

Speaker-Wise Dementia Screening from Naturalistic Speech with Lexical Entropy, Semantic Drift, and Multimodal Fusion

Royce Salah

rsalah3@gatech.edu

Georgia Institute of Technology

Research Article

Keywords:

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9795204/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Speaker-Wise Dementia Screening from Naturalistic Speech with Lexical Entropy, Semantic Drift, and Multimodal Fusion

Abstract

Although advances have been made in dementia symptom management, early detection during stages when interventions are most effective remains a significant challenge. Most speech-based screening methods are evaluated using controlled elicitation tasks, often without examining how feature strength changes in the years preceding formal diagnosis. This study introduces a multimodal approach for dementia risk screening from naturalistic speech, integrating four traditional lexical features, two computational markers (lexical entropy from a probabilistic language model and semantic drift from sentence-level embeddings), and eGeMAPSv02 acoustic biomarkers. The framework is assessed on the DementiaNet corpus using speaker-disjoint nested group cross-validation. The optimal configuration, late fusion with stacked blending, achieves an AUC of 0.607 at the speaker level. Time-to-diagnosis stratified analysis demonstrates statistically significant discrimination in near-diagnostic windows, with text features peaking in the 5-year bin (AUC = 0.678) and acoustic features at the year of diagnosis (AUC = 0.640, $p = 0.024$). Exploratory analysis of more distant temporal bins (10 and 15 years) indicates that computational features may retain discriminative value where traditional features do not, although these patterns do not reach statistical significance with the current sample size. These findings support further investigation of computational language features alongside traditional and acoustic markers in larger longitudinal studies of speech-based dementia screening.

1 Introduction

Dementia affects an estimated 57 million people worldwide, with projections suggesting it will nearly triple by 2050 as populations age ([GBD 2019 Dementia Forecasting Collaborators 2022](#)). Alzheimer’s disease, the most common form, is characterized by progressive cognitive degeneration that begins years, sometimes decades, prior to observable clinical symptoms. Early identification of individuals at risk is critical as interventions during the preclinical period may offer the greatest opportunity to slow disease progression and preserve quality of life. This is especially important given the irreversible nature of neurodegenerative cell loss in dementia-causing diseases, as

well as the applications of lifestyle adjustments and emerging therapies in slowing the progression of these diseases.

Spontaneous speech has emerged as a non-invasive, cost-effective source of patient biomarkers for cognitive decline, particularly with recent advances in language modeling. Speech production engages multiple cognitive systems, including lexical retrieval, executive planning, and motor control, rendering it highly sensitive to the cognitive decline characteristic of dementia. Previous research has demonstrated that features derived from spoken language, such as lexical diversity, syntactic complexity, and speech patterns, can distinguish individuals with dementia from healthy controls (Roark et al. 2011; Luz et al. 2020). More recent studies have utilized pretrained language models and self-supervised speech representations to extract richer features (Agbavor and Liang 2022).

Despite progress in language modeling, several limitations persist in the field. Most studies rely on controlled elicitation tasks, such as the Cookie Theft approach from the DementiaBank Pitt Corpus (Becker et al. 1994), which do not capture the variability inherent in naturalistic speech and thus limit broader applicability. Additionally, evaluations are typically conducted at the time of diagnosis, without investigating how feature strength evolves over time to diagnosis. Given the gradual onset and variability of symptom development, biomarkers identified in previous research may have limited applicability during the earliest stages of detection.

The present study addresses these limitations by introducing a multimodal analysis of the DementiaNet corpus (Gite n.d.), a longitudinal dataset of naturalistic spontaneous speech from public figures with retrospectively assigned dementia labels. The framework integrates six text features (four traditional lexical markers and two computational features) with acoustic biomarkers and evaluates multiple classifiers and fusion strategies using speaker-group cross-validation. The DementiaNet corpus includes time-to-diagnosis annotations, enabling stratified analysis across temporal bins ranging from the year of diagnosis to 5–15 years prior to symptom onset.

This study employs two exploratory computational features: lexical entropy and semantic drift. Lexical entropy is defined as the mean per-token Shannon entropy of the next-word probability distributions generated by a GPT-2 language model, capturing the uncertainty in word-sequence predictability. Semantic drift was captured with Sentence-BERT, a sentence embedding model, to compute the mean cosine distance between adjacent sentence embeddings, thereby quantifying discourse coherence within the speaker’s dialogue. These features are hypothesized to detect subtle cognitive impairment prior to formal diagnosis, manifesting as increased model surprisal and reduced semantic organization before traditional lexical deficits become apparent.

The primary contributions of this study are as follows:

1. A multimodal feature framework combining traditional lexical features (MATTR, TTR, mean sentence length, pronoun ratio), computational features (language model lexical entropy, embedding-based semantic drift), and eGeMAPSv02 acoustic features for dementia screening from naturalistic speech.
2. Time-to-diagnosis stratified evaluation demonstrates statistically significant discrimination in time-based windows and descriptive evidence of increased discriminatory power across feature families under different windows.

3. Systematic ablation and classifier comparison demonstrate that feature selection has a greater impact than classifier architecture in lower signal settings.

2 Related Work

2.1 Benchmark Datasets and Evaluation Norms

The ADReSS Challenge (Luz et al. 2020) introduced a demographically balanced benchmark for Alzheimer’s detection from spontaneous speech, establishing speaker-stratified evaluation and common metrics as baselines. ADReSSo (Luz et al. 2021) expanded this to speech-only pipelines and cognitive decline prediction, whilst ADReSS-M (Luz et al. 2024) reframed the problem in terms of cross-lingual generalization. These benchmarks have improved cross-study comparability but still rely on the Cookie Theft picture-description approach, which limits broader generalizability. The present work evaluates naturalistic spontaneous speech with longitudinal time-to-diagnosis annotations.

2.2 Interpretable Linguistic Markers

Prior research documents meaningful fluctuations in lexical diversity, sentence complexity, and pronoun usage in dementia speech (Fraser et al. 2016; Covington and McFall 2010), which support the usage of features such as MATTR and pronoun ratios as candidate screening markers. Roark et al. (2011) established that automatically derived speech and language markers can discriminate healthy controls from individuals with mild cognitive impairment. This analysis uses these traditional features as components in an ablation analysis pipeline to identify metrics with the highest discriminative performance.

2.3 Language-Model and Embedding-Based Approaches

The GPT-D approach (Li et al. 2022) introduced paired perplexity using intentionally degraded language models to identify dementia-related linguistic anomalies, with generalization across conversational and picture-description data. Further work has extended perplexity-based scoring to larger language models (Xiao et al. 2025). Text-based deep learning approaches, including fastText and CNN-based approaches, have been shown to support dementia recognition from spontaneous-speech transcriptions (Meghanani et al. 2021).

The approach presented in this paper differs by computing entropy directly from next-token distributions as an interpretable numerical feature, and fusing it with other biomarkers rather than using it in isolation. Dialogue cohesion is captured as mean adjacent-sentence cosine differences (semantic drift) to model discourse progression rather than summarization.

Table 1 DementiaNet corpus summary.

Property	Dementia	Control	Total
Speakers	84	131	215
Clips	131	324	455

2.4 Acoustic and Multimodal Approaches

[Haider et al. \(2020\)](#) evaluated acoustic feature sets for dementia detection, establishing eGeMAPS-style features as strong baselines, while expanded systems have investigated diverse fusion architectures ([Cummins et al. 2020](#); [Rohanian et al. 2021](#)). The contribution of the architecture presented in this paper is a transparent fusion comparison between early concatenation and late stacked blending, focused primarily on understanding which modalities contribute most strongly to each timespan.

2.5 Time-to-Diagnosis and Preclinical Detection

Research suggests linguistic biomarkers can provide valuable information years before a formal diagnosis ([Petti et al. 2023](#)). Despite this, most evaluations have focused only on data at the time of, or following, diagnosis, notably on highly structured and elicited data. This framework models time-to-diagnosis using structured, time-bound bins, enabling temporally stratified testing with quantifiable results on how different feature families behave at various distances from diagnosis.

3 Dataset

3.1 The DementiaNet Corpus

The DementiaNet corpus ([Gite n.d.](#)) is a publicly available longitudinal dataset of spontaneous speech from public figures and celebrities. The dementia group is comprised of individuals with confirmed dementia diagnoses (not limited to Alzheimer’s Dementia), with recordings spanning from the year of diagnosis to approximately 15 years before symptom onset. The control group consists of individuals who survived past age 85 without documented cognitive impairment, with recordings sampled across comparable age ranges.

Unlike Cookie Theft approaches, serving as the foundation for DementiaBank and ADReSS, DementiaNet captures naturalistic speech primarily through interviews and public appearances. This provides a stronger baseline for more typical speech, mirroring the low-signal reality of everyday life at the expense of greater variability in recording conditions and content.

3.2 Dataset Statistics

Table 1 summarises the corpus composition while Table 2 shows the distribution of dementia clips across temporal bins. All dementia clips fall within 0–15 years before diagnosis. Control clips serve as the shared negative reference across all temporal bins.

Table 2 Temporal bin distribution (dementia clips).

Bin	0	5	10	15
Clips	27	45	32	27

3.3 Data Limitations

Several properties of DementiaNet serve as limitations. Since all subjects are public figures, they are more likely to have access to higher education, media training, and healthcare access, limiting the generalizability of the data to broader clinical contexts and the general population. Second, dementia labels are assigned based on publicly available information rather than clinical assessments, which introduces potential imprecision in the timing of diagnoses and in pre-diagnosis clip labeling. Third, mirroring the real-world challenges of naturalistic speech, recording conditions vary substantially (microphones, environments, interviewers, time periods), making it increasingly difficult to create a reliable baseline.

3.4 Ethical Considerations

This study involved secondary analysis of publicly available speech recordings. No new data were collected, and no interactions with participants were conducted. Results are reported in aggregate only. In acknowledgment of the limitations of retrospective labeling, the results reported here reflect only statistical associations in a research setting, not clinically.

4 Methods

4.1 Pipeline Overview

The pipeline proceeds in five stages: automatic speech recognition, text feature extraction, acoustic feature extraction, classification with single-modality or fused features, and time-to-diagnosis stratified evaluation. Figure 1 illustrates the end-to-end architecture.

4.2 Transcription

Audio recordings are transcribed using OpenAI Whisper (Radford et al. 2023). Transcripts are plain-text files used as input for text feature extraction. ASR-generated transcripts were manually inspected to verify that each file contained coherent natural language output. This check was not a word-level accuracy audit, but rather a confirmation that there were no major transcription failures, such as garbled output or empty files.

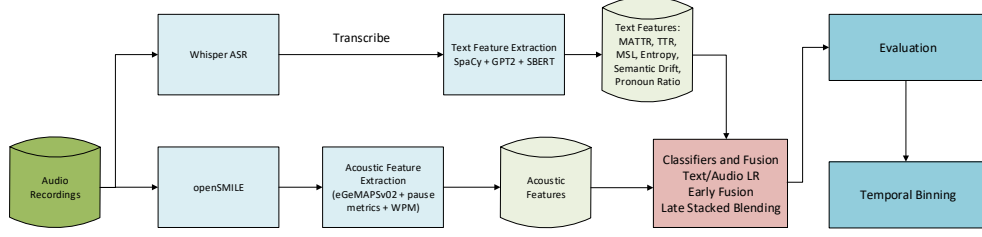


Fig. 1 End-to-end pipeline architecture. Audio recordings are processed through parallel text and acoustic feature extraction paths, then evaluated under single-modality and fused classifiers with temporal bin stratification.

4.3 Text Feature Extraction

Six features are extracted per transcript using spaCy (Honnibal et al. 2020), Sentence-BERT (Reimers and Gurevych 2019), and GPT-2 (Radford et al. 2019).

- **MATTR:** Moving-average type-token ratio (with a window of 50 tokens) over lowercased alphabetic tokens.
- **TTR:** Global type-token ratio (unique tokens / total tokens).
- **Mean Sentence Length:** Mean token count per sentence, excluding punctuation and whitespace.
- **Pronoun Ratio:** The ratio of the total number of pronouns used to the total number of nouns.
- **Lexical Entropy:** Mean per-token Shannon entropy of language model next-word distributions (raw text, truncated to 512 tokens).
- **Semantic Drift:** Mean cosine distance between adjacent sentence embeddings.

Lexical entropy quantifies the predictability of word choices relative to a reference language model. Higher entropy indicates a greater next-word uncertainty, potentially reflecting disrupted language planning. GPT-2 (117M parameters) is used without fine-tuning as a fixed reference model. Semantic drift quantifies discourse coherence with higher values indicating more erratic discourse between sentences, potentially reflecting a more nuanced indication of impaired cognitive function.

4.4 Acoustic Feature Extraction

Acoustic features are extracted using openSMILE (Eyben et al. 2010) with the eGeMAPSv02 configuration (Eyben et al. 2016), producing 88 functional descriptors. Five additional features are computed, which include voiced clip duration, words per minute, silence percentage, mean silence duration, and silence count. Three feature configurations were evaluated: full eGeMAPS (~101 features), core audio (~13 curated temporal/prosodic features), and PCA-reduced audio (50 principal components of the standardized full set).

4.5 Classification Models

Four classifier families were evaluated to assess whether discriminative performance is primarily a function of feature quality or model architecture. Logistic Regression serves as the primary model throughout, as it is interpretable and stable under small-sample conditions. The remaining three classifiers are included for comparison; if performance is similar across architectures, this confirms that feature selection is a stronger driver than model complexity. The following models involved in this analysis were integrated within the same cross-validation procedure and fusion strategies:

- **Logistic Regression (LR)** L2-regularised logistic regression with standard scaling. Serves as the primary model across all text-only, audio-only, and fusion configurations due to its interpretability and stability under small-sample conditions.
- **Calibrated SVM (SVM-cal)** LinearSVC with Platt-scaled probability calibration. Calibration is required to produce well-formed probabilities for late fusion blending.
- **Multi-Layer Perceptron (MLP)** Two-layer neural network (64–32 units) with early stopping. Included to assess whether a non-linear classifier extracts additional signal from the feature set.
- **CatBoost** Gradient-boosted trees applied to raw features without preprocessing. Included as a representative ensemble method to test whether model complexity improves over linear approaches.

4.6 Fusion Strategies

As part of testing, both early and late fusion strategies were employed to evaluate the effectiveness of combining insights from the lexical and acoustic modalities. Early fusion testing combined text and audio feature sets and used a single classifier trained on the joint representation. Late fusion combines the separate text and audio model outputs and trains a logistic regression blender on the stacked predictions. For test-time inference, the heads and blender were refit on the full outer training fold.

4.7 Evaluation Protocol

All models were evaluated under nested 5×5 StratifiedGroupKFold cross-validation with speakers as the grouping variable. This ensures that speakers do not appear in both the training and test sets within any fold, mitigating leakage and model overfitting to individual speakers.

The inner loop was used to select hyperparameters and generate out-of-fold (OOF) probabilities for threshold optimization. Rather than applying a fixed 0.5 decision threshold, a target-sensitivity policy was used, setting the threshold to the 20th percentile of positive-class OOF scores in the training partition, targeting approximately 80% sensitivity. This policy is meant to reflect the clinical priority of screening tools, in which false negatives carry a greater cost than false positives.

Predictions are reported at both the clip level and speaker level. At the clip level, each audio segment is treated as a probabilistically independent observation, while at

Table 3 Main results at speaker level (mean $\pm$ SD across 5 folds, with 95% CI and significance vs. chance).

Model	Modality	AUC $\pm$ SD	95% CI	p	F1	Sens.	Spec.	Brier
late_fusion_lr	Fusion	0.607 $\pm$ 0.011	[0.593, 0.622]	<0.001	0.531	0.786	0.246	0.243
early_fusion_lr	Fusion	0.596 $\pm$ 0.072	[0.506, 0.686]	0.041	0.552	0.868	0.183	0.239
text_lr	Text	0.585 $\pm$ 0.070	[0.498, 0.672]	0.053	0.571	0.823	0.329	0.243
audio_lr	Audio	0.580 $\pm$ 0.104	[0.452, 0.709]	0.158	0.515	0.820	0.130	0.240
text_svm_cal	Text	0.564 $\pm$ 0.060	[0.489, 0.639]	0.077	0.551	0.833	0.238	0.246
audio_core_lr	Audio	0.507 $\pm$ 0.070	[0.420, 0.593]	0.837	0.522	0.881	0.045	0.251

the speaker level, clip probabilities are averaged within each speaker, with the speaker label taken as the maximum label across their clips. In real-world applications, clips would be screened for a specific person rather than isolated instances, which is why results are reported at the speaker level.

Model performance was evaluated using the area under the ROC curve, with F1 score, sensitivity, specificity, and Brier score reported as supplementary metrics. Confidence intervals were calculated across cross-validation folds using the standard error of the mean and the appropriate critical value from the t -distribution. Statistical significance was assessed by testing whether the observed AUC differed significantly from chance-level performance.

4.8 Temporal Bin Analysis

Dementia clips are stratified into four bins: 0 (year of diagnosis), 5, 10, and 15 years prior to diagnosis. For each bin, AUC and sensitivity at 10% false-positive rate are computed by pairing the bin’s positive clips with all control clips; this analysis was performed per fold and averaged.

5 Results

5.1 Main Model Comparison

Table 3 presents the aggregate performance of six configurations at the speaker level under 5-fold nested CV with target-sensitivity thresholding. Figure 2 shows the corresponding ROC curves.

Late fusion is the only model to achieve both the highest AUC and statistical significance, with notably low cross-fold variance. Text-only and audio-only logistic regression models yield comparable point estimates (0.585 and 0.580, respectively) but do not reach significance at $\alpha = 0.05$, reflecting high fold-level variance (SDs of 0.070 and 0.104). Early fusion achieves statistical significance ($p = 0.041$). The core audio feature set does not discriminate above chance ($p = 0.837$).

The overall performance range (AUC 0.51–0.61) aligns with the inherently low-signal nature of pre-diagnostic screening from naturalistic speech, where many positive-class samples were recorded years before clinical manifestation.

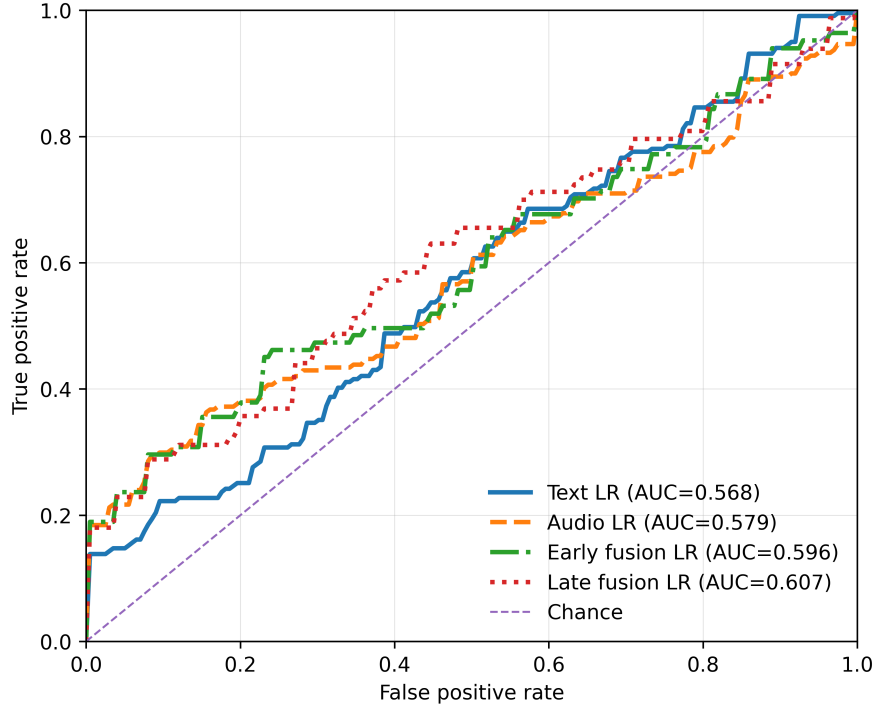


Fig. 2 Speaker-level ROC curves (mean across folds). AUC values reflect averaged per-fold curves and may differ slightly from the per-fold means in Table 3.

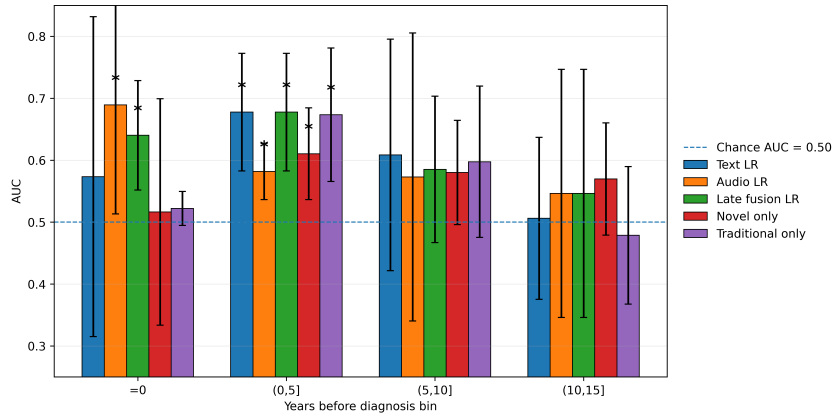


Fig. 3 AUC by temporal bin and feature family. Error bars show fold-level standard deviation. Asterisks denote $p < 0.05$ vs. chance (one-sample t -test). Significant discrimination is confined to the $= 0$ and $(0, 5]$ bins; patterns in more distant bins are not statistically confirmed.

Table 4 AUC by temporal bin (5-fold), with 95% CI, significance (one-sample t -test vs. 0.50), and Cohen’s d .

Model	Bin	AUC $\pm$ SD	95% CI	p	d	n_{pos}
text_lr	5	0.678 $\pm$ 0.095	[0.560, 0.795]	0.014	1.87	45
text_lr	10	0.585 $\pm$ 0.118	[0.438, 0.732]	0.182	0.72	32
text_lr	0	0.532 $\pm$ 0.056	[0.463, 0.601]	0.266	0.58	27
text_lr	15	0.487 $\pm$ 0.113	[0.347, 0.627]	0.812	−0.11	27
audio_lr	0	0.640 $\pm$ 0.088	[0.530, 0.750]	0.024	1.59	27
audio_lr	5	0.582 $\pm$ 0.045	[0.526, 0.638]	0.016	1.81	45
audio_lr	10	0.547 $\pm$ 0.090	[0.435, 0.660]	0.308	0.52	32
audio_lr	15	0.546 $\pm$ 0.200	[0.298, 0.795]	0.632	0.23	27
late_fusion	5	0.652 $\pm$ 0.055	[0.584, 0.720]	0.003	2.78	45
late_fusion	0	0.584 $\pm$ 0.051	[0.520, 0.647]	0.022	1.64	27
late_fusion	10	0.579 $\pm$ 0.110	[0.443, 0.715]	0.182	0.72	32
late_fusion	15	0.462 $\pm$ 0.136	[0.293, 0.631]	0.565	−0.28	27
novel_only	5	0.610 $\pm$ 0.074	[0.518, 0.702]	0.029	1.49	45
novel_only	15	0.570 $\pm$ 0.091	[0.457, 0.682]	0.161	0.77	27
traditional_only	5	0.673 $\pm$ 0.108	[0.540, 0.807]	0.023	1.61	45
traditional_only	15	0.479 $\pm$ 0.111	[0.341, 0.617]	0.690	−0.19	27

5.2 Time-to-Diagnosis Stratification

Table 4 presents the temporal bin analysis. Figure 3 visualises the AUC values across temporal bins with significance annotations.

Statistically significant discrimination ($p < 0.05$) is observed exclusively in the = 0 and (0, 5] temporal bins. In the (5, 10] and (10, 15] bins, no model achieves significance at $\alpha = 0.05$. This is partially a consequence of both diminished effect sizes and small positive-class samples (27–32 clips per bin), which yield wide confidence intervals. Despite the lack of statistical significance in distant bins, a pattern is observable across feature families.

5.2.1 At diagnosis.

Acoustic features provide the strongest point estimate (AUC = 0.640, $p = 0.024$), while text features hover near chance (0.532, $p = 0.266$). These results are consistent with late-stage motor and articulatory degradation being the primary acoustic signature.

5.2.2 Near diagnosis (5 years).

Text features achieved their peak (text_lr: 0.678, $p = 0.014$), with a large effect size ($d = 1.87$). Late fusion also achieves its strongest performance (0.652, $p = 0.003$). Traditional features alone reach 0.673 ($p = 0.023$), suggesting that lexical diversity and pronoun-ratio changes are the dominant near-diagnostic signal.

Table 5 Text feature ablation (speaker level, 5-fold).

Feature set	AUC $\pm$ SD	95% CI	p
Full (6 features)	0.585 $\pm$ 0.070	[0.498, 0.672]	0.053
Traditional only (4)	0.602 $\pm$ 0.092	[0.487, 0.716]	0.070
No entropy (5)	0.583 $\pm$ 0.086	[0.476, 0.690]	0.097
No drift (5)	0.586 $\pm$ 0.079	[0.488, 0.684]	0.071
Novel only (2)	0.575 $\pm$ 0.029	[0.539, 0.611]	0.005

Table 6 Audio dimensionality (10-fold, clip level).

Configuration	Dims	AUC $\pm$ SD
Full eGeMAPS	~ 101	0.575 $\pm$ 0.089
Core audio	~ 13	0.493 $\pm$ 0.109
Text + PCA-50 audio	56	0.569 $\pm$ 0.079

5.2.3 Mid-range and preclinical (10 and 15 years).

No model achieved statistical significance. Text feature performance declined across the farther temporal windows, while audio features remained relatively stable but non-significant. The novel-only feature set produced the highest point estimate in the 15-year bin compared with the other text-based configurations, although the current sample size does not allow this difference to be confirmed statistically. This pattern provides descriptive evidence that the novel language features may capture a different signal farther from diagnosis, but larger samples would be needed to determine whether this effect is reliable.

5.3 Feature Ablation

Table 5 presents the text feature ablation results. At the aggregate level, removing the novel features individually or in combination has minimal impact on AUC, with changes ranging from -0.010 to $+0.017$.

Notably, the novel-only model (AUC = 0.575) is the only ablation to achieve statistical significance ($p = 0.005$) against chance, with low cross-fold variance (SD = 0.029). This finding suggests that the novel features provide modest yet stable discriminative performance consistently above chance across all folds, whereas the traditional features exhibit higher variance, indicating greater sensitivity to fold composition.

In conjunction with the bin analysis, these results indicate that the novel features contribute a qualitatively distinct signal compared to traditional features. While the novel features were less effective in the near-diagnostic bin, where traditional features performed most strongly, they demonstrated greater consistency and potential informativeness in more distant preclinical windows.

Table 6 summarises the audio dimensionality comparison. The full eGeMAPS feature set outperforms the curated core subset, indicating that reducing to 13 hand-selected temporal features eliminates discriminative acoustic information rather than

Table 7 Text-feature classifier family comparison under 10-fold evaluation.

Classifier	AUC $\pm$ SD	95% CI	p
LR	0.585 $\pm$ 0.086	[0.523, 0.647]	0.013
SVM-cal	0.581 $\pm$ 0.096	[0.512, 0.650]	0.026
MLP	0.613 $\pm$ 0.118	[0.529, 0.697]	0.014
CatBoost	0.519 $\pm$ 0.153	[0.410, 0.628]	0.702

merely removing noise. PCA-reduced fusion does not improve upon the raw full audio feature set, suggesting that the performance limitations of acoustic features in this context are not primarily due to dimensionality. Instead, these limitations likely reflect the variability of recording conditions and speakers inherent in naturalistic speech, consistent with findings in Section 5.1.

5.4 Classifier Comparison

Table 7 reports classifier performance across all four families. In 10-fold evaluation, logistic regression, SVM, and MLP all perform above chance, whereas CatBoost did not.

6 Discussion

6.1 Feature Discrimination Across Temporal Windows

The temporal bin analysis is most informative in the near-diagnostic windows, where sample sizes support statistical testing. In the 0–5-year bin, text features discriminate at an AUC of 0.678 ($p = 0.014$, $d = 1.87$), which is consistent with the lexical-retrieval difficulties and syntactic simplification that are typically the first clinically recognized language symptoms. At the point of diagnosis, acoustic features take over (AUC = 0.640, $p = 0.024$), plausibly reflecting motor and articulatory changes that become pronounced late in the disease course. Beyond 5 years, both modalities lose statistical power, though whether this reflects genuinely weaker signal or simply the 27–32 clips per bin is an open question.

6.2 Novel Features: Stability and Preclinical Potential

What stands out about lexical entropy and semantic drift is not their discriminative power, at a modest 0.575 AUC, but their consistency. The novel-only ablation is the only configuration to reach significance against chance across all folds ($p = 0.005$, $SD = 0.029$), whereas the traditional features fluctuate substantially ($SD = 0.070$ – 0.092). One interpretation is that entropy and drift model a diffuse cognitive process (language-planning uncertainty and discourse-maintenance difficulty) that is present at a low level across all temporal windows, rather than more apparent lexical-retrieval failures that traditional features pick up near diagnosis. In the 15-year bin, the novel

features achieve the highest text-based AUC (0.570 vs. 0.479–0.487), but with only 27 positive clips and $p = 0.161$, this comparison cannot be resolved without more data.

6.3 Multimodal Fusion

Late fusion achieves the highest overall AUC (0.607) with low variance, indicating that stacked blending stabilizes predictions by integrating signals from both modalities. Although the fusion gain is modest (0.022 AUC over text alone), it is consistent, and the low variance distinguishes it from all single-modality models. While both modalities ultimately reflect cognitive decline, the fusion gain points to partially independent information across modalities.

6.4 Feature Engineering Versus Model Complexity

The classifier comparison shows that LR, SVM, and MLP achieve similar AUCs on the same features, whereas CatBoost did not perform above chance. This indicates that the discriminative structure of the six text features is approximately linear, and that the primary contribution of this work lies in the feature set rather than model architecture. With 215 speakers, regularized linear models were preferred for stability.

6.5 Clinical Context

The overall AUC of 0.607 does not justify deployment as a standalone diagnostic tool. However, the system is not intended to replace clinical assessment, but rather to function as one component within a broader multimodal screening pipeline. Statistically significant discrimination in near-diagnostic windows, with AUCs of 0.640 for acoustic features at diagnosis and 0.678 for text features in the 5-year bin, indicates that speech-based features may have practical value at the stage when clinical concern is typically first raised. The target-sensitivity threshold policy prioritizes recall over precision, reflecting the clinical reality that missing at-risk individuals carries greater cost than generating false alarms.

6.6 Sources of Error

Several sources of error could contribute to the observed effects and should be considered when interpreting these results.

Recording conditions. DementiaNet draws from YouTube interviews spanning different decades, microphones, environments, and audio quality levels. Acoustic features may partially capture recording-era characteristics rather than genuine vocal biomarkers.

Demographic confounds. Age, gender, education, and media experience were not modeled as covariates. Public figures with dementia may differ systematically from controls in ways correlated with speech patterns but unrelated to cognitive status.

ASR artifacts. Transcripts were manually inspected for major failures, but word-level ASR accuracy was not verified against ground-truth speech. Whisper errors may

introduce differences between clear speech and less fluent or degraded speech, potentially inflating apparent discrimination. Recent evidence suggests that ASR errors themselves may be discriminative (Li et al. 2024), further complicating conclusions.

Retrospective labeling. Diagnosis timing is based on public reports rather than clinical records. This introduces uncertainty in the years-pre-diagnosis variable, making temporal bin assignments less certain.

7 Limitations

Sample size. The dataset includes 215 speakers and 455 clips, limiting the statistical power of the analysis. This is especially important for the temporal-bin results, where the dementia-positive bins contain only 27–45 clips. As a result, the confidence intervals are wide, further supporting an exploratory interpretation of the analysis.

Representativeness. DementiaNet is composed of public figures who differ from the general population in education, media training, healthcare access, and recording context.

Single-corpus evaluation. No external validation was performed. Although speaker-disjoint cross-validation reduces speaker leakage within the corpus, the results should be interpreted as within-corpus generalization only. Validation on an independent clinical or naturalistic speech dataset would be needed before making stronger claims about broader applicability.

Fixed model settings. The primary logistic regression models used fixed hyperparameters rather than fully tuning model settings within each fold. This choice was made to keep the evaluation simple and comparable across feature configurations, but it understates the best achievable performance for some feature sets.

8 Conclusion

This paper presented a multimodal pipeline for dementia screening from naturalistic speech, evaluated on the DementiaNet corpus. Late fusion of text and acoustic features achieved the strongest aggregate performance ($AUC = 0.607$, $p < 0.001$), with statistically significant discrimination concentrated in the near-diagnostic temporal windows. Text features were strongest in the 5-year bin ($AUC = 0.678$, $p = 0.014$) while acoustic features showed their greatest utility at diagnosis ($AUC = 0.640$, $p = 0.024$). Across classifier families, feature selection had a greater impact on performance than model architecture, with linear models being preferred for this setting. The novel features, lexical entropy and semantic drift, demonstrated stable discrimination across folds and a trend toward discriminatory strength in the furthest temporal windows, though bin sizes (27 clips in the 15-year window) do not provide statistical confirmation.

These results do not support a claim of reliable preclinical dementia detection from speech, and findings from preclinical bins should be interpreted as exploratory rather than definitive. However, the near-diagnostic findings demonstrate that multimodal speech features possess discriminative value in near-diagnostic windows, and that traditional and computational language features exhibit distinct temporal profiles

of utility. If replicated in larger longitudinal datasets, this distinction could have significant implications for the design of speech-based screening tools and for determining the optimal stage of disease progression for their deployment.

Statements and Declarations

- **Funding:** No funding was received to assist with the preparation, analysis, or publication of this research.
- **Ethics declaration:** Not applicable.

References

- Agbavor F, Liang H (2022) Predicting dementia from spontaneous speech using large language models. *PLOS Digital Health* 1(12):e0000168. <https://doi.org/10.1371/journal.pdig.0000168>
- Becker JT, Boller F, Lopez OL, et al (1994) The natural history of Alzheimer’s disease. *Archives of Neurology* 51(6):585–594
- Covington MA, McFall JD (2010) Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics* 17(2):94–100. <https://doi.org/10.1080/09296171003643098>
- Cummins N, Pan Y, Ren Z, et al (2020) A comparison of acoustic and linguistics methodologies for Alzheimer’s dementia recognition. In: *Proceedings of Interspeech 2020*, pp 2182–2186, <https://doi.org/10.21437/Interspeech.2020-2635>
- Eyben F, Wöllmer M, Schuller B (2010) openSMILE: The Munich versatile and fast open-source audio feature extractor. In: *Proceedings of the 18th ACM International Conference on Multimedia (MM ’10)*, pp 1459–1462, <https://doi.org/10.1145/1873951.1874246>
- Eyben F, Scherer KR, Schuller BW, et al (2016) The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7(2):190–202
- Fraser KC, Meltzer JA, Rudzicz F (2016) Linguistic features identify Alzheimer’s disease in narrative speech. *Journal of Alzheimer’s Disease* 49(2):407–422. <https://doi.org/10.3233/JAD-150520>
- GBD 2019 Dementia Forecasting Collaborators (2022) Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. *The Lancet Public Health* 7(2):e105–e125. [https://doi.org/10.1016/S2468-2667\(21\)00249-8](https://doi.org/10.1016/S2468-2667(21)00249-8)
- Gite S (n.d.) DementiaNet: A longitudinal spontaneous speech audio dataset for dementia screening. GitHub repository, URL <https://github.com/shreyasgite/>

dementianet, accessed 2025

- Haider F, de la Fuente Garcia S, Luz S (2020) An assessment of paralinguistic acoustic features for detection of Alzheimer’s dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing* 14(2):272–281. <https://doi.org/10.1109/JSTSP.2019.2955022>
- Honnibal M, Montani I, Van Landeghem S, et al (2020) spaCy: Industrial-strength natural language processing in Python. <https://doi.org/10.5281/zenodo.1212303>
- Li C, Knopman D, Xu W, et al (2022) GPT-D: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp 1866–1877, <https://doi.org/10.18653/v1/2022.acl-long.131>
- Li C, Xu W, Cohen T, et al (2024) Useful blunders: Can automated speech recognition errors improve downstream dementia classification? *Journal of Biomedical Informatics* 150:104598. <https://doi.org/10.1016/j.jbi.2024.104598>
- Luz S, Haider F, de la Fuente S, et al (2020) Alzheimer’s dementia recognition through spontaneous speech: The ADReSS Challenge. In: *Proceedings of Interspeech 2020*, pp 2172–2176, <https://doi.org/10.21437/Interspeech.2020-2571>
- Luz S, Haider F, de la Fuente S, et al (2021) Detecting cognitive decline using speech only: The ADReSSo Challenge. In: *Proceedings of Interspeech 2021*, pp 3780–3784, <https://doi.org/10.21437/Interspeech.2021-1220>
- Luz S, Haider F, Fromm D, et al (2024) An overview of the ADReSS-M signal processing grand challenge on multilingual Alzheimer’s dementia recognition through spontaneous speech. *IEEE Open Journal of Signal Processing* 5:738–749. <https://doi.org/10.1109/ojsp.2024.3378595>
- Meghanani A, Anoop CS, Ramakrishnan AG (2021) Recognition of Alzheimer’s dementia from the transcriptions of spontaneous speech using fastText and CNN models. *Frontiers in Computer Science* 3:624558. <https://doi.org/10.3389/fcomp.2021.624558>
- Petti U, Baker S, Korhonen A, et al (2023) The generalizability of longitudinal changes in speech before Alzheimer’s disease diagnosis. *Journal of Alzheimer’s Disease* 92(2):547–564. <https://doi.org/10.3233/JAD-220847>
- Radford A, Wu J, Child R, et al (2019) Language models are unsupervised multitask learners. Tech. rep., OpenAI
- Radford A, Kim JW, Xu T, et al (2023) Robust speech recognition via large-scale weak supervision. In: *Proceedings of the 40th International Conference on Machine*

- Learning (ICML), URL <https://arxiv.org/abs/2212.04356>
- Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp 3982–3992, <https://doi.org/10.18653/v1/D19-1410>
- Roark B, Mitchell M, Hosom JP, et al (2011) Spoken language derived measures for detecting mild cognitive impairment. IEEE Transactions on Audio, Speech, and Language Processing 19(7):2081–2090. <https://doi.org/10.1109/TASL.2011.2112351>
- Rohanian M, Hough J, Purver M (2021) Alzheimer’s dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. In: Proceedings of Interspeech 2021, pp 3820–3824, <https://doi.org/10.21437/Interspeech.2021-1633>
- Xiao Y, Christensen H, Goetze S (2025) Alzheimer’s dementia detection using perplexity from paired large language models. In: Proceedings of Interspeech 2025, pp 1423–1427, <https://doi.org/10.21437/Interspeech.2025-2428>