Pattern Recognition 182 (2027) 114818

Contents lists available at ScienceDirect

Pattern Recognition

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

Cross-task generative augmentation of multimodal speech features for task-incomplete MCI detection☆

Hao Li [a], Bo Wang [a,*], Qiang He [c], Jun Mou [d], Junxin Chen [b,*]

[a] School of Information and Communication Engineering, Dalian University of Technology, Dalian, 116024, China
[b] School of Software, Dalian University of Technology, Dalian, 116621, China
[c] School of Computer Science and Engineering, Northeastern University, Shenyang, 110004, China
[d] School of Information Science and Engineering, Dalian Polytechnic University, Dalian, 116034, China

ARTICLE INFO

Keywords:
Mild cognitive impairment
Speech analysis
Generative modeling
Multimodal learning
Deep learning
Machine learning

ABSTRACT

Speech-based mild cognitive impairment (MCI) detection is constrained by limited data and incomplete language-task availability. Existing generative methods mainly augment observed tasks, leaving complementary information from unavailable tasks underexplored. We propose a cross-task generative augmentation framework that encodes linguistic, semantic, and acoustic features and generates missing target-task representations through a shared-private conditional variational autoencoder and a target-task-conditioned latent translator. Experiments used the Cookie Theft picture description (CT), Cinderella story recall (CIN), and Sandwich procedural narrative (SAN) tasks from the DementiaBank English Protocol Delaware Corpus. The Delaware evaluation comprised 250 subject-visit recordings from 202 unique participants, with CT, CIN, and SAN evaluated separately as the observed source task. External testing on the DementiaBank English Pitt Corpus used only CT and included 72 subject-visit recordings from 42 unique participants. Across the three Delaware source-task settings, adding two generated target-task feature sets increased mean balanced accuracy (BACC) from 62.41% to 69.19% and mean area under the receiver operating characteristic curve (AUC) from 64.84% to 73.44%. Without retraining on Pitt, the same augmentation increased BACC from 58.33% to 67.86% and AUC from 59.69% to 73.28%. The augmentation benefit varied with source-task informativeness, target-task complementarity, and classifier robustness. These results support cross-task feature generation for multimodal MCI detection when only one language task is observed. Source code is available at: https://github.com/DrBibobobi/speech-MCI.git.

1. Introduction

Alzheimer's disease (AD) and related dementias are prevalent neurodegenerative disorders associated with declines in memory, language, executive function, and daily functioning [1,2]. Mild cognitive impairment (MCI) is an intermediate stage between normal aging and dementia, and its early identification supports risk stratification, follow-up management, and potential intervention [3,4]. Conventional assessment relies on neuropsychological tests, clinical interviews, neuroimaging, and fluid biomarkers, but remains limited by cost, invasiveness, time, professional requirements, and scalability [5–7]. Language, speech, and other acoustic signals have shown promising performance in automated pattern-recognition applications, including affective analysis, communication understanding, and health monitoring [8,9]. Against this background, speech and language analysis has been explored as a non-invasive, low-cost, and repeatable digital biomarker for cognitive impairment detection and monitoring [10].

Current speech and language based studies often rely on limited elicitation tasks, especially picture description paradigms such as the Cookie Theft task (CT), which may restrict the characterization of heterogeneous MCI manifestations [11]. Multi task assessment provides complementary cognitive linguistic information but increases administration time, participant burden, transcription cost, and quality control requirements [12]. Existing multimodal fusion, cross-modality translation, domain adaptation, transfer learning, and synthetic augmentation methods mainly address modality integration, cross domain generalization, or sample level augmentation, with less attention to cross-task complementarity [13–15]. Moreover, excessive generation may introduce distribution shift and degrade classification performance [16].

☆ This article is part of a Special issue entitled: 'PR_Multimodal Perception and Pattern Recognition' published in Pattern Recognition.

* Corresponding authors.
E-mail addresses: Hao_Li@ieee.org (H. Li), bowang@dlut.edu.cn (B. Wang), heqiang@bmie.neu.edu.cn (Q. He), moujun@dlpu.edu.cn (J. Mou), junxinchen@ieee.org (J. Chen).

https://doi.org/10.1016/j.patcog.2026.114818
Received 30 May 2026; Received in revised form 31 July 2026; Accepted 24 August 2026
Available online 1 September 2026

Therefore, generating complementary discriminative information from unobserved task perspectives remains challenging when only a single real task is available.

To address this problem, this paper proposes a cross-task generative augmentation framework for MCI detection. The proposed framework learns source to target task mappings in multimodal feature space and generates complementary target task features from an observed real language task. Experiments are conducted on the Delaware Corpus from the DementiaBank English Protocol, using CT, the Cinderella story recall task (CIN), and the Sandwich procedural narrative task (SAN) as three representative connected speech paradigms [12]. This study aims to learn cross-task mappings, generate missing-task multimodal representations from a single observed task, and evaluate their effectiveness and cross-corpus generalizability for MCI detection. Different source and target task combinations are evaluated to examine whether generated target task features improve MCI detection when only a single real task is available.

The main contributions of this paper are summarized as follows:

- Propose a cross-task generative augmentation framework that generates missing task features from a single observed language task.
- Design a classification-oriented feature generator that learns cross-task mappings and produces complementary multimodal representations.
- Establish a validation pipeline covering three speech tasks, four feature types, and multiple machine learning and deep learning classifiers.
- Demonstrate that augmentation gains depend on source-task informativeness, target-task complementarity, and classifier robustness.

2. Related work

2.1. Speech-based MCI detection

Existing speech-based AD and MCI studies can first be grouped by feature type. Acoustic-prosodic work examines pauses, speech rate, pitch, spectral or cepstral measures, and voice quality, whereas textual-linguistic work quantifies lexical, syntactic, semantic, and discourse changes [11,17]. Examples include multivariate word properties in verbal-fluency tasks [18] and a transformer-derived Language Informativeness Index for reduced informativeness across English and Persian speech [19]. Multimodal studies combine acoustic and linguistic evidence to capture complementary manifestations of impairment [20, 21].

Studies can also be grouped by task paradigm. Pitt and ADReSS mainly use CT picture description, whereas Delaware spans picture description, story recall, procedural narration, and personal narratives [12,20]. Verbal fluency and scene construction provide other elicitation settings, while ADReSS-M and MultiConAD extend evaluation across languages, datasets, and conversational tasks [11,18]. Because cognitive demands and observable biomarkers vary by elicitation condition, findings from one task may not transfer directly to another task or language.

Modeling approaches progress from handcrafted features with conventional classifiers to self-supervised representations and multimodal deep networks [22]. Dual-dropout ranking selects compact biomarker subsets [23]; transformer systems learn acoustic and textual representations and fuse them through gated units or cross-modal attention [20]; and CogniAlign performs word-level audio-text alignment before gated cross-attention [21]. However, small cohorts, task and language heterogeneity, modality imbalance, and limited interpretability remain concerns [17], motivating the use of explainable models in multilingual classification [24].

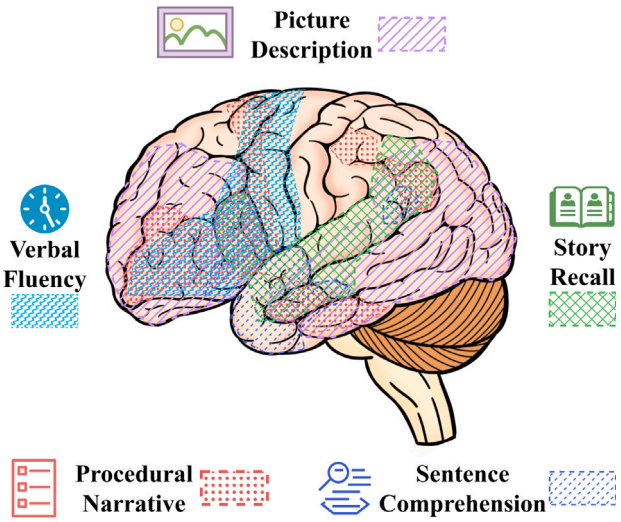


Fig. 1. Schematic illustration of language paradigms and neurocognitive systems.

2.2. Task effects and cross-task modeling

Language processing coordinates auditory perception, language production, semantic and syntactic processing, executive control, and memory [25]. Neuroimaging studies suggest that different language activities engage distinct but overlapping networks [25]. As shown in Fig. 1, connected speech paradigms impose different neurocognitive demands [12]: picture description emphasizes visual-semantic mapping, story recall involves episodic memory and discourse organization, and procedural narration requires planning and sequential organization. These task differences may influence observable language behavior and diagnostic sensitivity.

In connected speech research on AD and MCI, task paradigms influence diagnostic sensitivity and observable language features [26]. Stark et al. found that core lexical items in the Cinderella Story better distinguished MCI from HC, potentially because story recall imposes greater memory demands without continuous visual support [26]. Bae et al. similarly showed that high-memory-load recall tasks enhanced AD-related acoustic differences and improved AD classification and MMSE prediction [27]. Task type, memory load, sample length, and elicitation method may shape the disease-related cues available to classifiers.

2.3. Generative feature augmentation

Generative augmentation has been applied to speech based AD and MCI analysis to mitigate data limitations by constructing additional samples or feature patterns. Han et al. proposed a large language model based data generation framework to address limited MCI samples and class imbalance, and examined its effects on MCI detection sensitivity and speech language markers [15]. Zolnour et al. proposed LLMCARE, which integrates Transformer based representations, handcrafted linguistic features, and large language model generated synthetic transcripts for early cognitive impairment detection [16]. Mirgbagheri et al. proposed a mimicry based transcript generation method by injecting atypical linguistic feature distributions into typical speech transcripts [28]. These studies suggest that generative augmentation can provide additional training information for speech based MCI detection under data limited conditions [15,29].

Existing generative augmentation methods mainly address data scarcity by expanding samples within the observed task or by simulating task specific language patterns, rather than completing classification related information from another language task perspective

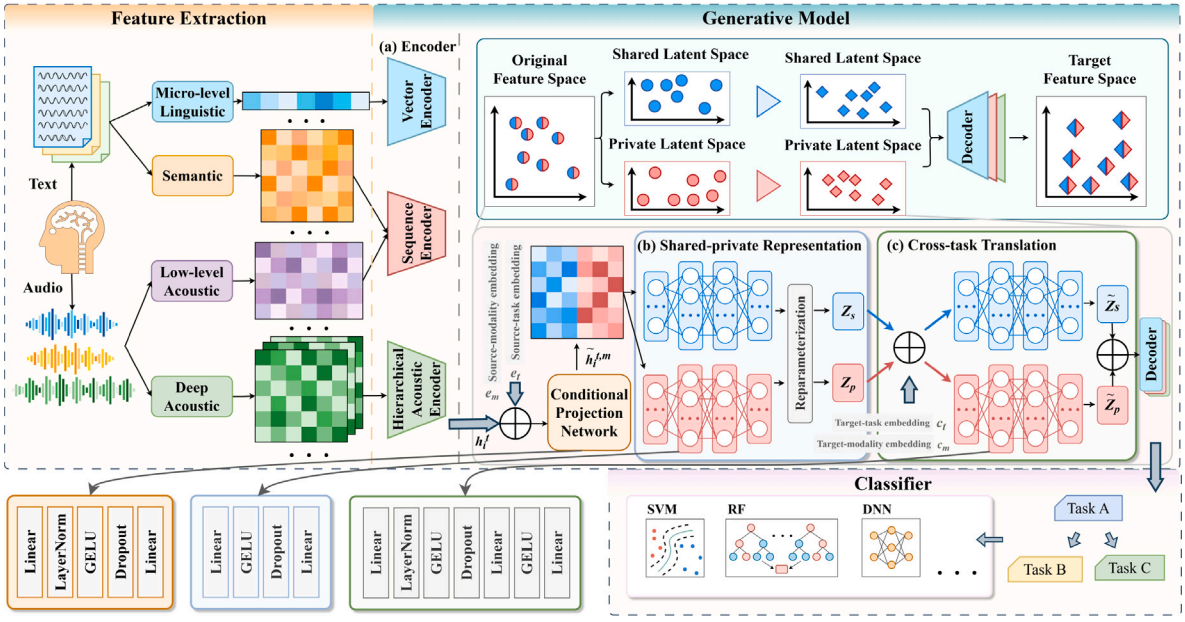


Fig. 2. Overall framework of the proposed cross-task multimodal feature generation and augmentation method for speech based MCI detection. (a) Modality-specific feature encoding. (b) Shared-private latent representation. (c) Cross-task translation.

when the target task is unavailable [28]. Therefore, this study focuses on cross-task feature generation and augmentation under task incomplete conditions, using the multimodal features of an observed source language task to generate target task features for MCI detection [12, 26].

3. Model architecture

As shown in Fig. 2, this study proposes a cross-task multimodal feature generation and augmentation framework for speech-based MCI detection. The framework contains four main components: modality-specific feature encoding, shared-private latent representation, target-task-conditioned cross-task translation, and downstream classification. Given one observed source language task, the framework performs feature-level generation to improve cross-task mapping stability and reduce nuisance variation associated with raw waveforms and full transcripts, while retaining disease-related acoustic and linguistic information in the small clinical cohort. The model encodes the heterogeneous features, decomposes them into shared and private latent representations, translates the latent variables into target-task-conditioned representations, and decodes them into complementary target-task features. The real source-task features and generated target-task features are then combined for MCI detection.

3.1. Task feature encoding

This study uses three connected speech paradigms, and four types of features are extracted for each task. The micro level linguistic features are extracted from CHAT transcripts and include fluency, turn organization, lexical, syntactic, and lexical-richness measures. RoBERTa is used to extract contextual semantic embeddings from participant turns [30], openSMILE is used to extract low level acoustic descriptors [31], and VGGish is used to obtain segment level deep acoustic representations [32].

For the ith sample in task t, the four features are denoted as $x^t_{i,\mathrm{mic}}$, $x^t_{i,\mathrm{sem}}$, $x^t_{i,\mathrm{lld}}$, and $x^t_{i,\mathrm{deep}}$. As shown in Fig. 3, they are encoded by separate paths according to their data structures. Here, mic denotes the 256-dimensional micro-level linguistic feature vector, sem denotes the 768-dimensional turn-level contextual semantic representation extracted by RoBERTa, lld denotes the 25-dimensional openSMILE low-level descriptor vector at each frame, and deep denotes the 128-dimensional VGGish deep acoustic embedding for each segment.

The micro level linguistic feature is a fixed dimensional vector and is encoded by a multilayer perceptron:

$$h^t_{i,\mathrm{mic}} = E_{\mathrm{mic}}(x^t_{i,\mathrm{mic}}). \tag{1}$$

In Eq. (1), $E_{\mathrm{mic}}(\cdot)$ consists of linear layers, layer normalization, GELU activation, and dropout, and maps micro level linguistic indicators into a unified latent representation.

The contextual semantic sequence consists of turn level RoBERTa representations [30]. A bidirectional gated recurrent unit followed by masked attention pooling is used:

$$U^t_{i,\mathrm{sem}} = \mathrm{BiGRU}_{\mathrm{sem}}(x^t_{i,\mathrm{sem}}), \quad h^t_{i,\mathrm{sem}} = \mathrm{AttPool}(U^t_{i,\mathrm{sem}}, r^t_{i,\mathrm{sem}}). \tag{2}$$

where $r^t_{i,\mathrm{sem}}$ denotes the turn level valid-position mask used to exclude padded positions.

The low level acoustic sequence is represented as $x^t_{i,\mathrm{lld}} \in \mathbb{R}^{W^t_i \times F^t_i \times d_{\mathrm{lld}}}$, where W^t_i and F^t_i denote the numbers of valid windows and frames, respectively. A hierarchical encoder is used to model its frame level and window level structures:

$$U^t_{i,w} = \mathrm{GRU}_f\left(\psi_f(x^t_{i,\mathrm{lld},w,1:F^t_i})\right), \quad w = 1, \ldots, W^t_i. \tag{3}$$

Masked attention pooling is then applied to the frame level hidden states within each window to obtain a window representation $g^t_{i,w}$. Finally, the sequence of window representations is processed by a task level gated recurrent unit and window level attention pooling to obtain the hierarchical acoustic representation:

$$h^t_{i,\mathrm{lld}} = \mathrm{AttPool}\left(\mathrm{GRU}_w(g^t_{i,1:W^t_i}), r^t_{i,\mathrm{lld}}\right). \tag{4}$$

In Eq. (4), $r^t_{i,\mathrm{lld}}$ denotes the window level valid-position mask. This hierarchical structure models both short-term acoustic dynamics within windows and cross-window temporal variations across the task.

The deep acoustic sequence consists of segment level VGGish representations [32], and is encoded using a bidirectional gated recurrent unit followed by masked attention pooling:

$$U^t_{i,\mathrm{deep}} = \mathrm{BiGRU}_{\mathrm{deep}}(x^t_{i,\mathrm{deep}}), \quad h^t_{i,\mathrm{deep}} = \mathrm{AttPool}(U^t_{i,\mathrm{deep}}, r^t_{i,\mathrm{deep}}). \tag{5}$$

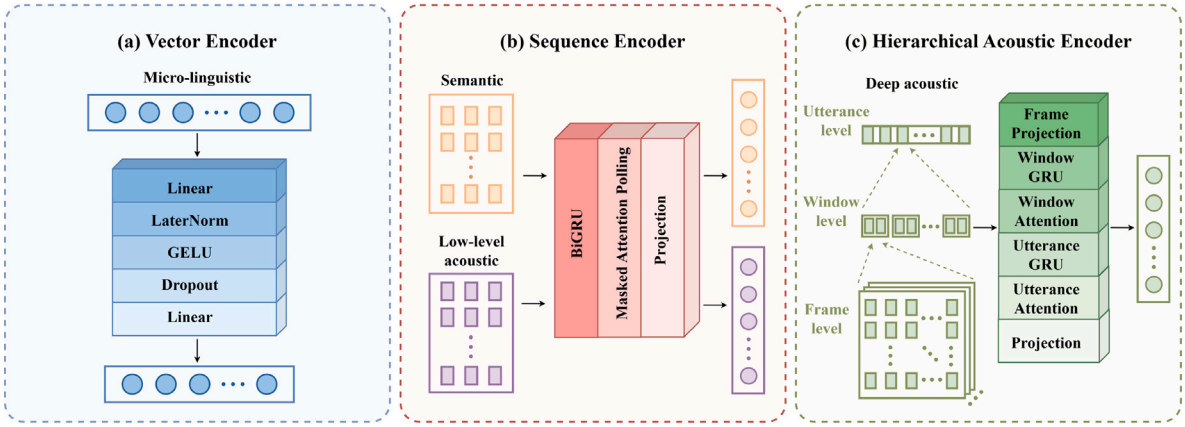


Fig. 3. Task feature encoding module. (a) Vector encoder for micro level linguistic features. (b) Sequence encoder for contextual semantic embeddings and deep acoustic embeddings. (c) Hierarchical acoustic encoder for low level acoustic sequences.

where $r^t_{i,\text{deep}}$ denotes the segment level valid-position mask.

The resulting modality specific representations, $h^t_{i,\text{mic}}$, $h^t_{i,\text{sem}}$, $h^t_{i,\text{lld}}$, and $h^t_{i,\text{deep}}$, are passed to the shared-private latent representation module. Masks are used for sequence features to exclude padded positions during pooling, reconstruction, and loss computation.

3.2. Shared-private representation

As illustrated in Fig. 2(b), the encoded task and modality representation is decomposed into shared and private latent variables. A standard variational autoencoder usually maps an input sample to a single latent variable, which may mix disease related information with task-, modality-, and individual-specific variations. To reduce this mixture, this study adopts a shared-private latent decomposition structure. The shared latent variable z_s is designed to capture relatively stable MCI detection related information across tasks and modalities, whereas the private latent variable z_p models non-shared variations related to task expression, modality structure, and individual differences.

For the encoded representation $h^{t,m}_i$ from task t and modality m, a task embedding e_t and a modality embedding e_m are introduced to indicate the current language task and feature modality. The encoded representation is concatenated with the 32-dimensional task embedding and the 16-dimensional modality embedding and fed into a modality-conditioned projection network:

$$\tilde{h}^{t,m}_i = \phi_m([h^{t,m}_i; e_t; e_m]). \tag{6}$$

In Eq. (6), $\phi_m(\cdot)$ denotes the conditional projection network after the modality specific encoder. It consists of linear layers, layer normalization, GELU activation, and dropout, and projects the conditional representation into a unified space for posterior estimation.

Given $\tilde{h}^{t,m}_i$, two posterior branches estimate the shared and private latent distributions, respectively. The shared branch outputs $\mu^{t,m}_s$ and $\log(\sigma^{t,m}_s)^2$, while the private branch outputs $\mu^{t,m}_p$ and $\log(\sigma^{t,m}_p)^2$. The two posterior distributions are defined as

$$\begin{aligned} q^s_\phi(z_s|x^{t,m}_i, t, m) &= \mathcal{N}\left(\mu^{t,m}_s, \text{diag}\left((\sigma^{t,m}_s)^2\right)\right), \\ q^p_\phi(z_p|x^{t,m}_i, t, m) &= \mathcal{N}\left(\mu^{t,m}_p, \text{diag}\left((\sigma^{t,m}_p)^2\right)\right). \end{aligned} \tag{7}$$

The shared variable z_s in Eq. (7) is constrained by disease classification supervision, shared latent alignment, and task and modality adversarial constraints, so that it retains disease related information while reducing task and modality separability. In contrast, z_p is constrained by task and modality classification heads to preserve task specific and modality specific variations. The private latent variable is passed through task and modality classification heads to minimize the corresponding cross-entropy losses and emphasize task- and modality-specific variations. Disease classification supervision is applied only to the shared latent variable to prevent direct disease-label gradients from entering the private branch and reduce redundant disease-related information in the private representation.

Direct sampling from the posterior distributions would block gradient propagation because latent variable sampling is stochastic. Following the reparameterization strategy of Auto-Encoding Variational Bayes [33], the shared and private latent variables are sampled as

$$z_s = \mu_s + \sigma_s \odot \epsilon_s, \quad z_p = \mu_p + \sigma_p \odot \epsilon_p, \quad \epsilon_s, \epsilon_p \sim \mathcal{N}(0, I). \tag{8}$$

Eq. (8) transfers stochasticity to the noise variables ϵ_s and ϵ_p, allowing the posterior parameters to be optimized through end to end backpropagation.

This shared-private structure provides explicit latent inputs for cross-task translation. The shared latent variable maintains MCI detection related information, while the private latent variable provides task- and modality-conditioned adaptation information.

3.3. Cross-task translation

As shown in Fig. 2(c), cross-task translation is performed in the latent space. After shared-private representation modeling, a source task sample is represented by z_s and z_p. Directly decoding these variables would mainly reconstruct source task-like features. Therefore, this study introduces a target task conditional translator to convert source task latent representations into target task conditioned representations.

Given a source task s, a target task t, and a modality m, the model first constructs a target condition vector: $c^{t,m} = [e_t; e_m]$, where e_t and e_m denote the target task embedding and modality embedding, respectively. This condition specifies both the target language paradigm and the feature modality.

Separate translators are used for the shared and private latent variables. The shared translator takes $z^{s,m}_{s,i}$ and $c^{t,m}$ as input and outputs the target task conditioned shared latent variable $\tilde{z}^{t,m}_{s,i}$. The private translator maps $z^{s,m}_{p,i}$ to the corresponding target task conditioned private latent variable $\tilde{z}^{t,m}_{p,i}$:

$$\tilde{z}^{t,m}_{s,i} = T_s([z^{s,m}_{s,i}; c^{t,m}]), \quad \tilde{z}^{t,m}_{p,i} = T_p([z^{s,m}_{p,i}; c^{t,m}]). \tag{9}$$

In Eq. (9), $T_s(\cdot)$ and $T_p(\cdot)$ denote the shared and private latent translators. Both are implemented as multilayer perceptrons composed of linear layers, layer normalization, GELU activation, and dropout.

The shared translator adjusts the target task expression of disease related information, whereas the private translator adapts task- and modality specific variations. The translated variables $\tilde{z}^{t,m}_{s,i}$ and $\tilde{z}^{t,m}_{p,i}$ from Eq. (9) are then concatenated and sent to the corresponding modality specific decoder to generate target task features.

Table 1
Participant demographics.

Group	Recordings	Participants	Age	Education			MoCA-Blind[a]
				E1	E2	E3	
HC	125(M:83 F:42)	101(M:70 F:31)	72.67 ± 6.11	9	26	90	20.17 ± 1.38
MCI	125(M:83 F:42)	101(M:69 F:32)	72.67 ± 7.43	8	29	88	17.37 ± 1.77

F: Female, M: Male. Age and MoCA-Blind: mean ± standard deviation. E1: high school or below; E2: some college, technical school, or associate degree; E3: bachelor's degree or above.
[a] MoCA-Blind scores were available for 89 HC and 68 MCI subject-visit samples.

Table 2
Summary statistics of the three connected speech paradigms.

Task	Recording duration (s)	Participant word count	Participant turns
CT	58.00 ± 41.77	118.97 ± 81.32	13.53 ± 8.84
CIN	204.40 ± 124.15	420.67 ± 264.87	41.35 ± 28.41
SAN	42.02 ± 27.62	103.38 ± 69.32	11.09 ± 6.60

Values are reported as mean ± standard deviation.

3.4. Decoding and optimization

As illustrated in Fig. 2, the target task conditioned latent variables are decoded into modality specific target task features:

$$\hat{x}_i^{t,m} = D_m([\tilde{z}_{s,i}^{t,m}; \tilde{z}_{p,i}^{t,m}]). \tag{10}$$

In Eq. (10), $D_m(\cdot)$ denotes the decoder for modality m. A multilayer perceptron decoder is used for fixed dimensional linguistic statistical features. Gated recurrent unit decoders are used for contextual semantic embedding sequences and deep acoustic embedding sequences. Window-level and frame level recurrent decoders are used for low level acoustic descriptor hierarchical sequences. For sequence features, target task sequence lengths and valid-position masks are used to preserve structural consistency.

The model is trained with a multi-objective loss consisting of reconstruction, variational regularization, disease supervision, shared latent alignment, task disentanglement, modality disentanglement, and adversarial feature distribution constraints:

$$\begin{aligned}\mathcal{L} = {} & \lambda_{rec}\mathcal{L}_{rec} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{dis}\mathcal{L}_{dis} + \lambda_{align}\mathcal{L}_{align} \\ & + \lambda_{task}\mathcal{L}_{task} + \lambda_{mod}\mathcal{L}_{mod} + \lambda_{adv}\mathcal{L}_{adv}.\end{aligned} \tag{11}$$

In Eq. (11), paired real target-task features from participants in each development fold training subset provide direct supervision for the generated features during training. $\mathcal{L}_{rec}$ is computed between the generated and corresponding real target-task features using mean squared error for fixed-dimensional vectors, masked mean squared error for two-dimensional sequences, and masked Smooth L1 loss for hierarchical low-level acoustic sequences. $\mathcal{L}_{KL}$ is computed as the Kullback–Leibler divergence between the approximate shared and private latent distributions and the standard Gaussian prior. $\mathcal{L}_{dis}$ is computed as the cross-entropy loss between the disease prediction obtained from the shared latent representation and the corresponding diagnostic label. $\mathcal{L}_{align}$ is computed as the mean squared difference between the source shared representation and its target-task-conditioned shared representation for the same participant. $\mathcal{L}_{task}$ and $\mathcal{L}_{mod}$ are computed using cross-entropy losses for task and modality prediction from the shared and private latent representations. For the shared representation, gradients from the task and modality classifiers are reversed through a gradient reversal layer. The private representations are optimized normally to predict the corresponding task and modality labels [34]. $\mathcal{L}_{adv}$ is computed using binary cross-entropy between the discriminator predictions for real and generated target-task features. The generator is optimized to make the generated features indistinguishable from the real target-task features. At inference, the trained generator uses only the observed source features and the specified target-task and modality identifiers, without access to real target-task features or class labels.

Since the training objective is not limited to reconstruction, generated features may have scale shifts relative to real features in the fixed dimensional statistical feature space. Therefore, this study applies modality-level feature space calibration:

$$\bar{x}_{gen}^{t,m} = \left(\frac{x_{gen}^{t,m} - \mu_{gen}^{t,m}}{\sigma_{gen}^{t,m} + \epsilon}\right)\sigma_{real}^{t,m} + \mu_{real}^{t,m}. \tag{12}$$

In Eq. (12), $\mu_{real}^{t,m}$ and $\sigma_{real}^{t,m}$ are estimated globally from the real feature blocks in the complete development set after cross-validation and model selection, while $\mu_{gen}^{t,m}$ and $\sigma_{gen}^{t,m}$ are estimated from generated feature blocks in the development set. These calibration parameters are fixed during test stage inference.

4. Experiments and results

4.1. Experimental setup

Dataset: This study used two datasets: the Delaware Corpus and the Pitt Corpus from the DementiaBank English Protocol, both of which contain audio recordings and CHAT transcripts from individuals with MCI and HC under connected speech protocols [12,35]. In the Delaware Corpus, three connected speech paradigms were used: the Cookie Theft picture description task (CT), the Cinderella story recall task (CIN), and the Sandwich procedural narrative task (SAN), corresponding to picture description, story recall, and procedural narration, respectively. In the Pitt Corpus, the CT task was used as an external test dataset to evaluate out-of-domain generalization under a source-only setting.

In the Delaware Corpus, samples were retained when complete task data, valid CHAT transcripts, and locatable timestamps were available. Participant demographics are summarized in Table 1. The Delaware dataset was split at the participant level into a development set and an independent test set. The development set contained 174 participants and 214 subject-visit records and was used for k-fold cross-validation, while the independent test set contained 28 participants and 36 subject-visit records. Task statistics for the three Delaware speech paradigms are reported in Table 2.

Implementation details: Participant speech was located using task markers, speaker labels, and time-alignment annotations in CHAT transcripts. Only participant turns were retained. Speech segments were split according to inter-turn pauses and examiner turns, and segments shorter than 2000 ms were excluded from acoustic extraction to remove mostly non-informative fragments, such as isolated fillers, brief breath sounds, background noise, and unstable vocalizations, thereby reducing acoustic noise and improving feature-extraction reliability. The remaining audio was enhanced using DeepFilterNet and converted to mono-channel, 16 kHz, PCM-16 WAV format.

The generator follows the architecture described in Section 3. It was trained with AdamW using an initial learning rate of 3×10^{-4}, a weight decay of 1×10^{-4}, a batch size of 4, and up to 180 epochs. Cosine learning rate scheduling, KL annealing from epoch 1 to 40, discriminator activation from epoch 15, and early stopping with a patience of 10 were applied. Linear KL warm-up prevented early regularization from overwhelming reconstruction; pilot runs showed poorer reconstruction stability with full early KL. Delaying the discriminator

Table 3
Hyperparameter settings for the seven downstream classifiers under different task configurations.

Model	CT	CIN	SAN	CT+CIN	CT+SAN	CIN+SAN	CT+CIN+SAN
LR	C:5 penalty:L1	C:0.01 penalty:L2	C:0.1 penalty:L1	C:0.01 penalty:L2	C:0.01 penalty:L2	C:0.1 penalty:L1	C:0.1 penalty:L1
RF	n:200 d:6, leaf:2 split:2 feat:log2	n:300 d:4, leaf:1 split:6 feat:sqrt	n:500 d:8, leaf:2 split:2 feat:log2	n:300 d:6, leaf:4 split:2 feat:log2	n:500 d:8, leaf:1 split:2 feat:sqrt	n:300 d:10, leaf:4 split:2 feat:log2	n:300 d:4, leaf:1 split:2 feat:log2
SVM	C:1, γ:scale	C:0.5, γ:scale	C:1, γ:scale	C:0.5, γ:auto	C:5, γ:auto	C:1, γ:scale	C:1, γ:auto
MLP	h: (256, 128) α: 10^{-3}, b:32 lr: 5×10^{-4}	h: (256, 128) α: 10^{-2}, b:32 lr: 5×10^{-4}	h: (256, 128) α: 10^{-4}, b:32 lr: 10^{-3}	h: (128) α: 10^{-4}, b:32 lr: 5×10^{-4}	h: (128) α: 10^{-2}, b:32 lr: 10^{-3}	h: (128) α: 10^{-4}, b:16 lr: 5×10^{-4}	h: (128) α: 10^{-4}, b:16 lr: 5×10^{-4}
AB	n:100, lr:0.5	n:50, lr:0.05	n:200, lr:0.1	n:200, lr:0.03	n:200, lr:0.03	n:100, lr:0.1	n:300, lr:0.05
GB	n:200, d:3 lr:0.03 sub:1.0	n:100, d:3 lr:0.03 sub:0.8	n:300, d:3 lr:0.1 sub:0.8	n:200, d:4 lr:0.05 sub:0.8	n:300, d:4 lr:0.03 sub:0.8	n:100, d:4 lr:0.1 sub:0.8	n:200, d:3 lr:0.03 sub:0.8
DNN	d_{in}:4040 h: (256, 128) drop:0.3 b:32, e:2 wd: 10^{-3}	d_{in}:4040 h: (512, 256, 128) drop:0.2 b:32, e:2 wd: 10^{-4}	d_{in}:4040 h: (256, 128) drop:0.3 b:16, e:1 wd: 10^{-4}	d_{in}:8080 h: (256, 128) drop:0.2 b:16, e:1 wd: 10^{-4}	d_{in}:8080 h: (256, 128) drop:0.3 b:16, e:1 wd: 10^{-4}	d_{in}:8080 h: (512, 256, 128) drop:0.2 b:16, e:1 wd: 10^{-3}	d_{in}:12120 h: (512, 256, 128) drop:0.3 b:16, e:1 wd: 10^{-3}

C, γ, α, and penalty denote inverse regularization strength, RBF-kernel coefficient, L2 regularization strength, and regularization type. n, d, feat, split, leaf, h, lr, b, sub, d_{in}, drop, wd, and e denote the number of estimators, maximum depth, maximum features, minimum split size, minimum leaf size, hidden-layer sizes, learning rate, batch size, subsampling ratio, input dimension, dropout, weight decay, and final training epochs, respectively. LR and SVM use balanced class weights; MLP and DNN use early stopping.

avoided early adversarial oscillation and introduced distribution matching after initial reconstruction convergence. Neither setting used test data. Loss weights were selected by participant-level cross-validation. During generator training, paired source- and target-task features from the same participant were used only within the development set to supervise the learning of cross-task mappings, including reconstruction and distribution-matching objectives. During inference, the generator synthesized missing target-task representations using only the observed source-task features and the specified target-task identifier, and did not have access to real target-task features or class labels.

Seven downstream classifiers were evaluated: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP), AdaBoost (AB), Gradient Boosting (GB), and a deep neural network (DNN). The selected model configurations and hyperparameters are reported in Table 3. Sequence features were summarized over valid positions, and all task and modality features were concatenated in a fixed order.

Evaluation protocol: Generator configurations were selected by 5-fold cross-validation. Downstream classifiers were tuned by 10-fold cross-validation within the development set, retrained on the full development set, and evaluated on the independent test set. All cross-validation folds and the independent test split were constructed at the participant level. On the Delaware dataset, each of CT, CIN, and SAN was used once as the real source task, with the other two tasks treated as generated target tasks. For each source setting, we evaluated Source only, source plus one generated target task, and source plus two generated target tasks. On the Pitt external dataset, only the CT task was used as the observed source task, and the missing target-task representations were generated from the available CT features to evaluate out-of-domain source-only generalization. Sensitivity (SEN), Specificity (SPE), F1-score (F1), Balanced Accuracy (BACC), and Area Under the ROC Curve (AUC) were reported.

4.2. Overall classification results

Table 4 shows consistent improvements from cross-task generative augmentation across all three source settings. Averaged across source tasks, BACC increased from 62.41% for Source only to 67.30% with G1 and 69.19% with G2, while AUC increased from 64.84% to 71.32% and 73.44%, respectively. All six augmented settings achieved statistically significant improvements in BACC and AUC over their corresponding source-only baselines, and G2 consistently outperformed G1. The magnitude of improvement remained source dependent, with SAN showing the largest relative gains and CIN achieving the strongest absolute performance. The stronger performance observed for CIN is associated with its greater demands on memory retrieval and discourse organization without continuous visual support, together with the richer speech samples shown in Table 2, which is consistent with previous findings that CIN is sensitive to MCI-related lexical-semantic changes and that high-memory-load recall enhances AD-related acoustic differences [26, 27].

To assess cross-corpus robustness, the Delaware-trained preprocessing pipeline, generator, and classifiers were applied to 72 CT subject-visit recordings from 42 unique participants in the Pitt cohort. As shown in Table 5, both augmented settings produced statistically significant improvements in BACC and AUC, while G2 achieved the strongest overall performance and improved all reported metrics relative to the CT-only setting. These results provide further evidence that the learned cross-task mappings and augmentation benefits can transfer to an independent cohort collected under different demographic and recording conditions.

The source-task-level gains are further summarized in Fig. 4. The relative improvements differ across source tasks, indicating that the effectiveness of cross-task generative augmentation depends on both source-task discriminability and source–target complementarity. Under the CT source setting, adding both generated CIN and SAN features yields relative improvements of 8.3% in BACC and 12.5% in AUC, with generated SAN providing slightly larger single-target gains than generated CIN. Under the CIN source setting, adding both generated CT and SAN features yields relative improvements of 8.4% in BACC and 9.1% in AUC, while generated SAN contributes more than generated CT in the single-target settings. Under the SAN source setting, adding both generated CT and CIN features produces the largest relative improvements, increasing BACC by 16.3% and AUC by 18.6%. Among the individual generated targets, CIN provides the greater benefit for SAN.

Classifier-level trends in Fig. 5 show that the augmentation benefit varies across decision models. Logistic Regression exhibits relatively consistent gains across source settings, suggesting that generated features enhance linear discriminative structure. DNN achieves strong performance but shows greater source-dependent variation, indicating higher sensitivity to generated feature quality. Ensemble methods are comparatively stable, whereas SVM, MLP, and AdaBoost show

Table 4
Overall classification results of cross-task generative feature augmentation under different source settings.

Tasks	BACC (%)	AUC (%)	SEN (%)	SPE (%)	F1 (%)
CT	60.71	62.20	**64.29**	57.14	55.41
CT + G1	63.40 ± 2.12*	67.30 ± 1.94*	59.04 ± 2.17	67.75 ± 2.08	55.96 ± 1.69
CT + G2	**65.72 ± 1.54***	**69.95 ± 1.39***	61.13 ± 1.61	**70.30 ± 1.47**	**58.91 ± 1.33**
CIN	66.84	69.39	55.10	78.57	58.20
CIN + G1	70.35 ± 2.17*	73.14 ± 1.87*	58.30 ± 2.23	82.39 ± 2.11	62.58 ± 1.76
CIN + G2	**72.46 ± 1.48***	**75.71 ± 1.36***	**60.31 ± 1.54**	**84.60 ± 1.43**	**64.92 ± 1.29**
SAN	59.69	62.94	62.24	57.14	54.20
SAN + G1	68.14 ± 2.23*	73.52 ± 1.84*	67.54 ± 2.27	68.73 ± 2.19	62.33 ± 1.93
SAN + G2	**69.40 ± 1.53***	**74.67 ± 1.32***	**68.51 ± 1.58**	**70.29 ± 1.49**	**63.76 ± 1.37**

* Indicates $p < 0.05$.

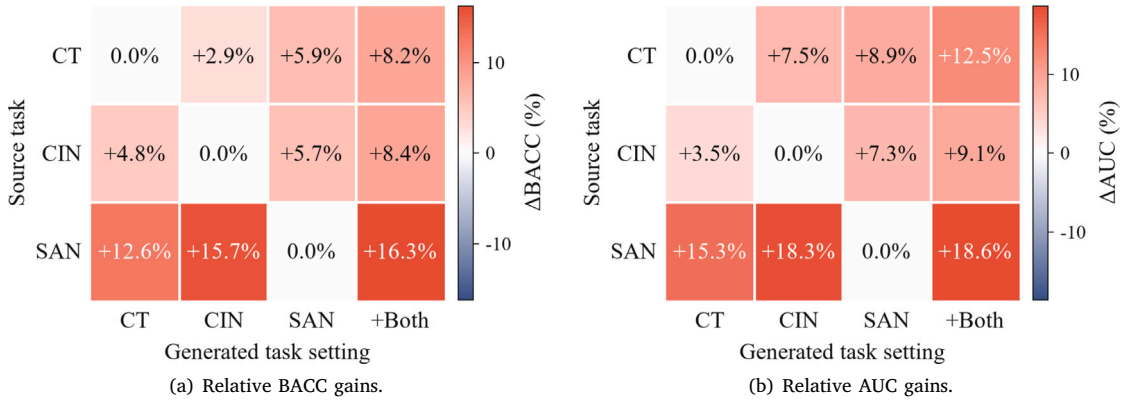
Source-only results are averages across seven classifiers. G1 and G2 denote one and two generated target-task feature sets, respectively, and are reported as mean ± sample standard deviation over ten runs. One-sided one-sample Wilcoxon signed-rank tests are applied to BACC and AUC to test whether G1 and G2 outperform the corresponding Source-only baseline. Bold and underlined values indicate the best and second-best results, respectively. SEN and SPE denote sensitivity and specificity.

Table 5
External classification results on the Pitt cohort.

Tasks	BACC (%)	AUC (%)	SEN (%)	SPE (%)	F1 (%)
CT	58.33	59.69	61.90	54.76	59.22
CT + G1	62.80 ± 3.21*	67.02 ± 3.92*	50.46 ± 3.27	**75.14 ± 3.15**	55.42 ± 3.93
CT + G2	**67.86 ± 3.53***	**73.28 ± 1.98***	**66.71 ± 3.48**	69.01 ± 3.58	**65.52 ± 3.75**

* Indicates $p < 0.05$.

Source-only results are averages across seven classifiers. G1 and G2 denote one and two generated target-task feature sets, respectively, and are reported as mean ± sample standard deviation over ten runs. One-sided one-sample Wilcoxon signed-rank tests are applied to BACC and AUC to test whether G1 and G2 outperform the corresponding Source-only baseline. Bold and underlined values indicate the best and second-best results, respectively. SEN and SPE denote sensitivity and specificity.



(a) Relative BACC gains. (b) Relative AUC gains.

Fig. 4. Relative performance gains under different source-task and generated target-task settings.

more variable trends. Overall, the effectiveness of generated target-task features depends on both task complementarity and classifier characteristics.

4.3. Visualization of generated features

Figs. 6–8 show that the generated representations generally overlap with the corresponding real target-task features while exhibiting source-to-target-dependent distributional differences. Under the CT source setting, the generated CIN and SAN features remain broadly compatible with the real target-task distributions, with local regions showing clearer separation between MCI and HC. The corresponding squared Maximum Mean Discrepancy (MMD^2) values are 0.0495 for CT→CIN and 0.0407 for CT→SAN, consistent with the moderate augmentation gains obtained from CT. Under the SAN source setting, SAN→CT shows stronger overlap between real and generated features, whereas the generated CIN features are more frequently distributed near the boundary of the real feature region. The MMD^2 values are 0.0368 for SAN→CT and 0.0396 for SAN→CIN. The generated representations also show more apparent MCI–HC separation in several regions, consistent with the relatively large gains obtained from SAN. Under the CIN source setting, the real and generated target-task features are broadly interleaved while retaining distinguishable MCI and HC structures. The MMD^2 values are 0.0405 for CIN→CT and 0.0450 for CIN→SAN. The improvement in class separation is less pronounced because CIN already provides the strongest source-only discrimination. Overall, the UMAP and MMD results indicate that the generated features remain compatible with the real target-task distributions while preserving or improving MCI–HC separability in a source- and target-dependent manner.

4.4. Ablation study

Seven ablation variants were considered. The first variant replaced the VGG-based acoustic encoder with the self-supervised wav2vec model, while the remaining six variants individually removed the

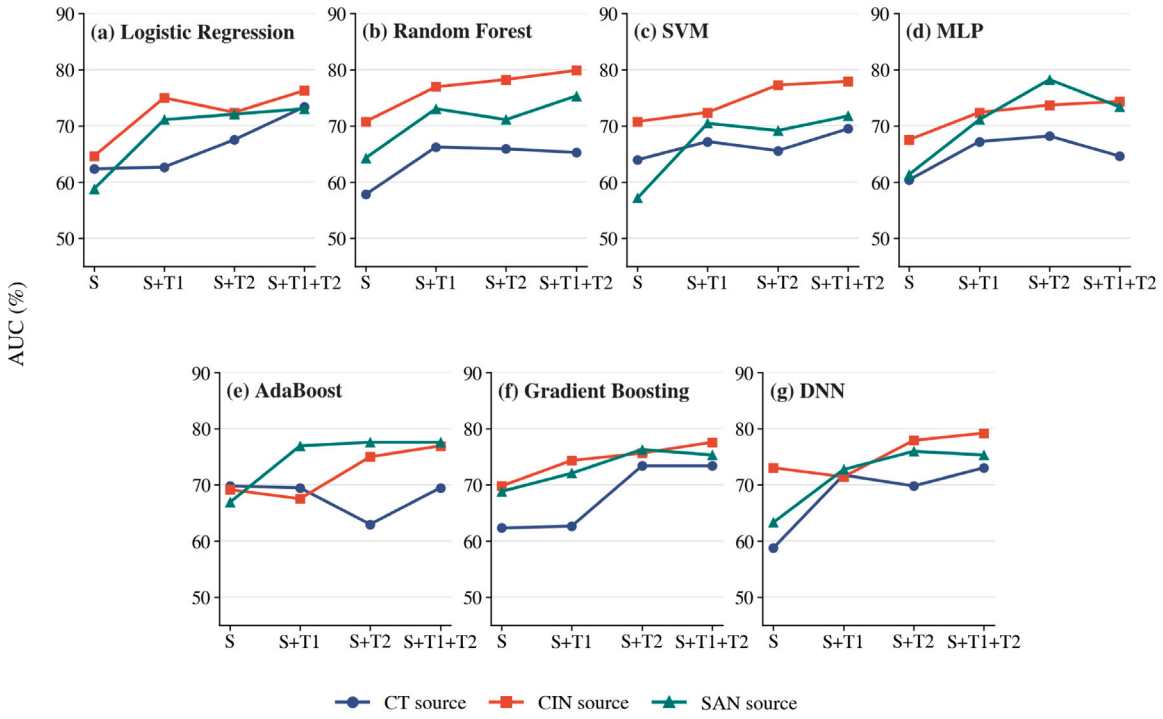


Fig. 5. Classifier-wise AUC trends across task feature settings. Each panel shows one classifier, with curves for CT, CIN, and SAN source settings. S denotes the source task, while T_1 and T_2 denote the first and second generated target tasks.

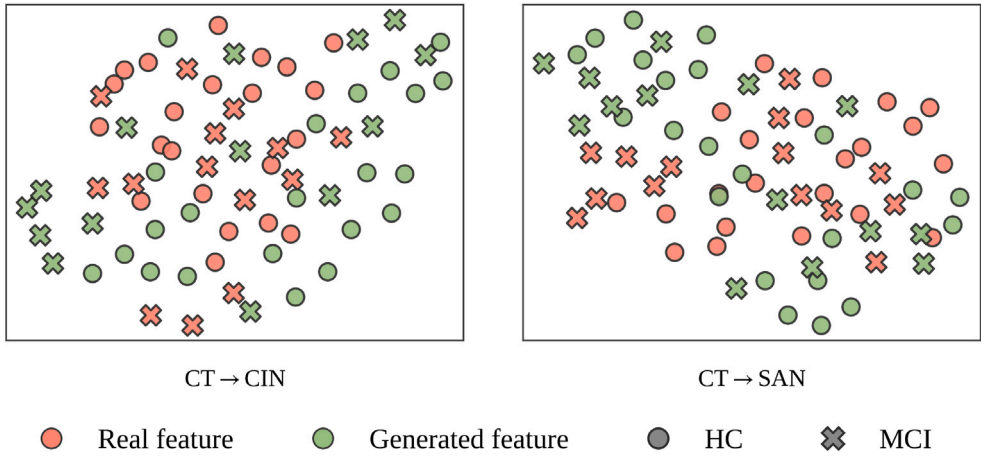


Fig. 6. UMAP visualization of real and generated target task features under the CT source setting.

latent translator, target-task conditioning, task and modality disentanglement, disease supervision, shared latent alignment, and the feature discriminator. All variants used the same generation, classification, and evaluation protocol as the main experiments. Table 6 summarizes the ablation results, with the corresponding source-only setting as the reference.

Wav2Vec 2.0. Replacing VGG with Wav2Vec 2.0 retained cross-task augmentation gains, but produced less consistent improvements, indicating weaker compatibility with the current generation setting. **Latent translator.** Removing the translator weakened augmentation for CT and reduced the gains for SAN, supporting explicit source-to-target translation. **Target-task conditioning.** Removing target-task conditioning produced inconsistent gains and reduced the improvement of CIN + G2 to a marginal level. **Task and modality disentanglement.** Removing the disentanglement constraints reduced augmentation stability, particularly for SAN. **Disease supervision.** Removing disease supervision degraded CT and reduced the gains for SAN, while CIN remained comparatively robust. **Shared latent alignment.** Removing shared alignment reduced the consistency of augmentation gains across source tasks. **Feature discriminator.** Removing the discriminator weakened CT and CIN + G2, while SAN retained moderate augmentation gains.

5. Discussion

5.1. Comparison with existing studies

Table 7 compares the proposed method with recent Delaware-related studies. Stark et al. used CIN core-lexicon features with LDA

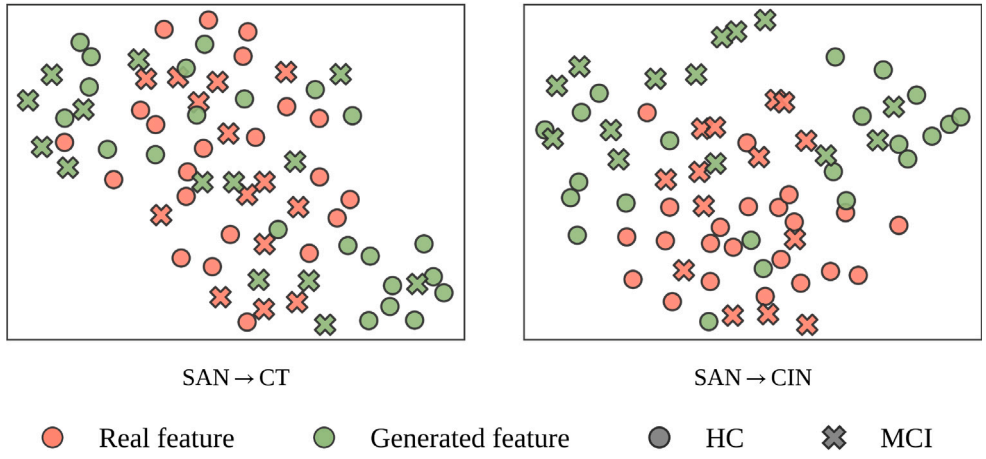


Fig. 7. UMAP visualization of real and generated target task features under the SAN source setting.

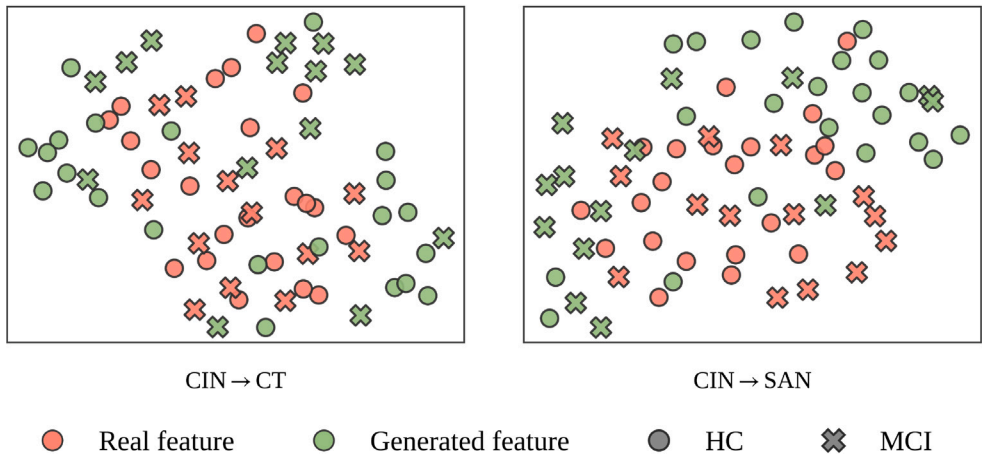


Fig. 8. UMAP visualization of real and generated target task features under the CIN source setting.

Table 6
Ablation results for balanced accuracy (%).

Tasks	w2v	Trans.	Cond.	Disent.	Disease	Align.	Disc.
CT	57.05 ± 5.61	60.71 ± 5.98	60.71 ± 5.98	60.71 ± 5.98	60.71 ± 5.98	60.71 ± 5.98	60.71 ± 5.98
CT + G1	59.88 ± 4.07	56.35 ± 3.97	57.42 ± 5.37	59.81 ± 5.06	53.39 ± 5.68	57.95 ± 4.75	53.46 ± 5.14
CT + G2	62.01 ± 3.65	56.03 ± 4.28	60.81 ± 3.78	63.31 ± 7.82	54.04 ± 7.67	62.62 ± 5.50	56.54 ± 5.99
CIN	63.68 ± 4.75	66.84 ± 3.95	66.84 ± 3.95	66.84 ± 3.95	66.84 ± 3.95	66.84 ± 3.95	66.84 ± 3.95
CIN + G1	66.49 ± 3.95	68.37 ± 3.03	68.81 ± 3.59	68.18 ± 6.38	72.01 ± 2.50	68.60 ± 3.64	67.18 ± 2.13
CIN + G2	68.83 ± 3.14	69.11 ± 2.86	67.53 ± 5.31	68.78 ± 3.74	75.93 ± 2.76	71.34 ± 5.27	65.77 ± 6.48
SAN	56.22 ± 1.89	59.69 ± 2.00	59.69 ± 2.00	59.69 ± 2.00	59.69 ± 2.00	59.69 ± 2.00	59.69 ± 2.00
SAN + G1	64.26 ± 5.84	64.91 ± 6.13	63.29 ± 6.37	60.78 ± 5.18	63.13 ± 4.41	64.61 ± 4.71	64.70 ± 5.99
SAN + G2	66.09 ± 5.11	64.80 ± 5.00	63.45 ± 3.72	57.98 ± 5.32	63.03 ± 5.59	64.80 ± 4.71	62.52 ± 5.20

w2v denotes the variant in which the original VGG-based acoustic encoder is replaced with the self-supervised wav2vec model. Trans., Cond., Disent., Disease, Align., and Disc. denote the variants without the latent translator, target-task conditioning, task and modality disentanglement constraints, disease-classification loss, shared-alignment loss, and feature discriminator, respectively. Values are reported as mean ± sample standard deviation across the seven classifiers.

and reported 67.44% ACC [26]. Lai et al. combined textual and acoustic CT features using LDA and SVM, achieving 57.94% ACC and 56.91% F1-score [36]. Ahangaran et al. evaluated pitch-shifted audio across multiple narrative tasks and obtained 63.70% ACC [37]. Zolnour et al. combined CT, CIN, and SAN through late fusion and synthetic augmentation, reporting 72.82% F1-score [16]. Chen et al. compared four text-based tasks and found that only CIN significantly distinguished MCI from HC, with 65.60% ACC [38]. Themistocleous et al. used linguistic biomarkers from five tasks with SVM and DynamicSMOTE, obtaining 60.00% F1-score [39]. Taherinezhad et al. evaluated prompting and fine-tuning on five-task transcripts, with fine-tuned GPT-4o achieving 82.00% F1-score [40]. The proposed method evaluates three task-incomplete settings using one observed task and two generated target-task feature sets, with CIN + G2 achieving 72.46 ± 1.48% BACC and 75.71 ± 1.36% AUC. Differences in tasks, inputs, modalities, participants, data partitions, classifiers, and metrics preclude a controlled head-to-head comparison; therefore, the table is presented only for contextual positioning.

Table 7
Comparison with existing Delaware-related studies and the proposed method for MCI detection.

Study	Year	Tasks	Modality	ACC (%)	F1 (%)	AUC (%)
Stark et al. [26]	2025	CIN	Text	67.44	–	–
Lai et al. [36]	2025	CT	Text+Audio	57.94	56.91	–
Ahangaran et al. [37]	2025	PD+SN+PROC+PN	Audio	63.70	–	–
Zolnour et al. [16]	2025	CT+CIN+SAN	Text	–	<u>72.82</u>	69.57
Chen et al. [38]	2025	CAT+RW+CIN+SAN	Text	65.60	–	–
Themistocleous et al. [39]	2026	CT+CAT+RW+CIN+SAN	Text	–	60.00	–
Taherinezhad et al. [40]	2026	CT+CAT+RW+CIN+SAN	Text	–	**82.00**	–
Proposed method	2026	CT+G2	Text+Audio	65.72 ± 1.54	58.91 ± 1.33	69.95 ± 1.39
Proposed method	2026	CIN+G2	Text+Audio	**72.46 ± 1.48**	64.92 ± 1.29	**75.71 ± 1.36**
Proposed method	2026	SAN+G2	Text+Audio	<u>69.40 ± 1.53</u>	63.76 ± 1.37	<u>74.67 ± 1.32</u>

CT: Cookie Theft picture description; CAT: Cat Rescue picture description; RW: Rockwell picture description; CIN: Cinderella story recall; SAN: Sandwich procedural narrative; PD: picture description; SN: story narrative; PROC: procedural narrative; PN: personal narrative. G2, two generated target-task feature set. "–" denotes not reported. Bold and underlined values indicate the best and second-best results.

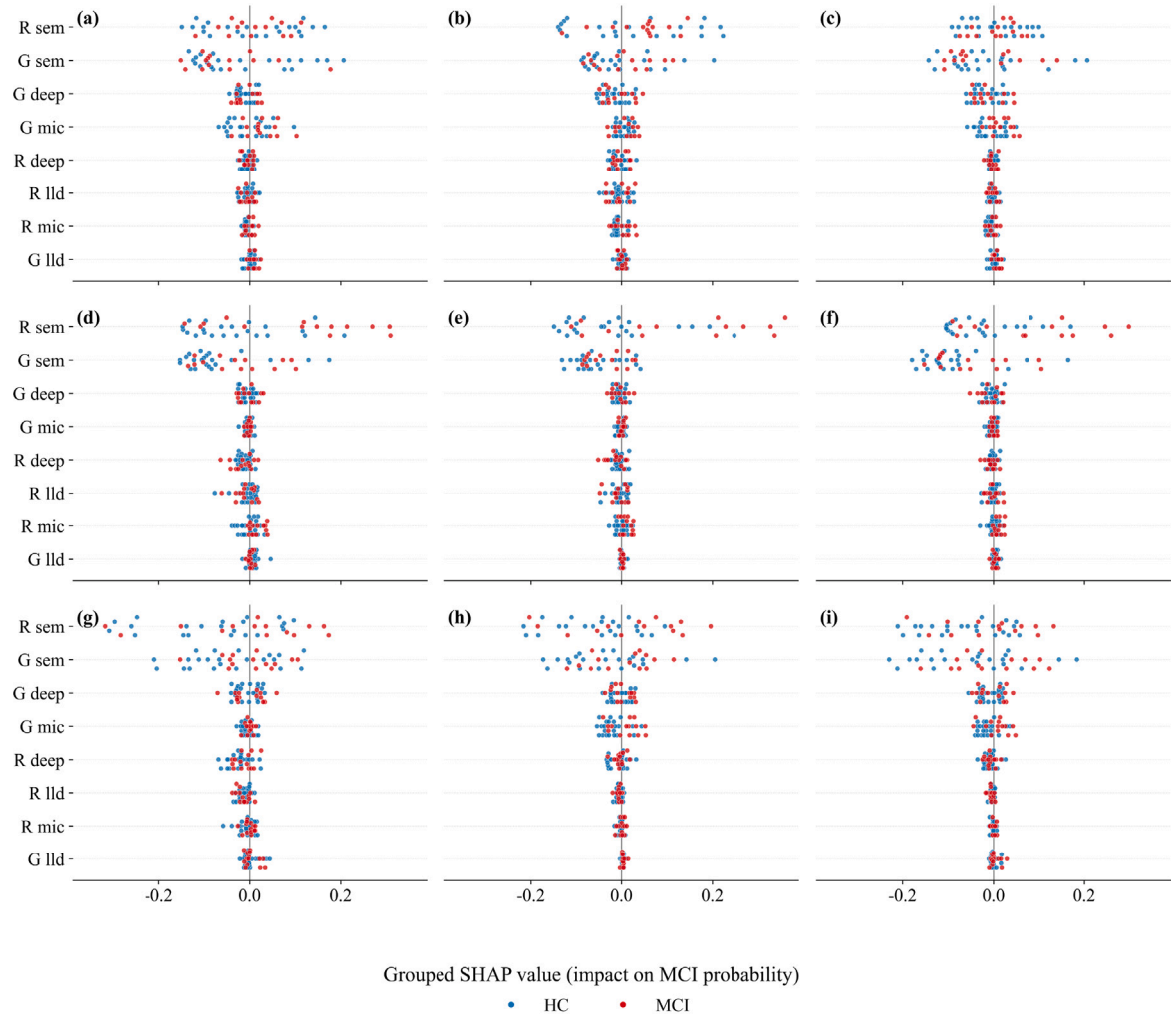


Fig. 9. SHAP analysis of real and generated modalities. Panels (a–c): CT+G(CIN), CT+G(SAN), and CT+G(CIN, SAN); (d–f): CIN+G(CT), CIN+G(SAN), and CIN+G(CT, SAN); (g–i): SAN+G(CT), SAN+G(CIN), and SAN+G(CT, CIN), where G(·) denotes generated task features. R and G denote real and generated modalities.

5.2. Modality importance analysis

Fig. 9 presents the grouped SHAP analysis of real and generated modalities, averaged across the seven classifiers. Generated semantic, deep acoustic, and micro-level linguistic features show clear contributions across task combinations, whereas generated low-level descriptors are generally concentrated near zero. These contributions indicate that the generated modalities provide complementary information underlying the observed classification gains.

5.3. Computational complexity

Let B denote the batch size, K the number of generated target tasks, and C_{enc}, C_{tr}, and C_{dec} the computational costs of multimodal encoding, latent translation, and feature decoding, respectively. The source representation is encoded once, whereas translation and decoding are repeated for each target task. The generator complexity is $\mathcal{O}\big(B[C_{\text{enc}} + K(C_{\text{tr}} + C_{\text{dec}})]\big)$. The additional cost therefore increases linearly with K and the feature-sequence lengths. On the Delaware

Corpus, one generator training run required approximately 1.3 min per source setting, while fitting single-, two-, and three-task classifiers required under 1, 1–2, and 2–3 min, respectively. At deployment, single-task classification averaged 0.12 ms per participant; generating one or two missing tasks required 2.76 or 3.69 ms, followed by 0.16 or 0.53 ms for two- or three-task classification. The resulting total inference times were 2.92 and 4.22 ms, respectively.

5.4. Limitations and future work

First, cross-dataset generalization was evaluated only on the Pitt CT data, which match the Delaware data in language and task. Extension to other English datasets requires paired target-domain data and independent participant-level validation to account for demographic, diagnostic, recording, transcription, and feature-extraction differences. Second, the translator was trained only on CT, CIN, and SAN. Extension to other connected-speech tasks requires paired recordings to learn task-specific embeddings and cross-task mappings while accommodating differences in elicitation, cognitive demand, response structure, and duration. Third, the framework was evaluated only on English speech. Cross-language extension requires language-appropriate feature extraction, paired target-language data, and model adaptation to account for phonological, lexical, syntactic, and cultural differences. Fourth, the generated features cannot be directly interpreted as speech or text, requiring further association with clinically meaningful acoustic and linguistic indicators. Fifth, the current binary MCI setting requires multiclass labels, longitudinal observations, and participant-wise temporal evaluation to support differential diagnosis and progression prediction. Finally, extension to Parkinson's disease (PD), amyotrophic lateral sclerosis, and Huntington's disease requires disease-relevant speech tasks, disorder-specific paired-task data, adaptation of the feature encoders and translator, and independent or longitudinal validation, together with privacy, security, and ethical governance considerations [41].

6. Conclusion

We proposed a cross-task generative feature augmentation framework for speech-based MCI detection under incomplete-task conditions. The framework generates classification-relevant multimodal representations for missing target tasks and combines them with observed source-task features for downstream classification. Experiments on the DementiaBank Delaware Corpus showed improvements in average BACC and AUC across CT, CIN, and SAN, although the gains varied with source-task informativeness, target-task complementarity, and classifier robustness. External testing on the independent Pitt cohort further showed that the Delaware-trained framework retained its augmentation benefit. These results support cross-task feature generation as a feasible approach to exploiting complementary cognitive-linguistic information in low-resource settings. In practice, augmentation is more suitable when the observed task contains reliable disease cues and paired training data support complementary mappings, while its effectiveness may decline for weak source tasks or substantial domain shifts.

CRediT authorship contribution statement

Hao Li: Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Bo Wang:** Supervision, Project administration. **Qiang He:** Supervision, Project administration. **Jun Mou:** Supervision, Project administration. **Junxin Chen:** Supervision, Project administration.

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 research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

Data will be made available on request.

References

[1] J. Heitz, I.M. Engler, N. Langer, Towards a speech-based digital biomarker for cognitive impairment: Speech as a proxy for cognitive assessment, Npj Digit. Med. 9 (2026) 179, http://dx.doi.org/10.1038/s41746-026-02360-8.

[2] S. Pang, Y. Chen, X. Shi, R. Wang, M. Dai, X. Zhu, B. Song, K. Li, Interpretable 2.5D network by hierarchical attention and consistency learning for 3D MRI classification, Pattern Recognit. 164 (2025) 111539, http://dx.doi.org/10.1016/j.patcog.2025.111539.

[3] P.M. Butler, et al., Smartwatch- and smartphone-based remote assessment of brain health and detection of mild cognitive impairment, Nature Med. 31 (2025) 829–839, http://dx.doi.org/10.1038/s41591-024-03475-9.

[4] L. Zhang, J. Wu, L. Wang, L. Wang, D.C. Steffens, S. Qiu, G.G. Potter, M. Liu, Brain anatomy prior modeling to forecast clinical progression of cognitive impairment with structural MRI, Pattern Recognit. 165 (2025) 111603, http://dx.doi.org/10.1016/j.patcog.2025.111603.

[5] B. Lei, Y. Liang, J. Xie, Y. Wu, E. Liang, Y. Liu, P. Yang, T. Wang, C. Liu, J. Du, X. Xiao, S. Wang, Hybrid federated learning with brain-region attention network for multi-center Alzheimer's disease detection, Pattern Recognit. 153 (2024) 110423, http://dx.doi.org/10.1016/j.patcog.2024.110423.

[6] Q. Li, Y. Wang, Y. Zhang, Z. Zuo, J. Chen, W. Wang, Fuzzy-ViT: A deep neuro-fuzzy system for cross-domain transfer learning from large-scale general data to medical image, IEEE Trans. Fuzzy Syst. 33 (1) (2025) 231–241, http://dx.doi.org/10.1109/TFUZZ.2024.3400861.

[7] B. Lei, Y. Zhu, S. Yu, H. Hu, Y. Xu, G. Yue, T. Wang, C. Zhao, S. Chen, P. Yang, X. Song, X. Xiao, S. Wang, Multi-scale enhanced graph convolutional network for mild cognitive impairment detection, Pattern Recognit. 134 (2023) 109106, http://dx.doi.org/10.1016/j.patcog.2022.109106.

[8] T. Sohail, A. Aiman, E. Hashmi, A.S. Imran, S.M. Daudpota, S.Y. Yayilgan, Hate speech detection in code-mixed datasets using pretrained embeddings and transformers, in: 2024 International Conference on Frontiers of Information Technology, FIT, 2024, pp. 1–6, http://dx.doi.org/10.1109/FIT63703.2024.10838452.

[9] J. Chen, S. Sun, L.-b. Zhang, B. Yang, W. Wang, Compressed sensing framework for heart sound acquisition in internet of medical things, IEEE Trans. Ind. Inform. 18 (3) (2022) 2000–2009, http://dx.doi.org/10.1109/TII.2021.3088465.

[10] J. Chen, Z. Guo, X. Xu, L.-b. Zhang, Y. Teng, Y. Chen, M. Woźniak, W. Wang, A robust deep learning framework based on spectrograms for heart sound classification, IEEE/ACM Trans. Comput. Biol. Bioinform. 21 (4) (2024) 936–947, http://dx.doi.org/10.1109/TCBB.2023.3247433.

[11] R. He, K. Chapin, J. Al-Tamimi, N. Bel, M. Marquié, M. Rosende-Roca, V. Pytel, J.P. Tartari, M. Alegret, A. Sanabria, A. Ruiz, M. Boada, S. Valero, W. Hinzen, Automated classification of cognitive decline and probable Alzheimer's dementia across multiple speech and language domains, Am. J. Speech-Language Pathol. 32 (5) (2023) 2075–2086, http://dx.doi.org/10.1044/2023_AJSLP-22-00403.

[12] A.M. Lanzi, A.K. Saylor, D. Fromm, H. Liu, B. MacWhinney, M.L. Cohen, DementiaBank: Theoretical rationale, protocol, and illustrative analyses, Am. J. Speech-Language Pathol. 32 (2) (2023) 426–438, http://dx.doi.org/10.1044/2022_AJSLP-22-00281.

[13] J. Chen, Z. Ye, R. Zhang, H. Li, B. Fang, L. bo Zhang, W. Wang, Medical image translation with deep learning: Advances, datasets and perspectives, Med. Image Anal. 103 (2025) 103605, http://dx.doi.org/10.1016/j.media.2025.103605.

[14] J. Chen, W. Wang, B. Fang, Y. Liu, K. Yu, V.C.M. Leung, X. Hu, Digital twin empowered wireless healthcare monitoring for smart home, IEEE J. Sel. Areas Commun. 41 (11) (2023) 3662–3676, http://dx.doi.org/10.1109/JSAC.2023.3310097.

[15] Y. Han, J.C.K. Lam, V.O.K. Li, L.Y.L. Cheung, A large language model based data generation framework to improve mild cognitive impairment detection sensitivity, Data Policy 7 (2025) e33, http://dx.doi.org/10.1017/dap.2025.8.

[16] A. Zolnour, H. Azadmaleki, Y. Haghbin, F. Taherinezhad, M.J.M. Nezhad, S. Rashidi, M. Khani, A. Taleban, S.M. Sani, M. Dadkhah, J.M. Noble, S. Bakken, Y. Yaghoobzadeh, A.-H. Vahabie, M. Rouhizadeh, M. Zolnoori, LLMCARE: Early detection of cognitive impairment via transformer models enhanced by LLM-generated synthetic data, Front. Artif. Intell. 8 (2025) 1669896, http://dx.doi.org/10.3389/frai.2025.1669896.

[17] Z. Jafari, M.K. Andrew, K.J. Rockwood, Diagnostic utility of speech-based biomarkers in mild cognitive impairment: A systematic review and meta-analysis, Age Ageing 54 (10) (2025) afaf316, http://dx.doi.org/10.1093/ageing/afaf316.

[18] F.J. Ferrante, J. Migeot, A. Birba, L. Amoruso, G. Pérez, E. Hesse, E. Tagliazucchi, C. Estienne, C. Serrano, A. Slachevsky, D. Matallana, P. Reyes, A. Ibáñez, S. Fittipaldi, C.G. Campo, A.M. García, Multivariate word properties in fluency tasks reveal markers of Alzheimer's dementia, Alzheimer's Dement. 20 (2) (2024) 925–940, http://dx.doi.org/10.1002/alz.13472.

[19] S. Bayat, M. Sanati, M. Mohammad-Panahi, A. Khodadadi, M. Ghasimi, S. Rezaee, S. Besharat, Z. Mahboubi-Fooladi, M. Almasi-Dooghaee, M. Sanei-Taheri, B.C. Dickerson, N. Rezaii, Language abnormalities in Alzheimer's disease indicate reduced informativeness, Ann. Clin. Transl. Neurol. 11 (11) (2024) 2946–2957, http://dx.doi.org/10.1002/acn3.52205.

[20] L. Ilias, D. Askounis, J. Psarras, Detecting dementia from speech and transcripts using transformers, Comput. Speech Lang. 79 (2023) 101485, http://dx.doi.org/10.1016/j.csl.2023.101485.

[21] D. Ortiz-Perez, M. Benavent-Lledo, J. Rodriguez-Juan, J. Garcia-Rodriguez, D. Tomás, CogniAlign: Word-level multimodal speech alignment with gated cross-attention for Alzheimer's detection, Knowl.-Based Syst. 329 (2025) 114264, http://dx.doi.org/10.1016/j.knosys.2025.114264.

[22] S. Hassan, E. Hashmi, G. Mujtaba, M.M. Yamin, M.H. Mughal, M. Ullah, SELFIE-Net: Self-supervised learning and feature-intensified emotion network for facial emotion recognition, IEEE Trans. Cogn. Dev. Syst. (2026) 1–16, http://dx.doi.org/10.1109/TCDS.2026.3681803, Early Access.

[23] X. Ke, M.W. Mak, H.M. Meng, Automatic selection of spoken language biomarkers for dementia detection, Neural Netw. 169 (2024) 191–204, http://dx.doi.org/10.1016/j.neunet.2023.10.018.

[24] E. Hashmi, S.Y. Yayilgan, M.M. Yamin, M. Ullah, Enhancing misogyny detection in bilingual texts using explainable AI and multilingual fine-tuned transformers, Complex Intell. Syst. 11 (1) (2025) 39, http://dx.doi.org/10.1007/s40747-024-01655-1.

[25] G. Hickok, D. Poeppel, The cortical organization of speech processing, Nature Rev. Neurosci. 8 (5) (2007) 393–402, http://dx.doi.org/10.1038/nrn2113.

[26] B.C. Stark, S.G. Dalton, A.M. Lanzi, Access to context-specific lexical-semantic information during discourse tasks differentiates speakers with latent aphasia, mild cognitive impairment, and cognitively healthy adults, Front. Hum. Neurosci. 18 (2025) 1500735, http://dx.doi.org/10.3389/fnhum.2024.1500735.

[27] M. Bae, M.-G. Seo, H. Ko, H. Ham, K.Y. Kim, J.-Y. Lee, The efficacy of memory load on speech-based detection of Alzheimer's disease, Front. Aging Neurosci. 15 (2023) 1186786, http://dx.doi.org/10.3389/fnagi.2023.1186786.

[28] M. Mirgbagheri, H. Saghir, T. Chau, A novel approach to mimic linguistic features of transcripts of atypical speech, IEEE Trans. Audio Speech Lang. Process. 33 (2025) 3690–3702, http://dx.doi.org/10.1109/TASLPRO.2025.3603921.

[29] W. Wang, X. Yu, B. Fang, Y. Zhao, Y. Chen, W. Wei, J. Chen, Cross-modality LGE-CMR segmentation using image-to-image translation based data augmentation, IEEE/ACM Trans. Comput. Biol. Bioinform. 20 (4) (2023) 2367–2375, http://dx.doi.org/10.1109/TCBB.2022.3140306.

[30] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019, arXiv preprint arXiv:1907.11692, arXiv preprint arXiv:1907.11692, URL https://arxiv.org/abs/1907.11692.

[31] F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, E. André, C. Busso, L.Y. Devillers, J. Epps, P. Laukka, S.S. Narayanan, K.P. Truong, The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing, IEEE Trans. Affect. Comput. 7 (2) (2016) 190–202, http://dx.doi.org/10.1109/TAFFC.2015.2457417.

[32] S. Hershey, S. Chaudhuri, D.P.W. Ellis, J.F. Gemmeke, A. Jansen, R.C. Moore, M. Plakal, D. Platt, R.A. Saurous, B. Seybold, M. Slaney, R.J. Weiss, K. Wilson, CNN architectures for large-scale audio classification, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2017, pp. 131–135, http://dx.doi.org/10.1109/ICASSP.2017.7952132.

[33] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, in: Proceedings of the 2nd International Conference on Learning Representations, ICLR, 2014, arXiv:1312.6114, URL https://arxiv.org/abs/1312.6114, Unpaginated conference paper.

[34] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, J. Mach. Learn. Res. 17 (59) (2016) 1–35, URL https://jmlr.org/papers/v17/15-239.html.

[35] J.T. Becker, F. Boiler, O.L. Lopez, J. Saxton, K.L. McGonigle, The natural history of alzheimer's disease: Description of study cohort and accuracy of diagnosis, Arch. Neurol. 51 (6) (1994) 585–594, http://dx.doi.org/10.1001/archneur.1994.00540180063015.

[36] Y.-L. Lai, E.-Y. Chang, Y.-W. Liu, J.L. Hsu, H.-C. Hsu, Mild cognitive impairment detection via linear discriminant analysis of picture description speech features: A cross corpus comparison, in: 2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC, 2025, pp. 1074–1079, http://dx.doi.org/10.1109/APSIPAASC65261.2025.11249167.

[37] M. Ahangaran, N. Dawalatabad, C. Karjadi, J. Glass, R. Au, V.B. Kolachalama, Obfuscation via pitch-shifting for balancing privacy and diagnostic utility in voice-based cognitive assessment, Alzheimer's Dement. 21 (3) (2025) e70032, http://dx.doi.org/10.1002/alz.70032.

[38] Y. Chen, R.J. Hartsuiker, A. Pistono, A comparison of different connected-speech tasks for detecting mild cognitive impairment using multivariate pattern analysis, Aphasiology 39 (4) (2025) 476–499, http://dx.doi.org/10.1080/02687038.2024.2358556.

[39] C. Themistocleous, B.C. Stark, Language biomarker screening using AI: A transdiagnostic approach to the brain, Sci. Rep. 16 (1) (2026) 4207, http://dx.doi.org/10.1038/s41598-025-34257-z.

[40] F. Taherinezhad, M.J. Momeni Nezhad, S. Karimi, S. Rashidi, A. Zolnour, M. Dadkhah, Y. Haghbin, H. Azadmaleki, M. Zolnoori, Large language model adaptation strategies in speech-based cognitive screening: Systematic evaluation, JMIR AI 5 (2026) e82608, http://dx.doi.org/10.2196/82608.

[41] E. Hashmi, M.M. Yamin, S.Y. Yayilgan, Securing tomorrow: a comprehensive survey on the synergy of Artificial Intelligence and information security, AI Ethics 5 (3) (2025) 1911–1929, http://dx.doi.org/10.1007/s43681-024-00529-z.

Hao Li is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering, Dalian University of Technology, China. Her research interests include speech and language processing, multimodal learning, generative artificial intelligence, and medical artificial intelligence.

Bo Wang (Member, IEEE) received the BS degree in electronic and information engineering, and the Ph.D. degree in signal and information processing from the Dalian University of Technology, Dalian, China, in 2003 and 2010, respectively. From 2010 to 2012, he was a postdoctoral research associate with the Faculty of Management and Economics, Dalian University of Technology. From 2017 to 2019, he was with the State of University of New York at Buffalo, as a visiting scholar. He is currently a professor with the School of Information and Communication Engineering, Dalian University of Technology. His research interests include artificial intelligence security and multimedia security.

Qiang He received the M.S. and Ph.D. degrees from Northeastern University, Shenyang, China, in 2016 and 2020, respectively. He is currently an Associate Professor with Northeastern University. He has published more than 90 journal articles and conference papers. His research interests include edge computing, machine learning, social network analytics, data mining, health care, and infectious diseases informatics.

Jun Mou (Senior Member, IEEE) received the BS, MS, and Ph.D. degrees in physics and electronics from Central South University, Changsha, China, in 2003, 2006, and 2017, respectively. He is currently a professor with the School of Information Science and Engineering, Dalian Polytechnic University, Dalian, China. His research interests include nonlinear system control, and secure communication.

Junxin Chen is currently a Professor at the School of Software, Dalian University of Technology, China. His research interests include AI Agent, wearable devices, artificial intelligence. He has authored/co-authored over 200+ scientific papers in international peer-reviewed journals and conferences.