Speech-Based Dementia Detection for Low-Resource Communities via Symbolic and Text Encoding

Toan Quoc Nguyen[1], Linh Le[2], and Nghia Duong-Trung[3]

[1] Research Center Trustworthy Data Science and Security
TU Dortmund University, Dortmund, Germany
toanquoc.nguyen@tu-dortmund.de
[2] Mila – Quebec AI Institute, Montreal, Canada
thai.linh.le@mila.quebec
[3] IU International University of Applied Sciences
Frankfurter Allee 73a, 10247, Berlin, Germany
nghia.duong-trung@iu.org

Abstract. Early detection of dementia, especially its most common form, Alzheimer's disease (AD), remains a critical challenge, particularly in low-resource communities where access to specialist assessment and neuroimaging is limited. The objective of this study is to develop a scalable, non-invasive, and cost-effective speech-based method for early detection of AD. The approach is evaluated on speech samples from 237 participants performing the Cookie Theft picture description task. Acoustic patterns are captured using Symbolic Aggregate Approximation (SAX), while linguistic information is represented through embeddings from a pre-trained language model. These multimodal features are integrated using a Gated Recurrent Unit (GRU) network with ensemble learning. The main results demonstrate an accuracy of 97.08%, outperforming existing state-of-the-art speech-based AD detection methods. The significance of this work lies in its proposed high-performing model, its accessibility, and its scalability, thereby offering a practical screening solution that can facilitate earlier intervention, reduce healthcare disparities, and support wider deployment of AD assessment in low-resource settings.

Keywords: Alzheimer's, Speech, Symbolic Aggregate Approximation, Text Embedding, Biomarkers, Deep Learning

1 Introduction

Dementia is recognised as the seventh leading cause of death worldwide and is a significant contributor to disability and dependence among older adults [1]. Among the different types of dementia, Alzheimer's Disease (AD) is the most common, representing approximately 60% to 80% of cases [2,3,4,5], with its prevalence markedly increasing among individuals aged 65 years and older

[6]. AD is marked by a gradual decline in cognitive abilities, memory loss, and neuronal degeneration, ultimately leading to brain atrophy and extensive tissue damage [7]. Given the absence of a definitive cure [2], early detection is crucial to slow disease progression and to improve individuals' Quality of Life (QoL) through timely interventions and supportive efforts [8,9].

The advancement of Artificial Intelligence (AI) technologies, including Machine Learning (ML) and Deep Learning (DL), has created new opportunities for early AD detection. Nevertheless, many current approaches rely on costly imaging techniques such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) [4,10,11,12,13], which are often unavailable or impractical in low-resource settings [4]. Although Electroencephalography (EEG) [16] offers a more affordable alternative, it can still pose financial challenges for many underprivileged communities due to its relatively high cost. Additionally, several methods depend on invasive biomarkers, such as Cerebrospinal Fluid (CSF) analysis [17], which may cause discomfort, reduce patient willingness to undergo testing, and limit broader clinical adoption.

Compared to those mentioned above, speech-based methods present a promising alternative, as they are more cost-effective, non-invasive, and highly adaptable for remote cognitive assessments and telemedicine systems [18,19,20,21,22], offering a scalable solution to reach and support a broader population in need. Language impairments can serve as early indicators of cognitive decline and act as biomarkers for AD [23]. Such impairments often manifest as changes in speech rate, pausing patterns, word retrieval difficulties, and coherence of expression [24]. Both acoustic features, including prosody, fluency, and pause distribution, and linguistic features, such as syntactic complexity, semantic richness, and discourse coherence, are capable of capturing these alterations, thereby enabling earlier and more accurate detection [25,26]. Previous studies have shown that these features can distinguish individuals with AD from Normal Controls (NC) and exhibit strong discriminative capacity in early-stage detection [27,28].

To support early AD detection in resource-constrained communities, this study introduces a scalable, non-invasive approach that leverages speech and AI. It combines Symbolic Aggregate Approximation (SAX) for time-series feature extraction [29,30], pre-trained Automatic Speech Recognition (ASR) transcripts generated using Whisper-large-v3 [31,32,33], and text embedding vectors from the Solon-embeddings-large-0.1 model [34,35] within a deep learning approach.

The key contribution of this study lies in providing a cost-effective and clinically applicable solution, enabling broad deployment in real-world settings, especially for low-resource communities, while jointly modelling acoustic and linguistic characteristics through a unified architecture that achieves high detection accuracy. To illustrate, the developed method attained 97.08% accuracy,

[4] Low-resource settings refer to healthcare environments with constrained financial resources, limited access to advanced medical equipment, insufficient specialist availability, or underdeveloped clinical infrastructure, such as rural regions or low- and middle-income countries [14,15].

achieving the model's fairness with performance consistency across genders[5] using speech samples from 237 participants describing the Cookie Theft picture from the Boston Diagnostic Aphasia Examination [40], a task shown to be effective for distinguishing NC from AD [41,40]. SAX features were processed through a Gate Recurrent Unit (GRU) network [42] and linear layer [43] to model temporal dependencies, while transcript embeddings were passed through a linear layer. Outputs from both branches were fused and classified with eXtreme Gradient Boosting (XGBoost) [44] (ensemble learning), enhancing performance.

2 Related Work

There have been various advances in using speech and AI for AD detection. Li *et al.* [32] applied transfer learning with Whisper, using full transcripts to enhance segment training, achieving 0.8451 accuracy and 0.8450 F1-score. Wang *et al.* proposed HSAF-DWAN [45], a multimodal AD detection model combining audio, transcripts, and attention-weighted images, achieving 0.8671 accuracy and 0.8815 F1-score. Gauder *et al.* [46] utilised pre-trained text embeddings from TRILL, Allosaurus, and wav2vec 2.0 with a segment-based neural network to detect AD, achieving 0.7887 accuracy. Pan *et al.* [47] developed the Tree Bagger (TB) model using acoustic and character path signatures extracted from wav2vec 2.0 embeddings and transcripts, also achieving 0.7887 accuracy. Pappagari *et al.* [48] combined acoustic (x-vectors, prosody) and linguistic (ASR-BERT) models, achieving 0.8451 accuracy. Jiawen *et al.* [49] proposed Soft Target Distillation (SoTD) and Instance-level Re-balancing (InRe), achieving 0.8464 accuracy with BERT features and 0.8380 accuracy with RoBERTa features. Finally, Yifan *et al.* [50] proposed DSTC-Net, a dual-stage time-context network that combines local acoustic features and global conversational patterns, achieving 0.8310 accuracy and 0.8315 F1-score.

Despite these advances, three critical gaps remain. First, fairness in model evaluation is often overlooked, limiting the reliability of speech-based AI models and potentially introducing biases, particularly gender-related, which can degrade performance and affect detection outcomes [37,38,39]. Second, detailed pattern analyses are frequently neglected, reducing the transparency of developed AI systems, even though understanding the data used for model development is vital for fostering user trust [51]. Third, most existing models achieve accuracies below 0.9, which is generally considered inadequate for deployment in medical applications, where higher detection performance standards are paramount. Therefore, this research seeks to address current limitations in the literature by proposing a high-performing and more reliable AI-based method for advanc-

[5] In this study, gender refers to biological sex, categorised as male (M) or female (F) [36]. Previous research has demonstrated that speech-based AI models are susceptible to gender-related biases, which may affect both fairness and predictive accuracy [37,38]. Furthermore, biases in word embeddings highlight the importance of incorporating gender considerations during model evaluation [39].

ing speech-based AD detection, with a particular focus on supporting broader community benefits in low-resource settings.

3 Methods

3.1 Material

This study utilises a widely recognised and extensively used public speech dataset for AD detection, comprising 237 English-speaking participants, including 115 individuals with normal cognition (NC; 40 M, 75 F; mean age 60.06 ± 6.30 years) and 122 individuals with Alzheimer's disease (AD; 43 M, 79 F; mean age 69.38 ± 6.90 years) [40]. This benchmark dataset has been adopted in numerous prior studies as a standard evaluation corpus for speech-based dementia detection (See Table 2), enabling direct comparison with existing approaches. Ground truth labels were provided in the original release and were assigned by domain professionals following established clinical procedures.

All participants completed the Cookie Theft picture description task from the Boston Diagnostic Aphasia Examination [40,41,40]. Speech recordings had an average duration of 61.60 ± 26.90 seconds for NC participants and 65.70 ± 38.30 seconds for AD participants. The recordings were sampled at 41 kHz to preserve fine-grained acoustic information, which is particularly important for medical speech analysis [52,53].

Before analysis, audio signals were amplitude-normalised to the range $[-1, 1]$ to ensure consistency across recordings. Spectral subtraction was applied to reduce stationary background noise, followed by Wiener filtering to mitigate residual artefacts and improve speech intelligibility before transcription. These preprocessing steps were selected based on established practices in speech enhancement to ensure noise reduction while preserving clinically relevant acoustic characteristics [54].

3.2 Proposed Model

Figure 1 illustrates the general workflow of the proposed method, including several components below:

Transcript Generation: Given a user's speech recording in the form of a raw audio waveform $\mathbf{y}^{(i)}$, the signal is processed and transcribed using the *Whisper-large-v3* [31,32,33], a pre-trained state-of-the-art ASR model. The resulting transcript $T^{(i)}$ is stored as plain text for downstream processing.

Text Embedding: The transcript $T^{(i)}$ is tokenised and embedded using the pre-trained *Solon-embeddings-large-0.1* model [34,35], resulting in a linear vector $\mathbf{x}_{\text{text}}^{(i)} \in R^{32}$. This embedding captures the linguistic semantics of the spoken content, providing a compact representation of the participant's language.

SAX Audio Encoding: In parallel, the same raw audio waveform $\mathbf{y}^{(i)}$ is z-normalised to ensure zero mean and unit variance. The signal is segmented into 50 ms frames, selected based on the lowest energy ratio, indicating effective capture of high-frequency speech dynamics. These raw signal segments are

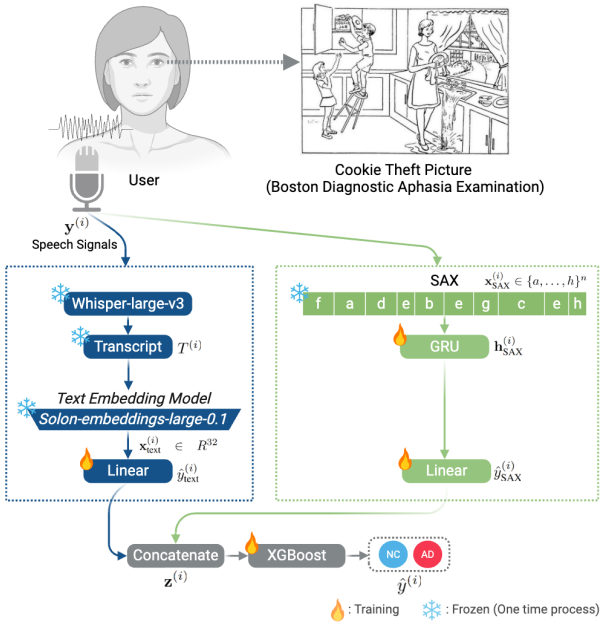

Fig. 1: Overview of the proposed speech-based AD detection framework. Given a speech signal $y^{(i)}$ from participant i describing the Cookie Theft picture, two parallel streams are constructed: (1) a text stream where Whisper-large-v3 generates the transcript $T^{(i)}$, which is encoded into a semantic embedding $x^{(i)}_{\text{text}} \in \mathbb{R}^{32}$ and passed through a linear layer; and (2) an acoustic stream where the signal is converted into a symbolic sequence $x^{(i)}_{\text{SAX}} \in \{a, \ldots, h\}^n$ using SAX and processed by a GRU to produce temporal features. The resulting representations are concatenated as $z^{(i)}$ and fed into an XGBoost classifier to obtain the final prediction $\hat{y}^{(i)} \in \{\text{NC}, \text{AD}\}$, where snowflake and flame icons denote frozen and trainable components, respectively.

summarised using Piecewise Aggregate Approximation (PAA) [29,30] and then discretised into symbolic form via SAX, yielding a symbolic sequence $\mathbf{x}^{(i)}_{\text{SAX}} \in \{a, \ldots, h\}^n$. The SAX alphabet size of 8 was selected before model training as it achieved the highest entropy score of 0.9266 on the training data, indicating a balanced and information-rich symbolic representation.

SAX-GRU Temporal Modelling Prediction: The symbolic sequence $\mathbf{x}^{(i)}_{\text{SAX}}$ is embedded into a continuous vector space and fed into a GRU network [43], producing a temporal representation $\mathbf{h}^{(i)}_{\text{SAX}}$. This representation is passed through a linear (fully connected) layer:

$$\hat{y}^{(i)}_{\text{SAX}} = \sigma\left(\mathbf{w}^{\top}_{\text{SAX}}\mathbf{h}^{(i)}_{\text{SAX}} + b_{\text{SAX}}\right) \tag{1}$$

where $\mathbf{w}_{\text{SAX}}$ and b_{SAX} are learnable parameters, and $\sigma(\cdot)$ denotes the sigmoid activation. The output $\hat{y}^{(i)}_{\text{SAX}}$ represents the probability of AD based on acoustic features.

Text-based Prediction: Similarly, the text embedding $\mathbf{x}^{(i)}_{\text{text}}$ is passed through a linear layer and sigmoid function to independently produce a probability score from the linguistic content:

$$\hat{y}^{(i)}_{\text{text}} = \sigma\left(\mathbf{w}^{\top}_{\text{text}}\mathbf{x}^{(i)}_{\text{text}} + b_{\text{text}}\right) \tag{2}$$

Feature Fusion and Final Decision: The output scores from the acoustic ($\hat{y}^{(i)}_{\text{SAX}}$) and linguistic ($\hat{y}^{(i)}_{\text{text}}$) branches are concatenated into a vector:

$$\mathbf{z}^{(i)} = \begin{bmatrix} \hat{y}^{(i)}_{\text{SAX}} \\ \hat{y}^{(i)}_{\text{text}} \end{bmatrix} \tag{3}$$

This vector is fed into an XGBoost classifier [44], which outputs the final prediction $\hat{y}^{(i)}$ along with an associated confidence score $\hat{p}^{(i)}$. If $\hat{y}^{(i)} = 1$, the model predicts the individual as AD.

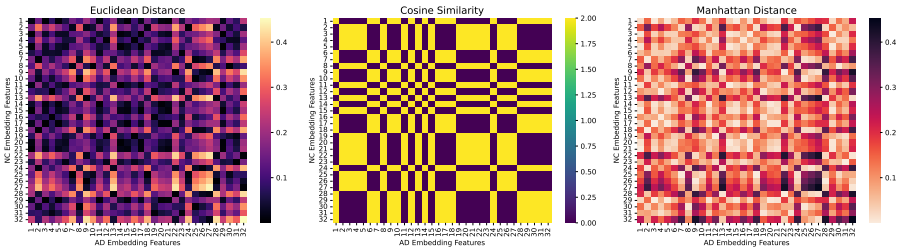

Fig. 2: Euclidean, Cosine, and Manhattan Distances of Embedding Features Between NC and AD.

4 Experiments

For model performance evaluation, the model was validated using k-fold cross-validation ($k = 10$) with all model settings, allowing evaluation across multiple data splits. Key metrics are reported: accuracy ($Acc \uparrow$), F1-score ($F1 \uparrow$), True Positive Rate ($TPR \uparrow$), and False Positive Rate ($FPR \downarrow$). These metrics range between 0 and 100%, with higher values ($\uparrow$) indicating better performance, except for FPR, where lower values ($\downarrow$) indicate better models.

To evaluate fairness, three widely-adopted metrics are employed, calculated as differences across demographic groups (A = males and B = females, as detailed in Section 3.1). Equalized Odds is measured using differences in True

Positive Rate (ΔTPR) and False Positive Rate (ΔFPR) [55,56], Overall Accuracy Equality (OAE) compares prediction accuracy across groups [55], and calibration fairness is assessed via group-specific Brier scores (Δ_{Brier}) [55,57]. In addition, Expected Calibration Error (ECE) is reported alongside the Brier score, with both its values and the gap between groups (Δ_{ECE}) across groups calculated. Fairness is satisfied when the group differences fall below the threshold $\gamma = 0.2$, ensuring that females achieve at least 80% of the performance of males (and vice versa) [58,59,60].

Regarding text-embedding models, the prominent ones in the literature and MTEB ranking [61,62] have been leveraged to develop the proposed method: Solon-embeddings-large-0.1, bge-m3, granite-embedding-278m-multilingual, gte-multilingual-base, multilingual-e5-large-instruct-scripts, snowflake-arctic-embed-l-v2.0, and text-embedding-3-small. Moreover, this study evaluates several deep learning architectures commonly used in time-series classification tasks. These include Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and GRU. Their inclusion is based on their widespread adoption, demonstrated effectiveness, and continued relevance in the field of sequential data analysis [43].

For reproducibility purposes, the transcribed speech was embedded using the *Solon-embeddings-large-0.1* model from HuggingFace [34]. Input texts were tokenized with *AutoTokenizer* and encoded with *AutoModel* using a maximum sequence length of 512, with *padding=True* and *truncation=True*. Linguistic embeddings were extracted using the *last_token_pool* strategy, followed by *L2* normalization, and the final representation was dimensionally reduced by truncating to a fixed size of 32. Concurrently, the dual-stream model architecture processes both modalities through independent, dedicated paths before fusion. In the acoustic stream, the discrete SAX symbolic sequences are projected into a continuous feature space using a trainable token embedding layer with an embedding dimension of 16. These continuous token vectors are then fed into a single-layer GRU with 32 hidden units, followed by a linear layer and a *sigmoid* activation. This recurrent sequence architecture is trained for 10 epochs using the *Adam* optimizer with a learning rate of 0.001 and binary cross-entropy loss (*BCELoss*). The linguistic stream independently routes the 32-dimensional truncated text embeddings through a separate linear classifier consisting of one hidden layer with 32 units optimized across a maximum of 500 training iterations. Finally, the extracted feature outputs from both streams are integrated via an ensemble meta-classifier utilizing *XGBoost* [44] configured with *n_estimators=100* and *min_samples_split=2*.

5 Results

The proposed method demonstrates high performance while generally ensuring fairness, good calibration, and statistically significant feature differences, positioning it as a reliable solution for speech-based AD detection in low-resource settings.

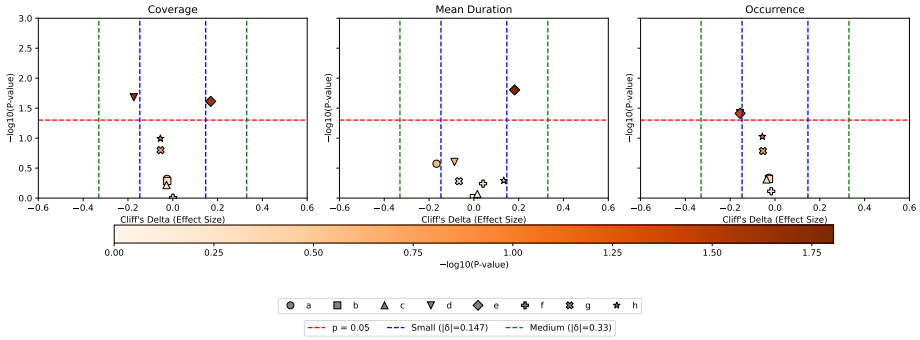

Fig. 3: Effect Size vs. Significant Difference for SAX Features.

5.1 Model Performance

The proposed approach achieved a noticeable high performance with the *Solon-embeddings-large-0.1 + GRU* configuration at embedding size 32 (See Table 1), achieving an accuracy of 97.08%±3.25%, F1-score of 97.23%±3.12%, TPR of 98.40%±3.21%, and FPR of 4.32%±4.33%. Especially, it also outperforms recent state-of-the-art methods in speech-based AD detection (See Table 2), proving its high effectiveness.

Moreover, several important insights can be drawn from the results. First, large-scale pretrained embeddings (Solon, gte-multilingual-base, multilingual-e5-large-instruct) consistently outperformed smaller alternatives (e.g., text-embedding-3-small), underscoring the importance of semantic richness in text-derived features. Second, GRU architectures generally delivered higher accuracy and stability compared to RNN, LSTM, and BiLSTM, with GRU combined with Solon or multilingual-e5 embeddings consistently producing leading results across embedding sizes. Finally, top-performing models maintained low FPR, with the lowest reported at 1.67%±3.33% (multilingual-e5-large-instruct + RNN, size 128), reducing detection risk and strengthening the clinical applicability of the proposed approach. Overall, these findings demonstrate that **embedding selection and temporal modelling are effectively integrated** for having high model performance in speech-based AD detection in low-resource communities.

5.2 Model Analysis

The proposed model demonstrates strong calibration, with an overall Brier score of 0.0530 ± 0.0204 and ECE of 0.1639 ± 0.0369, values that fall within the range regarded as well-calibrated in clinical machine learning systems [64]. Regarding fairness analysis (See Table 3), it shows minimal disparities across gender groups, with $\Delta_{TPR} = 0.0333 \pm 0.0703$, $\Delta_{FPR} = 0.0900 \pm 0.1015$, and $OAE = 0.0435 \pm 0.0572$, all substantially below the accepted threshold of $\gamma = 0.2$ [58,59,60]; calibration fairness metrics are also low ($\Delta_{Brier} = 0.0174 \pm 0.0209$, $\Delta_{ECE} = 0.0362 \pm 0.0278$), proving consistent probabilistic reliability across subgroups.

Table 1: Performance of the Proposed Method Across Deep Learning Time-Series Models.

Text-embedding	Time-series	Acc ↑	F1 ↑	TPR ↑	FPR ↓
Solon-embeddings-large-0.1	GRU	97.08±3.25	97.23±3.12	98.40±3.21	4.32±4.33
	RNN	94.95±2.48	95.05±2.56	95.13±5.46	5.23±4.28
	LSTM	93.66±2.84	93.86±2.70	93.46±3.28	6.14±4.03
	BiLSTM	94.93±2.60	95.09±2.44	95.06±4.03	5.30±4.34
bge-m3	GRU	94.08±3.43	94.34±3.09	95.00±4.08	6.97±7.69
	RNN	95.33±3.01	95.61±2.70	97.50±3.82	7.05±6.75
	LSTM	95.74±2.75	95.89±2.65	95.83±4.17	4.47±4.47
	BiLSTM	94.09±3.34	94.34±3.17	95.00±4.08	6.97±6.53
granite-embedding-278m-multilingual	GRU	94.51±2.69	94.59±2.68	93.40±3.31	4.32±4.33
	RNN	94.08±4.29	93.99±4.54	91.79±8.21	3.64±4.45
	LSTM	94.95±2.54	95.05±2.43	94.23±3.78	4.39±4.40
	BiLSTM	94.51±2.69	94.59±2.58	92.69±4.22	3.48±4.27
gte-multilingual-base	GRU	95.38±3.46	95.42±3.53	94.29±5.28	3.33±5.53
	RNN	95.76±3.28	95.82±3.26	94.36±6.28	2.73±4.17
	LSTM	96.18±2.98	96.33±2.82	96.79±5.22	4.39±5.99
	BiLSTM	93.68±2.09	93.82±2.20	94.29±6.30	6.97±5.11
multilingual-e5-large-instruct-scripts	GRU	96.20±2.32	96.36±2.23	97.56±3.72	5.15±4.22
	RNN	94.95±3.10	95.23±2.96	97.56±3.72	7.80±5.89
	LSTM	95.36±1.23	95.42±1.52	95.06±5.49	4.39±4.40
	BiLSTM	91.52±4.28	91.91±4.01	93.40±4.98	10.53±6.72
snowflake-arctic-embed-l-v2.0	GRU	95.33±2.36	95.39±2.39	94.29±5.28	3.41±4.18
	RNN	93.64±3.47	93.90±3.25	94.29±3.74	7.05±5.39
	LSTM	94.47±3.88	94.69±3.61	94.29±3.74	5.30±5.95
	BiLSTM	94.91±3.19	95.03±3.05	94.23±3.78	4.39±5.99
text-embedding-3-small	GRU	94.47±3.35	94.73±3.15	95.13±3.98	6.21±4.08
	RNN	93.22±4.33	93.27±4.50	92.56±8.70	6.14±6.83
	LSTM	94.91±3.70	95.23±3.47	96.79±3.93	7.12±5.41
	BiLSTM	95.34±2.96	95.45±2.96	95.06±5.49	4.39±4.40

Table 2: Summary of State-of-the-Art Speech-Based AI Methods for AD Detection in the Literature using the Same Dataset.

Method	Acc ↑ (%)
Proposed Method	**97.08**
DEMENTIA Model [63]	89.58
Whisper Transfer Learning [32]	84.51
HSAF-DWAN [45]	86.71
Trill + Allosaurus + Wav2Vec 2.0 (Segment NN) [46]	78.87
Tree Bagger (wav2vec 2.0 + Transcripts) [47]	78.87
Acoustic (x-vectors, prosody) + Linguistic (ASR-BERT) [48]	84.51
SoTD + InRe (BERT features) [49]	84.64
SoTD + InRe (RoBERTa features) [49]	83.80
DSTC-Net [50]	83.10

Table 3: Fairness Results of the Proposed Method.

Metric	Value
Δ_{TPR}	0.0333 ± 0.0703
Δ_{FPR}	0.0900 ± 0.1015
OAE	0.0435 ± 0.0572
Δ_{Brier}	0.0174 ± 0.0209
Δ_{ECE}	0.0362 ± 0.0278

5.3 Feature Analysis

This section investigates the interpretability of the proposed method by analysing symbolic acoustic patterns and text embedding representations. The aim is to understand whether the learned features capture meaningful and discriminative characteristics between NC and AD speech.

Table 4: Mann-Whitney U Test Results for Embedding and SAX Features Between NC and AD Groups.

(a) Embedding Features: Raw Values and Pairwise Distance Metrics

Feature	Raw	Euclidean	Cosine	Manhattan
1	0.9358	0.4232	0.2131	0.4232
2	<0.05	0.5322	0.7987	0.5322
3	0.6054	<0.05	0.5537	<0.05
4	0.134	0.2327	0.4526	0.2327
5	0.7095	0.4365	0.6682	0.4365
6	0.5949	0.2268	0.768	0.2268
7	<0.01	0.6161	0.8916	0.6161
8	0.053	<0.01	0.3465	<0.01
9	<0.01	0.0826	0.69	0.0826
10	<0.05	0.4026	0.7137	0.4026
11	0.7885	<0.05	0.9293	<0.05
12	0.294	0.3332	0.6352	0.3332
13	<0.01	0.8638	0.69	0.8638
14	0.9887	0.1249	0.4912	0.1249
15	0.9403	0.8459	0.3359	0.8459
16	0.1684	<0.001	0.7214	<0.001
17	0.5831	<0.05	0.9598	<0.05
18	<0.05	0.5498	0.9867	0.5498
19	0.1212	0.2502	0.4956	0.2502
20	0.9705	0.7123	0.9354	0.7123
21	0.7827	0.6444	0.9231	0.6444
22	<0.001	0.6773	0.7961	0.6773
23	0.3155	0.6773	0.83	0.6773
24	<0.01	0.7653	0.7923	0.7653
25	0.0542	0.9403	0.898	0.9403
26	<0.05	<0.05	0.9231	<0.05
27	<0.001	0.5422	0.2936	0.5422
28	<0.01	0.6663	0.2172	0.6663
29	0.4454	0.0666	0.419	0.0666
30	0.081	0.8341	0.4629	0.8341
31	0.3495	0.2888	0.7334	0.2888
32	<0.001	0.6389	0.2396	0.6389

(b) SAX Symbolic Statistics: Coverage, Duration, and Occurrence

Metric	Symbolic	p-value
Coverage	a	0.48
	b	0.52
	c	0.61
	d	< 0.05
	e	< 0.05
	f	0.99
	g	0.16
	h	0.10
Duration	a	0.27
	b	0.98
	c	0.86
	d	0.25
	e	< 0.05
	f	0.57
	g	0.52
	h	0.51
Occurrence	a	0.46
	b	0.48
	c	0.49
	d	< 0.05
	e	< 0.05
	f	0.77
	g	0.16
	h	0.09

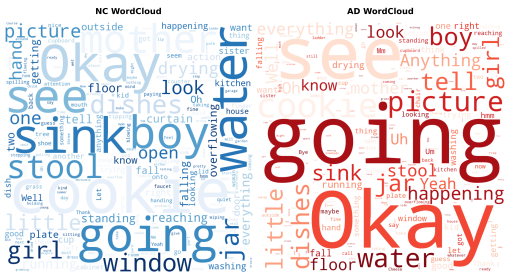

Fig. 4: Word Clouds for NC and AD Groups.

Firstly, the SAX-based pattern analysis reveals significant group-level differences in symbolic dynamics, particularly in terms of coverage, mean duration, and occurrence (see Table 4b and Figure 3). Several symbolic states exhibit statistically significant differences ($p < 0.05$) with small-to-medium effect sizes, indicating systematic shifts in temporal distribution and persistence patterns

between groups. Specifically, differences in coverage reflect changes in the proportion of time spent in particular symbolic states, while alterations in mean duration and occurrence suggest variations in state stability and transition frequency. Collectively, these results demonstrate that group differences are expressed not only in overall state frequency but also in the temporal structure of symbolic sequences. [6] These findings suggest that AD speech exhibits altered temporal organisation and symbolic state usage compared to NC speech.

Secondly, feature-level analyses further substantiate these differences. Several text embedding dimensions show statistically significant group separation under the Mann-Whitney U test (Table 4a). In addition, distance-based comparisons across Euclidean, cosine, and Manhattan metrics (Figure 2) consistently indicate separability between NC and AD samples in the embedding space. Figure 4 illustrates the most frequently occurring words in NC and AD speech. Notably, AD speech exhibits more disfluencies and conversational fillers (e.g., *uh, um, yeah*), whereas NC speech reflects richer and more descriptive vocabulary (e.g., *children, kitchen, curtains*).

5.4 Limitations and Conclusions

While the proposed framework achieves state-of-the-art performance, several limitations remain to be addressed in future work. First, the evaluation is constrained by a single, English-speaking baseline dataset, necessitating broader validation across diverse multilingual populations, regional dialects, and varied cultural contexts to ensure global generalization. Second, the algorithmic fairness assessment focused primarily on biological sex. Future research must extend this evaluation to encompass other critical demographic factors, such as age, educational attainment, and ethnicity, to mitigate systemic bias and ensure equitable clinical deployment across populations. Third, relying on a pretrained ASR model like Whisper-large-v3 introduces an operational dependency; high background noise, overlapping speech, or non-standard accents can cause transcription degradation, which cascades into downstream text embedding errors. Finally, from a methodological perspective, the current dataset captures only cross-sectional, single-session descriptive tasks, limiting the model's ability to assess long-term temporal dynamics. Incorporating longitudinal data will be essential to track continuous cognitive decline and disease progression, while integrating complementary modalities-such as acoustic prosody, self-supervised speech representations, or facial dynamics-presents a promising avenue to enhance overall diagnostic robustness.

In conclusion, this study presents a non-invasive, cost-effective speech-based Alzheimer's disease screening solution tailored for low-resource communities. By

[6] Coverage reflects the proportion of time that symbolic states are expressed within the sequence, highlighting differences in how long NC and AD subjects maintain specific states. Duration captures the average length of consecutive symbolic segments, indicating altered temporal stability of states between groups. Occurrence measures the frequency with which states reappear, reflecting distinctive repetition patterns in NC versus AD sequences.

pairing pre-trained transcript embeddings with SAX-discretized acoustic features in an ensemble GRU-XGBoost pipeline, the method achieves a state-of-the-art classification accuracy of 97.08%, outperforming existing benchmarks in the literature. Beyond raw predictive performance, statistical analysis and distance-based metrics confirm that the model extracts transparent, clinically meaningful linguistic and temporal biomarkers while maintaining calibration and minimal demographic fairness gaps. This balance of high performance, equity, and interpretability may be a vital requirement for trustworthy medical AI deployment.

References

1. World Health Organization, "Dementia," *World Health Organization*, 2023, accessed: 2025-04-19. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia
2. A. Association, "2023 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 19, no. 4, pp. 1598–1695, 2023.
3. Q.-T. Nguyen, "Standardising number of eeg sensors for ai-driven dementia detection," *IEEE Sensors Letters*, 2025.
4. M. Wang, W. Shao, S. Huang, and D. Zhang, "Identifying longitudinal intermediate phenotypes between genotypes and clinical score via exclusive relationship-induced association analysis in alzheimer's disease," *IEEE Transactions on Cognitive and Developmental Systems*, 2024.
5. Q.-T. Nguyen, L. Le, D. Williams-King, Q.-H. Tang, and N. Duong-Trung, "Explainable ai for dementia detection using eeg microstates," in *Companion Proceedings of the 31st International Conference on Intelligent User Interfaces*, 2026, pp. 39–42.
6. O. Alewijn *et al.*, "Prevalence of alzheimer's disease and vascular dementia: association with education. the rotterdam study," *Bmj*, vol. 310, no. 6985, pp. 970–973, 1995.
7. W. M. van der Flier, M. E. de Vugt, E. M. Smets, M. Blom, and C. E. Teunissen, "Towards a future where alzheimer's disease pathology is stopped before the onset of dementia," *Nature aging*, vol. 3, no. 5, pp. 494–505, 2023.
8. B. Dubois, A. Padovani, P. Scheltens, A. Rossi, and G. Dell'Agnello, "Timely diagnosis for alzheimer's disease: a literature review on benefits and challenges," *Journal of Alzheimer's disease*, vol. 49, no. 3, pp. 617–631, 2016.
9. E. W. S *et al.*, "Early recognition and treatment of neuropsychiatric symptoms to improve quality of life in early alzheimer's disease: Protocol of the beat-it study," *Alzheimer's research & therapy*, vol. 11, pp. 1–12, 2019.
10. H. Givian and J.-P. Calbimonte, "A review on machine learning approaches for diagnosis of alzheimer's disease and mild cognitive impairment based on brain mri," *IEEE Access*, 2024.
11. Z. Dong, R. Li, Y. Wu, T. T. Nguyen, J. Chong, F. Ji, N. Tong, C. Chen, and J. H. Zhou, "Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 86 048–86 073, 2024.
12. Z. Ou, C. Jiang, Y. Pan, Y. Zhang, Z. Cui, and D. Shen, "A prior-information-guided residual diffusion model for multi-modal pet synthesis from mri," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, 2024, pp. 4769–4777.

13. V. S. Rallabandi and K. Seetharaman, "Deep learning-based classification of healthy aging controls, mild cognitive impairment and alzheimer's disease using fusion of mri-pet imaging," *Biomedical Signal Processing and Control*, vol. 80, p. 104312, 2023.
14. C. Van Zyl, M. Badenhorst, S. Hanekom, and M. Heine, "Unravelling 'low-resource settings': a systematic scoping review with qualitative content analysis," *BMJ global health*, vol. 6, no. 6, 2021.
15. S. Murthy and N. K. Adhikari, "Global health care of the critically ill in low-resource settings," *Annals of the American Thoracic Society*, vol. 10, no. 5, pp. 509–513, 2013.
16. A. T. Adebisi, H.-W. Lee, and K. C. Veluvolu, "Eeg-based brain functional network analysis for differential identification of dementia-related disorders and their onset," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024.
17. Gogishvili and others., "Discovery of novel csf biomarkers to predict progression in dementia using machine learning," *Scientific reports*, vol. 13, no. 1, p. 6531, 2023.
18. A. König, R. Zeghari, R. Guerchouche, *et al.*, "Remote cognitive assessment of older adults in rural areas by telemedicine and automatic speech and video analysis: protocol for a cross-over feasibility study," *BMJ open*, vol. 11, no. 9, p. e047083, 2021.
19. S. Zhao, L. Zhang, I. Cohen, and Y. Yang, "Multimodal sensor fusion of bone and air-conducted speech for robust emotion recognition," *IEEE Sensors Journal*, 2025.
20. H. Surapaneni, A. T. Chalackal, and S. Kanakambaran, "Spectral similarity estimation and ml based classification of speech using pdms embedded fbg sensor-based contact microphone," *IEEE Sensors Journal*, 2024.
21. Z. Han, Y. Shang, Z. Shao, J. Liu, G. Guo, T. Liu, H. Ding, and Q. Hu, "Spatial–temporal feature network for speech-based depression recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 16, no. 1, pp. 308–318, 2023.
22. J. Wang, K. Lv, C. Liu, X. Nie, D. Gowda, and S. Luan, "Automatic assessment for severe self-reported depressive symptoms using speech cues," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 4, pp. 875–884, 2020.
23. U. Petti and A. Korhonen, "Losst-ad: A longitudinal corpus for tracking alzheimer's disease related changes in spontaneous speech," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 10 813–10 821.
24. J. Appell, A. Kertesz, and M. Fisman, "A study of language functioning in alzheimer patients," *Brain and language*, vol. 17, no. 1, pp. 73–91, 1982.
25. S. Luz, F. Haider, D. Fromm, B. MacWhinney *et al.*, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," in *INTERSPEECH 2020*, 2020, pp. 2172–2176.
26. K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.
27. M. Rohanian, J. Hough, and M. Purver, "Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs," in *INTERSPEECH 2021*, 2021, pp. 3820–3824.
28. F. Haider, S. De La Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2019.
29. L. Pappa, P. Karvelis, and C. Stylios, "Exploring the diverse world of sax-based methodologies," *Data Mining and Knowledge Discovery*, vol. 39, no. 1, pp. 1–78, 2025.

30. S. Cohen, O. Katz, D. Presil, O. Arbili, and L. Rokach, "Ensemble learning for alcoholism classification using eeg signals," *IEEE Sensors Journal*, vol. 23, no. 15, pp. 17 714–17 724, 2023.
31. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning (ICML)*. PMLR, 2023, pp. 28 492–28 518.
32. J. Li and W.-Q. Zhang, "Whisper-based transfer learning for alzheimer disease classification: Leveraging speech segments with full transcripts as prompts," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 211–11 215.
33. OpenAI, "Whisper large v3," https://huggingface.co/openai/whisper-large-v3, 2023, accessed: 2025-04-28.
34. OrdalieTech, "Solon-embeddings-large-0.1," https://huggingface.co/OrdalieTech/Solon-embeddings-large-0.1, 2024, mIT License.
35. B. Icard, E. Zve, L. Sainero, A. Breton, and J.-G. Ganascia, "Embedding style beyond topics: Analyzing dispersion effects across different language models," *arXiv preprint arXiv:2501.00828*, 2025.
36. J. A. Mong and D. M. Cusmano, "Sex differences in sleep: impact of biological sex and sex steroids," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 371, no. 1688, p. 20150110, 2016.
37. X. Gu, W. Zeng, and Y. Wang, "Elucidate gender fairness in singing voice transcription," in *Proceedings of the 31st ACM International Conference on Multimedia (MM)*, 2023, pp. 8760–8769.
38. T. Garg, S. Masud, T. Suresh, and T. Chakraborty, "Handling bias in toxic speech detection: A survey," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–32, 2023.
39. A. Caliskan, P. P. Ajay, T. Charlesworth, R. Wolfe, and M. R. Banaji, "Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2022, pp. 156–170.
40. S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The adresso challenge," in *INTERSPEECH 2021*. ISCA, 2021.
41. Y. Wang, A. Favaro, T. Thebaud, J. Villalba, N. Dehak, and L. Moro-Velazquez, "Exploring the complementary nature of speech and eye movements for profiling neurological disorders," in *INTERSPEECH*, 2024, pp. 1475–1479.
42. H.-B. Kim, S. Lee, B.-I. Ham, K.-H. Choi, and K.-S. Kim, "Temperature compensation method for a six-axis force/torque sensor using a gated recurrent unit," *IEEE Sensors Journal*, 2025.
43. N. Mohammadi Foumani, L. Miller, C. W. Tan, G. I. Webb, G. Forestier, and M. Salehi, "Deep learning for time series classification and extrinsic regression: A current survey," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–45, 2024.
44. Z. Lu, S. Chen, J. Yang, C. Liu, and H. Zhao, "Prediction of lower limb joint angles from surface electromyography using xgboost," *Expert Systems with Applications*, vol. 264, p. 125930, 2025.
45. N. Wang, M. Wu, W. Gu, and Z. Chai, "Hybrid self-aligned fusion with dual-weight attention network for alzheimer's detection," *IEEE Signal Processing Letters*, 2024.
46. L. Gauder, L. Pepino, L. Ferrer, and P. Riera, "Alzheimer disease recognition using speech-based embeddings from pre-trained models," in *INTERSPEECH 2021*, 2021, pp. 3795–3799.
47. Y. Pan, M. Lu, Y. Shi, and H. Zhang, "A path signature approach for speech-based dementia detection," *IEEE Signal Processing Letters*, 2023.

48. R. Pappagari, J. Cho, S. Joshi, L. Moro-Velázquez, P. Zelasko, J. Villalba, and N. Dehak, "Automatic detection and assessment of alzheimer disease using speech and language technologies in low-resource scenarios." in *INTERSPEECH*, vol. 2021, 2021, pp. 3825–3829.
49. J. Kang, D. Han, L. Meng, J. Zhou, J. Li, X. Wu, and H. Meng, "Towards within-class variation in alzheimer's disease detection from spontaneous speech," *arXiv preprint arXiv:2409.16322*, 2024.
50. Y. Gao, L. Guo, and H. Liu, "A dual-stage time-context network for speech-based alzheimer's disease detection," *arXiv preprint arXiv:2502.13064*, 2025.
51. B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, "Trustworthy ai: From principles to practices," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–46, 2023.
52. A. H. Sfayyih, N. Sulaiman, and A. H. Sabry, "A review on lung disease recognition by acoustic signal analysis with deep learning networks," *Journal of big Data*, vol. 10, no. 1, p. 101, 2023.
53. S. Cohen, R. Perez, and L. Kishon-Rabin, "Auditory sequence learning with degraded input: children with cochlear implants ('nature effect') compared to children from low and high socio-economic backgrounds ('nurture effect')," *Scientific Reports*, vol. 15, no. 1, p. 7872, 2025.
54. D. O'Shaughnessy, "Speech enhancement—a review of modern methods," *IEEE Transactions on Human-Machine Systems*, vol. 54, no. 1, pp. 110–120, 2024.
55. S. Caton and C. Haas, "Fairness in machine learning: A survey," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–38, 2024.
56. C. Long, H. Hsu, W. Alghamdi, and F. Calmon, "Individual arbitrariness and group fairness," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
57. D. Wei, K. N. Ramamurthy, and F. P. Calmon, "Optimized score transformation for fair classification," *Proceedings of Machine Learning Research*, vol. 108, 2020.
58. Q.-T. Nguyen, L. Le, X.-T. Tran, T. Do, and C.-T. Lin, "Fairad-xai: Evaluation framework for explainable ai methods in alzheimer's disease detection with fairness-in-the-loop," in *Proceeding of The 2024 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2024, pp. 870–876.
59. M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.
60. D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–44, 2022.
61. M. Darrin, P. Formont, I. Ayed, J. C. Cheung, and P. Piantanida, "When is an embedding model more promising than another?" *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 68 330–68 379, 2024.
62. K. Enevoldsen, M. Kardos, N. Muennighoff, and K. L. Nielbo, "The scandinavian embedding benchmarks: Comprehensive assessment of multilingual and monolingual text embedding," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 40 336–40 358, 2024.
63. Z. Zhang, T. Wang, Z. Hu, L.-Z. Yang, and H. Li, "Dementia: A hybrid attention-based multimodal and multi-task learning framework with expert knowledge for alzheimer's disease assessment from speech," *IEEE Journal of Biomedical and Health Informatics*, 2024.
64. M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 29, no. 1, 2015.