1. Executive Summary

The DEMENTIA framework represents a significant advancement in the assisted assessment of Alzheimer's Disease (AD) from speech. This multi-task learning (MTL) model is specifically engineered for the simultaneous detection of AD and the prediction of cognitive state, as measured by the Mini-Mental State Examination (MMSE) score. A core innovation of this framework lies in its integration of hybrid attention mechanisms with multimodal representations, encompassing audio, text, and expert-derived knowledge. This design allows the model to capture intricate interactions both within and between different data modalities.

The DEMENTIA model has demonstrated robust performance in AD assessment. On the ADRESS dataset, it achieved an accuracy of 89.58% and a recall of 91.67% for the classification task, alongside a Root Mean Square Error (RMSE) of 4.31 for the regression task. Furthermore, its generalizability has been validated on an external dataset, demonstrating consistent performance. 1 Beyond quantitative metrics, the framework emphasizes explainability. Comprehensive analyses reveal specific speech patterns characteristic of AD patients, such as slower speech rates, reduced syntactic complexity, and a greater propensity to use pause fillers and pronouns. Such interpretability is crucial for fostering clinical trust and facilitating the adoption of AI-assisted diagnostic tools. 1 By addressing the limitations of previous studies, particularly those that overlooked inter-modal interactions and lacked model explainability, DEMENTIA paves the way for more reliable and interpretable AI solutions in clinical practice for AD assessment. 1

DEMENTIA\_A...owledge\_fo q

PDF

**2. Introduction to Alzheimer's Disease and Speech Biomarkers**

**2.1. Alzheimer's Disease: Pathophysiology and Cognitive Impact**

Alzheimer's Disease (AD) stands as the most prevalent neurodegenerative disorder and the primary cause of dementia, imposing a substantial burden on affected individuals, their caregivers, and society at large. The disease is characterized by a profound and progressive deterioration in cognitive functions. This includes a notable slowing of processing speeds and significant declines in attention, working memory, executive function, and language abilities. The damage inflicted by AD is both progressive and irreversible, with no effective treatments currently available that can halt its progression. Consequently, the timely detection and early intervention in the initial stages of AD are of paramount importance. Such early action is crucial for controlling symptoms, potentially slowing the rate of disease progression, and ultimately enhancing the quality of life for patients.

**2.2. The Rationale for Speech as a Non-Invasive Biomarker**

Language is a fundamental human ability, forming the very foundation of cognition and daily communication with the surrounding world. While seemingly simple, language production is a high-level cognitive function in humans, relying on complex multidimensional skills that involve various and interdependent cognitive domains. These include memory, planning, and the integration of perceptual and motor functions. A critical aspect of AD progression is that early cognitive decline can manifest through subtle changes in speech patterns. These changes often go unnoticed by the human ear but offer a unique and early window for observing an individual's cognitive state.

The practical advantages of utilizing speech data for diagnostic purposes are compelling. Speech data is easily accessible and recordable, allowing for collection anytime in natural settings. It possesses favorable ecological validity, meaning it reflects real-world communication, and is both non-invasive and cost-effective. These attributes make artificial intelligence (AI)-based speech signal processing and natural language processing (NLP) techniques highly promising research directions for AD detection. The speech representations derived from these analyses can serve as effective biomarkers for assessing cognitive decline.

The utilization of AI and NLP in this context is particularly profound. While human observation might perceive speech as a "window into the mind," AI-driven analysis transforms it into a "microscope" for revealing an individual's cognitive state. This transformation implies that AI/NLP techniques enable the detection and quantification of subtle, micro-level changes in speech that are often imperceptible to the human ear or traditional clinical assessments. These minute linguistic and acoustic deviations are frequently precursors to overt cognitive decline. The DEMENTIA model's reported high sensitivity, with an 89.58% accuracy and 91.67% recall for the classification task, directly supports this capability to discern such subtle indicators. This positions AI-driven speech analysis not merely as an alternative diagnostic method, but as a potentially superior tool for early AD detection. Its enhanced sensitivity to subtle changes allows for objective, data-driven biomarker identification, shifting the paradigm from subjective clinical observation to quantifiable, actionable assessments. This capability is vital given the progressive and irreversible nature of AD and the importance of early intervention.

**2.3. Limitations of Prior Research and DEMENTIA's Contribution**

Despite the promising results achieved by current studies utilizing speech analyses for AD detection, several limitations persist in existing research. Unimodal studies, which rely solely on acoustic or linguistic features, often employ structured handcrafted features that demand significant domain expertise and time-consuming feature engineering for optimal results. Conversely, neural network-based unimodal approaches frequently lack the integration of prior knowledge from human experts.

Furthermore, most multimodal fusion methods, which combine audio and text information, typically rely on simpler early, late, or mixed fusion techniques. These approaches often overlook the intrinsic correlations between different modalities and the complex interactions between model representations, leading to suboptimal performance and increased training costs. A significant challenge identified in existing complex AI models for AD detection is the lack of explainability. Addressing the explainability of AI decisions and fostering clinicians' trust in these decisions is an essential prerequisite for the successful application of AI in clinical practice.

To overcome these constraints, the DEMENTIA model was proposed. It is a joint hybrid attention and multimodal representation model with multi-task learning for AD assessment. By fusing temporal and semantic relationships in audio signals, contextual representations in text signals, and prior knowledge from human experts, the model adequately captures inter- and intra-modal interrelationships across modalities, enhancing both performance and clinical explainability.

**3. The DEMENTIA Framework: A Deep Dive into its AI Architecture**

**3.1. Overall Multimodal and Multi-Task Learning Design**

The DEMENTIA model is conceptualized as a multimodal fusion framework, meticulously designed to process three distinct input modalities: raw audio signals, their corresponding textual transcripts, and expert-derived knowledge features. This architecture is engineered to deliver two primary outputs: a binary classification for Alzheimer's Disease (AD) detection (distinguishing AD patients from healthy controls) and a regression task for predicting Mini-Mental State Examination (MMSE) scores, which offer a continuous measure of cognitive state.

The framework integrates four main computational components, critically leveraging a hybrid attention mechanism. This mechanism is central to thoroughly capturing both intra-modal (relationships within a single modality) and inter-modal (relationships between different modalities) interactions, a key differentiator from simpler fusion approaches. The multi-task learning (MTL) paradigm is foundational to the model, allowing for shared learning across these related tasks. This approach is designed to enhance overall efficiency, improve generalization ability, and increase robustness by enabling the model to leverage common underlying patterns and representations across diagnostic and cognitive assessment tasks.

**3.2. Audio Encoder: Processing Acoustic Signals**

The audio encoder component of the DEMENTIA model is responsible for extracting and processing relevant information from raw speech signals. The initial step in this process involves the extraction of **Mel-frequency Cepstral Coefficients (MFCCs)** from the original audio signal. MFCCs are a widely adopted feature in speech processing, renowned for their ability to represent the short-term power spectrum of sound in a manner that closely approximates human auditory perception. This process focuses on the frequencies most discernible to the human ear, applying a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz to align with the non-linear human auditory system. The calculation of MFCCs involves several stages: pre-emphasis to amplify high-frequency components, framing and windowing (typically using a Hamming window) to segment the signal into overlapping frames and reduce discontinuities, Fast Fourier Transform (FFT) to convert the signal into the frequency domain, application of a Mel-filterbank, logarithmic compression of the power spectrum to match human perception of loudness, and finally, Discrete Cosine Transform (DCT) to obtain the cepstral coefficients.

Following MFCC extraction, the processed audio features are fed into a **Bi-directional Long Short-Term Memory (BiLSTM) network**. BiLSTMs are a sophisticated type of recurrent neural network (RNN) designed to handle sequence-based data. Unlike standard LSTMs that process data in a single direction, a BiLSTM consists of two LSTMs: one processing the input sequence in a forward direction (from beginning to end) and the other in a backward direction (from end to beginning). This dual processing allows the network to capture contextual dependencies from both past and future information within the speech sequence, which is crucial for understanding the complex temporal dynamics of speech patterns and their subtle changes in AD.

Subsequently, a **One-dimensional Average Pooling (AP-ID) layer** is applied. This temporal pooling step serves a dual purpose: it reduces the dimensionality of the feature sequences, thereby decreasing computational costs during subsequent training, and it helps to prevent model overfitting, which is particularly beneficial given the variable lengths of audio samples in real-world datasets.

The final component of the audio encoder is an **Audio Multi-Head Attention (MHA) layer**. MHA is a key component of the Transformer architecture, enabling the model to simultaneously attend to multiple different parts of the input sequence. By projecting the input into multiple smaller-dimensional subspaces, each head independently computes its own self-attention. In this context, Audio MHA extracts and emphasizes the most relevant correlations within the multidimensional temporal variables derived from the audio, allowing the model to focus on salient acoustic cues that are indicative of cognitive impairment.

**3.3. Text Encoder: Leveraging Contextual Language Models**

For the text modality, the DEMENTIA model leverages the power of the large language model **DistilBERT** to obtain embedding representations for each token in the text. DistilBERT is a "distilled" version of the more extensive BERT model. This means it is a smaller, faster, and more computationally efficient neural network trained to mimic the behavior of its larger "teacher" model through a sophisticated knowledge distillation process. This technique allows DistilBERT to retain approximately 97% of BERT's performance while being 40% smaller and 60% faster.

The strategic choice of DistilBERT is particularly relevant for clinical applications. Its optimized balance of performance and efficiency makes it significantly more practical for deployment in real-world clinical environments, which often have limited computational resources, or on edge devices. The "smaller, faster, cheaper and lighter" nature of DistilBERT directly translates to lower operational costs for healthcare providers and quicker turnaround times for patient assessments, factors that are critical for integrating AI into routine clinical workflows. DistilBERT generates rich embedding representations for each token in the text, effectively capturing contextually relevant information and semantic relationships within the linguistic data. This capability is critical for downstream Natural Language Processing (NLP) tasks related to AD assessment, such as identifying subtle linguistic markers of cognitive decline. Its training involves a triple loss function that combines distillation loss, masked language modeling loss, and cosine embedding loss, ensuring it maintains strong language understanding capabilities.

**3.4. Cross-Modal (CM) Attention Module: Fusing Audio and Text**

To adequately fuse the valid information between the audio and text modalities, the DEMENTIA framework incorporates a dedicated Cross-Modal (CM) Attention Module. The outputs of the audio and text encoders first undergo processing through

**one-dimensional convolutional kernels (Conv-1D)**. This step is crucial for transforming the features from both modalities into a common, compatible feature space, ensuring that they can be effectively compared and combined. While the Conv-1D settings for the text encoder are fixed, the convolutional kernel size for the audio encoder's output is adaptively computed to ensure proper dimension matching.

Following dimension alignment, **inter-modal fusion** is performed using cross-modal attention mechanisms. This involves two main attention flows, allowing each modality to inform and refine the representation of the other :

* **Audio-to-Text Attention (ATa→t​):** In this mechanism, the Query matrix is derived from the text vector, while the Key and Value matrices are derived from the audio vector. This allows the model to interactively learn and refine the high-level text representation by attending to relevant information extracted from the audio signals, such as prosodic cues or speech rate variations.
* **Text-to-Audio Attention (ATt→a​):** Conversely, the Query matrix originates from the audio vector, and the Key and Value matrices come from the text vector. This enables the audio representation to be enhanced and informed by contextual linguistic information, such as the semantic content of the words being spoken.

Both attention outputs are processed through Multi-Head Attention (MHA) and Layer Normalization (LN) to create MHAa→t​ and MHAt→a​ respectively, which are then added to the original modality outputs to enrich them.

Next, the outputs from these cross-modal attention mechanisms are concatenated to form a fused representation (Fat​). This combined feature is then fed into a

**global attention layer**, which calculates Luong's general attention scores. This process involves applying a softmax function to obtain attention weights (

αt​) for each time step, followed by a weighted summation to produce a context vector. This context vector, combined with the last time step output, is passed through an FC layer with a tanh activation to yield the final attention vector. This global attention mechanism enables the model to selectively focus on the most salient information across the deeply fused audio-text representation, ensuring that the most diagnostically relevant features are emphasized.

**3.5. Hybrid Attention and Concrete Dropout for Robustness**

The DEMENTIA model's "hybrid attention" is a sophisticated combination of three distinct attention mechanisms: the Audio-MHA within the audio encoder, the Cross-Modal MHA for inter-modal fusion, and the Global Attention layer that aggregates the fused audio-text representation. This multi-layered attention strategy is designed to comprehensively capture both intra-modal and inter-modal relationships, enhancing the model's ability to extract and prioritize relevant features from complex multimodal data.

The synergy of these attention mechanisms facilitates a sophisticated "dialogue" between different data types. The Audio-MHA refines the audio modality by capturing intra-modal temporal relationships, while DistilBERT (in the text encoder) provides contextual embeddings, effectively performing an implicit intra-modal attention for text. The Cross-Modal MHA then explicitly models inter-modal interactions, allowing the refined audio features to inform the text representation and vice-versa. This is crucial for understanding how acoustic cues (e.g., pauses) relate to linguistic content (e.g., word choice). Finally, Global Attention aggregates these deeply fused representations, focusing on the most salient features across both modalities. This layered approach ensures comprehensive capture of relationships: within each modality, between modalities, and overall salience for the diagnostic task. This architectural choice directly addresses the limitation of previous multimodal fusion methods that "overlook intrinsic correlations between modalities and interactions between model representations". It suggests that for complex biomedical data, simply concatenating features is insufficient; explicit, hierarchical mechanisms for cross-modal interaction are crucial for superior performance, deeper understanding of interdependencies, and robust feature learning.

The output of this hybrid attention module is then passed through a **Concrete Dropout (CD) layer** before being fed into a fully connected (FC) layer. Concrete Dropout is an advanced regularization technique that optimizes the dropout probability through gradient descent. This adaptive approach helps in preventing model overfitting, which is particularly crucial in scenarios with limited dataset sizes, such as those often encountered in biomedical research. It also improves the model's convergence speed and overall generalization ability by avoiding issues like gradient vanishing or exploding. This architectural choice reflects a pragmatic and robust approach to model design given common constraints in medical datasets. It suggests that advanced, adaptive regularization techniques are not just for marginal performance gains but are critical enablers for reliably deploying deep learning in data-scarce or sensitive domains, directly contributing to the model's good generalization performance.

**3.6. Expert Knowledge Integration and Final Output Layers**

The DEMENTIA model strategically integrates 14-dimensional speech features, which are meticulously designed and selected by domain experts. These features are chosen based on two primary criteria: their proven effectiveness for cognitive assessment in prior studies and their demonstrated good reliability. This expert knowledge is treated as a distinct and crucial modality within the framework. These handcrafted features provide complementary information that the deep learning network might not automatically capture from raw audio or text, thereby enriching the model with valuable domain-specific insights.

For the final fusion and prediction, this expert knowledge feature vector is concatenated with the fusion representation derived from the audio and text modalities (AT). This combined, enriched representation is then fed into subsequent Concrete Dropout and FC layers. Batch Normalization (BN) and a Rectified Linear Unit (ReLU) activation function are applied before the final FC layer, which then outputs the model's predicted probability for the AD detection task and the scaled MMSE score for the cognitive assessment task.

**3.7. Multi-Task Learning (MTL) and Loss Function Optimization**

The overall training of the DEMENTIA model is optimized using a stochastic gradient descent with momentum optimizer, initialized with a learning rate of 0.003 and a momentum of 0.9. This optimizer is responsible for minimizing the overall multi-task loss and updating the neural network parameters during the training process.

For the **AD detection task**, which is a binary classification problem, a standard binary classification cross-entropy loss (Lcls​) is utilized. This loss function quantifies the difference between the true AD/HC label (

t1​) and the model's predicted probability (p1′​) after a sigmoid activation, guiding the model to accurately distinguish between the two classes.

For the **cognitive assessment task** of predicting MMSE scores, the mean square error (MSE) loss (Lreg​) is employed. This loss function measures the average squared difference between the predicted MMSE scores and the actual MMSE values. The model's raw output for this task (

pn′​) is scaled by multiplying it by 30 (using a sigmoid activation) to map it to the standard 0-30 range of MMSE scores, ensuring the predictions are clinically meaningful.

The final multi-task loss (L) is defined as a weighted sum of the classification and regression losses: L=αLcls​+100α​Lreg​. The weight

α for the classification task is optimized as a hyperparameter. A crucial aspect of this weighting is that the regression loss is scaled down by a factor of 100 relative to the classification loss. This adjustment is made because the magnitude of the regression loss values was observed to be significantly larger (two orders of magnitude) than the classification loss during training. This careful balancing ensures that both tasks contribute appropriately to the overall model optimization and prevents one task from dominating the learning process, thereby facilitating effective multi-task learning.

**4. Expert Knowledge Integration: Bridging AI with Clinical Insights**

**4.1. Selection and Description of Expert-Derived Speech Features**

The DEMENTIA model strategically incorporates 14 specific speech features, carefully chosen based on two primary criteria: their proven effectiveness for cognitive assessment in prior studies and their demonstrated good reliability. This integration ensures that the model is grounded in established clinical and linguistic understanding of AD. These features are categorized into two groups: acoustic and linguistic, each providing distinct yet complementary information about speech patterns in AD.

**Acoustic Features (4):**

* **F0 Standard Deviation (SD):** This measures the variability of the fundamental frequency of vocal fold vibration, expressed in semitones. It reflects pitch stability and vocal control.
* **Duration Pause Intervals (DPI):** This represents the median duration of silent pauses within speech. It serves as an indicator of speech fluency and the cognitive processing load required for speech production.
* **Voiced Rate:** This quantifies the number of voiced segments per second. It reflects the overall speech production speed and fluency.
* **Hesitation Ratio:** This calculates the ratio of the total duration of hesitations to the total speech time, where hesitations are defined as voiceless segments lasting longer than 30 milliseconds. It provides a measure of disfluency and the cognitive effort involved in speech planning.

**Linguistic Features (10):**

* **Empty Word Frequency (EWF):** This is a specific metric defined in the study, counting occurrences of common stop words and disfluencies such as "oh," "uh," "=laughs," "well," "some," "she," "he," "him," "hm," "it," "down," "what," "fall," and nonsensical utterances marked as "xxx" in transcripts. This feature was motivated by observations that AD patients use these words more frequently than healthy controls.
* **Word Rate:** This is the number of words produced per second, reflecting overall speech fluency and speed.
* **Function Word Ratio:** This is the ratio of function words (including adverbs, prepositions, conjunctions, particles, interjections, and onomatopoeia) to the total number of words. Function words primarily express grammatical relationships.
* **Lexical Density:** This is the ratio of content words (lexical verbs, nouns, and adjectives) to the total number of words, indicating the informational richness and complexity of speech.
* **Mean Length of Utterance (MLU):** This is the ratio of words to the total number of sentences, reflecting syntactic complexity and sentence formulation ability.
* **Noun Phrase Rate:** This is the ratio of noun phrases to the total number of sentences, indicating the complexity of nominal structures.
* **Verb Phrase Rate:** This is the ratio of verb phrases to the total number of sentences, indicating the complexity of verbal structures.
* **Parse Tree Height:** This represents the average height of the syntactic subtree across all sentences, serving as a measure of overall syntactic complexity.
* **Yngve Depth Total:** This is the total Yngve depth of the syntactic subtree across all sentences, another measure of grammatical complexity that reflects sentence processing load.
* **Dependency Distance Total:** This is the average dependency distance per sentence, calculated as the sum of all dependency distances in each sentence divided by the total number of sentences, reflecting syntactic complexity based on word relationships.

**4.2. Statistical Validation and Clinical Significance of Expert Features**

Independent samples t-tests confirmed that all 14 expert knowledge-based speech features exhibited statistically significant differences between AD patients and healthy controls (HC). These findings are highly consistent with established clinical observations and previous research on how AD impacts speech and language, providing a strong empirical basis for the model's reliance on these features.

**Acoustic Manifestations:** AD patients showed a significantly slower voiced rate during the speech task compared to HC (p < 0.0001). Conversely, their Fundamental Frequency (F0) Standard Deviation, Duration Pause Intervals (DPI), and Hesitation Ratio were significantly higher (p < 0.05, p < 0.01, and p < 0.001, respectively). These acoustic patterns indicate increased vocal variability, longer and more frequent pauses, and a greater number of disfluencies, reflecting impaired motor speech control and increased cognitive effort during speech production.

**Linguistic Manifestations:** AD patients exhibited significantly higher Empty Word Frequency (EWF) and function word ratio (p < 0.001 and p < 0.01, respectively) compared to HC. This suggests a greater reliance on non-specific terms and grammatical connectors, often indicative of word-finding difficulties. In contrast, other linguistic features, including Word Rate, Lexical Density, Mean Length of Utterance (MLU), Noun Phrase Rate, Verb Phrase Rate, Parse Tree Height, Yngve Depth Total, and Dependency Distance Total, were significantly lower in AD patients. This reflects reduced linguistic complexity, diminished content richness, and impaired higher-order language planning.

Spearman correlation analysis further revealed linear relationships between most speech features and Mini-Mental State Examination (MMSE) scores, which measure global cognitive state.

* **Acoustic Correlations:** F0 SD and Hesitation Ratio showed a significant negative correlation with MMSE. This means that higher vocal variability and more hesitations were associated with lower cognitive scores.
* **Linguistic Correlations:** EWF and function word ratio were significantly negatively correlated with MMSE , indicating that increased use of empty words and grammatical connectors correlates with greater cognitive impairment. All other linguistic features (e.g., Word Rate, Lexical Density, MLU, syntactic complexity measures), with the exception of Yngve Depth Total, showed a significant positive correlation with MMSE. This implies that higher values of these features are associated with better cognitive function.

The explicit fusion of expert-derived knowledge features provides a crucial "biological anchor" to the AI's learning process. The comprehensive statistical validation for all 14 features quantitatively confirms these expert features as robust, clinically meaningful biomarkers. For instance, increased hesitation ratio and Empty Word Frequency directly reflect word-finding difficulties and increased cognitive load during speech production, which are core symptoms of AD. Similarly, reduced syntactic complexity, as measured by Parse Tree Height, Yngve Depth Total, and Dependency Distance Total, reflects the breakdown of higher-order language planning and executive function. This integration means the AI model is not merely learning abstract patterns from raw data; it is guided by and leverages features already known to be clinically relevant and reflective of underlying neurological changes in AD. This approach makes its decisions inherently more interpretable and trustworthy for clinicians.

The distinct changes observed across both acoustic and linguistic features provide a comprehensive "linguistic fingerprint" of AD, reflecting different facets of the disease's widespread impact on the brain. Acoustic changes, such as a slower voiced rate and increased Duration Pause Intervals, primarily reflect impairments in motor speech control, reduced speech fluency, and increased cognitive effort during speech production. The increased F0 SD specifically points to less stable pitch, potentially due to impaired vocal motor control. These are often early, subtle indicators of neurological decline affecting motor pathways. Linguistic changes, including higher Empty Word Frequency and function word ratio, and lower word rate, lexical density, MLU, noun/verb phrase rates, and syntactic complexity measures, reflect higher-level cognitive and language impairments. These include word-finding difficulties (leading to increased use of empty words and fillers), reduced lexical access, simplified sentence structures, and impaired semantic processing. The observation that AD patients use more "uh," "oh," "yeah," "well," "some," "she," "he" and less content words directly supports this. This multi-dimensional analysis suggests that speech is not just one biomarker, but a rich source of multiple, complementary biomarkers. The interplay of these features across acoustic and linguistic domains provides a robust and nuanced assessment of AD. This multi-faceted decline observed in speech aligns perfectly with AD's widespread impact on various cognitive and motor functions, making speech analysis a powerful, holistic diagnostic tool.

**Table 1: Expert Knowledge Features and Their Clinical Significance in AD**

| Feature Name | Category | Description | Observed Difference in AD vs. HC | Correlation with MMSE | Clinical/Biological Implication in AD |
| --- | --- | --- | --- | --- | --- |
| F0 Standard Deviation | Acoustic | Standard deviation of the fundamental frequency of vocal fold vibration, measured in semitones. | Significantly Higher in AD (p < 0.05) | Significant Negative Correlation | Reflects impaired vocal motor control and reduced pitch stability. |
| Duration Pause Intervals | Acoustic | Median duration of pause intervals. | Significantly Higher in AD (p < 0.01) | Significant Negative Correlation | Indicates increased cognitive load during speech planning and word-finding difficulties. |
| Voiced Rate | Acoustic | Number of voiced segments per second. | Significantly Slower in AD (p < 0.0001) | Significant Positive Correlation | Suggests reduced speech fluency and overall slowing of speech production. |
| Hesitation Ratio | Acoustic | Ratio of total hesitation duration to total speech time, where hesitation is defined as voiceless segments lasting longer than 30 ms. | Significantly Higher in AD (p < 0.001) | Significant Negative Correlation | Reflects increased disfluency and cognitive effort in speech planning. |
| Empty Word Frequency | Linguistic | Total count of words including "oh," "uh," "=laughs," "down," "well," "some," "what," "fall," "she," "he," "him," "hm," "it," and "xxx." | Significantly Higher in AD (p < 0.001) | Significant Negative Correlation | Indicates word-finding difficulties and increased reliance on non-specific fillers. |
| Word Rate | Linguistic | Number of words per second. | Significantly Lower in AD | Significant Positive Correlation | Shows reduced speech output and overall fluency. |
| Function Word Ratio | Linguistic | Ratio of function words (adverbs, prepositions, conjunctions, particles, interjections, onomatopoeia) to total words. | Significantly Higher in AD (p < 0.01) | Significant Negative Correlation | Suggests a shift from content-rich to grammatically-focused speech, often due to lexical access issues. |
| Lexical Density | Linguistic | Ratio of content words (lexical verbs, nouns, adjectives) to total words. | Significantly Lower in AD | Significant Positive Correlation | Reflects reduced informational richness and semantic complexity of speech. |
| Mean Length of Utterance | Linguistic | Ratio of words to total number of sentences. | Significantly Lower in AD | Significant Positive Correlation | Indicates simplified sentence structures and reduced syntactic complexity. |
| Noun Phrase Rate | Linguistic | Ratio of noun phrases to total number of sentences. | Significantly Lower in AD | Significant Positive Correlation | Suggests reduced complexity of nominal structures and potential noun retrieval difficulties. |
| Verb Phrase Rate | Linguistic | Ratio of verb phrases to total number of sentences. | Significantly Lower in AD | Significant Positive Correlation | Indicates reduced complexity of verbal structures. |
| Parse Tree Height | Linguistic | Average height of the syntactic subtree across all sentences. | Significantly Lower in AD | Significant Positive Correlation | Measures overall syntactic complexity, indicating simplification in AD speech. |
| Yngve Depth Total | Linguistic | Total Yngve depth of syntactic subtree across all sentences. | Significantly Lower in AD | Not Significant | Reflects grammatical complexity and sentence processing load. |
| Dependency Distance Total | Linguistic | Average dependency distance per sentence. | Significantly Lower in AD | Significant Positive Correlation | Indicates reduced syntactic complexity based on word relationships. |

**5. Multi-Task Learning Approach: Simultaneous Detection and Cognitive Assessment**

**5.1. Principles and Advantages of Multi-Task Learning (MTL)**

Multi-Task Learning (MTL) is a machine learning paradigm that involves training a single model to perform multiple related tasks simultaneously, rather than developing separate, independent models for each task. The core principle underlying MTL in the DEMENTIA framework is to enhance model efficiency, improve generalization ability, and increase robustness by enabling knowledge sharing across these related tasks. By allowing different tasks to share the same model architecture and parameters, the model can leverage common underlying patterns and representations that are beneficial for all tasks. This approach facilitates a mutual enhancement of performance across tasks, meaning that the learning process for one task can positively influence and improve the model's ability to perform another related task. Additionally, MTL can lead to reduced training time and computational costs compared to training individual models for each task, making it a more efficient approach for complex diagnostic systems.

**5.2. Dual Tasks: AD Detection and MMSE Prediction**

The DEMENTIA framework implements two distinct, yet inherently related, tasks within its Multi-Task Learning (MTL) architecture: Alzheimer's Disease (AD) detection and Mini-Mental State Examination (MMSE) score prediction.

The **Alzheimer's Disease Detection** is formulated as a binary classification task. Its primary objective is to distinguish between individuals diagnosed with AD and healthy controls (HC) based on their speech patterns. The model learns to classify speech samples into one of these two categories by processing the extracted multimodal features. The training objective for this classification task is minimized using a standard binary classification cross-entropy loss (Lcls​), which quantifies the dissimilarity between the model's predicted probabilities and the true diagnostic labels.

The **Mini-Mental State Examination (MMSE) Score Prediction** is implemented as a regression task. This task focuses on predicting the continuous MMSE score, a widely used clinical measure that provides a quantitative assessment of global cognitive function. The model learns to estimate a numerical score, typically ranging from 0 to 30, which reflects the severity of cognitive impairment. The training objective for this regression task is minimized using the mean square error (MSE) loss (Lreg​), which quantifies the average squared difference between the predicted MMSE scores and the actual MMSE values. To ensure the predicted outputs align with the clinical range, the model's raw output for this task is scaled by multiplying it by 30 after applying a sigmoid activation function.

The overall training of the DEMENTIA model is guided by a **combined multi-task loss function** (L), which is defined as a weighted sum of the classification and regression losses: L=αLcls​+100α​Lreg​. The weight

α for the classification task is treated as a tunable hyperparameter, allowing for adjustment of its relative importance. A critical aspect of this weighting scheme is that the regression loss is scaled down by a factor of 100 relative to the classification loss. This adjustment is necessary because the magnitude of the regression loss values was observed to be significantly larger (approximately two orders of magnitude) than the classification loss during training. This careful balancing ensures that both tasks contribute appropriately to the overall model optimization and prevents one task from dominating the learning process, thereby facilitating effective multi-task learning.

The use of this MTL paradigm creates a mutual reinforcement effect that is highly beneficial for clinical AD assessment. AD detection (binary classification) and MMSE prediction (continuous regression) are inherently linked in clinical practice. The features learned for distinguishing AD from healthy controls, such as the presence of disfluencies or simplified syntax, are highly relevant for predicting the degree of cognitive impairment, as reflected by the MMSE score, and vice-versa. By training these tasks together, the model learns a more generalized and robust representation of "cognitive impairment from speech." This mutual reinforcement means the model gains a richer, more nuanced understanding of the disease continuum rather than just a binary state. This is particularly valuable in clinical practice, where both a diagnosis and a severity score are often needed for comprehensive patient management and treatment planning. This MTL paradigm suggests that for diseases with a spectrum of severity, or where multiple diagnostic/prognostic outputs are required, MTL can yield more clinically useful and integrated models. Instead of separate, potentially inconsistent models for diagnosis and staging, a single MTL model can provide a cohesive assessment, potentially reducing computational overhead and improving consistency across different diagnostic outputs. This approach mirrors how clinicians holistically assess patients, considering both diagnosis and severity concurrently.

**6. Performance Evaluation and Generalizability**

**6.1. Performance on the ADRESS Dataset**

The DEMENTIA model's performance was rigorously evaluated on the ADRESS challenge dataset, a widely recognized benchmark for speech-based AD assessment.

For the **classification task** of AD detection, the model achieved an accuracy of 89.58%, a recall of 91.67%, an F1 score of 89.80%, and a precision of 88.00% on the test set. This performance places DEMENTIA tied for second among the top state-of-the-art methods reviewed in a recent survey. For instance, Wang et al. (2022) achieved 93.75% accuracy, while Yuan et al. (2020) also obtained 89.58% accuracy, and Liu et al. (2022) reported 87.50% accuracy. These results demonstrate DEMENTIA's highly competitive diagnostic capability.

For the **regression task** of MMSE prediction, the model achieved a Root Mean Square Error (RMSE) of 4.31 on the test set. This result positions DEMENTIA as the second-best performer in the regression task among the compared methods. The current best performance for MMSE regression was achieved by Koo et al. (2020) with an RMSE of 3.74, while Searle et al. (2020) and Farzana et al. (2020) reported RMSEs of 4.32 and 4.34, respectively. This indicates DEMENTIA's strong ability to quantitatively assess cognitive state.

**6.2. Insights from Comprehensive Ablation Experiments**

To rigorously validate the contribution of each module within the DEMENTIA model, comprehensive ablation experiments were conducted by systematically removing key components. The results, summarized in Table 2, underscore the critical role of each design choice.

**Impact of Attention Mechanisms:** The hybrid attention mechanism proved indispensable. Removing the Audio Multi-Head Attention (MHA) individually led to a drastic drop in classification accuracy to 45.83%. Similarly, removing the Cross-Modal MHA resulted in an accuracy of 81.25%, and removing the Global Attention mechanism yielded 83.33% accuracy. When all attention mechanisms were removed, both classification and regression performance significantly declined, highlighting the indispensable role of the hybrid attention in capturing complex relationships within and across modalities.

**Impact of Multi-Task Learning (MTL):** The benefits of the multi-task learning approach were empirically validated. Training single-task models (either regression or classification only) consistently yielded poorer results compared to the integrated MTL approach. For instance, training only the MMSE regression task resulted in an RMSE of 4.61 (compared to 4.31 with MTL), and training only the AD classification task achieved an accuracy of 83.33% (compared to 89.58% with MTL). This empirically validates the mutual reinforcement effect of MTL, where shared learning parameters enhance the performance of both tasks.

**Impact of Modalities:** Ablation studies demonstrated that removing any single modality (audio, text, or expert knowledge) or retaining only a single modality consistently led to a degradation in the model's overall performance. For example, removing audio, text, or expert knowledge resulted in accuracy drops to 79.17%, 80.00%, and 77.08% respectively, from the full DEMENTIA model's 89.58%. This provides strong empirical evidence for the effectiveness of multimodal fusion, indicating that each modality contributes unique and complementary information essential for robust AD assessment.

The observed performance degradation when modalities are removed suggests that each modality provides unique and complementary information that the others cannot fully capture on their own. Audio captures prosodic and acoustic cues (e.g., speech rate, pauses, F0 variability) that reflect motor speech and fluency. Text captures semantic content and syntactic complexity, reflecting higher-level language processing. Expert knowledge features, being handcrafted and clinically validated, likely provide a robust, low-dimensional representation of known AD markers that might be harder for deep learning to reliably extract from raw signals alone. The fusion allows the model to build a more complete and resilient picture of the patient's cognitive state by integrating these diverse signals. This redundancy and complementarity make the overall model more robust to noise or ambiguities present in any single modality. This reinforces the concept that a holistic approach, integrating data from different physiological and cognitive manifestations of a disease, leads to more accurate, comprehensive, and robust diagnostic tools. It also validates the continued relevance and synergistic effect of human-engineered features even within deep learning paradigms, especially where clinical interpretability and established biological links are desired.

**6.3. Generalizability on External Pitt Dataset**

To assess the model's ability to perform reliably on unseen data from different populations, the trained DEMENTIA model was directly applied to the external Pitt dataset. This dataset comprises 443 valid samples, including 272 dementia patients and 171 healthy controls. Unlike the ADRESS dataset, the Pitt dataset was not matched for age and sex, presenting a more challenging and realistic test of generalizability.

For **classification performance** on the Pitt dataset, the model achieved an accuracy of 80.81%, an F1 score of 83.23%, a recall of 77.57%, and a precision of 89.79% for AD detection. For

**regression performance**, the RMSE for predicting MMSE scores on the Pitt dataset was 4.38.

While these results may not always surpass the absolute state-of-the-art performance reported by models specifically tuned or augmented for the Pitt dataset (e.g., 84.78% accuracy with text, timestamp, and data augmentation, or 97.42% accuracy with an attention-based hybrid network) , they are highly significant. The consistent performance on an external, more diverse dataset sufficiently demonstrates the DEMENTIA model's robust generalization ability, which is a critical prerequisite for clinical applicability. For any AI diagnostic tool to be clinically useful, it must perform reliably on diverse, real-world patient populations, which often differ from controlled research datasets in demographics, recording conditions, and disease presentation. The Pitt dataset, with its unmatched demographics, represents a more challenging and realistic test of generalizability. The fact that DEMENTIA maintains strong performance on this external, less controlled dataset is a significant indicator of its potential for real-world application. Furthermore, the paper notes that there is currently no published research on MMSE prediction using the Pitt dataset, highlighting the novelty of DEMENTIA's contribution in this area. This emphasizes that beyond achieving peak performance on a single benchmark, robustness and generalizability across diverse datasets are paramount for clinical utility and building trust. A model that performs slightly less optimally but generalizes well is often more valuable in practice than one that achieves state-of-the-art on a single dataset but struggles with external validation. This directly addresses the "trust issues in clinical practice" and positions the model closer to real-world deployment.

**Table 2: DEMENTIA Model Performance Summary and Ablation Insights**

| Model Configuration | Acc (%) | F1 (%) | Rec (%) | Pre (%) | RMSE |
| --- | --- | --- | --- | --- | --- |
| **DEMENTIA (ADRESS Test Set)** | **89.58** | **89.80** | **91.67** | **88.00** | **4.31** |
| DEMENTIA (Pitt External Test Set) | 80.81 | 83.23 | 77.57 | 89.79 | 4.38 |
| w/o Audio-MHA | 45.83 | 13.33 | 8.33 | 85.71 | 5.40 |
| w/o CM-MHA | 81.25 | 80.00 | 75.00 | 86.36 | 5.25 |
| w/o Global Attention | 83.33 | 82.61 | 79.17 | 86.36 | 4.07 |
| w/o Hybrid Attention | 79.17 | 66.67 | 88.89 | 76.19 | 5.21 |
| only Regression | NA | NA | NA | NA | 4.61 |
| only Classification | 83.33 | 81.82 | 75.00 | 90.00 | NA |
| w/o Audio | 79.17 | 85.42 | 84.44 | 90.48 | 3.88 |
| w/o Text | 80.00 | 81.25 | 80.00 | 75.00 | 4.66 |
| w/o Expert Knowledge | 77.08 | 71.79 | 58.33 | 93.33 | 4.98 |
| only Audio | 50.00 | 63.16 | 60.42 | 55.81 | 5.48 |
| only Text | 85.42 | 83.72 | 75.00 | 94.74 | 4.67 |
| only Expert Knowledge | 85.42 | 84.44 | 79.17 | 90.48 | 5.06 |

Export to Sheets

*Note: Performance metrics for ablation experiments are based on the ADRESS test set. NA indicates "Not Applicable."*

**7. Explainability Analyses: Unveiling Model Decisions and Clinical Relevance**

The DEMENTIA model's design prioritizes transparency, and a series of post-hoc explainability analyses were conducted to unveil its internal decision-making processes across various modalities, thereby fostering clinical trust.

**7.1. Global Explainability of Expert Knowledge Modality (SHAP)**

The SHapley Additive exPlanations (SHAP) algorithm was applied to perform a post-hoc global explainability analysis specifically on the expert knowledge modality within the cognitive assessment (MMSE regression) task. This analysis quantifies the contribution of each feature to the model's output, providing a clear, clinically interpretable ranking of their impact on cognitive assessment.

The analysis revealed that **noun phrase rate** and **word rate** were the two most influential positive speech features. This indicates that higher values for these features, reflecting more complex nominal structures and faster, more fluent speech, significantly contributed to the model predicting higher MMSE scores, signifying better cognitive function. Conversely,

**Empty Word Frequency (EWF)** and **function word ratio** were identified as the two most significant negative features. Higher values for these features, indicating more disfluencies, filler words, and reliance on grammatical connectors, were strongly associated with the model predicting more severe dementia (lower MMSE scores). This analysis quantifies the relative importance of each expert-derived feature in the model's decision-making, providing a clear, clinically interpretable ranking of their impact on cognitive assessment.

**7.2. Local Explainability of Text Modality (LIME)**

The Local Interpretable Model-agnostic Explanations (LIME) method was employed to provide local, instance-specific explanations for the text modality, revealing which specific words or phrases influenced the model's prediction for individual speech samples.

For **healthy controls (HC)**, the model primarily focused on and assigned higher importance weights to meaningful **content words**, such as concrete nouns like "cookie," "curtains," "window," and "standing," as well as the definite article "the". This indicates that the model associates rich, specific vocabulary with healthy cognitive function. In contrast, for

**AD patients**, the model assigned greater importance to **pause fillers** (e.g., "mhm," "hm"), **pronouns** (e.g., "she," "he"), and **conjunctions** (e.g., "and"). This observation aligns with clinical understanding of AD patients exhibiting word-finding difficulties, which often leads to increased use of non-specific terms and disfluencies to maintain conversational flow. These findings are consistent with the word cloud maps presented in the study, which visually highlight the differential usage of words between AD and HC groups, and with prior linguistic studies on AD speech.

**7.3. Attention Patterns in Audio Modality (Mel-spectrogram Visualization)**

The Multi-Head Attention (MHA) output from the audio encoder was aligned and visualized with the Mel-spectrograms of the speech signals. This allowed for a direct observation of the model's emphasis on specific speech frames over time, providing insights into which acoustic segments were most influential in its decisions.

For **healthy samples**, the Mel-spectrograms typically showed a faster speech rate with a more uniform frequency and energy distribution across the audio. The model's attention scores were distributed across these segments, indicating a focus on coherent and fluent speech production. Conversely, for

**AD patients**, the Mel-spectrograms highlighted a slower speech rate and lower, less uniform energy levels. Crucially, the model assigned higher attention scores to these specific time frames, particularly focusing on segments corresponding to pauses and areas of reduced energy. This directly aligns with the statistical analyses of acoustic features, which showed higher Duration Pause Intervals and hesitation ratios, and a lower voiced rate in AD patients. This consistency indicates that the model effectively identifies and leverages these acoustic markers of cognitive impairment.

**7.4. Multimodal Feature Fusion Visualization (t-SNE)**

To explore the discriminative power of the model's internal representations, the high-dimensional features extracted from the ReLU activation layer (just before the final FC layer of the DEMENTIA model) were nonlinearly downsized to two dimensions using t-distributed stochastic neighbor embedding (t-SNE) for visualization.

The t-SNE visualizations of models trained on only audio, text, or expert knowledge modalities showed a tendency for AD and HC samples to overlap significantly. This indicates that no single modality alone provides sufficient information for robust and clear discrimination between the two groups. In stark contrast, the t-SNE visualization for the full multimodal DEMENTIA model demonstrated much better clustering and a clearer separation of AD and HC samples. This powerful visual evidence confirms that the proposed multimodal representation method, by effectively leveraging a hybrid attention network to capture advanced shared representations across all three modalities, offers significantly greater discriminative power for identifying AD.

This multi-pronged explainability strategy moves beyond a general understanding to provide specific, granular interpretability for clinical actionability. By providing detailed insights into which specific words, acoustic patterns, or expert features drive a diagnosis, the model transitions from a "black box" to a "glass box." This level of transparency is essential for healthcare professionals to understand not just *what* the model predicts, but *why* it makes those predictions, thereby fostering greater confidence and trust in AI-assisted decision-making in clinical settings.

**8. Conclusion**

Speech serves as both a "window into the mind" and a "microscope" for revealing an individual's cognitive state, offering an easy, non-invasive, and low-cost method for Alzheimer's Disease (AD) assessment. The DEMENTIA model represents a significant advancement in this domain, combining hybrid attention with multimodal representations to achieve simultaneous AD detection and cognitive assessment. The model demonstrated strong performance on the ADRESS dataset, achieving an accuracy of 89.58% and a recall of 91.67% for AD detection, along with an RMSE of 4.31 for cognitive assessment, results comparable to current state-of-the-art methods.

Comprehensive ablation experiments confirmed that the proposed method effectively captures intra- and inter-modal interactions, and that the combination of expert knowledge-based multimodal representations and multi-task learning significantly enhances performance. Testing on an external Pitt dataset further validated the model's robust generalization ability, a critical factor for real-world clinical applicability. Additionally, the post-hoc explainability analyses of the model and each modality provide trustworthy clinical support for assisted decision-making by revealing specific acoustic and linguistic markers of AD and demonstrating the discriminative power of multimodal fusion.

Despite these promising results, the study acknowledges several limitations. The dataset size is relatively small, and participants are predominantly white and English-speaking, necessitating further validation with larger and more diverse ethnic groups to ensure broader applicability. Furthermore, the full potential of all speech features within the model was not exhaustively explored, suggesting that additional a priori features based on expert knowledge could potentially enhance performance. Finally, while the model's internal decision-making processes were thoroughly examined, the current explanation at the neurobiological level requires further refinement. Future research should aim to incorporate speech analysis with neuroimaging or cerebrospinal fluid biomarkers to establish a clearer link between speech patterns, model decisions, and the underlying biological changes associated with AD