Alzheimer's & Dementia: The Journal of the Alzheimer's Association

An Automatic and Speech-based Cross-Lingual Classification Framework for Early Screening of Cognitive Impairment

--Manuscript Draft--

Manuscript Number:	ADJ-D-24-02241
Article Type:	Research Article
Keywords:	Alzheimer's disease; cognitive impairment; Large language model; NLP; machine learning
Corresponding Author:	Xiang Fan Peking University Shenzhen Hospital Shenzhen, CHINA
First Author:	Yue Wu
Order of Authors:	Yue Wu
	Yining Liao, Msc
	Keyan Yu, Msc
	Lele Chen, Bsc
	Zhuonan Wei, Bsc
	Lin Hu, Bsc
	Gaigai Lu, Bsc
	Hui Chen, Bsc
	Guanxu Cheng, MD
	Kai Wang, PhD
	Xiang Fan, PhD
Abstract:	INTRODUCTION The use of speech data for distinguishing cognitive impairment (CI) is efficient and convenient for early screening of potential AD. However, few studies have developed available automated frameworks with external cross-lingual Chinese validation. METHODS This study utilized speech data from the Cookie Theft description task, employing the ADReSSo dataset and the local Chinese dataset of the STAR cohort. We constructed an automated framework for CI screening, leveraging AI methods, including ASR, LLMs, and multiple types of machine learning classifiers. We used datasets in multiple languages and addressed the issue of language inconsistency. RESULTS Our framework achieved 74% in accuracy and 75% in AUC in the external cross-lingual Chinese validation experiment. We conducted an ablation study to demonstrate the necessity of each module within the framework. DISCUSSION The proposed framework provides fully automated assessments in distinguishing CI, making it highly beneficial for large-scale early screening and self-testing.

Dear Editor, 31 October, 2024

I am writing to submit a research article entitled " An Automatic and Speech-based Cross-Lingual Classification Framework for Early Screening of Cognitive Impairment " to the special issue named *Spotlight on Alzheimer's disease and related dementias research in East Asia* for your kind consideration for publication in *Alzheimer's & Dementia*.

In this paper, we construct a novel framework that leverages several AI methods for automatically screening cognitive impairment (CI) based on the Cookie Theft picture description task with a multilingual dataset. It holds a high potential for clinical application in early AD detection as it's fully automatic and has achieved high performance with 74% in accuracy and 75% in AUC in the external cross-lingual Chinese validation experiment, excels in distinguishing CI, and is beneficial for large-scale screening and self-testing of CI, which will remind potential AD patients to undergo timely hospital-based examinations and therapies.

We believe this manuscript is appropriate for publication by *Alzheimer's Disease and Related Dementias Research in East Asia* because it is highly relevant to the issue, highlighting the potential of non-invasive early screening through cross-lingual validation based on a 1-minute speech task.

We confirm that this work is original and has not been published elsewhere, nor is it currently under consideration for publication elsewhere.

Thank you for considering our manuscript for publication in this special issue. We look forward to the opportunity to share our research with your readers and contribute to the global dialogue on dementia research.

Sincerely,

Dr Xiang Fan
Peking University Shenzhen Hospital
Shenzhen, China
fiona@link.cuhk.edu.hk

An Automatic and Speech-based Cross-Lingual Classification Framework for Early Screening of Cognitive Impairment

Yue Wu[1,2#], Yining Liao[3#], Keyan Yu[1], Lele Chen[1], Zhuonan Wei[1], Lin Hu[1], Gaigai Lu[1], Hui Chen[1], Guanxu Cheng[1]*, Kai Wang[4]*, Xiang Fan[1]*

[1](Department of Medical Imaging, Peking University Shenzhen Hospital, Shenzhen, Guangdong, 518172, P.R. China)

[2](Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, Guangdong, 518055, P.R. China)

[3](School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Guangdong, 518172, P.R. China)

[4](School of Computer Science and Technology, Dongguan University of Technology, Dongguan, Guangdong, 523808, P.R. China)

Corresponding author:

chengguanxun@outlook.com,

kevin.w@dgut.edu.cn,

fiona@link.cuhk.edu.hk

[#]: Contribution equally

*: Corresponding author

Abstract

INTRODUCTION

The use of speech data for distinguishing cognitive impairment (CI) is efficient and convenient for early screening of potential AD. However, few studies have developed available automated frameworks with the external cross-lingual Chinese validation.

METHODS

This study utilized speech data from the Cookie Theft description task, employing the ADReSSo dataset and the local Chinese dataset of the STAR cohort. We constructed an automated framework for CI screening, leveraging AI methods, including ASR, LLMs, and multiple types of machine learning classifiers. We used datasets in multiple languages and addressed the issue of language inconsistency.

RESULTS

Our framework achieved 74% in accuracy and 75% in AUC in the external cross-lingual Chinese validation experiment. We conducted an ablation study to demonstrate the necessity of each module within the framework.

DISCUSSION

The proposed framework provides fully automated assessments in distinguishing CI, making it highly beneficial for large-scale early screening and self-testing.

Background

Alzheimer's disease (AD) is the most common type of dementia in aging people. According to the World Health Organization, the number of patients with dementia will rise from 55 million in 2019 to 139 million in 2050, increasing rapidly with the global population age[1,2]. The typical characteristics of AD include the progressive degradation of memory, cognition, and motor skills, as well as the decline of speech, language, and logistics[3]. Though there is no effective cure for AD currently, clinical research has figured out that early detection could delay disease progression and provide preventive care[4]. Therefore, evaluating cognitive ability and detecting cognitive impairment (CI) on a widespread scale is crucial to help early screening.

Traditional AD diagnostic methods concentrate on various biomarkers from cognitive assessments, cerebrospinal fluid (CSF), and neuroimaging techniques (e.g., magnetic resonance imaging (MRI) and positron emission tomography(PET))[5–8]. These methods are often expensive, time-consuming, invasive, and require specialized equipment, making them impractical for widespread use[9]. It is neither feasible nor affordable to use each of the above examinations for large-scale screenings. An easy, quick, automatic, and user-friendly framework is necessary for screening and selecting potential AD individuals to undergo timely hospital-based examinations and therapies.

Speech reflects cognitive functions, including attention, memory[10], idea formation, and the translation of thoughts into coherent articulation[11]. It shows significant potential for analyzing and understanding cognitive processes in the early stages[12]. In various language tasks, patients with AD demonstrate distinct performance patterns compared to healthy individuals[13], notably

marked by a reduction in discourse complexity and connected speech—key symptoms of AD [14–16]. Language impairment often emerges in the early stages of the disease[17], making it a promising indicator for early diagnosis. Consequently, utilizing audio analysis to detect AD is particularly valuable, with the advantage of being non-intrusive, cost-effective, and more scalable. Speech-based automatic AD diagnosis has risen and developed quickly recently, represented by ADReSSo challenges[18,19]. Many researchers have extracted various acoustic and linguistic features [20,21] and built speech-based AD detection systems using machine learning algorithms from spontaneous speech [22–32]. Previous studies have achieved available performance in detection frameworks based on vocal features [22,23], lexical features [24] , and vocal-lexical fusion features[25]. Kong et al. [22] developed an automatic detection pipeline based on vocal features and achieved an accuracy of 80%. Fraser et al.[23] developed a binary classification model with 35 top-ranked features and achieved 82% average accuracy. Another approach used a speech recognition and translation model and achieved 84.41% accuracy in the ADReSSo test set [24]. Linguistic and acoustic features were combined to predict cognitive impairment and achieved an AUC of 94%[25], and are combined to build an explainable classification model[26].

As Natural Language Preprocess (NLP) has shown reliable performance in health care[33], and Automatic Speech Recognition (ASR) has strong performance in transcription, several researchers used language models combined with ASR to extract linguistic features automatically. In a previous study, ASR and fine-tuned embedding models achieved high accuracy in the ADReSS20 challenge[27,28]. Many researchers also leveraged ASR, language model for embedding, and different classifiers to build the whole framework[29–32]. However, most of these studies were

limited by using relatively small sample sizes of clinical populations, using single-language datasets, and often faced data scarcity issues[27]. Some studies relied on expensive handcrafted features (including manual transcription)[23], which hinder its implication in a large-scale setting. Besides, very few researchers included mild cognitive impairment (MCI) cases in their studies[15,34], which restricts the application in early screening.

In this study, we constructed an automatic framework for early screening CI based on voice recording without manual feature extraction. We utilized datasets in multiple languages to address the data deficiency problem. This study aims to develop a robust, automatic early screening framework of CI that is easy to deploy, requires no specialized equipment, and is suitable for widespread use. It focuses on detecting CI and facilitating large-scale early screening. The proposed framework uses machine learning techniques to take Cookie Theft picture description task speech as input and predicts the likelihood of CI. We leveraged ASR for transcription, LLM for translation and embedding, and different classical machine learning classifiers and deep learning neural networks for classification. We rigorously evaluated our approach on public and local datasets and achieved promising performance. The results highlighted the challenges and opportunities in building a cross-linguistic diagnostic framework for large-scale early screening and self-testing CI.

Methods

2.1 Datasets

This study utilized two datasets for CI screening: the ADReSSo[19] Challenge dataset and the dataset from the Shenzhen Multimodal Aging Research (STAR) cohort. The ADReSSo dataset is

extracted from the Cookie Theft picture description task of the Pitt Corpus in the DementiaBank database [35]. It ensures balanced distributions across age, sex, and diagnosis. Each entry in the dataset includes a Cookie Theft storytelling speech and an associated binary label. The training set comprises 166 participants, of which 87 are diagnosed with AD and 79 are non-AD. The local participants in the STAR cohort were recruited from memory clinics and poster advertisements at Peking University Shenzhen Hospital, China. Participants in the STAR cohort included those diagnosed with CI and cognitively unimpaired (CU) as controls. The local voice dataset mirrors the same structure with speech recordings of the Cookie Theft picture description task, binary clinical labels (i.e., CI and CU), demographic information such as age and sex, neuropsychology scale assessments, and so on. All procedures in this research were conducted following the Declaration of Helsinki and were approved by the Ethics Committee of Peking University Hospital (No.2022-160-01). Written informed consent was obtained from each participant before their inclusion in the study. Additionally, this study was registered with the Chinese Clinical Trial Registry (ChiCTR; ChiCTR2200066700). **Table 1** presents the participant's characteristics, including self-reported sex, age statistics, education status, MMSE, and MoCA.

2.2 Framework Construction

The proposed CI screening framework was based on the audio recordings from the Cookie Theft picture description task. An ASR system was incorporated to transcribe all speech recordings to text in data preparation. Then, we translated the text into Chinese based on the LLM GLM-4 model. After that, we used a large language embedding model to extract linguistic features from the text for embedding generation. At last, the discriminative embedding features were fed into multiple

classification components. The pipeline of our proposed CI screening framework is shown in **Figure 1.**

2.2.1 ASR For Automatic Transcription

To enable automatic transcription and enhance the framework's applicability, we adopted the Whisper model[36] for speech recognition. Whisper is a transformer-based multilingual speech recognition and translation model trained on 680,000 hours of supervised audio data, enabling robust performance in diverse languages like English and Chinese. We selected the medium-sized model (769M parameters) to balance accuracy and performance. The Whisper model was used to transcribe the storytelling speech into text automatically as the foundation for subsequent processing steps.

2.2.2 LLM for Translation

To address the cross-lingual challenges across datasets, we translated the ADReSSo transcriptions into Chinese using GLM-4 [37], a transformer model pre-trained on ten trillion tokens in both Chinese and English. GLM-4 employs techniques such as Supervised Fine-Tuning (SFT) and Reinforcement Learning with Human Feedback (RLHF), which have achieved remarkable performance in the Chinese language alignment task. This capability ensured linguistic consistency between the ADReSSo and local Chinese datasets, facilitating effective feature extraction and robust classification. During the translation process, we set the top_p parameter to 0.1 and the temperature to 0.15.

2.2.3 LLM for Embedding Feature Extraction

To obtain discriminative text features automatically, we utilized the large language model, Embedding-3 from Zhipu AI[38], to convert entire sentences into dense vector representations (i.e., embeddings). The extracted text embeddings implied essential linguistic features, the meaning, and the semantics within the text. By encoding complex language structures and subtle semantic nuances, we found that the Embedding-3 model can be particularly useful for CI screening as early diagnosis hinges on a detailed understanding of linguistic patterns.

We mapped each speech segment to a 512-dimensional vector, encapsulating the linguistic attributes critical for screening CI. The high-dimensional embeddings enabled discriminative text feature extraction and thus ensured the comprehensive inputs for the following classifications.

2.2.4 Classifiers

We formulated the CI screening task as a binary classification problem, distinguishing between participants with AD (label 1) and those without it (label 0). To this end, we explored a range of machine learning classifiers, including:

- Logistic Regression (LR), a classical binary classification model;
- Random Forest (RF)[39] aggregates predictions from multiple decision trees with strong robustness and generalization. We configured the model with 100 trees, a minimum sample split of 2, a minimum sample leaf of 1, and no restriction on maximum depth;
- Support Vector Machine (SVM)[40] maximizes the margin between classes, making it particularly effective for small sample sizes. In this work, we performed SVM with the radial basis function kernel (i.e., Gaussian kernel) with regularization parameter as 1.0 and kernel

parameter $\gamma = {}^{1}/_{2\sigma^2}$, where σ indicates standard deviation.

- Extreme Gradient Boosting (XGBoost)[41] utilizes boosting techniques with decision trees, achieving both efficiency and high accuracy. The model parameters were set to 100 trees, a learning rate of 0.3, a maximum depth of 6, and a subsample ratio of 1.
- Multi-Layer Perceptron (MLP) uses embeddings feature extracted from LLM as an input layer, two hidden layers with 128 and 64 units using ReLU activation, and an output layer with a Sigmoid activation for probability scoring.
- Multi-Layer Perceptron with a Transformer encoder block (MLP-Trans) uses embeddings feature extracted from LLM as an input layer, a hidden layer with 128 units using ReLU activation, a Transformer encoder block with 8 attention heads, a second hidden layer with 64 units, and an output layer with a Sigmoid activation for probability scoring.

We conducted **two sets of experiments**:

1. **Local validation**, where both training and evaluation were performed on local data using cross-validation ($k = 5$). This experiment indicated the effectiveness of our approach in a single-language setting. To ensure robustness, we randomly split the data 100 times for each cross-validation with different random seeds and calculated the mean result as the final measure of the model performance.
2. **Cross-lingual validation**, where we trained the models on the ADReSSo (English) and evaluated performance on the local dataset (Chinese). This experiment aimed to assess the robustness of the models across different languages, which is crucial for broader applicability and coping with data scarcity.

We reported six metrics of the experimental results for quantitative measurement, including accuracy (ACC), precision (PRE), recall (REC), F1 score, the receiver operating characteristic (ROC) curve, and area under the curve (AUC) for the positive class (CI). This thorough approach enabled the comparison between classical machine learning methods and deep neural networks and thus revealed the model strengths and limitations of different methods based on monolingual and cross-lingual contexts.

2.3 Ablation Study

We investigated the necessity of translation and the performance of the LLM within the entire cross-lingual framework. We performed an ablation experiment to compare the performance of our framework with/without the translation procedure and illustrate the necessity of language translation for early AD screening. Moreover, we conducted an experiment by replacing the LLM Embedding-3 with a much smaller pre-trained embedding model to reveal the importance of incorporating a large language embedding model for discriminative AD-related text feature extraction on both local and cross-lingual data.

Results

3.1 Monolingual Experiment: Performance on the Local Dataset with K-fold Cross-Validation

Table 2 summarizes the performance of different classifiers on the local dataset under 5-fold cross-validation. Each reported metric represents the mean across 100 random splits to ensure robust

evaluation. We also tried to evaluate different classifiers with the train-test split in **Appendix A (Table A1)**. The first value in each cell indicates performance on the training set, while the second value corresponds to the test set. The distribution of accuracy and AUC for different classifiers are also shown in **Figure 2**.

The results suggested that RF and XGBoost achieved best performance in nearly all metrics, especially in precision, recall, and F1 scores, and show signs of overfitting by achieving perfect performance in training and a significant decline in testing. LR, SVM, and MLP demonstrated stable performance in accuracy during training and testing without suffering from overfitting or underfitting. However, the near-zero results for precision, recall, and F1 score across these three methods indicated the class imbalance issue and the challenge of learning comparative patterns from the dataset. Moreover, although MLP-Trans performed better in testing than training in all evaluation metrics, the performance is still far from ideal, with underfitting during training.

3.2 Cross-Lingual Experiment: Training on Public English Data and Testing on Local Chinese Data

We conducted experiments where models were trained on the ADReSSo public English dataset and tested on the local Chinese dataset with different classifiers. We visualized the embedding using t-SNE[42] and shown in **Appendix B (Figure B1)**. The whole outcomes are summarized in **Table 3**, and accuracy and AUC for different classifiers are shown in **Figure 2**.

From the results, we can see that RF consistently achieved the best overall performance across key

metrics: accuracy, precision, F1 score, and AUC. These results indicated that RF maintained a strong balance between identifying actual CI cases and minimizing false positives. While not excelling in precision, LR achieved the highest recall at 0.77, showing it was sensitive to CI cases. However, this higher sensitivity led to a drop in precision, resulting in a moderate F1 score of 0.52, reflecting the model's tendency to over-predict CI cases. In contrast, models such as SVM, XGBoost, and MLP demonstrated relatively consistent but slightly lower performances in accuracy, F1 score, and AUC compared to RF. These models offered similar trade-offs between recall and precision, though they did not match the same level of balance across the metrics as RF. Specifically, SVM and XGBoost performed well in the recall, but their lower precision led to a more modest overall F1 performance. The MLP models, including the transformer variant, exhibited comparable results but did not outperform RF across most metrics.

Overall, ensemble models that combine the predictions from multiple decision trees are capable of effectively reducing both bias and variance in individual models and enhancing robustness in capturing complex patterns. This approach enabled the identification of a broader range of data features, particularly in cases of class imbalance, where the diversity of different trees assists in recognizing potential patterns in minority class samples. Furthermore, the introduction of randomness in ensemble learning helps to improve generalization capabilities. For CI screening, the data is often scarce and imbalanced, while generated embeddings contain rich textual information. Thus, RF is particularly suitable for CI screening in our task.

3.3 Ablation Study

In this experiment, we compared the framework's performance with and without translation to determine the necessity of translation. The results demonstrated that when language consistency was not maintained—i.e., when the ADReSSo dataset remained in English—there was a marked degradation in performance. To explore the impact of different language models, we experimented with MiniLM[43], a much smaller and lightweight embedding model, for cross-lingual datasets. MiniLM has fewer parameters and a smaller architecture, making it computationally efficient, but lacks capacity for complex feature extraction.

The performance of our framework with RF classifier without the translation procedure and with MiniLM instead of Embedding-3 is shown in **Figure 3.** The detailed performance of each classifier without translation and with MiniLM for embedding is shown in **Table 4.** The first value in each cell indicates performance without the translation, while the second value corresponds to using MiniLM.

3.3.1 Comparison of Framework Performance with and without Translation

Across all six models, the results demonstrated that using the translated data led to a substantial improvement in performance. Specifically, the accuracy of models trained on the translated data was approximately twice that of models trained on the non-translated data, with precision increasing by about 1.5 times. It suggested that models trained on translated data were better at correctly identifying AD patients while reducing the number of false positives (non-AD cases incorrectly classified as AD). A decline in recall was observed when using the translated dataset, indicating that the translated models may miss some AD cases, leading to a higher number of false negatives. This

decline was likely due to discrepancies in linguistic structure and semantic features between the two languages, which affected the framework's ability to generalize across datasets. In contrast, translating the English text into Chinese ensured that both datasets were processed uniformly, improving feature extraction and enhancing classification accuracy. This comparison highlighted the importance of language alignment in cross-linguistic tasks, particularly in methods that rely on subtle linguistic features for the CI screening task.

3.3.2 Comparison of Framework Performance with MiniLM and Embedding-3

From the results, Embedding-3 has demonstrated its ability to improve precision, F1, and AUC across most classifiers, particularly with RF and XGBoost. The remarkable performance of Embedding-3 can be attributed to its superior ability to extract linguistic information and retain more intricate details from the input data. This level of detail is crucial when dealing with complex language structures as it allows the model to better understand nuances and contextual meanings that may be lost with less effective embeddings.

Although MiniLM achieved higher recall with several classifiers (e.g., LR), its increases in sensitivity came at the cost of precision and led to moderate F1 scores. In contrast, the superior AUC of Embedding-3 underscores its capacity to generalize more effectively across diverse datasets, especially for ensemble-based classifiers. This capability is significant, as ensemble methods like RF leverage the strengths of multiple models, combining their predictions to enhance overall performance.

The combination of the powerful ensemble classifiers (e.g., RF) and the detailed embeddings from

Embedding-3 results in a robust framework for capturing relevant patterns in translated texts. By effectively integrating detailed linguistic features with the decision-making capabilities of RF, the proposed framework can identify subtle patterns that contribute to improved classification performance. This combination not only addressed the challenges posed by cross-linguistic variations but also enhanced the model's ability to accurately classify instances, resulting in higher overall performance metrics. Hence, this method can be shown as a promising approach to tackling the complexities of language processing in different languages for the CI screening task.

Discussion

This study explored an innovative approach, using a cross-lingual dataset to automatically screen CI through a speech analysis framework by leveraging ASR and LLM for translation, the large language embedding model, and various classifiers. Due to its automatic ability and great performance in validation tests, the proposed framework is suitable for large-scale population-based screening and is available for self-testing of CI, which is beneficial for the eventual timely therapy of AD patients [44]. This framework could be widely implemented in resource-limited regions, significantly aiding populations needing accessible diagnostic tools and bringing substantial economic benefits. AI-based CI detection tools can be greatly cost-saving, and our work has broad prospects for future applications. This section discusses our approach's key findings, strengths, limitations, and implications, offering insight into future research directions.

4.1 Impact of Translation on Performance

Our experiments revealed that maintaining linguistic consistency between datasets significantly

improved classification performance. Specifically, translating the ADReSSo dataset from English to Chinese led to better feature extraction and classification accuracy. In contrast, the framework performance declined markedly when the ADReSSo dataset remained in English. Though it has been shown that multilingual sentence encoders could achieve great performance,[45]the differences in linguistic structure, syntax, and semantics between English and Chinese still exist and cause degradation, which hinders the framework's ability to generalize across datasets. These findings highlight the importance of aligning linguistic features in multilingual contexts, especially when models rely on subtle language nuances for binary classification tasks.

4.2 Impact of LLM Ability on Performance

We observed that Embedding-3, a sentence embedding model from Zhipu AI, consistently outperformed MiniLM in capturing the linguistic features necessary for CI screening. MiniLM, although computationally efficient with a smaller architecture, struggled to retain the complex semantic and contextual nuances required for early diagnosis. In contrast, Embedding-3's high-dimensional (512) vectors encapsulated essential linguistic information, contributing to improved classifier performance. This comparison emphasizes the need for powerful embeddings when detecting cognitive impairments like AD.

4.3 Classifier Performance and Framework Robustness

Among the classifiers tested, RF exhibited the most balanced performance across multiple metrics, achieving the highest AUC (0.74) and a reliable balance between precision (0.56) and F1 score (0.62). On the other hand, LR showed superior sensitivity, with a recall of 0.77, making it effective

for identifying CI cases, though at the cost of reduced precision. The neural network models, including the transformer-based architecture, demonstrated promising results but did not surpass the traditional machine learning models in this setting. These results suggest that ensemble learning techniques such as RF effectively capture patterns in linguistic data for CI screening.

4.4 Cross-Dataset Evaluation and Generalization

Training on the ADReSSo dataset and testing on the local dataset allowed us to evaluate the generalization capability of our framework. The cross-dataset evaluation underscored the challenges associated with dataset heterogeneity, especially when demographic factors and linguistic features differ. Despite these challenges, our framework maintained reasonable performance, affirming that combining powerful embeddings and robust classifiers can enhance generalization.

4.5 Limitations and Future Directions

While our approach shows promise, several limitations should be noted. First, although translating the ADReSSo dataset into Chinese improved performance, translation may introduce minor semantic shifts, potentially impacting subtle linguistic patterns. Second, the limited availability of AD patient data restricts the development of more sophisticated models, leaving significant room for performance improvements. Expanding the dataset with more AD samples and diverse linguistic inputs will be essential to enhance the robustness and generalization of the whole framework. Besides, this framework only uses linguistic features; we could design ways to combine acoustic features to achieve better performance. Expanding research across languages and dialects will also be crucial for building more inclusive and reliable diagnostic tools.

4.6 Conclusion

Our study highlights the effectiveness of combining machine learning and the LLM as natural language processing techniques for CI screening. We emphasize the importance of linguistic alignment across datasets, the role of high-quality embeddings, and the utility of ensemble classifiers for robust performance. Addressing the limitations identified will pave the way for more efficient and reliable speech-based diagnostic tools in the future. Our findings contribute to the research on automatic and speech-based quick tasks, offering new cross-lingual framework for AI-driven large-scale early screening, self-testing, and healthcare solutions.

Appendix A: Performance on the Local Dataset with Train-Test Split

Table A1 summarizes the performance of different classifiers on the local dataset under train-test split. Each reported metric represents the mean across 100 random splits to ensure robust evaluation. The first value in each cell indicates performance on the training set, while the second value corresponds to the test set.

Appendix B: Embedding Visualization

To gain insights into the distribution of different labels in the embedding space, we applied t-SNE to reduce the high-dimensional embeddings to two dimensions and plotted the resulting data points as a scatter plot (**Figure B1**). The visualization showed that while there is some overlap between the different labels, the overall structure revealed two distinct clusters. This indicated that the

embeddings generated by the model generally capture meaningful semantic differences between the classes, providing a solid foundation for the subsequent classification tasks.

References

[1] International AD. World Alzheimer Report 2023: Reducing Dementia Risk: Never too early, never too late 2023.

[2] WHO Dementia Data https://www.who.int/news-room/facts-in-pictures/detail/dementia (accessed October 28, 2024).

[3] Sachdev PS, Blacker D, Blazer DG, Ganguli M, Jeste DV, Paulsen JS, et al. Classifying neurocognitive disorders: the DSM-5 approach. Nat Rev Neurol 2014;10:634–42. https://doi.org/10.1038/nrneurol.2014.181.

[4] De Roeck EE, De Deyn PP, Dierckx E, Engelborghs S. Brief cognitive screening instruments for early detection of Alzheimer's disease: a systematic review. Alzheimers Res Ther 2019;11:21. https://doi.org/10.1186/s13195-019-0474-3.

[5] Waldemar G, Dubois B, Emre M, Georges J, McKeith IG, Rossor M, et al. Recommendations for the diagnosis and management of Alzheimer's disease and other disorders associated with dementia: EFNS guideline. Eur J Neurol 2007;14:e1-26. https://doi.org/10.1111/j.1468-1331.2006.01605.x.

[6] Scheltens P, Blennow K, Breteler MMB, de Strooper B, Frisoni GB, Salloway S, et al. Alzheimer's disease. Lancet Lond Engl 2016;388:505–17. https://doi.org/10.1016/S0140-6736(15)01124-1.

[7] Weiner MW, Veitch DP, Miller MJ, Aisen PS, Albala B, Beckett LA, et al. Increasing participant diversity in AD research: Plans for digital screening, blood testing, and a community-engaged approach in the Alzheimer's Disease Neuroimaging Initiative 4. Alzheimers Dement J Alzheimers Assoc 2023;19:307–17. https://doi.org/10.1002/alz.12797.

[8] Turner RS, Stubbs T, Davies DA, Albensi BC. Potential New Approaches for Diagnosis of Alzheimer's Disease and Related Dementias. Front Neurol 2020;11:496. https://doi.org/10.3389/fneur.2020.00496.

[9] Clute-Reinig N, Jayadev S, Rhoads K, Ny A-LL. Alzheimer's Disease Diagnostics Must Be Globally Accessible. J Alzheimers Dis 2021;84:1453. https://doi.org/10.3233/JAD-210663.

[10] Roger E, Banjac S, Thiebaut de Schotten M, Baciu M. Missing links: The functional unification of language and memory (L ∪ M). Neurosci Biobehav Rev 2022;133:104489. https://doi.org/10.1016/j.neubiorev.2021.12.012.

[11] Dronkers N, Ogar J. Brain areas involved in speech production. Brain J Neurol 2004;127:1461–2. https://doi.org/10.1093/brain/awh233.

[12] Stück D, Signorini A, Hanai T, Sandoval M, Lemke C, Glass J, et al. NOVEL DIGITAL VOICE BIOMARKERS OF DEMENTIA FROM THE FRAMINGHAM STUDY. Alzheimers Dement 2018;14:P778–9. https://doi.org/10.1016/j.jalz.2018.06.956.

[13] Clarke N, Barrick TR, Garrard P. P2-515: Characterising Spoken Language Deficits in Mild Alzheimer's Disease and Mild Cognitive Impairment. Alzheimers Dement 2018;14:P931–P931. https://doi.org/10.1016/j.jalz.2018.06.1209.

[14] Da S, Sj K, Ja M, Lh G, Dr W, Wr M. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. Findings from the Nun Study. JAMA 1996;275.

[15] Filiou R-P, Bier N, Slegers A, Houzé B, Belchior P, Brambati S. Connected speech assessment in the early detection of Alzheimer's disease and mild cognitive impairment: a scoping review. Aphasiology 2019;34:1–33. https://doi.org/10.1080/02687038.2019.1608502.

[16] S P, D C, M W, J G. Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing. Behav Res Methods 2011;43. https://doi.org/10.3758/s13428-010-0037-9.

[17] The cognitive neuropsychology of Alzheimer-type dementia. New York, NY, US: Oxford University Press; 1996.

[18] Luz S, Haider F, Fuente SDL, Fromm D, MacWhinney B. Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. Interspeech 2020, ISCA; 2020, p. 2172–6. https://doi.org/10.21437/Interspeech.2020-2571.

[19] Luz S, Haider F, Fuente SDL, Fromm D, MacWhinney B. Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge. Interspeech 2021, ISCA; 2021, p. 3780–4. https://doi.org/10.21437/Interspeech.2021-1220.

[20] Li J, Yu J, Ye Z, Wong S, Mak M, Mak B, et al. A Comparative Study of Acoustic and Linguistic Features Classification for Alzheimer's Disease Detection. ICASSP 2021 - 2021 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP, 2021, p. 6423–7. https://doi.org/10.1109/ICASSP39728.2021.9414147.

[21] Martinc M, Haider F, Pollak S, Luz S. Temporal Integration of Text Transcripts and Acoustic Features for Alzheimer's Diagnosis Based on Spontaneous Speech. Front Aging Neurosci 2021;13:642647. https://doi.org/10.3389/fnagi.2021.642647.

[22] König A, Satt A, Sorin A, Hoory R, Toledo-Ronen O, Derreumaux A, et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimers Dement Amst Neth 2015;1:112–24. https://doi.org/10.1016/j.dadm.2014.11.012.

[23] Fraser KC, Meltzer JA, Rudzicz F. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. J Alzheimers Dis JAD 2016;49:407–22. https://doi.org/10.3233/JAD-150520.

[24] Li J, Zhang W-Q. Whisper-Based Transfer Learning for Alzheimer Disease Classification: Leveraging Speech Segments with Full Transcripts as Prompts. ICASSP 2024 - 2024 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP, 2024, p. 11211–5. https://doi.org/10.1109/ICASSP48485.2024.10448004.

[25] Thomas JA, Burkhardt HA, Chaudhry S, Ngo AD, Sharma S, Zhang L, et al. Assessing the Utility of Language and Voice Biomarkers to Predict Cognitive Impairment in the Framingham Heart Study Cognitive Aging Cohort Data. J Alzheimers Dis JAD 2020;76:905–22. https://doi.org/10.3233/JAD-190783.

[26] Chandler C, Diaz-Asper C, Turner RS, Reynolds B, Elvevåg B. An explainable machine learning model of cognitive decline derived from speech. Alzheimers Dement Diagn Assess Dis Monit 2023;15:e12516. https://doi.org/10.1002/dad2.12516.

[27] Wang Y, Wang T, Ye Z, Meng L, Hu S, Wu X, et al. Exploring linguistic feature and model combination for speech recognition based automatic AD detection 2022. https://doi.org/10.48550/arXiv.2206.13758.

[28] Wang Y, Deng J, Wang T, Zheng B, Hu S, Liu X, et al. Exploiting prompt learning with pre-trained language models for Alzheimer's Disease detection 2023. https://doi.org/10.48550/arXiv.2210.16539.

[29] Pan Y, Mirheidari B, Harris JM, Thompson JC, Jones M, Snowden JS, et al. Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic- and BERT-Based Alzheimer’s Dementia Detection Through Spontaneous Speech. Interspeech 2021, ISCA; 2021, p. 3810–4. https://doi.org/10.21437/Interspeech.2021-1519.

[30] Gómez-Zaragozá L, Wills S, Tejedor-Garcia C, Marín-Morales J, Alcañiz M, Strik H. Alzheimer Disease Classification through ASR-based Transcriptions: Exploring the Impact of Punctuation and Pauses. INTERSPEECH 2023, ISCA; 2023, p. 2403–7. https://doi.org/10.21437/Interspeech.2023-1734.

[31] Guo Z, Liu Z, Ling Z, Wang S, Jin L, Li Y. Text Classification by Contrastive Learning and Cross-lingual Data Augmentation for Alzheimer's Disease Detection. In: Scott D, Bel N, Zong C, editors. Proc. 28th Int. Conf. Comput. Linguist., Barcelona, Spain (Online): International Committee on Computational Linguistics; 2020, p. 6161–71. https://doi.org/10.18653/v1/2020.coling-main.542.

[32] Amini S, Hao B, Yang J, Karjadi C, Kolachalama VB, Au R, et al. Prediction of Alzheimer's disease progression within 6 years using speech: A novel approach leveraging language models. Alzheimers Dement 2024;20:5262–70. https://doi.org/10.1002/alz.13886.

[33] Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. Int J Med Inf 2019;125:37–46. https://doi.org/10.1016/j.ijmedinf.2019.02.008.

[34] Amini S, Hao B, Zhang L, Song M, Gupta A, Karjadi C, et al. Automated detection of mild cognitive impairment and dementia from voice recordings: A natural language processing approach. Alzheimers Dement 2023;19:946–55. https://doi.org/10.1002/alz.12721.

[35] Becker JT, Boller F, Lopez OL, Saxton J, McGonigle KL. The natural history of Alzheimer's disease. Description of study cohort and accuracy of diagnosis. Arch Neurol 1994;51:585–94. https://doi.org/10.1001/archneur.1994.00540180063015.

[36] Radford A, Kim JW, Xu T, Brockman G, Mcleavey C, Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision. Proc. 40th Int. Conf. Mach. Learn., PMLR; 2023, p. 28492–518.

[37] GLM T, Zeng A, Xu B, Wang B, Zhang C, Yin D, et al. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools 2024. https://doi.org/10.48550/arXiv.2406.12793.

[38] Embedding-3 Large Language Embedding Model https://open.bigmodel.cn/dev/api/vector/embedding-3 (accessed November 1, 2024).

[39] Breiman L. Random Forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.

[40] Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Their Appl 1998;13:18–28. https://doi.org/10.1109/5254.708428.
[41] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System 2016. https://doi.org/10.48550/arXiv.1603.02754.
[42] Maaten L van der, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res 2008;9:2579–605.
[43] nreimers/MiniLM-L6-H384-uncased · Hugging Face https://huggingface.co/nreimers/MiniLM-L6-H384-uncased (accessed October 28, 2024).
[44] Eskildsen SF, Coupé P, García-Lorenzo D, Fonov V, Pruessner JC, Collins DL, et al. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. NeuroImage 2013;65:511–21. https://doi.org/10.1016/j.neuroimage.2012.09.058.
[45] Yang Y, Cer D, Ahmad A, Guo M, Law J, Constant N, et al. Multilingual Universal Sentence Encoder for Semantic Retrieval. In: Celikyilmaz A, Wen T-H, editors. Proc. 58th Annu. Meet. Assoc. Comput. Linguist. Syst. Demonstr., Online: Association for Computational Linguistics; 2020, p. 87–94. https://doi.org/10.18653/v1/2020.acl-demos.12.

Figure 1: The pipeline of our framework. ASR, automatic speech recognition; LLM, large language model; XGBoost, extreme gradient boosting; RF, random forest; MLP, multiple-layer perception, CU, cognitively unimpaired; CI, cognitive impairment.

Figure 2: The performance of different classifiers in Monolingual Experiment and Cross-Lingual Experiment. ACC, accuracy; AUC, area under the curve. LR, Logistic Regression; SVM, Support Vector Machine; RF, Random Forest; XGBoost, a Boosting; MLP, Multi-Layer Perceptron; MLP-Trans, Multi-Layer Perceptron with Transformer architecture.

Figure 3: Results of our ablation study for measuring the effectiveness of different settings using RF as the classifier. ACC, accuracy; PRE, precision; REC, recall; F1, F1 score; AUC, area under the curve.

Figure B1: Scatter plot of the embedding space, the distribution of CI and CU. t-SNE, t-Distributed Stochastic Neighbor Embedding; CU, cognitively unimpaired; CI, cognitive impairment.

Acknowledgment

This study was partly supported by the Shenzhen Science and Technology Program (KCXFZ20211020163408012). The funder played an essential role in the data collection and the writing of this manuscript. The authors would like to thank the participants for their time.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

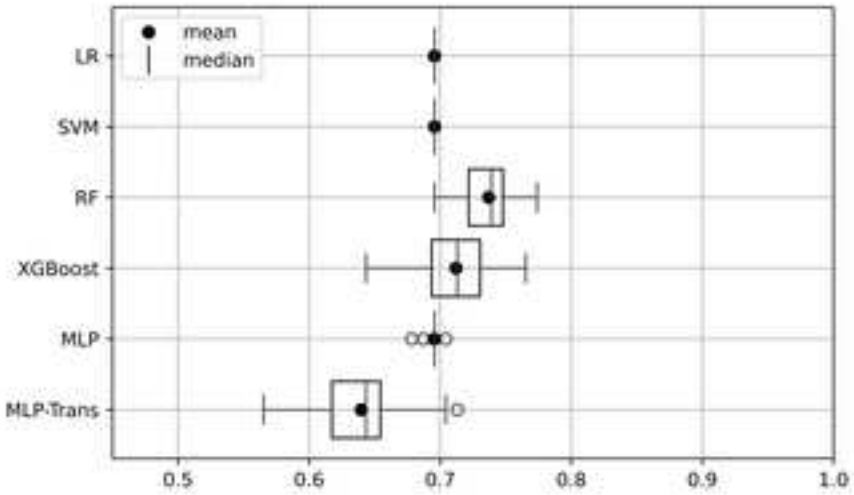
Consent Statement

All human objects provided informed consent.

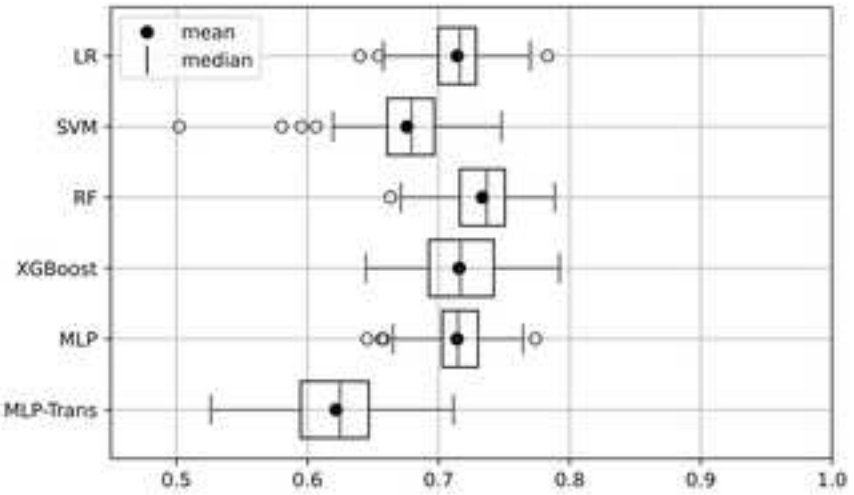
KEYWORDS

Alzheimer’s disease, cognitive impairment, Large language model, NLP, machine learning

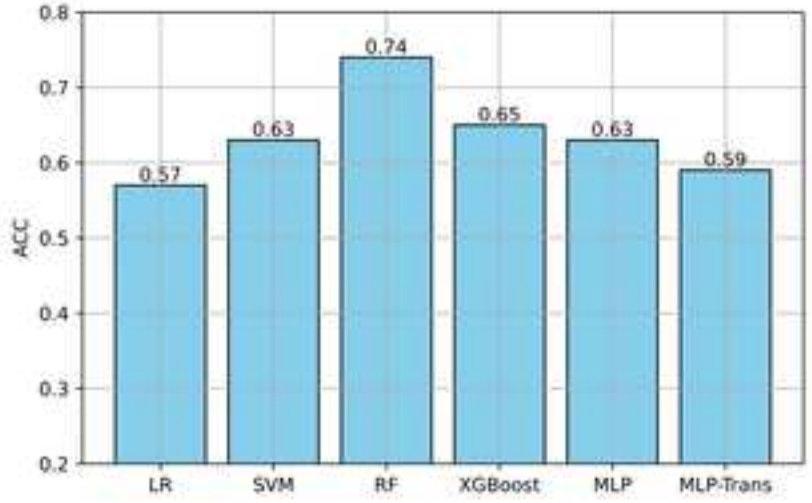
Figure 2



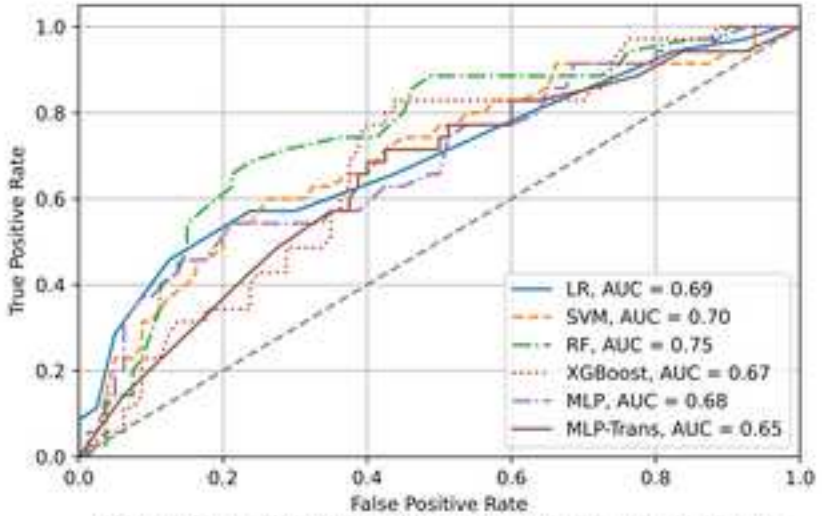
(A) The distribution of ACC for different classifiers in Monolingual Experiment



(B) The distribution of AUC for different classifiers in Monolingual Experiment



(C) ACC for different classifiers in Cross-Lingual Experiment



(D) AUC for different classifiers in Cross-Lingual Experiment

Figure 3

Click here to access/download;Figure;Figure 3_Yue Wu.jpeg

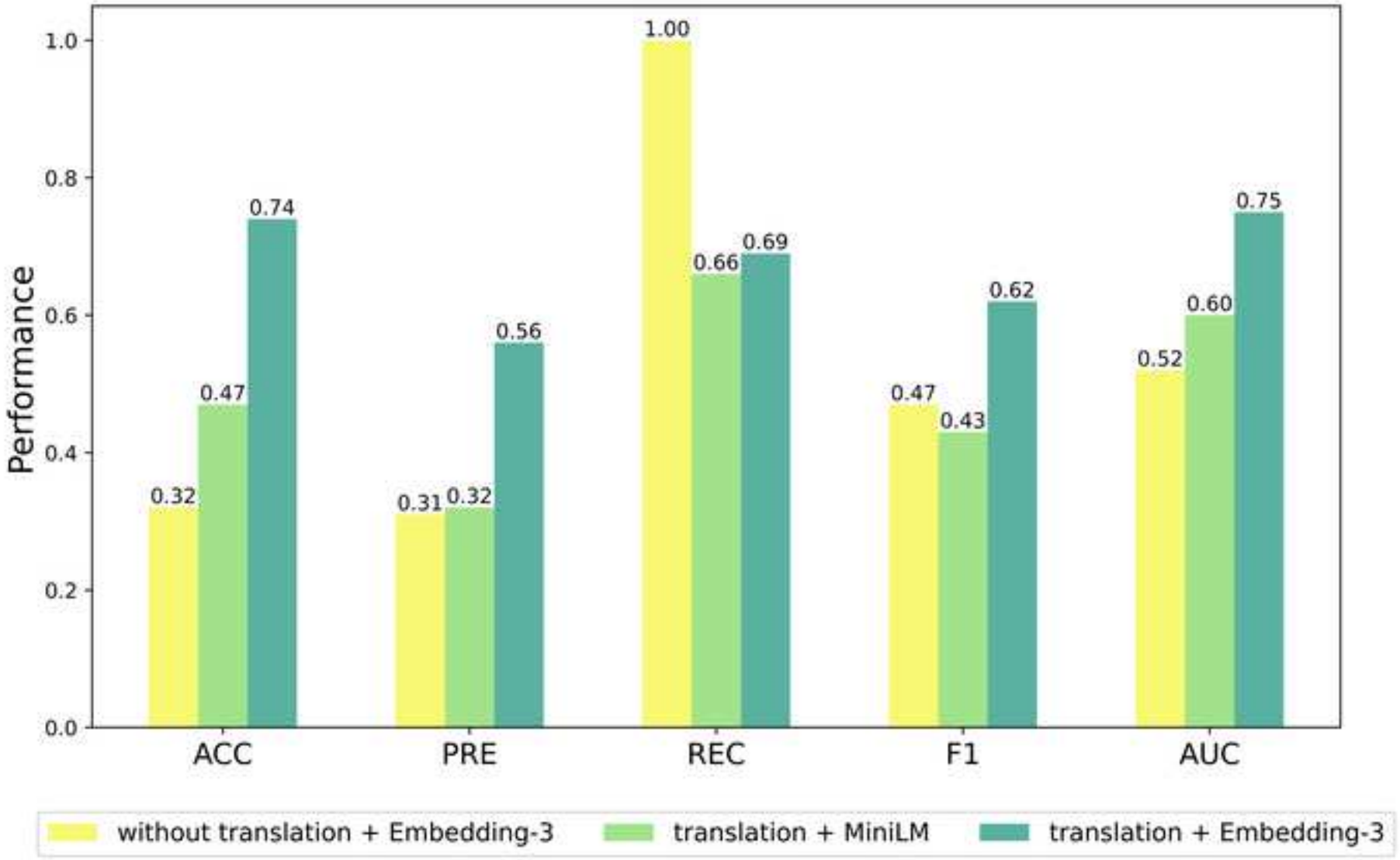


Figure B1

Click here to access/download;Figure;Figure B1_Yue Wu.jpeg

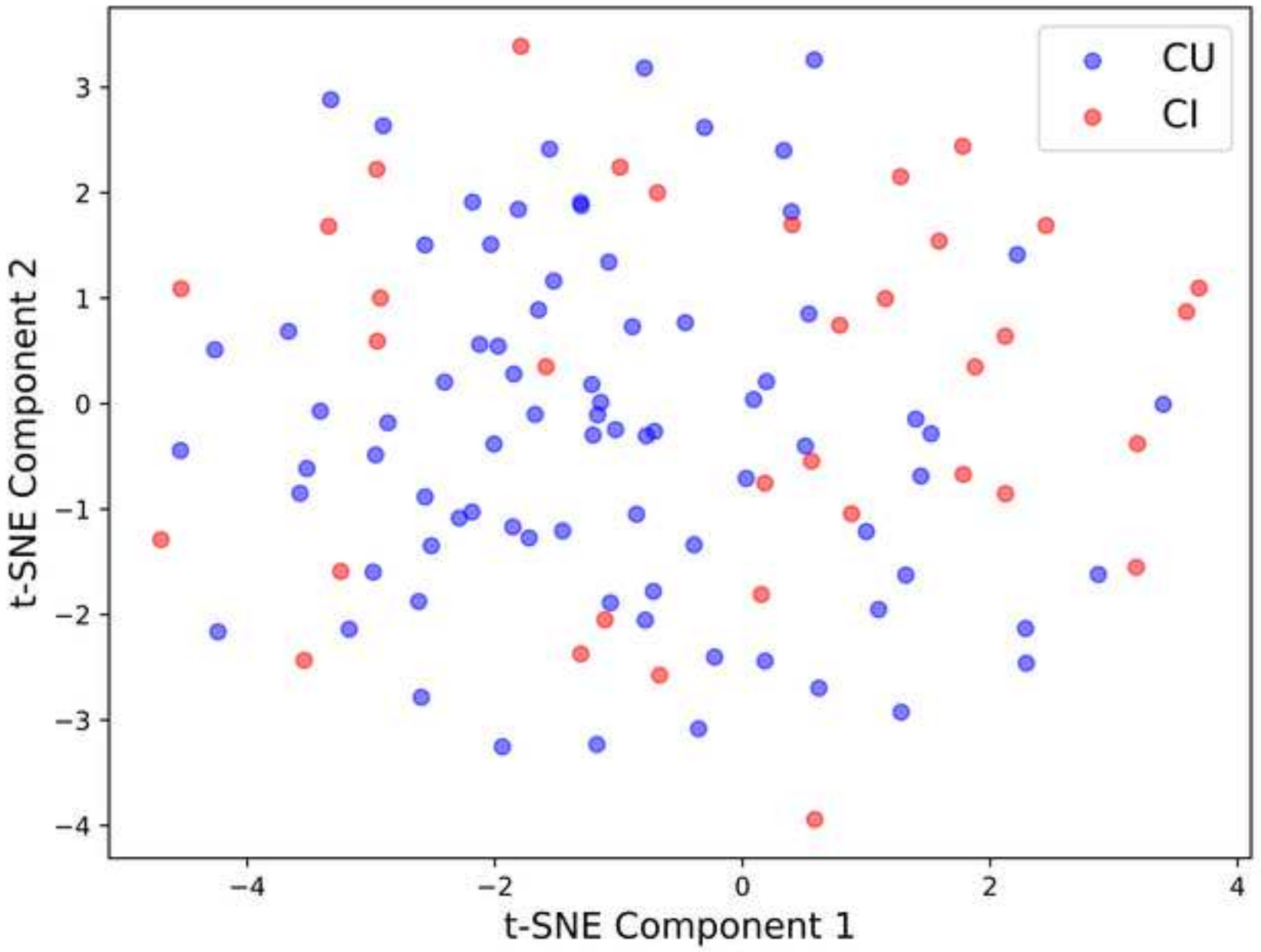


Figure 1

Table 2

Table 2 Average performance of classifiers on local data evaluated using 5-fold cross-validation, left for train and right for test.

	ACC	PRE	REC	F1	AUC
LR	0.70/0.70	0.00/0.00	0.00/0.00	0.00/0.00	0.86/0.71
SVM	0.70/0.70	0.00/0.00	0.00/0.00	0.00/0.00	0.94/0.68
RF	**1.00**/**0.74**	**1.00**/**0.66**	**1.00**/0.31	**1.00**/**0.40**	**1.00**/**0.73**
XGBoost	**1.00**/0.71	**1.00**/0.57	**1.00**/**0.36**	**1.00**/**0.40**	**1.00**/0.72
MLP	0.70/0.70	0.07/0.00	0.01/0.00	0.01/0.00	0.50/0.71
MLP-Trans	0.58/0.64	0.30/0.41	0.30/0.35	0.30/0.35	0.50/0.62

Abbreviations: LR, Logistic Regression; SVM, Support Vector Machine; RF, Random Forest; XGBoost, a Boosting; MLP, Multi-Layer Perceptron; MLP-Trans, Multi-Layer Perceptron with Transformer architecture; ACC, accuracy; PRE, precision; REC, recall; F1, F1 score, AUC, area under the curve

Table 3

Table 3 The performance of classifiers on local data trained on public datasets.

	ACC	PRE	REC	F1	AUC
LR	0.57	0.39	**0.77**	0.52	0.69
SVM	0.62	0.42	0.66	0.51	0.70
RF	**0.74**	**0.56**	0.69	**0.62**	**0.75**
XGBoost	0.65	0.45	0.71	0.56	0.67
MLP	0.63	0.42	0.57	0.48	0.68
MLP-Trans	0.59	0.41	0.74	0.53	0.67

Abbreviations: LR, Logistic Regression; SVM, Support Vector Machine; RF, Random Forest; XGBoost, a Boosting; MLP, Multi-Layer Perceptron; MLP-Trans, Multi-Layer Perceptron with Transformer architecture; ACC, accuracy; PRE, precision; REC, recall; F1, F1 score, AUC, area under the curve

Table 4

Table 4 The performance of classifiers on local data trained on public datasets, without translation(left) and using MiniLM(right).

	ACC	PRE	REC	F1	AUC
LR	0.30/0.66	0.30/0.46	**1.00/0.66**	**0.47/0.54**	**0.69/0.69**
SVM	0.30/0.6	0.30/0.42	**1.00/0.69**	**0.47/0.52**	0.65/0.69
RF	0.32/0.47	0.31/0.32	**1.00/0.66**	**0.47/0.43**	0.52/0.60
XGBoost	0.30/0.47	0.30/0.31	0.97/0.63	0.46/0.42	0.55/0.51
MLP	0.30/0.66	0.30/0.46	0.97/0.63	0.46/0.53	0.65/0.66
MLP-Trans	**0.43/0.50**	**0.33/0.29**	0.83/0.46	**0.47/0.36**	0.64/054

Abbreviations: LR, Logistic Regression; SVM, Support Vector Machine; RF, Random Forest; XGBoost, a Boosting; MLP, Multi-Layer Perceptron; MLP-Trans, Multi-Layer Perceptron with Transformer architecture; ACC, accuracy; PRE, precision; REC, recall; F1, F1 score, AUC, area under the curve

Table A1 Average performance of classifiers on local data, evaluated using a train-test split. Results are presented with training performance on the left and testing performance on the right.

	ACC	PRE	REC	F1	AUC
LR	0.70/0.69	0.00/0.00	0.00/0.00	0.00/0.00	0.87/0.70
SVM	0.70/0.69	0.00/0.00	0.00/0.00	0.00/0.00	0.94/0.66
RF	**1.00**/**0.72**	**1.00**/**0.69**	**1.00**/0.27	**1.00**/0.36	**1.00**/**0.72**
XGBoost	**1.00**/0.70	**1.00**/0.53	**1.00**/**0.34**	**1.00**/**0.39**	**1.00**/0.69
MLP	0.70/0.69	0.07/0.01	0.01/0.00	0.02/0.00	0.50/0.70
MLP-Trans	0.58/0.63	0.30/0.41	0.30/0.34	0.30/0.35	0.50/0.61

Abbreviations: LR, Logistic Regression; SVM, Support Vector Machine; RF, Random Forest; XGBoost, a Boosting; MLP, Multi-Layer Perceptron; MLP-Trans, Multi-Layer Perceptron with Transformer architecture; ACC, accuracy; PRE, precision; REC, recall; F1, F1 score, AUC, area under the curve

Table 1

TABLE1 Characteristics in the STAR cohort

	CU	CI
Number of subjects	80	35
Sex(Male/Female)	31/49	12/23
Age(Mean±SD)	62.09±0.64	64.73±1.17
Education(Mean±SD)	13.70±0.27	10.08±0.69
MMSE(Mean±SD)	28.68±0.14	25.36±0.60
MoCA(Mean±SD)	26.30±0.22	18.24±0.73

Abbreviations: MMSE, Mini-Mental State Examination; MoCA, Montreal Cognitive Assessment; CU, cognitively unimpaired; CI, cognitive impairment.

Highlights:

- Constructed a novel classification framework for distinguishing cognitive impairment (CI) from cognitively unimpaired (CU) using speech data from the Cookie Theft picture description task across different languages, achieving 74% in accuracy and 75% in AUC in the external cross-lingual validation experiment.
- Leveraged AI methods, including Automatic Speech Recognition (ASR), Large Language Models (LLMs), and various machine learning classifiers, to develop an automatic framework for CI screening.
- Developed a fully automatic assessment framework that excels in distinguishing CI, providing a valuable tool for large-scale early screening and self-testing of cognitive impairment.

Research in Context:

Systematic review:

Few studies have investigated available automated frameworks by confusing acoustic and linguistic features for early screening cognitive impartment (CI). Additionally, it also lacks cross-language attempts, caused by data scarcity and insufficient sample diversity.

Interpretation:

We developed a novel automated framework that combines the ASR method for speech transcription and large language model (LLM)-based module for linguistic feature extraction. Our findings, based on a multilingual dataset combined ADReSSo dataset ($n = 166$) and the STAR Cohort ($n = 115$), demonstrate superior performance in machine learning (ML) classifiers, achieving 74% in accuracy and 75% in AUC in the external cross-lingual Chinese validation experiment.

Future direction:

This study highlights the effectiveness of a framework consisting of LLMs and MLs for CI screening in multilingual scenarios. It contributes to developing a cost-effective, widely accessible way for large-scale early screening and self-testing of CI, offering some inspiration for AI-driven healthcare solutions.

This PowerPoint is the original file of the Figure 1

Click here to access/download

Video Files

Figure 1_Yue Wu.pptx