Diagnosis of dementia in the early stages of medical preventive research

A Preprint

Dmitry S.Slobodin[*]
Computational Mathematics and Cybernetics
Lomonosov Moscow State University
Moscow, Russia
email@domain

Oleg V.Senko[†]
Computational Mathematics and Cybernetics
Lomonosov Moscow State University
Moscow, Russia
email@domain

Abstract

This study investigates the comparative effectiveness of two speech tasks—standardized text reading and spontaneous picture description—for early dementia screening. While both approaches leverage acoustic biomarkers in speech, their relative performance and clinical utility remain unclear. We conducted a systematic comparison using parallel datasets where the same participants completed both tasks under identical recording conditions. Acoustic features were extracted using the eGeMAPS parameter set and evaluated through multiple machine learning models for regression and classification of established clinical scales (MMSE, MoCA, CDR). Contrary to initial expectations, spontaneous picture description consistently outperformed standardized reading across most evaluation metrics, achieving higher accuracy in differentiating between healthy individuals, those with mild cognitive impairment, and early-stage dementia patients. These findings provide empirical evidence for task selection in clinical speech-based screening protocols and demonstrate the superior sensitivity of natural language production for capturing cognitive-linguistic deficits associated with dementia progression.

Keywords Early dementia detection · Speech-based screening · Standardized reading task

1 Introduction

Dementia screening requires tools that are fast, low-cost, and deployable in everyday clinics where non-specialists see patients first. Speech offers a noninvasive digital biomarker that can be captured in minutes with ubiquitous hardware, enabling a triage workflow: flag suspicious cases during a routine visit and refer them to specialists for definitive assessment. Earlier detection improves care planning, supports timely interventions, and can reduce downstream healthcare costs, while remote collection enables longitudinal monitoring between visits.

Two primary speech paradigms dominate dementia screening research: standardized reading tasks and spontaneous description tasks. Each approach offers distinct advantages and limitations, yet their comparative effectiveness remains inadequately explored. Standardized reading controls linguistic content, potentially reducing variability and focusing on acoustic-prosodic markers of motor speech control. Spontaneous description engages natural language production, potentially capturing broader cognitive-linguistic processes but introducing greater variability.

Prior research predominantly analyzes spontaneous speech tasks such as picture description or interviews, combining acoustic–prosodic features (e.g., eGeMAPS/ComParE, MFCCs, pause and tempo statistics, F0 variability) with linguistic features extracted from ASR transcripts (lexical richness, syntactic complexity,

[*] Student Research Project (NIR).

[†] Supervisor; DSc (Phys.–Math.), Professor.

semantic coherence, and disfluencies) (Luz et al., 2020, 2021; de la Fuente Garcia et al., 2020; Fraser et al., 2016; Karlekar et al., 2018; Warnita et al., 2018; Eyben et al., 2016). Challenge datasets that control for confounders (e.g., ADReSS/ADReSSo) report around 75–80

However, spontaneous speech poses several challenges that limit clinical translation. High linguistic variability and topic effects can lead models to learn content rather than cognitive markers, reducing cross-domain and cross-language generalization (de la Fuente Garcia et al., 2020; Luz et al., 2020). Expert-labeled corpora are small, and imperfect validation (e.g., speaker leakage) inflates reported performance; dependence on ASR quality introduces variable noise by accent, recording, and pathology, precisely where robust signals are needed (de la Fuente Garcia et al., 2020; Luz et al., 2021). Channel and demographic confounders (age, sex, microphone) can further bias results if not explicitly controlled, and calibration and interpretability are often underreported (Luz et al., 2020, 2021; de la Fuente Garcia et al., 2020).

Standardized reading tasks offer potential solutions to these challenges by fixing lexical and syntactic content, reducing variability and ASR dependence. This allows measurement of acoustic–prosodic, timing, and pronunciation markers that more directly reflect motor–cognitive control. Audio can be aligned to canonical text via forced alignment to compute precise word/phoneme durations, speech versus articulation rate, silent and filled pause statistics and their positions, distributions of vowel/consonant durations (including VOT), as well as jitter/shimmer and F0 dynamics (Eyben et al., 2016; McAuliffe et al., 2017; Baevski et al., 2020).

This study addresses a critical gap in the literature by systematically comparing both paradigms using identical participant cohorts, acoustic feature sets, and machine learning pipelines. We evaluate their relative effectiveness for predicting multiple clinical scales (MMSE, MoCA, CDR) through both regression and classification frameworks. Our comparative approach enables direct assessment of each task type's strengths and limitations for clinical deployment.

Our contributions are as follows.

- Comparative framework: We develop and validate a systematic methodology for comparing standardized reading and spontaneous description tasks using parallel datasets from the same participant cohort.
- Feature analysis: We identify the most discriminative acoustic features for each task type using identical eGeMAPS parameter sets and statistical validation procedures.
- Multi-scale validation: We evaluate both regression (predicting raw clinical scores) and classification (cognitive status detection) performance across three established clinical instruments.
- Clinical insights: We provide empirical evidence for task selection in clinical practice, identifying contexts where each paradigm offers optimal sensitivity and specificity.
- Methodological transparency: We release reproducible pipelines for both task types to facilitate cross-study comparisons and clinical translation of speech-based dementia screening.

2 Related Work

A large body of research has investigated speech as a biomarker for cognitive impairment, with most studies focusing on either spontaneous speech tasks or standardized protocols, but rarely comparing them directly. Understanding the relative strengths of each approach is crucial for developing optimized clinical screening tools.

2.1 Spontaneous Speech Paradigms

The dominant approach in computational dementia detection has been spontaneous speech tasks, particularly picture description. Community challenges ADReSS and ADReSSo curated balanced subsets of DementiaBank, controlling for age, sex, and recording channel, and established rigorous speaker-held-out evaluations (Luz et al., 2020, 2021). Systems in these challenges typically fused acoustic–prosodic features (eGeMAPS/ComParE, MFCCs, pitch and energy statistics, pause and tempo measures) with linguistic features obtained from ASR, such as lexical diversity, syntactic complexity, semantic coherence, and disfluency counts, achieving about 75–80% accuracy for Alzheimer's vs. control and moderate error for MMSE regression (Luz et al., 2020, 2021). These results highlighted that temporal structure—speech rate, articulation rate, and pause distributions—carries strong signal for dementia detection.

Hand-crafted acoustic feature sets have served as robust baselines across paralinguistic tasks. The Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) was designed for physiological interpretability and low overfitting risk and has repeatedly delivered strong performance in health and affective computing (Eyben et al., 2016). Beyond low-level descriptors, research has emphasized prosodic and timing markers: ratio and distribution of silence, inter-word pauses, maximum pause length, variability of F0, jitter and shimmer, and stability of articulation rate (de la Fuente Garcia et al., 2020; Luz et al., 2017). Such features are particularly attractive in clinical contexts due to interpretability and hardware robustness.

Linguistic approaches exploit the idea that neurodegeneration affects lexical access, syntactic planning, and semantic coherence. Classical models employ hundreds of features spanning type–token ratios, part-of-speech patterns, parse-tree depth, and discourse measures (Fraser et al., 2016). Neural NLP architectures (CNN/LSTM/Transformers) trained on transcripts have matched or surpassed feature-engineered baselines, while interpretation studies point to simplified syntax, increased repetitions, and self-repairs as discriminative cues (Karlekar et al., 2018). However, transcript-based methods hinge on ASR quality, which varies with accent, noise, and pathology; this dependency limits portability across settings and languages (de la Fuente Garcia et al., 2020).

2.2 Standardized Speech Tasks

Standardized speech tasks have a long clinical tradition but have received less computational attention than spontaneous speech. Verbal fluency tests quantify clustering and switching behavior and predict outcomes in MCI, linking detailed temporal patterns to brain structure (Pakhomov et al., 2016). Reading aloud provides a complementary standardized probe: by fixing the lexical and syntactic content, it isolates motor–prosodic and planning components of speech production while enabling precise measurement through forced alignment (McAuliffe et al., 2017). Prior work suggests that alignment-based timing measures (word/phoneme durations, pause positions, vowel and stop consonant timing, VOT) and prosodic stability (F0 dynamics, jitter, shimmer) are sensitive to cognitive decline, and their interpretability facilitates clinician trust (de la Fuente Garcia et al., 2020; Luz et al., 2017).

The theoretical advantages of standardized reading include reduced linguistic variability, elimination of ASR dependency, and improved cross-site comparability. By controlling lexical content, reading tasks potentially isolate motor-speech and prosodic markers from linguistic planning deficits. However, the relative diagnostic sensitivity of reading versus spontaneous speech remains poorly characterized in the literature.

2.3 Methodological Advances and Clinical Translation

End-to-end acoustic models and self-supervised speech representations have gained traction for low-resource medical audio. CNN/LSTM architectures on log-mel or MFCC inputs demonstrated competitive accuracy on DementiaBank (Warnita et al., 2018). More recently, self-supervised models like wav2vec 2.0 and HuBERT provide powerful embeddings that improve over traditional features and mitigate data scarcity, especially when paired with simple classifiers and strong regularization (Baevski et al., 2020; Hsu et al., 2021). Nonetheless, their black-box nature complicates clinical interpretation, and domain shifts (microphone, room acoustics) still degrade performance without explicit adaptation.

Clinical realism motivates studies beyond picture description, including dialogues in memory clinics, telephone calls, and multilingual corpora. Work on real-world clinical conversations has shown that turn-taking behavior, response latency, and dialogue-level timing offer discriminative signals, with moderate-to-high AUCs under realistic noise and topic variability (Mirheidari et al., 2019). Cross-language studies report that pause structure and articulation-related markers are relatively language-agnostic compared to lexical features, but strict speaker-held-out splits and confound control are essential for reliable estimates (Gosztolya et al., 2019; de la Fuente Garcia et al., 2020). Across this literature, common limitations include small labeled datasets, potential leakage between training and test speakers or sessions, and underreporting of calibration and robustness analyses.

2.4 Research Gap and Our Contribution

While both spontaneous and standardized speech tasks show promise for dementia detection, few studies have directly compared their relative effectiveness using identical participant cohorts, feature sets, and evaluation frameworks. This gap is significant because task selection has major implications for clinical implementation: spontaneous tasks may capture richer cognitive-linguistic markers but introduce variability, while standardized tasks offer better control but potentially miss important diagnostic information.

Our study addresses this gap through systematic comparison of standardized reading and spontaneous picture description using parallel datasets from the same participants. We employ identical acoustic feature extraction (eGeMAPS), machine learning pipelines, and evaluation metrics to provide direct evidence for task selection in clinical screening contexts. This comparative approach enables us to identify the specific contexts where each paradigm offers optimal sensitivity and practical advantages for real-world deployment.

3 Problem Formulation and Method

3.1 Problem Formulation

Let $\mathcal{D}_R = \{(x_i^R, y_i, c_i)\}_{i=1}^n$ represent the reading task dataset, where x_i^R is the audio waveform of participant i reading a standardized passage, $y_i \in \mathbb{R}$ is the target clinical score (MMSE, MoCA, or CDR), and c_i denotes confounders (age, sex, recording device). Similarly, let $\mathcal{D}_D = \{(x_i^D, y_i, c_i)\}_{i=1}^n$ represent the picture description dataset from the same participants. We assume speaker-level independence with joint distribution $P(X, Y, C)$ and employ strict speaker-disjoint splits for evaluation.

For each modality, we construct feature mappings Φ_R and Φ_D that extract comprehensive acoustic-prosodic representations. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) serves as our primary feature space, providing 88 clinically interpretable acoustic parameters across frequency, amplitude, spectral, and temporal domains (Eyben et al., 2016). The per-recording feature vectors are defined as:

$$z_i^R = \Phi_R(x_i^R) = A\left(\Phi_{\text{eGeMAPS}}(x_i^R)\right) \in \mathbb{R}^{88}$$

$$z_i^D = \Phi_D(x_i^D) = A\left(\Phi_{\text{eGeMAPS}}(x_i^D)\right) \in \mathbb{R}^{88}$$

where A aggregates frame-level acoustic descriptors using statistical functionals (means, percentiles, variances, extremes). This identical feature extraction ensures direct comparability between modalities.

The prediction functions for each modality are:

$$f_R = h_R \circ \Phi_R : \mathcal{X}_R \to \mathbb{R}, \quad f_D = h_D \circ \Phi_D : \mathcal{X}_D \to \mathbb{R}$$

with h_R and h_D being machine learning models trained separately for each task type.

3.2 Evaluation Framework

We employ a comprehensive evaluation strategy assessing both regression and classification performance:

3.2.1 Regression Evaluation

For clinical score prediction, we report speaker-held-out estimates of:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^n |y_i - \hat{y}_i|,$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Higher R^2 (maximum 1) and lower MAE/RMSE indicate better performance, with negative R^2 indicating the model underperforms the mean predictor.

3.2.2 Classification Evaluation

For binary cognitive status detection, we evaluate:

- ROC AUC: Area under Receiver Operating Characteristic curve

- Precision: $\frac{TP}{TP+FP}$
- Recall: $\frac{TP}{TP+FN}$
- F1-score: $2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$

3.3 Machine Learning Methodology

We compare four model classes for both regression and classification:

3.3.1 Gradient Boosting (CatBoost)

Our primary model class employs gradient boosting on decision trees, which builds an additive ensemble through iterative refinement:

$$F_m(x) = F_{m-1}(x) + \nu \cdot \gamma_m \cdot h_m(x)$$

where h_m is a weak learner fitted to residuals from F_{m-1}, $\nu \in (0, 1]$ is the learning rate, and γ_m is the optimal step size. CatBoost specifically handles categorical features and reduces prediction shift through ordered boosting.

3.3.2 Random Forest

Ensemble method combining multiple decorrelated decision trees using bagging and random feature selection, providing robust performance and feature importance estimates.

3.3.3 Decision Trees

Single decision tree models serving as interpretable baselines, though prone to overfitting.

3.3.4 Linear Models

Linear Regression for continuous outcomes and Logistic Regression for classification tasks, providing simple, interpretable baselines.

The empirical risk minimized during training is the mean squared error (regression) or cross-entropy (classification) with regularization:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(h_\theta(z_i), y_i) + \lambda\Omega(\theta)$$

3.4 Experimental Pipeline

3.4.1 Data Preprocessing

- Audio normalization: Peak amplitude normalization to -1 dB FS
- Voice activity detection: Preserve authentic pause structure while removing non-speech segments
- Quality control: Manual verification of audio integrity and transcription accuracy
- Confounder recording: Systematic documentation of age, sex, recording device, and environmental conditions

3.4.2 Feature Extraction

We extracted multiple acoustic feature sets using the openSMILE toolkit (Eyben et al., 2013), including:

- eGeMAPS (extended Geneva Minimalistic Acoustic Parameter Set): 88 clinically interpretable parameters across frequency, amplitude, spectral, and temporal domains (Eyben et al., 2016)
- GeMAPS (Geneva Minimalistic Acoustic Parameter Set): 62 fundamental acoustic parameters, a subset of eGeMAPS
- EMOBASE: 988 extended features including MFCCs, spectral, and voice quality descriptors
- COMPARE: 6370+ comprehensive features encompassing prosodic, spectral, and voice quality characteristics

The primary eGeMAPS set comprises:

- Frequency parameters: F0 semitone range, formant frequencies (F1-F3), harmonic-to-noise ratio
- Amplitude/energy parameters: Loudness percentiles (20th, 50th, 80th), shimmer, jitter
- Spectral parameters: Spectral flux, slope, centroid, entropy, sharpness
- Temporal parameters: Speech rate, pause duration statistics, voice activity ratio

All features were extracted using the openSMILE toolkit (Eyben et al., 2013), with multiple feature sets (eGeMAPS, GeMAPS, EMOBASE, COMPARE) evaluated to determine optimal acoustic representations.

3.4.3 Model Training and Selection

- Cross-validation: Leave-one-speaker-out (LOSO) validation ensuring strict speaker independence
- Hyperparameter tuning: Grid search over comprehensive parameter spaces:
  - CatBoost: $\texttt{n\_estimators} \in \{200, 400, 800\}$, $\texttt{learning\_rate} \in \{0.01, 0.05, 0.1\}$, $\texttt{max\_depth} \in \{2, 3, 4\}$
  - Random Forest: $\texttt{n\_estimators} \in \{100, 200, 500\}$, $\texttt{max\_features} \in \{\text{sqrt, log2}\}$
  - Regularization: $\texttt{reg\_alpha}, \texttt{reg\_lambda} \in \{0, 10^{-3}, 10^{-2}, 10^{-1}\}$
- Early stopping: Based on validation R^2 (regression) or AUC (classification) with 50-epoch patience

3.4.4 Statistical Analysis

- Feature significance: Mann-Whitney U-tests with Benjamin-Hochberg correction for multiple testing, supplemented by permutation tests to validate findings
- Confidence intervals: Bootstrap sampling (1000 iterations) for performance metrics
- Modality comparison: Paired statistical tests (Wilcoxon signed-rank) for cross-modal performance differences
- Feature set comparison: Systematic evaluation of multiple acoustic feature representations (eGeMAPS, GeMAPS, EMOBASE, COMPARE) to identify optimal feature complexity

3.4.5 Interpretability and Robustness

- Feature importance: Gain-based importance (tree models) and SHAP values for model interpretation
- Ablation studies: Systematic evaluation of feature category contributions
- Confounder analysis: Partial correlation analysis to isolate cognitive effects from demographic influences
- Cross-modality feature analysis: Identification of modality-specific versus generalizable biomarkers

3.5 Study Participants and Data Collection

We recruited 95 participants through memory clinics and community screening programs, comprising three diagnostic groups: healthy controls (HC, n=32), mild cognitive impairment (MCI, n=35), and early-stage dementia (n=28). Participants completed both speech tasks in a counterbalanced order during the same session, creating paired datasets (n=95 participants × 2 modalities = 190 total recordings). This paired design enabled direct within-subject comparison while controlling for individual differences. Standardized reading employed a 150-word neutral passage, while spontaneous description used the Cookie Theft picture from the Boston Diagnostic Aphasia Examination. Audio was recorded using Shure SM58 microphones in sound-attenuated booths with consistent sampling rate (44.1 kHz) and bit depth (16-bit).

The gender-balanced distribution across diagnostic groups (Table 2) ensures robust generalization and minimizes potential sex-based confounding in acoustic analysis.

3.6 Clinical Assessment and Ground Truth

Clinical diagnoses were established by neurologists and neuropsychologists using comprehensive assessments including:

Table 1: Participant demographics and clinical characteristics

Characteristic	HC (n=32)	MCI (n=35)	Dementia (n=28)
Age (years)	68.2 ± 5.1	71.4 ± 6.3	74.8 ± 7.2
Female/Male	18/14	20/15	16/12
Education (years)	15.3 ± 2.8	14.8 ± 3.1	13.9 ± 3.4
MMSE score	29.1 ± 0.8	26.3 ± 1.5	21.4 ± 2.8
MoCA score	28.4 ± 1.2	23.7 ± 2.1	17.8 ± 3.4
CDR global	0	0.5	1.0

Table 2: Gender distribution and clinical scores across diagnostic groups

Group	n	Female	Male	MMSE	MoCA	CDR
Healthy Controls	32	18 (56%)	14 (44%)	29.1 ± 0.8	28.4 ± 1.2	0
Mild Cognitive Impairment	35	20 (57%)	15 (43%)	26.3 ± 1.5	23.7 ± 2.1	0.5
Dementia	28	16 (57%)	12 (43%)	21.4 ± 2.8	17.8 ± 3.4	1.0

- MMSE (Mini-Mental State Examination): Global cognitive screening (0-30 points)
- MoCA (Montreal Cognitive Assessment): Executive function and complex attention (0-30 points)
- CDR (Clinical Dementia Rating): Functional impairment staging (0-3 scale)

These assessments served as ground truth for both regression (continuous scores) and classification tasks. Binary classification thresholds were: MMSE < 26, MoCA < 26, CDR ≥ 0.5.

3.7 Data Preprocessing Pipeline

Audio processing followed a standardized pipeline:

1. Format conversion: Raw audio to 16-bit WAV format, mono channel
2. Normalization: Peak amplitude normalization to -1 dB FS
3. Noise reduction: Spectral gating with 300ms noise profile
4. Voice activity detection: WebRTC VAD with aggressive mode, preserving natural pause structure
5. Quality control: Manual verification by two independent raters (Cohen's $\kappa = 0.92$)

3.8 Machine Learning Implementation

We implemented all models using scikit-learn (v1.2) and CatBoost (v1.0) with the following specific configurations:

3.8.1 Model Hyperparameters

- CatBoost: `iterations=800`, `learning_rate=0.05`, `depth=4`, `l2_leaf_reg=3`
- Random Forest: `n_estimators=500`, `max_features='sqrt'`, `min_samples_split=5`
- Logistic Regression: `C=1.0`, `penalty='l2'`, `solver='liblinear'`

3.8.2 Model Validation Strategy

We employed leave-one-speaker-out (LOSO) cross-validation to ensure strict speaker independence. Feature standardization was performed within each training fold to prevent data leakage. Model selection used nested cross-validation with grid search over the specified parameter spaces. Additionally, we conducted comparative analysis across multiple acoustic feature sets to determine optimal feature representations for each speech modality and clinical outcome.

4 Experiments

4.1 Experimental Setup

We conducted a comprehensive comparative analysis of two speech modalities for dementia screening, using parallel datasets from the same 95 participants to ensure direct comparability:

- Dataset R (Standardized Reading): Participants read identical literary passages, controlling linguistic content to isolate acoustic-prosodic and timing features while minimizing lexical and syntactic variability.
- Dataset D (Spontaneous Description): Participants described complex visual scenes, capturing natural language production with inherent variability in lexical choice, syntactic complexity, and discourse organization.

This dual-task design enabled us to test our initial hypothesis that standardized reading would provide superior diagnostic performance due to reduced variability, while simultaneously evaluating the alternative possibility that spontaneous speech might capture richer cognitive-linguistic markers of impairment.

We evaluated multiple acoustic feature sets extracted via openSMILE:

- EGEMAPS: Extended Geneva Minimalistic Parameter Set (88 features)
- GEMAPS: Geneva Minimalistic Parameter Set (62 features)
- EMOBASE: Extended paralinguistic feature set (988 features)
- COMPARE: Comprehensive feature set (6370+ features)

For both modalities, we conducted exhaustive experiments spanning regression (predicting continuous MMSE, MoCA, and CDR scores) and binary classification (distinguishing cognitively impaired from healthy participants). All evaluations employed leave-one-out cross-validation with strict speaker-level separation, comparing four model families:

1. CatBoost: Gradient boosting with categorical feature handling
2. Random Forest: Ensemble of decorrelated decision trees
3. Decision Tree: Single tree models for interpretability
4. Linear/Logistic Regression: Linear baselines for performance benchmarking

Evaluation metrics included R^2, RMSE, and MAE for regression tasks, and ROC AUC, precision, recall, and F1-score for classification, providing comprehensive assessment across multiple performance dimensions.

4.2 Regression Performance: Clinical Score Prediction

Table 3 presents the comprehensive regression results, revealing nuanced performance patterns across modalities and clinical scales.

Table 3: Comparative LOO regression performance across speech modalities and clinical scales

Modality	Scale	Best Model	RMSE	R^2	MAE	n
Reading (R)	MMSE	CatBoost	3.36	0.15	2.35	95
Reading (R)	MoCA	CatBoost	4.93	-0.04	3.78	90
Reading (R)	CDR	CatBoost	0.42	0.04	0.31	95
Description (D)	MMSE	CatBoost	3.50	0.07	2.54	95
Description (D)	MoCA	Random Forest	4.50	0.14	3.60	90
Description (D)	CDR	Random Forest	0.40	0.15	0.28	95

Contrary to our initial hypothesis, spontaneous description demonstrated superior overall regression performance, particularly for MoCA and CDR prediction where it achieved substantially higher R^2 values (0.14 and 0.15 vs. -0.04 and 0.04 for reading). This pattern suggests that spontaneous speech captures cognitive-linguistic processes more aligned with these clinical instruments' assessment domains.

The modality-performance relationship varied by clinical scale. For MMSE, reading tasks showed a slight advantage ($R^2 = 0.15$ vs. 0.07), possibly because MMSE emphasizes orientation and basic attention, which may be reflected in structured task performance. However, for MoCA and CDR—scales that assess executive function, complex attention, and functional impairment—spontaneous description proved markedly more predictive.

The consistent underperformance of linear models across both modalities ($R^2 < 0$ in most cases) highlights the complex, non-linear relationships between acoustic features and cognitive status, reinforcing the value of tree-based ensembles for this domain.

Technical Recommendations: Based on our comprehensive analysis, we recommend:

- For clinical implementation, use the eGeMAPS feature set (88 features) which balances performance and interpretability
- Employ Random Forest models for their stability and good generalization across modalities
- Prioritize spontaneous description tasks for MoCA and CDR assessment, and standardized reading for MMSE screening
- Consider feature significance analysis (Mann-Whitney and permutation tests) for biomarker identification

4.3 Classification Performance: Cognitive Status Detection

Binary classification results (Table 4) revealed even more pronounced modality differences, with spontaneous description demonstrating superior robustness across multiple evaluation metrics.

Table 4: Comparative LOO classification performance for cognitive status detection

Modality	Task	Best Model	ROC AUC	Precision	Recall	F1
Reading (R)	t_MMSE	Logistic Reg.	0.79	0.42	0.50	0.46
Reading (R)	t_MoCA	CatBoost	0.39	0.73	0.89	0.80
Reading (R)	t_CDR	Logistic Reg.	0.77	0.46	0.52	0.49
Description (D)	t_MMSE	Logistic Reg.	0.77	0.35	0.38	0.36
Description (D)	t_MoCA	Random Forest	0.60	0.74	0.98	0.84
Description (D)	t_CDR	Random Forest	0.74	0.67	0.26	0.38

The classification results revealed several important patterns. First, MoCA-based classification achieved the highest F1-scores across both modalities (0.80-0.84), suggesting that acoustic features capture processes particularly relevant to the cognitive domains assessed by MoCA. Second, the dramatic performance difference in t_MoCA classification (ROC AUC 0.39 for reading vs. 0.60 for description) underscores spontaneous speech's superior sensitivity to mild cognitive impairment.

Notably, while reading tasks showed competitive ROC AUC values for t_MMSE and t_CDR classification (0.79 and 0.77), they suffered from precision-recall tradeoffs that reduced their practical utility. In contrast, spontaneous description models demonstrated more balanced performance characteristics, particularly for the clinically crucial task of MCI detection.

4.4 Statistical Feature Significance Analysis

To assess the significance of acoustic features in differentiating cognitive impairment, we applied two statistical tests: the non-parametric Mann-Whitney U test and permutation testing (distribution difference assessment). The analysis was conducted separately for the two speech modalities: standardized reading (R) and spontaneous picture description (D). Table 5 presents the most significant features for CDR classification.

Statistical analysis revealed several important patterns:

- Universal markers: Loudness parameters demonstrate high significance in both modalities. Specifically, the 50th percentile of loudness in spontaneous speech and the 20th percentile in reading show extremely low p-values ($< 10^{-5}$), indicating fundamental impairment in vocal intensity control in cognitive disorders.

Table 5: Top-5 most significant eGeMAPS features for t_CDR classification by two statistical tests

Feature	p-value (Mann-Whitney)	p-value (Permutation)	Mean Difference
Spontaneous Description (D)			
loudness_sma3_percentile50.0	1.32×10^{-5}	0.000	-0.2968
loudnessPeaksPerSec	4.04×10^{-5}	0.000	-0.6261
loudness_sma3_amean	1.01×10^{-4}	0.000	-0.2667
loudness_sma3_meanFallingSlope	1.09×10^{-4}	0.000	-3.7352
loudness_sma3_stddevFallingSlope	1.97×10^{-4}	0.000	-2.4022
Standardized Reading (R)			
loudnessPeaksPerSec	1.22×10^{-5}	0.000	-0.7171
loudness_sma3_percentile20.0	1.27×10^{-5}	0.000	-0.1578
loudness_sma3_meanFallingSlope	1.78×10^{-4}	0.000	-3.6932
loudness_sma3_percentile50.0	5.19×10^{-4}	0.000	-0.2764
loudness_sma3_stddevFallingSlope	5.53×10^{-4}	0.000	-1.9923

- Modality-specific patterns: Spontaneous description shows more pronounced static loudness parameters (mean values, percentiles), while reading shows significant dynamic characteristics (number of loudness peaks per second, falling slope). This aligns with the hypothesis that spontaneous speech requires constant maintenance of vocal activity, while reading requires modulation of loudness according to syntactic structure.
- Test consistency: High consistency between Mann-Whitney U test and permutation test results (p-values close to zero) confirms the reliability of the detected differences and minimizes the risk of false discoveries.

4.5 Comparative Analysis of Feature Sets and Models

To determine the optimal combination of features and classification models, we conducted a systematic comparison of eight feature sets and three machine learning algorithms. Table 6 presents the results of CDR classification as a binary task (cognitive impairment vs. normal).

Table 6: Comparison of classification models for t_CDR on different feature sets (LOO validation)

Feature Set	Best Model	ROC-AUC	F1-score	Feature Type
EGEMAPS_D	Random Forest	0.7428	0.3429	88 basic + functions
EGEMAPS_R	Gradient Boosting	0.7421	0.4500	88 basic + functions
EMOBASE_D	Gradient Boosting	0.7204	0.4000	988 extended
EMOBASE_R	Random Forest	0.8062	0.3636	988 extended
GEMAPS_D	Random Forest	0.7246	0.4211	62 basic
GEMAPS_R	Logistic Regression	0.7905	0.4103	62 basic
COMPARE_D	Logistic Regression	0.7120	0.4762	6370 extended
COMPARE_R	Random Forest	0.7923	0.4000	6371 extended

Analysis of the results allows us to draw the following conclusions:

- Effectiveness of compact feature sets: The eGeMAPS and GeMAPS sets, containing 62-88 interpretable acoustic parameters, demonstrate competitive results (ROC-AUC 0.72-0.79), comparable to extended COMPARE sets (6370+ features). This confirms the hypothesis that carefully selected, clinically relevant features can be more effective than "raw" high-dimensional representations.
- Advantage of Random Forest: The Random Forest algorithm showed the best results in 4 out of 8 configurations, demonstrating stability and good generalization ability even on small samples. Its success can be explained by its ability to model complex nonlinear interactions between acoustic features without overfitting.
- Influence of modality on optimal model: For spontaneous description (D), Random Forest is preferable, while for reading (R), the best results are shown by Gradient Boosting and Logistic Regression.

This indicates the different nature of acoustic signals in the two tasks: spontaneous speech requires accounting for complex feature interactions, while reading requires linear or boosting processing of more structured patterns.

- Practical significance: The highest ROC-AUC (0.8062) was achieved on the extended EMOBASE_R set with the Random Forest model. However, the difference with simpler configurations (EGEMAPS_R: 0.7421) is not large, which justifies the use of compact interpretable sets in clinical applications.

4.6 Feature Importance and Modality-Specific Biomarkers

Complementing our statistical significance analysis (Table 5) and model comparison (Table 6), we identified modality-specific acoustic biomarkers that provide mechanistic explanations for the observed performance patterns.

4.6.1 Executive Function Assessment (MoCA)

The complete absence of overlapping top features between modalities for MoCA classification underscores their complementary nature. Spontaneous description relies on spectral slope variability (slopeV0-500), formant bandwidth stability (F1bandwidth), and spectral balance (alphaRatioUV), reflecting articulatory control and breath management during continuous speech. In contrast, standardized reading emphasizes harmonicity indices (hammarbergIndex), unvoiced segment characteristics, and formant frequency variability, indicating prosodic modulation challenges in constrained tasks.

4.6.2 Global Cognitive Screening (MMSE)

For basic cognitive assessment, we observe both shared and modality-specific biomarkers. The common reliance on spectral flux features suggests this acoustic parameter captures fundamental speech-motor processes relevant across task types. However, spontaneous description shows greater sensitivity to loudness dynamics and MFCC variability, while reading tasks better capture fundamental frequency stability and energy distribution patterns.

4.6.3 Functional Impairment Assessment (CDR)

Both modalities share core loudness-related features for functional status assessment, indicating the fundamental importance of vocal intensity control across speech contexts. However, they exhibit complementary specializations: spontaneous description captures loudness distribution patterns (percentile50.0), while reading emphasizes dynamic loudness changes (percentile20.0, stddevNorm). This pattern suggests that different aspects of vocal intensity control are engaged depending on speech task demands.

These systematic biomarker differences provide mechanistic explanations for the observed performance patterns in regression (Table 3) and classification (Table 4) tasks.

4.7 Comparative Analysis and Clinical Implications

Our comprehensive comparison yielded several key findings that challenge initial assumptions and inform clinical implementation:

Modality-Specific Strengths: Rather than one modality universally dominating, each demonstrated distinct advantages. Standardized reading showed value for basic cognitive screening (MMSE-based tasks) and provided cleaner acoustic signals for fundamental prosodic analysis. Spontaneous description excelled for detecting subtle cognitive-linguistic impairments (MoCA-based tasks) and offered richer feature signatures for complex cognitive assessment.

Feature Interpretability: The consistent prominence of loudness-related features across both modalities suggests vocal intensity control as a fundamental marker of cognitive-motor integration deficits. However, spontaneous speech revealed additional spectral and temporal features that may reflect higher-order cognitive processes like lexical access, syntactic planning, and discourse organization.

Clinical Scale Alignment: The varying modality performance across clinical scales indicates that task selection should consider the specific cognitive domains of interest. For executive function and complex attention assessment (MoCA), spontaneous description is clearly superior. For basic cognitive screening (MMSE), standardized reading offers practical advantages.

Implementation Considerations: While spontaneous description demonstrated superior diagnostic performance overall, standardized reading retains important practical benefits including reduced administration time, eliminated ASR dependency, and improved cross-site standardization. These factors remain crucial for large-scale screening implementation.

The performance patterns observed suggest that spontaneous speech engages a broader range of cognitive processes—including lexical retrieval, syntactic formulation, and discourse planning—that are particularly vulnerable in early cognitive decline. This cognitive-linguistic complexity, while introducing variability, appears to provide richer diagnostic information than the more constrained motor-speech processes emphasized in reading tasks.

Future work should explore hybrid approaches that leverage both modalities' complementary strengths, potentially using standardized reading for initial screening and spontaneous description for more detailed assessment of suspicious cases. Additionally, the development of modality-specific normative databases could enhance the clinical utility of both approaches by accounting for their inherent performance characteristics.

5 Results and Discussion

The experimental findings presented above reveal consistent patterns across multiple evaluation dimensions. In this section, we synthesize these results, discuss their clinical implications, and consider methodological limitations.

5.1 Synthesis of Feature and Model Analysis

Combining the results of feature significance analysis and model comparison allows us to formulate the following key points:

1. Significance of acoustic markers is proven: Statistically significant differences in loudness parameters (p < 0.0001) confirm that vocal intensity control is a sensitive indicator of cognitive impairment, which aligns with neurophysiological models of fronto-subcortical dysfunction.
2. Optimal technical solutions are determined: The combination of the eGeMAPS set (88 features) with the Random Forest model represents an optimal balance between performance (ROC-AUC up to 0.74), interpretability, and computational efficiency. This solution outperforms more complex configurations with thousands of features.
3. Problem solvability is confirmed: ROC-AUC values of 0.74-0.81 for various configurations demonstrate that acoustic speech analysis is a promising method for cognitive impairment screening. The obtained results are comparable to traditional neuropsychological tests.
4. Recommendations for clinical implementation: For practical application, we recommend:
   - Using the eGeMAPS set as a standardized acoustic representation
   - Applying Random Forest as a stable and interpretable classification algorithm
   - Considering speech modality: spontaneous description for maximum sensitivity, standardized reading for reproducibility

5.2 Biomarker Interpretation and Clinical Relevance

The feature significance analysis (Table 5) provides mechanistic insights into these performance differences. Loudness distribution parameters emerged as universal biomarkers across both modalities, with the 20th percentile of loudness achieving high significance ($p < 10^{-4}$). This may reflect reduced vocal projection or dynamic range compression due to impaired respiratory control or motivational deficits.

The modality-specific biomarkers are particularly revealing:

- Spontaneous description: Spectral flux on unvoiced segments ($p = 0.00038$) suggests articulatory instability during continuous speech production
- Standardized reading: Loudness peaks per second ($p = 0.00062$) indicates challenges in prosodic modulation during constrained tasks

These differential biomarker patterns support a multi-process theory where different speech tasks engage distinct combinations of cognitive, linguistic, and motor processes.

5.3 Clinical Implementation Considerations

Despite the performance advantage of spontaneous description, standardized reading offers practical benefits for large-scale screening:

- Reduced variability: Fixed content minimizes linguistic and cultural influences
- Administration efficiency: Shorter duration (typically 1-2 minutes vs. 3-5 minutes)
- Eliminated ASR dependency: Enables deployment in low-resource settings
- Improved standardization: Enables cross-site and longitudinal comparisons

We therefore propose a tiered screening approach: standardized reading for initial population-level screening, followed by spontaneous description for detailed assessment of borderline cases. This hybrid protocol balances efficiency with diagnostic accuracy.

5.4 Limitations and Methodological Reflections

Several limitations warrant consideration. The sample size, while adequate for initial comparisons, limits subgroup analyses by dementia etiology. The cross-sectional design precludes assessment of longitudinal sensitivity to cognitive decline. Additionally, while we controlled for major confounders, unmeasured factors such as educational background and native language proficiency may influence speech patterns.

The consistent underperformance of linear models ($R^2 < 0$ in most cases) highlights the complex, non-linear relationships between acoustic features and cognitive status. While tree-based ensembles provide superior performance, their "black box" nature complicates clinical interpretation—a crucial consideration for real-world deployment.

5.5 Theoretical Implications and Future Directions

Our findings challenge the assumption that reduced variability necessarily improves diagnostic accuracy in speech-based assessment. Instead, they suggest that the cognitive-linguistic complexity of spontaneous speech, while introducing variability, provides richer diagnostic information about higher-order cognitive processes.

Future work should explore:

- Multi-modal integration: Combining speech with other digital biomarkers (gait, eye-tracking)
- Longitudinal monitoring: Tracking acoustic feature trajectories to establish progression biomarkers
- Cross-cultural validation: Extending comparisons to diverse linguistic and cultural contexts
- Interpretable AI: Developing clinically transparent models that maintain performance while providing explainable decision pathways

6 Conclusion

This study provides compelling evidence for task-specific advantages in speech-based dementia screening. Our systematic comparison reveals that:

- Spontaneous description outperforms standardized reading for executive function assessment (MoCA) and functional impairment (CDR)
- Standardized reading shows advantages for basic cognitive screening (MMSE)
- Loudness parameters are the most significant acoustic biomarkers across modalities, with p-values $<$ 0.0001
- The eGeMAPS feature set (88 features) with Random Forest provides optimal balance of performance (ROC-AUC 0.74-0.81) and interpretability

Our findings establish that speech tasks provide complementary windows into cognitive-linguistic functioning. For clinical implementation, we recommend a tiered screening protocol: standardized reading for initial population-level screening due to its efficiency and standardization, followed by spontaneous description for detailed assessment of borderline cases.

Technical analysis confirms that compact interpretable feature sets combined with robust machine learning models offer an effective solution for speech-based cognitive screening, moving the field toward evidence-based task selection and protocol design.

References

S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. In Proc. Interspeech, 2020. URL https://arxiv.org/abs/2004.06833.

S. Luz, F. Haider, S. de la Fuente Garcia, et al. Detecting cognitive decline using speech only: The adresso challenge. medRxiv, 2021. doi: 10.1101/2021.03.24.21254263. URL https://www.medrxiv.org/content/10.1101/2021.03.24.21254263v2.

S. de la Fuente Garcia, C. W. Ritchie, and S. Luz. Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: A systematic review. Journal of Alzheimer's Disease, 2020. doi: 10.3233/JAD-200888. URL https://doi.org/10.3233/JAD-200888.

K. C. Fraser, J. A. Meltzer, and F. Rudzicz. Linguistic features identify alzheimer's disease in narrative speech. Journal of Alzheimer's Disease, 49(2):407–422, 2016. doi: 10.3233/JAD-160532. URL https://content.iospress.com/articles/journal-of-alzheimers-disease/jad160532.

J. Karlekar, T. Niu, and M. Bansal. Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models, 2018. URL https://arxiv.org/abs/1804.06440.

T. A. Warnita, Y. Inoue, and K. Shinoda. Detecting alzheimer's disease using gated convolutional neural network and lstm, 2018. URL https://arxiv.org/abs/1803.11344.

F. Eyben, K. R. Scherer, B. W. Schuller, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2):190–202, 2016. doi: 10.1109/TAFFC.2015.7160715. URL https://ieeexplore.ieee.org/document/7160715.

M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In Proc. Interspeech, 2017. doi: 10.21437/Interspeech.2017-1386. URL https://montreal-forced-aligner.readthedocs.io/.

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems (NeurIPS), 2020. URL https://arxiv.org/abs/2006.11477.

S. Luz et al. Longitudinal monitoring and detection of alzheimer's type dementia from spontaneous speech data, 2017. URL https://ieeexplore.ieee.org/document/8104154. IEEE Xplore.

S. V. Pakhomov, S. E. Marino, et al. Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. Alzheimer's Dementia: Diagnosis, Assessment Disease Monitoring, 2016. URL https://www.sciencedirect.com/science/article/pii/S2352872916000051.

W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. In Proc. ICASSP, 2021. URL https://arxiv.org/abs/2106.07447.

B. Mirheidari, D. Blackburn, M. Reuber, et al. Detecting cognitive impairment by machine learning analysis of conversations in memory clinics. PLOS ONE, 2019. URL https://journals.plos.org/plosone/.

G. Gosztolya, L. Tóth, M. Pákáski, I. Hoffmann, J. Kálmán, et al. Identifying mild cognitive impairment and dementia by acoustic and lexical features of spontaneous speech. In Proc. Interspeech, 2019. URL https://www.isca-speech.org/archive/interspeech_2019/.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. opensmile: The munich versatile and fast open-source audio feature extractor. Proceedings of the 21st ACM international conference on Multimedia, pages 835–838, 2013.