Early Detection of Dementia Using Multimodal Speech and Text Analysis: A Cognitive Science-Informed Machine Learning Approach

Royce Salah
College of Computing
Georgia Institute of Technology
San Jose, CA, United States of America
https://orcid.org/0009-0004-7415-7083

***Abstract*—Early detection of dementia remains a critical public health challenge due to gradual onset and high rates of underdiagnosis. Traditional clinical assessments often fail to capture subtle cognitive markers until the disease is well-advanced. This paper presents a multimodal machine learning pipeline for detecting early signs of dementia through analysis of both transcribed speech and acoustic features in audio files. Utilizing the DementiaNet dataset, composed of public interview audio clips from celebrities later diagnosed with dementia and healthy controls, features such as lexical entropy, semantic drift, and acoustic prosody were extracted. The resulting late-stage fusion model achieved an AUC of 0.605 ± 0.07 across 10-fold speaker-wise cross-validation, indicating modest but consistent predictive power. Feature importance analysis revealed lexical diversity and semantic metrics as more predictive than acoustic counterparts. Findings suggest that even lightweight large language models (LLMs), when paired with speech analysis, can capture subtle linguistic and paralinguistic cues indicative of cognitive decline. This work provides a foundation for scalable, privacy-preserving early detection systems embedded in consumer devices.**

I. Introduction

A. Context and Motivation

Dementia remains drastically underdiagnosed, with studies estimating that nearly 59% of individuals over the age of 65 are unaware of their condition [1]. This underdiagnosis stems from a combination of systemic barriers including limited access to healthcare, disparities in patient-provider communication, educational gaps, and resource constraints which collectively contribute to delayed or missed diagnoses [2]. As a result, many patients are not identified until the disease has progressed beyond the point at which treatment is most effective.

This delay is particularly consequential in the case of Alzheimer's disease (AD), where pharmaceutical interventions such as lecanemab-irmb (Leqembi) and donanemab-azbt (Kisunla) have shown efficacy only during the early stages of disease progression. These interventions do not reverse, but instead aim to slow, cognitive decline thus emphasizing the necessity of early and accurate detection. Similarly, non-pharmaceutical interventions such as diet modification, physical activity, and social engagement have been shown to mitigate risk but are most beneficial when impleme nted proactively.

In this context, early detection is not just advantageous, but critical. Addressing this gap could tap the potential for timely clinical interventions, improved quality of life, and reduced long-term care burdens. Emerging computational methods, particularly those leveraging natural language processing, offer a promising pathway toward scalable, non-invasive screening tools capable of identifying subtle cognitive changes well before traditional diagnosis.

B. Problem Statement

Subtle linguistic and acoustic changes that signal early cognitive decline are often too nuanced to be reliably detected through traditional clinical assessments. As a result, these early markers are frequently overlooked, delaying formal diagnosis and intervention during the critical early stages of dementia. Whilst prior research has demonstrated the potential of lexical and acoustic analysis, particularly in Alzheimer's Disease, for identifying such early signs, most approaches have relied on hand-crafted features or rule-based modeling techniques.

Recent advances in natural language processing, especially through large language models (LLMs), offer new opportunities to detect these deviations with greater sensitivity. This work seeks not only to replicate the predictive capabilities of previous approaches, but to extend them by leveraging multimodal analysis to identify individuals at risk of dementia as much as 5 to 15 years before clinical diagnosis.

C. Cognitive Science Justification

This work is grounded in well-established cognitive theories of semantic memory degradation and lexical retrieval impairment in dementia. Prior research has found that longer between-utterance pauses, and reduced speech rate are correlated with early tau pathology throughout adulthood [3]. Similarly, progressive declines in the cognitive task of lexical retrieval in dementia patients have been found to be associated with grammatical and syntactic deficits, particularly reduced verb inflection and lexical variation are prevalent in AD [4] [5].

These impairments are consistent with cognitive science literature on semantic drift and lexical entropy, which have

emerged as algorithmic frameworks for cognitive load, coherence, and semantic memory issues. As other research has shown, computational natural-language-based, linguistic analysis can serve as a viable early diagnostic tool with the correct feature extraction [6].

II. Model Design and Tool Selection

A. Data and Tooling

1) Dataset: Dementia, as a well-documented form of cognitive decline, lies at the intersection of a wide range of cognitive science concepts like language, memory, and executive function. In approaching this research question, the focus was on identifying subtle linguistic and acoustic markers that might precede clinical diagnosis by several years. Given the limited availability of open-source clinical datasets, the DementiaNet corpus was selected for its relevance to this goal. It contains audio recordings from public interviews with celebrities later diagnosed with dementia, spanning 5 to 15 years prior to formal diagnosis. These limited longitudinal samples, alongside control interviews from healthy individuals, provide a rare opportunity to examine early signs of cognitive decline in natural speech.

2) Tool Selection: Due to the dataset's audio corpus, this allows for the opportunity to perform a multimodal analysis on both the audio itself and transcribed text. Accordingly, the tool selection process was guided by a structured pipeline aligned with the central research goal. The pipeline was divided into three key stages: data preparation, feature extraction, and model training.

To analyze lexical features, speech audio needed to be accurately transcribed into text. Given the importance of preserving natural speech variability including disfluencies and pauses, OpenAI's *Whisper* automatic speech recognition (ASR) system was selected. Trained on 680,000 hours of multilingual and noisy data, Whisper offered high transcription accuracy, robustness to varied recording conditions, and minimal need for manual correction which made it ideal for downstream analysis.

The transcribed text and original audio formed the basis for multimodal feature extraction. Lexical features were derived from two primary sources: algorithmically calculated metrics (e.g., type-token ratio, pronoun ratio), and large language model (LLM)-based representations. To compute semantic drift, the *all-MiniLM-L6-v2* sentence transformer was used due to its efficiency and ability to encode semantic meaning in numerical 384-dimension vectors. Cosine similarity between adjacent sentence embeddings quantified shifts in meaning, offering a estimation for coherence loss typically observed in early cognitive decline. For lexical entropy, OpenAI's GPT-2 small was employed to generate next-token probability distributions across transcripts. These probabilities, when passed through an information theory entropy formula, quantified linguistic unpredictability. This served as an indicator of semantic disorganization or impaired coherence.

On the acoustic side, features were extracted using the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), which includes prosodic markers commonly linked to cognitive and emotional states. The openSMILE toolkit was chosen to implement this, offering a reproducible and well-supported method for extracting eGeMAPS features such as pitch, jitter, loudness, and speech rate, all of which have been associated with neurodegenerative changes. With both lexical and acoustic features extracted, modality-specific classifiers were trained and later fused via a meta-classifier, the tools of which were selected based on input data and predictive performance.

B. Concept Derivation and Mapping

Several cognitive science concepts serve as the foundation for the model pipeline and in the identification of extractable markers of cognitive deterioration. Primary markers and concepts were identified through a literature review of several studies of dementia-related symptoms, their linkage to cognitive impairment, and their manifestations in speech. Some of the most telling signs of cognitive decline in dementia can be detected through changes in speech content and vocal delivery, though they can be elusive, especially 5 to 15 years prior to a conventional diagnosis. Studies have found that many features of cognitive strain and decline can be observed in individuals already diagnosed with mild cognitive impairment (MCI) with traditional natural language processing techniques such as verb and noun extraction or sentence counts. Recent LLMs enable detection of more subtle cognitive disruptions, particularly in cases where semantic complexity obscures topic coherence or abstract concept linkage. Cognitive strain in patients with dementia can manifest as anomia [12] and tangentiality [6] among others. These are the main principles which will attempt to be addressed with this model pipeline and its integration of modern LLMs.

C. Pipeline Architecture and Impemlentation

The architecture of the model pipeline was structured into three main stages: data pre-processing, feature extraction, and model training, as outlined in Fig. 1.

1) Data Pre-Processing: The pipeline begins with ingestion and organization of the DementiaNet dataset, which includes 84 dementia cases and 62 controls, totaling 131 and 153 audio clips respectively. Audio files were transcribed using the Whisper automatic speech recognition (ASR) system to enable downstream lexical feature extraction.

2) Text Feature Extraction: Transcribed text was used to derive both advanced and traditional linguistic features. Primary features included semantic drift and lexical entropy, supported by several additional natural language metrics.

Semantic drift was computed using the HuggingFace Sentence Transformer framework with the all-MiniLM-L6-v2 model to convert each sentence to its vector-space representation. Each sentence was transformed into a 384-dimensional vector. Mean cosine similarity, as defined in Equation (1), was calculated between consecutive non-zero vectors to approximate coherence and conceptual continuity.

$$\text{cosine similarity} = S_C(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

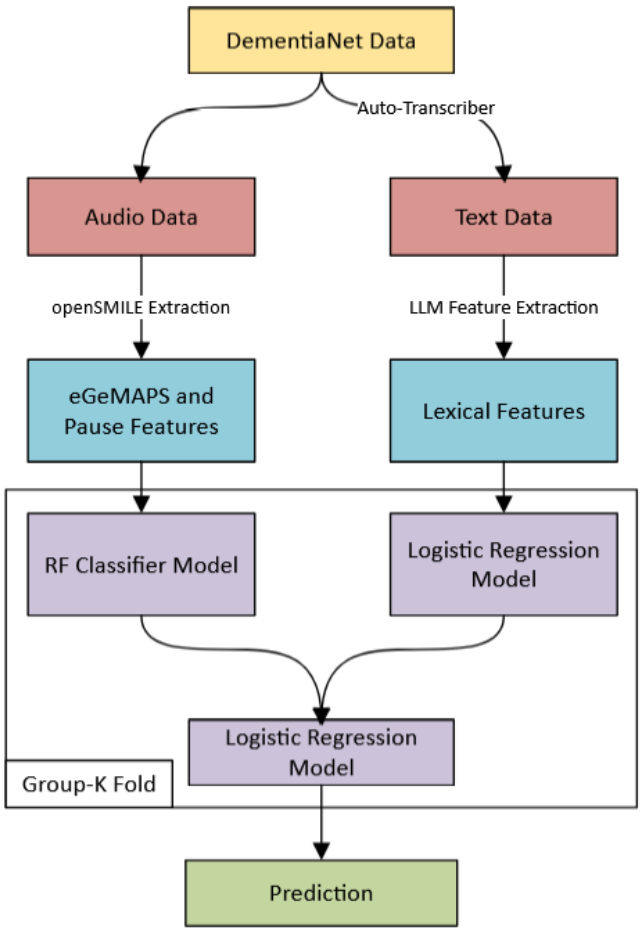

Fig. 1. Pipeline architecture detailing each stage of processing

To assess lexical entropy, a measure of dialogue disorganization, the GPT-2 small next-token probability functionality was leveraged to compute a probabilistic distribution by comparing adjacent tokens. Resulting lists of next-token probabilities were fed into Shannon's equation (2) to compute entropy in the dialogue; high entropy indicating increased unpredictability and disordered thought patterns.

$$\mathsf{H}(X) := -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{2}$$

3) Traditional NLP Features: Several traditional natural language processing features were used to supplement the computed metrics of entropy and drift. The three core supplementary features were moving-average type-token ratio (MATTR), mean sentence length, and pronoun ratio. Dementia patients typically exhibit a decreased vocabulary and shorter, more incomplete, sentences [13]. To capture this MATTR slides a 50-token window across the text and computes the moving average of unique words, with the choice of a moving window being to avoid longer dialogues skewing results lower. Quantifying sentence length was trivial and was computed using a word-count average of each sentence in the dialogue. Additionally, studies have shown that AD patients exhibited higher usage of pronouns with respect to total noun usage [9]. This was computed as a ratio of pronoun and total noun counts in the transcript text.

D. Acoustic Feature Extraction

Like physiological markers of stress and cognitive load, prosodic elements of acoustic data served as a powerful resource for this estimation, as they have been linked with brain volumes and cognitive load [11]. In this stage, the OpenSMILE framework was used to extract features in alignment with the eGeMAPSv02 features, such as tone, pitch, loudness, and jitter. Along with the eGeMAPSv02 features, several manually-derived speech-pause statistics, such as the number of pauses, the duration of pauses, and overall speech rate were computed.

E. Late-Stage Fusion

Once lexical and acoustic features have been extracted, the pipeline passes this data into a late-stage fusion model architecture comprising separate models trained on each modality. The outputs of these models are subsequently provided to a meta-classifier, which integrates the results from both the lexical and acoustic components. This entire training pipeline is cross-validated using Group-K Fold to ensure that individuals with multiple audio clips do not have their clips in both the train and test sets. The mapping for the chosen models and validation strategy is described below.

- **Text Classifier:** Logistic Regression
- **Audio Classifier:** Random Forest
- **Fusion Strategy:** Logistic meta-classifier combining text and audio modalities
- **Validation:** 10-fold Group-K Fold (speaker-wise)

III. Results

A. Performance Metrics

The final model pipeline emerged through multiple iterations involving different classifier architectures and feature-set combinations. The late-stage fusion model achieved an AUC of 0.605 ± 0.077, indicating modest but consistent predictive performance. Prior to fusion, the individual lexical and acoustic classifiers achieved AUCs of 0.602 ± 0.106 and 0.593 ± 0.073, respectively. While all models performed only slightly above chance (AUC = 0.5), they consistently outperformed the baseline across validation folds. A pooled ROC curve (Fig. 2), aggregating predictions across folds, further confirmed this, demonstrating stable performance across the dataset and varying thresholds.

B. Feature Importance

Lexical features demonstrated stronger predictive power than their acoustic counterparts, with the top five ranked features, MATTR, pronoun ratio, semantic drift, type-token ratio, and lexical entropy, all originating from the text modality. This trend emphasizes the relevance of psycholinguistic markers in capturing early signs of cognitive decline.

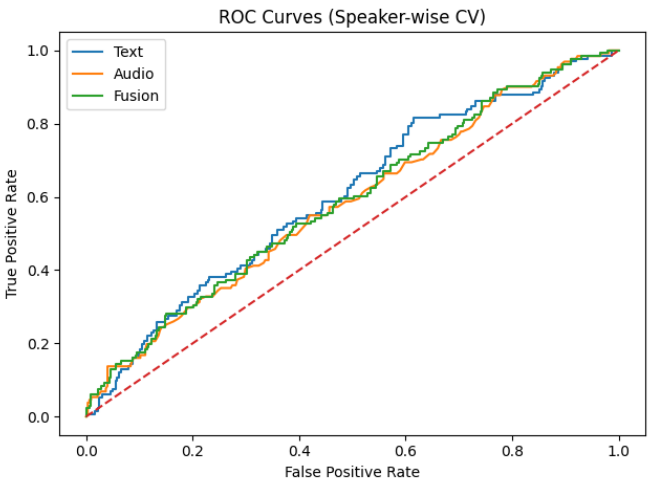

Fig. 2. ROC plot of individual text and audio classifiers and the finalized fusion classifier with respect to a random-choice baseline.

IV. Discussion

A. Interpretation and Implications

Results of pipeline testing indicate that the late-fusion model offered complementary insights by integrating both acoustic and lexical modalities, though performance did not substantially exceed that of the individual classifiers outside of making its predictions more precise. Each modality independently demonstrated modest predictive power, particularly the lexical branch, which utilized large language models to compute features like lexical entropy and semantic drift. This is noteworthy given limited data, lightweight models, and minimal preprocessing on acoustic features.

Despite these constraints, the success of lexical features alone as non-invasive indicators of cognitive decline emphasizes the potential of machine learning, especially multimodal architectures, to detect subtle markers of dementia and mild cognitive impairment (MCI) up to 15 years before clinical diagnosis. This proof-of-concept highlights not only the feasibility of such models in real-world settings but also the value of integrating novel language-based markers in early screening tools.

B. Real-World Applications

The rapid advancement of large-language models, in addition to the growing ubiquity of voice-enabled platforms, presents compelling opportunities for real-time, non-invasive cognitive health monitoring. As predictive performance continues to improve, these models could serve not only as risk-screening tools for clinicians but also as integral components of wearable and smartphone-based health technologies. By leveraging longitudinal voice data collected passively through everyday interactions, such systems could detect early signs of cognitive decline with precision.

Moreover, the implementation of federated learning would enable on-device model refinement while preserving user privacy, facilitating scalable, population-level adaptation. This holds significant promise for democratizing access to dementia risk assessment, improving both the generalizability and accuracy of machine learning models without centralized data aggregation.

V. Conclusion

This research demonstrates the feasibility of a cognitively grounded, multimodal machine learning approach for early dementia detection using speech data. By integrating lexical and acoustic features from public interview recordings, the proposed pipeline captures subtle markers associated with cognitive decline. Lexical features, particularly those derived from large language models such as semantic drift and lexical entropy, emerged as especially informative, consistently outperforming acoustic features in predictive strength. Although the overall classification performance remains modest, this study offers a compelling proof-of-concept. It highlights that even with lightweight models and limited preprocessing, it is possible to identify early indicators of dementia several years prior to clinical diagnosis. The implications for real-world deployment are substantial; future systems built on this foundation could enable passive, privacy-preserving risk screening through everyday devices such as smartphones and wearables, transforming dementia detection from a symptom-reactive clinical process into a proactive and accessible public health tool. As LLMs continue to evolve and as multimodal data becomes more readily available, there is strong potential for these models to become integral to next-generation cognitive health monitoring platforms. This work provides a meaningful step toward that vision by bridging cognitive science, speech analysis, and modern machine learning.

VI. Limitations

While the results are promising, several important limitations must be acknowledged.

A. Dataset Scope and Representativeness

The DementiaNet dataset, while novel and longitudinal to some degree, is composed exclusively of public interviews with celebrities. This introduces potential sampling biases such as scripted speech or atypical expression which may not generalize to broader populations with spontaneous speech or in clinical contexts. Additionally, the dataset remains relatively small, with limited demographic diversity, constraining the robustness and external validity of the model.

B. Modeling Constraints and Simplifications

Due to computational limitations, the models used in this study were relatively lightweight. The LLM-based features were derived from GPT-2 small and all-MiniLM-L6-v2 models that may not fully capture the rich semantic patterns available to more powerful model architectures. Furthermore, the acoustic analysis relied on standardized prosodic features with minimal tuning, limiting the opportunity to capture more nuanced speech interactions.

C. Modality Limitations

The current pipeline is limited to lexical and acoustic modalities. Although informative, this omits valuable contextual cues from other sources such as facial expression, gesture, and conversational dynamics.

D. Diagnostic Ambiguity

Because many dementia-labeled samples were taken 5–15 years prior to a formal diagnosis, the precise timing and severity of underlying cognitive changes remain uncertain. While being a practical necessity for this model, reliance on retrospective labeling introduces noise into the target variable, potentially diluting model performance and interpretability.

REFERENCES

[1] H. Amjad, K. D. Roth, E. Sheehan, and D. L. Wolff, "Underdiagnosis of dementia: An observational study of patterns in diagnosis and awareness in US older adults," *J. Gen. Intern. Med.*, vol. 33, no. 7, pp. 1131–1138, 2018.

[2] A. Bradford, L. Kunik, C. Schulz, S. Williams, and M. Singh, "Missed and delayed diagnosis of dementia in primary care: Prevalence and contributing factors," *Alzheimer Dis. Assoc. Disord.*, vol. 23, no. 4, pp. 306–314, 2009.

[3] C. B. Young, S. R. Becker, D. W. Harrison, and A. L. D'Andrea, "Speech patterns during memory recall relates to early tau burden across adulthood," *Nat. Aging*, vol. 4, no. 3, 2024.

[4] Frontiers in Aging Neuroscience, "Language markers of dementia," *Front. Aging Neurosci.*, Article: PMC11305841.

[5] V. Rentoumi, J. Charalabopoulou, D. Drakopoulou, and G. Tsoulfas, "Automatic detection of linguistic indicators of Alzheimer's disease," in *Proc. IEEE Int. Conf. Bioinform. Biomed.*, 2017.

[6] M. F. Mendez, M. M. Shapira, and J. H. Miller, "Psychotic-like speech in frontotemporal dementia," *J. Neuropsychiatry Clin. Neurosci.*, vol. 18, no. 3, pp. 295–303, 2006.

[7] Wikipedia, "Entropy (information theory)." [Online]. Available: https://en.wikipedia.org/wiki/Entropy_(information_theory)

[8] Y. Bestgen, "Estimating lexical diversity using the moving average type-token ratio," *J. Lang. Cogn.*, vol. 6, no. 3, 2024.

[9] D. Bittner, S. M. Torres, and E. R. Grossman, "Changes in pronoun use a decade before clinical diagnosis of Alzheimer's dementia," *Front. Aging Neurosci.*, vol. 13, 2022.

[10] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016.

[11] H. Ding, J. Zhao, K. Murabito, and R. S. Vasan, "Association between acoustic features and brain volumes: The Framingham Heart Study," *Front. Dement.*, vol. 1, 2023.

[12] M. L. Henry, M. A. Beeson, M. E. Stark, and B. J. Rapcsak, "Treatment for anomia in semantic dementia," *Neuropsychol. Rev.*, vol. 19, no. 3, 2009.

[13] S. Banovic, L. Junuzovic Zunic, and O. Sinanovic, "Communication difficulties as a result of dementia," *Front. Aging Neurosci.*, vol. 10, Art. no. 127, 2018. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6195406/

VII. APPENDIX

TABLE I
INDIVIDUAL PROJECT TIMELINE AND TASK BREAKDOWN

Week	Task	Task Description	Hours	Done (Y/N)
2–3	1	Create the template task list.	0.5	Y
2–3	2	Choose research question.	0.25	Y
2–3	3	Prepare and send data request emails.	0.75	Y
2–3	4	Draft ML pipeline workflows.	2	Y
2–3	5	Prepare project pitch.	5	Y
		PROJECT PITCH DUE		
4–6	6	Research feature extraction tools.	10	Y
4–6	7	Clean audio data.	5	Y
4–6	8	Build audio feature extraction pipeline.	5	Y
4–6	9	Transcribe audio to extract textual cues.	5	Y
4–6	10	Use pretrained models to assess lexical diversity	5	Y
4–6	11	Clean visual data.	5	Y
4–6	12	Build visual cue extraction pipeline.	5	Y
4–6	13	Develop baseline multimodal fusion model.	10	Y
4–6	14	Prepare for midpoint check-in.	5	Y
		OPTIONAL MIDPOINT CHECK-IN		
7–9	15	Evaluate model.	5	Y
7–9	16	Finetune prototyped model.	15	Y
7–9	17	Prepare final report.	12	Y
7–9	18	Final presentation prep.	5	Y
		FINAL REPORT DUE		
10	19	Prepare presentation.	5	N
10	20	Conduct final presentation.	3	N
		FINAL PRESENTATION DUE		
Total			108.5 hrs	