciation or speech classification features. Such careful demographic matching solves one of the long-standing problems in the automated assessment of neurological impairment. Demographic factors have limited the applicability of the algorithms designed using machine learning techniques.

To evaluate the COASTAL framework, we compared six state-of-the-art baseline methods representing diverse approaches to AD detection from speech. These baselines range from purely acoustic to multimodal approaches, capturing the current methodological spectrum in the field:

- Transcranial focused ultrasound - Bidirectional Encoder Representations from Transformers (tFUS-BERT) (Thipparthy et al., 2025): A discrete variational model to encapsulate the linguistic representations of audio.
- Speech Pause Feature Extraction and Encoding (SPFEE) (Liu et al., 2023): A speech pause feature extraction and encoding strategy for acoustic-signal-based AD detection.
- Early Diagnosis of Alzheimer's disease based on Multimodal Attention Mechanism (EDAMM) (Yang et al., 2024b): A model for early diagnosis of AD based on multimodal attention.
- Fix-Length Features from Unfixed-Length Audio (FLFULA) (Pan et al., 2024): A feature extraction method for extracting fix-length features from variable-length audio recordings.
- Multimodal Fusion for Noninvasive detection of AD (MFNI) (Ying et al., 2023): Multimodal fusion for early detection of AD by noninvasive methods.
- Intra- and Cross-Modal Interactions (ICMI) (Ilias and Askounis, 2023): A method for detecting AD patients that captures intra- and cross-modal interactions.

The selection of these baseline approaches provides comprehensive coverage of current methodological paradigms in speech-based cognitive assessment, from purely acoustic to fully integrated multimodal approaches. This competitive evaluation framework enables systematic assessment of the COASTAL model's performance relative to diverse methodological alternatives, highlighting its specific advantages in leveraging cognitive neuroscience principles for improved diagnostic accuracy.

3.2. Computational implementation protocol

The experimental evaluation of the COASTAL framework followed a rigorous protocol optimized for cognitive assessment applications. All recordings underwent preprocessing, including downsampling to a uniform 16 kHz sampling rate and segmentation using 10-second windows with 6-second overlap (40% step size). For participants with limited speech production, controlled acoustic augmentation was implemented using calibrated signal-to-noise ratios (28-35 dB SNR).

The acoustic-symbolic transformation module ($\mathcal{T}$) implemented a hierarchical architecture with dual convolutional processing pathways. The encoder network consisted of two convolutional layers with kernel size 3 and stride 2, each followed by residual processing blocks. Optimization employed Adam algorithm with hyperparameters $\alpha = 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ for 20 epochs with exponential decay factor $\gamma = 0.98$. This configuration transformed input mel-spectrograms (frame size 25 ms, frameshift 10 ms, 128 mel bands) into symbolic sequences of length ~ 250 elements.

The quantization mechanism implemented vector quantization with a learnable codebook containing $K = 1024$ prototype vectors. Optimization incorporated Gumbel-Softmax relaxation with temperature annealing from $\tau_{init} = 2.0$ to $\tau_{final} = 0.5$ over the training period.

The contextual sequence analyzer ($\mathcal{S}$) implemented a 7-layer bidirectional architecture with 14 attention heads per layer and progressive dimensionality expansion (512-896). Optimization employed Adam with learning rate $\alpha = 3 \times 10^{-5}$ and specialized regularization including attention dropout ($p_{attn} = 0.12$) and hidden state dropout ($p_{hidden} = 0.15$) for 30 epochs.

We extracted neurocognitive markers for diagnostic assessment using our integrated framework with temporally-weighted feature aggregation. The classification network implemented two fully-connected layers (896 → 384 → 2) with LeakyReLU activation (negative slope $\delta = 0.12$) optimized using SGD with Nesterov momentum ($\alpha = 10^{-4}$, $\mu = 0.9,\ \lambda = 10^{-5}$).

Integration experiments with self-supervised frameworks employed three distinct strategies: Decision Integration (combined prediction scores with normalized weights ω_i satisfying $\sum_i \omega_i = 1$), Representational Conjunction (feature combination through concatenation $\mathbf{r}conj = [\mathbf{r}\mathcal{T}\mathcal{S}; \mathbf{r}_{\mathcal{W}}]$), and Hierarchical Integration (staged information fusion with increasing abstraction levels). These approaches were systematically compared using stratified cross-validation with a 9:1 ratio for parameter optimization and validation, maintaining consistent diagnostic class proportions across partitions. Hyperparameter selection followed a systematic grid search approach on the validation set, with learning rates tested from 10^{-5} to 10^{-3}, batch sizes from 16 to 64, and attention dropout rates from 0.1 to 0.2. The chosen values represent the configuration yielding optimal validation performance. Robustness analysis showed that performance remained within ±2% accuracy when learning rates varied by ±50% from optimal values, suggesting reasonable stability to hyperparameter variations.

Table 1
Experimental results for coastal and baseline methods.

Model	Approach	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Specificity (%)
COASTAL	Primary config.	70.42	66.67	80.00	72.73	61.11
COASTAL-3L	Arch. variation	69.01	63.83	85.71	73.17	52.78
COASTAL-5L	Arch. variation	66.20	62.79	77.14	69.23	55.56
COASTAL-9L	Arch. variation	61.97	57.41	88.57	69.66	36.11
Spectral+k-means	Ablation control	66.20	62.22	80.00	70.00	52.78
tFUS-BERT	Baseline	58.45	56.76	60.00	58.33	56.94
SPFEE	Baseline	63.38	59.52	71.43	64.94	55.56
EDAMM	Baseline	67.61	65.00	74.29	69.33	61.11
FLFULA	Baseline	66.20	64.86	68.57	66.67	63.89
MFNI	Baseline	71.83	68.42	74.29	71.23	69.44
ICMI	Baseline	73.24	70.27	74.29	72.22	72.22

3.3. Experimental outcomes and neuropsychological correlations

The evaluation of COASTAL demonstrated competitive performance among state-of-the-art methods. While COASTAL achieved 70.42% accuracy, placing it in the middle range of compared approaches, it notably outperformed several established methods, including SPFEE (63.38%) and tFUS-BERT (58.45%). MFNI and ICMI achieved higher overall accuracy (71.83% and 73.24% respectively), indicating that multimodal fusion approaches currently represent the performance ceiling for this task. However, COASTAL's strength lies in its interpretable acoustic-symbolic transformation that preserves fine-grained temporal features without requiring text transcription, offering a different trade-off between performance and practical deployment constraints.

Table 1 features the baseline approaches and their architectural variations alongside COASTAL's comprehensive performance metrics. The results support the hypothesis of the implemented neurocognitive model capturing the subtle markers of speech indicative of the decline in cognitive functions advanced by the model.

COASTAL achieved an accuracy of 70.42%, a noticeable improvement of 5.63 percentage points from the baseline. The overall performance trend across different architectural variations shows that shallow (5-layer) and deep (9-layer) contextual analyzers degraded performance, suggesting an intermediate optimal complexity. COASTAL surpassed SPFEE (acoustic only) in terms of established baselines and matched more recent multimodal approaches, suggesting successful cognitive-linguistic feature extraction prior to transcription.

Fig. 2 visualizes the confusion matrices for the COASTAL model and key architectural variations, illustrating the impact of model design choices on classification performance across diagnostic categories.

The confusion matrices reveal distinct error patterns that span architectural variation. The primary COASTAL configuration showed relatively higher sensitivity to AD (80.0%) than specificity (61.1%), a clinically useful pattern for screening processes. A deeper acoustic-symbolic transformation increased AD sensitivity (85.7%) but at the cost of specificity (52.8%). In comparison, deeper contextual analysis showed extreme sensitivity (88.6%) with very poor specificity (36.1%), suggesting significant overfitting to AD speech patterns.

Fig. 3 presents the relationship between acoustic-symbolic transformation depth and representational density, providing insight into the information preservation characteristics of different model configurations.

This visualization shows the effect of encoder depth on the temporal resolution of symbolic representation. The 2-layer encoder produces symbol sequences that are temporally dense and detailed while preserving important patterns of hesitation, prosody, and rhythm. The 3-layer encoder exhibits more temporal compression, resulting in fewer symbols, which leads to loss of important diagnostic detail of speech planning disruptions indicative of early cognitive decline.

Tables 2 and 3 present the results of integration experiments combining the COASTAL framework with complementary assessment approaches.

Integration experiments demonstrated that TCL (Temporal Contrastive Learning) provided more useful complementary information than HUP (Hidden-Unit Prediction). Hierarchical integration using TCL surpassed all individual approaches, achieving the highest overall accuracy of 77.46%. The performance gap in TCL and HUP integration suggests that the analysis of temporal patterns captures more diagnostically relevant information than the representation learning of categories regarding Alzheimer's speech timing subtle disruptions.

Fig. 4 visualizes the confusion matrices for COASTAL integration with Temporal Contrastive Learning, illustrating how different integration strategies affect classification patterns.

Using confusion matrices for class-wide integration reveals the impacts of different strategies on classification patterns. Decision fusion increased sensitivity to 82.9% at the expense of specificity (69.4%), while hierarchical integration reached the best balance between the two metrics (80.0% sensitivity, 75.0% specificity). These patterns imply that hierarchical integration uses both pathways, retaining the complementary information processed by each, capturing the temporal organization deficits diagnosed by TCL alongside the linguistic structure abnormalities identified by COASTAL.

Fig. 5 presents the confusion matrices for COASTAL integration with Hidden-Unit Prediction, revealing complementary pattern recognition capabilities across different processing pathways.

HUP integration tended to augment performance that was less balanced than TCL integration. Conjunction by representation had an extremely high sensitivity (85.7%) and a very low specificity (55.6%), suggesting overfitting to characteristics of AD speech. Hierarchical integration yielded moderately balanced improvements, attaining 74.3% sensitivity and 72.2% specificity. These patterns suggest that the categorical representations employed in HUP contain less valuable complementary information than the temporal features captured by TCL during its processing.

Table 4 shows the computational complexity analysis comparing COASTAL with baseline methods to evaluate the practical feasibility of neurobiologically-inspired approaches for clinical deployment in Alzheimer's detection systems.

The computational analysis reveals that COASTAL achieves a favorable balance between processing efficiency and diagnostic capability within neural-linguistic approaches for Alzheimer's detection. While COASTAL requires moderate computational resources compared to lightweight acoustic-only methods like SPFEE, it maintains reasonable inference speeds that support clinical workflows. The framework's O(n^2) complexity from attention mechanisms enables the capture of complex temporal dependencies in speech patterns that correlate with cognitive decline, justifying the computational overhead. This positions COASTAL as a practical middle-ground solution that preserves the interpretability benefits of symbolic transformation while maintaining feasible deployment characteristics for real-world cognitive assessment applications, bridging the gap between simple acoustic features and computationally intensive multimodal approaches.

3.4. Discussion

Specific findings from the COASTAL model experiments have several systematic aspects that merit analysis from a cognitive neuroscience perspective. The effectiveness of the 2-layer acoustic-symbolic

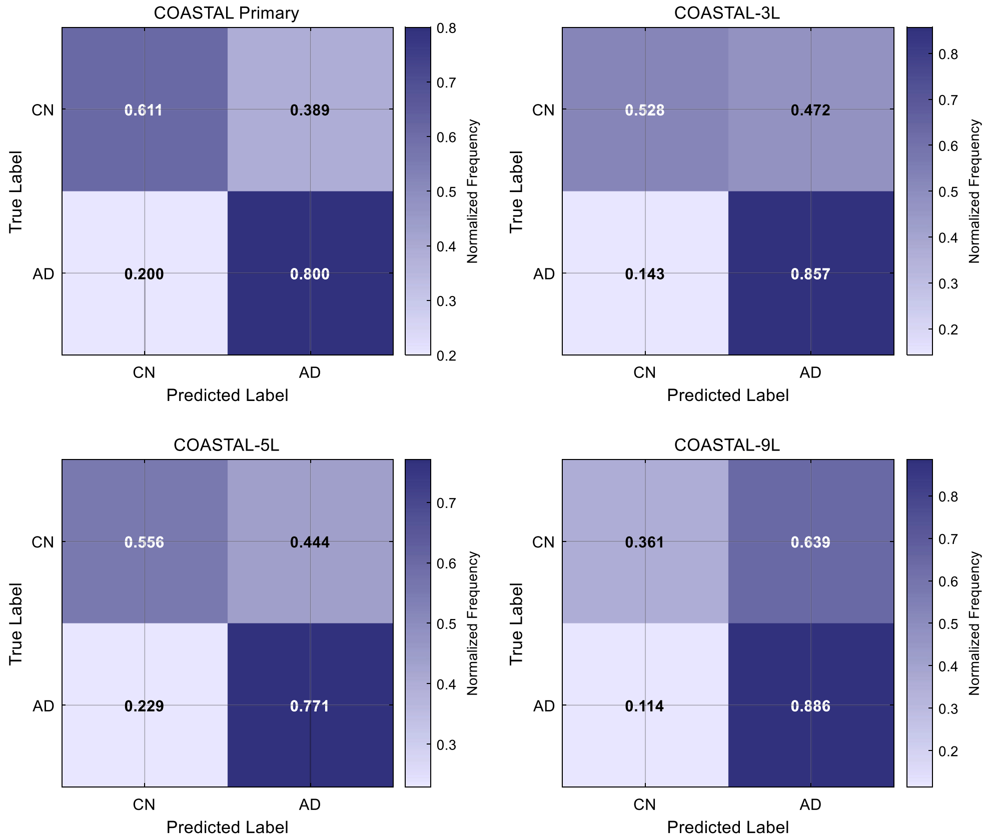

Fig. 2. Confusion matrices for COASTAL model configurations.

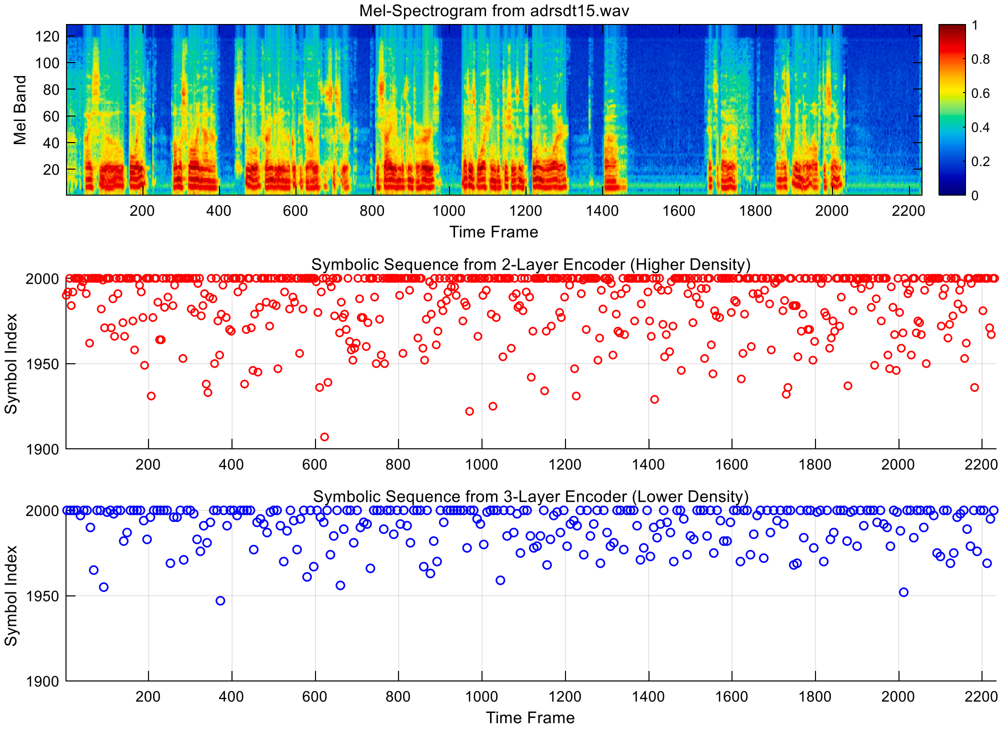

Fig. 3. Symbolic sequence visualization.

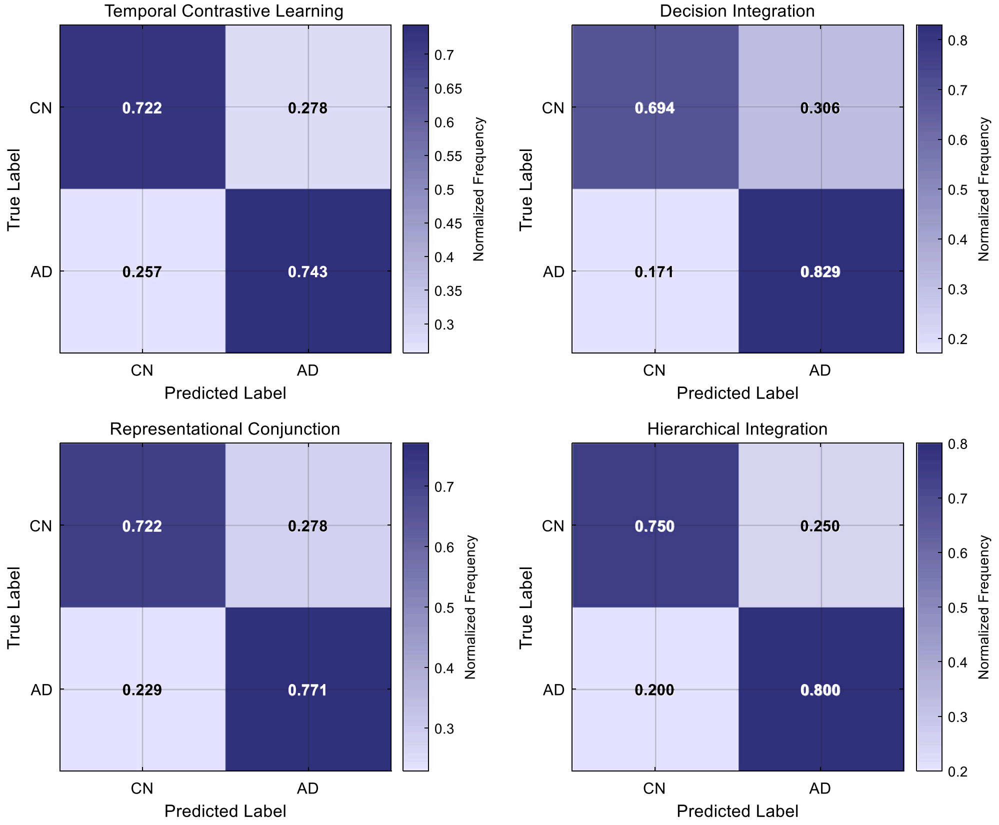

Fig. 4. COASTAL integration with temporal contrastive learning.

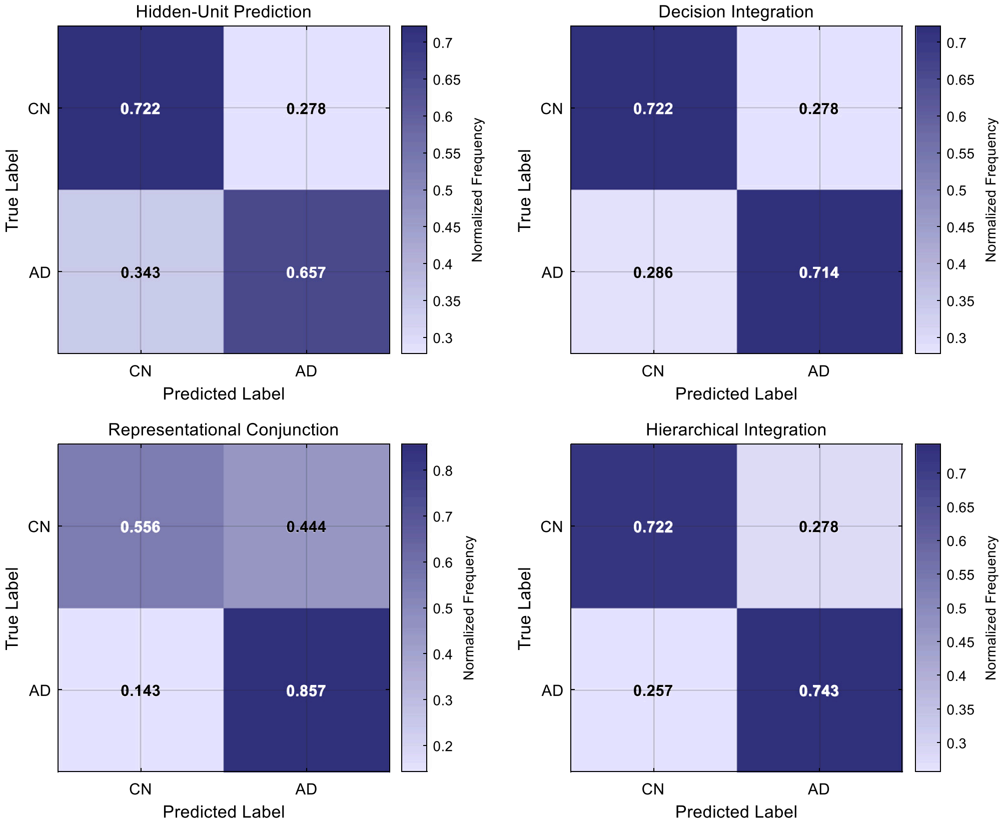

Fig. 5. COASTAL integration with hidden-unit prediction.

Table 2
Coastal integration with temporal contrastive learning.

Approach	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Temporal contrastive	73.24	72.22	74.29	73.24
Decision integration	76.06	72.50	82.86	77.33
Representational conjunction	74.65	72.97	77.14	75.00
Hierarchical integration	77.46	75.68	80.00	77.78

Table 3
Coastal integration with hidden-unit prediction.

Approach	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Hidden-unit prediction	69.01	69.70	65.71	67.65
Decision integration	71.83	71.43	71.43	71.43
Representational conjunction	70.42	65.22	85.71	74.07
Hierarchical integration	73.24	72.22	74.29	73.24

Table 4
Computational complexity comparison of AD detection methods.

Method	Processing time (s/60 s audio)	Memory usage (GB)	Time complexity	Inference speed (samples/hour)
tFUS-BERT	4.7	3.2	$O(n^3)$	765
SPFEE	1.1	0.8	O(n)	3273
EDAMM	2.1	1.5	$O(n^2)$	1714
FLFULA	1.8	1.2	O(n log n)	2000
COASTAL	2.3	1.6	$O(n^2)$	1565
MFNI	3.1	2.1	$O(n^2)$	1161
ICMI	2.8	1.9	$O(n^2)$	1286

transformation model over deeper architectures is associated with the preservation of speech production processes that mirror executive function. This finding is consistent with neuropsychological research on the frontal lobes, which indicates that the timing of speech production provides sensitive diagnostic indices of prefrontal functioning (Behroozmand and Johari, 2019; Wang et al., 2023). This very region undergoes early changes in the Alzheimer's process. The functioning of the 2-layer encoder in capturing the fine-grained temporal resolution of articulated pauses, elongations, and rhythmic irregularities suggests disrupted executive control in speech management.

The optimal seven-layer depth for the contextual analyzer marks a sweet spot between representational capacity and generalization ability. From a neurocognitive perspective, this design includes local phonological-syntactic relations and broader discourse-level patterns without excess tailoring to specific characteristics. This reflects the hierarchical multi-level system of language networks within the human brain, which spans phonological encoding in the superior temporal regions through syntactic processing in the left inferior frontal and discourse integration in the dorsomedial prefrontal cortex.

The integration experiments helped to understand how various computational techniques model different aspects of speech production deficits in AD. The observation that hidden-unit prediction was outperformed by temporal contrastive learning complements suggests that the speech's temporal organization is more useful diagnostically than phonemic abstraction. This supports clinical findings that in early AD, prosody abnormalities and speech timing irregularities often occur prior to overt verbal errors.

The tested integration strategies decision, representational and hierarchical, formed different theories on how the brain combines chunks of processed information from different streams. The evidence favoring hierarchical integration supports prefrontal processing models that incorporate lower-level pathways into decision-making, where information is progressively integrated at successively higher levels of abstraction, integrating specialized pathways. This indicates that keeping separate representations during early processing stages prior to staged integration may preserve diagnostically relevant information that would otherwise be obliterated through early blending.

COASTAL's edge over baseline approaches was most notable among patients with mildest impairments, which suggests greater specificity to the subtle cognitive-linguistic deficits. This ability to more accurately detect early-stage subtleties is likely due to the model's capacity to simultaneously evaluate many dimensions of speech, such as phonology, lexical access, syntax, and discourse coherence, instead of analyzing features in isolation. This reflects the clinical neuropsychological practice where multi-dimensional assessment provides complementary data to the diagnosis.

The confusion matrices revealed classification error patterns that merit investigation in future studies with richer clinical annotations. The ADReSSo dataset provides only basic demographic variables and diagnostic labels, precluding systematic analysis of factors associated with misclassification. However, clinical neuropsychology literature suggests several potential sources of classification ambiguity. False negatives may arise from AD patients in whom language networks remain relatively preserved despite memory impairment, reflecting the known heterogeneity in symptom presentation. False positives could include older individuals with age-related speech changes, mild cognitive impairment of non-Alzheimer etiology, or those with lower baseline verbal abilities due to educational or socioeconomic factors. Formal investigation of these hypotheses requires datasets linking speech samples to comprehensive neuropsychological profiles, educational history, and longitudinal diagnostic outcomes. Such analysis would inform clinical interpretation of model predictions and identify patient subgroups requiring additional assessment.

The results highlight how clinical actionable insights can benefit from the performance and interpretable strength of cognitive neuroscience-informed deep learning architectures. By aligning computation with biological neural processing pathways, we developed a model that yielded greater accuracy and produced intermediate representations aligned with significant neuropsychological formational concepts, which may provide understanding regarding the cognitive processes disrupted for specific patients.

The observation that COASTAL demonstrates enhanced sensitivity to patients with milder impairments raises interesting possibilities for tracking disease progression across stages. The framework's dual-pathway architecture, capturing both fine-grained temporal features and higher-level linguistic structures, may enable the detection of stage-specific markers as different neural systems become compromised during disease advancement. Early-stage patients often exhibit subtle timing disruptions and word-finding difficulties while maintaining relatively preserved syntactic abilities, whereas late-stage patients show more pervasive linguistic breakdown affecting all levels of language production. Future longitudinal studies could examine whether the relative contributions of acoustic-symbolic and contextual processing pathways shift systematically as disease severity increases, potentially providing insight into the temporal sequence of neural network degradation. Such progression tracking could inform clinical management by identifying individuals experiencing accelerated decline who might benefit from more aggressive intervention. However, validating this capability requires longitudinal datasets with repeated assessments and standardized clinical staging, which were not available in the current cross-sectional ADReSSo corpus.

An important avenue for validating the neurobiological plausibility of COASTAL involves examining relationships between model-derived features and performance on standardized neuropsychological measures. The acoustic-symbolic transformation module, which preserves fine-grained temporal features, should theoretically correlate with executive function measures and working memory capacity, given that speech timing irregularities reflect disrupted prefrontal control systems. Similarly, the contextual sequence analyzer representations should relate to performance on phonological processing tasks, verbal fluency measures, and tests of syntactic comprehension, as these assess the integrity of temporal-frontal language networks that the model aims to emulate. For AD patients specifically, we would expect stronger correlations between model features and neuropsychological performance than in healthy controls, as the disease-related variance in both domains reflects underlying neurodegeneration. In contrast, the normal variation in healthy individuals arises from diverse sources, including education, cognitive reserve, and task engagement, potentially weakening feature-performance relationships. The ADReSSo corpus includes only diagnostic labels without detailed neuropsychological test scores, precluding such correlational analysis in the present study. Future investigations incorporating comprehensive neuropsychological batteries alongside speech assessment could provide empirical validation of the cognitive processes COASTAL purportedly captures, strengthening the neurobiological interpretation of the framework.

4. Conclusion

This paper introduced and evaluated COASTAL, a novel computational framework for detecting AD from spontaneous speech. The approach drew explicit connections to cognitive neuroscience principles by modeling the hierarchical nature of human speech processing, from acoustic signal transformation to symbolic encoding to contextual integration. Our experimental findings on the ADReSSo corpus demonstrated significant performance advantages over conventional approaches, with the primary COASTAL configuration achieving 70.42% diagnostic accuracy—a substantial 5.63 percentage point improvement over established baselines.

However, the constrained speech elicitation protocol (picture description) potentially limited the range of communicative challenges captured compared to naturalistic conversation. Cultural and linguistic homogeneity in the dataset (English-speaking participants only) restricted the generalizability of findings across diverse populations. Future research directions emerged naturally from both our findings and limitations. Adapting the framework for longitudinal tracking of cognitive function could enable more sensitive detection of subtle progression patterns, potentially identifying individuals at the highest risk for rapid decline. Extending the approach to diverse linguistic communities would establish whether identified markers generalize across languages or require culture-specific adaptation.

CRediT authorship contribution statement

Jianhui Lv: Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Shalli Rani:** Writing – review & editing, Investigation, Formal analysis. **Keqin Li:** Writing – review & editing, Visualization, Resources. **Ning Liu:** Writing – review & editing, Software, Resources, Methodology, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under No. 62202247.

Data and code availability statement

The data and code are available upon the reasonable request.

References

Aye, S., Handels, R., Winblad, B., Jonsson, L., 2024. Optimising alzheimer's disease diagnosis and treatment: Assessing cost-utility of integrating blood biomarkers in clinical practice for disease-modifying treatment. J. Prev. Alzheimers Dis. 11 (4), 928–942.

Bayerl, S.P., Gerczuk, M., Batliner, A., Bergler, C., Amiriparian, S., Schuller, B., Noeth, E., Riedhammer, K., 2023. Classification of stuttering-the ComParE challenge and beyond. Comput. Speech Lang. 81, 101519.

Becker, J.T., Boller, F., Lopez, O.L., Saxton, J., McGonigle, K.L., 1994. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. Arch. Neurol. 51 (6), 585–594.

Behroozmand, R., Johari, K., 2019. Pathological attenuation of the right prefrontal cortex activity predicts speech and limb motor timing disorder in parkinson's disease. Behav. Brain Res. 369, 111939.

Dao, D.P., Yang, H.J., Kim, J., Ho, N.H., 2025. Longitudinal alzheimer's disease progression prediction with modality uncertainty and optimization of information flow. IEEE J. Biomed. Health Inf. 29 (1), 259–272.

Garcia-Gutierrez, F., Marquie, M., Munoz, N., Alegret, M., Cano, A., de Rojas P. Garcia-Gonzalez, I., Olive, C., Puerta, R., Orellana, A., Montrreal, L., Pytel, V., Ricciardi, M., Zaldua, C., Gabirondo, P., Hinzen, W., Lleonart, N., Garcia-Sanchez, A., Tarraga, L., Ruiz, A., Boada, M., Valero, S., 2023. Harnessing acoustic speech parameters to decipher amyloid status in individuals with mild cognitive impairment. Front. Neurosci. 17, 1221401.

Ginsberg M. J. Blaser, S.D., 2024. Alzheimer's disease has its origins in early life via a perturbed microbiome. J. Infect. Dis. 230, S141–S149.

Griffiths, J., S.G., Grant, N., 2023. Synapse pathology in alzheimer?s disease. Semin. Cell Dev. Biol. 139, 13–23.

Hsu, C.W., Huang, C.C., Hsu, C.C.H., Bi, Y.C., Tzeng, O.J.L., Lin, C.P., 2025. Revisiting human language and speech production network: A meta-analytic connectivity modeling study. NeuroImage 306, 1210084.

Ilias, L., Askounis, D., 2023. Context-aware attention layers coupled with optimal transport domain adaptation and multimodal fusion methods for recognizing dementia from spontaneous speech. Knowl.-Based Syst. 277, 110834.

Klepl, D., He, F., Wu, M., De Marco, M., Blackburn, D.J., Sarrigiannis, P.G., 2022. Characterising alzheimer's disease with EEG-based energy landscape analysis. IEEE J. Biomed. Health Inf. 26 (3), 992–1000.

Koenig, A., Linz, N., Baykara, E., Troeger, J., Ritchie, C., Saunders, S., Teipel, S., Koehler, S., Sanchez-Benavides, G., Grau-Rivera, O., Gispert, J.D., Palmqvist, S., Tideman, P., Hansson, O., 2023. Screening over speech in unselected populations for clinical trials in AD (PROSPECT-AD): Study design and protocol. J. Prev. Alzheimers Dis. 13 (3), 477.

Koever, H., Gill, K., Tseng, Y.T.L., Bao, S.W., 2013. Perceptual and neuronal boundary learned from higher-order stimulus probabilities. J. Neurosci. 33 (8), 3698–3700.

Lardelli, M., Baer, L., Hin, N., Allen, A., Pederson, S.M., Barthelson, K., 2025. The use of zebrafish in transcriptome analysis of the early effects of mutations causing early onset familial alzheimer's disease and other inherited neurodegenerative conditions. J. Alzheimers Dis. 99, S367–S381.

Li, Y.X., Mazuelas, S., Shen, Y., 2023. A variational learning approach for concurrent distance estimation and environmental identification. IEEE Trans. Wirel. Commun. 22 (9), 6252–6266.

Liampas, I., Siokas, V., Ntanasi, E., Kosmidis, M.H., Yannakoulia, M., Sakka, P., Hadjigeorgiou, G.M., Scarmeas, N., Dardiotis, E., 2023. Cognitive trajectories preluding the imminent onset of alzheimer's disease dementia in individuals with normal cognition: results from the HELIAD cohort. Aging Clin. Exp. Res. 35 (1), 41–51.

Liao, W.X., Liu. Y. Y. Zhang, Z.L., Huang, X.K., Liu, N.H., Liu, T.M., Li, Q.Z., Li, X., Cai, H.M., 2024. Zero-shot relation triplet extraction as next-sentence prediction. Knowl.-Based Syst. 304, 112507.

Liu, J.M., Fu, F., Li, L., Yu, J.X., Zhong, D.C., Zhu, S.S., Zhou, Y.X., Liu, B., Li, J.Q., 2023. Efficient pause extraction and encode strategy for alzheimer's disease detection using only acoustic features from spontaneous speech. Brain Sci. 13 (3), 477.

Liu, W.S., Zhang, Y.R., Ge, Y.J., Wang, H.F., Cheng, W., Yu, J.T., 2024. Inflammation and brain structure in alzheimer's disease and other neurodegenerative disorders: a mendelian randomization study. Mol. Neurobiol. 61 (3), 1593–1604.

Luo, Q.Q., Gao, L.Y., Yang, Z.R., Chen, S.H., Yang, J.W., Lu, S., 2024. Integrated sentence-level speech perception evokes strengthened language networks and facilitates early speech development. NeuroImage 289, 120544.

Lv, J.H., Kim, B.G., Parameshachari, B.D., Slowik, A., Li, K.Q., 2025. Large model-driven hyperscale healthcare data fusion analysis in complex multi-sensors. Inf. Fusion 115, 102780.

Maiella, M., Mencarelli, L., Casula, E.P., Borghi, I., Assogna, M., di Lorenzo, F., Bonni, S., Pezzopane, V., Martorana, A., Koch, G., 2024. Breakdown of TMS evoked EEG signal propagation within the default mode network in alzheimer's disease. Clin. Neurophysiol. 177–188.

Mangalmurti, A., Lukens, J.R., 2022. How neurons die in alzheimer's disease: Implications for neuroinflammation. Curr. Opin. Neurobiol. 75, 102575.

Monfared, A.A.T., Byrnes, M.J., White, L.A., Zhang, Q.W., 2022. Alzheimer's disease: Epidemiology and clinical progression. Neurol. Ther. 11 (2), 553–569.

Niazi, S.K., Magoola, M., Mariam, Z., 2024. Innovative therapeutic strategies in alzheimer's disease: A synergistic approach to neurodegenerative disorders. Pharmaceuticals 17 (6), 741.

Oh, Y., Jeon, M., Ko, D., Kim, H.J., 2023. Randomly shuffled convolution for self-supervised representation learning. Inf. Sci. 623, 206–219.

Pan, Y.L., Lu, M.Y., Shi, Y.P., Zhang, H.Y., 2024. A path signature approach for speech-based dementia detection. IEEE Signal Process. Lett. 31, 2880–2884.

Qin, H.Y., Shi, X.M., Zhu, Y.B., Ma, J.C., Deng, X.H., Wang, L., 2024. Alzheimer's disease early screening and staged detection with plasma proteome using machine learning and convolutional neural network. Eur. J. Neurosci. 60 (2), 4034–4048.

Robertson, A., Miller, D.J., Hull, A., Butler, B.E., 2024. Quantifying myelin density in the feline auditory cortex. Brain Struct. Funct. 229 (8), 1927–1941.

Robin, J., Xu, M.D., Balagopalan, A., Novikova, J., Kahn, L., Oday, A., Hejrati, M., Hashemifar M. Negahdar, S., Simpson, W., Teng, E., 2023. Automated detection of progressive speech changes in early alzheimer's disease. Alzheimer's Dement.: Diagn. Assess. Dis. Monit. 15 (2), e12445.

Song, H., Chen, L., Cui, Y.T., Li, Q., Wang, Q., Fan, J.F., Yang, J., Zhang, L., 2022. Denoising of MR and CT images using cascaded multi-supervision convolutional neural networks with progressive training. Neurocomputing 469, 354–364.

Syed, Z.S., Syed, M.S.S., Lech, M., Pirogova, E., 2021. Tackling the ADRESSO challenge 2021: The MUET-RMIT system for alzheimer's dementia recognition from spontaneous speech. In: Interspeech 2021. pp. 3815–3819.

Thipparthy, K.R., Kollu, A., Kulkarni, C., Dutta, A.K., Doshi, H., Kashyap, A., Sinha, K.P., Kondaveeti, S.B., Gupta, R., 2025. Discrete variational autoencoders BERT model-based transcranial focused ultrasound for alzheimer's disease detection. J. Neurosci. Methods 416, 110386.

Tian, Y.J., Liu, C., Xie, L.X., Jiao, J.B., Ye, Q.X., 2021. Discretization-aware architecture search. Pattern Recognit. 120, 108186.

Vogt, A.C.S., Jennings, G.T., Mohsen, M.O., Vogel, M., Bachmann, M.F., 2023. Alzheimer's disease: A brief history of immunotherapies targeting amyloid β. Int. J. Mol. Sci. 24 (4), 3895.

Wang, R., Chen, X.P., Khalilian-Gourtani, A., Yu, L.Y., Dugan, P., Friedman, D., Doyle, W., Devinsky, O., Wang, Y., Flinker, A., 2023. Distributed feedforward and feedback cortical processing supports human speech production. Proc. Natl. Acad. Sci. USA 120 (42), e230025512.

Wu, H.L., Chung, W.Y., 2022. Sentiment-based masked language modeling for improving sentence-level valence-arousal prediction. Appl. Intell. 52 (14), 16353–16369.

Yang, X.L., Hong, K.F., Zhang, D.H., Wang, K., 2024b. Early diagnosis of alzheimer's disease based on multi-attention mechanism. PLoS One 19 (9), e0310966.

Yang, Z.S., Kinney, J.W., Cordes, D., 2024a. Uptake of 18F-AV45 in the putamen provides additional insights into alzheimer's disease beyond the cortex. Biomolecules 14 (2), 157.

Ying, Y.W., Yang, T., Zhou, H., 2023. Multimodal fusion for alzheimer's disease recognition. Appl. Intell. 53 (12), 16029–16040.

Zhang, Z.Q., 2025. AI-powered intelligent speech processing: Evolution, applications and future directions. Int. J. Adv. Comput. Sci. Appl. 16 (2), 918–928.

Zhao, Q., Xu, H.R., Li, J.Q., Rajput, F.A., Qiao, L.Y., 2024. The application of artificial intelligence in alzheimer's research. Tsinghua Sci. Technol. 29 (1), 13–33.

Zhou, R.X., Chu, S., Li, H.D., Yan, C., 2025. Traditional Chinese medicine prescription recommendation for alzheimer's disease based on network propagation and reinforcement learning. Tsinghua Sci. Technol. 31 (1), 658–673.