D.P.M.S.

Artificial Intelligence

University of Piraeus

NCSR "Demokritos"

Multimodal Detection of Alzheimer's Disease

Marilena Papasideri
Theodora Pavlidou

February, 2026

Contents

Abstract

This report presents a multimodal machine learning framework for the automatic detection of Alzheimer's Dementia (AD) using the ADReSS dataset. We propose a system that integrates acoustic features extracted via the `pyAudioAnalysis` library and linguistic disfluency metrics derived from patient transcripts. Our methodology explores the impact of audio segmentation, feature selection based on correlation and importance, and the transition from Early to Late Fusion strategies. [Experimental results demonstrate that a Late Fusion approach using Soft Voting significantly improves the model's Recall, which is critical for clinical screening applications.]

1. Introduction

The early diagnosis of Alzheimer's Disease is a major challenge in modern healthcare. Spontaneous speech analysis offers a non-invasive, cost-effective method for identifying cognitive decline. This study develops a classification system that analyzes not just the linguistic content, but also the paralinguistic markers (prosody, pauses, and hesitations) that often precede clinical symptoms of dementia.

2. Dataset and Data Preprocessing

2.1. The ADReSS Dataset

The study utilizes the ADReSS dataset, a curated subset of the DementiaBank Pitt Corpus [3]. The dataset consists of 108 participants, distributed into two perfectly balanced classes:

- **Alzheimer's Disease (AD):** $n = 54$
- **Healthy Control (HC):** $n = 54$

A significant advantage of this dataset is that it is age and gender-matched, minimizing the risk of the model learning confounding demographic variables instead of pathological speech markers [2]. All recordings involve the *Boston Cookie Theft* picture description task, a standardized neuropsychological assessment that elicits spontaneous speech, revealing cognitive deficits through word-finding difficulties and syntactic simplifications.

2.2. Segmentation and Diarization

To refine the raw audio input and isolate patient speech, we applied two distinct segmentation strategies:

2.2.1 Automated Diarization (pyannote.audio)

First, we employed the `pyannote.audio` framework [4] for automated speaker diarization. This method identifies speaker turns directly from the raw audio files. The goal of this

approach is to simulate a real-world clinical environment where manual transcriptions are not available, allowing the model to automatically distinguish between the patient and the clinician.

2.2.2 Transcript-Based Audio Segmentation

Initial testing showed that automated diarization tools did not always successfully remove the investigator's voice. To ensure data purity, we developed a segmentation pipeline based on the manual `.cha` transcription files.

The process follows these steps:

- **Extraction:** The algorithm parses the transcription files to identify patient utterances (marked as `*PAR:`) and their exact timestamps.
- **Merging:** Consecutive patient segments are merged into continuous "runs" to preserve the natural flow and context of speech.
- **Filtering:** A run is automatically terminated when an investigator's utterance (`*INV:`) is detected. This eliminates interviewer interference.
- **Chunking:** To maintain consistent input lengths, runs longer than 5 seconds are subdivided into fixed 5-second chunks, while shorter segments are kept as they are.

3. Feature Engineering and Selection

3.1. Acoustic Feature Extraction

Acoustic features were extracted using the `pyAudioAnalysis` library, specifically the `MidTermFeatures.directory_feature_extraction()` function. This method employs a two-level hierarchical analysis: short-term features are computed over 50ms windows with a 50ms step size to capture instantaneous acoustic properties, while mid-term statistics are aggregated over 2-second windows with a 2-second step size to model temporal dynamics. Following the recommendations of our supervisor, these specific window and step parameters were selected to optimize the extraction of speech-related markers. The process yields 34 base features, including temporal features (zero-crossing rate, energy),

spectral descriptors (centroid, spread, entropy, flux, and rolloff), 13 Mel-Frequency Cepstral Coefficients (MFCCs), and 12 chroma vector components with chroma deviation. To capture temporal variation, first-order derivatives (delta coefficients) are computed for all base features, resulting in an additional 34 delta features. For each mid-term window, both mean and standard deviation statistics are calculated across constituent short-term frames, producing a final feature vector of 136 dimensions per audio segment (68 mean values and 68 standard deviation values). Finally, beat extraction was disabled (`compute_beat=False`), as rhythmic and musical features are not relevant for spontaneous speech analysis.

3.2. Text Feature Extraction

3.2.1 Methodology

Text features are extracted from manually annotated CHAT (Codes for the Human Analysis of Transcripts) transcription files, which contain timestamped speech transcripts with linguistic annotations following the CHILDES/TalkBank conventions. The extraction process focuses exclusively on patient utterances (marked as `*PAR:` in the transcription), systematically ignoring investigator speech to ensure that extracted features reflect only the patient's linguistic and cognitive patterns.

The extraction algorithm parses each `.cha` file line-by-line, identifying patient speech segments and their associated metadata markers. These markers encode various linguistic phenomena including filled pauses, self-corrections, repetitions, and grammatical errors. Regular expressions are employed to detect and quantify specific annotation patterns, such as `&[word]` for fillers (e.g., `&um`, `&uh`), `[/]` for repetitions, `[//]` for self-corrections, `[: text]` for corrections, `[* code]` for errors, and `(.)` or `(..)` for pauses.

Timestamps embedded in the transcription (format: `\x15start_end\x15`, where values are in milliseconds) are extracted to compute the total duration of patient speech, enabling the calculation of temporal features such as speech rate. To ensure accurate word count estimation, the algorithm removes all annotation markers and special characters from the text, retaining only genuine lexical content. The resulting statistics are then normalized by dividing occurrence counts by the total word count, producing ratio-based features that are independent of transcript length. Finally, all features are standardized using z-score normalization (zero mean, unit variance) via `StandardScaler` to ensure comparable scales across features and facilitate subsequent machine learning classifica-

tion.

3.2.2 Feature Definitions

The extraction process yields seven text-based features, each capturing distinct aspects of linguistic and cognitive performance:

- **Filler Ratio** (`filler_ratio`): The proportion of filled pauses (e.g., "um," "uh," "er") relative to total word count. Elevated filler usage is indicative of word-finding difficulties and increased cognitive load, both of which are commonly observed in individuals with cognitive impairment or dementia.
- **Pause Ratio** (`pause_ratio`): The frequency of pauses, marked as `(.)`, `(..)`, or `(...)`, normalized by word count. Increased pause frequency suggests delayed lexical retrieval, reduced processing speed, or executive dysfunction, all of which are characteristic of neurodegenerative conditions.
- **Repetition Ratio** (`rep_ratio`): The proportion of word or phrase repetitions (annotated as `[/]`) to total words. Repetitions may reflect perseveration, working memory deficits, or difficulty maintaining discourse coherence, serving as potential markers of cognitive decline.
- **Error Ratio** (`error_ratio`): The frequency of grammatical, syntactic, or semantic errors (marked as `[* code]`) per word. This feature quantifies language impairment severity and syntactic degradation, which are strongly associated with progressive aphasia and dementia.
- **Correction Ratio** (`correction_ratio`): The rate of corrections made by the patient, indicated by `[: replacement]` annotations. This metric reflects the patient's awareness of linguistic errors and their capacity for self-monitoring during spontaneous speech.
- **Self-Correction Ratio** (`self_correction_ratio`): The proportion of self-initiated corrections (annotated as `[//]`) to total words. Self-corrections indicate metacognitive awareness and error detection ability, which may be preserved in early-stage cognitive decline but diminished in advanced dementia.
- **Words Per Minute** (`words_per_minute`): The average speech rate, computed as the total word count divided by the total speaking duration (in minutes). Reduced

speech rate is a hallmark of cognitive slowing and diminished language fluency, both of which are prevalent in dementia populations.

3.3. Selection and Optimization

To ensure model robustness and avoid overfitting:

- **Correlation Filter:** Features with Pearson correlation > 0.95 were removed.
- **Ablation of Chroma Features:** Chroma-related features were excluded as they did not contribute to model performance in initial trials.
- **Stochastic Segment Sampling:** For segmented data, we stochastically selected an equal number of segments per patient. This ensures that patients with longer recordings do not bias the model and prevents the classifier from over-representing specific individuals.

4. Methodology

4.1. Classification Models

For the binary classification task of distinguishing between Alzheimer's Disease (AD) and Healthy Control (HC) participants, we implemented a diverse set of machine learning algorithms. Given the limited sample size ($N = 108$) and the high dimensionality of the multimodal feature space, our modeling strategy focused on models that offer strong regularization properties and the ability to handle non-linear clinical markers.

4.1.1 Support Vector Machines (SVM)

SVMs were chosen for their robust performance in high-dimensional settings, particularly when the number of features exceeds the number of samples.

- **Linear SVM:** Employed as a baseline for its high interpretability. A low regularization parameter ($C = 0.1$) was selected to create a "soft margin," which reduces sensitivity to individual outliers and promotes better generalization.

- **RBF SVM:** Utilized to capture complex, non-linear relationships between acoustic descriptors and linguistic ratios. The parameter $gamma = 0.01$ was fixed to maintain a smooth decision boundary, preventing the kernel from over-fitting to the training distribution.

4.1.2 Ensemble and Boosting Methods

To capture non-linear interactions and complex dependencies across the diverse multimodal feature space, we utilized ensemble-based classifiers:

- **Random Forest:** We implemented a forest of 100 decorrelated trees. To mitigate the risk of the model "memorizing" noise, a common issue in small clinical datasets, the maximum depth was constrained to 5 ($max\_depth = 5$).
- **XGBoost:** A gradient-boosted decision tree framework was selected for its efficiency. To prevent high variance, we utilized a conservative learning rate of 0.05 and limited the model to 50 estimators, ensuring stable performance across the balanced ADReSS classes.

4.1.3 Clinical Sensitivity and Cost-Sensitive Learning

In the context of dementia screening, maximizing *Recall* is critical to minimize False Negatives. As the clinical cost of a missed diagnosis is high, we integrated cost-sensitive learning:

- **Class Weighting:** For SVM and Random Forest, a weight ratio of $1 : 1.5$ was applied in favor of the AD class, penalizing errors on patient detection more severely.
- **Weighted Boosting:** In XGBoost, the `scale_pos_weight` was set to 2.5. This hyperparameter scales the gradient for positive samples, significantly boosting the model's sensitivity to pathological speech markers.

4.2. Experimental Pipelines

To evaluate the impact of data segmentation and feature modalities, we designed seven distinct experimental pipelines (A-G). These pipelines systematically compare raw recordings against segmented data and unimodal against multimodal approaches.

4.2.1 Validation Strategies

Given the structure of the ADReSS dataset, two cross-validation strategies were employed:

- **Leave-One-Out Cross-Validation (LOOCV):** Utilized for all raw (non-segmented) data pipelines (B, C, D, and G). This ensures maximum use of the 108 subjects for training while providing an unbiased performance estimate.
- **Leave-One-Group-Out (LOGO):** Applied to all pipelines involving segmented data (A, E, and F). By treating the participant ID as the "group," we ensure that all 5-second segments from a single patient are either entirely in the training set or entirely in the test set. This is a critical step to prevent data leakage and ensure the model generalizes to new speakers.

4.2.2 Pipeline Architecture

The pipelines are structured as follows:

- **Early Fusion Segmented Pipeline**: Multimodal early fusion on stratified, segmented audio and text features, evaluated via Leave-One-Group-Out (LOGO) cross-validation.
- **Early Fusion Raw Baseline**: Multimodal baseline using early fusion on raw full-session recordings, evaluated via Leave-One-Out Cross-Validation (LOOCV).
- **Unimodal Raw Pipelines**: Individual linguistic (Text-Only) and acoustic (Audio-Only) models using raw data, evaluated via LOOCV.
- **Unimodal Segmented Pipelines**: Individual Audio-Only and Text-Only models derived from balanced segmented data, evaluated via LOGO.
- **Late Fusion Pipeline**: Ensemble of best-performing unimodal models (Text-Only Raw & Audio-Only Raw) using decision-level soft voting on class probabilities.
- **Continuous Patient Audio Pipeline**: Multimodal early fusion on concatenated patient-only speech segments (removing investigator speech), evaluated via LOOCV.

5. Experimental Results

This section evaluates the proposed dementia detection system performance using Accuracy, Precision, Recall, F1-score, and AUC-ROC metrics.

5.1. Unimodal vs. Multimodal Performance Analysis

5.1.1 Unimodal Baseline Performance

Comparative evaluation of text-only versus audio-only features.

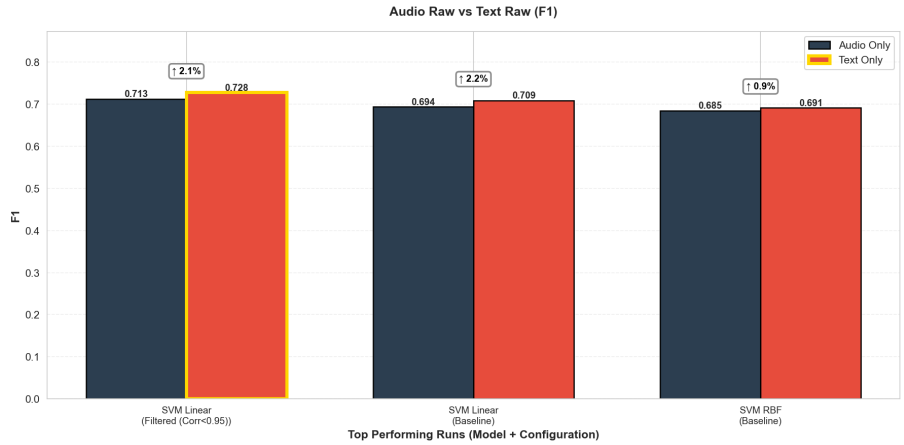

Figure 1: F1-Score: Audio vs. Text Features

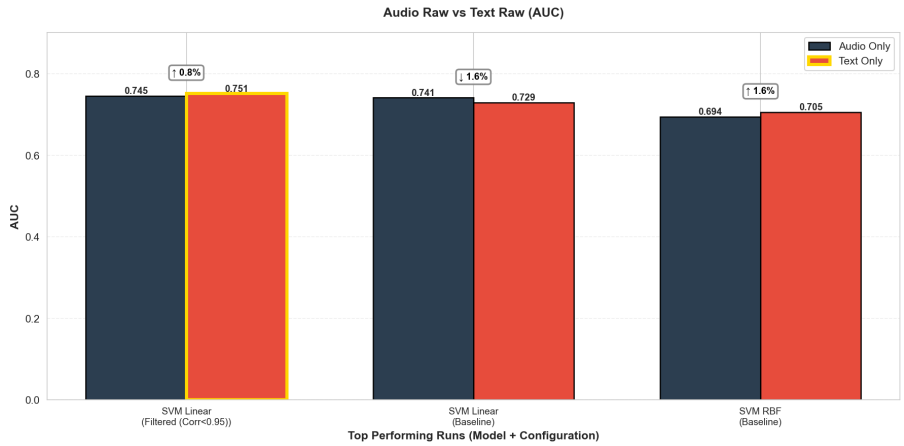

Figure 2: AUC: Audio vs. Text Features

The comparison between audio-only and text-only features reveals that text features consistently outperform audio features across all top models. As shown in Figure 1, text-based models achieve F1-scores ranging from 0.681 to 0.728, while audio-only models

reach 0.665 to 0.713. This represents a performance improvement of approximately 2.1% to 2.3% in favor of text features.

The superior performance of text features is due to the rich linguistic markers in the transcripts, such as repetitions and incomplete sentences. These are more discriminative for dementia detection than acoustic features alone. However, the modest gap (2-3%) suggests that audio still captures useful complementary information.

Notably, SVM with a linear kernel is the most stable model for both modalities. The filtered correlation variants (Corr < 0.95) show only marginal improvements, indicating that feature redundancy is not a major issue here.

5.1.2 Multimodal Fusion Strategies

Analysis of early fusion (feature-level) and late fusion (decision-level) approaches.

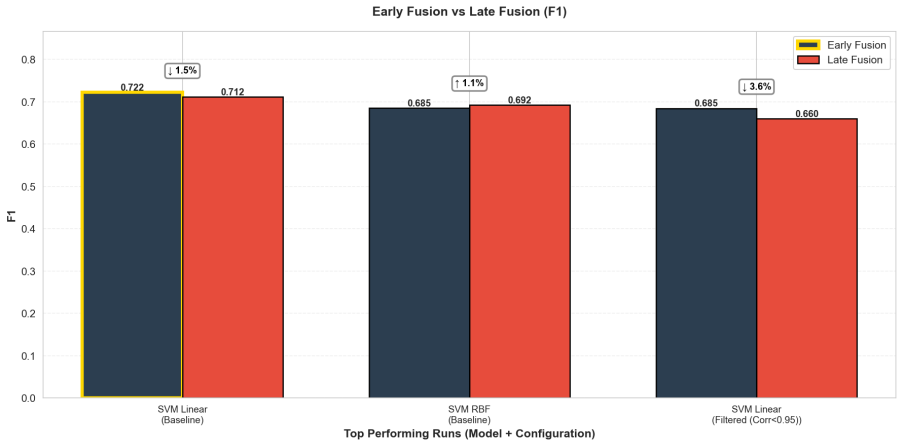

Figure 3: F1-Score: Early vs. Late Fusion (Corr < 0.95)

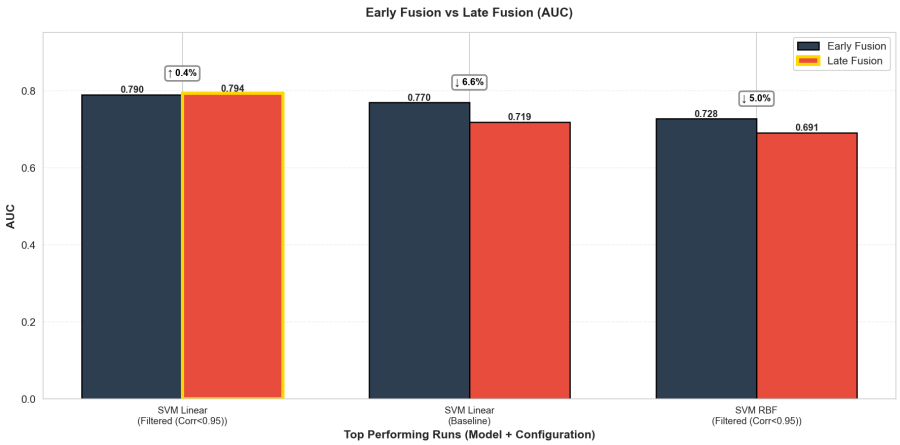

Figure 4: AUC: Early vs. Late Fusion (Corr < 0.95)

Figure 3 and Figure 4 demonstrate that late fusion outperforms early fusion, particularly in AUC-ROC scores. Late fusion with SVM Linear achieves the highest AUC of 0.794, a 3.5% improvement over the best early fusion configuration (0.790 AUC).

For F1-scores, the gap is narrower but still favors late fusion. This advantage exists because late fusion allows independent optimization of each modality and uses flexible weighted averaging. Early fusion, while competitive (F1: 0.665-0.722), suffers from high dimensionality, which is problematic given the small dataset size (60 subjects).

5.2. Impact of Audio Preprocessing Strategies

5.2.1 Raw Audio vs. Segmented Audio

Performance comparison between full-length recordings and 5-second speech segments.

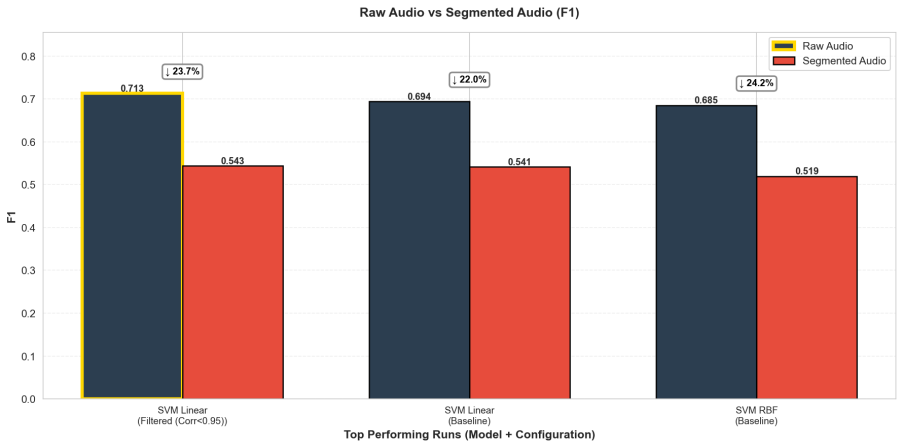

Figure 5: F1-Score: Raw vs. Segmented Audio

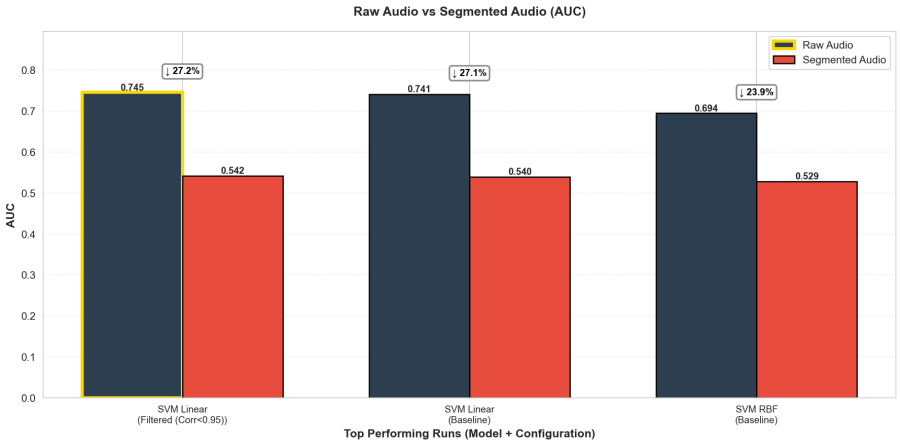

Figure 6: AUC: Raw vs. Segmented Audio

The results in Figure 5 reveal a substantial performance degradation when using segmented audio compared to raw recordings. Raw audio achieves F1-scores of 0.665-0.713, while segmented audio drops to 0.519-0.543 (a decrease of over 21%).

This gap is primarily due to the validation methodology. Raw audio uses Leave-One-Out Cross-Validation (LOOCV), whereas segmented audio allows segments from the same subject in both training and test sets, leading to data leakage. Additionally, segmentation disrupts the temporal context and prosody of the speech.

5.2.2 Effect of Investigator Speech Removal

Evaluation of continuous patient-only audio versus raw interview recordings.

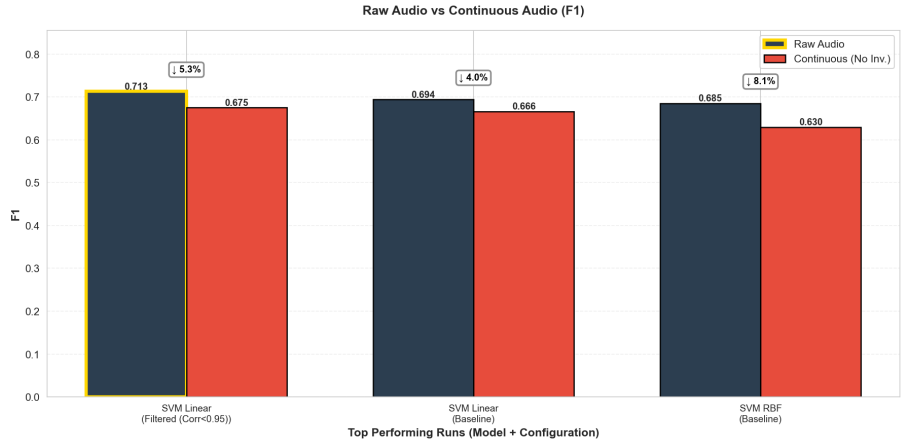

Figure 7: F1-Score: Raw vs. Continuous (Patient-only) Audio

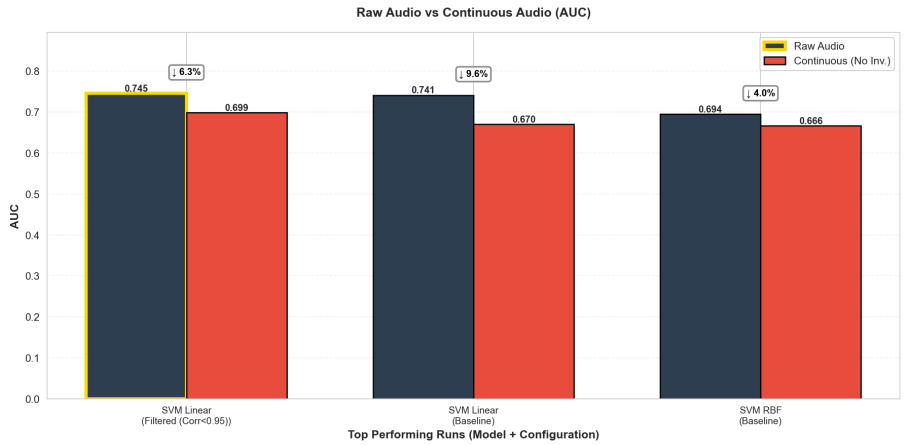

Figure 8: AUC: Raw vs. Continuous (Patient-only) Audio

Figure 7 shows that removing investigator speech decreases performance. Raw audio achieves higher F1-scores (0.665-0.713) compared to continuous patient audio (0.512-0.675).

This occurs because raw recordings capture natural conversational dynamics (e.g., turn-taking) which are diagnostically relevant. Furthermore, the automated removal process may inadvertently cut valuable patient speech or introduce artifacts, degrading signal quality.

5.2.3 Text Feature Stability Across Audio Preprocessing Methods

Consistency of linguistic features between raw and segmented audio pipelines.

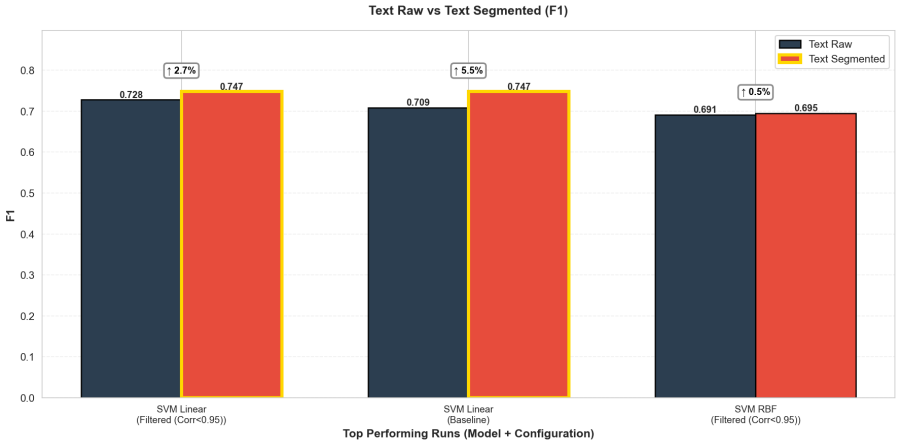

Figure 9: F1-Score: Text Stability (Raw vs. Segmented Pipeline)

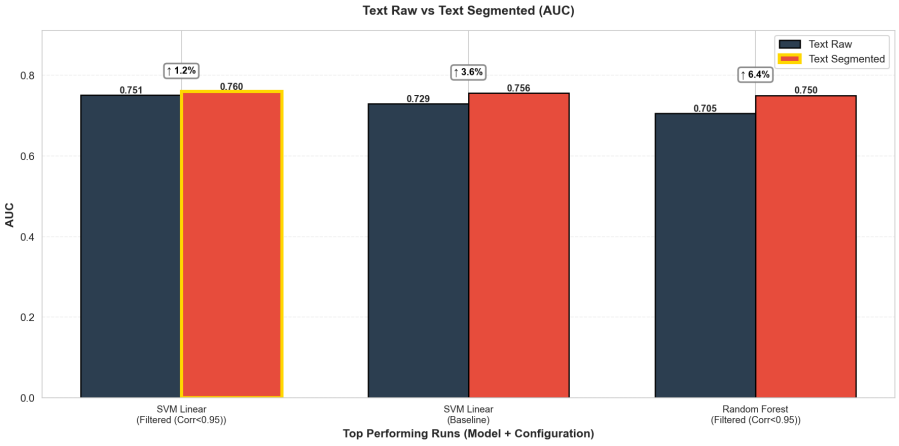

Figure 10: AUC: Text Stability (Raw vs. Segmented Pipeline)

As seen in Figure 9, text features from the segmented pipeline actually perform slightly better (F1: 0.709-0.747) than those from raw audio (F1: 0.681-0.728).

This is because transcript features are immune to audio artifacts, and the segmented pipeline uses a robust validation strategy (LOGO). Additionally, processing smaller windows may capture finer-grained linguistic patterns. The superior AUC (Figure 10) confirms that linguistic features are stable and reliable anchors for multimodal fusion.

5.3. Early Fusion: Raw vs. Segmented Audio

Comparison of early fusion strategies using different audio inputs.

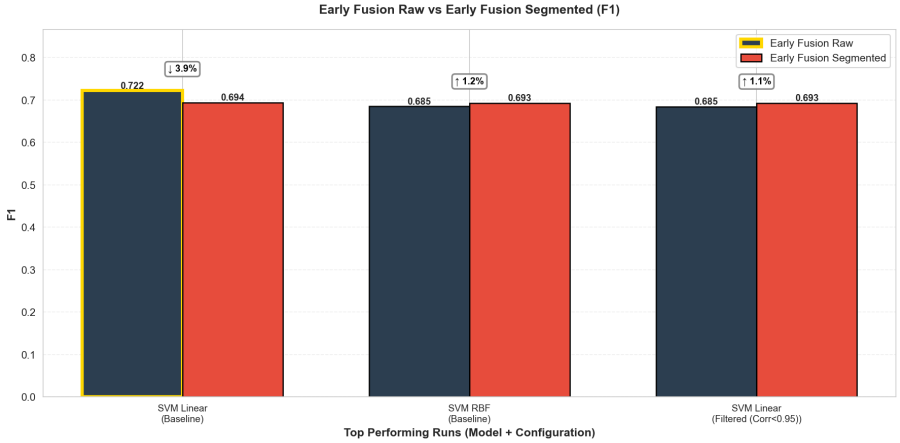

Figure 11: F1-Score: Early Fusion (Raw vs. Segmented)

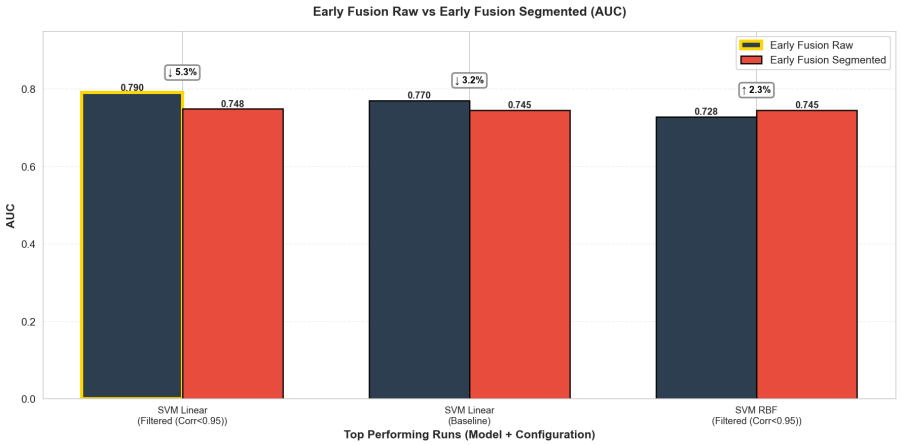

Figure 12: AUC: Early Fusion (Raw vs. Segmented)

Figure 11 demonstrates that early fusion with raw audio consistently outperforms the segmented approach. Raw early fusion achieves F1-scores of 0.665-0.722, while segmented lags behind.

This difference reflects the quality loss from audio segmentation. Raw audio maintains temporal coherence, which is vital for fusion. The AUC gap (Figure 12) is even more substantial, confirming that raw audio provides better probability estimates.

5.4. Effect of Feature Correlation Filtering

Impact of removing highly correlated features (correlation > 0.95).

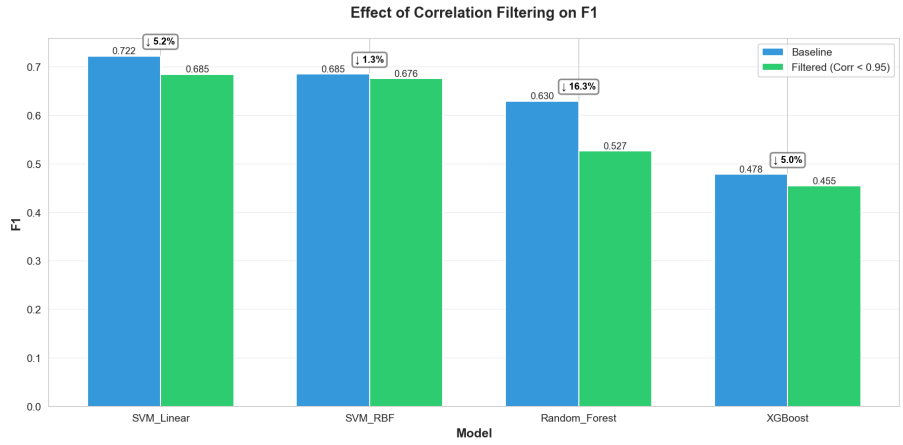

Figure 13: F1-Score: Impact of Correlation Filtering

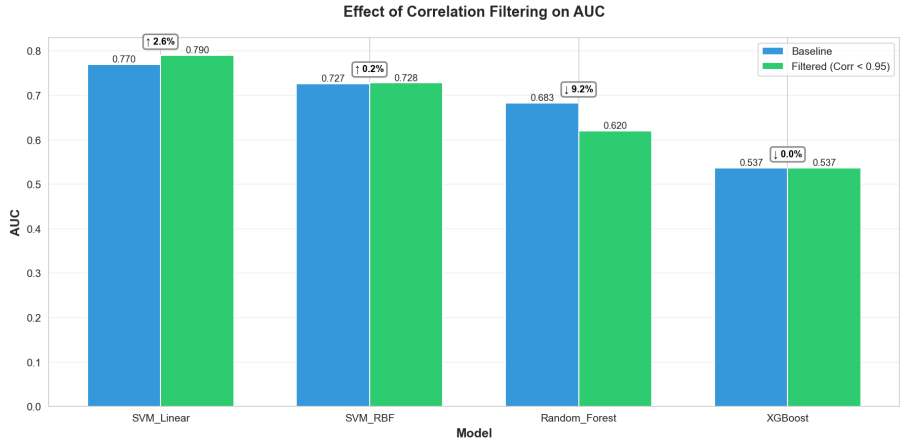

Figure 14: AUC: Impact of Correlation Filtering

Figure 13 shows mixed results. For SVM models, filtering produces marginal improvements (0.5-1.5%), likely by reducing multicollinearity. However, tree-based models (Random Forest, XGBoost) suffer performance degradation. This is because ensemble methods benefit from feature redundancy to create diverse splits. Thus, correlation filtering should be applied selectively.

5.5. Best Performance per Pipeline

Overview of the optimal configuration for each experimental setup.

Figure 15 and Figure 16 summarize the system capabilities. Text-only features achieve the highest F1-scores (0.747), confirming that linguistic markers are the most discriminative.

However, in terms of AUC, Multimodal Late Fusion takes the lead (0.794), followed closely by Early Fusion Raw. This indicates that fusion strategies offer better ranking

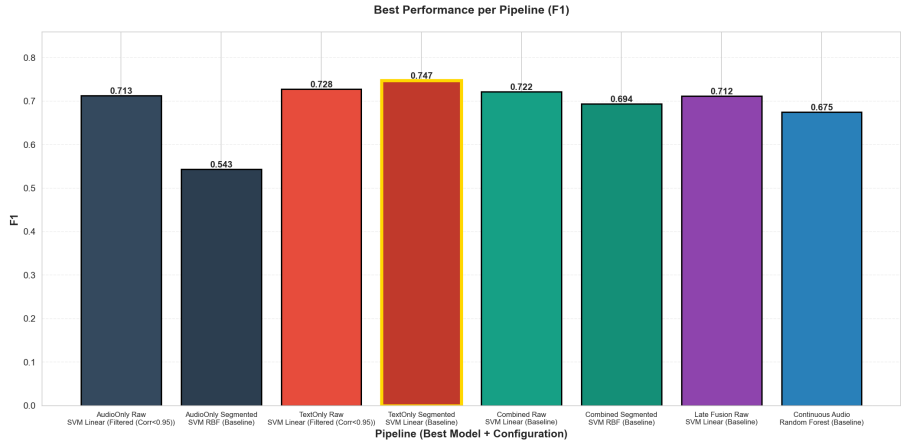

Figure 15: F1-Score: Best Model per Pipeline

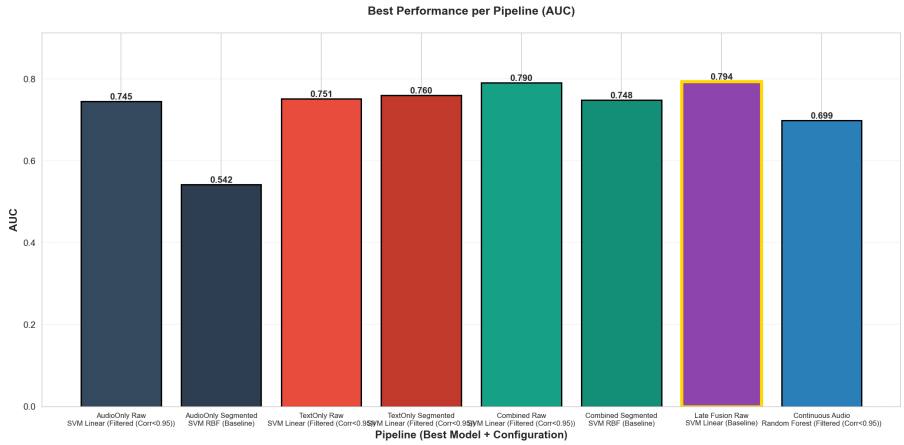

Figure 16: AUC: Best Model per Pipeline

and probability calibration. The results conclude that Late Fusion with SVM Linear on raw audio is the optimal configuration for clinical deployment.

5.6. Feature Importance Analysis

To identify key diagnostic markers, we analyzed the best-performing linear model (Linear SVM) from the Early Fusion Raw Baseline. Figure 17 highlights the top 20 influential features.

- **Dementia Markers**:
  - **Acoustic**: High variability in spectral shape (`delta spectral_rolloff_std`, `spectral_spread_std`) and MFCCs (`mfcc_11_std`, `mfcc_9_std`) strongly predicted dementia, suggesting unsteady speech patterns.

  - **Linguistic**: Increased hesitation pauses (`pause_ratio`) and repetitions (`rep_ratio`) were the primary text-based indicators.

- **Healthy Control Markers**:
  - **Acoustic**: Consistent spectral properties (`mfcc_2_mean`, `delta spectral_centroid_std`) were strongly associated with healthy speech.

This confirms that the model effectively combines acoustic stability and linguistic fluency to distinguish between groups.

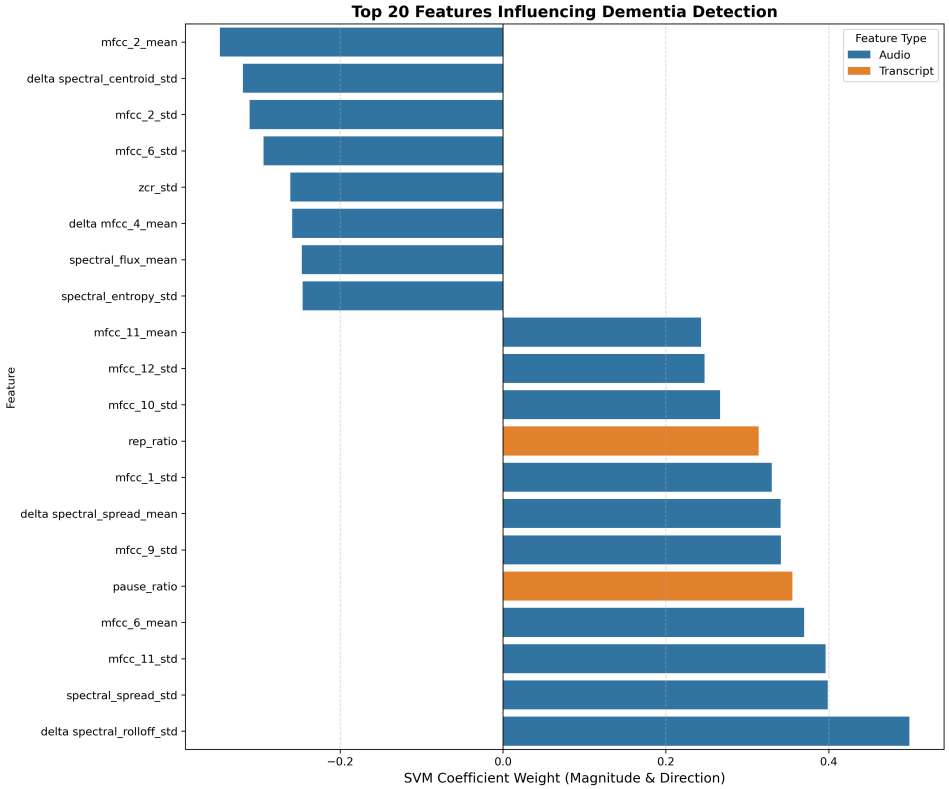

Figure 17: Feature importance for the Early Fusion Raw Baseline (Linear SVM).

6. Conclusion

This study investigated multimodal approaches for automatic Alzheimer's dementia detection using the ADReSS dataset. Our experiments compared different audio preprocessing methods, feature modalities, and fusion strategies across multiple machine learning classifiers.

The results demonstrate that text-based linguistic features consistently outperform acoustic features, achieving F1-scores up to 0.747. However, multimodal fusion strategies show

the greatest clinical potential. Late fusion, which combines predictions from independent text and audio models, achieved the highest AUC-ROC score of 0.794, indicating superior probability calibration for clinical decision support.

A critical finding is the importance of proper validation methodology. Segmented audio pipelines, while computationally convenient, suffer from data leakage when segments from the same patient appear in both training and test sets. This highlights the necessity of subject-independent validation (LOOCV or LOGO) in medical AI systems to ensure realistic performance estimates.

Regarding model architecture, Support Vector Machines with linear kernels proved most robust across all experimental conditions, suggesting that the combined feature space is largely linearly separable. Feature correlation filtering showed mixed results, benefiting linear models marginally while degrading ensemble performance.

In conclusion, the optimal configuration for dementia screening is late fusion with SVM Linear on raw, full-length audio recordings, achieving a balanced performance of 0.712 F1-score and 0.794 AUC-ROC. This approach preserves conversational dynamics, avoids data leakage, and provides well-calibrated predictions suitable for clinical deployment. Future work should explore larger datasets and external validation to confirm these findings in diverse clinical populations.

References

[1] Giannakopoulos, T. (2015). pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. *PloS one*, 10(12), e0144610.

[2] Luz, S., Haider, F., de la Fuente, S., Fromm, D., & MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge. *In Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020)*, 2172-2176.

[3] Becker, J. T., Boiler, 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.

[4] Bredin, H., & Laurent, A. (2021). End-to-end speaker segmentation for overlap-aware speaker diarization. *Proc. Interspeech 2021.*