

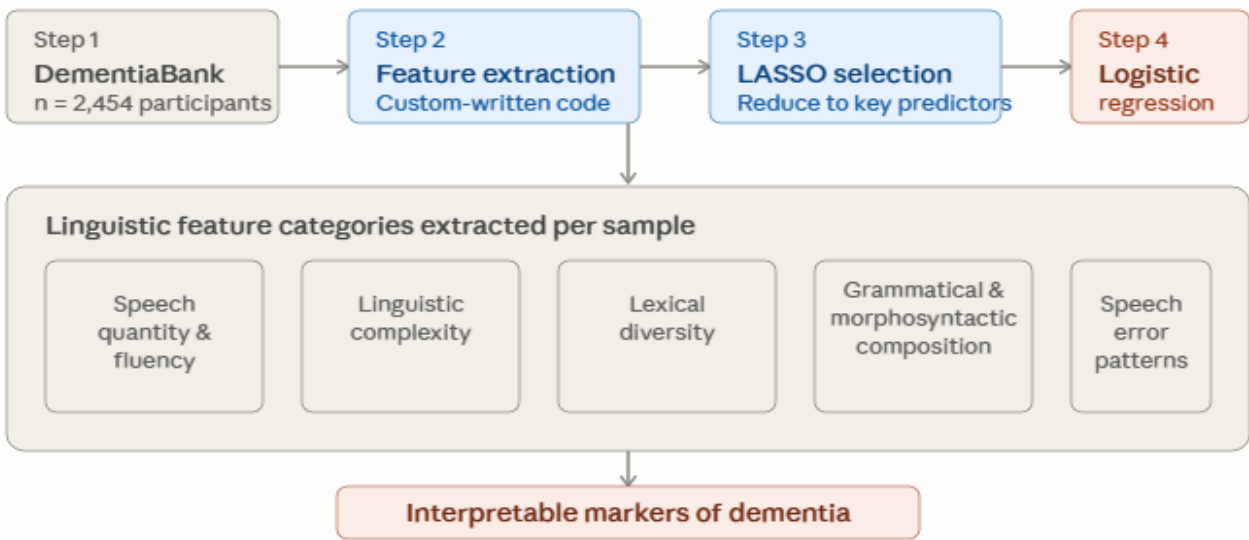
Arjun Venkateswaran¹, Vikesh Patel², and Auriel Willette³ ¹Rutgers University, Department of Neurology, Kendall Park, NJ, USA ; ² Woodbridge Academy Magnet School, Edison, NJ, USA; ³Rutgers University Department of Neurology, New Brunswick, NJ, USA

- 1.Reduced lexical diversity emerged as the strongest speech-based predictor of dementia diagnosis.
- 2.Fewer utterance-level speech errors were associated with increased dementia likelihood, an unexpected inverse relationship.
- 3.Elevated word-level error rates were modestly associated with greater odds of dementia.
- 4.The ratio of content to function words showed no significant association with dementia status.

INTRODUCTION

Speech and language impairments are among the earliest observable indicators of dementia, making computational analysis of speech a promising, non-invasive marker for identifying cognitive decline. All-cause dementia is known to affect lexical diversity, grammatical structure, and speech fluency, yet many prior studies rely on limited linguistic features or small sample sizes, restricting how well their findings generalize. This study investigated three key questions:

- Can large-scale computational analysis predict dementia?** This study tested whether a comprehensive, automated analysis of speech recordings could predict the likelihood of an all-cause dementia diagnosis using linguistic features as predictors.
- Which speech features carry the most information?** Speech recordings from 2,454 participants drawn from the DementiaBank database (TalkBank) — including individuals with all-cause dementia and cognitively unimpaired controls — were analyzed using custom-written code to extract a comprehensive set of linguistic measures, including speech quantity and fluency, linguistic complexity, lexical diversity, grammatical and morphosyntactic composition, and speech error patterns.
- How can the most informative predictors be isolated?** Machine learning–based feature selection (LASSO regularization) was applied to reduce this large feature set down to the linguistic markers most strongly associated with dementia diagnosis, allowing for an interpretable, data-driven model.



Developing scalable tools for early dementia detection is an especially urgent public health priority as populations age and the burden of Alzheimer’s disease and related dementias continues to grow. While clinical diagnosis often depends on costly neuroimaging or invasive biomarker testing, speech is fast, inexpensive, and can be collected remotely. This study leverages a large, publicly available speech database to identify which linguistic features most reliably distinguish dementia from normal cognitive aging, laying the groundwork for speech-based screening tools that could complement — not replace — existing diagnostic approaches.

METHODS

A. Data Source & Sample
Speech recordings and transcripts were obtained from the DementiaBank database, a publicly available section of TalkBank containing spontaneous speech samples from older adults. The final sample included 2,454 participants, comprising individuals with a clinical diagnosis of all-cause dementia and cognitively unimpaired controls. All transcripts were de-identified prior to analysis.

B. Speech & Language Feature Extraction
Custom-written code was used to extract a comprehensive set of speech and language measures from each transcript, spanning five domains:

- Speech quantity and fluency** — total utterances, word counts, pause patterns, and speech rate.
- Linguistic complexity** — sentence length, syntactic complexity, and use of subordinate clauses.
- Lexical diversity** — variety of vocabulary used relative to speech length, including frequency-weighted type-token ratio (FREQ_TTR).
- Grammatical and morphosyntactic composition** — proportion of open-class (nouns, verbs) versus closed-class (function) words, and grammatical structure.
- Speech error patterns** — utterance-level errors, word-level errors, and revision/repair behaviors during speech production.

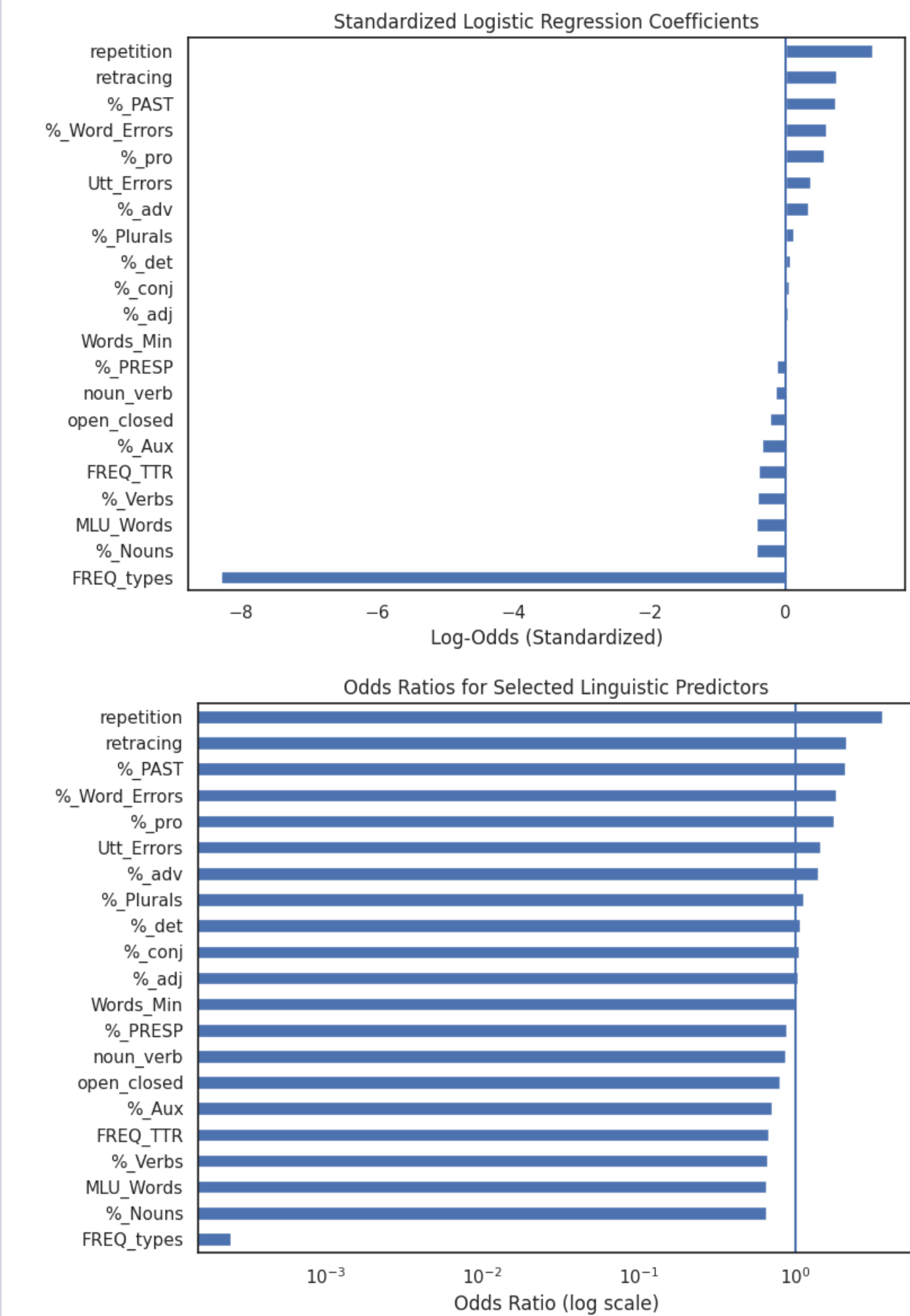
C. Feature Selection & Statistical Analysis
Because the extracted feature set was large and many measures were correlated, machine learning–based feature selection (LASSO regularization) was applied to identify the subset of linguistic features most informative for predicting dementia diagnosis. Selected features were then entered into a logistic regression model to estimate odds ratios (OR) and 95% confidence intervals for each predictor’s association with dementia status. Model fit was assessed using McFadden’s Pseudo R², and multicollinearity among predictors was evaluated using variance inflation factors (VIF).

Feature domain	Example variables
Speech quantity & fluency	Total utterances, word count, speech rate, pause frequency
Linguistic complexity	Mean sentence length, subordinate clause use, syntactic depth
Lexical diversity	Frequency-weighted type-token ratio (FREQ_TTR)
Grammatical & morphosyntactic	Open/closed-class word ratio, part-of-speech distribution
Speech error patterns	Utterance-level errors, word-level error percentage

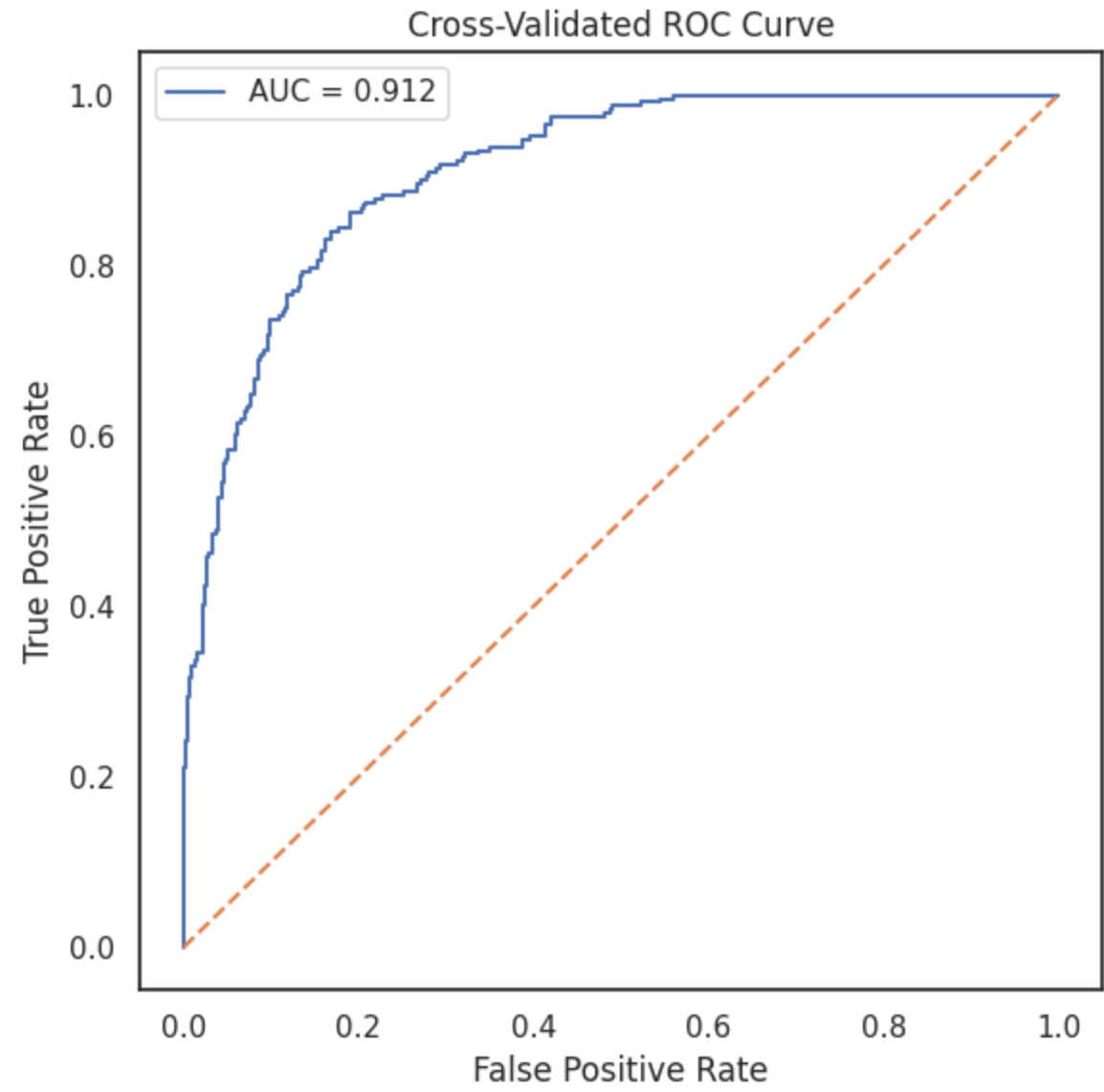
RESULTS

A. Feature Selection & Predictor Rankings
LASSO regularization reduced the full feature set to a subset of the most informative predictors of dementia diagnosis. Standardized logistic regression coefficients revealed that:

- Repetition, retracing, and past-tense usage (%_PAST)** showed the strongest positive associations with dementia
- Word-level and utterance-level error rates** were also meaningfully associated with diagnosis
- Type-token frequency (FREQ_types)** showed the largest negative coefficient by a wide margin, reflecting its strong inverse relationship with dementia likelihood
- Several morphosyntactic features (%_det, %_conj, %_adj) contributed minimally, consistent with their exclusion from the final reduced model

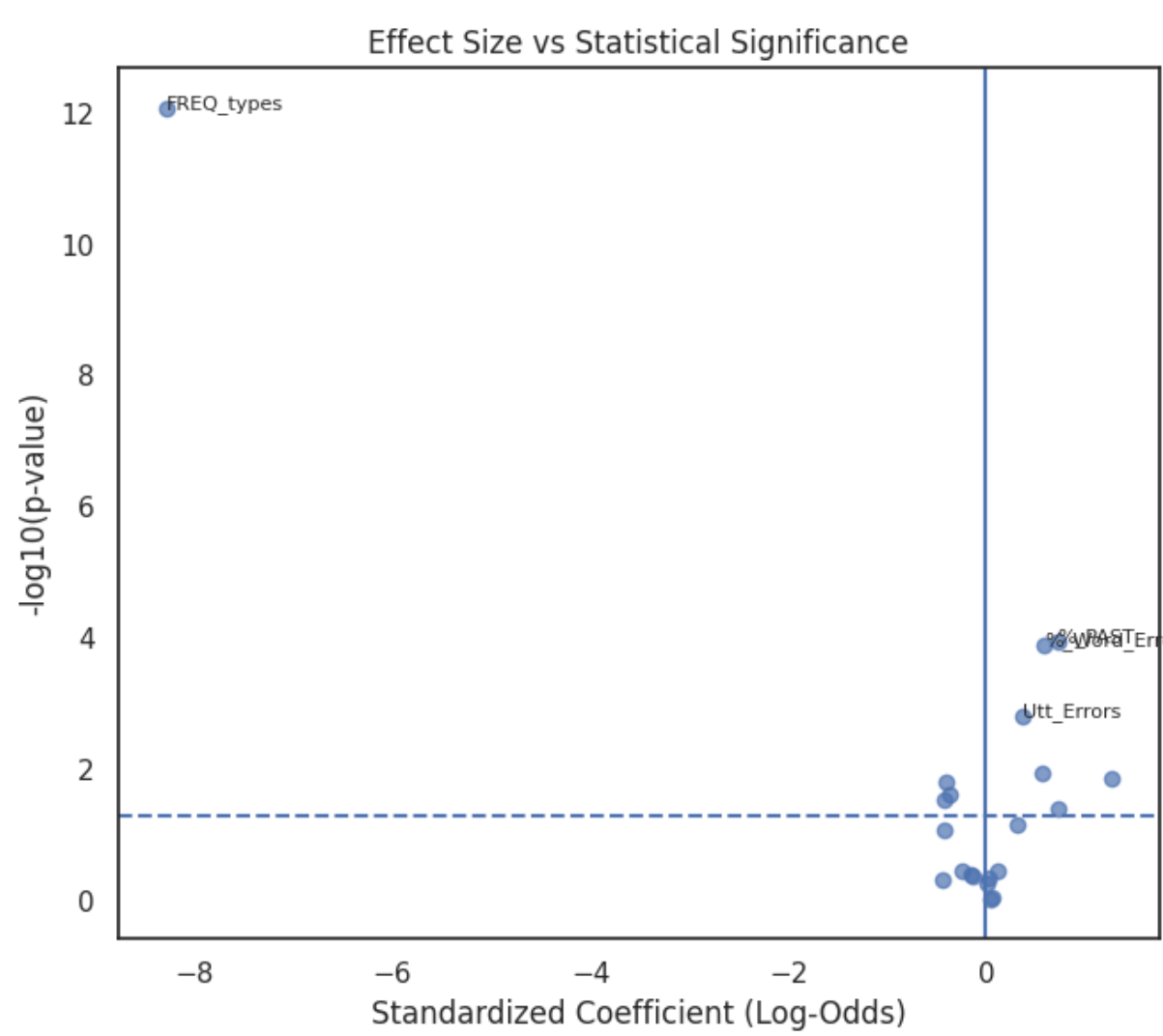


- B. Model Performance**
- To evaluate how well the selected linguistic features distinguished dementia cases from cognitively unimpaired controls, model performance was assessed using cross-validation.
- The model achieved strong discriminative ability, with a cross-validated **AUC of 0.912**, indicating that speech-based linguistic features reliably differentiate dementia status
 - Predicted probabilities were consistently higher among individuals with a true dementia diagnosis compared to unimpaired controls, though some overlap was observed, reflecting natural heterogeneity in speech presentation

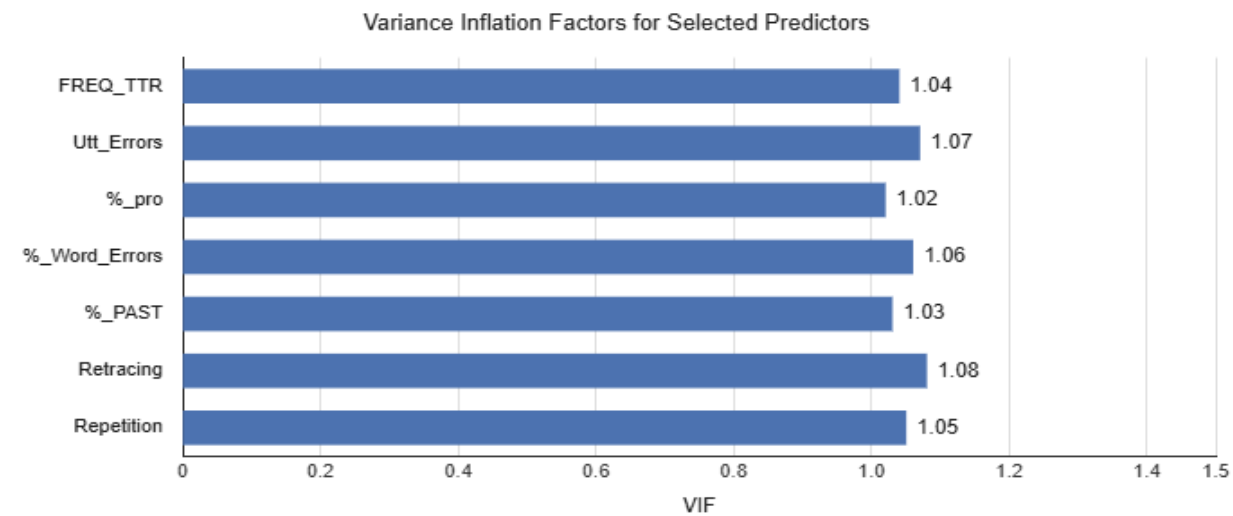


RESULTS (CONT.)

- C. Effect Size & Statistical Significance**
To jointly evaluate the strength and reliability of each predictor, standardized coefficients were plotted against their statistical significance.
- FREQ_types** stood out as both the strongest and most statistically significant predictor, with a highly negative coefficient and the lowest p-value in the model
 - %_PAST, %_Word_Errors, and Utt_Errors** were also statistically significant, falling above the significance threshold line
 - Most remaining features clustered near zero effect size and did not reach statistical significance, supporting the model’s reliance on a small set of robust predictors rather than diffuse contributions across many features



- D. Model Fit and Multicollinearity Diagnostics**
Beyond discrimination and predictor significance, the reliability of a regression model depends on whether its predictors are independent of one another and whether the model captures a meaningful share of the outcome’s variability.
- Multicollinearity was minimal across all retained predictors.** Variance inflation factors (VIFs) for every feature in the final model were below 1.1 — far under the commonly used threshold of 5 used to flag problematic collinearity. This confirms that each selected linguistic feature (e.g., repetition, FREQ_TTR, %_Word_Errors) contributes distinct, non-overlapping information rather than several features capturing the same underlying signal.
 - Explained variance was modest but consistent with typical logistic regression models of this kind.** McFadden’s Pseudo R² was 0.133, which — as discussed in the Discussion section — should be interpreted alongside AUC rather than in isolation. A modest Pseudo R² alongside a high AUC (0.912) indicates a model that classifies cases accurately even though no single feature (or the feature set overall) accounts for a large proportion of outcome variance — a common and expected pattern in behavioral and speech-based prediction models.



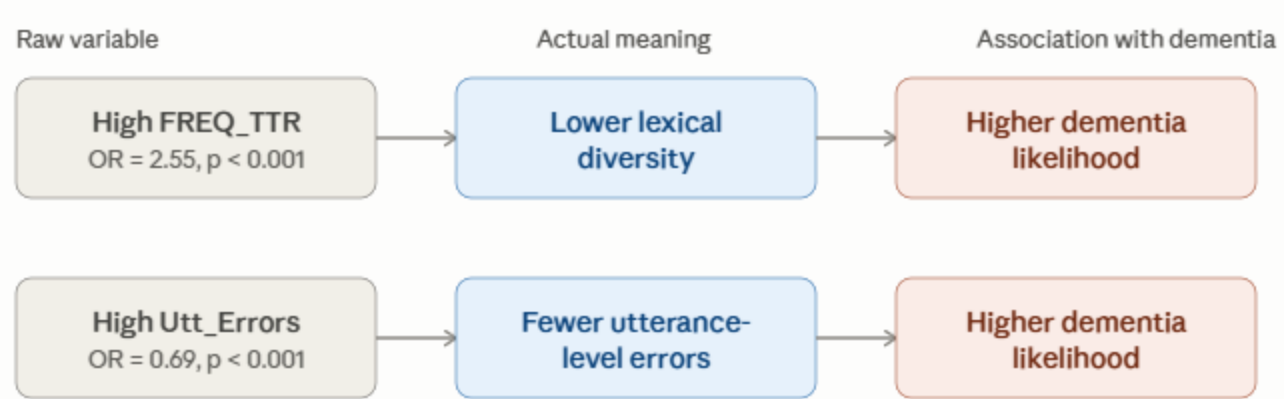
DISSCUSSION

This study demonstrates that large-scale computational analysis of spontaneous speech can identify interpretable linguistic markers associated with all-cause dementia, supporting speech as a viable, scalable, and non-invasive screening modality. By analyzing 2,454 participants — a substantially larger sample than most prior speech-based dementia studies — this work extends earlier findings that were often limited by small cohorts or a narrow set of linguistic features, offering a more robust and generalizable picture of how speech changes with cognitive decline.

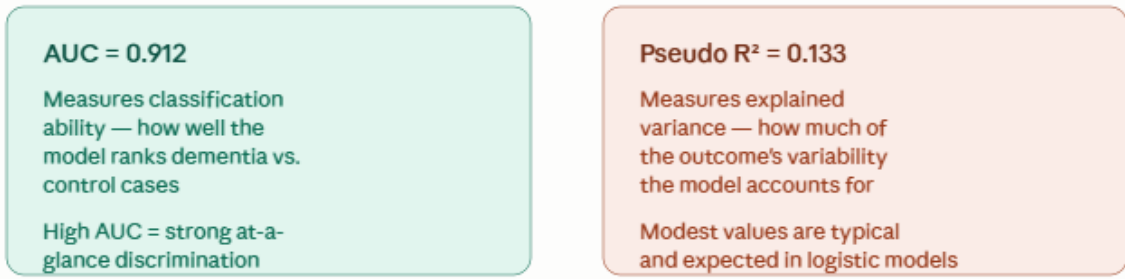
Interpreting the key predictors

- Lexical diversity was the strongest signal in the model.** Frequency-weighted lexical diversity (FREQ_TTR) showed the strongest association with dementia risk (OR = 2.55, 95% CI: 2.26–2.88, p < 0.001). Due to legacy coding conventions, a *higher* FREQ_TTR value reflects a *lower* lexical diversity — meaning participants who relied on a narrower, more repetitive vocabulary were substantially more likely to have a dementia diagnosis. This is consistent with prior literature linking vocabulary impoverishment and word-retrieval difficulty to early cognitive decline.
- Fewer utterance-level errors were linked to higher dementia odds — an unexpected but explainable pattern.** Because of the same coding convention, a higher Utt_Errors value corresponds to fewer errors, and this was associated with increased dementia likelihood (OR = 0.69, 95% CI: 0.62–0.76, p < 0.001). This likely reflects a shift in speech style rather than improved language ability — individuals with dementia may produce shorter, simpler, more formulaic utterances that are structurally less complex and therefore present fewer opportunities for detectable errors. “Fewer errors” in this context may signal reduced linguistic ambition rather than preserved linguistic skill.

DISCUSSION (CONT.)



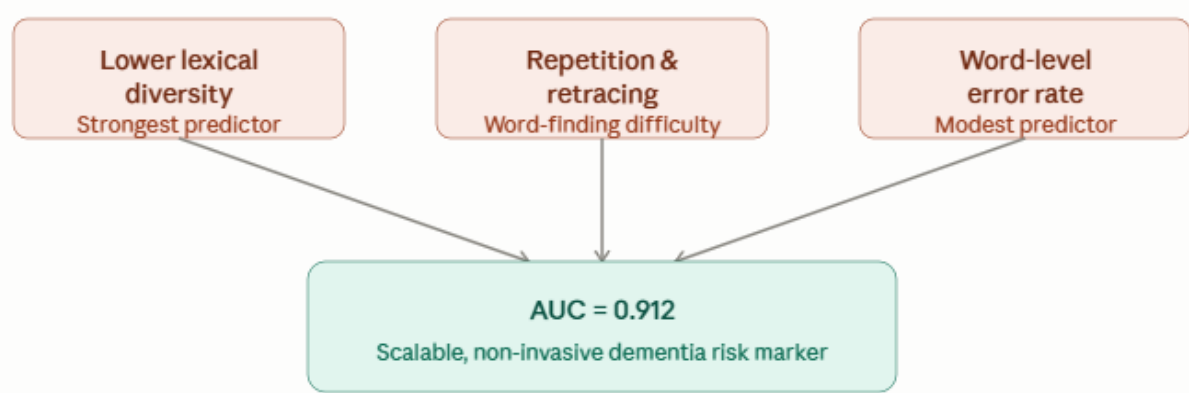
- Word-level errors showed the expected, more intuitive direction.** A higher percentage of word-level errors was modestly associated with greater dementia likelihood (OR = 1.22, 95% CI: 1.05–1.43, p = 0.012), consistent with the idea that lexical substitution and word-finding errors increase as cognitive decline progresses. Unlike utterance-level errors, this measure was not subject to the same coding reversal, making it more directly interpretable.
 - Repetition and retracing emerged as strong, clinically consistent markers.** Increased repetition and speech retracing were among the largest positive coefficients in the standardized model (see Image 1, Results section). Clinically, this aligns well with known dementia presentations — word-finding difficulty and working memory limitations often manifest as repeated words, phrases, or restated sentences during spontaneous speech.
 - Broad grammatical composition was largely uninformative.** The open/closed-class word ratio was not significantly associated with dementia (OR = 0.93, 95% CI: 0.80–1.08, p = 0.366), suggesting that coarse grammatical category usage is a less sensitive marker than finer-grained lexical and fluency measures. This is a useful negative finding — it helps narrow down which categories of speech features are worth prioritizing in future screening tools, rather than treating all linguistic features as equally informative.
- Model performance in context**
- The model demonstrated strong discriminative ability, with a cross-validated AUC of 0.912 (Image 3), indicating that a relatively small set of linguistic features can meaningfully separate dementia cases from cognitively unimpaired controls. This is a notably strong result for a model built entirely from automatically extracted speech features, without any neuroimaging or biomarker input.
 - Despite this strong discrimination, the model’s McFadden’s Pseudo R² was comparatively modest (0.133). This is not a contradiction — AUC and Pseudo R² measure different things, and it’s worth clarifying this distinction explicitly on the poster, since viewers unfamiliar with logistic regression metrics may otherwise read the modest R² as a weakness.



- Both metrics are valid — they simply answer different questions**
AUC and Pseudo R² are complementary metrics that answer different questions. AUC measures how well the model ranks dementia cases against controls, while Pseudo R² measures the proportion of outcome variance explained by the model. Both are valid, but they answer different questions.
- Markers each contribute unique, non-redundant information to the model, rather than several features capturing the same underlying signal.** This strengthens confidence that each retained predictor reflects a genuinely distinct aspect of speech affected by dementia.
- Limitations**
 - The DementiaBank sample, while large relative to prior speech-based dementia studies, still reflects a specific population, language, and recording context, which may limit generalizability to other languages, dialects, or clinical settings.
 - All-cause dementia was treated as a single diagnostic category in this analysis. Future work should examine whether these linguistic markers differ across dementia subtypes (e.g., Alzheimer’s disease vs. vascular dementia vs. frontotemporal dementia), since underlying speech and language deficits may vary by etiology.
 - The data are cross-sectional, meaning this model captures a snapshot association between speech and current dementia status. It cannot establish whether these linguistic patterns predict *future* decline or only reflect impairment severity at the time of recording.
 - As with any feature-selection approach, LASSO regularization prioritizes predictive strength and may exclude features that are clinically meaningful but statistically weaker or collinear with stronger predictors — meaning the excluded features are not necessarily uninformative, only less independently informative in this particular model.
- Clinical and research implications**
 - These findings support the feasibility of speech-based digital biomarkers as a low-cost, scalable complement to existing diagnostic workflows, which often depend on costly neuroimaging or invasive biomarker testing.
 - Because speech samples can be collected remotely — over the phone, through a mobile app, or during routine telehealth visits — this approach could support more frequent, longitudinal monitoring of at-risk populations than is currently feasible with imaging-based methods.
 - The interpretability of the selected features (e.g., lexical diversity, repetition, error rates) means this approach could eventually be paired with clinician-facing tools that explain *why* a given speech sample was flagged as high-risk, rather than functioning as an opaque black-box classifier.

CONCLUSION

- This study shows that computational analysis of naturalistic speech can reveal robust, interpretable linguistic markers of all-cause dementia, most notably reduced lexical diversity, increased repetition and retracing, and altered error patterns
- The model achieved strong discriminative performance (AUC = 0.912), suggesting these features carry meaningful diagnostic signal even when derived from unstructured, spontaneous speech
- Speech-based screening approaches are unlikely to replace clinical or biomarker-based diagnosis, but may offer a scalable, non-invasive tool for early risk identification when used alongside other assessment methods
- Future work should validate these markers in independent samples, explore differences across dementia subtypes, and test longitudinal speech change as a predictor of cognitive decline over time



REFERENCES

Badal, V. D., Reinen, J., Twamley, E. W., et al. (2024). Investigating acoustic and psycholinguistic predictors of cognitive impairment in older adults: Modeling study. *JMIR Aging*, 7, e52145. <https://doi.org/10.2196/52145>

Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer’s disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6), 585–594. [Data retrieved via TalkBank DementiaBank]

BioRender. (2026). *BioRender* (Version 1.0) [Computer software]. <https://www.biorender.com>

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12), 2767–2796. <https://doi.org/10.1093/cercor/bhp026>

Braak, H., & Braak, E. (1991). Neuropathological staging of Alzheimer-related changes. *Acta Neuropathologica*, 82(4), 239–259. <https://doi.org/10.1007/BF00308809>

Chen, X., De Bodt, M., & Vandenbosche, J. (2024). Connected-speech tasks in mild cognitive impairment detection: A comparative study. *Journal of Clinical Medicine*, 13(4), 1102. <https://doi.org/10.3390/jcm13041102>

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, 15(6), 254–262. <https://doi.org/10.1016/j.tics.2011.04.003>

Eyigoz, E., Mathur, S., Santamaria, M., Cecchi, G., & Naylor, M. (2020). Linguistic markers predict onset of Alzheimer’s disease. *EBioMedicine*, 62, 103065. <https://doi.org/10.1016/j.ebiom.2020.103065>

Hickok, G. (2009). The functional neuroanatomy of language. *Physics of Life Reviews*, 6(3), 121–143. <https://doi.org/10.1016/j.phlrev.2009.06.001>

Huang, Y., Li, X., Zhang, Y., & Wang, J. (2024). Automated speech analysis for the detection of mild cognitive impairment: A systematic review. *Ageing Research Reviews*, 94, 102187. <https://doi.org/10.1016/j.arr.2024.102187>

Kljajevic, V. (2014). *Comprehension of complex sentences in Alzheimer’s disease*. Springer. <https://doi.org/10.1007/978-3-319-02634-3>

Lanzi, A. M., Saylor, A. K., Fromm, D., Liu, H., MacWhinney, B., & Cohen, M. (2023). DementiaBank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*, 32(1), 12–25. https://doi.org/10.1044/2022_AJSLP-22-00281

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Lawrence

CONTACT

Arjun Venkateswaran
arjunvenkateswaran329@gmail.com
732-823-2314
Vikesh Patel
vikesh1212@hotmail.com
201-450-8545

