npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02026-x

SpeechCARE: dynamic multimodal modeling for cognitive screening in diverse linguistic and speech task contexts

Check for updates

Hossein Azadmaleki[1], Yasaman Haghbin[1], Sina Rashidi[1], Mohammad Javad Momeni Nezhad[1], Ali Zolnour[1] & Maryam Zolnoori[1,2]

SpeechCARE is a multimodal transformer pipeline designed to detect cognitive impairment from brief speech recordings through multiclass classification of Alzheimer's Disease and Related Dementias (ADRD), Mild Cognitive Impairment (MCI), and healthy controls. It integrates an advanced preprocessing pipeline that includes LLM-based audio anomaly detection, speech-task identification, noise reduction, and transcription. Its core architecture fuses mHuBERT (acoustic) and mGTE (linguistic) embeddings with demographic information using a novel Adaptive Gating Fusion mechanism. Additionally, a specialized encoding component further processes mHuBERT outputs to capture global temporal patterns across segmented audio, addressing key limitations of speech transformers in modeling long-range dependencies in extended recordings. Trained on the National Institute on Aging's PREPARE challenge dataset (1655 participants in English, Spanish, and Mandarin), SpeechCARE achieved an average F1-score of 72.11% on the held-out test set ($n = 412$), earning a special recognition award from NIA. Threshold optimization improved MCI recall. While fairness analysis showed moderate disparities (particularly for Spanish speakers), the model demonstrated strong multilingual generalizability. SpeechCARE complements blood-based biomarkers by capturing functional speech deficits, supporting early, scalable detection.

Alzheimer's disease and related dementias (ADRD) affect one in five adults over 60, severely impacting quality of life, healthcare use, and costs[1–3]. Despite nationwide efforts, over half of individuals with cognitive decline, including mild cognitive impairment (MCI) and ADRD, remain undiagnosed due to symptoms being unrecognized by patients, limited biomarker availability (e.g., MRI), and clinicians' time constraints[4–6]. With 11 to 16 million cognitive impairment cases projected by 2050, developing a robust screening tool that can expand to diverse populations for early identification has been recognized as a research priority by the National Institute on Aging (NIA)[7].

Acoustic and linguistic cues in speech can indicate cognitive decline, reflecting deficits in phonetic motor planning, linguistic organization, executive functioning, and semantic memory. For example, compromised vocal control may lead to difficulties with prosody, articulation, and speech rhythm, while memory lapses can disrupt syntax, semantics, and conversational fluency[8,9].

Existing speech-processing pipelines for detecting cognitive impairment primarily rely on hand-crafted acoustic features (e.g., eGeMAPS[10]) and general-purpose deep learning models (e.g., YAMNet[11], VGGish[12]). These approaches often yield suboptimal performance and limited generalizability, missing subtle nuances in patient speech or requiring extensive labeled datasets to achieve reliable performance. Particularly, in cognitive impairment research, assembling large training datasets is hindered by privacy concerns, logistical barriers, and high associated costs. Speech and linguistic transformer[13] models have revolutionized speech processing by leveraging attention mechanisms to capture long-range dependencies. Pretrained on large corpora, these models can be fine-tuned on relatively small datasets (sometimes with only several hundred samples) and learn to encode acoustic subtleties (e.g., prosody, rhythm, verbal fluency) and linguistic nuances (e.g., lexical richness, syntactic errors) to produce context-aware representations[14,15]. Despite their potential, transformer models have not

[1]Columbia University Irving Medical Center, New York, NY, USA. [2]Columbia University School of Nursing, New York, NY, USA.
e-mail: mz2825@cumc.columbia.edu

been fully explored in cognitive impairment-related studies, mainly due to the difficulty of collecting even small datasets in this domain and the absence of fine-tuning pipelines capable of handling data scarcity.

One strategy to address data scarcity is consolidating small datasets from studies where participants have consented to future use. These typically include speech from standard neurological tasks (e.g., picture description, story recall, verbal fluency) across various languages. This aggregation approach underlies the NIA PREPARE challenge[16], launched in 2023 to create the PREPARE dataset, followed in 2024 by a challenge to develop multilingual screening tools for early detection. Developing such tools requires a pipeline capable of efficiently learning patterns and transferring knowledge across varied (neurological) speech tasks and languages.

In response, we developed SpeechCARE, a multimodal, multilingual speech-processing pipeline for cognitive impairment detection, which received NIA's special recognition and explainability prizes. Inspired by the Mixture of Experts paradigm[17], SpeechCARE uses a novel fusion architecture that dynamically weights transformer-derived acoustic and linguistic features, along with demographic features, to enhance detection performance and generalizability across speech tasks and languages. This study aims to (1) provide an overview of the NIA challenge dataset; (2) outline the preprocessing pipeline for integrating speech data from different sources, particularly in the absence of standardized collection guidelines; (3) describe the SpeechCARE architecture for integrating linguistic, acoustic, and demographic features; and (4) evaluate the model's performance and bias across languages and demographic subgroups.

Results

Performance analysis

Incorporating age solely as a demographic modality with the acoustic and linguistic transformers resulted in the highest predictive accuracy, achieving a micro Area Under the Curve (AUC) of 86.83 ± 0.46% and an F1-score of 72.11 ± 0.44%, with 95% confidence intervals across 10 runs. This highlights age as a particularly relevant demographic factor in cognitive impairment detection. We used this best-performing combination of modalities as the final SpeechCARE setting, and all remaining analyses are based on this setting.

Figure 1 illustrates the robust discriminative performance of SpeechCARE, showing a micro AUC of 86.83 ± 0.46% and a weighted AUC of 80.67 ± 0.65% on the test set. Precision-recall analysis further validated model performance, achieving micro and weighted precision (AP) scores of 74.73 ± 1.21% and 73.50 ± 0.66%, respectively. These findings confirm the model's balanced effectiveness in accurately distinguishing cognitive impairment, demonstrating its potential utility in clinical settings.

In comparison with the PREPARE Challenge Phase 2 results, the PREPARE Phase 2 leaderboard was dominated by solutions that leveraged large, pre-trained speech transformer encoders (e.g., Wav2Vec 2.0, HuBERT) to generate embeddings, followed by different classification layers[18]. Some teams also combined these embeddings with traditional acoustic features or machine learning algorithms like support vector machine (SVM). Multi-class log-loss (cross-entropy) across 67 teams ranged from 1.5151 (lowest) to 0.6299 (best), with the top three scores being 0.6299, 0.6343, and 0.6523. SpeechCARE achieved a log-loss of 0.6553, ranking fourth on the challenge leaderboard[19]. However, for the final Model Report submission, SpeechCARE improved to 0.6460 after further hyperparameter tuning, earning special recognition for multiple contributions: (i) a robust audio preprocessing pipeline with Large Language Model (LLM)-based noise filtering and task-type identification; (ii) the SpeechCARE architecture for adaptive multimodal fusion; and (iii) model outcome highlighting modality importance across tasks.

Threshold optimization

Given the inherent difficulty in identifying MCI cases due to their subtle clinical presentation compared to Alzheimer's disease, we conducted threshold optimization for the model output probabilities to achieve a better precision–recall balance across diagnostic classes (Control, MCI, AD). We report threshold optimization results from the model checkpoint that showed performance close to the average obtained from the 10 training runs. As demonstrated in Table 1, this optimization modestly improved the overall F1-score from 72.57% to 73.54% and achieved better alignment between precision and recall across all diagnostic classes, particularly strengthening the detection of early-stage cognitive impairment (MCI).

Specifically, the Control class exhibited a slight increase in precision (from 72.60% to 73.82%), while recall remained consistently high (from 89.08% to 88.65%), reflecting fewer false positives without significant loss of true positives. In the MCI class, recall substantially improved (from 43.14% to 54.90%) with a slight increase in precision (from 61.11% to 62.22%), greatly enhancing the model's ability to identify MCI cases. For the AD class, precision improved (from 76.84% to 78.26%), thus reducing false positives. Conversely, recall slightly decreased (from 55.30% to 54.55%); however, specificity remained strong, ensuring reliable AD identification. The

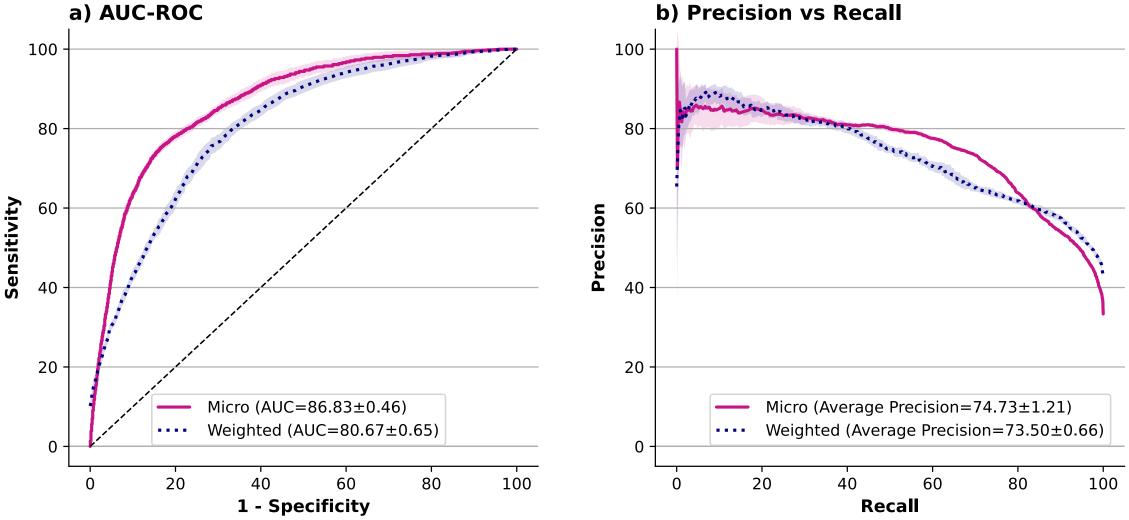

Fig. 1 | Model performance evaluated using AUC and precision–recall curves. **a** AUC-ROC curves (micro vs. weighted averaging). The micro average aggregates outcomes across all classes, yielding an AUC of 86.83%, while the weighted average, reflecting class frequency, achieves an AUC of 80.67%. **b** Precision–recall curves. The micro average achieves an average precision of 74.73%, and the weighted average yields an average precision of 73.50%. Overall, these metrics highlight how SpeechCARE balances sensitivity and specificity (in AUC) as well as precision and recall across the three diagnostic categories.

Table 1 | Effects of threshold optimization on precision and recall for Control, MCI, and AD classes

Result	Precision			Recall			F1-score
	Control	MCI	AD	Control	MCI	AD	
Original result	72.60	61.11	76.84	89.08	43.14	55.30	72.57
After threshold optimization	73.82	62.22	78.26	88.65	54.90	54.55	73.54

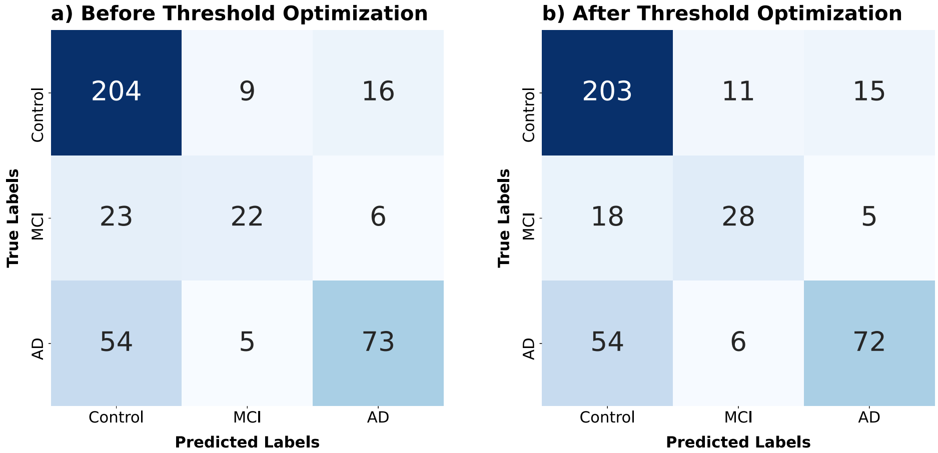

Fig. 2 | Confusion matrices before and after threshold optimization. **a** Rows represent actual class labels (Control, MCI, AD) and columns represent predicted labels. The diagonal cells show correct classifications for each diagnostic category, with off-diagonal cells indicating misclassifications. **b** Comparing panel a and panel b highlights how adjusting classification thresholds improves MCI detection rates while minimally affecting Control and AD predictions.

confusion matrices of classification outcomes before and after optimization are illustrated in Fig. 2.

Fairness analysis

We evaluated model fairness across demographic and linguistic subgroups—before and after applying bias mitigation strategies—using Equality of Opportunity (EOO), which measures consistency in true positive rates (TPRs), and Average Odds (AO), which captures disparities in both true positive and false positive rates (FPRs). All results were reported as averages with 95% confidence intervals across 10 runs with different random seeds (see Fig. 3a, b). Prior to mitigation, age-related disparities were the most pronounced: participants aged 80 and older had the lowest fairness scores (EOO: 51.33%; AO: 37.13%), while the 40–65 age group showed the highest (EOO: 80.04%; AO: 45.69%). Education-based variation was also evident, with AO ranging from 36.65% to 42.51% and EOO from 47.12% to 63.20% across education levels. Gender-based differences were relatively minor, with females (EOO: 61.54%; AO: 39.44%) slightly lower than males (EOO: 61.65%; AO: 40.36%). Linguistic disparities were notable: Spanish speakers exhibited the lowest fairness scores (EOO: 50.14%; AO: 37.05%), while Mandarin (EOO: 91.11%; AO: 45.56%) and English speakers (EOO: 62.86%; AO: 40.11%) had higher scores.

Following the application of frequency-based reweighting and threshold optimization, we observed substantial reductions in subgroup disparities. Among age groups, fairness improved for the 80+ group (EOO: 60.56%; AO: 41.37%), and the overall EOO range narrowed to 60.56–74.91%, reflecting more balanced TPRs. Education-level disparities were also mitigated: AO scores increased and became more consistent across levels (post-mitigation range: 40.59–44.28%), and EOO scores became more aligned across education subgroups. Gender fairness improved modestly, with AO rising to 42.78% for females and 41.63% for males. Language-based improvements were limited; English speakers' AO improved to 42.58%, Mandarin speakers' AO improved to 50.00%, and Spanish speakers' AO decreased slightly to 34.25%. Overall, bias mitigation strategies were most effective in improving fairness across age and education groups, with moderate gains for gender and minimal impact for language subgroups.

Modality weights analysis

To assess the relative contribution of each input modality, we computed the average gating weights assigned to the acoustic, linguistic, and demographic streams—where demographic here refers specifically to participant age in categorical form—across all speech tasks in the test set (Fig. 4).

Across the various speech tasks, the model dynamically shifted its reliance on acoustic, linguistic, and demographic inputs in ways that reflected the cognitive and structural demands of each task. For more lexically driven tasks (e.g., semantic verbal fluency), linguistic features dominated, indicating a strong dependence on vocabulary and semantic organization. In contrast, tasks involving spontaneous speech and structured narratives (e.g., picture description) showed a more balanced contribution from both linguistic and acoustic features, suggesting that prosodic cues and speech delivery complemented lexical content in these contexts.

When speech content was relatively brief or highly constrained (e.g., voice assistant interactions), acoustic features became the primary driver, and demographic input (participant age) also played a relatively larger role—likely capturing speaker-specific vocal characteristics and traits in the absence of rich lexical cues. Across most tasks, demographic input contributed modestly beyond the combined acoustic and linguistic inputs, highlighting how the model weights each modality in direct response to the communicative and cognitive demands of the task.

External generalizability evaluation- ADReSSo challenge 2021

Under the ADReSSo 2021 benchmark, SpeechCARE achieved an AUC of 91.93 ± 0.92%, an F1-score of 85.08 ± 1.24%, and an accuracy of 86.62 ± 0.86%, with 95% confidence intervals calculated across 10 training runs, indicating stable performance and strong generalization on this widely used English-language benchmark.

According to the ADReSSo-2021 Challenge report[20], participants employed a variety of acoustic, linguistic, and fusion strategies for cognitive impairment detection. Acoustic features ranged from hand-crafted descriptors (e.g., Mel-Frequency Cepstral Coeficients[21], prosodic features) to pre-trained transformer embeddings from models like Wav2vec 2.0, which often outperformed handcrafted features. Linguistic features included both hand-crafted metrics (e.g., Linguistic Inquiry and Word Count[22],

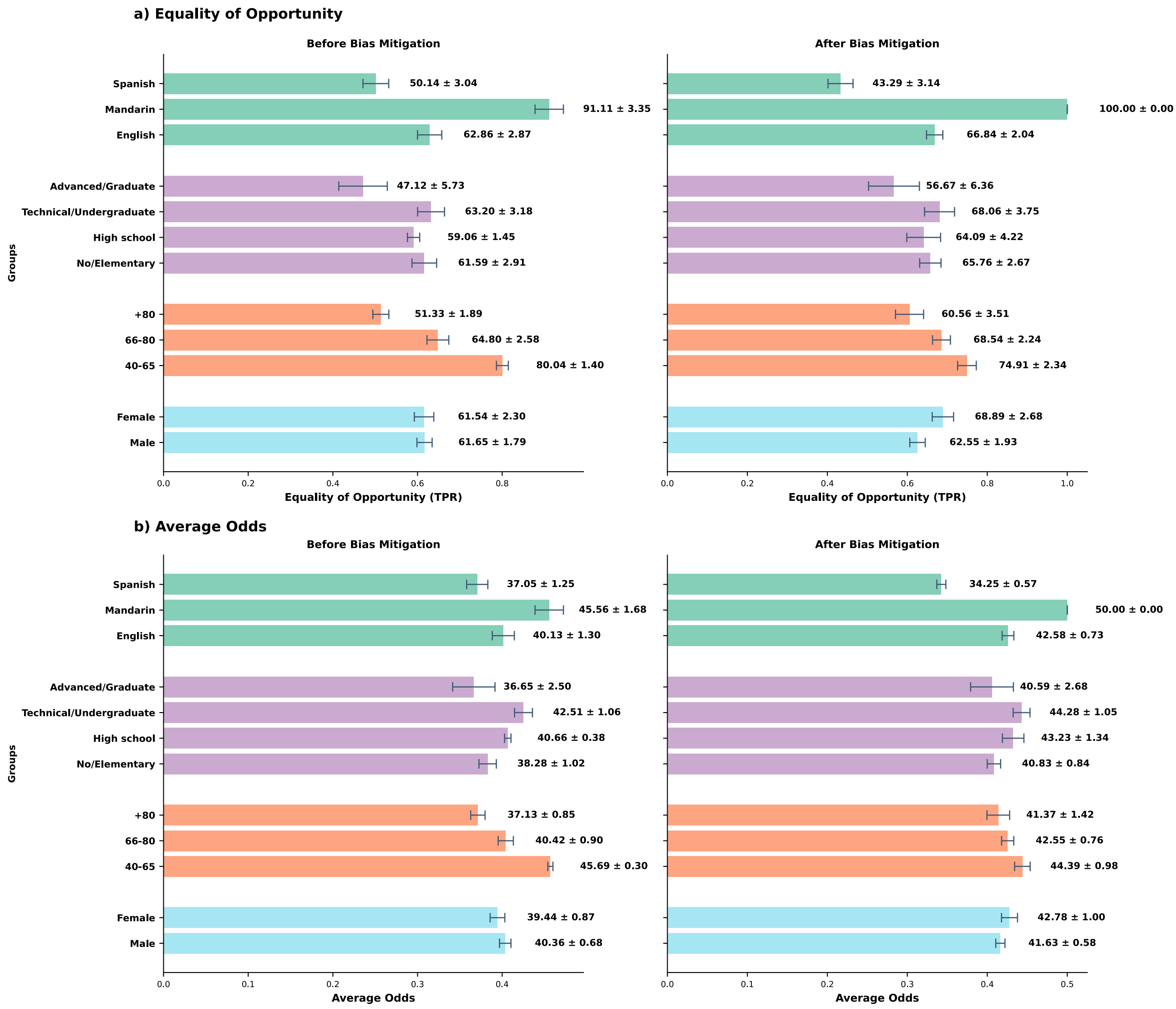

Fig. 3 | Fairness metrics by subgroup before and after bias mitigation. **a** Average Odds (AO). Horizontal bars show average ± 95% CI (across 10 random seeds) for each age, gender, education, and language subgroup before (left panel) and after (right panel) frequency-based reweighting plus threshold optimization. Mitigation narrows AO gaps—most noticeably across age (e.g., AO drops from 45.69% to 44.39% in the 40–65 age group and rises from 37.13% to 41.37% in the 80+ group) and education levels—while language disparities persist, with Spanish remaining the lowest. **b** Equality of Opportunity (EOO). The same subgroup layout is shown for EOO. Post-mitigation, true-positive-rate differences shrink: the +80-age group rises from 51.33% to 60.56%, and EOO values for all age brackets converge to the 61.56–75.91% range. Education and gender subgroups exhibit smaller—but still positive—gains, whereas language subgroups show limited improvement.

Part of Speech tags) and contextual embeddings from BERT-based models. Fusion strategies included early fusion (feature concatenation), late fusion (decision-level aggregation), and ensemble fusion (e.g., majority or weighted voting). Reported F1-scores ranged from 70.89% to 84.93%, and accuracies from 67.61% to 84.93%[20]. However, as F1-score computations varied across submissions, accuracy was viewed as a more consistent basis for comparison.

Top-performing pipelines in the challenge employed various fusion strategies and pre-trained transformers. Notable examples include: Pan et al.[23] achieved an F1-score of 84.93% and an accuracy of 84.51% by fine-tuning a BERT-Large and employing automatic speech recognition (ASR) confidence scores for final classification. Syed et al.[24] extracted embeddings (without fine-tuning) from five different BERT-derivative transformers (e.g., RoBERTa[25], DistilBERT) and used an SVM classifier. They combined labels (using majority voting) derived from each transformer's embeddings and achieved an F1-score of 84.45% and an accuracy of 84.51%. Qiao et al.[26] separately trained two logistic regression models on hand-crafted linguistic features and fine-tuned BERT and ERNIE[27] transformers. They then ensembled all models using Stacked Generalization[28] and achieved an F1-score of 82% and an accuracy of 83%.

Additionally, several recent works have validated models on this dataset. Examples of top-performing studies include: (1) Ilias and Askounis[29] proposed a multimodal approach by fine-tuning BERT and DeiT[30] (operating on audio spectrograms) transformer models. By incorporating context-based self-attention[31], optimal transport domain adaptation[32], and label smoothing[33], and integrating modalities using Co-attention fusion[34], they achieved an F1-score of 85.27% and an accuracy of 85.35%. (2) Bang et al. [35] extracted acoustic embeddings from Wav2Vec 2.0, transcription-based embeddings from BERT, and used another BERT to extract embeddings from opinions generated by ChatGPT as a third modality. They used transformer blocks to enrich modality-specific embeddings, fused the resulting embeddings by intermediate fusion, and classified them using a single linear layer, achieving an F1-score of 87.25% and an accuracy of 87.32%. (3) Shao et al. [36] (2025) achieved an F1-score of 84.27% and an accuracy of 84.30% using cross-modal attentions on Wav2vec 2.0 and BERT transformers.

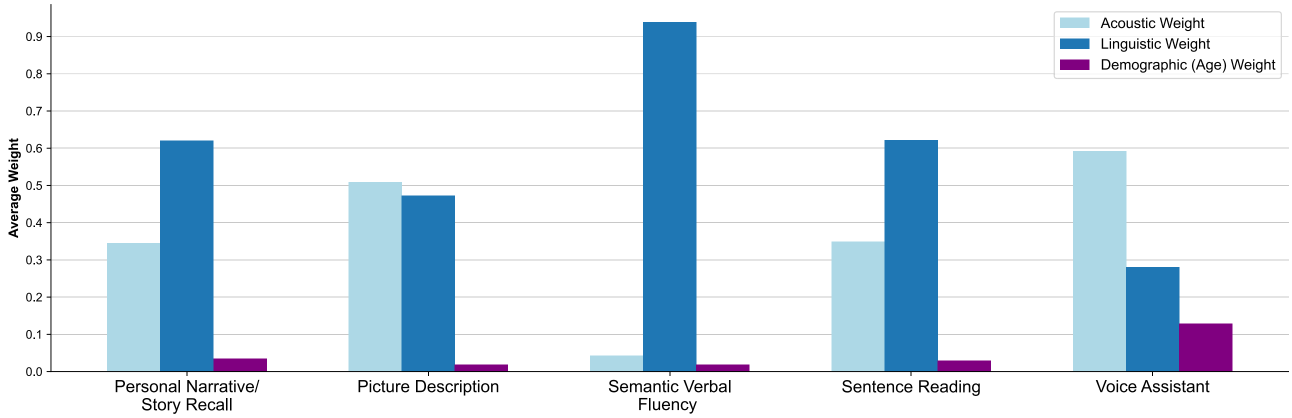

Fig. 4 | Modality weights (acoustic, linguistic, demographic) assigned by the SpeechCARE Adaptive Gating Fusion (AGF) network across various speech tasks in the test set. Each cluster of bars represents one task, showing how the model differentially weights acoustic (light blue), linguistic (dark blue), and demographic (purple) features depending on the cognitive demands of the task.

Compared to these studies, our results show that SpeechCARE offers strong adaptability to small monolingual speech datasets, achieving performance that matches or exceeds existing pipelines on the ADReSSo benchmark.

External generalizability evaluation- DementiaBank Chou Corpus (Mandarin)

Using the transfer learning approach, fine-tuning SpeechCARE on the Mandarin-speaking Chou dataset resulted in an AUC of 92.67 ± 4.54% and an F1-score of 85.47 ± 3.65%, with 95% confidence intervals calculated across 10 runs. These results demonstrate that SpeechCARE is capable of generalizing to multilingual contexts when exposed to representative training data. This finding highlights the critical role of linguistic diversity in model development to ensure robust and equitable performance across languages in cross-lingual cognitive impairment screening applications.

Comparative and ablation analysis of SpeechCARE components

Table 2 summarizes the results of all comparative and ablation studies performed on SpeechCARE, evaluated on both the validation and test sets.

We began by examining acoustic-only refinements. According to Table 2, adding a learnable [CLS] embedding to the mHuBERT baseline increased the test F1-score from 66.80% to 67.77%. Introducing 5-s segments with 20% overlapping segmentation recovered the slight AUC loss and yielded an additional ≈0.5-point F1-score increase (68.23%). Together, these modifications confirm that global contextualization and audio segmentation are both able to improve the performance of speech transformer models.

We next evaluated single-modality baselines. On the official Test set, the speech-only model achieved an F1-score of 68.23%, and an AUC of 84.85%, and the transcription-only model reached an F1-score of 68.88% with an AUC of 85.00%. In contrast, the demographics-only model performed substantially worse, with a reduction of about 13 points in F1-score and AUC. These results indicate that while both speech and transcription modalities independently capture meaningful information, demographic features alone offer limited predictive utility (Table 2).

We then explored modalities integration. Fusing speech and transcription using Adaptive Gating Fusion (AGF) improved the test F1-score by ≈3 points over the best unimodal model, reaching 70.51%. Adding all demographic variables diluted this benefit, but including age alone produced the highest overall performance: F1-score of 72.11% and AUC of 86.83%, indicating that age provides a complementary signal when dynamically weighted by AGF (Table 2).

We further compared fusion strategies. AGF outperformed Intermediate Fusion, Scaled Late Fusion, and Cross-Modal Attention by 1–2 points in F1-score, while maintaining comparable AUC, demonstrating superior precision–recall trade-offs through adaptive modality weighting (Table 2).

Finally, we assessed the impact of audio noise reduction. Relative to the low-pass–filtered baseline (AUC = 86.83%a; F1-score =72.11%), SpeechCARE-AGF trained on raw audio reduced AUC by ≈1.1 points and F1-score by ≈3 points. Applying the CMGAN neural denoiser resulted in comparable AUC but the largest decline in F1-score (−3.1 points to 69.0%), suggesting that aggressive enhancement can distort cognitively relevant cues. In contrast, a simple 8 kHz low-pass filter—which attenuates only high-frequency noise while preserving the human voice frequency components—achieved the highest performance, underscoring that targeted spectral trimming is more effective and less risky than either no filtering or complex denoising in cognitive impairment studies (Table 2).

As shown in Table 3, paired *t*-tests across all ablation analyses confirm that the performance gains (measured by F1-score) reported in Table 2 for SpeechCARE components are statistically significant. The effect sizes (Cohen's d) further indicate that most differences across these comparisons are of medium to large magnitude.

Discussion

In this study, we introduce SpeechCARE, a multimodal, multilingual pipeline developed to detect cognitive impairment using brief speech recordings. The system begins with an advanced preprocessing stage that integrates demographic data preparation and imputation, noise reduction, transcription, and LLM-based data anomaly detection and speech-task identification (e.g., semantic fluency, sentence reading). The core architecture combines linguistic (mGTE) and acoustic (mHuBERT) transformer models, enhanced by a specialized encoding component that captures global temporal patterns across the input audio. A novel Adaptive Gating Fusion (AGF) mechanism dynamically weights acoustic, linguistic, and demographic modalities, optimizing classification performance. When evaluated on the NIA PREPARE challenge dataset—comprising multilingual speech samples (English, Spanish, Mandarin) across diverse speech tasks—SpeechCARE achieved an average F1-score of 72.11%, receiving a special recognition prize from the NIA.

External validation demonstrated the generalizability of SpeechCARE across datasets and languages. Without any architectural modifications, the model achieved strong performance on the English-only ADReSSo 2021 dataset (AUC = 91.93%; F1-score = 85.08%), performing comparably to or better than previously published systems. On the Mandarin Chou Corpus, where all Mandarin speech in the PREPARE dataset was labeled as MCI, zero-shot inference initially failed due to a language–label confound. However, after fine-tuning on the small Chou training set ($n = 51$), performance was substantially recovered (AUC = 92.67%; F1-score = 85.47%). These results indicate that SpeechCARE captures language-agnostic

Table 2 | Results of comparative and ablation analysis of SpeechCARE components

Model	Validation		Test	
	AUC	F1-score	AUC	F1-score
Acoustic-only refinements				
mHuBERT (Base Model)	84.03 ± 1.09	66.78 ± 2.27	84.07 ± 0.60	66.80 ± 1.25
mHuBERT + CLS Embedding	84.92 ± 1.23	68.06 ± 1.83	84.99 ± 0.60	67.77 ± 1.06
mHuBERT + CLS Embedding + Segmentation	84.55 ± 1.14	67.60 ± 1.56	84.85 ± 0.70	**68.23 ± 1.11**
Single-modality baselines and modalities integration				
All Demographics (Age, Gender, Education)	72.78 ± 0.79	55.82 ± 0.80	72.31 ± 0.71	55.70 ± 0.43
Voice (mHuBERT + CLS Embedding + Segmentation)	84.55 ± 1.14	67.60 ± 1.56	84.85 ± 0.70	68.23 ± 1.11
Transcription (mGTE)	81.26 ± 1.17	63.70 ± 1.26	85.00 ± 0.40	68.88 ± 0.78
Fusion-AGF: Voice + Transcription	84.42 ± 1.95	67.57 ± 2.36	86.57 ± 0.45	70.51 ± 0.93
Fusion-AGF: Voice + Transcription + All demographics	83.49 ± 0.96	66.14 ± 1.74	85.49 ± 0.69	68.42 ± 0.78
Fusion-AGF: Voice + Transcription + Education	83.19 ± 2.09	66.02 ± 2.87	85.99 ± 0.68	69.20 ± 1.00
Fusion-AGF: Voice + Transcription + Gender	84.02 ± 1.67	66.78 ± 2.18	86.35 ± 0.45	69.95 ± 0.68
Fusion-AGF: Voice + Transcription + Age	84.97 ± 1.57	68.12 ± 2.69	86.83 ± 0.46	**72.11 ± 0.44**
Fusion strategies				
Intermediate Fusion	85.07 ± 1.35	67.94 ± 2.09	86.28 ± 0.48	70.10 ± 0.78
Scaled Late Fusion	85.19 ± 1.38	68.62 ± 2.77	86.21 ± 0.57	70.29 ± 1.13
Cross-Modal Attention (+Intermediate Fusion)	85.49 ± 1.50	68.58 ± 2.05	86.61 ± 0.56	70.51 ± 0.71
Adaptive Gating Fusion (AGF)	84.97 ± 1.57	68.12 ± 2.69	86.83 ± 0.46	**72.11 ± 0.44**
Noise reduction (audio preprocessing)				
SpeechCARE-AGF: Raw Audio	84.25 ± 1.33	85.76 ± 0.80	85.76 ± 0.80	69.15 ± 1.01
SpeechCARE-AGF: CMGAN-Enhanced Audio	83.22 ± 1.89	65.34 ± 2.26	85.80 ± 1.10	69.00 ± 1.12
SpeechCARE-AGF: Low-Pass Filtered Audio	84.97 ± 1.57	68.12 ± 2.69	86.83 ± 0.46	**72.11 ± 0.44**

best F1-score of each category is presented in bold.

markers of cognitive impairment while maintaining the flexibility to adapt to language-specific characteristics, supporting its applicability across diverse clinical and linguistic settings.

Growing evidence indicates that subtle speech alterations are early indicators of cognitive decline. However, assembling sufficient and diverse speech data poses significant hurdles due to privacy constraints, high costs, and logistical issues. Earlier competitions, such as the 2021 ADReSSo challenge, relied on just 237 English-only samples—limiting wider applicability. In 2023, the NIA addressed these shortcomings by curating a multilingual dataset of 2058 participants for the 2024 NIA challenge, encouraging the development of generalizable algorithms. Among the top ten submissions, classification losses ranged from 0.629 to 0.688, with our approach achieving a loss of 0.642. Teams employed diverse methods, including LLM-based feature extraction, transformer-based models, and hand-crafted acoustic features. Our analysis revealed that purely LLM-based and hand-crafted approaches struggled with limited datasets. Conversely, multilingual transformer models exhibited greater robustness, superior generalization across languages, and sensitivity to nuanced speech patterns.

Fairness in speech processing algorithms, especially for cognitive impairment detection, is an essential consideration in healthcare applications. In our study, we explicitly measured potential biases using metrics such as EOO and Average Odds, revealing modest but notable performance disparities across language and demographic groups—particularly for Spanish speakers, likely due to smaller sample sizes and higher transcription error rates. We also observed mild differences across gender subgroups, which may reflect acoustic profile variations (e.g., pitch, phonation) or disparities in sample sizes. Although bias measurement has been discussed in prior machine learning work, standardized reporting remains uncommon for healthcare-oriented speech processing systems, making our efforts an important step toward transparency in such applications. Previous studies have rarely reported these approaches in speech-based algorithms for cognitive impairment detection, underscoring the novelty and clinical relevance of our work. In ongoing work, we are incorporating clinically accepted bias-mitigation strategies—such as oversampling, adversarial learning, and post-hoc calibration—to further reduce demographic disparities.

Recent regulatory advancements also highlight the growing importance of multi-modal approaches. In May 2025, the FDA approved Fujirebio's Lumipulse G pTau217/β-amyloid 1-42 blood test for Alzheimer's disease, offering a minimally invasive way to detect biological changes associated with the condition. While this represents a significant advance, it does not provide information about how cognitive decline affects a person's ability to communicate in daily life. Language and speech changes—such as reduced fluency, disorganized sentences, or altered vocal patterns—often emerge early and can reflect real-world functional decline that biological tests alone cannot capture. SpeechCARE fills this critical gap by analyzing short voice recordings to identify subtle communication changes in real time. Its Mixture-of-Experts–inspired gating fusion is designed to incorporate additional inputs (e.g., age, brain imaging) to generate a more complete clinical picture. By combining biological confirmation through blood-based testing with speech-based insights, clinicians may gain a fuller understanding of both the underlying disease and its impact on everyday functioning—supporting more informed decisions about imaging, referrals, and early intervention.

High-quality speech data are essential to detect subtle acoustic and linguistic cues of early cognitive decline, particularly for MCI detection. However, non-stationary background noise—such as overlapping conversations—complicated model training in the NIA PREPARE challenge, as most recordings were collected in uncontrolled clinical settings. Although sophisticated noise reduction systems exist, removing overlapping speech without affecting nuanced participant cues remains challenging and can

Table 3 | Statistical comparisons of SpeechCARE model variants using paired *t*-tests on F1-scores

Comparison	Mean	SD	*P*-value	Effect size[b]
Acoustic-only refinements				
mHuBERT (Base Model) vs. mHuBERT + CLS Embedding + Segmentation	68.23 vs. 66.80	1.55 vs. 1.74	0.0358[a]	0.78 (medium)
Single-modality baselines and modalities integration				
Fusion-AGF: Voice + Transcription + Age vs. Transcription-mGTE (best unimodal model)	72.11 vs. 68.88	0.62 vs. 1.10	0.0000[a]	2.619 (large)
Fusion strategies				
Adaptive Gating Fusion vs. Intermediate Fusion	72.11 vs. 70.10	0.62 vs. 1.10	0.0006[a]	1.627 (large)
Adaptive Gating Fusion vs. Scaled Late Fusion	72.11 vs. 70.29	0.62 vs. 1.59	0.0068[a]	1.105 (large)
Adaptive Gating Fusion vs. Cross-Modal Attention	72.11 vs. 70.51	0.62 vs. 0.99	0.0003[a]	1.79 (large)
Noise reduction (audio preprocessing)				
SpeechCARE-AGF: Low-Pass Filtered Audio vs. SpeechCARE-AGF: Raw Audio	72.11 vs. 69.15	0.62 vs. 1.42	0.0003[a]	1.77 (large)
SpeechCARE-AGF: Low-Pass Filtered Audio vs. SpeechCARE-AGF: CMGAN-Enhanced Audio	72.11 vs. 69.00	0.62 vs. 1.56	0.0004[a]	1.721 (large)

[a] Statistically significant difference at $p < 0.05$.

[b] Effect sizes are reported as Cohen's d, calculated as the mean of the paired differences relative to their standard deviation.

impair algorithm performance. Therefore, using overhead microphones and minimizing background speech are recommended for this task.

Additionally, although tasks like picture description and reading elicited spontaneous speech, they were often prompted by clinicians and may not fully capture cues present in naturalistic communication. Previous research indicates that communication cues such as disrupted turn-taking, delayed responses, and repair sequences, which are strongly associated with cognitive impairment. In our ongoing research, we audio-record naturally occurring patient–nurse interactions in home healthcare settings, enabling the use of speech-processing algorithms without requiring structured prompts. This work aims to integrate real-world communication data with home healthcare workflows, clinical records, and social determinants of health to support automated identification of speech-related impairments and associated clinical and social risk factors. This approach not only improves early detection of MCI and early dementia but also enhances clinical documentation and enables automated notifications to clinicians regarding patients' potential cognitive impairment—facilitating timely decision-making and more targeted interventions. Together, these elements support the real-world adoption of automated speech-based screening tools in home healthcare settings.

Despite these promising findings, several limitations should be acknowledged. SpeechCARE was developed to distinguish among healthy controls, individuals with MCI, and those with Alzheimer's disease, as these were the only diagnostic categories available in the NIA PREPARE challenge dataset. Other neurological or psychiatric conditions that may exhibit similar cognitive or speech-related symptoms, such as depression, delirium, or vascular dementia were not included. This restricted diagnostic scope limits the generalizability of our findings to real-world clinical populations. Future work should consider incorporating datasets with a broader range of diagnostic categories or applying multi-task learning approaches to improve differential diagnosis across overlapping cognitive conditions.

While the PREPARE challenge dataset included English, Spanish, and Mandarin speech samples, it lacked systematic representation of dialectal variation within each language, such as regional accents or race-associated vernaculars for the English language. These within-language differences can influence both transcription accuracy and model performance, potentially introducing bias, particularly for underrepresented dialects. Addressing this limitation will require more granular subgroup analyses and the inclusion of additional speech data reflecting diverse dialects to ensure equitable performance across sociolinguistic subgroups.

Speech samples in this study were collected through structured tasks (e.g., picture description, sentence reading), which may not fully reflect the dynamics of naturalistic conversation or capture spontaneous communication challenges encountered in everyday settings. Additionally, variability in recording quality—such as background noise and inconsistencies in speaker diarization—was observed across speech corpora of the NIA PREPARE challenge dataset, potentially affecting the reliability of acoustic and linguistic analyses. Future studies should prioritize collecting speech data in real-world clinical environments using standardized recording protocols and task-free interactions to improve ecological validity and robustness of speech processing algorithms.

Overall, this work demonstrates that SpeechCARE's multimodal, multilingual pipeline offers a practical and accurate approach to cognitive impairment screening. By intelligently fusing acoustic, linguistic, and demographic features, along with adopting robust data preprocessing and threshold calibration methods, our system achieved high performance even with limited training samples. These findings pave the way for broader clinical applications, including remote screening and telemedicine platforms, which may facilitate earlier detection and intervention for MCI and dementia in diverse populations.

Methods

Figure 5 presents an overview of the study's methodology.

Data: PREPARE challenge dataset

The PREPARE challenge dataset includes 2058 participants, of whom 1646 were assigned to training and 412 to testing, drawn from 10 speech corpora featuring various tasks across three languages. English accounted for 80.42% ($n = 1655$) of the recordings, Spanish for 17.49% ($n = 360$), and Mandarin for 2.09% ($n = 43$). Clinically, 1140 participants were cognitively healthy, 268 had MCI, and 650 had Alzheimer's disease. Female participants comprised 58.2%, and ages ranged from 46 to 99 years (mean = 75.13, SD = 8.65). Education was recorded either numerically (1–20 years) or categorically (e.g., elementary school, associate degree), with 35.2% missing (see section "Demographic information preparation"). Race was reported for fewer than 8% of participants, resulting in 92.61% of entries missing (see Table 4 for further details).

Audio recordings averaged 27 s in length. Most recordings (66.5%) were exactly 30 s, followed by 20.7% between 20 and 30 s, 11.9% between 10 and 20 s, and 0.9% under 10 s. The distribution was similar across diagnostic classes. According to the challenge guidelines, all recordings were truncated to a maximum of 30 s, but the specific truncation procedure was not described. Very short recordings were mostly due to issues such as missing audio or technical failures.

We created a validation set by randomly selecting 20% of the training data (329 participants), while maintaining a stratified split across diagnostic classes (Control, MCI, AD) and demographic variables (age, gender, education). This ensures fair performance evaluation and prevent overfitting. This approach preserved the overall distribution of key characteristics present in the original training set. Throughout the training process for all experiments, we used this validation set for early stopping and hyperparameter tuning to avoid data leakage from the final test set (see section "Model training and hyperparameter tuning" for details).

Preprocessing components

To incorporate demographic variables into our model and enable bias analysis across demographic groups, we applied the following procedures. Age was grouped into mid-life adults (46–65 years, 12%), older adults

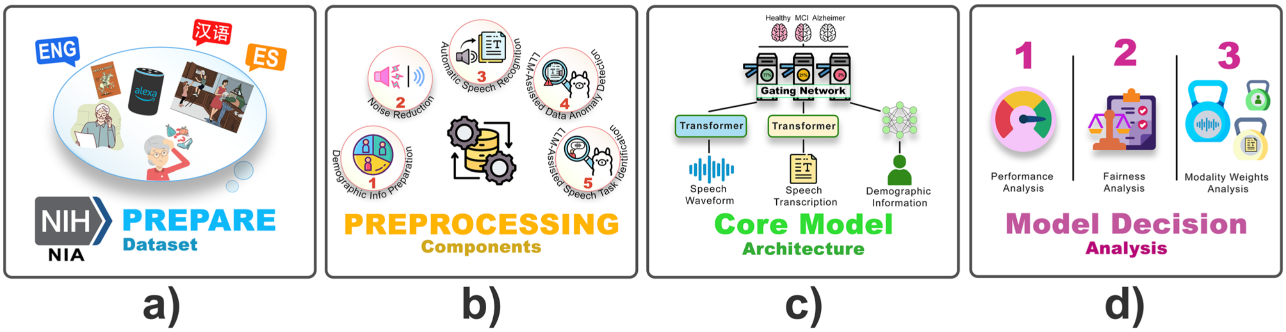

Fig. 5 | Overview of the SpeechCARE Screening Pipeline. a The pipeline begins with multilingual speech data from the NIH/NIA PREPARE dataset, including English, Spanish, and Mandarin recordings from various sources; **b** Preprocessing involves demographic data extraction, noise reduction, automatic speech recognition, and LLM-assisted detection of anomalies and task types; **c** The core model combines acoustic, linguistic, and demographic features using transformer-based encoders and an adaptive gating mechanism to classify cognitive status; **d** Model decision analysis includes performance evaluation, fairness assessment, and analysis of modality contributions.

Table 4 | Summary of participant characteristics of the PREPARE dataset shared by the challenge organizers

Attribute	Training set			Test set		
	Control	MCI	AD	Control	MCI	AD
Cognitive status	911 (55.3%)	217 (13.2%)	518 (31.5%)	229 (55.6%)	51 (12.4%)	132 (32.0%)
Gender-Female	540 (59.3%)	123 (56.7%)	297 (57.3%)	150 (65.5%)	29 (56.9%)	80 (60.6%)
Age category						
–65	111 (12.2%)	40 (18.4%)	57 (11.0%)	25 (10.9%)	4 (7.8%)	10 (7.6%)
65–80	362 (39.7%)	121 (55.8%)	214 (41.3%)	93 (40.6%)	26 (51.0%)	65 (49.2%)
+ 80	438 (48.1%)	56 (25.8%)	247 (47.7%)	111 (48.5%)	21 (41.2%)	57 (43.2%)
Language						
English	760 (83.4%)	115 (53.0%)	456 (88.0%)	184 (80.3%)	20 (39.2%)	120 (90.9%)
Spanish	151 (16.6%)	68 (31.3%)	62 (12.0%)	45 (19.7%)	22 (43.1%)	12 (9.1%)
Mandarin	0 (0.0%)	34 (15.7%)	0 (0.0%)	0 (0.0%)	9 (17.6%)	0 (0.0%)
Education level						
No/Elementary	92 (10.1%)	52 (24.0%)	39 (7.5%)	29 (12.7%)	18 (35.3%)	9 (6.8%)
High school	276 (30.3%)	40 (18.4%)	92 (17.8%)	72 (31.4%)	10 (19.6%)	24 (18.2%)
Technical/ Undergraduate	157 (17.2%)	38 (17.5%)	38 (17.5%)	44 (19.2%)	10 (19.6%)	20 (15.2%)
Advanced/Graduate	132 (14.5%)	21 (9.7%)	49 (9.5%)	22 (9.6%)	2 (3.9%)	10 (7.6%)
Missing	254 (27.9%)	66 (30.4%)	263 (50.8%)	62 (27.1%)	11 (21.6%)	69 (52.3%)

(66–80 years, 52.7%), and elderly individuals (80+ years, 35.3%) to reflect key stages of cognitive aging[37,38].

Education was categorized into four levels: elementary, high school, technical/undergraduate, and advanced/graduate, approximately corresponding to the International Standard Classification of Education ISCED–11 levels 0–1, 2–3, 4–6, and 7–8[39]. These categories were designed to capture differences in cognitive reserve, a known modifier of dementia risk[40]. Because 35.23% of the education data was missing, we used the Iterative Imputer method[41] (available in the Scikit-learn package), which models each incomplete variable as a function of other features and iteratively estimates missing values. Using age, gender, and language as predictors, this approach helped preserve the integrity of the education data by leveraging relationships among observed variables.

To evaluate and improve data quality, we assessed noise in the PREPARE dataset to guide preprocessing and ensure real-world generalizability. Recordings contained both stationary (device hum) and dynamic (background chatter) noise, varying across the 10 corpora. To evaluate perceived noise severity, we computed metrics like Signal-to-Noise Ratio[42] and Spectral Flatness Measure[43]. However, inconsistent preprocessing by various corpus developers led to outcomes not strongly correlated with human perception of noise. Consequently, we performed a Mean Opinion Score[44] assessment on a 5% stratified subset ($N = 103$), balancing diagnosis, language, and corpus. Two expert raters (HA, MZ) independently used a 1–5 Likert scale, finding 62.4% severely noisy recordings (scores 1–2), 14.7% moderately noisy (score 3), and 22.9% slightly noisy (scores 4–5).

For noise reduction, we applied a uniform low-pass filter at 8 kHz to eliminate high-frequency noise while preserving speech components[45]. Although advanced neural speech enhancement models can handle more complex noise patterns (e.g., human conversation noises), they may compromise critical acoustic cues vital for detecting cognitive impairment by introducing artifacts or over suppress frequency and spectral features. Therefore, we chose more reliable techniques. To empirically demonstrate the limitations of neural enhancement models, we conducted an ablation study (see *Results: Comparative and Ablation Analysis of SpeechCARE Components*) assessing their impact on SpeechCARE's performance.

We transcribed audio data using Whisper-Large[46], a multilingual ASR system with state-of-the-art performance among open-source models. This allows us to integrate linguistic cues into the SpeechCARE pipeline. To evaluate Whisper's transcription accuracy, we measured the word error rate[47] (WER) on a stratified random sample (by diagnostic class, corpus, gender, and age), comprising 5% of English ($N = 83$), 10% of Spanish ($N = 36$), and 15% of Mandarin ($N = 9$) recordings. Comparing Whisper's output with human transcriptions yielded WERs of 0.12 (English), 0.27 (Spanish), and 0.32 (Mandarin), indicating high accuracy for English but more challenges for Spanish and Mandarin, likely due to phonetic and grammatical variations and limited training data in those languages.

We employed an open-weight LLM (LLaMA-3.1-70b-instruct[48]) for data anomaly detect. While evaluating Whisper transcriptions, we found some audio files contained only clinician speech, likely due to speaker-diarization errors by corpus developers. To identify these systematically, we employed LLaMA-3.1-70b-instruct using prompt engineering and few-shot learning. We began by selecting and labeling a random sample comprising 100 transcripts from participants and 10 transcripts containing only clinician speech. From these, 50 participant and 5 clinician-only transcripts were used as candidate examples for few-shot prompt construction, while the remaining 55 served as a validation set to evaluate prompt performance. We then iteratively refined and validated our prompt until the LLM achieved 100% accuracy on the validation set.

Applying the final prompt (available in this GitHub link) to the remaining dataset revealed 12 additional clinician-only transcripts, bringing the total to 22. These were removed from the data for model training to prevent confusion during model learning.

We employed an open-weight LLM (LLaMA-3.1-70b-instruct) for speech task identification. The PREPARE dataset features speech samples derived from multiple speech tasks; however, it lacks metadata specifying which task each recording represents. Identifying the task type thus helps clinicians contextualize model outputs and understand the cognitive processes behind observed speech changes. For instance, picture description tasks reveal deficits in semantic and syntactic organization, reading tasks expose articulation and prosodic patterns, and verbal fluency tasks highlight lexical access and executive function.

To classify the task type of each sample, we used LLaMA-3.1-70b-instruct with prompt engineering and few-shot learning[49]. We began by creating an annotated dataset of 150 randomly selected recordings, stratified by language, corpus, and gender. Two experts in cognitive assessment manually labeled each recording by task type. Fifty samples were used for selecting few-shot demonstrations, and the remaining 100 formed the validation set for evaluating prompt performance. We iteratively refined the prompt and demonstrations (few shots) by reviewing misclassifications and adjusting the instructions after each round. For example, we introduced an "uncategorized" label for transcripts that were too short to interpret, often due to recording errors. After each update, we evaluated the prompt on the validation set and continued refining until the model achieved 100% accuracy on the validation dataset.

When applied to the entire dataset, the final prompt (available in this GitHub link) labeled six task types: personal narrative, story recall, picture description, semantic verbal fluency, sentence reading, and voice assistant interactions (see Fig. 6). Only 2% of recordings were labeled "unrecognized".

Model architecture

Given the multilingual and multi-task nature of the PREPARE challenge, SpeechCARE core architecture is built on multilingual transformer models and has two core components:

1. Feature network: Generates representations from pre-trained linguistic and acoustic transformers and can incorporate additional modalities—such as demographic information in this study.
2. Fusion network: Combines these features using a novel integration method that enhances overall prediction accuracy.

Feature network

Pre-trained linguistic transformers (e.g., BERT[50]) have often been used in cognitive impairment research, achieving variable results depending on dataset size, language, speech task, and fine-tuning strategies[51–53]. In this study, SpeechCARE's core architecture includes mGTE[54] (*multilingual Generative Text Encoder*), an encoder-only transformer with 305 million parameters released by Alibaba in 2024. Pre-trained on roughly 1028 billion tokens across 75 languages, mGTE produces 769-dimensional embeddings and can process sequences up to 8192 tokens—roughly equivalent to 20–30 min of transcribed speech— well beyond BERT's 512-token limit. This extensive context length allows mGTE to effectively encode the disfluencies and syntactic errors often associated with cognitive decline.

To extract linguistic features, we followed standard transformer fine-tuning practices and used the [CLS] embedding from mGTE's last layer as a summary representation of each transcript. Specifically, given a tokenized sequence of N tokens $L = \{t_{\text{CLS}}, t_1, t_2, \ldots t_N\}$, mGTE produces contextualized embeddings $\{e_{\text{CLS}}, e_1, e_2, \ldots e_i, \ldots, e_N\} \in R^{768}$, where e_i corresponds to the embedding vector of token t_i. The embedding associated with the special [CLS] token, e_{CLS}, was used as the linguistic representation from the feature network:

$$mGTE : e_{CLS} \rightarrow \bar{x}_L$$

Recent self-supervised speech models (e.g., Wav2vec 2.0[55], HuBERT[56]) have delivered state-of-the-art results on various speech tasks, but they remain underutilized in cognitive impairment detection for two key reasons: (1) Issues with long audio. These transformer models are typically trained on short audio segments, and their performance degrades when processing longer recordings than those seen during pre-training. Additionally, because they generate embeddings every 20 ms, even short recordings result in very long sequences. Therefore, fine-tuning speech transformer models on longer inputs (e.g., 30 s or more) becomes computationally intensive and memory-demanding. (2) Lack of global representation. These speech transformer models lack a dedicated global representation, such as the [CLS] embedding used in many linguistic transformer models that can summarize the entire sequence. This absence limits their ability to capture long-range temporal dependencies across the full recording, which is essential for detecting patterns associated with cognitive decline.

In SpeechCARE, we used mHuBERT[57] (with 98 million parameters, released in 2024) as the base model because of its extensive multilingual pretraining (90,000 h of speech in 147 languages), leading to robust, linguistically diverse phonetic representations, making it particularly well-suited for cognitive impairment detection in multilingual speech data. To address the challenges of fine-tuning acoustic transformer models including mHuBERT on longer audio, we introduce two key modifications: (1) a segmentation strategy that handles 30-s recordings in shorter segments, and (2) a learnable [CLS] embedding to capture global context.

Segmentation step. Let A denote the raw audio waveform, up to 30 s in length (shorter recordings were zero-padded to 30 s). We split A into 5 s segments with 25% overlap: $\{s_1, s_2, \ldots s_7\}$. Each segment s_i is processed by mHuBERT to produce 250 frame-level embeddings corresponding to a 25-ms window:

$$f_{mHubert}(s_i) = \{e_1^i, s_2^i, \ldots s_{250}^i\} \in R^{768}$$

The segment embeddings are concatenated into a unified sequence of 1750 vectors:

$$\{e_1^1, \ldots e_1^{250}, e_2^1, \ldots e_2^{250}, \ldots, e_7^1, \ldots e_7^{250}\} \rightarrow \{e_k\}_{k=1}^{1750}$$

This segmentation strategy was motivated by architectural constraints of mHuBERT. Similar to other speech transformer models (e.g., Wav2vec 2.0, HuBERT), mHuBERT applies convolutional positional embeddings to encode the position of each unit in the encoder input sequence. This design enables the model to extrapolate to audio sequences longer than those seen during pre-training. However, such extrapolation is inherently limited and may distort attention patterns and reduce precision. The segmentation strategy ensures that input audio remains within the temporal range of the pre-training audio data, enabling more effective and reliable fine-tuning. Additionally, since the complexity of the self-attention mechanism in mHuBERT grows quadratically with input length, shorter input segments reduce both memory usage and processing time.

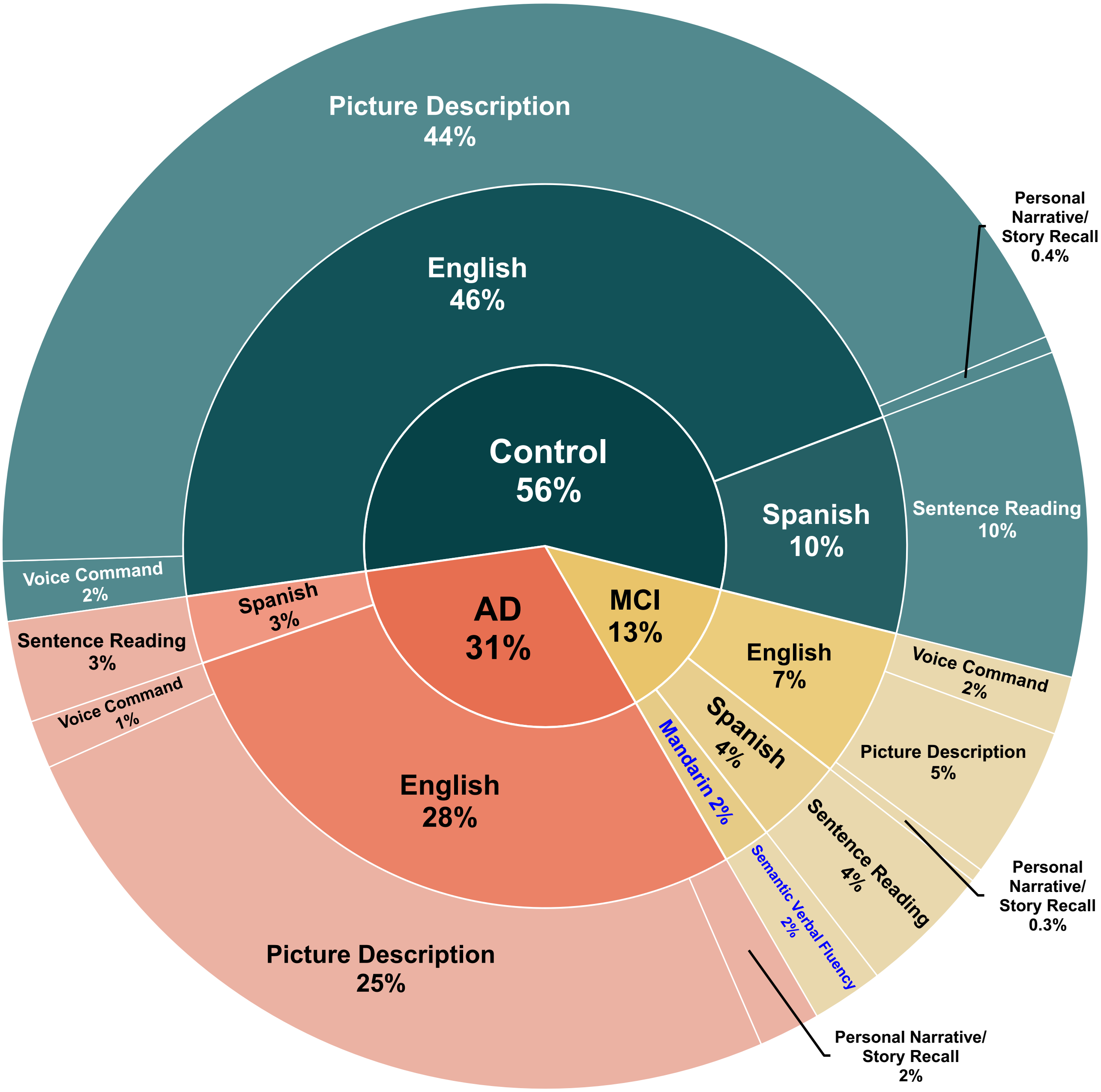

Fig. 6 | Dataset distribution by diagnosis (center), language (middle ring), and speech task (outer ring). The center circle shows the proportion of participants diagnosed as Control (56%), MCI (13%), or AD (31%). Each slice then expands into language groups—English, Spanish, and Mandarin—and further into specific speech tasks (e.g., picture description and sentence reading). Percentages at each ring reflect the relative share within that category, illustrating how participants are distributed across clinical status, language, and task types.

Incorporating CLS embedding. To summarize the full audio and capture dependencies across segments, we introduce a learnable [CLS] embedding into the mHuBERT output. Specifically, we prepend a randomly initialized embedding vector e_{CLS} to the concatenated sequence of segment embeddings:

$$[CLS] + \{e_k\}_{k=1}^{1750} \rightarrow \{e_{CLS}, e_1, e_2, \ldots, e_{1750}\}$$

This sequence was then passed through a Customized Self-Attention Encoder (CSE), consisting of two stacked blocks with four attention heads, dropout, residual connections, and normalization (see Fig. 7). In our preliminary analysis, this architecture performed well on the relatively small PREPARE challenge dataset (~10 h training data). During training, the [CLS] embedding evolves to encode a global summary of the input, capturing temporal dependencies across segments and local frames. The output [CLS] embedding e_{CLS} served as the acoustic representation from the feature network:

$$mHuBERT : e_{CLS} \rightarrow \bar{x}_A$$

To evaluate the contributions of segmentation and the [CLS] embedding, we conducted an ablation study to measure their impact on SpeechCARE performance (see *Results: Comparative and Ablation Analysis of SpeechCARE Components*).

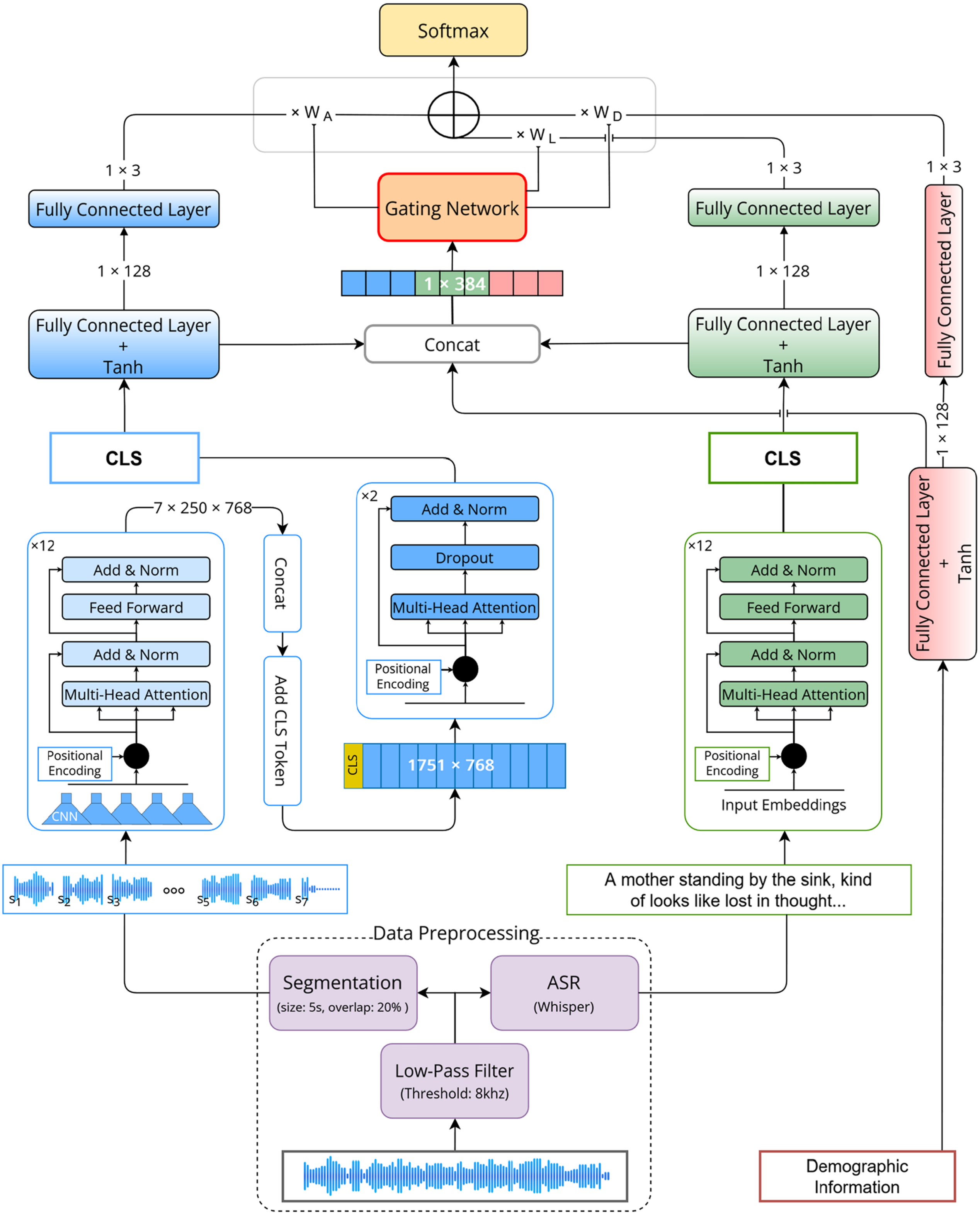

Fig. 7 | SpeechCARE architecture and data preprocessing pipeline. The diagram shows how SpeechCARE processes raw audio (up to 30 s) via low-pass filtering (8 kHz cutoff) and segmentation (5-s chunks, 25% overlap), followed by automatic speech recognition (ASR) to produce text. The acoustic pathway (blue boxes) uses a multilingual HuBERT encoder plus a Customized Self-Attention Encoder to extract a global [CLS] embedding. Meanwhile, the linguistic pathway (green boxes) processes the transcript using the multilingual GTE encoder. Demographic data (e.g., age, gender, education) enters as a third modality (pink). Each modality's embedding passes through its own fully connected layer, and the Gating Network (red box) adaptively fuses these representations before the final softmax prediction.

We treated age, gender, and education as categorical variables and one-hot encoded them to form $\bar{x}_D$. Preliminary experiments showed that categorizing age performed better than using it as a numeric value. We then explored these demographics both individually and in combination, treating them as a third modality within SpeechCARE.

Fusion network: adaptive gating fusion (AGF)

Fusion networks underpin multimodal systems, particularly in cognitive impairment studies integrating acoustic, linguistic, and other signals. Existing fusion methods range from intermediate fusion[58–60] (simple concatenation of feature vectors) to advanced cross-modal attention[36,61], yet they are typically not able to dynamically assess each modality's importance in final predictions. Such modality-level weighting is vital in speech datasets involving heterogeneous tasks (e.g., picture description vs. sentence reading), where distinct cognitive demands differentially affect acoustic and linguistic cues. Thus, selectively emphasizing each modality becomes critical for accurate cognitive impairment detection.

Inspired by the Mixture-of-Experts (MoE) paradigm, we developed a novel fusion approach—the AGF network—that dynamically weights modalities by adaptively attending to the most informative feature vectors (Fig. 7).

AGF operates on three feature vectors (see section "Feature network"): linguistic ($\bar{x}_L$), acoustic ($\bar{x}_A$); and demographic ($\bar{x}_D$), collectively denoted $\bar{x}_m$ for $m \in \{L, A, D\}$.

Each feature vector passes through modality-specific fully connected layers with a Tanh activation function to yield a hidden representation h_m:

$$h_m = tanh\left(W_m \bar{x}_m + b_m\right), \forall m \in \{L, A, D\}$$

These hidden representations are processed in three stages, which occur concurrently:

Stage One (Gating Network). To dynamically assign modality weights, hidden representations are concatenated and passed through a gating network (Fig. 7). The gating network employs a soft-attention mechanism, projecting the concatenated vector onto a three-dimensional space (corresponding to modalities) via a linear transformation and a Softmax function. This produces modality attention weights α_m for each $m \in \{L, A, D\}$:

$$\{\alpha_L, \alpha_A, \alpha_D\} = softmax\left(W_g[h_L; h_A; h_D] + b_g\right)$$

Stage two (modality-specific score). Concurrently with stage one, each modality's hidden representation separately passes through a fully connected layer to compute the modality-specific score o_m:

$$o_m = W_o h_m + B_o, \forall\, m \in \{L, A, D\}$$

Stage three (model's prediction). Finally, the model classification probabilities $\hat{y}$ are computed as the weighted sum of the modality-specific outcome o_m, using their corresponding attention weights α_m:

$$\hat{y} = Softmax\left(\sum_{m \in \{L, A, D\}} a_m o_m\right)$$

We evaluated the performance of AGF in comparison with commonly used fusion strategies, including intermediate fusion, scaled late fusion, and cross-modal attention[62]. For details, refer to the "Comparative and Ablation Analysis of SpeechCARE Components" subsection in the Methods and Results sections.

Model training and hyperparameter tuning

To train the SpeechCARE multimodal model, both the mHuBERT and mGTE transformer encoders were fine-tuned concurrently within a unified architecture (Fig. 7). To better control the optimization process, we adopted a multi–learning rate strategy: one learning rate was assigned exclusively to mGTE parameters, and a separate learning rate was used for the remaining components of the model, including mHuBERT, the Customized Self-Attention Encoder (CSE; see section "Acoustic Features"), the gating network, and fully connected layers. Preliminary experiments indicated that this strategy improved model performance compared to using a single learning rate for the entire model.

To achieve optimal model performance, we conducted a hyperparameter search that included both optimization-related parameters (e.g., learning rate, weight decay) and model-specific settings (e.g., number of hidden neurons, number of CSE attention heads). Each configuration was trained for 15 epochs, and model performance was evaluated on the validation set after each epoch. The checkpoint with the highest validation F1-score was selected for final evaluation on the test set.

Under the best hyperparameter setting, we used a learning rate of 1×10^{-6} for mGTE and 1×10^{-5} for the remaining components of the model, with a weight decay of 1×10^{-3} and a batch size of 4. The CSE module comprised two blocks with four attention heads and a dropout rate of 0.1. All fully connected layers used to encode input feature vectors employed 128 neurons with Tanh activation. The gating network included a fully connected layer with 384 neurons (corresponding to three modalities × 128 neurons each).

Model evaluation

We began our evaluation by analyzing model performance. Following training, we selected the model checkpoint that achieved the highest micro-averaged F1-score on the validation set and used this checkpoint to evaluate performance on the official Test set released by the PREPARE challenge organizers.

Due to the inherent randomness in deep learning training, we evaluated model stability by training SpeechCARE under 10 distinct random seeds. For each run, the official Test set remained fixed, while the random seed controlled the train–validation split, data loading order, and model weight initialization. This approach provided insight into the variability introduced by stochastic processes during training. We report the average evaluation metrics across the 10 runs on the official Test set, along with 95% confidence intervals, reflecting variation in model initialization and data sampling.

To assess model effectiveness, we employed several evaluation metrics. Evaluation metrics included the area under the receiver operating characteristic curve (AUC-ROC), computed using a one-vs-rest scheme across the three diagnostic classes (Control, MCI, and AD), with both micro-average and weighted-average scores. We additionally computed micro F1-score, as well as class-specific precision and recall capturing both overall and per-class performance. To further assess class separability, we computed precision–recall (PR) curves using the same one-vs-rest approach.

To further refine class-specific predictions, we applied a post hoc threshold adjustment to the model's predicted probabilities. Instead of using the default maximum probability rule, we independently calibrated decision thresholds for each diagnostic class (Control, MCI, AD). Thresholds were optimized on the validation set to maximize the class-specific F1-score, aiming to improve precision-recall balance, particularly for minority classes. This adjustment was performed after model training and did not affect the model parameters.

Fairness analysis

We also conducted a fairness analysis to examine potential bias across demographic and language subgroups. We assessed potential model bias across demographic (age, gender, and education) and language subgroups using two fairness metrics: EOO[63], which compares TPRs across groups, and Average Odds (AO)[64], which compares both TPRs and FPRs. These metrics were selected because they explicitly capture disparities in correctly identifying positive cases (true positives) and incorrectly classifying negative cases as positive (false positives), both critical in clinical settings. Other metrics, such as demographic parity and predictive equality, were not used as they do not distinctly measure disparities in these clinically important error types.

To reduce potential disparities, we applied three widely used bias mitigation strategies frequency-based reweighting[65], calibration[66], and threshold optimization[67], applied both individually and in combination. Frequency-based reweighting is an in-processing method that adjusts the importance of subgroups (e.g., age, education) by accounting for disparities in the joint distribution of sensitive attributes and class labels. This approach enhances the representation of underrepresented groups without modifying the original features or labels. Calibration and threshold optimization are post-processing techniques designed to improve fairness by refining the model outputs. For calibration, we trained group-wise isotonic regression models in a one-vs-rest setup, allowing predicted probabilities to better align with empirical likelihoods and improving subgroup-level calibration. Threshold optimization, in contrast, adjusts decision boundaries using group- and class-specific TPR and FPR thresholds to reduce disparities in fairness metrics.

Modality weights analysis

To understand how the model integrates acoustic, linguistic, and demographic inputs, we analyzed the AGF weights assigned to each modality for different speech tasks. Specifically, we grouped the test set samples by task type and computed the average gating weight for each modality. By comparing these averages across tasks, we could see which tasks relied more on acoustic cues (e.g., articulation or prosody), linguistic content (e.g., syntax or semantics), or demographic factors (e.g., age or education). This analysis provides insight into the model's adaptive strategy for weighting each modality based on the cognitive demands of the task.

External generalizability evaluation- ADReSSo challenge 2021

To evaluate SpeechCARE's generalizability beyond the multilingual, three-class PREPARE setting, we conducted external validation using the Pitt corpus[68] from DementiaBank[69], the dataset used in the ADReSSo challenge 2021. This corpus differs from PREPARE in three key ways: (i) it features a single speech task (Cookie-Theft picture description), (ii) all recordings are in English, and (iii) the diagnostic labels are binary: cognitively impaired (AD and MCI) vs. control (cognitively normal).

The ADReSSo dataset comprises 237 participants—122 with cognitive impairment (ADRD or MCI) and 115 control. The training ($n = 166$; 87 cognitive impairment, 79 control) and test ($n = 71$; 35 cognitive impairment, 36 control) splits were predefined by the ADReSSo challenge organizers. We stratified 20% of the training set for validation to support early stopping and hyperparameter tuning.

Since the ADReSSo dataset is about ten times smaller than the PREPARE dataset and contains only English speakers, we replaced mGTE with a more lightweight linguistic transformer model, DistilBERT[70], which outperformed mGTE on the validation set. The acoustic transformer remained mHuBERT with a learnable [CLS] embedding and 5-s/20% overlap segmentation. Demographic features were excluded due to missing data. The model was fine-tuned with multiple learning rates (2×10^{-6} for DistilBERT and 1×10^{-5} for other parameters) and early stopping based on F1-score on the validation set.

Given the binary classification task, we report the AUC and F1-score for the cognitive impairment class. Results were averaged over 10 runs with different random seeds and reported with 95% confidence intervals, reflecting variability due to randomness in model initialization and data sampling.

To ensure a fair comparison with prior studies on the ADReSSo challenge, we fine-tuned SpeechCARE exclusively on the ADReSSo training set without transfer learning, following the official challenge protocol.

External generalizability evaluation- DementiaBank Chou Corpus (Mandarin)

We further evaluated SpeechCARE on the Chou[71] corpus from DementiaBank. This corpus includes: (i) three picture description tasks—a father taking care of his baby, a night market, and park activities, (ii) recordings in Mandarin Chinese, and (iii) binary labels: MCI vs. control. The dataset includes 87 participants (47 MCI, 40 controls), with no information on age or education.

We performed a participant-level split, assigning ~60% of speakers to training ($n = 51$), ~20% to validation ($n = 17$) and the remaining ~20% ($n = 19$) to a held-out test set, ensuring all recordings from an individual appear in only one partition.

To adapt SpeechCARE for the Chou Corpus, we used two approaches. (1) *Zero-shot inference*. We first ran the SpeechCARE checkpoint fine-tuned on the PREPARE dataset directly on the entire Chou Corpus. Because all Mandarin recordings in the PREPARE dataset belonged to the MCI class, the model labeled every Chou sample as MCI, illustrating a language–label confound in the source data. (2) *Transfer learning*. SpeechCARE (without demographic modality) was first fine-tuned on the PREPARE dataset, then further fine-tuned on the Chou training set using the same hyperparameters. The F1-score on the Chou validation set was used for early stopping. AUC and F1-score were computed on the held-out Chou test set, with confidence intervals obtained from 10 runs with different random seeds.

Comparative and ablation analysis of SpeechCARE components

To quantify the contribution of each modality, we compared the multimodal SpeechCARE model—which integrates acoustic, linguistic, and demographic modalities via the AGF network—against unimodal baselines. Each unimodal model was fine-tuned separately on the training set, with hyperparameters optimized based on performance measured using the validation set. Specifically, we tested:

1. Acoustic-only model. We implemented three versions of the acoustic model to isolate the effect of the modifications we applied to the speech transformer:

a. First, as a baseline model, we fine-tuned mHuBERT by averaging its embedding vectors (each representing 25 ms of audio) and passing the resulting vector into an MLP with one fully connected layer (128 neurons, Tanh activation), omitting both segmentation and the learnable [CLS] embedding. The learning rate was set to 1×10^{-5}.
b. Second, we fine-tuned mHuBERT by incorporating the Customized Self-Attention Encoder (CSE, with two blocks and four attention heads and a dropout rate of 0.1) module and a learnable [CLS] embedding.
c. Third, we split the input audio into 5-s segments with 20% overlap, processed segments with mHuBERT, concatenated the resulting embeddings, prepended a [CLS] embedding, and processed the sequence using the CSE and finally passed the processed [CLS] to an MLP—replicating the full configuration used in SpeechCARE for generating acoustic features.

1. Linguistic-only model. We fine-tuned mGTE by passing its [CLS] embedding to an MLP (128 neurons, Tanh activation), using a learning rate of 1×10^{-6}.
2. Demographic-only model. We one-hot encoded demographic features (age, gender, and education), fed them into an MLP, and trained it with a learning rate of 1×10^{-3}.

To assess the added value of each modality in a multimodal configuration, we conducted the following analyses:

1. We combined acoustic and linguistic modalities using the AGF network to evaluate the benefit of integrating voice and transcription data.
2. To further evaluate the contribution of demographic features, we separately integrated age, gender, and education into the acoustic–linguistic model, and then combined all three, using the same network architecture and training setup.

To illustrate the effectiveness of our AGF network, we also compared its performance with the widely used fusion strategies:

1. Intermediate fusion. This is one of the most common fusion strategies, where feature vectors from each modality are concatenated into a single representation that is then used for classification.

2. Scaled late fusion. It employs learnable scalar weights for each modality. Prediction scores are computed independently for each modality and subsequently combined via a weighted sum, where the weights remain fixed during inference.
3. Cross-modal attention. This is a more complex, two-stage fusion strategy. In the first stage, each modality's representation is enriched via multi-head cross-attention, where embeddings from one modality attend to those of another; no fusion occurs at this stage. In the second stage, the enriched embeddings are concatenated to form a representation for classification. The complexity of the multi-head attention mechanism demands large datasets for effective fine-tuning, limiting its utility for small datasets.

Finally, to evaluate low-pass filtering as an effective noise reduction method, we tested its impact on SpeechCARE's performance against two alternatives:

1. Raw audio files without any noise reduction were used for both transcription and model training.
2. Audio files were processed using CMGAN[72] (Conformer-based Metric-GAN), a neural speech enhancement model that operates in the time-frequency domain and ranks among the top performers in speech enhancement benchmarks, and then used for both transcription and model training.

To evaluate the statistical significance of differences in model performance across ablation analyses, we conducted paired *t*-tests between the performance of individual SpeechCARE components and their corresponding baseline or comparative variants. Effect sizes were reported as Cohen's d, calculated as the mean of paired differences relative to their standard deviation.

Data availability

The data used in this research is from the 2024 NIA PREPARE challenge, provided exclusively to our team as participants in the challenge. This dataset is not publicly available, and participants are prohibited from sharing it. To request access, please contact the challenge organizers [[here](https://www.drivendata.org/competitions/group/nih-nia-alzheimers-adrd-competition)]. Additionally, the data is now part of DementiaBank; access may also be requested by contacting the DementiaBank administrators directly.

Code availability

The code used in this study is available at: https://github.com/SpeechCARE/SpeechCARE-NIA-Phase2.

Received: 30 May 2025; Accepted: 22 September 2025;
Published online: 17 November 2025

References

1. Busse, A., Hensel, A., Gühne, U., Angermeyer, M. C. & Riedel-Heller, S. G. Mild cognitive impairment: long-term course of four clinical subtypes. *Neurology* **67**, 2176–2185 (2006).
2. Ganguli, M. et al. Prevalence of mild cognitive impairment by multiple classifications: the Monongahela-Youghiogheny Healthy Aging Team (MYHAT) project. *Am. J. Geriatr. Psychiatry* **18**, 674–683 (2010).
3. Gagnon-Roy, M. et al. Preventing emergency department (ED) visits and hospitalisations of older adults with cognitive impairment compared with the general senior population: what do we know about avoidable incidents? Results from a scoping review. *BMJ Open* **8** (2018).
4. Boise, L., Neal, M. B. & Kaye, J. Dementia assessment in primary care: results from a study in three managed care systems. *J. Gerontol. A Biol. Sci. Med. Sci.* **59**, M621–M626 (2004).
5. Tóth, L. et al. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Curr. Alzheimer Res.* **15**, 130–138 (2018).
6. National Institute on Aging. *Assessing Cognitive Impairment in Older Patients* https://www.nia.nih.gov/health/assessing-cognitive-impairment-older-patients (2023).
7. Assessing Cognitive Impairment in Older Patients | National Institute on Aging. https://www.nia.nih.gov/health/health-care-professionals-information/assessing-cognitive-impairment-older-patients (2023).
8. Meilán, J. J. G., Martínez-Sánchez, F., Martínez-Nicolás, I., Llorente, T. E. & Carro, J. Changes in the rhythm of speech difference between people with nondegenerative mild cognitive impairment and with preclinical dementia. *Behav. Neurol.* **2020** (2020).
9. Petersen, R. C. et al. Practice guideline update summary: Mild cognitive impairment [RETIRED]: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology. *Neurology* **90**, 126–135 (2018).
10. Eyben, F. et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. In *Proc. IEEE Transactions on Affective Computing* (eds. Epps, J., Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y.) (IEEE, 2015)•ieeexplore.ieee.org. (2015).
11. models/research/audioset/yamnet at master · tensorflow/models. https://github.com/tensorflow/models/tree/master/research/audioset/yamnet.
12. Hershey, S. et al. CNN architectures for large-scale audio classification. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 131–135 https://doi.org/10.1109/ICASSP.2017.7952132 (2016).
13. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, 5999–6009 (2017).
14. Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R. & Poria, S. A review of deep learning techniques for speech processing. *Inf. Fusion* **99**, 101869 (2023).
15. Wolf, T. et al. Transformers: state-of-the-art natural language processing. In *EMNLP 2020 - Conference on Empirical Methods in Natural Language Processing, Proceedings of Systems Demonstrations* 38–45 https://doi.org/10.18653/V1/2020.EMNLP-DEMOS.6 (2020).
16. PREPARE: Pioneering Research for Early Prediction of Alzheimer's and Related Dementias EUREKA Challenge. https://www.drivendata.org/competitions/group/nih-nia-alzheimers-adrd-competition/ (2024).
17. Cai, W. et al. A survey on mixture of experts in large language models. *IEEE Trans. Knowl. Data Eng*. (2025).
18. Meet the winners of Phase 2 of the PREPARE Challenge - DrivenData Labs. https://drivendata.co/blog/prepare-phase2-winners (2025).
19. Competition: PREPARE Challenge - Phase 2: Model Arena (Acoustic Track). https://www.drivendata.org/competitions/299/competition-nih-alzheimers-acoustic-2/leaderboard/ (2024).
20. DementiaBank. https://talkbank.org/dementia/ADReSSo-2021/index.html (2021).
21. Han, W., Chan, C.-F., Choy, C.-S. & Pun, K.-P. An efficient MFCC extraction method in speech recognition. In *Proc. 2006 IEEE International Symposium on Circuits and Systems* 4 (IEEE, 2006).
22. Tausczik, Y. R. & Pennebaker, J. W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *J. Lang.Soc. Psychol.* 29 (2010).
23. Pan, Y. et al. Using the outputs of different automatic speech recognition paradigms for acoustic- and BERT-based Alzheimer's dementia detection through spontaneous speech. *Proc. Interspeech* 3810–3814 (2021).
24. Syed, Z. S., Syed, M. S. S., Lech, M. & Pirogova, E. Tackling the ADRESSO Challenge 2021: the MUET-RMIT system for Alzheimer's dementia recognition from spontaneous speech. In *Proc. Annual Conference of the International Speech Communication Association INTERSPEECH* **6**, 3815–3819 (2021).

25. Liu, Y. et al. *RoBERTa: A Robustly Optimized Bert Pretraining Approach* (2019). arXiv preprint arXiv:1907.11692 (2019).
26. Qiao, Y., Yin, X., Wiechmann, D., & Kerz, E. Alzheimer's disease detection from spontaneous speech through combining linguistic complexity and (Dis)fluency features with pretrained language models. In *Proc. Annual Conference of the International Speech Communication Association INTERSPEECH* **6**, 4226–4230 (2021).
27. Sun, Y. et al. ERNIE 2.0: A continual pre-training framework for language understanding. *Proc. AAAI Conf. Artif. Intell.* **34**, 8968–8975 (2020).
28. Wolpert, D. H. Stacked generalization. *Neural Netw.* **5**, 241–259 (1992).
29. Ilias, L. & Askounis, D. 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 (2023).
30. Touvron, H. et al. Training data-efficient image transformers & distillation through attention. In *Proc. 38th International Conference on Machine Learning* (eds. Meila, M. & Zhang, T.) **139**, 10347–10357 (PMLR, 2021).
31. Chen, C., Han, D. & Chang, C.-C. CAAN: context-aware attention network for visual question answering. *Pattern Recognit.* **132**, 108980 (2022).
32. Courty, N., Flamary, R., Tuia, D. & Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1853–1865 (2017).
33. Müller, R., Kornblith, S. & Hinton, G. E. When does label smoothing help? Advances in neural information processing systems, 32 (2019).
34. Li, P. & Li, X. Multimodal fusion with co-attention mechanism. In *Proc. 2020 23rd International Conference on Information Fusion, FUSION 2020* https://doi.org/10.23919/FUSION45008.2020.9190483= (2020).
35. Bang, J., Han, S. & Kang, B. Alzheimer's disease recognition from spontaneous speech using large language models. *ETRI J.* **46**, 96–105 (2024).
36. Shao, H., Pan, Y., Wang, Y. & Zhang, Y. Modality fusion using auxiliary tasks for dementia detection. *Comput. Speech Lang.* **95**, 101814 (2026).
37. Kramarow, E. A. Diagnosed dementia in adults age 65 and older: United States, 2022 (2022).
38. Petersen, R. C. et al. Practice guideline update summary: Mild cognitive impairment: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology. *Neurology* **90**, 126 (2018).
39. Lutz, W. & KC, S. Global human capital: Integrating education and population. *Science* **333**, 587–592 (2011).
40. Darwish, H., Farran, N., Assaad, S. & Chaaya, M. Cognitive reserve factors in a developing country: education and occupational attainment lower the risk of dementia in a sample of lebanese older adults. *Front. Aging Neurosci.* **10**, 277 (2018).
41. Van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).
42. Johnson, D. H. Signal-to-noise ratio. *Scholarpedia* **1**, 2088 (2006).
43. Madhu, N. Note on measures for spectral flatness. *Electronics letters.* **45**, 1195–1196 (2009).
44. Streijl, R. C., Winkler, S. & Hands, D. S. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimed. Syst.* **22**, 213–227 (2016).
45. Remove High-Frequency Noise in Measured Data - MATLAB & Simulink. https://www.mathworks.com/help/thingspeak/remove-high-frequency-noise-in-measured-data.html (2025).
46. Introducing Whisper | OpenAI. https://openai.com/index/whisper/ (2022).
47. Ali, A. & Renals, S. Word error rate estimation for speech recognition: e-WER. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* 20–24 (2018).
48. Dubey, A. et al. The Llama 3 herd of models. Preprint at http://arxiv.org/abs/2407.21783 (2024).
49. Pornprasit, C. & Tantithamthavorn, C. Fine-tuning and prompt engineering for large language models-based code review automation. *Inf. Softw. Technol.* **175**, 107523 (2024).
50. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings Conference* **1**, 4171–4186 (2018).
51. Pappagari, R., Cho, J., Moro-Velazquez, L. & Dehak, N. Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity. In *Proc. Annual Conference of the International Speech Communication Association, INTERSPEECH* **2020**, 2177–2181 (2020).
52. Zhu, Y., Liang, X., Batsis, J. A. & Roth, R. M. Exploring deep transfer learning techniques for Alzheimer's dementia detection. *Front. Comput. Sci.* **3**, 624683 (2021).
53. Koo, J., Lee, J. H., Pyo, J., Jo, Y. & Lee, K. Exploiting multi-modal features from pre-trained networks for Alzheimer's dementia recognition. In *Proc. Annual Conference of the International Speech Communication Association, INTERSPEECH* **2020**, 2217–2221 (2020).
54. Zhang, X. et al. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. Preprint at https://arxiv.org/abs/2407.19669 (2024).
55. Schneider, S., Baevski, A., Collobert, R. & Auli, M. wav2vec: Unsupervised pre-training for speech recognition. Preprint at https://arxiv.org/abs/1904.05862 (2019).
56. Hsu, W.-N. et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio Speech Lang. Process* **29**, 3451–3460 (2021).
57. Boito, M. Z., Iyer, V., Lagos, N., Besacier, L. & Calapodescu, I. mhubert-147: A compact multilingual hubert model. Preprint at https://arxiv.org/abs/2406.06371 (2024).
58. Ilias, L. & Askounis, D. Multimodal deep learning models for detecting dementia from speech and transcripts. *Front. Aging Neurosci.* **14**, 830943 (2022).
59. Yamada, Y. et al. Combining multimodal behavioral data of gait, speech, and drawing for classification of Alzheimer's disease and mild cognitive impairment. *J. Alzheimer's Dis.* **84**, 315–327 (2021).
60. Roesler, O. et al. Towards scalable remote assessment of mild cognitive impairment via multimodal dialog. In *Proceedings of Interspeech* (2024).
61. Poor, F. F., Dodge, H. H. & Mahoor, M. H. A multimodal cross-transformer-based model to predict mild cognitive impairment using speech, language and vision. *Comput. Biol. Med.* **182**, 109199 (2024).
62. Jiao, T., Guo, C., Feng, X., Chen, Y. & Song, J. A comprehensive survey on deep learning multi-modal fusion: methods, technologies and applications. *Comput. Mater. Contin.* **80**, 1–35 (2024).
63. Roemer, J. E. & Trannoy, A. Equality of opportunity. in *Handbook of Income Distribution* **2** 217–300 (Elsevier, 2015).
64. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.* **169**, 866–872 (2018).
65. Chai, J. & Wang, X. Fairness with Adaptive Weights. In *Proc. 39th International Conference on* Machine Learning (eds. Chaudhuri, K. et al.) **162** 2853–2866 (PMLR, 2022).
66. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. & Weinberger, K. Q. On Fairness and calibration. in *Advances in Neural Information Processing Systems* (eds Guyon, I. et al.) 30 (Curran Associates, Inc., 2017).
67. Yang, J., Soltan, A. A. S., Eyre, D. W., Yang, Y. & Clifton, D. A. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digit. Med.* **6**, 55 (2023).

68. DementiaBank English Pitt Corpus. https://talkbank.org/dementia/access/English/Pitt.html (2025).
69. DementiaBank. https://talkbank.org/dementia/ (2025).
70. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Preprint at http://arxiv.org/abs/1910.01108 (2019).
71. DementiaBank Mandarin Chou Corpus. https://talkbank.org/dementia/access/Mandarin/Chou.html (2025).
72. Abdulatif, S., Cao, R. & Yang, B. CMGAN: conformer-based metric-GAN for monaural speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **32**, 2477–2493 (2024).

Acknowledgements

The authors have no acknowledgements to declare.

Author contributions

Hossein Azadmaleki: Led model development, evaluation, and contributed to manuscript drafting. Yasaman Haghbin: Contributed to model development and evaluation. Sina Rashidi: Contributed to model development. Mohammad Javad Momeni Nezhad: Contributed to model development. Ali Zolnouri: Contributed to model evaluation. Maryam Zolnoori: Conceived the study, supervised all aspects of model development and evaluation, and led manuscript drafting and revision.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Maryam Zolnoori.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.