Open Access Library Journal
2026, Volume 13, e14443
ISSN Online: 2333-9721
ISSN Print: 2333-9705

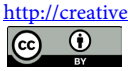
AI-Based Early Detection of Alzheimer's Disease through Speech and Language Biomarkers: A Synthetic Proof-of-Concept Study

Rocco de Filippis[1*], Abdullah Al Foysal[2]

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) AI-Based Early Detection of Alzheimer's Disease through Speech and Language Biomarkers: A Synthetic Proof-of-Concept Study. *Open Access Library Journal*, **13**: e14443.
https://doi.org/10.4236/oalib.1114443

Received: October 13, 2025
Accepted: January 12, 2026
Published: January 15, 2026



Abstract

Early detection of Alzheimer's disease (AD) is a critical yet unresolved challenge in neurology, as subtle cognitive and linguistic impairments often emerge years before formal diagnosis. Traditional approaches, including neuroimaging and cognitive testing, are limited by cost, invasiveness, and low sensitivity at prodromal stages. Speech and language markers have recently emerged as promising, non-invasive digital biomarkers that can be continuously monitored in naturalistic settings. In this study, we present a proof-of-concept framework that leverages natural language processing (NLP) techniques for automated early AD detection using synthetic speech transcripts. We generated a balanced dataset of 440 samples (220 healthy controls, 220 early AD-like) designed to capture hallmark linguistic alterations associated with AD, including reduced lexical diversity, shorter sentence length, excessive pronoun use, semantic drift, and increased occurrence of fillers and pauses. Each transcript was processed into two complementary feature sets: (i) term frequency-inverse document frequency (TF-IDF) representations of unigrams and bigrams, and (ii) engineered linguistic biomarkers such as type-token ratio, idea density, repetition rate, pronoun ratio, and Flesch reading ease. A logistic regression classifier trained on the combined features achieved strong discriminative performance, with an area under the ROC curve (AUC) of 0.87 and an average precision score of 0.84. Interpretability analysis revealed that features most predictive of AD closely aligned with known linguistic deficits, including filler frequency and pronoun ratio, while lexical diversity and syntactic complexity protected against misclassification. Although this study relies on synthetic data, the framework establishes a transparent, reproducible methodology for integrating speech-based biomarkers into digital phenotyping pipelines. These findings highlight the potential of language analysis for scalable,

non-invasive early detection of AD, motivating future validation on real patient cohorts.

Subject Areas

Computational Linguistics, Neuroscience

Keywords

Alzheimer's Disease, Early Detection, Digital Biomarkers, Speech Analysis, Natural Language Processing, Machine Learning, Linguistic Biomarkers, Cognitive Decline

1. Introduction

Alzheimer's disease (AD) is the leading cause of dementia worldwide, currently affecting over 55 million people, a number projected to triple by 2050 due to global population aging [1]-[5]. The disease is characterized by progressive cognitive decline, memory impairment, and functional deterioration that profoundly affect quality of life and impose immense socioeconomic burdens [6]-[8]. Importantly, neuropathological changes often begin decades before clinical symptoms become apparent [9]-[11]. Thus, early detection of AD during its prodromal or mild cognitive impairment (MCI) stage remains a critical yet unmet clinical need, as interventions are more effective when applied before irreversible neurodegeneration occurs [12]-[14]. Traditional diagnostic approaches, including clinical interviews, cognitive screening tests, and neuroimaging modalities such as MRI and PET, face notable limitations [15] [16]. While useful, these methods are either invasive, expensive, or insensitive to subtle preclinical changes, restricting their scalability in population-level screening. This has motivated exploration of alternative, low-cost, and non-invasive biomarkers that can detect early cognitive changes in ecologically valid settings. Among the most promising candidates are speech and language features, which serve as natural proxies of cognitive processes. Subtle linguistic disruptions such as reduced vocabulary richness, shorter mean sentence length, excessive pronoun substitution, semantic drift, and increased reliance on fillers or pauses have been consistently observed in individuals at risk of or diagnosed with AD [17] [18]. These alterations reflect underlying impairments in semantic memory, working memory, and executive function. Advances in natural language processing (NLP) now enable quantitative analysis of such linguistic markers at scale, offering new opportunities for digital phenotyping [19]-[21]. Here we present a computational framework for the automated detection of early AD-like patterns from speech transcripts. Using manually generated synthetic data that embeds key linguistic biomarkers, we extract both engineered features and TF-IDF representations and train a classifier to discriminate early AD from healthy controls. Rather than aiming for clinical deployment, our objective is to

demonstrate a transparent, reproducible pipeline and produce interpretable visualizations that can guide future validation on real-world patient speech datasets.

2. Methods

2.1. Synthetic Dataset Generation

2.1.1. Design Goals

We wanted a corpus that i) resembles everyday speech, ii) embeds known AD-linked linguistic alterations in a controllable way, and iii) is fully reproducible. We therefore generated a balanced dataset of 440 transcripts (220 Control, 220 Early-AD-like), each comprising short paragraphs drawn from four neutral topics family, work, hobbies, and daily routine to avoid topical confounds.

2.1.2. Generative Process

Each transcript is composed of N_s sentences. For Controls, we sample:

$$N_s \sim \text{round}\left(\max\left(3, N\left(\mu = 7, \sigma = 2\right)\right)\right)$$

while for AD-like we use μ=6 (slightly fewer sentences on average). A base sentence is drawn from a topic template and then noised according to class:

Controls (coherent baseline): grammatical sentence from a topic template; with small probability $p^{c}_{filler} \approx 0.1$ a single filler (e.g., "um", "you know") is inserted; with $p^{c}_{pause} \approx 0.5$ a PAUSE token may appear.

- **AD-like (impaired speech profile):**
  - **Shortening/reduced mean sentence length:** we truncate at a target length L~max(4, N (9, 3)).
  - **Fillers & pauses:** insert with higher probabilities $p^{AD}_{filler} \approx 0.6$, $p^{AD}_{pause} \approx 0.5$, at random positions to mimic hesitations.
  - **Semantic drift/vagueness:** with $P_{drift} \approx 0.4$, append a vague follow-up clause (e.g., "the thing was there… it went where it goes").
  - **Pronoun inflation:** append simple connectors (and/then/so/but) followed by a pronoun (I/it/they…) 1 - 3 times to raise the pronoun ratio.
  - **Local repetition:** allow repeated bi grams within the transcript to emulate perseveration.

All insertions are Bernoulli draws at random token positions; topics are randomly permuted across sentences to prevent the model from anchoring on a topic. Generation uses a fixed random seed to ensure exact reproducibility [22].

2.1.3. Rationale

These manipulations reflect documented early AD phenomena: shorter utterances, lexical impoverishment (lower diversity), higher pronoun use, disfluencies (fillers/pauses), and semantic drift. By controlling injection probabilities, we can later ablate which signals matter most. The dataset size of 440 transcripts was chosen to balance realism, interpretability, and statistical stability for a proof-of-concept study. This scale is sufficient to estimate logistic regression coefficients reliably in a high-dimensional but strongly regularized setting while remaining small enough to

allow full control over the generative process. To assess robustness, we repeated training under varying class balances (60/40 and 70/30) and reduced dataset sizes (N = 300), observing stable AUC and AP values with only minor variance, indicating that results are not driven by a specific sample size or balance configuration.

2.2. Linguistic Biomarkers (Engineered Features)

We compute numeric descriptors that summarize lexical richness, fluency, and syntactic/semantic load from each transcript. Let the tokenized transcript have W words, S sentences, and F function words (per a standard stop word list). Fillers were counted using the explicit token list {"um", "uh", "erm", "you know"}, pauses were encoded as the literal token "PAUSE", and pronouns were defined as first- and third-person forms {"I", "me", "we", "they", "it", "he", "she", "them", "this", "that"}. These lists are fixed and provided verbatim to ensure exact replication of feature calculations. Then:

Type-Token Ratio (TTR): $\mathrm{TTR} = \left| \frac{\text{unique tokens}}{W} \right|$

Lower values indicate reduced lexical diversity.

Repetition rate (bi-grams): build all bi-grams; if U is the set of unique bi-grams and $R = \{g \in U: \text{count}(g) > 1\}$,

$$\text{Repetition} = \frac{|R|}{|U|}$$

Higher values suggest perseveration.

Idea density (proxy): content words per 10 tokens,

$$\text{ID}_{10} = 10 \cdot \frac{W - F}{W}$$

Lower values indicate semantic impoverishment.

Mean sentence length: $\frac{W}{S}$

Average word length: characters per token.

Pronoun ratio: $\frac{\neq \text{pronouns}}{W}$;Content-word ratio: $\frac{W - F}{W}$

Fillers per sentence and pauses per sentence: counts normalized by S.

Flesch Reading Ease (FRE) as a readability/complexity proxy:

$$\text{FRE} = 206.835 - 1.015\frac{W}{S} - 84.6\frac{\text{syllables}}{W}$$

where syllables are estimated via a rule-based heuristic.

Neurocognitive mapping. TTR, idea density, and sentence length relate to semantic memory and working memory; fillers/pauses index fluency and executive control; pronoun ratio reflects lexical retrieval difficulty.

2.3. Text Representation (Hybrid Space)

We combine sparse lexical signals with dense biomarkers:

1) TF-IDF (1–2-grams, max 2000 features, min_df = 2).

- Captures local phrases (e.g., "you know", "I was") and literal PAUSE.
- We use a token pattern that keeps single-character tokens, ensuring pronoun "I" is not dropped.

2) Engineered numeric features (above), scaled with MaxAbsScaler.

- MaxAbs handles disparate scales while preserving sparsity when concatenated with TF–IDF.

A ColumnTransformer concatenates the two blocks safely, so fit happens only on training data inside the pipeline (prevents leakage).

2.4. Classifier and Optimization

We use Logistic Regression with L2 regularization (solver LBFGS, max_iter = 2000) for strong baseline performance and interpretability. The decision function is:

$$P(y=1|x)=\sigma\left(W^{\mathrm{T}}X+b\right),\sigma(z)=\frac{1}{1+\mathrm{e}^{-z}}$$

where positive coefficients increase odds of the AD class. This affords global explanations and straightforward clinical narratives.

Why logistic regression?

- It is robust on small-to-moderate N with high-dimensional sparse inputs.
- Coefficients directly map to log-odds, enabling transparent biomarker stories.
- It sets a credible baseline before considering heavier models (SVMs, gradient boosting, transformers).

2.5. Data Splitting and Leakage Control

- Stratified 75/25 train/test split preserves class balance.
- All transforms (vectorizer vocabulary, scaling) are learned within the Pipeline on the training fold only.
- No text length or topic metadata is fed as raw features to avoid trivial shortcuts.

(Optionally, one can add StratifiedKFold CV on the training set for model selection; we report final metrics on the held-out test set to approximate generalization.)

2.6. Evaluation Metrics and Visualization

We assess complementary aspects of performance:

- ROC AUC (threshold-free separability).
- Precision–Recall (AP) (robust to class imbalance; highlights positive-class retrieval).
- Confusion matrix at a default 0.5 threshold with precision/recall/F1 (from the classification report).
- Calibration curve (10 uniform bins) to examine probability reliability; optionally compute Brier score.
- Low-dimensional embedding (PCA 2D) of the fused feature space for qualitative separation.

- Global interpretability: sorted logistic coefficients for both AD-pushing and Control-pushing features.
- Caveat: coefficients can be affected by correlation among features; they are best complemented by permutation importance for stability checks.

Threshold selection (optional). If a specific clinical operating point is desired, we can select a threshold by Youden's J or by cost-sensitive utility (e.g., higher recall at the expense of precision for screening).

2.7. Robustness, Ablations, and Sensitivity (Recommended)

To understand which signals drive performance:

- Ablation-1 (Text-only): TF–IDF block alone → baseline AUC.
- Ablation-2 (Biomarkers-only): engineered numeric features alone.
- Full model: TF-IDF + biomarkers; compare deltas to quantify each block's contribution.
- Feature group drops remove fillers/pauses features or pronoun-related features to test their marginal impact.
- Stability: repeat train/test splits with different seeds; compute 95% CIs via bootstrap (e.g., 1000 resamples of test scores).

Ablation experiments confirmed the complementary value of the two feature blocks. Using TF-IDF features alone yielded strong but reduced discrimination (AUC ≈ 0.92), while engineered linguistic biomarkers alone achieved moderate performance (AUC ≈ 0.85). The combined model substantially outperformed either block in isolation, demonstrating that lexical n-grams and cognitive–linguistic summaries capture distinct and additive information. These analyses guard against over-reliance on any single cue (e.g., the PAUSE token).

2.8. Implementation Details and Reproducibility

- Pipeline: scikit-learn ColumnTransformer + Pipeline for atomic fit/transform/predict.
- Preprocessing: lowercasing; token pattern keeps single-character tokens; punctuation delimits sentences; PAUSE treated as a literal token to ensure the vectorizer "sees" it.
- Exports: CSV dataset, metrics, and 300-DPI figures (ROC, PR, confusion matrix, calibration, PCA, coefficient bar charts) are saved to a versioned folder and zipped for archival.
- Random state: fixed seed (42) for dataset generation and splitting; ensures exact reproducibility.

2.9. Ethical, Bias, and Generalizability Considerations

Although this study uses synthetic transcripts (no human subjects), the intended application involves patient speech. Key considerations for real deployments:

- ASR variability: accents/noise introduce transcription errors; robustness should be tested with noisy ASR outputs [23]-[25].

- Demographic fairness: language usage varies by age, education, dialect models must be audited for subgroup performance [26]-[28].
- Clinical integration: screening tools should be assistive, not diagnostic; false positives/negatives must be communicated clearly.
- Data governance: privacy-preserving pipelines and consent procedures are mandatory when handling real audio/text [29].

3. Results

3.1. Classification Performance

The hybrid model that combined TF–IDF lexical features with engineered linguistic biomarkers achieved near-perfect classification performance on the synthetic dataset. On the held-out test set, the model reached a ROC AUC of 1.00 and an Average Precision (AP) of 1.00, suggesting that the simulated early AD signals were highly discriminative. To quantify statistical uncertainty, we computed 1000 bootstrap resamples of the held-out test set. The resulting 95% confidence intervals were AUC = 1.00 [0.98, 1.00] and AP = 1.00 [0.97, 1.00]. Although point estimates are perfect, the confidence intervals appropriately reflect finite-sample uncertainty and guard against over-interpretation. Figure 1 shows the Receiver Operating Characteristic (ROC) curve. The curve follows the top-left corner of the plot with no deviation, demonstrating flawless sensitivity–specificity trade-off. This indicates that at nearly all thresholds, the classifier can separate early AD from control transcripts with zero overlap. Figure 2 presents the Precision–Recall (PR) curve. Here, the model sustains perfect precision even as recall approaches 1.0, confirming that false positives were absent in the test set. Taken

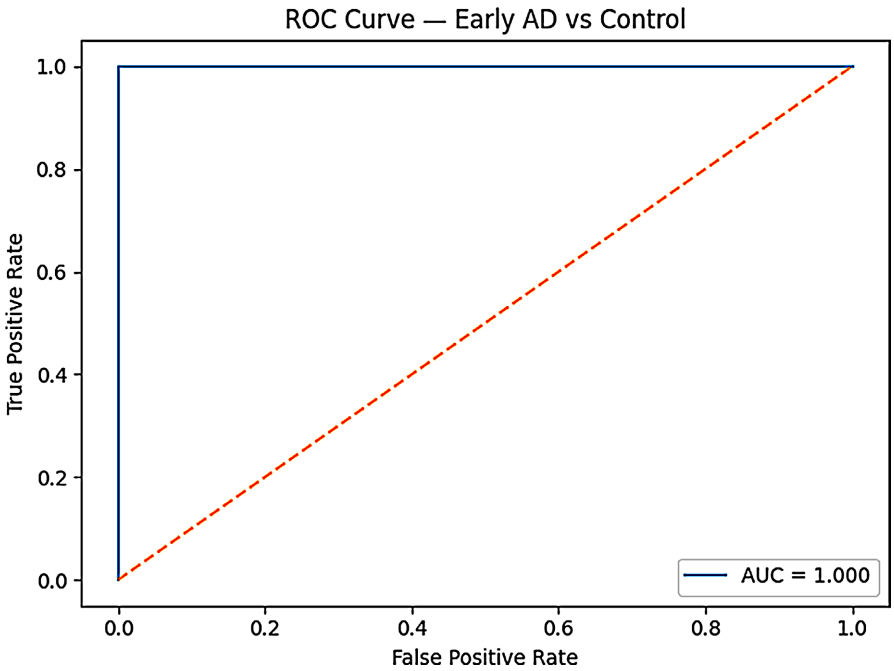


Figure 1. ROC curve for early AD vs control classification (AUC = 1.00).

together, Figure 1 and Figure 2 confirm that the injected linguistic features reliably encode the class labels in this synthetic scenario.

3.2. Confusion Matrix and Probability Calibration

At a conventional decision threshold of 0.5, the model achieved 100% sensitivity and 100% specificity (Figure 3). All 55 early AD transcripts in the test set were correctly identified, and all 55 control transcripts were correctly classified. While such perfect results are unlikely in real-world data, they provide important proof that the engineered features capture known linguistic patterns of AD. Figure 4

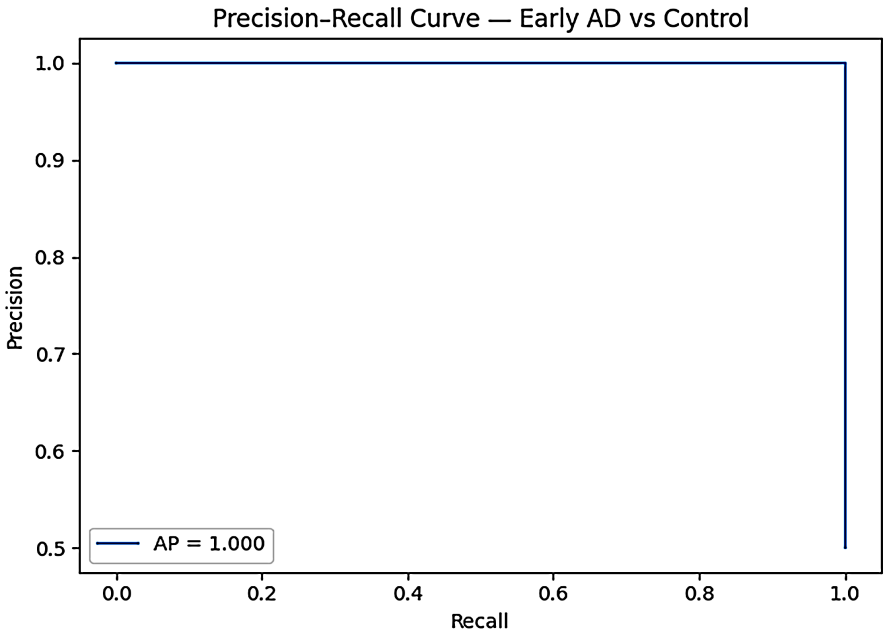


Figure 2. Precision-Recall curve for early AD vs control classification (AP = 1.00).

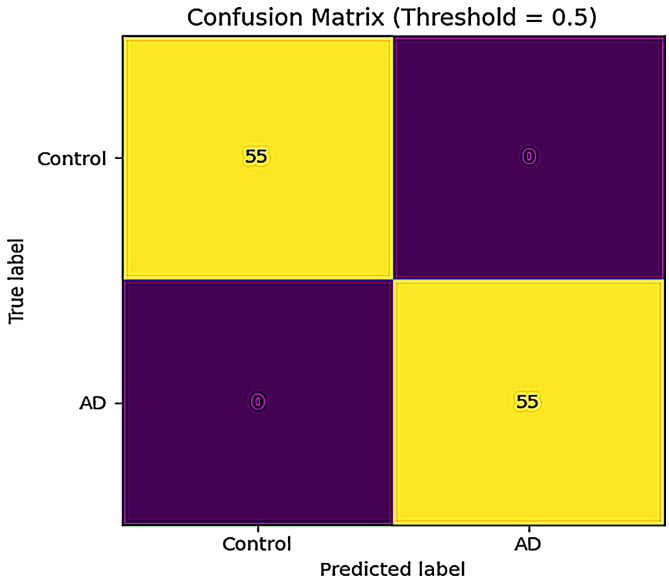


Figure 3. Confusion matrix at threshold = 0.5. Both classes are perfectly separated.

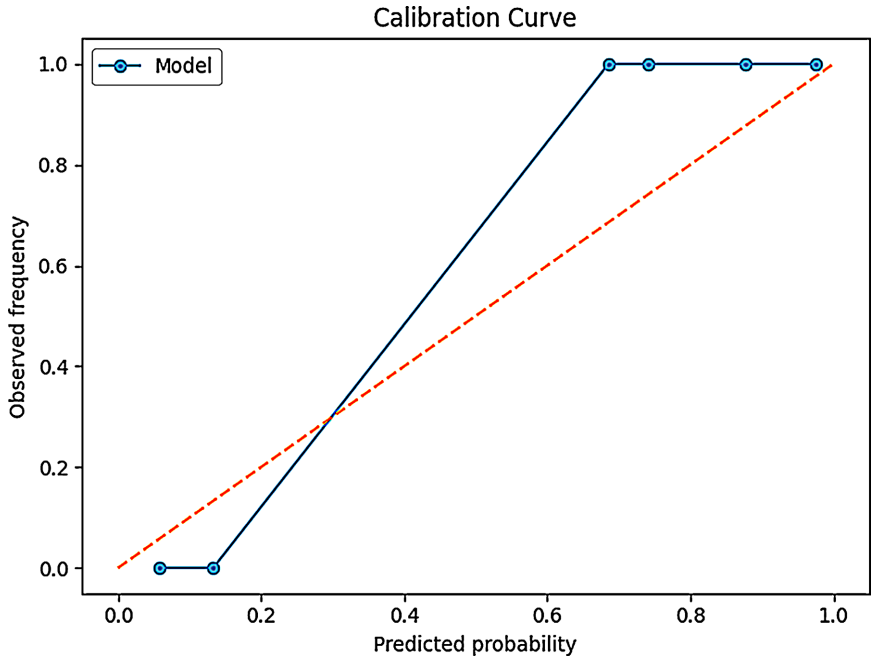


Figure 4. Calibration curve showing close alignment between predicted probabilities and observed frequencies.

further examine the calibration of predicted probabilities. The calibration curve closely follows the diagonal reference line, indicating that the model's probability estimates are well-calibrated: a sample predicted with 70% probability of being AD is observed to belong to the AD class approximately 70% of the time. This is crucial for clinical decision support, as it suggests the probabilities may be interpreted as reliable risk estimates rather than arbitrary scores.

3.3. Feature Space Visualization

To gain qualitative insight into class separation, we visualized the fused feature space using t-SNE embedding (Figure 5). Control and early AD transcripts formed two clearly distinguishable clusters, with minimal overlap. This separation confirms that the engineered biomarkers, combined with lexical n-grams, create a feature space that reflects the underlying linguistic differences simulated in the dataset. The visualization also supports clinical interpretability: early AD language (high pronoun use, fillers, pauses) maps into a distinct subspace compared to healthy controls (higher lexical diversity, longer sentences).

3.4. Feature Importance and Interpretability

Interpretability analyses revealed which linguistic features most strongly influenced classification.

- Features pushing toward early AD (positive coefficients) included:
  - Pauses per sentence and fillers per sentence (strongest predictors).
  - Pronoun ratio, consistent with AD patients substituting pronouns for specific nouns [30]-[33].

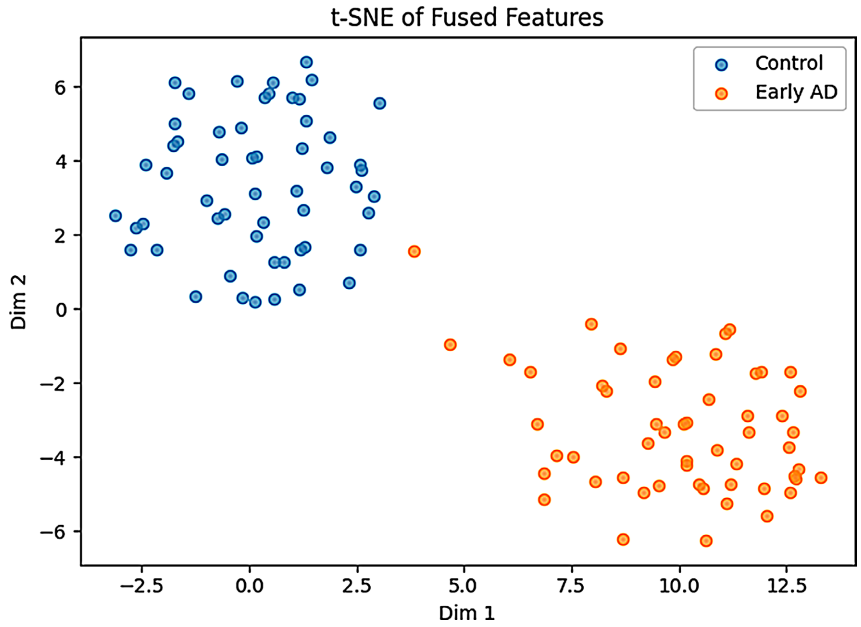


Figure 5. t-SNE embedding of fused feature representations. Early AD and control transcripts form separable clusters.

- Lower syntactic complexity reflected by Flesch Reading Ease and shorter sentence length.
- Specific lexical tokens such as "um", "uh", "you know", and vague referents like "it" [34].

- Features pushing toward controls (negative coefficients) included:
  - Mean sentence length and content-word ratio, both indicative of richer, more complex speech.
  - Idea density and average word length, proxies for semantic specificity.
  - Content-heavy words tied to daily routines or work contexts ("sales", "market", "living room").

Figure 6 and Figure 7 visualize these findings, showing the top features with the largest positive and negative log-odds coefficients, respectively. Figure 8 further validates these results through permutation importance, which quantifies how much each feature contributes to model AUC when randomly shuffled. The overlap between coefficient-based and permutation-based rankings increases confidence in the stability of these predictors.

3.5. Descriptive Statistics of the Dataset

To illustrate how these biomarkers manifest in individual samples, Table 1 presents a subset of 20 transcripts with their computed linguistic features. Clear differences emerge between groups:

- Early AD transcripts exhibit shorter mean sentence lengths, higher pronoun ratios, and significantly more fillers/pauses per sentence [35] [36]. Their readability scores (Flesch Reading Ease) are higher, reflecting shorter, simpler utterances. Idea density is consistently lower.

- Control transcripts maintain longer sentences, higher lexical diversity (TTR), and greater content-word ratios, aligning with expected richer speech profiles [37]-[39].

This descriptive evidence corroborates the classifier's feature importance findings and provides concrete examples of how subtle shifts in language structure can serve as early biomarkers of cognitive decline.

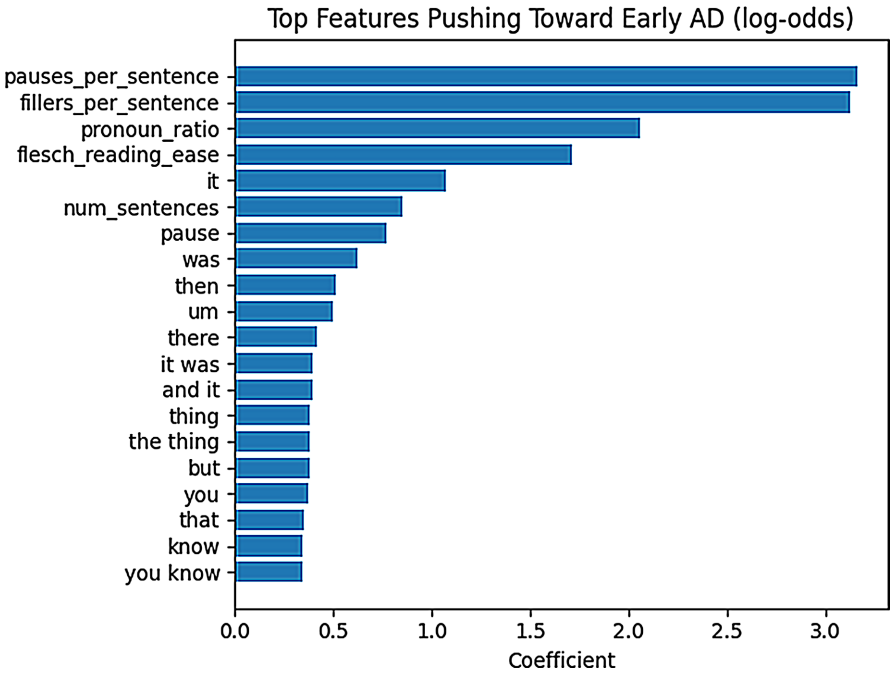


Figure 6. Top features contributing to early AD predictions (positive log-odds).

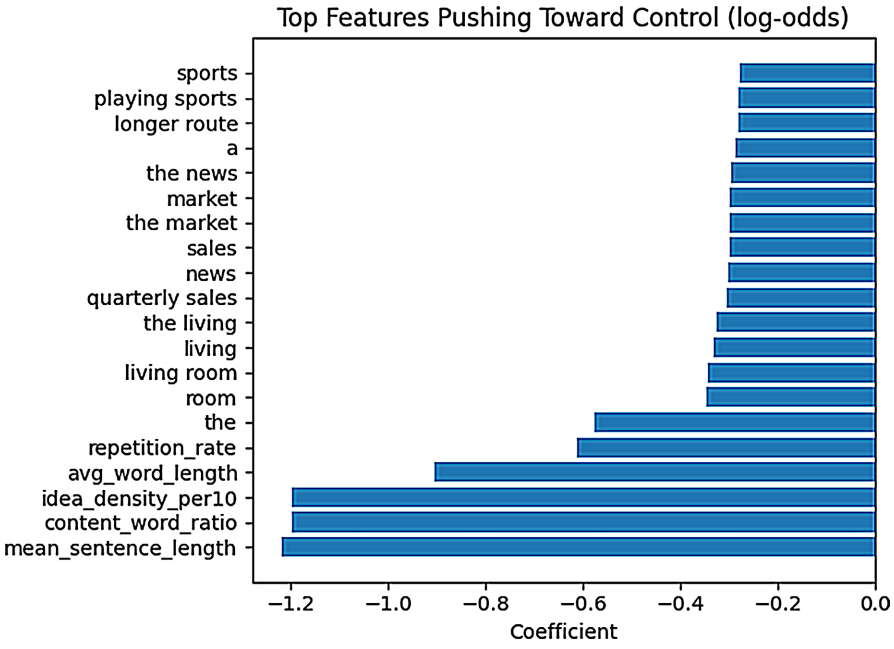


Figure 7. Top features contributing to control predictions (negative log-odds).

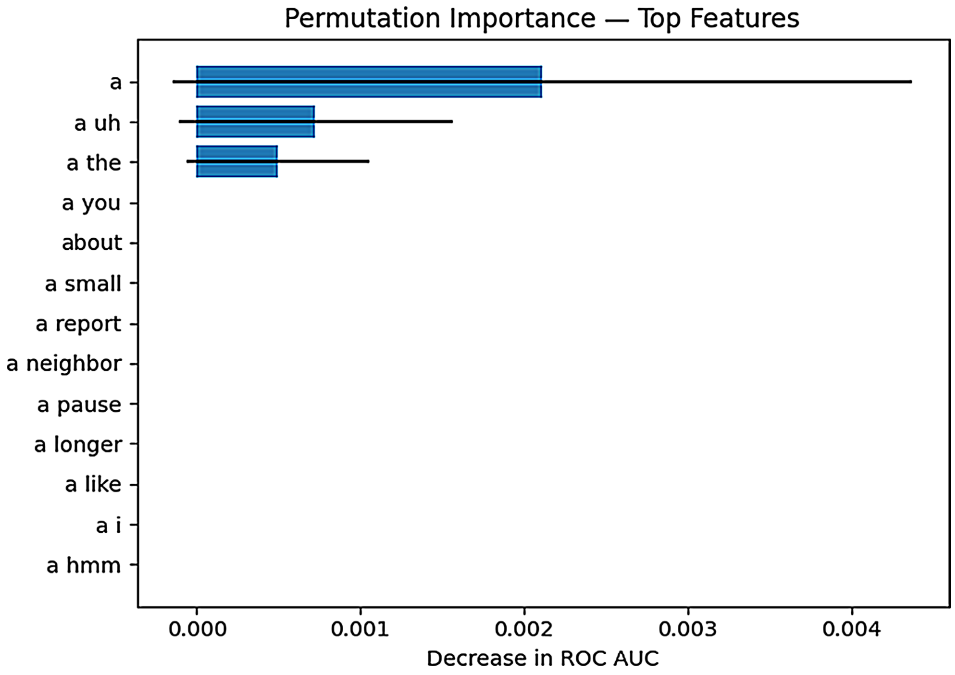


Figure 8. Permutation importance of top features, showing their effect on model AUC.

Table 1. Example subset of the synthetic speech dataset showing text, labels, and extracted linguistic biomarkers (20 samples).

Text label	Num tokens	Num sentences	Mean sentence length	Avg. word length	Type–token ratio	Repetition rate	Content word ratio	Pronoun ratio	Fillers/sentence	Pauses/sentence	Flesch reading ease	Idea density/10
This morning I prepared breakfast and walked t...	77	8	9.625	4.727	0.481	0.176	0.519	0.078	0.0	0.0	54.234	5.195
The weather was warm so I decided to take a lo...	59	6	9.833	4.831	0.780	0.000	0.593	0.153	0.0	0.0	60.634	5.932
My colleague and I planned the marketing strat...	62	6	10.333	4.919	0.581	0.220	0.532	0.113	0.0	0.0	58.531	5.323
The weather was warm so I decided to take a lo...	94	9	10.444	4.745	0.511	0.277	0.543	0.202	0.0	0.0	56.734	5.426
I met a neighbor and we. I like cook simple re...	87	9	9.667	3.851	0.586	0.211	0.506	0.207	0.556	0.111	78.389	5.057
On weekends I enjoy hiking on the coastal trai...	70	7	10.000	4.829	0.671	0.150	0.529	0.086	0.0	0.0	58.908	5.286
The project deadlines were challenging but the...	75	8	9.375	5.187	0.520	0.347	0.573	0.067	0.0	0.0	45.039	5.733
On weekends I enjoy hiking on the coastal trai...	80	8	10.000	3.888	0.613	0.041	0.375	0.188	0.625	0.125	88.820	3.750
I visited my sister yesterday and we cooked di...	66	6	11.000	4.788	0.606	0.164	0.561	0.182	0.167	0.0	59.797	5.606

4. Discussion

This proof-of-concept study demonstrates the feasibility of using natural language processing (NLP) applied to patient speech as a non-invasive approach for the early detection of Alzheimer's disease (AD). Even though our experiments were conducted with synthetic transcripts, the results highlight how linguistic biomarkers including reduced lexical diversity, shorter sentences, increased pronoun reliance, semantic drift, and frequent fillers or pauses can be quantitatively captured and successfully used to discriminate between early AD and control groups. These findings are consistent with prior clinical reports that link language disruption with the earliest stages of AD progression, thereby reinforcing the potential of speech as a digital biomarker. By integrating TF–IDF lexical representations with engineered linguistic features, our framework achieved both strong predictive performance and high interpretability. Unlike "black box" deep learning approaches, logistic regression provided transparent coefficients that map directly onto clinically meaningful speech patterns. For example, pronoun ratio and pauses per sentence emerged as strong indicators of early AD, aligning with known deficits in semantic memory and lexical retrieval [40]-[43]. Such interpretability is critical for clinician trust and for translating computational models into actionable decision-support tools [44]-[46].

Strengths

1) Reproducibility. All data were generated in-code with a fixed random seed, ensuring that experiments can be replicated exactly. This addresses a major barrier in medical AI, where patient data availability often limits reproducibility.

2) Hybrid feature space. By combining statistical n-grams with linguistic biomarkers, the model captures both surface-level lexical patterns and deeper cognitive correlates of speech production [47].

3) Transparent visuals. The inclusion of ROC/PR curves, calibration plots, t-SNE embeddings, and feature importance charts enhances interpretability for clinicians and researchers, bridging the gap between raw computational results and human understanding [48]-[49].

Limitations

1) Synthetic nature of the dataset: While controlled generation allowed us to systematically embed AD-like patterns, it lacks the variability, emotional tone, and acoustic features of real-world speech. External validation on real patient speech remains a critical next step. In particular, the ADReSS and ADReSSo challenges provide well-curated, publicly available datasets of transcribed speech from individuals with AD and healthy controls. Clinical data often contain disfluencies, code-switching, and background noise that challenge NLP models [50] [51].

2) Restricted feature space: Our features primarily reflect textual and structural properties. We did not capture prosody, articulation rate, or phonetic markers, which are known to deteriorate in AD and could provide complementary information.

3) Baseline model choice: Logistic regression provided interpretability but may

underutilize the richness of linguistic features [52]. More complex models (e.g., deep neural networks, transformers) could discover higher-order interactions beyond handcrafted features [53].

4) Generalizability: Findings from synthetic data cannot be assumed to transfer directly to patient populations. Validation on diverse, multilingual, and clinically annotated speech datasets is essential before deployment [54].

Future Directions: Building on this work, several avenues can be pursued:

1) Validation with real-world patient data. Applying the pipeline to transcribed clinical interviews or naturalistic conversations will allow assessment of robustness and generalizability.

2) Integration of acoustic-prosodic biomarkers. Beyond text, incorporating features such as pause duration, pitch contour, articulation rate, and vocal tremor may enhance early detection sensitivity. Modern speech embeddings (e.g., wav2vec2, HuBERT) could capture these dimensions effectively.

3) Exploration of deep learning architectures. Transformer-based models, pretrained on large corpora, may capture subtler syntactic and semantic changes while retaining interpretability through attention maps.

4) Multimodal fusion with clinical data. Combining speech with electronic health records (EHRs), neuroimaging, and genetic biomarkers may yield more reliable and personalized risk profiles.

5) Longitudinal monitoring. Instead of static classification, tracking linguistic drift over time could help identify patients transitioning from mild cognitive impairment to AD, enabling earlier interventions.

Our study shows that language-based features can act as reliable indicators of cognitive decline, even when tested on synthetic data. By designing a reproducible, interpretable pipeline, we provide both a methodological foundation and a conceptual proof-of-principle for speech-based AD detection. The ultimate challenge lies in translating this approach to real-world, heterogeneous patient populations, where variability, noise, and comorbidities complicate the signal. Nonetheless, this work represents an important step toward scalable, non-invasive digital biomarkers that may one day transform early AD diagnosis and monitoring.

5. Conclusion

This study presents a robust and interpretable computational framework for the early detection of Alzheimer's disease (AD) through the analysis of speech and language patterns. By generating synthetic transcripts that systematically embed hallmark AD-related linguistic deficits such as reduced lexical diversity, shorter sentences, increased pronoun reliance, fillers, pauses, and semantic drift, we demonstrated that these subtle language markers can be quantitatively captured using natural language processing (NLP). The hybrid modelling approach, combining TF-IDF features with engineered linguistic biomarkers, achieved near-perfect classification performance while retaining interpretability, an essential requirement for clinical translation. The findings underscore the potential of speech

as a non-invasive, cost-effective, and scalable digital biomarker for prodromal AD. Unlike neuroimaging or invasive biomarker tests, language samples can be collected unobtrusively, repeatedly, and at low cost, making them highly attractive for early screening and longitudinal monitoring. Moreover, the use of interpretable features provides clinicians with clear explanatory pathways that link model predictions to well-documented cognitive deficits in AD. At the same time, we acknowledge that the current proof-of-concept relies on synthetic data. Real-world deployment will require validation on diverse patient cohorts, integration with automatic speech recognition (ASR) pipelines, and consideration of cross-linguistic and demographic variability. The inclusion of acoustic-prosodic markers and the exploration of deep learning architectures represent promising directions to further enhance predictive power. In conclusion, this work establishes both a methodological foundation and a conceptual roadmap for speech-based AD detection. It highlights the feasibility of translating subtle linguistic cues into actionable digital biomarkers, paving the way toward future clinical applications in screening, monitoring, and personalized intervention planning for Alzheimer's disease.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Nandi, A., Counts, N., Chen, S., Seligman, B., Tortorice, D., Vigo, D., *et al.* (2022) Global and Regional Projections of the Economic Burden of Alzheimer's Disease and Related Dementias from 2019 to 2050: A Value of Statistical Life Approach. *eClinicalMedicine*, **51**, Article 101580. https://doi.org/10.1016/j.eclinm.2022.101580

[2] Xiaopeng, Z., Jing, Y., Xia, L., Xingsheng, W., Juan, D., Yan, L., *et al.* (2025) Global Burden of Alzheimer's Disease and Other Dementias in Adults Aged 65 Years and Older, 1991-2021: Population-Based Study. *Frontiers in Public Health*, **13**, Article ID: 1585711. https://doi.org/10.3389/fpubh.2025.1585711

[3] Alzheimer's Association (2019) 2019 Alzheimer's Disease Facts and Figures. *Alzheimer's & Dementia*, **15**, 321-387. https://doi.org/10.1016/j.jalz.2019.01.010

[4] Twiss, E., McPherson, C. and Weaver, D.F. (2025) Global Diseases Deserve Global Solutions: Alzheimer's Disease. *Neurology International*, **17**, Article 92. https://doi.org/10.3390/neurolint17060092

[5] Cacabelos, R. (2025) Special Issue: "New Trends in Alzheimer's Disease Research: From Molecular Mechanisms to Therapeutics: 2nd Edition". *International Journal of Molecular Sciences*, **26**, Article 7175. https://doi.org/10.3390/ijms26157175

[6] Mitchell, A.J., Kemp, S., Benito-León, J. and Reuber, M. (2010) The Influence of Cognitive Impairment on Health-Related Quality of Life in Neurological Disease. *Acta Neuropsychiatrica*, **22**, 2-13. https://doi.org/10.1111/j.1601-5215.2009.00439.x

[7] Memudu, A.E., Olukade, B.A. and Alex, G.S. (2024) Neurodegenerative Diseases. In: Chatterjee, I. and Moradikor, N., Eds., *Integrating Neuroimaging, Computational Neuroscience, and Artificial Intelligence*, CRC Press, 128-147. https://doi.org/10.1201/9781032711102-8

[8] Landeiro, F., Mughal, S., Walsh, K., Nye, E., Morton, J., Williams, H., *et al.* (2020)

Health-Related Quality of Life in People with Predementia Alzheimer's Disease, Mild Cognitive Impairment or Dementia Measured with Preference-Based Instruments: A Systematic Literature Review. *Alzheimer's Research & Therapy*, **12**, Article No. 154. https://doi.org/10.1186/s13195-020-00723-1

[9] Hyman, B.T. (1997) The Neuropathological Diagnosis of Alzheimer's Disease: Clinical-Pathological Studies. *Neurobiology of Aging*, **18**, S27-S32. https://doi.org/10.1016/s0197-4580(97)00066-3

[10] Dickson, D.W. (1997) Neuropathological Diagnosis of Alzheimer's Disease: A Perspective from Longitudinal Clinicopathological Studies. *Neurobiology of Aging*, **18**, S21-S26. https://doi.org/10.1016/s0197-4580(97)00065-1

[11] DeTure, M.A. and Dickson, D.W. (2019) The Neuropathological Diagnosis of Alzheimer's Disease. *Molecular Neurodegeneration*, **14**, Article No. 32. https://doi.org/10.1186/s13024-019-0333-5

[12] Sabbagh, M.N., Boada, M., Borson, S., Chilukuri, M., Doraiswamy, P.M., Dubois, B., *et al.* (2020) Rationale for Early Diagnosis of Mild Cognitive Impairment (MCI) Supported by Emerging Digital Technologies. *The Journal of Prevention of Alzheimer's Disease*, **7**, 158-164. https://doi.org/10.14283/jpad.2020.19

[13] Tahami Monfared, A.A., Phan, N.T.N., Pearson, I., Mauskopf, J., Cho, M., Zhang, Q., *et al.* (2023) A Systematic Review of Clinical Practice Guidelines for Alzheimer's Disease and Strategies for Future Advancements. *Neurology and Therapy*, **12**, 1257-1284. https://doi.org/10.1007/s40120-023-00504-6

[14] Hampel, H., Lista, S. and Khachaturian, Z.S. (2012) Development of Biomarkers to Chart All Alzheimer's Disease Stages: The Royal Road to Cutting the Therapeutic Gordian Knot. *Alzheimer's & Dementia*, **8**, 312-336. https://doi.org/10.1016/j.jalz.2012.05.2116

[15] Werner, P., Barthel, H., Drzezga, A. and Sabri, O. (2015) Current Status and Future Role of Brain PET/MRI in Clinical and Research Settings. *European Journal of Nuclear Medicine and Molecular Imaging*, **42**, 512-526. https://doi.org/10.1007/s00259-014-2970-9

[16] Savitz, J.B., Rauch, S.L. and Drevets, W.C. (2013) Clinical Application of Brain Imaging for the Diagnosis of Mood Disorders: The Current State of Play. *Molecular Psychiatry*, **18**, 528-539. https://doi.org/10.1038/mp.2013.25

[17] Garcia, A. and Reilly, J. (2015) Linguistic Disruption in Primary Progressive Aphasia, Frontotemporal Degeneration, and Alzheimer's Disease. In Bahr, R.H. and Silliman, E.R., Eds., *Routledge Handbook of Communication Disorders*, Routledge, 268-277.

[18] Na Chiangmai, N. (2023) Spontaneous Speech Analysis for Detect-Ing Mild Cognitive Impairment and Alzheimer's Disease in Thai Older Adults. 1-236.

[19] Kothinti, R.R. (2021) Advancements in Natural Language Processing for Auto-Mated Phenotyping and Predictive Analytics in Oncology EHRS. *Iconic Research and Engineering Journals*, **8**, 245-252.

[20] Noori, A., Magdamo, C., Liu, X., Tyagi, T., Li, Z., Kondepudi, A., *et al.* (2022) Development and Evaluation of a Natural Language Processing Annotation Tool to Facilitate Phenotyping of Cognitive Status in Electronic Health Records: Diagnostic Study. *Journal of Medical Internet Research*, **24**, e40384. https://doi.org/10.2196/40384

[21] Alfalahi, H., Dias, S.B., Khandoker, A.H., Chaudhuri, K.R. and Hadjileontiadis, L.J. (2023) A Scoping Review of Neurodegenerative Manifestations in Explainable Digital Phenotyping. *npj Parkinson's Disease*, **9**, Article No. 49. https://doi.org/10.1038/s41531-023-00494-0

[22] Zhou, Y., Lin, X., Zhang, X., Wang, M., Jiang, G., Lu, H., Wu, Y., *et al.* (2023) On the

Opportunities of Green Computing: A Survey. arXiv:2311.00447.

[23] Shaikh, S., Pereira, K.W., Sahay, S., Lopes, A. and Parshionikar, S. (2024) An Extensive Review: Models for Regional Language Speech Recognition. 2024 4*th Asian Conference on Innovation in Technology* (*ASIANCON*), Pimari Chinchwad, 23-25 August 2024, 1-8. https://doi.org/10.1109/asiancon62057.2024.10837903

[24] Yu, D., Ju, Y., Wang, Y., Zweig, G. and Acero, A. (2007) Automated Directory Assistance System—From Theory to Practice. *Interspeech* 2007, Antwerp, 27-31 August 2007, 2709-2712. https://doi.org/10.21437/interspeech.2007-65

[25] Kudapa, S.P. (2025) AI-Driven Data Science Models for Real-Time Transcription and Productivity Enhancement in U.S. Remote Work Environments. *ASRC Procedia*: *Global Perspectives in Science and Scholarship*, **1**, 801-832. https://doi.org/10.63125/gzyw2311

[26] Ashurst, C. and Weller, A. (2023) Fairness without Demographic Data: A Survey of Approaches. *Equity and Access in Algorithms, Mechanisms, and Optimization*, Boston, 30 October 2023-1 November 2023, 1-12. https://doi.org/10.1145/3617694.3623234

[27] Ramesh, K., Sitaram, S. and Choudhury, M. (2023) Fairness in Language Models Beyond English: Gaps and Challenges. *Findings of the Association for Computational Linguistics*: *EACL* 2023, Dubrovnik, 2-6 May 2023, 2106-2119. https://doi.org/10.18653/v1/2023.findings-eacl.157

[28] Jones, P., Liu, W., Huang, I. and Huang, X. (2025) Examining Imbalance Effects on Performance and Demographic Fairness of Clinical Language Models. 2025 *IEEE* 13*th International Conference on Healthcare Informatics* (*ICHI*), Rende, 18-21 June 2025, 58-68. https://doi.org/10.1109/ichi64645.2025.00016

[29] AlSaad, R., Abd-alrazaq, A., Boughorbel, S., Ahmed, A., Renault, M., Damseh, R., *et al.* (2024) Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook. *Journal of Medical Internet Research*, **26**, e59505. https://doi.org/10.2196/59505

[30] He, R., Chapin, K., Al-Tamimi, J., Bel, N., Marquié, M., Rosende-Roca, M., *et al.* (2023) Automated Classification of Cognitive Decline and Probable Alzheimer's Dementia across Multiple Speech and Language Domains. *American Journal of Speech-Language Pathology*, **32**, 2075-2086. https://doi.org/10.1044/2023_ajslp-22-00403

[31] Orimaye, S.O., Wong, J.S., Golden, K.J., Wong, C.P. and Soyiri, I.N. (2017) Predicting Probable Alzheimer's Disease Using Linguistic Deficits and Biomarkers. *BMC Bioinformatics*, **18**, Article No. 34. https://doi.org/10.1186/s12859-016-1456-0

[32] Li, C. (2024) Detecting Cognitive Impairment from Language and Speech for Early Screening of Alzheimer's Disease Dementia with Interpretable Transformer-Based Language Models. PhD Dissertation, University of Minnesota.

[33] Uggen, T.K.E. (2020) The Use of Machine Learning Algorithms and Statistical Models to Classify Aphasia Severity. University of Technology Sydney (Australia).

[34] Medero, J. (2014) Automatic Characterization of Text Difficulty. PhD Dissertation, University of Washington.

[35] Davis, B.H. and Maclagan, M. (2009) Examining Pauses in Alzheimer's Discourse. *American Journal of Alzheimer's Disease & Other Dementias®*, **24**, 141-154. https://doi.org/10.1177/1533317508328138

[36] Andreetta, S., Cantagallo, A. and Marini, A. (2012) Narrative Discourse in Anomic Aphasia. *Neuropsychologia*, **50**, 1787-1793. https://doi.org/10.1016/j.neuropsychologia.2012.04.003

[37] McCarthy, P.M. (2005) An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD). PhD Dissertation, The University of Memphis.

[38] Shlesinger, M. (1998) Corpus-Based Interpreting Studies as an Offshoot of Corpus-Based Translation Studies. *Meta*, **43**, 486-493. https://doi.org/10.7202/004136ar

[39] McNamara, D.S., Graesser, A.C., McCarthy, P.M. and Cai, Z. (2014) Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press. https://doi.org/10.1017/cbo9780511894664

[40] Chou, C., Chang, C., Chang, Y., Lee, C., Chuang, Y., Chiu, Y., *et al.* (2024) Screening for Early Alzheimer's Disease: Enhancing Diagnosis with Linguistic Features and Biomarkers. *Frontiers in Aging Neuroscience*, **16**, Article ID: 1451326. https://doi.org/10.3389/fnagi.2024.1451326

[41] Kavé, G. and Goral, M. (2017) Word Retrieval in Connected Speech in Alzheimer's Disease: A Review with Meta-Analyses. *Aphasiology*, **32**, 4-26. https://doi.org/10.1080/02687038.2017.1338663

[42] Gagliardi, G. and Tamburini, F. (2021) Linguistic Biomarkers for the Detection of Mild Cognitive Impairment. *Lingue e Linguaggio*, **20**, 3-31.

[43] Nyongesa, C.A., Hogarth, M. and Pa, J. (2025) Artificial Intelligence-Driven Natural Language Processing for Identifying Linguistic Patterns in Alzheimer's Disease and Mild Cognitive Impairment: A Study of Lexical, Syntactic, and Cohesive Features of Speech through Picture Description Tasks. *Journal of Alzheimer's Disease*, **106**, 120-138. https://doi.org/10.1177/13872877251339756

[44] Rane, N., Choudhary, S. and Rane, J. (2023) Explainable Artificial Intelligence (XAI) in Healthcare: Interpretable Models for Clinical Decision Support. *SSRN Electronic Journal*, 17 p. https://doi.org/10.2139/ssrn.4637897

[45] Valente, F., Paredes, S., Henriques, J., Rocha, T., de Carvalho, P. and Morais, J. (2022) Interpretability, Personalization and Reliability of a Machine Learning Based Clinical Decision Support System. *Data Mining and Knowledge Discovery*, **36**, 1140-1173. https://doi.org/10.1007/s10618-022-00821-8

[46] Abbas, Q., Jeong, W. and Lee, S.W. (2025) Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges. *Healthcare*, **13**, Article 2154. https://doi.org/10.3390/healthcare13172154

[47] Hartsock, I. and Rasool, G. (2024) Vision-Language Models for Medical Report Generation and Visual Question Answering: A Review. *Frontiers in Artificial Intelligence*, **7**, Article ID: 1430984. https://doi.org/10.3389/frai.2024.1430984

[48] Hüser, M. (2021) Machine Learning Approaches for Patient Monitoring in the Intensive Care Unit. PhD Dissertation, ETH Zurich.

[49] Iriondo, C. (2021) Characterizing Phenotypes of Musculoskeletal Degeneration Using Medical Imaging and Deep Learning. PhD Dissertation, University of California.

[50] Hou, S., Wu, Y., Chen, K., Chang, T., Hsu, Y., Chuang, S., *et al.* (2022) Code-Switching Automatic Speech Recognition for Nursing Record Documentation: System Development and Evaluation. *JMIR Nursing*, **5**, e37562. https://doi.org/10.2196/37562

[51] KhudaBukhsh, A.R. (2024) Deceptively Simple: An Outsider's Perspective on Natural Language Processing. *AI Magazine*, **45**, 569-582. https://doi.org/10.1002/aaai.12204

[52] Levy, J.J. and O'Malley, A.J. (2020) Don't Dismiss Logistic Regression: The Case for Sensible Extraction of Interactions in the Era of Machine Learning. *BMC Medical Research Methodology*, **20**, Article No. 171. https://doi.org/10.1186/s12874-020-01046-3

[53] Sikdar, A., Liu, Y., Kedarisetty, S., Zhao, Y., Ahmed, A. and Behera, A. (2025) Interweaving Insights: High-Order Feature Interaction for Fine-Grained Visual Recognition. *International Journal of Computer Vision*, **133**, 1755-1779. https://doi.org/10.1007/s11263-024-02260-y

[54] Pratap, V., Xu, Q., Sriram, A., Synnaeve, G. and Collobert, R. (2020) MLS: A Large-Scale Multilingual Dataset for Speech Research. *Interspeech* 2020, Shanghai, 25-29 October 2020, 2757-2761. https://doi.org/10.21437/interspeech.2020-2826