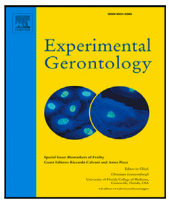
Experimental Gerontology 208 (2025) 112812

Contents lists available at ScienceDirect

Experimental Gerontology

journal homepage: www.elsevier.com/locate/exger



Multimodal LLM for enhanced Alzheimer's Disease diagnosis: Interpretable feature extraction from Mini-Mental State Examination data☆

Meiwei Zhang [a,1], Yuwei Pan [a,1], Qiushi Cui [a,*], Yang Lü [b], Weihua Yu [c]

[a] College of Electrical Engineering, Chongqing University, Chongqing, 400030, China
[b] Department of Geriatrics, The First Affiliated Hospital of Chongqing Medical University, Chongqing, 400016, China
[c] Institute of Neuroscience, Chongqing Medical University, Chongqing, 400016, China

ARTICLE INFO

Section Editor: Tibor Hortobagyi

Keywords:
Multimodal Large language model
MMSE
AD screening
Decision tree
Explainable AI

ABSTRACT

Alzheimer's Disease (AD) poses a considerable global health challenge, necessitating early and accurate diagnostics. The Mini-Mental State Examination (MMSE) is widely used for initial screening, but its traditional application often underutilizes the rich multimodal data it generates, such as videos, images, and speech. Integrating these modalities with modern Large Language Models (LLMs) offers untapped potential for improved diagnostics. In this study, we propose a multimodal LLM framework fundamentally reinterprets MMSE data. Instead of relying on conventional, often limited MMSE features, proposed LLM acts as a sophisticated cognitive analyst, directly processing MMSE modalities. This deep multimodal understanding allows for the extraction of novel, high-level features that transcend traditional metrics. These are not merely visual or acoustic signals, but rich semantic representations imbued with cognitive insights gleaned by the LLM. We then construct an interpretable decision tree classifier and derive a succinct rule list, yielding transparent diagnostic pathways readily understandable by clinicians. Finally, framework integrates a counterfactual explanation module to provide individualized "what-if" analyses, illuminating how minimal feature changes could alter model outputs. Our empirical study on real-world clinical data achieves a diagnostic accuracy of approximately 6% percentage points improvements with diagnosing explanation, reinforcing the viability of our framework as a promising, interpretable, and scalable solution for early AD detection.

1. Introduction

Alzheimer's Disease (AD) has become a global health crisis of unprecedented scale, significantly impacting aging populations and healthcare systems worldwide (Breijyeh and Karaman, 2020). The imperative for early and accurate diagnosis of AD is paramount, as timely intervention and management strategies can substantially improve patient outcomes and quality of life (Knopman et al., 2021). Clinical guidelines recommend using neuropsychological assessments, cerebrospinal fluid biomarkers, and multimodal imaging for AD diagnosis (Jack Jr. et al., 2018). Among these, the Mini-Mental State Examination (MMSE) remains the cornerstone of initial cognitive assessment (Cockrell and Folstein, 2002), widely used for its simplicity and ability to screen for cognitive impairment. MMSE inherently gathers diverse multimodal data, encompassing video recordings of tasks like paper folding and eye closure, image-based assessments of pentagon and handwriting, and spoken language responses (Zhang et al., 2025). However, in traditional clinical practice, most clinicians often face challenges in integrating and comprehensively judging heterogeneous medical data from multiple sources due to limitations in professional background, technological equipment constraints, or time pressure (Deng et al., 2023). The full diagnostic potential of the MMSE's rich multimodal information remains largely untapped.

Traditional approaches to analyzing MMSE data often rely on limited, hand-engineered features that may fail to capture the subtle yet crucial cognitive nuances embedded within its multimodal components (Calzà et al., 2021; Wen et al., 2020). Current methodologies frequently focus on isolated aspects of MMSE performance, potentially overlooking the synergistic insights that could be derived from a holistic, integrated analysis of video, image, and speech data (Zhang

☆ This study was supported by grants from Science Innovation Programs Led by the Academicians in Chongqing under Project (2022YSZX-JSX0002CSTB), Key Project of Technological Innovation and Application Development of Chongqing Science & Technology Bureau (CSTC2021jscx-gksb-N0020).

* Corresponding author.
E-mail addresses: zhangmw_play@163.com (M. Zhang), 2725361737@qq.com (Y. Pan), qcui@cqu.edu.cn (Q. Cui), yanglyu@hospital.cqmu.edu.cn (Y. Lü), yuweihua@cqmu.edu.cn (W. Yu).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.exger.2025.112812
Received 30 March 2025; Received in revised form 13 June 2025; Accepted 14 June 2025
Available online 3 July 2025

et al., 2024). This fragmented approach may limit diagnostic sensitivity, hindering the ability to detect early AD indicators present across multiple modalities. Therefore, there is a pressing need for innovative feature extraction techniques that can fully unlock the diagnostic potential of MMSE's multimodal data, moving beyond the constraints of conventional feature engineering.

Decision trees have long been valued in medical diagnostics for their interpretability and straightforward decision pathways (Moreno-Sanchez, 2020). However, while decision tree algorithms offer inherent interpretability, their traditional construction in complex medical domains often relies heavily on data-driven heuristics, potentially lacking the incorporation of expert knowledge and nuanced clinical reasoning (Murthy, 1998). Existing decision tree methods applied to AD diagnosis not fully leverage the wealth of pre-existing knowledge about cognitive decline and disease progression (Zhang et al., 2014). This limitation can hinder the ability of decision trees to capture subtle disease patterns and generalize effectively across diverse patient populations. This necessitates the exploration of novel decision tree construction methods that can integrate predefined knowledge and advanced reasoning capabilities to enhance diagnostic accuracy and robustness in AD screening.

Despite the increasing adoption of Artificial Intelligence (AI) in healthcare (Manne and Kantheti, 2021), a critical barrier to widespread clinical translation remains the limited interpretability and transparency of many AI-driven diagnostic tools (London, 2019). "Black box" models, while exhibiting excellent performance, neglect the need for model interpretability, leading to insufficient trust in AI decisions among clinicians (Volkov and Averkin, 2023), especially in high-stakes domains such as AD diagnosis (Bloch et al., 2022). The absence of transparent decision-making processes can impede clinical adoption and limit the ability to validate and refine AI-based diagnostic strategies in real-world settings. Thus, enhancing the interpretability and transparency of AI-driven AD diagnostic methods is crucial for fostering clinical trust, facilitating effective validation, and ultimately improving patient care.

The contributions of this paper are summarized as follows:

- The paper introduces a novel feature extraction pipeline that leverages a multimodal Large Language Model (LLM) to analyze diverse data modalities from traditional MMSE tasks (videos, images, and audio). This approach moves beyond conventional hand-crafted features, enabling a more comprehensive and nuanced capture of cognitive profiles relevant to AD diagnosis.
- An inherently interpretable diagnostic model utilizing a decision tree classifier was developed. Furthermore, the extraction of a compact rule list from the decision tree enhances the transparency and clinical utility of the model, facilitating understanding of the decision-making process for clinicians.
- Incorporating a counterfactual explanation module to provide clinically meaningful justifications for individual diagnostic predictions. By generating "what-if" scenarios, this module offers actionable insights into the factors influencing the diagnosis, thereby increasing the clinical relevance and interpretability of the AI system.
- The research culminates in a multi-faceted and integrated AI system designed for robust, explainable, and high-resolution diagnosis of AD. This holistic approach, combining novel feature extraction, interpretable modeling, and counterfactual explanations, contributes to the advancement of transparent and clinically applicable AI in the domain of neurodegenerative disease detection.

The rest of this article is framed as follows: The background of the AI-based AD diagnosing related work is provided in Section 2. The proposed methods is elaborated in Section 3. The experiment details are presented in Section 4, followed by the results and discussions in Section 5, and the conclusions and future work in Section 6.

2. Related work

The rapid advancements in Artificial Intelligence, particularly LLMs, offer transformative technological pathways for precise AD diagnosis, holding the potential to compensate for, and even partially replace, the limitations of human experts in multimodal data analysis (Feng et al., 2023). Multimodal LLMs have recently emerged as powerful tools for extracting integrated representations from diverse data sources, such as text, audio, and images (Baltrušaitis et al., 2018; Wang et al., 2024a,b; Thapa and Adhikari). Specifically, models like CLIP (Contrastive Language-Image Pre-training) have demonstrated strong capabilities in aligning visual and textual data (Radford et al., 2021), which could be beneficial for integrating MMSE-related imagery and verbal responses. In the context of MMSE, these models enable the ingestion of video-based tasks like paper folding, audio speech samples, and static images as a cohesive input. By generating context-aware embeddings, multimodal LLMs can capture nuanced cognitive cues and subtle clinical markers that might be missed by hand-engineered features (Vigo et al., 2022). Furthermore, ongoing advancements in GPT-like architectures, which handle visual and linguistic streams in a unified framework, promise further improvements in capturing intricate cross-modal relationships (Achiam et al., 2023; Ma et al., 2024), thus offering richer potential for early AD detection.

Decision trees have long been employed in medical diagnostics for their transparency and ease of rule interpretation, including in AD screening tasks (Chen et al., 2023). In typical implementations, a tree iteratively splits the data on meaningful features — such as memory scores or certain visuospatial tasks — leading to classification rules that can be readily interpreted by clinicians. However, standard tree-building heuristics may overlook domain-specific nuances, particularly when faced with heterogeneous AD datasets containing complex relationships among cognitive, behavioral, and biological variables (Geurts et al., 2006). Recent research has thus begun to explore tree-based methods combined with expert knowledge or machine-learned representations for enhanced accuracy, highlighting the potential for integrating advanced feature extraction with interpretable tree structures (Vyas et al., 2022; Jahan et al., 2023; Ahmadi et al., 2024; Mitra and Rehman, 2024).

Beyond model building and predictive accuracy, the interpretability of clinical AI models is essential for building trust and facilitating their adoption in real-world healthcare settings (Verma et al., 2020). Counterfactual explanations, in particular, provide "what-if" scenarios — illustrating minimal feature changes necessary to alter a prediction — thereby revealing decision boundaries and key decision-driving attributes (Dwivedi et al., 2023). In AD diagnosis, such counterfactuals can highlight whether small adjustments in metrics like speech fluency, motor coordination, or visuospatial accuracy could swing a borderline classification, potentially guiding clinicians toward more targeted follow-up assessments or interventions (Stepin et al., 2021). By bridging advanced multimodal feature extraction with tree-based decision rules and counterfactual insights, researchers aim to deliver models that are not only accurate but also aligned with the interpretability needs of real-world healthcare settings (Hsieh et al., 2023; Bohn et al., 2023; Bloch et al., 2024; Viswan et al., 2024).

3. Proposed method

In this section, we detail our proposed methodology for enhancing AD diagnosis through interpretable artificial intelligence. Our approach is threefold: first, we introduce a transformative feature extraction pipeline leveraging a multimodal LLM to analyze diverse data from MMSE tasks, moving beyond traditional metrics to capture a richer cognitive profile. Second, we deploy an interpretable decision tree classifier and derive a compact rule list, ensuring transparency and clinical utility in our diagnostic model. Finally, we integrate a counterfactual explanation module to provide insights into individual predictions,

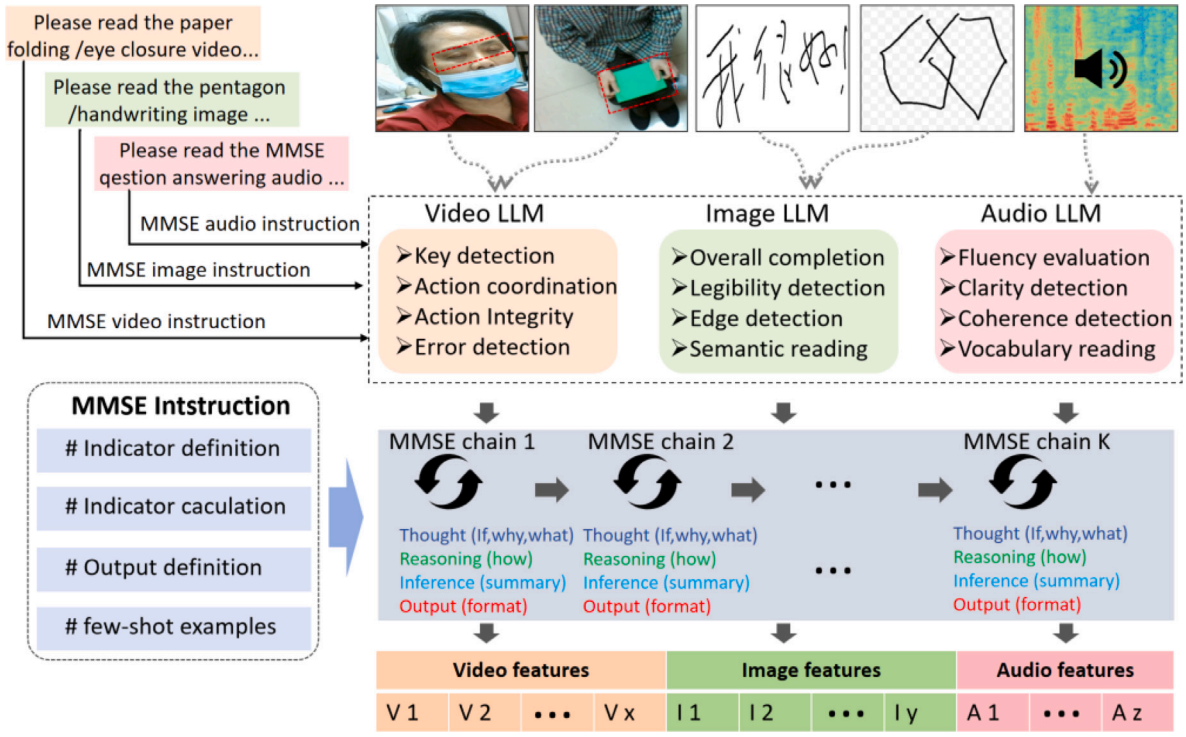


Fig. 1. Transformative feature extraction working diagram.

offering clinically relevant justifications and "what-if" scenarios to further understand and refine the diagnostic process. This multi-faceted approach aims to deliver a robust, explainable, and high-resolution diagnostic system for AD.

3.1. Transformative feature extraction via multimodal LLM

To accommodate the limitations of traditional MMSE tasks and to leverage the sophisticated reasoning capabilities of modern LLMs, we design a feature extraction framework spanning five MMSE-like tasks: *(1) Paper-folding video*, *(2) Eye-closing video*, *(3) Pentagon drawing*, *(4) Handwriting*, and *(5) Speech audio*. Whole process refers to Fig. 1. This pipeline is powered by a suite of specialized models from the Alibaba Cloud Qwen family to handle the different data modalities. Specifically, for visual tasks involving video and image analysis, we utilize the Qwen-VL-Max model. For the audio-based task, the Qwen-Audio-Turbo model is employed to process speech data. Numeric features are typically normalized to the range ([0,1]), whereas categorical features consist of a fixed set of discrete labels. Compared with conventional methods relying on hand-crafted rules or human raters, this LLM-driven pipeline provides:

1. **Comprehensive Analysis:** By ingesting raw multimodal inputs (videos of patient actions, images of drawings, speech audio), the LLM captures both subtle and explicit indicators of cognitive state.
2. **Consistent Scoring:** Proposed approach mitigates subjective variance and inter-rater reliability issues by relying on a unified model-based evaluation scheme.
3. **Extensible Metrics:** Beyond the rigid boundaries of standard MMSE, new dimensions can be seamlessly integrated.

Paper-Folding Video Features We define five features to measure how a subject folds paper from start to finish.

MotorCoordination (Numeric [0,1]). Assesses the smoothness of folding actions. A higher value indicates fewer tremors or uncoordinated motions. For instance:

$$\text{MotorCoordination} = 1 - \frac{\sum(\text{UnstableMotionTime})}{T_{\text{Total}}}. \tag{1}$$

FollowingInstruction (Numeric [0,1]). Reflects the ratio of correctly executed folding steps over the total number of steps:

$$\text{FollowingInstruction} = \frac{\text{CorrectSteps}}{\text{TotalSteps}}. \tag{2}$$

FinalShapeCorrectness (Numeric [0,1]). Evaluates how closely the final folded paper matches a reference shape:

$$\text{FinalShapeCorrectness} = \text{IoU}(\text{FoldedShape}, \text{ReferenceShape}). \tag{3}$$

ExecutionTime (Numeric, in seconds). The total time taken from the task start to completion:

$$\text{ExecutionTime} = t_{\text{end}} - t_{\text{start}}. \tag{4}$$

ObservedErrorType (Categorical). Classifies mistakes observed: {NoError, MinorError, MajorError}. For example, missing a minor fold might be MinorError, while folding an entirely wrong shape would be MajorError.

Eye-Closing Video Features Here, we focus on the subject's ability to understand and execute the "close your eyes" and "open your eyes" instructions, including reaction times.

ResponseLatency (Numeric, in seconds). The time interval between the prompt to close the eyes and the actual eye closure:

$$\text{ResponseLatency} = t_{\text{close\_observed}} - t_{\text{cmd}}. \tag{5}$$

ClosureStability (Numeric [0,1]). Measures how steadily the subject keeps eyes closed. Fewer re-openings or "fluttering" eyes raise the score. One possible formulation:

$$\text{ClosureStability} = 1 - \frac{\text{ReopenCount}}{\text{PossibleReopenLimit}}. \tag{6}$$

EyeOpeningDelay (Numeric, in seconds). The time from the "open your eyes" command to fully opening the eyes:

$$\text{EyeOpeningDelay} = t_{\text{open\_observed}} - t_{\text{cmd}}. \tag{7}$$

InstructionComprehension (Categorical). Reflects whether the subject fully understood and responded to the instructions: {Understood, Partial, NotUnderstood}.

Pentagon Drawing Features Visuospatial ability is often tested by having subjects draw a pentagon. These features quantify completion and geometric accuracy.

ShapeCompleteness (Numeric [0,1]). The fraction of edges drawn relative to the five expected sides of a pentagon:

$$\text{ShapeCompleteness} = \frac{\text{DrawnEdges}}{5}. \tag{8}$$

AngleConsistency (Numeric [0,1]). Evaluates how close each of the five angles is to the ideal angles of a regular pentagon:

$$\text{AngleConsistency} = 1 - \frac{\sum_{i=1}^{5} \left|\theta_i - \theta_{\text{ideal}}\right|}{5 \times \theta_{\text{ideal}}}. \tag{9}$$

LineOverlap (Categorical). Denotes the presence of retracing lines: {NoOverlap, PartialOverlap, SevereOverlap}.

GlobalRecognition (Numeric [0,1]). Indicates overall similarity to a standard pentagon. One might, for example, invert the Hausdorff distance:

$$\text{GlobalRecognition} = 1 - d_{\text{haus}}(\text{Drawing}, \text{Pentagon}). \tag{10}$$

Handwriting Features Handwriting tasks help assess fine motor skills and language abilities. We examine clarity, spacing, strokes, and completion.

Legibility (Numeric [0,1]). The proportion of discernible characters or words:

$$\text{Legibility} = \frac{\text{ReadableChars}}{\text{TotalChars}}. \tag{11}$$

SpacingConsistency (Numeric [0,1]). Measures how uniform the inter-character or inter-line spacing is:

$$\text{SpacingConsistency} = 1 - \frac{\sum_{i=1}^{n} \left|\delta_{\text{space},i} - \bar{\delta}\right|}{n \times \bar{\delta}}, \tag{12}$$

where $\delta_{\text{space},i}$ is the spacing in the ith region, and $\bar{\delta}$ is the mean spacing.

StrokeControl (Numeric [0,1]). Captures the fluidity of pen strokes:

$$\text{StrokeControl} = 1 - \frac{\text{TremorCount}}{\text{TotalStrokes}}. \tag{13}$$

OverallCompletion (Categorical). Indicates the writing task was accomplished: {Complete, Partial, Incomplete}.

Speech Audio Features Lastly, audio clips can reveal linguistic and articulatory aspects, which are critical for detecting early cognitive impairment.

SpeechFluency (Numeric [0,1]). Reflects the degree of continuous, uninterrupted speech:

$$\text{SpeechFluency} = 1 - \frac{\text{PauseTime}}{T_{\text{Speech}}}. \tag{14}$$

VocabularyRichness (Numeric [0,1]). Represents lexical diversity:

$$\text{VocabularyRichness} = \frac{\text{UniqueWords}}{\text{TotalWords}}. \tag{15}$$

PronunciationClarity (Numeric [0,1]). Indicates how distinctly words are articulated:

$$\text{PronunciationClarity} = 1 - \frac{\text{UnclearWordCount}}{\text{TotalWords}}. \tag{16}$$

Coherence (Categorical). Rates the logical or semantic flow of speech: {Consistent, PartiallyConsistent, Incoherent}.

By unifying these automatically extracted metrics, we seek to construct a robust, high-resolution cognitive profile that surpasses traditional MMSE summaries, thereby facilitating earlier or more precise detection of AD-related impairments.

3.2. Interpretable diagnosis tree

We deploy an interpretable classifier decision tree. The decision tree is learned by maximizing information gain $I(G)$ at each split. For a node containing subset S of the data, and a candidate split that partitions S into $\{S_L, S_R\}$, we compute:

$$I(G) = H(S) \; - \; \frac{|S_L|}{|S|} H(S_L) \; - \; \frac{|S_R|}{|S|} H(S_R), \tag{17}$$

where $H(S)$ is the entropy $H(S) = -\sum_{c\in\{0,1\}} p_c \log_2 p_c$ for class proportions p_c in S. The split yielding the highest $I(G)$ is chosen (greedy optimal). To avoid overfitting, we prune the tree by limiting the depth D and requiring a minimum leaf size m. Our final tree has $D = 5$ levels, providing a balance between complexity and interpretability. Each leaf of the tree corresponds to a decision rule (conjunction of conditions on features) leading to a predicted class.

In parallel, we derive a compact *rule list* model. This is essentially a sequence of if-then rules optimized for accuracy. For instance, the first rule might capture patients with very low speech fluency as AD ($\texttt{ifSpeechFluency} < 0.7 \rightarrow y = 1$). Subsequent rules cover other patterns (e.g., $\texttt{ifSpeechFluency} \geq 0.7\texttt{andFollowingInstruction} < 0.8 \rightarrow y = 1$), and a final default rule handles remaining cases. We extract such rules from the pruned decision tree, ensuring the list is short.

3.3. Counterfactual explanation module

For each subject i, let $\hat{y}_i = f(\tilde{\mathbf{x}}_i)$ be the model's prediction. If $\hat{y}_i \neq y_i$ (an error) or if an explanation is desired, we seek a counterfactual $\mathbf{x}_i' = \tilde{\mathbf{x}}_i + \Delta\mathbf{x}$ that changes the prediction: $f(\mathbf{x}_i') \neq f(\tilde{\mathbf{x}}_i)$. We formulate this as the constrained minimization problem stated above. In practice, we solve:

$$\min_{\Delta\mathbf{x}} \|\Delta\mathbf{x}\|_1 \;\; \text{s.t.} \;\; f(\tilde{\mathbf{x}}_i + \Delta\mathbf{x}) = y_{\text{target}}, \tag{18}$$

using a small subset of features for $\Delta\mathbf{x}$ (ensuring interpretability and feasibility). We use an iterative greedy algorithm to test feature changes: at each step, choose the feature j that most reduces the prediction error $|f(\mathbf{x}) - y_{\text{target}}|$, and adjust x_j toward the target boundary (for continuous features, by a small delta; for categorical, by flipping the category). This process repeats until the prediction flips. The result is a sparse $\Delta\mathbf{x}$ highlighting only a few changes—e.g., *"if Coherence were 'PartiallyConsistent' instead of 'Consistent'"*. These counterfactuals provide insight into model decisions and align with clinical reasoning by pinpointing what factor would need to change to alter the diagnosis.

All parameter settings within the proposed method and related algorithms were fixed for the baseline comparisons during the pre-experiment shown in Table 1.

4. Experimental setup

This study employed a real-world clinical dataset sourced from the First Affiliated Hospital of Chongqing Medical University. The dataset consists of clinical data from individuals who underwent evaluation at the hospital. Participant classification into AD and Cognitively Normal (CN) groups was determined based on a combination of education level and MMSE scores, aligning with expert clinical consensus. Specifically, the diagnostic criteria for AD were defined as follows: an MMSE score below 18 for illiterate participants, an MMSE score less than 22 for participants with 1–11 years of education, and an MMSE score below

Table 1
Key parameters for baseline machine learning Algorithms.

Model	Parameter	Value
Decision Tree (Baseline)	criterion splitter	'gini' 'best'
K-Nearest Neighbors (KNN)	n_neighbors (k) metric	7 'euclidean'
Gradient Boosting	n_estimators learning_rate	150 0.1
SVM (RBF Kernel)	C gamma	10 'scale'
Multi-Layer Perceptron (MLP)	hidden_layer_sizes activation	(100,) 'relu'
Quadratic Discriminant Analysis (QDA)	reg_param (Standard implementation)	0.0 –

24 for participants with 12 or more years of education. Individuals not meeting these criteria were classified as CN. While Mild Cognitive Impairment (MCI) represents a crucial stage for early detection, its inclusion was limited in this study primarily due to constraints on the availability and robust classification of MCI cases within the collected clinical dataset. Participant selection was exclusively based on these diagnostic criteria, without considering factors such as age or gender, to ensure a focused investigation on core cognitive indicators. The final dataset comprised 160 participants, equally divided into AD and CN groups, with 80 participants in each group. Notably, each participant's data included a comprehensive multimodal dataset encompassing origami video recordings, eye-closed video recordings, pentagon drawing images, handwriting images, and voice recordings.

The demographic characteristics of the 160 participants were analyzed to provide context for the study cohort. The mean age of the overall group was 71.1 ± 9.1 years (range: 46–90), with 115 (71.9%) females and an average of 10.5 ± 4.6 years of education. When comparing the diagnostic groups, the Alzheimer's Disease (AD) group (n=80) was significantly older (73.5 ± 8.9 vs. 68.6 ± 8.8 years, $p < .001$) and had fewer years of education (7.9 ± 4.5 vs. 13.0 ± 3.4 years, $p < .001$) than the Cognitively Normal (CN) group (n=80). Gender distribution did not differ significantly between the groups ($p = 0.368$). These baseline differences are consistent with known risk factors for AD and underscore the importance of our analytical approach, which focuses on cognitive performance features.

In the experimental framework, all transformative feature extraction procedures were conducted using multimodal LLMs from the Alibaba Cloud Qwen series. It is crucial to clarify that these LLMs were used exclusively for zero-shot inference to extract features, and no fine-tuning was performed with our dataset. The extracted features were then used to train and test the downstream decision tree classifier. Specifically, we utilized the Qwen-VL-Max model, a large-scale multimodal model accessed via Alibaba Cloud's AI services, for processing video, image, and audio inputs. While the exact parameter count for this specific API-based tier is not publicly disclosed, it is known to be a substantial model designed for complex multimodal understanding tasks, indicative of a large scale. The detailed prompts and configurations employed for feature extraction from each modality are provided in Appendix to ensure reproducibility. Key inference parameters included a low temperature (0.3) to ensure deterministic outputs and a max token limit of 2048 to prevent truncation. The performance of the proposed methodology was evaluated using a standard suite of classification metrics. These metrics included accuracy, F1-score, and recall. These metrics collectively provide a robust evaluation of the model's capability to accurately and reliably differentiate between AD and CN individuals based on the multimodal data.

5. Results and discussion

Our experimental analysis systematically evaluates the proposed framework through three critical dimensions: (1) the discriminative power of LLM-derived cognitive metrics, (2) diagnostic accuracy-complexity tradeoffs in interpretable modeling, and (3) clinical utility enhancement via counterfactual reasoning. We first validate the information richness of our multimodal feature space by comparing its diagnostic resolution against conventional MMSE scoring paradigms. Subsequently, we dissect the performance characteristics of the knowledge-infused decision tree across varying architectural configurations, contrasting its operation against both traditional machine learning baselines and clinical interpretability requirements. Finally, we quantify how counterfactual explanation mechanisms refine diagnostic precision while preserving model transparency. This tripartite evaluation strategy ensures comprehensive validation of our framework's technical innovation and clinical applicability.

5.1. Proposed diagnosing tree performance

The proposed diagnosis tree as Fig. 2, it reveals how the model hierarchically distinguishes patients with likely AD from CN controls. Notably, `SpeechFluency` emerges as the primary splitting criterion at the root, suggesting that a threshold around 0.72 is highly discriminative for AD risk. Among those with sufficiently high `SpeechFluency`, `ClosureStability` and `EyeOpeningDelay` further refine the prediction by gauging the smoothness and timing of eye-closing tasks, reflecting important markers of motor and attentional control. Meanwhile, for the subtree where `SpeechFluency` remains relatively high, additional features like `Legibility` and `StrokeControl` in handwriting tasks help isolate subtle deficits in visuomotor skills. This layered structure provides a transparent decision path, offering clinically interpretable cutoffs — such as whether `FinalShapeCorrectness` exceeds 0.81 — thereby aligning with common cognitive assessments. Overall, the tree highlights a clear diagnostic flow: from global speech patterns to more nuanced measures of motor execution and visual–spatial accuracy.

Table 2 summarizes the classification performances of several common machine learning techniques juxtaposed with the newly proposed decision tree. Notably, our decision tree achieves an overall accuracy of 0.76 and a recall of 0.82, reflecting a balanced ability to correctly identify AD cases while maintaining a strong overall predictive power. In contrast, methods such as Support Vector Machines (SVM) with the RBF kernel and Quadratic Discriminant Analysis (QDA) yield perfect recall (1.00) at the cost of considerable lower accuracy (0.55 and 0.52, respectively), suggesting they excessively label samples as positive for AD. While this high recall can be beneficial for sensitive screening, the accompanying drop in accuracy potentially reduces clinical utility by inflating false positives. To ensure a robust evaluation and mitigate bias from a single data split, the performance metrics for all algorithms presented in Table 2 were calculated using 5-fold cross-validation across the dataset.

Meanwhile, K-Nearest Neighbors (KNN) and Gradient Boosting (GB) present moderate performance profiles (accuracy and recall both

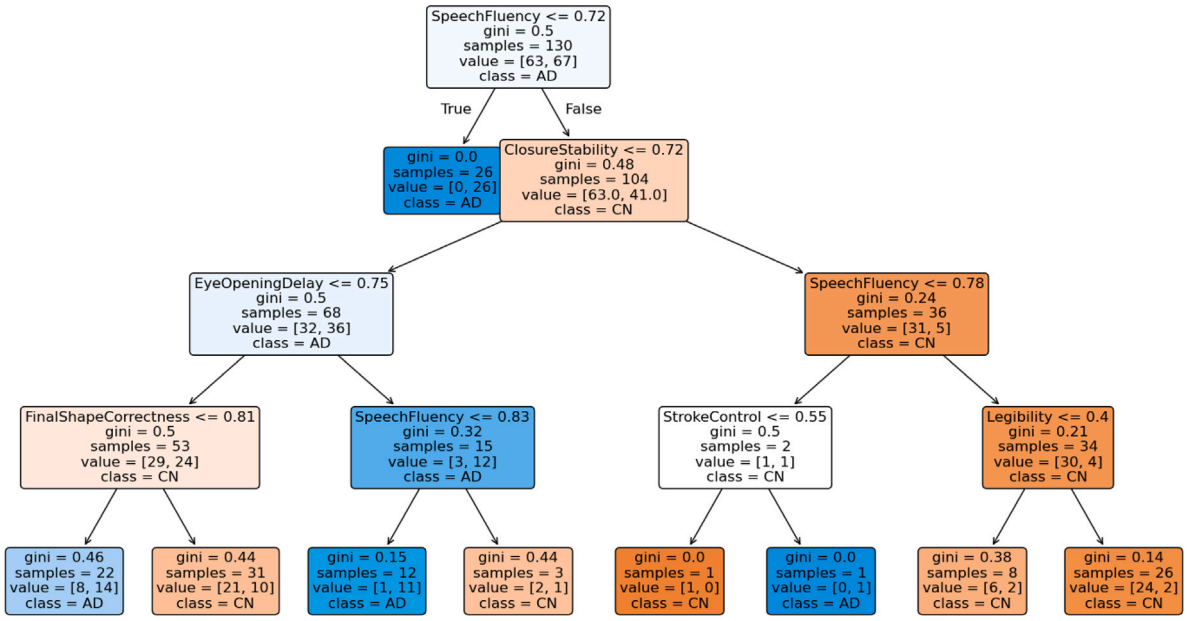


Fig. 2. Diagnosing tree structure.

Table 2
Diagnosing results within different algorithms.

Algorithm	Accuracy	Recall	F1-score
Decision Tree (New)	0.76 ± 0.03	0.82 ± 0.06	0.76 ± 0.12
LR (MMSE Score)	0.74 ± 0.03	0.80 ± 0.05	0.74 ± 0.11
K-Nearest Neighbors	0.70 ± 0.07	0.71 ± 0.08	0.71 ± 0.09
Gradient Boosting	0.70 ± 0.07	0.71 ± 0.02	0.71 ± 0.03
SVM (RBF)	0.55 ± 0.07	1.00 ± 0	0.69 ± 0.05
MLP	0.73 ± 0.03	0.65 ± 0.07	0.71 ± 0.08
QDA	0.52 ± 0.09	1.00 ± 0	0.68 ± 0.02

around 0.70–0.71), indicating relatively balanced yet less competitive performance. A Multi-Layer Perceptron (MLP) offers improved accuracy (0.73), but recall decreases to 0.65, indicating the model may overlook a substantial fraction of AD-positive cases. By contrast, the proposed decision tree demonstrates a higher recall without sacrificing overall accuracy or F1-score, thereby striking a desirable balance between sensitivity and specificity. Furthermore, from a clinical standpoint, the interpretability of the decision tree provides clear diagnostic pathways and rule-based insights—an advantage that purely black-box methods cannot offer. These results underscore not only the empirical efficacy of the decision tree but also its practical alignment with diagnostic objectives, namely, robust detection of positive cases alongside transparent clinical reasoning.

To establish a clinically relevant benchmark, we trained a Logistic Regression model using only the standard numerical MMSE score, which achieved a competitive accuracy of 0.74 and a recall of 0.80. Despite this strong baseline, our proposed decision tree, leveraging rich LLM-extracted features, surpasses these results with an accuracy of 0.76 and a recall of 0.82. This improvement demonstrates that our framework captures subtle, diagnostically crucial information from multimodal behaviors that is inherently lost when relying solely on the single, aggregated MMSE score.

5.2. Counterfactual explanations influence for diagnosing tree

Tables 4 and 5 examine the diagnostic accuracy of the proposed decision tree under varying tree depths and minimum leaf sizes, both before and after incorporating counterfactual explanations. Prior to counterfactual refinement, we observe that a shallow depth of 3 generally yields suboptimal performance, with a maximum accuracy of only 0.64 and recall of 0.41, suggesting an underfit model that fails to capture sufficiently discriminative partitions. Increasing the tree depth to 4 or 5 improves the balance between accuracy and recall (up to 0.73), indicating that deeper trees better separate subtle variations in cognitive features. Nevertheless, past a depth of 5, improvements plateau or fluctuate, hinting at potential overfitting or diminishing returns from further complexity. Notably, the choice of minimum leaf size has limited influence on overall metrics within the same depth level, possibly due to the modest dataset size.

Following counterfactual explanation integration, we see more pronounced variability across the same range of parameters. In particular, specific configurations (e.g., depth = 5, leaf = 2; depth = 7, leaf = 2) exhibit larger gains, achieving peak accuracies of 0.76 and recalls above 0.80. These improvements underscore how counterfactual insights can lead to a more refined decision boundary—by revisiting misclassified samples and allowing the model to better account for meaningful feature adjustments, the tree can effectively reduce false negatives while maintaining or improving overall accuracy. Conversely, some parameter combinations (e.g., depth = 5, leaf = 5) appear to revert to lower scores after counterfactual refinement, implying that not all tuned configurations adapt equally well to the newly introduced corrective feedback. This observation suggests that properly matching model complexity to the quantity and quality of counterfactual corrections is critical for maximizing gains.

Collectively, these results highlight the delicate interplay between model depth, leaf size, and counterfactual-based corrections. While deeper trees tend to capture more complex decision boundaries, they may also risk overfitting without an appropriate regularization scheme. However, when coupled with targeted counterfactual adjustments, well-chosen parameters (such as depth = 5 or 7 with a modest leaf size) deliver a more robust and interpretable AD diagnostic tool. This synergy of parameter tuning and counterfactual refinement offers a promising route for balancing specificity and sensitivity—particularly crucial in clinical screening where the cost of misclassification is high and the need for model transparency is paramount.

To ensure the methodological transparency of this process, it is important to note that the hyperparameters for the decision tree were not chosen arbitrarily. The values presented in 4 and 5 represent a systematic exploration. Specifically, parameters such as max_depth and min_samples_leaf were tuned using a grid search combined with 5-fold cross-validation. The final configuration of max_depth=5 and

Table 3
Wrong prediction samples and updated predictions.

Wrong prediction sample id	Current prediction	Suggestion	New prediction
[55, 29, 19, 30, 18, 12, 9, 31, 56, 78]	Actual=1, Predicted=0	SpeechFluency: 0.8 →≤ 0.725	AD (1)
[159, 131]	Actual=0, Predicted=1	ClosureStability: 0.7 →> 0.725	CN (0)

Table 4
Diagnosing tee performance with different parameters before counterfactual explanation.

Depth	Leaf	Accuracy	Recall	F1-score
3	2	0.64	0.41	0.54
3	5	0.64	0.41	0.54
3	7	0.64	0.41	0.54
3	10	0.64	0.41	0.54
4	2	0.73	0.71	0.73
4	5	0.73	0.71	0.73
4	7	0.73	0.71	0.73
4	10	0.73	0.71	0.73
5	2	0.70	0.71	0.71
5	5	0.67	0.65	0.67
5	7	0.67	0.65	0.67
5	10	0.67	0.65	0.67
7	2	0.67	0.71	0.69
7	5	0.67	0.65	0.67
7	7	0.67	0.65	0.67
7	10	0.67	0.65	0.67

Table 5
Diagnosing tee performance with different parameters after counterfactual explanation.

Depth	Leaf	Accuracy	Recall	F1-score
3	2	0.64	0.41	0.54
3	5	0.64	0.41	0.54
3	7	0.64	0.41	0.54
3	10	0.64	0.41	0.54
4	2	0.70	0.65	0.69
4	5	0.70	0.65	0.69
4	7	0.73	0.71	0.73
4	10	0.73	0.71	0.73
5	2	0.76	0.82	0.78
5	5	0.64	0.59	0.63
5	7	0.67	0.65	0.67
5	10	0.67	0.65	0.67
7	2	0.73	0.82	0.76
7	5	0.64	0.59	0.63
7	7	0.67	0.65	0.67
7	10	0.67	0.65	0.67

min_samples_leaf=2 was selected as it provided the optimal trade-off between cross-validated performance (accuracy and recall) and the need for clinical interpretability, ensuring the model is both effective and understandable.

To further investigate the discriminative capability of our LLM-extracted features, we visualized the feature space defined by the two most influential predictors from our decision tree model: SpeechFluency and ClosureStability. Fig. 3 presents a scatter plot of these features, with each point representing a participant, colored by their diagnostic group.

Crucially, we have overlaid the primary decision boundaries from our tree model as dashed lines on the plot. The vertical red line at SpeechFluency = 0.72 corresponds to the root node split, while the horizontal green line at ClosureStability = 0.72 represents the subsequent key decision. The plot clearly shows that these two features effectively partition the data. A high concentration of CN participants is clustered in the upper-right quadrant, characterized by high scores in both fluency and stability. Conversely, AD participants predominantly occupy the other regions, either exhibiting low SpeechFluency (left of the red line) or displaying impaired ClosureStability even with adequate speech performance (below the green line). This visualization provides a powerful and intuitive confirmation that our LLM-based features create a separable feature space, and it visually validates the logic of our interpretable decision tree model.

5.3. Explanation analysis

In examining the impact of counterfactual modifications on individual misclassifications, we identified several samples where minor feature adjustments flipped the model's prediction. As shown in Table 3, most of these cases involved tuning `SpeechFluency` down from around 0.80 to a threshold of about 0.72 for patients actually labeled AD but wrongly predicted as cognitively normal (CN). Similarly, the misclassified CN cases typically required slightly improving `ClosureStability` (e.g., from 0.70 to above 0.725) to regain the correct prediction. In Fig. 4, each sub-plot illustrates how a small shift in the identified key feature suffices to cross the model's decision boundary—underscoring both the sensitivity of the classifier to crucial dimensions (such as speech or motor-control attributes) and the potential for interpretable, localized corrections. These counterfactual insights not only highlight which features most strongly influence the diagnostic outcome, but also demonstrate the model's capacity for nuanced, clinically relevant adjustments that directly connect to patient behaviors observed in MMSE-like tasks.

6. Conclusions and future work

In this study, we introduced a multimodal LLM-based framework for AD diagnosis that reinterprets conventional MMSE tasks by extracting a high-resolution cognitive profile from videos, images, and speech data. Through an interpretable decision tree complemented by a succinct rule list, our proposed method demonstrates a balanced trade-off between accuracy and recall, while offering rule-based transparency for clinical contexts. Moreover, the integration of counterfactual explanations enables individualized insights into the key features driving each prediction, supporting both model validation and potential refinement of cognitive assessments. Together, these contributions highlight the framework's capacity to more fully leverage MMSE's multimodal richness to detect early-stage AD with improved interpretability.

Despite these advances, several challenges merit further attention. First, the overall diagnostic performance remains constrained by the scale and variability of the available dataset, as well as by current LLM capabilities in capturing subtle cognitive signals from diverse media. Second, the feature transformation process, although more granular than traditional MMSE scoring, can be extended to include finer-grained motor, linguistic, and visual cues that might yield higher diagnostic sensitivity. Furthermore, while cross-validation provides a robust internal evaluation, validation on a large, independent test set is a crucial step for confirming generalizability, which is currently limited by data availability. A primary limitation of this study is that it is based on data from a single clinical center, potentially limiting the direct generalizability of our findings to different populations, healthcare settings, or variations in MMSE administration protocols. While our methodology intentionally focused on cognitive performance features to mitigate the influence of these demographic confounders, further validation on more diverse datasets is needed to fully address this limitation. Future work will thus focus on curating larger, more heterogeneous data repositories and refining the feature extraction pipeline to incorporate additional clinical insights, ultimately aiming to enhance both accuracy and generalizability of the proposed methodology. As part of this effort, we plan to perform formal bias and fairness analyses to ensure the model performs equitably across different demographic subgroups. Exploring domain adaptation techniques could further enhance the model's applicability to new clinical environments.

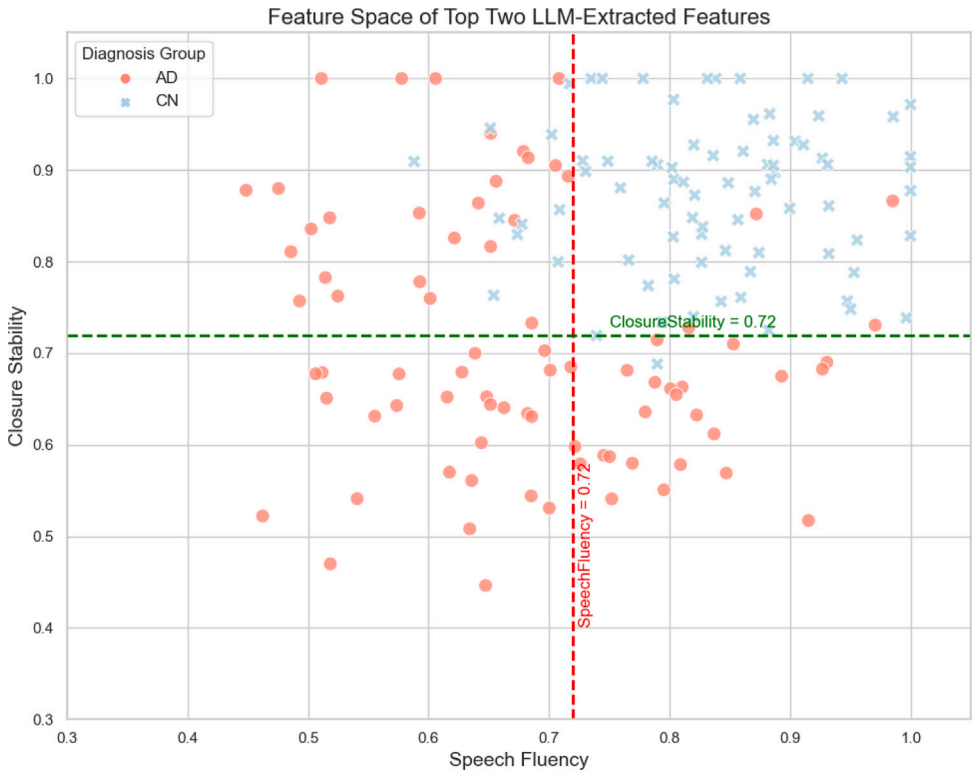


Fig. 3. Top two LLM-extracted features.

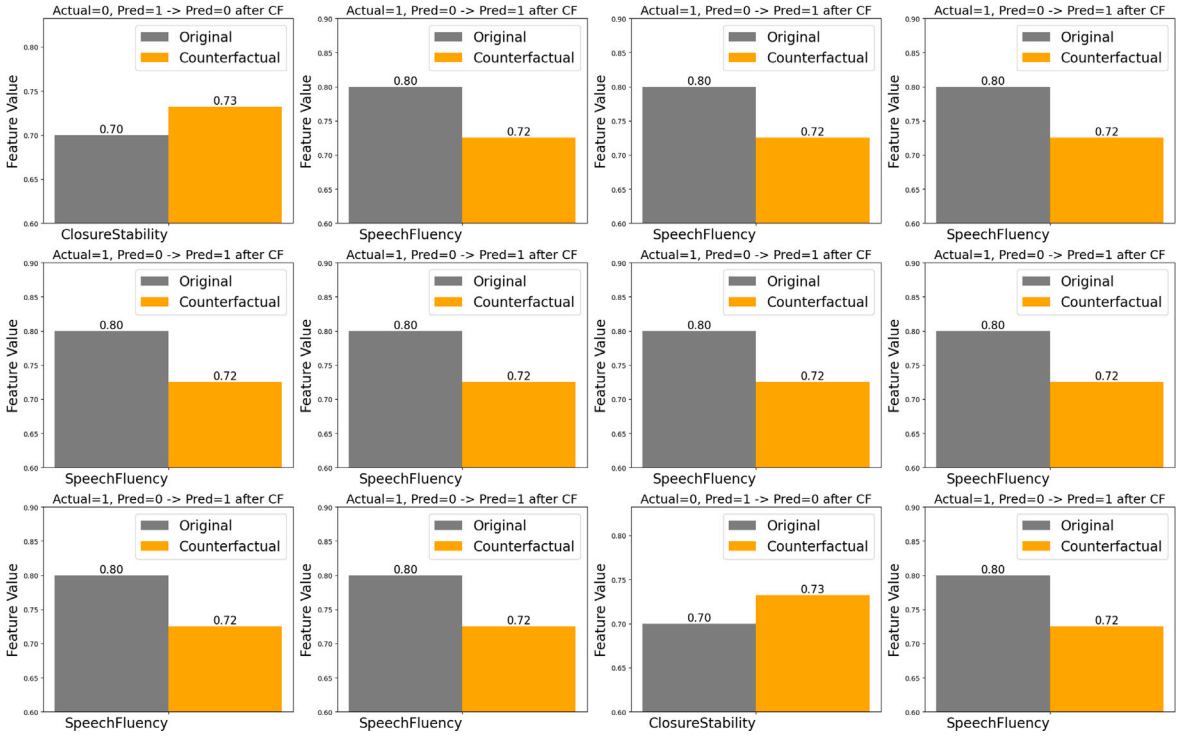


Fig. 4. Counterfactual visuals.

CRediT authorship contribution statement

Meiwei Zhang: Writing – review & editing, Writing – original draft, Supervision, Investigation, Formal analysis, Data curation. **Yuwei Pan:** Writing – original draft, Validation, Supervision, Investigation, Formal analysis, Data curation, Conceptualization. **Qiushi Cui:** Visualization, Validation, Software, Methodology. **Yang Lü:** Supervision, Investigation, Funding acquisition, Data curation. **Weihua Yu:** Funding acquisition, Data curation.

Data and prompts statement

The NSA data and the prompts supporting the findings of this study are available from the corresponding author Qiushi Cui, upon reasonable request.

Ethical statement

This study was conducted in accordance with the ethical principles of the Declaration of Helsinki. The research protocol was reviewed and approved by the Ethics Committee of The First Affiliated Hospital of Chongqing Medical University (Approval No: 2021-492). All participants, or their legally authorized representatives, provided written informed consent prior to their inclusion in the study. All data was fully anonymized to protect patient confidentiality.

Declaration of generative AI in scientific writing

The authors retain full responsibility and accountability for the entirety of this work, affirming the absence of generative AI in its conception and development.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Prompts settings

A.1. Pentagon drawing prompt

Role setting. You are an experienced cognitive-assessment expert, especially skilled in using the MMSE scale to screen for Alzheimer's disease (AD).

Task description. I will provide an image of a pentagon drawn by a patient in the MMSE pentagon-copying task. From a cognitive-assessment perspective, analyse this image and, for each feature dimension below, give a score in the range 0–1 (higher values indicate better cognitive ability in that aspect). (During training you may supply example images.)

Output features (0–1 unless otherwise noted):

- *ShapeCompleteness*: Whether all five sides are fully drawn.
- *AngleConsistency*: Match between the five angles and a regular pentagon.
- *LineOverlap*: "NoOverlap", "PartialOverlap", or "SevereOverlap".
- *GlobalRecognition*: Overall resemblance to a standard pentagon.

Example output:

{ "ShapeCompleteness": 0.75, "AngleConsistency": 0.80, "LineOverlap": "NoOverlap", "GlobalRecognition": 0.78 }

A.2. Handwriting prompt

Role setting. You are an experienced cognitive-assessment expert, especially skilled in using the MMSE scale to screen for AD.

Task description. I will provide a handwriting image produced by a patient in the MMSE sentence-writing task. Analyse it and score each feature dimension below in the range 0–1 (higher means better).

Output features:

- *Legibility*: Clarity of handwriting.
- *SpacingConsistency*: Uniformity of inter-word and inter-line spacing.
- *StrokeControl*: Smoothness of strokes.
- *OverallCompletion*: "Complete", "Partial", or "Incomplete".

Example output:

{ "Legibility": 0.90, "SpacingConsistency": 0.65, "StrokeControl": 0.85, "OverallCompletion": "Complete" }

A.3. Audio prompt

Role setting. You are an experienced cognitive-assessment expert for MMSE-based AD screening.

Task description. I will provide an audio clip of a patient's spoken responses during an MMSE task. Score each feature below (0–1 unless stated).

Output features:

- *SpeechFluency*
- *VocabularyRichness*
- *PronunciationClarity*
- *Coherence*: "Consistent", "PartiallyConsistent", or "Incoherent"

Example output:

{ "SpeechFluency": 0.85, "VocabularyRichness": 0.70, "PronunciationClarity": 0.88, "Coherence": "Consistent" }

A.4. Paper-folding video prompt

Role setting. You are an experienced cognitive-assessment expert.

Task description. A video shows a patient following the instruction "Pick up this sheet with your right hand, fold it in half, and place it on your thigh". Score:

- *MotorCoordination*
- *FollowingInstruction*
- *FinalShapeCorrectness*
- *ExecutionTime*: shorter is better (0–1)
- *ObservedErrorType*: "NoError", "MinorError", "MajorError"

Example output:

{ "MotorCoordination": 0.82, "FollowingInstruction": 0.90, "FinalShapeCorrectness": 0.78, "ExecutionTime": 95, "ObservedErrorType": "MinorError" }

A.5. Eye-closure video prompt

Role setting. Experienced MMSE assessor.

Task description. A video shows a patient performing the eye-closure task. Provide:

- *ResponseLatency*: time to close eyes (0–1)
- *ClosureStability*
- *EyeOpeningDelay*
- *Attention*

- *InstructionComprehension*: "Understood", "Partial", "NotUnderstood"

Example output:

{ "ResponseLatency": 1.2, "ClosureStability": 0.95, "EyeOpeningDelay": 2.7, "InstructionComprehension": "Understood" }

Data availability

The authors do not have permission to share data.

References

Achiam, Josh, Adler, Steven, Agarwal, Sandhini, Ahmad, Lama, Akkaya, Ilge, Aleman, Florencia Leoni, Almeida, Diogo, Altenschmidt, Janko, Altman, Sam, Anadkat, Shyamal, et al., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Ahmadi, Mohsen, Javaheri, Danial, Khajavi, Matin, Danesh, Kasra, Hur, Junbeom, 2024. A deeply supervised adaptable neural network for diagnosis and classification of Alzheimer's severity using multitask feature extraction. Plos One 19 (3), e0297996.

Baltrušaitis, Tadas, Ahuja, Chaitanya, Morency, Louis-Philippe, 2018. Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 41 (2), 423–443.

Bloch, Louise, Friedrich, Christoph M., Alzheimer's Disease Neuroimaging Initiative, 2022. Machine learning workflow to explain black-box models for early Alzheimer's disease classification evaluated for multiple datasets. SN Comput. Sci. 3 (6), 509.

Bloch, Louise, Friedrich, Christoph M., Alzheimer's Disease Neuroimaging Initiative, et al., 2024. Systematic comparison of 3D deep learning and classical machine learning explanations for Alzheimer's disease detection. Comput. Biol. Med. 170, 108029.

Bohn, Linzy, Drouin, Shannon M., McFall, G. Peggy, Rolfson, Darryl B., Andrew, Melissa K, Dixon, Roger A., 2023. Machine learning analyses identify multi-modal frailty factors that selectively discriminate four cohorts in the Alzheimer's disease spectrum: a COMPASS-ND study. BMC Geriatr. 23 (1), 837.

Breijyeh, Zeinab, Karaman, Rafik, 2020. Comprehensive review on Alzheimer's disease: causes and treatment. Molecules 25 (24), 5789.

Calzà, Laura, Gagliardi, Gloria, Favretti, Rema Rossini, Tamburini, Fabio, 2021. Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. Comput. Speech & Lang. 65, 101113.

Chen, Kaile, Abtahi, Farhad, Carrero, Juan-Jesus, Fernandez-Llatas, Carlos, Seoane, Fernando, 2023. Process mining and data mining applications in the domain of chronic diseases: A systematic review. Artif. Intell. Med. 144, 102645.

Cockrell, Joseph R., Folstein, Marshal F., 2002. Mini-mental state examination. Princ. Pr. Geriatr. Psychiatry 140–141.

Deng, Yi, Wang, Haiyin, Gu, Kaicheng, Song, Peipei, 2023. Alzheimer's disease with frailty: Prevalence, screening, assessment, intervention strategies and challenges. Biosci. Trends 17 (4), 283–292.

Dwivedi, Rudresh, Dave, Devam, Naik, Het, Singhal, Smiti, Omer, Rana, Patel, Pankesh, Qian, Bin, Wen, Zhenyu, Shah, Tejal, Morgan, Graham, et al., 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. ACM Comput. Surv. 55 (9), 1–33.

Feng, Yingjie, Xu, Xiaoyin, Zhuang, Yueting, Zhang, Min, 2023. Large language models improve Alzheimer's disease diagnosis using multi-modality data. In: 2023 IEEE International Conference on Medical Artificial Intelligence (MedAI). IEEE, pp. 61–66.

Geurts, Pierre, Ernst, Damien, Wehenkel, Louis, 2006. Extremely randomized trees. Mach. Learn. 63, 3–42.

Hsieh, Kang-Lin, Plascencia-Villa, German, Lin, Ko-Hong, Perry, George, Jiang, Xiaoqian, Kim, Yejin, 2023. Synthesize heterogeneous biological knowledge via representation learning for Alzheimer's disease drug repurposing. Iscience 26 (1).

Jack Jr., Clifford R., Bennett, David A., Blennow, Kaj, Carrillo, Maria C, Dunn, Billy, Haeberlein, Samantha Budd, Holtzman, David M., Jagust, William, Jessen, Frank, Karlawish, Jason, et al., 2018. NIA-AA research framework: toward a biological definition of Alzheimer's disease. Alzheimer' s & Dement. 14 (4), 535–562.

Jahan, Sobhana, Abu Taher, Kazi, Kaiser, M. Shamim, Mahmud, Mufti, Rahman, Md Sazzadur, Hosen, ASM Sanwar, Ra, In-Ho, 2023. Explainable AI-based Alzheimer's prediction and management using multimodal data. Plos One 18 (11), e0294253.

Knopman, David S., Amieva, Helene, Petersen, Ronald C., Chételat, Gäel, Holtzman, David M., Hyman, Bradley T., Nixon, Ralph A., Jones, David T., 2021. Alzheimer disease. Nat. Rev. Dis. Prim. 7 (1), 33.

London, Alex John, 2019. Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent. Rep. 49 (1), 15–21.

Ma, Luyi, Li, Xiaohan, Fan, Zezhong, Zhao, Kai, Xu, Jianpeng, Cho, Jason, Kanumala, Praveen, Nag, Kaushiki, Kumar, Sushant, Achan, Kannan, 2024. Triple modality fusion: Aligning visual, textual, and graph data with large language models for multi-behavior recommendations. arXiv preprint arXiv:2410.12228.

Manne, Ravi, Kantheti, Sneha C., 2021. Application of artificial intelligence in healthcare: chances and challenges. Curr. J. Appl. Sci. Technol. 40 (6), 78–89.

Mitra, Uddalak, Rehman, Shafiq Ul, 2024. Ml-powered handwriting analysis for early detection of Alzheimer's disease. IEEE Access.

Moreno-Sanchez, Pedro A., 2020. Development of an explainable prediction model of heart failure survival by using ensemble trees. In: 2020 IEEE International Conference on Big Data (Big Data). IEEE, pp. 4902–4910.

Murthy, Sreerama K., 1998. Automatic construction of decision trees from data: A multi-disciplinary survey. Data Min. Knowl. Discov. 2, 345–389.

Radford, Alec, Kim, Jong Wook, Hallacy, Chris, Ramesh, Aditya, Goh, Gabriel, Agarwal, Sandhini, Sastry, Girish, Askell, Amanda, Mishkin, Pamela, Clark, Jack, et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PmLR, pp. 8748–8763.

Stepin, Ilia, Alonso, Jose M, Catala, Alejandro, Pereira-Fariña, Martín, 2021. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. Ieee Access 9, 11974–12001.

Thapa, Surendrabikram, Adhikari, Surabhi, Leveraging ChatGPT-like large language models for Alzheimer's disease: Enhancing care, advancing research, and overcoming challenges. In: Smart Healthcare Systems. CRC Press, pp. 265–275.

Verma, S., Boonsanong, V., Hoang, M., Hines, K.E., Dickerson, J.P., Shah, C., 2020. Counterfactual explanations and algorithmic recourses for machine learning: A review. ACM Comput. Surv. 56, 1–42. http://dx.doi.org/10.1145/3677119.

Vigo, Ines, Coelho, Luis, Reis, Sara, 2022. Speech-and language-based classification of Alzheimer's disease: a systematic review. Bioengineering 9 (1), 27.

Viswan, Vimbi, Shaffi, Noushath, Mahmud, Mufti, Subramanian, Karthikeyan, Hajamohideen, Faizal, 2024. Explainable artificial intelligence in Alzheimer's disease classification: A systematic review. Cogn. Comput. 16 (1), 1–44.

Volkov, Egor N., Averkin, Aleksej N., 2023. Explainable artificial intelligence in medical image analysis: State of the art and prospects. In: 2023 XXVI International Conference on Soft Computing and Measurements. SCM, IEEE, pp. 134–137.

Vyas, Akhilesh, Aisopos, Fotis, Vidal, Maria-Esther, Garrard, Peter, Paliouras, Georgios, 2022. Identifying the presence and severity of dementia by applying interpretable machine learning techniques on structured clinical records. BMC Med. Inform. Decis. Mak. 22 (1), 271.

Wang, Jiankun, Ahn, Sumyeong, Dalal, Taykhoom, Zhang, Xiaodan, Pan, Weishen, Zhang, Qiannan, Chen, Bin, Dodge, Hiroko H, Wang, Fei, Zhou, Jiayu, 2024a. Augmented risk prediction for the onset of Alzheimer's disease from electronic health records with large language models. arXiv preprint arXiv:2405.16413.

Wang, Zhepeng, Bao, Runxue, Wu, Yawen, Liu, Guodong, Yang, Lei, Zhan, Liang, Zheng, Feng, Jiang, Weiwen, Zhang, Yanfu, 2024b. A self-guided multimodal approach to enhancing graph representation learning for Alzheimer's diseases. arXiv preprint arXiv:2412.06212.

Wen, Junhao, Thibeau-Sutre, Elina, Diaz-Melo, Mauricio, Samper-González, Jorge, Routier, Alexandre, Bottani, Simona, Dormont, Didier, Durrleman, Stanley, Burgos, Ninon, Colliot, Olivier, et al., 2020. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. Med. Image Anal. 63, 101694.

Zhang, Meiwei, Cui, Qiushi, Li, Wenyuan, Yu, Weihua, Chen, Lihua, Li, Wenjie, Zhu, Chenzhe, Lü, Yang, 2025. Augmented dialectal speech recognition for AI-based neuropsychological scale assessment in Alzheimer's disease. Biomed. Signal Process. Control. 99, 106821.

Zhang, Meiwei, Cui, Qiushi, Lü, Yang, Yu, Weihua, Li, Wenyuan, 2024. A multimodal learning machine framework for Alzheimer's disease diagnosis based on neuropsychological and neuroimaging data. Comput. Ind. Eng. 197, 110625.

Zhang, Yu-Dong, Wang, Shuihua, Dong, Zhengchao, 2014. Classification of Alzheimer disease based on structural magnetic resonance imaging by kernel support vector machine decision tree. Prog. Electromagn. Res. 144, 171–184.

npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01940-4

CARE-AD: a multi-agent large language model framework for Alzheimer's disease prediction using longitudinal clinical notes

Rumeng Li[1,2], Xun Wang[3], Dan Berlowitz[2,4,5], Jesse Mez[6], Honghuang Lin[7] & Hong Yu[1,2,5,8]

Large language models (LLMs) have shown promising capabilities across diverse domains, yet their application to complex clinical prediction tasks remains limited. In this study, we present CARE-AD (Collaborative Analysis and Risk Evaluation for Alzheimer's Disease), a multi-agent LLM-based framework for forecasting Alzheimer's disease (AD) onset by analyzing longitudinal electronic health record (EHR) notes. CARE-AD assigns specialized LLM agents to extract signs and symptoms relevant to AD and conduct domain-specific evaluations—emulating a collaborative diagnostic process. In a retrospective evaluation, CARE-AD achieved higher accuracy (0.53 vs. 0.26–0.45) than baseline single-model approaches in predicting AD risk 10 years prior to the first recorded diagnosis code. These findings highlight the feasibility of using multi-agent LLM systems to support early risk assessment for AD and motivate further research on their integration into clinical decision support workflows.

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by cognitive decline, memory impairment, and functional deterioration, ultimately leading to loss of independence in affected individuals[1]. Being the most common cause of dementia worldwide, AD imposes a significant burden on patients, caregivers, and healthcare systems[2]. With the aging global population, the prevalence of AD is expected to rise substantially in the coming decades, underscoring the urgent need for early detection and effective management strategies[2].

Although formal diagnosis of AD typically involves cognitive assessments and biomarker-based tests, these procedures are often costly, invasive, and impractical for large-scale screening and expensive, limiting their widespread adoption in clinical practice[3–5]. Meanwhile, studies have identified early indicators of AD risk that emerge well before formal diagnosis[6]. Subjective cognitive decline and prodromal symptoms of AD dementia frequently manifest years in advance, involving subtle and often neglected changes in memory, cognition, and behavior[6–9]. Recognizing these indicators is crucial for early AD prediction and intervention[10]. Nevertheless, such signs are often overlooked because they are frequently described within unstructured electronic health record (EHR) notes rather than documented in standardized fields such as International Classification of Diseases (ICD) codes or lab results[2,11]. As a result, much of the critical pre-diagnostic information remains underutilized.

Previous research has explored the use of structured EHR data for early AD prediction[12–19], but relatively few studies have incorporated unstructured narratives. Existing NLP efforts have largely focused on isolated symptoms or specific note types, limiting their generalizability across longitudinal clinical records[11,20–27]. Recent advances in large language models (LLMs), such as OpenAI's GPT-4 and Meta's LLaMA family, offer new opportunities to extract complex patterns from free-text data[28–30]. However, significant challenges remain for healthcare applications, including data privacy, model scalability, and the limitations of single-model reasoning in capturing the multidimensional nature of clinical decision-making[31].

To address these challenges, we drew inspiration from the clinical diagnostic process for AD, which relies on a rigorous multidisciplinary approach. In clinical practice, specialists in neurology, psychiatry, geriatrics, primary care, and other relevant fields, each contribute complementary expertise to comprehensively assess patient risk[32–34]. This collaborative model is essential for evaluating multifactorial conditions like AD, where diverse symptom domains must be integrated for an accurate and nuanced assessment.

[1]Manning College of Information & Computer Sciences, University of Massachusetts Amherst, Amherst, MA, USA. [2]Center for Health Organization & Implementation Research, VA Bedford Health Care System, Bedford, MA, USA. [3]Microsoft Corporation, Redmond, WA, USA. [4]Department of Public Health, University of Massachusetts Lowell, Lowell, MA, USA. [5]Center of Biomedical and Health Research in Data Sciences, University of Massachusetts Lowell, Lowell, MA, USA. [6]Chobanian & Avedisian School of Medicine, Boston University, Boston, MA, USA. [7]Department of Medicine, UMass Chan Medical School, Worcester, MA, USA. [8]Miner School of Computer and Information Sciences, University of Massachusetts Lowell, Lowell, MA, USA. e-mail: hong_yu@uml.edu

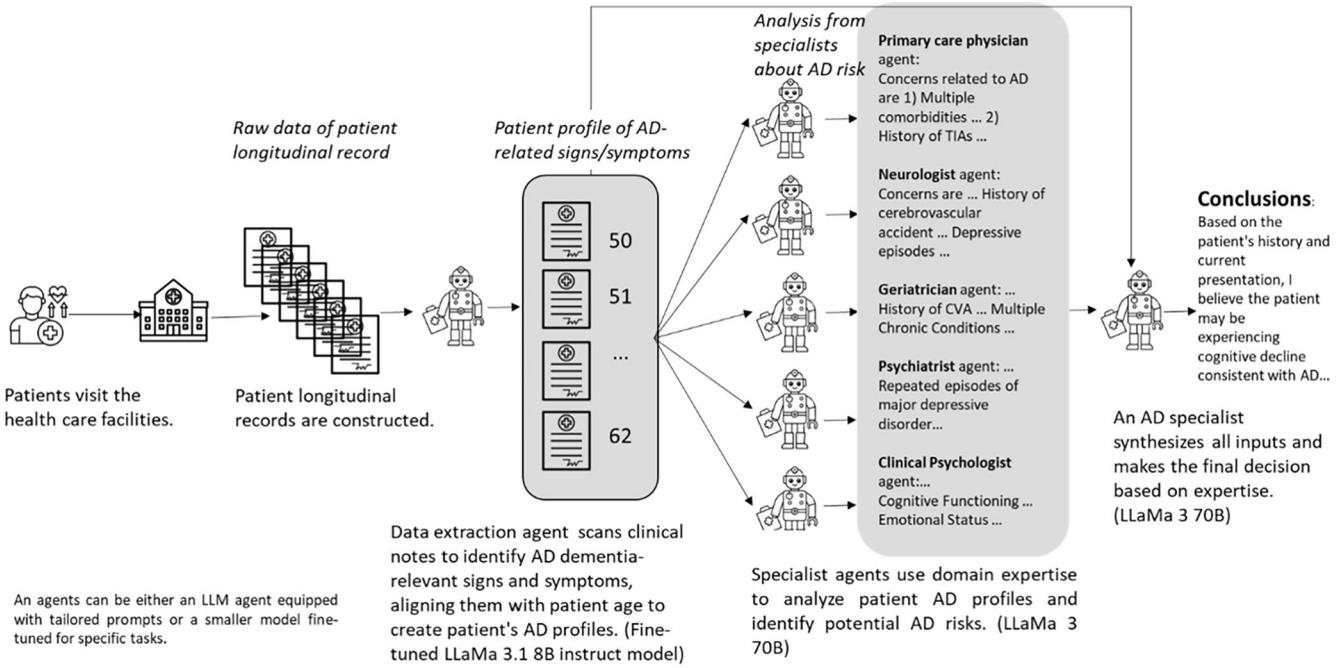


Fig. 1 | Illustration of an example patient processed by the CARE-AD framework for early AD prediction using longitudinal EHRs.

Table 1 | Demographics of the evaluation set

Characteristic	AD case	Control
Patients, No.	1000	3627
AD onset/index time age (SD)	79.0 (8.4)	78.6 (8.6)
Sex, *n* (%)		
Female	18 (1.8%)	57 (1.6%)
Male	982 (98.2%)	3570 (98.4%)
Race no. (%)		
White	793 (79%)	2913 (80.3%)
Black or African American	131 (13.1%)	400 (11.0%)
Other	76 (7.6%)	314 (8.7%)
Ethnicity no. (%)		
Hispanic/Latino	66 (6.6%)	253 (7.0%)
Non-Hispanic/Latino	903 (90.3%)	3282 (90.5%)
Others/Unknown	31 (3.1%)	92 (2.5%)
AD-relevant sentences per year (average)	99	60

We propose to simulate this clinical procedure through a multi-agent framework, with each agent representing a specialist domain. By mimicking the collaboration of clinicians, this design aims to enhance predictive performance and interpretability. Multi-agent methods have shown promise in healthcare tasks such as medical question answering[35] and mitigating cognitive biases in clinical decision-making[36]. Coordinating specialized agents not only improves prediction accuracy but also yields clearer intermediate reasoning steps—an important factor for clinical transparency and trust. This approach is also conceptually aligned with the Mixture of Experts (MoE) paradigm[37], which demonstrates that specialization across expert components can improve performance on complex tasks.

Building on these insights from clinical practice and model specialization, we developed CARE-AD (Collaborative Analysis & Risk Evaluation for Alzheimer's Disease)—a multi-agent LLM framework designed to predict AD risk from longitudinal unstructured EHR data. CARE-AD simulates a virtual multidisciplinary consultation: agents representing clinical domains such as primary care, neurology, psychiatry, geriatrics, and psychology analyze a patient's symptom trajectory and provide domain-specific assessments. These are then synthesized by an AD specialist agent into an individualized risk prediction. By modeling temporal symptom patterns and incorporating diverse clinical perspectives, CARE-AD aims to improve sensitivity to early AD-related signs—especially those often underrepresented in structured records—while enhancing interpretability through agent-specific contributions that clinicians can review.

While further validation in real-world clinical workflows is needed, this study presents the design and evaluation of CARE-AD on a large dataset from the U.S. Veterans Health Administration (VHA), demonstrating its potential to improve early AD risk stratification and support more informed clinical decision-making.

Results

The CARE-AD prediction framework involves three steps to assess AD risk. First, a data extraction agent identifies AD-related signs and symptoms from EHR notes organizing them into age-aware patient profiles categorized by specific symptom types. Second, a multidisciplinary team of specialist agents —including a primary care physician agent for holistic assessment, neurologist and psychiatrist agents for neurological and psychiatric evaluation, a geriatrician agent for assessing daily living and independence, and a clinical psychologist agent for behavioral and psychological analysis—conducts a coordinated, domain-specific evaluation. Finally, an AD-focused specialist agent synthesizes these insights to generate a robust AD risk assessment. The framework is illustrated in Fig. 1. A detailed description of the system overview and technical architecture is provided in Supplementary Note 1.

Study sample

Our cohort consists of 17,488 AD cases and 64,691 controls from the VHA. Supplementary Fig. 1 illustrates the cohort creation process. We assessed the CARE-AD prediction framework using a randomly sampled evaluation set of 1000 AD cases and 3627 controls. Demographic details of the evaluation set are presented in Table 1.

Performance of data extraction agent

Our data extraction agent is designed to identify signs and symptoms of AD from longitudinal EHR notes using a two-step classification process. First, we perform a binary classification to determine whether a candidate sentence contains any AD-related signs or symptoms. Second, for those sentences deemed relevant, we apply a multi-class classification to assign each instance to one of the five expert-defined AD categories: cognitive impairment, notice/concern by others, requiring assistance/functional impairment, physiological changes, and neuropsychiatric symptoms. We trained separate LLaMA 3.1 8B models for each classification step on a dataset derived from previous work (Supplementary Table 1)[38]. Unlike earlier approaches, we excluded categories involving cognitive assessments or diagnostic tests, because our method relies strictly on symptom-based evidence rather than formal clinical investigations of AD. Comprehensive descriptions of each category are available in the Methods section and in Supplementary Note 2.

Table 2 presents the performance of our first fine-tuned LLaMA 3.1 8B model on the binary classification step, demonstrating whether sentences indicate AD-related signs or symptoms. We compare this model against a strong ensemble baseline[38], which integrates three pretrained language models—BERT (bert-base-uncased), RoBERTa (roberta-base), and ClinicalBERT—fine-tuned on our dataset. Our LLaMA 3.1 8B model outperforms this ensemble model, highlighting its effectiveness on the initial binary decision. Sentences identified as relevant are then processed by our second fine-tuned LLaMA 3.1 8B model, which performs multi-class classification to assign each instance to one of five symptom categories. These category-specific outputs help generate detailed inputs for subsequent LLM agents. The classifier's evaluation results are reported in Table 3.

Based on the classified sentences, we constructed longitudinal, AD-specific patient profiles by mapping identified signs and symptoms from EHR notes chronologically to symptom categories and the patient's age, forming a time series of disease-relevant manifestations. Supplementary Note 3 details the construction process, and Supplementary Note 4 provides an example of an aggregated patient profile. This example illustrates how diverse AD-related symptoms—such as cognitive impairments and physiological changes—are captured and tracked across the patient's clinical history.

Multi-agent risk prediction across time points

We evaluated our methodology by predicting AD risk at seven distinct time points: 1 day, 1 year, 2 years, 3 years, 5 years, 7 years, and 10 years prior to the formal ICD-based diagnosis. A multidisciplinary team of specialist agents—including a primary care physician, neurologist, psychiatrist, geriatrician, clinical psychologist, and an AD specialist—collaboratively analyzed patients' AD profiles within a specific observation window generated by the data extraction agent. Detailed setup and prompts for the agents are pro-

Table 2 | Data extraction agent performance as a binary classifier

	Precision (Positive)	Recall (Positive)	F-1(Positive)	Overall accuracy
Fine-tuned LLaMa 3.1 8B instruct	0.74 (0.72, 0.76)	0.89 (0.88, 0.91)	0.81 (0.79, 0.83)	0.93 (0.91, 0.95)
Ensemble (Li et al. 2023)	0.71 (0.69, 0.73)	0.84 (0.82, 0.86)	0.77 (0.75, 0.79)	0.91 (0.89, 0.93)

Table 3 | Data extraction agent performance as a multi-class classifier

Symptom category	Precision	Recall	F1-score
Cognitive impairment	0.77 (0.75, 0.78)	0.82 (0.83, 0.84)	0.79 (0.77, 0.81)
Notice/Concern by others	0.84 (0.83, 0.84)	0.39 (0.37, 0.41)	0.53 (0.52, 0.55)
Requires assistance	0.69 (0.67, 0.70)	0.65 (0.63, 0.67)	0.67 (0.65, 0.68)
Physiological changes	0.74 (0.73, 0.75)	0.76 (0.74, 0.77)	0.75 (0.73, 0.77)
Neuropsychiatric symptoms	0.78 (0.77, 0.80)	0.83 (0.81, 0.85)	0.8 (0.78, 0.82)
Overall accuracy	Micro-average	0.75 (0.73, 0.77)	
	Macro-average	0.76 (0.74, 0.78)	

Table 4 | AD prediction performance of our proposed CARE-AD approach

Prediction Y ear	AD cases			Controls			Accuracy
	Precision	Recall	F1 score	Precision	Recall	F1 score	
−1 day	0.59 (0.57, 0.61)	0.73 (0.70, 0.76)	0.65 (0.63, 0.67)	0.92 (0.91, 0.93)	0.86 (0.85, 0.87)	0.89 (0.88, 0.90)	0.83 (0.82, 0.84)
−1 year	0.47 (0.45, 0.49)	0.66 (0.63, 0.69)	0.55 (0.53, 0.57)	0.89 (0.89, 0.90)	0.79 (0.78, 0.80)	0.84 (0.83, 0.85)	0.76 (0.75, 0.78)
−2 year	0.39 (0.37, 0.41)	0.59 (0.56, 0.62)	0.47 (0.45, 0.49)	0.87 (0.86, 0.88)	0.75 (0.73, 0.76)	0.80 (0.79, 0.81)	0.71 (0.70, 0.73)
−3 year	0.34 (0.32, 0.35)	0.54 (0.51, 0.57)	0.41 (0.39, 0.43)	0.85 (0.84, 0.87)	0.71 (0.69, 0.72)	0.77 (0.76, 0.79)	0.67 (0.66, 0.68)
−5 year	0.26 (0.25, 0.28)	0.46 (0.43, 0.49)	0.33 (0.31, 0.35)	0.81 (0.80, 0.82)	0.65 (0.64, 0.67)	0.72 (0.71, 0.73)	0.61 (0.59, 0.62)
−7 year	0.24 (0.23, 0.26)	0.43 (0.40, 0.46)	0.31 (0.29, 0.33)	0.80 (0.79, 0.81)	0.62 (0.61, 0.64)	0.70 (0.69, 0.71)	0.58 (0.57, 0.60)
−10 year	0.20 (0.18, 0.21)	0.38 (0.35, 0.41)	0.26 (0.24, 0.28)	0.77 (0.76, 0.78)	0.57 (0.55, 0.59)	0.65 (0.64, 0.67)	0.53 (0.51, 0.54)

vided in Supplementary Note 5. As shown in Table 4, our multi-agent system demonstrated consistent performance across all time points, with an accuracy of 0.83 at −1 day and 0.53 at −10 years. These results suggest the model's potential to identify both near-term and earlier indicators of AD risk based on longitudinal clinical narratives.

Comparison with single-model baselines

For comparison, we also evaluated four baseline methods, each using the same LLaMA 3 70B model: (1) a zero-shot approach with a single LLM call; (2) a Chain of Thought (CoT) approach[39] that guides language models to reason step by step by generating intermediate reasoning steps before producing a final answer; (3) a self-consistency approach[40] that generates multiple responses and selects the most consistent output through majority voting; and (4) a self-refine approach[41] that iteratively revises its outputs to improve clarity and correctness. As shown in Table 5, with an equal number of LLM calls (six), our CARE-AD method consistently outperformed these baselines, demonstrating the benefits of collaborative, domain-specialized reasoning.

Multi-agent conversation baseline

To further strengthen the comparison, we implemented a multi-agent conversational baseline using the AutoGen framework[42]. This setup mirrors the structure of CARE-AD, in which a supervisor agent (AD specialist) engages in multi-round dialogue with five domain-specific doctor agents. As shown in Table 5, the AutoGen-based configuration achieved comparable performance to CARE-AD when using 12 or more LLM calls, but required greater computational cost to match the performance of our more efficient prompt-based design.

Ablation study

In reviewing these ablation results for prediction at 10 years prior, CARE-AD (the full multi-agent configuration) achieved the highest overall accuracy (0.53). As shown in Table 6, the zero-shot baseline (no agent roles) performed poorly, particularly in identifying AD cases (F-score of 0.13), resulting in the lowest accuracy (0.26). When each specialty doctor agent was individually excluded, performance dropped below that of the full CARE-AD model, indicating that all agent roles contributed positively to classification. Notably, removing the neurologist role reduced accuracy to 0.50, suggesting that neurologist expertise is especially informative for distinguishing AD symptoms. Similarly, excluding the psychiatrist role lowered accuracy to 0.49, underscoring the importance of psychiatric insights in detecting mental health disorders associated with AD. Removing other roles (clinical psychologist, primary care physician, or geriatrician) also resulted in performance declines, though these decreases were comparatively smaller. Overall, the results in Table 6 highlight the value of incorporating multiple complementary clinical perspectives to improve AD vs. control classification accuracy.

Table 5 | Performance comparison of the proposed CARE-AD method with baseline models at -10-year prediction

Method	LLM calls	AD cases (P/R/F)	Controls (P/R/F)	Accuracy
Zero-shot	1	0.09 (0.08, 0.09)/0.25 (0.23, 0.28)/0.13 (0.11, 0.14)	0.56 (0.54, 0.57)/0.26 (0.25, 0.27)/0.35 (0.34, 0.37)	0.26 (0.25, 0.27)
Chain of thought (CoT)	1	0.11 (0.10, 0.12)/0.27 (0.25, 0.29)/0.15 (0.13, 0.17)	0.65 (0.63, 0.67)/0.37 (0.35, 0.39)/0.47 (0.45, 0.49)	0.35 (0.33, 0.37)
Self-consistency	6 reasoning paths	0.13 (0.11, 0.15)/0.29 (0.26, 0.32)/0.18 (0.16, 0.20)	0.70 (0.69, 0.71)/0.47 (0.45, 0.49)/0.56 (0.54, 0.58)	0.43 (0.42, 0.44)
Self-refine	6 refine rounds	0.16 (0.14, 0.18)/0.36 (0.33, 0.39)/0.22 (0.19, 0.25)	0.73 (0.72, 0.74)/0.47 (0.45, 0.49)/0.57 (0.55, 0.59)	0.45 (0.44, 0.46)
AutoGen multi-agent (1 round)	6 doctor agents (6 LLM calls)	0.16 (0.15, 0.17)/0.36 (0.33, 0.39)/0.22 (0.20, 0.24)	0.73 (0.72, 0.74)/0.48 (0.47, 0.50)/0.58 (0.57, 0.59)	0.45 (0.44, 0.47)
AutoGen multi-agent (2 rounds)	6 doctor agents (12 LLM calls)	0.20 (0.18, 0.21)/0.38 (0.35, 0.41)/0.26 (0.23, 0.28)	0.77 (0.76, 0.78)/0.58 (0.56, 0.59)/0.66 (0.65, 0.67)	0.53 (0.52, 0.55)
AutoGen multi-agent (3 rounds)	6 doctor agents (18 LLM calls)	0.20 (0.18, 0.21)/0.38 (0.35, 0.41)/0.26 (0.24, 0.28)	0.77 (0.76, 0.78)/0.58 (0.56, 0.59)/0.66 (0.64, 0.67)	0.53 (0.52, 0.55)
CARE-AD	6 doctor agents	0.20 (0.18, 0.21)/0.38 (0.35, 0.41)/0.26 (0.24, 0.28)	0.77 (0.76, 0.78)/0.57 (0.55, 0.59)/0.65 (0.64, 0.67)	0.53 (0.51, 0.54)

Table 6 | Ablation study showing the impact of agent roles on AD prediction at -10-year prediction

Agents	AD cases (P/R/F)	Controls (P/R/F)	Accuracy
CARE-AD	0.20 (0.18, 0.21)/0.38 (0.35, 0.41)/0.26 (0.24, 0.28)	0.77 (0.76, 0.78)/0.57 (0.55, 0.59)/0.65 (0.64, 0.67)	0.53 (0.51, 0.54)
Zero-shot (No agents)	0.09 (0.08, 0.09)/0.25 (0.23, 0.28)/0.13 (0.11, 0.14)	0.56 (0.54, 0.57)/0.26 (0.25, 0.27)/0.35 (0.34, 0.37)	0.26 (0.25, 0.27)
Exclude neurologist	0.17 (0.16, 0.18)/0.33 (0.30, 0.36)/0.23 (0.21, 0.24)	0.75 (0.74, 0.76)/0.55 (0.53, 0.57)/0.64 (0.62, 0.65)	0.50 (0.49, 0.52)
Exclude clinical psychologist	0.19 (0.17, 0.20)/0.36 (0.33, 0.39)/0.25 (0.23, 0.27)	0.76 (0.75, 0.77)/0.57 (0.55, 0.59)/0.65 (0.64, 0.67)	0.52 (0.51, 0.54)
Exclude psychiatrist	0.17 (0.15, 0.18)/0.34 (0.31, 0.37)/0.23 (0.21, 0.24)	0.75 (0.73, 0.75)/0.53 (0.51, 0.55)/0.62 (0.61, 0.63)	0.49 (0.47, 0.50)
Exclude primary care physician	0.18 (0.17, 0.20)/0.37 (0.34, 0.40)/0.25 (0.23, 0.27)	0.76 (0.75, 0.77)/0.55 (0.53, 0.57)/0.64 (0.63, 0.65)	0.51 (0.50, 0.53)
Exclude geriatrician	0.18 (0.16, 0.19)/0.34 (0.31, 0.37)/0.23 (0.21, 0.25)	0.75 (0.74, 0.76)/0.56 (0.54, 0.58)/0.64 (0.63, 0.66)	0.51 (0.50, 0.53)

Table 7 | Random forest prediction results using structured data features

Prediction year	AD cases			Controls			Accuracy
	Precision	Recall	F1 score	Precision	Recall	F1 score	
−1 day	0.45 (0.42, 0.48)	0.65 (0.62, 0.68)	0.53 (0.50, 0.56)	0.89 (0.87, 0.91)	0.78 (0.75, 0.81)	0.83 (0.81, 0.85)	0.75 (0.72, 0.78)
−1 year	0.37 (0.34, 0.40)	0.57 (0.54, 0.60)	0.45 (0.42, 0.48)	0.86 (0.84, 0.88)	0.73 (0.70, 0.76)	0.79 (0.76, 0.82)	0.70 (0.67, 0.73)
−2 year	0.31 (0.28, 0.34)	0.51 (0.48, 0.54)	0.39 (0.36, 0.42)	0.84 (0.82, 0.86)	0.69 (0.66, 0.72)	0.76 (0.73, 0.79)	0.65 (0.62, 0.68)
−3 year	0.20 (0.18, 0.22)	0.38 (0.35, 0.41)	0.26 (0.23, 0.29)	0.77 (0.74, 0.80)	0.58 (0.55, 0.61)	0.66 (0.63, 0.69)	0.53 (0.50, 0.56)
−5 year	0.18 (0.16, 0.20)	0.35 (0.32, 0.38)	0.24 (0.21, 0.27)	0.76 (0.73, 0.79)	0.55 (0.52, 0.58)	0.64 (0.61, 0.67)	0.51 (0.48, 0.54)
−7 year	0.14 (0.12, 0.16)	0.31 (0.28, 0.34)	0.20 (0.18, 0.22)	0.72 (0.69, 0.75)	0.49 (0.46, 0.52)	0.58 (0.55, 0.61)	0.45 (0.42, 0.48)
−10 year	0.11 (0.09, 0.13)	0.26 (0.23, 0.29)	0.16 (0.14, 0.18)	0.68 (0.65, 0.71)	0.44 (0.41, 0.47)	0.53 (0.50, 0.56)	0.40 (0.37, 0.43)

Structured data baseline

To establish a structured-data baseline, we implemented a random forest classifier trained on ICD codes, medications, and abnormal lab measurements, following prior work[12]. All features were processed using term frequency–inverse document frequency (TF-IDF) representations. The model was trained and tuned on a 90%/10% split of the full cohort after first holding out the 1,000 patients as an independent evaluation set. As shown in Table 7, our LLM-based CARE-AD framework consistently outperformed the structured-data model across all prediction horizons, achieving higher F1 scores for both AD cases and controls—particularly at earlier time points.

Discussion

In this study, we propose CARE-AD, a novel and feasible multi-agent LLM-based framework for early AD prediction using real-world longitudinal clinical notes. Building on advancements in multi-agent systems such as MEDAGENTS[35], our approach simulates a multidisciplinary diagnostic process where specialized agents analyze distinct aspects of AD-related signs and symptoms—cognitive impairment, physiological changes, neuropsychiatric symptoms, and other subtle indicators—extracted from clinical narratives. By dividing responsibilities across agents, the system identifies domain-specific markers that may be overlooked by a single general-purpose model. To our knowledge, this is one of the first applications of LLMs that not only extract AD-relevant indicators exclusively from unstructured clinical text but also employ a multi-agent workflow for early AD detection. Evaluations on retrospective clinical data suggest that CARE-AD offers improvements in predictive performance, helping bridge the gap between general-purpose LLM capabilities and the specialized requirements of AD-focused clinical applications.

CARE-AD outperformed single-model zero-shot approaches in our retrospective evaluation. With an accuracy of 0.53 at 10 years prior to ICD-based diagnosis, these findings suggest that relevant risk indicators may appear earlier than traditionally recognized, potentially offering a window for earlier clinical attention. While iterative single-model methods, such as self-consistency and self-refine, exceeded the zero-shot baseline, they still underperformed compared to the multi-agent strategy. The strength of CARE-AD lies in its distributed expertise and collaborative decision-making framework. Unlike self-refine and self-consistency methods, which constrain multiple reasoning paths within a single model, CARE-AD assigns distinct roles to specialized "doctor" agents, each leveraging domain-specific knowledge, and integrates their assessments through an AD specialist agent. This structure emulates real-world clinical collaboration and supports more comprehensive risk evaluation. For example, as detailed in Supplementary Note 6, the primary care physician agent identified comorbidities and the absence of cognitive screening; the neurologist agent emphasized past transient ischemic attacks and medication interactions; the geriatrician agent noted age-related vulnerabilities and polypharmacy; the psychiatrist agent highlighted how depressive symptoms could mask early cognitive decline; and the clinical psychologist agent recommended further mood and cognitive monitoring. The AD specialist agent then synthesized these insights and proposed that the patient may be experiencing cognitive decline consistent with early-stage AD, recommending confirmatory evaluations. By integrating complementary perspectives across clinical domains, CARE-AD offers an approach for evaluating early cognitive risk in a manner inspired by multidisciplinary consultation. The multi-agent design also enhances interpretability by revealing intermediate reasoning steps and showing how differing viewpoints are synthesized. While further prospective validation is needed, this approach offers a potential pathway for improving early detection, supporting longitudinal monitoring, and informing targeted interventions.

We also compared CARE-AD with an AutoGen-based multi-agent setup, which offers a general-purpose framework for inter-agent dialogue. When constrained to the same number of LLM calls (six), AutoGen underperformed relative to CARE-AD. This may be due to AutoGen's generalized architecture, which includes predefined system messages and automated coordination mechanisms that introduce additional reasoning steps or role negotiations that are less aligned with the streamlined requirements of clinical inference. In contrast, CARE-AD's role-specific prompting explicitly enforces task specialization, enabling more efficient extraction and synthesis of patient information. Increasing the number of LLM calls in AutoGen to 12 or 18 yielded comparable performance to CARE-AD, though improvements plateaued beyond 12 calls, indicating diminishing returns with further computation. These results suggest that CARE-AD offers a more resource-efficient alternative for early AD risk prediction.

In comparison with a traditional random forest model trained on structured EHR data, CARE-AD demonstrated the value of analyzing unstructured clinical narratives for identifying early AD risk indicators. Structured data, such as ICD codes, medications, and lab results, typically capture downstream diagnoses or late-stage manifestations, potentially missing earlier behavioral or cognitive changes. In contrast, narrative notes often contain subtle, pre-diagnostic observations that precede formal diagnosis by years. By leveraging this unstructured information, CARE-AD detected early symptom patterns more effectively than the structured-data model, particularly at longer prediction horizons. These findings reinforce the potential of LLMs in mining free-text EHR data for early disease signal detection.

This study has several important limitations. First, we relied on VHA data, which may not fully represent the broader population, as VHA patients often have distinct demographic characteristics, including a

significant sex imbalance, socioeconomic challenges, and higher rates of post-traumatic stress disorder and traumatic brain injury. Consequently, our findings require validation in non-VHA populations. Second, to ensure sufficient information for prediction, we required a minimum of 5 years of longitudinal notes in the observation window. This requirement may have introduced selection bias, as patients with lower hospital utilization and fewer clinical visits—those who could benefit most from large-scale screening—were underrepresented. In future work, we plan to include additional data sources to capture this group and improve our predictive models. Third, we defined the diagnosis date using the first recorded AD-related ICD code and included a −1 day prediction window, consistent with prior studies[12]. However, manual review revealed that in some cases, the actual diagnosis may have preceded the ICD code date, potentially inflating performance estimates, particularly at the −1 day window. Including a broader range of earlier time points like −1 year, −2 years, and −3 years before diagnosis, helps better assess the model's predictive performance across different stages of disease progression. Fourth, despite leveraging extensive baselines for comparison, privacy constraints prevented us from evaluating our approach using other cutting-edge LLMs (e.g., the GPT family[29]), limiting our ability to examine its generalizability to larger models. Nevertheless, our findings offer meaningful insights into how well the method adapts when data confidentiality is strictly enforced. In future work, we will explore publicly available datasets to more thoroughly assess how the model can scale and perform with other LLMs.

Expert-level performance in complex medical tasks like AD diagnosis will likely require collaborative, multi-agent systems. CARE-AD illustrates this approach's potential by leveraging coordinated specialized LLM agents to extract symptoms, assess risk, and predict AD onset up to 10 years before diagnosis, achieving higher accuracy than single-model baselines in our evaluation. The data extraction agent is designed to operate with longitudinal EHR data and could, with further validation, support symptom tracking and trend analysis for clinical decision-making. By incorporating domain-specific expertise, specialist agents enhance clinical decision-making, ensuring a more comprehensive and accurate assessment that may improve diagnostic interpretability. While this work focuses on AD, the underlying framework demonstrates the potential of multi-agent LLM solutions for addressing other complex medical conditions. It may be adaptable to other multifactorial diseases that require multidisciplinary expertise for diagnosis and management, providing a foundation for further exploration of AI-assisted clinical decision support.

Methods

Data sources and ethical approval

This study used the EHR database from the VHA Corporate Data Warehouse (CDW), covering the period from 2000 to 2022. The VHA is the largest integrated healthcare network in the U.S., comprising over 1200 medical centers and clinics, with extensive data on demographics, medications, diagnoses, procedures, clinical notes, and billing information, making it a valuable resource for large-scale health research. This study was approved by the Institutional Review Board of the US Veterans Affairs (VA) Bedford Health Care and conducted in accordance with the principles of the Declaration of Helsinki. A waiver of informed consent was obtained due to minimal risk to participants.

Cohort design

To construct the study cohort, we adopted a case-control design prioritizing diagnostic specificity to capture biologically homogeneous AD cases suitable for identifying early predictive markers. AD cases were defined based on the presence of AD-specific ICD codes (Supplementary Table 2) between October 1, 2015 (ICD-10 implementation), and September 30, 2022. We required at least two AD diagnoses on separate occasions, with one diagnosis recorded in a specialty clinic such as neurology, geriatrics, geriatric patient aligned care team (GeriPACT), mental health, psychology, psychiatry, or geriatric psychiatry—provided by a provider specializing in neurology, vascular neurology, psychiatry, neuropsychology, or geriatric medicine. These clinic types are identified by Stop Codes (Supplementary Table 3), which the VHA uses to specify the type of outpatient care and the workload associated with a visit[43]. These stringent criteria exclude patients with non-AD dementia, ensuring our cohort captures true AD trajectories essential for studying decade-long preclinical predictors.

Observation windows for each AD case began at the later of the patient's EHR initiation date or the study start date and ended at predetermined prediction time points prior to the first AD diagnosis (1 day, 1, 2, 3, 5, 7, and 10 years). A minimum observation period of 5 years was required, yielding 17,488 AD cases.

Controls were selected from VHA patients without any dementia diagnosis codes (Supplementary Table 4). Each AD case was matched with up to four controls based on age, sex, race/ethnicity, clinical utilization, Charlson Comorbidity Index (CCI), and Area Deprivation Index (ADI), following established methods[43]. The ADI was included to account for socioeconomic and environmental factors that shape health outcomes in AD, consistent with existing studies[44]. The final control cohort comprised 64,691 patients. Supplementary Fig. 1 details cohort inclusion and exclusion criteria.

AD diagnoses in this study reflect clinical practice, where diagnoses are based on cognitive and functional symptoms rather than biomarker confirmation. Thus, we use the terms "Alzheimer's disease (AD)" and "AD dementia" interchangeably.

Evaluation sampling

Because we employed LLMs for zero-shot evaluation and analyzing longitudinal notes is computationally intensive, we randomly selected 1000 AD cases and 3627 matched controls from the full cohort for evaluation. This subset approach aligns with conventional machine learning practices, where a portion of the data is reserved for testing, though no dedicated training set was needed in our zero-shot setting.

Multi-agent framework

Taxonomy development and annotation. AD dementia exhibits a complex continuum of cognitive, behavioral, and functional signs that evolve over many years[2]. Accurately interpreting these signs in large volumes of longitudinal EHR notes is challenging. By focusing solely on real-world clinical observations in EHRs—rather than specialized cognitive assessments or AD-specific diagnostic tests—this work identifies subtle early indicators, such as forgetfulness, behavioral shifts, and functional difficulties, that might otherwise go unnoticed. We intend to seek those insights to uncover overlooked aspects of patient histories and enhance predictive accuracy. Building on existing literature[38], domain experts crafted a novel, pragmatic taxonomy of five categories to capture the full spectrum of AD dementia signs and symptoms.

- Cognitive impairment: Captures the initial cognitive decline associated with AD, including subtle memory lapses, reduced problem-solving abilities, and difficulties in language comprehension, etc. These symptoms represent early indicators of neurodegeneration.
- Notice/of concern to others: Encompasses alterations in behavior and cognition that are noticeable and concerning to family members, close friends, neighbors, etc. Such changes signal deviations from the individual's typical functioning and may include increased confusion, disorientation, or withdrawal from social activities, etc.
- Requiring assistance/Functional impairment: Indicates a progressive loss of independence in daily activities. Patients begin to require assistance with tasks of instrumental activities of daily living (iADLs) such as managing finances, taking medications properly, or handling household chores. As their functional abilities decline further, they may also require support with activities of daily living (ADLs), for example, personal hygiene and other basic self-care tasks.
- Physiological changes: Includes physical symptoms indicative of AD progression, such as hearing/smelling loss, disrupted sleep patterns (e.g., insomnia or excessive sleepiness), inability to combine muscle movements, etc.

- Neuropsychiatric symptoms: Encompasses a range of psychiatric and behavioral manifestations seen in AD. While many of these symptoms —such as mood disturbances (depression, anxiety), psychotic features (hallucinations, delusions), agitation, and aggression—tend to become more pronounced in the later stages, certain issues like depression can emerge even before an AD diagnosis is formally made.

Detailed definitions for each category are provided in the expert-curated annotation guidelines in Supplementary Note 2. Detecting these signs and symptoms from EHRs is a crucial task for early diagnosis, treatment, and care planning of AD.

To create a gold-standard dataset, we applied our proposed taxonomy by systematically annotating 5112 longitudinal EHR notes from 76 individuals with AD (excluded from the evaluation set). Under two physicians' supervision, two trained medical professionals identified relevant sentences and assigned taxonomy-based labels. First, both annotators independently labeled all notes from six patients to assess inter-annotator reliability, achieving a high Cohen's κ (0.868) once disagreements were resolved through discussion. They then split the remaining patients between them for annotation, consulting their supervising physicians for any ambiguous cases. This process yielded a gold-standard dataset of 11,571 sentences, demonstrating the taxonomy's consistent applicability to clinical text.

Building on previously validated synthetic data resources shown to enhance model performance[38], we employed a subset of an existing synthetic dataset, selecting only the symptom categories relevant to AD. The synthetic data was originally generated using two established methods: (1) a data-to-label approach, in which sentences were randomly sampled from MIMIC-III discharge summaries and annotated by a LLM guided by clinical annotation guidelines; and (2) a label-to-data approach, where GPT-4 was prompted with predefined symptom category definitions to generate synthetic clinical note sentences paired with corresponding labels. These approaches enabled the creation of diverse and high-quality training samples without manual annotation. Statistics of the dataset used in this study are provided in Supplementary Table 1.

LLM fine-tuning for data extraction agent. Using both annotated and synthetic datasets, we fine-tuned the LLaMA 3.1 8B Instruct model with Low-Rank Adaptation (LoRA) to develop a specialized data extraction agent[45]. This LoRA strategy substantially decreases the number of trainable parameters, thereby improving efficiency and reducing costs—key factors in large-scale, aging-focused research. At inference, LoRA's lightweight parameter updates merge seamlessly with the base model to yield the final adapted system. We employed the Parameter-Efficient Fine-Tuning (PEFT) package[46] to complete the fine-tuning process using 8× NVIDIA A6000 (48 GB) GPUs over approximately 10 h. Parameter settings are provided in Supplementary Table 5.

Fine-tuning was performed for logit-based classification tasks. For binary classification, the input was a single sentence, and the output was a logit-based prediction indicating whether it was AD-relevant. We used a combination of annotated and synthetic AD-relevant sentences as positive samples, and randomly sampled non-AD-relevant sentences from the longitudinal notes of the same 76 patients to form negative samples, using a 5:1 negative-to-positive ratio[38]. For multi-label classification, we used only AD-relevant sentences, and the model produced a probability distribution over predefined AD symptom categories, with the predicted category selected via an argmax over logits.

We developed the CARE-AD framework using the LLaMA 3.1 8B and 70B models. The data extraction agent was fine-tuned using the LLaMA 3.1 8B model to balance performance and computational efficiency, enabling training on clinical data with manageable resource demands. The specialty doctor agents and the AD specialist agent were implemented using the LLaMA 3 70B model, selected for its strong zero-shot and in-context reasoning capabilities, scalability, and open-source availability—allowing secure deployment within the VINCI environment in compliance with VA data governance policies. Proprietary models such as GPT-4 were excluded due to VHA privacy restrictions prohibiting data transfer outside the VINCI system. While medical-domain LLMs (e.g., BioGPT[47], MedAlpaca[48], PMC-LLaMA[49], Clinical Camel[50]) may offer domain-specific advantages, they were not adopted due to limitations in scale, training data (mostly biomedical literature rather than real-world EHR notes), or deployment restrictions.

Patient time-series construction. To generate patient profiles suitable for temporal modeling of AD progression, we aligned each patient's clinical notes to their age at the time of each visit. We applied the fine-tuned LLaMA 3.1 8B model to classify sentences into one of the predefined AD symptom categories. The categorized sentences were then aggregated chronologically to create structured, time-stamped profiles capturing symptom evolution over time. These profiles enabled the specialist agents to assess patients' longitudinal trajectories rather than isolated encounters, facilitating temporally informed risk assessments.

Domain-specific and AD specialist agents. To emulate expert clinical reasoning without additional fine-tuning, we implemented structured, role-specific prompts within a multi-agent framework. Five domain-specific agents—a primary care physician, neurologist, geriatrician, psychiatrist, and clinical psychologist—were each guided by prompts reflecting their respective clinical expertise. These agents evaluated patient symptom profiles and provided domain-specific assessments. An AD specialist agent then integrated these evaluations with the extracted evidence to estimate the likelihood of AD development. Supplementary Table 6 outlines the agent configurations within the CARE-AD framework, and the full set of prompts is provided in Supplementary Note 5.

Baseline comparisons. To establish a conversational multi-agent baseline, we implemented the AutoGen framework using the LLaMA 3 70B model. The system comprised a supervisor agent (AD specialist) and five domain-specific agents—primary care physician, neurologist, psychiatrist, geriatrician, and clinical psychologist—each guided by structured prompts reflecting their clinical expertise. Agents engaged in multi-round dialogues to assess shared patient profiles, critique each other's reasoning, and iteratively refine their outputs under the supervision of the AD specialist. We evaluated the system's performance across varying numbers of dialogue rounds. The prompts used for the LLM-based baselines are provided in Supplementary Note 7, and Supplementary Table 7 summarizes all baseline model configurations and comparisons.

For the structured-data baseline, we implemented a random forest classifier using scikit-learn[51,52]. This model was chosen based on prior evidence that random forests outperform logistic regression for structured-data-based AD prediction[12]. We used structured EHR features--ICD diagnosis codes, medications, and abnormal lab measurements--processed with term frequency–inverse document frequency (TF-IDF) representations to enhance discriminative power. Additional implementation details are provided in Supplementary Note 8.

Data preprocessing. During the study period (2000–2022), we examined unstructured EHR notes from each patient's EHR initiation date or the study start date, whichever was later, up to their first ICD-coded AD diagnosis (the AD index date). To manage computational demands across this 20-year span, we first restricted analysis to notes from clinically relevant encounter types, including primary care, emergency visits, home-based primary care (HBPC), memory clinics, neurology, neuropsychology, geriatrics, psychiatry, psychology, cognitive care nursing, mental health clinics, compensation and pension examinations, and consultation visits.

To prepare unstructured text for sentence-level classification, we applied standard pre-processing steps. Sentence segmentation was performed using spaCy[53], which parsed clinical narratives into individual sentences. We then applied basic sentence filtering heuristics to remove low-information or noisy inputs, such as those with fewer than three tokens, numeric-only content, or more than 125 tokens. These pre-processing steps

ensured cleaner inputs and more consistent inference performance when using LLMs.

Evaluation and performance metrics. For evaluation, in cases where the AD specialist agent did not provide a definitive "Yes" or "No" response—typically recommending further clinical evaluation instead—we applied a consistent evaluation rule. Specifically, if the agent explicitly stated there was no AD risk or that symptoms were not related to AD, the case was classified as non-AD. All other responses, including expressions of uncertainty or deferrals for further testing, were classified as AD-positive, aligning with the study's goal of identifying early, pre-diagnostic risk indicators. This protocol reflects the clinical reality that early signs of AD often emerge before formal diagnostic confirmation.

To quantify model performance, we used stratified bootstrapping with 5000 iterations to estimate 95% confidence intervals (CIs) for CARE-AD and all baseline models. In each iteration, we resampled the test set with replacement while preserving the original AD/control class distribution and computed performance metrics. The 95% CI was calculated by taking the 2.5th and 97.5th percentiles of the resulting metric distribution.

Data availability

The data used in the preparation of this article are from VHA. Approval by the Department of Veterans Affairs is required for data access.

Code availability

Code for prediction models can be made available upon request. Relevant packages include PEFT, LLaMa 3.1 8B instruct[53,54], LLaMa 3 70B[55], Autogen, and Scikit-learn[52]. All prompts used in this work are in the supplementary.

Received: 25 February 2025; Accepted: 6 August 2025;
Published online: 24 August 2025

References

1. Mucke, L. Alzheimer's disease. *Nature* **461**, 895–897 (2009).
2. Alzheimer's Association, 2024 Alzheimer's disease facts and figures. *Alzheimers Dement.* **20**, 3708–3821 (2024).
3. Bateman, R. J. et al. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *N. Engl. J. Med.* **367**, 795–804 (2012).
4. Frisoni, G. B. et al. Strategic roadmap for an early diagnosis of Alzheimer's disease based on biomarkers. *Lancet Neurol.* **16**, 661–676 (2017).
5. Nam, E., Lee, Y.-B., Moon, C. & Chang, K.-A. Serum Tau proteins as potential biomarkers for the assessment of Alzheimer's disease progression. *Int. J. Mol. Sci.* **21**, 5007 (2020).
6. Rajan, K. B., Wilson, R. S., Weuve, J., Barnes, L. L. & Evans, D. A. Cognitive impairment 18 years before clinical diagnosis of Alzheimer's disease dementia. *Neurology* **85**, 898–904 (2015).
7. Riley, K. P., Snowdon, D. A., Desrosiers, M. F. & Markesbery, W. R. Early life linguistic ability, late life cognitive function, and neuropathology: findings from the Nun Study. *Neurobiol. Aging* **26**, 341–347 (2005).
8. Bature, F., Guinn, B., Pang, D. & Pappas, Y. Signs and symptoms preceding the diagnosis of Alzheimer's disease: a systematic scoping review of literature from 1937 to 2016. *BMJ Open* **7**, e015746 (2017).
9. Swaddiwudhipong, N. et al. Pre-diagnostic cognitive and functional impairment in multiple sporadic neurodegenerative diseases. *Alzheimers Dement*. **19**, 1752–1763 (2023).
10. van der Flier, W. M., de Vugt, M. E., Smets, E. M. A., Blom, M. & Teunissen, C. E. Towards a future where Alzheimer's disease pathology is stopped before the onset of dementia. *Nat. Aging* **3**, 494–505 (2023).
11. Wang, L. et al. Development and validation of a deep learning model for earlier detection of cognitive decline from clinical notes in electronic health records. *JAMA Netw. Open* **4**, e2135174 (2021).
12. Tang, A. S. et al. Leveraging electronic health records and knowledge networks for Alzheimer's disease prediction and sex-specific biological insights. *Nat. Aging* **4**, 379–395 (2024).
13. Mohammed, B. A. et al. Multi-method analysis of medical records and MRI images for early diagnosis of dementia and Alzheimer's disease based on deep learning and hybrid methods. *Electronics* **10**, 2860 (2021).
14. Tjandra, D., Migrino, R. Q., Giordani, B. & Wiens, J. Cohort discovery and risk stratification for Alzheimer's disease: an electronic health record-based approach. *Alzheimers Dement. Transl. Res. Clin. Interv.* **6**, e12035 (2020).
15. Li, Q. et al. Early prediction of Alzheimer's disease and related dementias using real-world electronic health records. *Alzheimers Dement.* **19**, 3506–3518 (2023).
16. Xu, J. et al. Data-driven discovery of probable Alzheimer's disease and related dementia subphenotypes using electronic health records. *Learn. Health Syst.* **4**, e10246 (2020).
17. Liu, Z. et al. AD-GPT: Large language models in Alzheimer's disease. Preprint at https://doi.org/10.48550/arXiv.2504.03071 (2025).
18. Almalki, H., Khadidos, A. O. & Alhebaishi, N. Enhancing Alzheimer's detection: leveraging ADNI data and large language models for high-accuracy diagnosis. *Int. J. Adv. Comput. Sci. Appl*. **15.11**, https://doi.org/10.14569/IJACSA.2024.01511134 (2024).
19. Zhang, M., Pan, Y., Cui, Q., Lü, Y. & Yu, W. Multimodal LLM for enhanced Alzheimer's disease diagnosis: Interpretable feature extraction from Mini-Mental State Examination data. *Exp. Gerontol.* **208**, 112812 (2025).
20. Du, X. et al. Enhancing early detection of cognitive decline in the elderly: a comparative study utilizing large language models in clinical notes. *eBioMedicine* **109**, 105401 (2024).
21. Tayefi, M. et al. Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Comput. Stat.* **13**, e1549 (2021).
22. Halpern, R. et al. Using electronic health records to estimate the prevalence of agitation in Alzheimer's disease/dementia. *Int. J. Geriatr. Psychiatry* **34**, 420–431 (2019).
23. Shao, Y. et al. Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. *BMC Med. Inform. Decis. Mak.* **19**, 128 (2019).
24. Hane, C. A., Nori, V. S., Crown, W. H., Sanghavi, D. M. & Bleicher, P. Predicting onset of dementia using clinical notes and machine learning: case-control study. *JMIR Med. Inform.* **8**, e17819 (2020).
25. Gilmore-Bykovskyi, A. L. et al. Unstructured clinical documentation reflecting cognitive and behavioral dysfunction: toward an EHR-based phenotype for cognitive impairment. *J. Am. Med. Inform. Assoc.* **25**, 1206–1212 (2018).
26. Noori, A. et al. Development and evaluation of a natural language processing annotation tool to facilitate phenotyping of cognitive status in electronic health records: diagnostic study. *J. Med. Internet Res.* **24**, e40384 (2022).
27. Prakash, R., Dupre, M. E., Østbye, T. & Xu, H. Extracting critical information from unstructured clinicians' notes data to identify dementia severity using a rule-based approach: feasibility study. *JMIR Aging* **7**, e57926 (2024).
28. DeepSeek-AI et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. Preprint at https://doi.org/10.48550/arXiv.2501.12948 (2025).
29. OpenAI et al. GPT-4 technical report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2024).
30. Dubey, A. et al. The Llama 3 herd of models. Preprint at http://arxiv.org/abs/2407.21783 (2024).
31. Karabacak, M. & Margetis, K. Embracing large language models for medical applications: opportunities and challenges. *Cureus* **15**, e39305 (2023).

32. Özge, A. et al. One patient, three providers: a multidisciplinary approach to managing common neuropsychiatric cases. *J. Clin. Med.* **12**, 5754 (2023).
33. Galvin, J. E. et al. Early stages of Alzheimer's disease: evolving the care team for optimal patient management. *Front. Neurol.* **11**, 592302 (2021).
34. Galvin, J. E., Valois, L. & Zweig, Y. Collaborative transdisciplinary team approach for dementia care. *Neurodegener. Dis. Manag.* **4**, 455–469 (2014).
35. Tang, X. et al. MedAgents: large language models as collaborators for zero-shot medical reasoning. In Findings of the Association for Computational Linguistics: ACL 2024, pages 599–621, Bangkok, Thailand. Association for Computational Linguistics.
36. Ke, Y. et al. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: simulation study. *J. Med. Internet Res.* **26**, e59439 (2024).
37. Cai, W. et al. A survey on mixture of experts in large language models. *IEEE Trans. Knowl. Data Eng.* **37**, 3896–3915 (2025).
38. Li, R., Wang, X. & Yu, H. Two directions for clinical data generation with large language models: data-to-label and label-to-data. in *Findings of the Association for Computational Linguistics: EMNLP 2023* (eds Bouamor, H., Pino, J. & Bali, K) 7129–7143 (Association for Computational Linguistics, 2023).
39. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models[J]. *Adv. Neural Inf. Process. Syst.* **35**, 24824–24837 (2022).
40. Wang, X. et al. Self-consistency improves chain of thought reasoning in language models. In The Eleventh International Conference on Learning Representations (2023).
41. Madaan, A. et al. SELF-REFINE: iterative refinement with self-feedback. *Adv. Neural Inf. Process. Syst* **37**, 46534–46594 (2023).
42. Wu, Q. et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation. In *COLM 2024* (2024).
43. U.S. Department of Veterans Affairs. ECX-3 user guide. https://www.va.gov/vdl/documents/Financial_Admin/Decision_Supp_Sys_(DSS)/ecx_3_ug.pdf (2024).
44. Vassilaki, M., Petersen, R. C. & Vemuri, P. Area deprivation index as a surrogate of resilience in aging and dementia. *Front. Psychol.* **13**, 930415 (2022).
45. Hayou, S., Ghosh, N. & Yu, B. LoRA+: efficient low rank adaptation of large models. *Proc. Int. Conf. Mach. Learn.* **41**, 17783–17806 (2024).
46. Parameter-Efficient Fine-Tuning using PEFT. https://huggingface.co/blog/peft (2025).
47. Luo, R. et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.* **23**, bbac409 (2022).
48. Han, T. et al. MedAlpaca – an open-source collection of medical conversational AI models and training data. Preprint at https://doi.org/10.48550/arXiv.2304.08247 (2025).
49. Wu, C. et al. PMC-LLaMA: towards building open-source language models for medicine. *J. Am. Med. Inform. Assoc.* **31**, 1833–1843 (2024).
50. Toma, A. et al. Clinical camel: an open expert-level medical language model with dialogue-based knowledge encoding. Preprint at https://doi.org/10.48550/arXiv.2305.12031 (2023).
51. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
52. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
53. spaCy · Industrial-strength Natural Language Processing in Python. https://spacy.io/.
54. meta-llama/Llama-3.1-8B-Instruct · Hugging Face. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct (2024).
55. meta-llama/Meta-Llama-3-70B · Hugging Face. https://huggingface.co/meta-llama/Meta-Llama-3-70B (2024).

Acknowledgements

This study was funded by the National Institute on Aging of the National Institutes of Health (NIH) under award number R01AG080670. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author contributions

R.L. and H.Y. conceived and designed the study. R.L. conducted the experiments, performed data analyses, and drafted the manuscript. X.W. contributed to baseline experiments and the AutoGen analysis, provided critical feedback, and assisted with manuscript revisions. H.Y., D.B., J.M., and H.L. offered critical feedback, helpful suggestions, and contributed to editing the manuscript. H.Y. provided overall research supervision. All authors contributed to manuscript editing, agreed with the results and conclusions, and approved the final draft. Authors had access to the study data and take responsibility for the submission.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01940-4.

Correspondence and requests for materials should be addressed to Hong Yu.

Journal of Biomedical Informatics 118 (2021) 103803

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Predicting Alzheimer's Disease from Spoken and Written Language Using Fusion-Based Stacked Generalization

Ahmed H. Alkenani [a,b], Yuefeng Li [a,*], Yue Xu [a], Qing Zhang [b]

[a] School of Computer Science, Queensland University of Technology, Brisbane 4001, Australia
[b] The Australian e-Health Research Centre, CSIRO, Brisbane 4029, Australia

ARTICLE INFO

Keywords:
Machine learning
Feature selection
Information fusion
Ensemble classifier
Cognitive decline
Alzheimer's disease
Clinical diagnosis
Neurolinguistics

ABSTRACT

The importance of automating the diagnosis of Alzheimer disease (AD) towards facilitating its early prediction has long been emphasized, hampered in part by lack of empirical support. Given the evident association of AD with age and the increasing aging population owing to the general well-being of individuals, there have been unprecedented estimated economic complications. Consequently, many recent studies have attempted to employ the language deficiency caused by cognitive decline in automating the diagnostic task via training machine learning (ML) algorithms with linguistic patterns and deficits. In this study, we aim to develop multiple heterogeneous stacked fusion models that harness the advantages of several base learning algorithms to improve the overall generalizability and robustness of AD diagnostic ML models, where we parallelly utilized two different written and spoken-based datasets to train our stacked fusion models. Further, we examined the effect of linking these two datasets to develop a hybrid stacked fusion model that can predict AD from written and spoken languages. Our feature spaces involved two widely used linguistic patterns: lexicosyntactics and character *n*-gram spaces. We firstly investigated lexicosyntactics of AD alongside healthy controls (HC), where we explored a few new lexicosyntactic features, then optimized the lexicosyntactic feature space by proposing a correlation feature selection technique that eliminates features based on their feature-feature inter-correlations and feature-target correlations according to a certain threshold. Our stacked fusion models establish benchmarks on both datasets with AUC of 98.1% and 99.47% for the spoken and written-based datasets, respectively, and corresponding accuracy and F1 score values around 95% on spoken-based dataset and around 97% on the written-based dataset. Likewise, the hybrid stacked fusion model on linked data presents an optimal performance with 99.2% AUC as well as accuracy and F1 score falling around 97%. In view of the achieved performance and enhanced generalizability of such fusion models over single classifiers, this study suggests replacing the initial traditional screening test with such models that can be embedded into an online format for a fully automated remote diagnosis.

1. Introduction and motivation

The recent digitalization waves have resulted in increased initiatives attempting to replace the traditional manual processes in various domains. One of the most attractive domains to many researchers is the automation of medical knowledge and decision support systems. Specifically, medical diagnosis is an appealing and active research area; yet considered an extremely complicated task given the involved processes and vital corresponding procedures and treatments [1]. The assessment and diagnosis of neurodegenerative syndromes, including Alzheimer disease (AD), remain primarily clinical and psychometric in spite of the exceptional advances in information and communication technologies [2]. Due to its established association with the age and to the fact that it currently has no cure, AD has been linked to unprecedented estimated economic complications in the near future aligning with the expected growth of AD cases given the increased quality of life [3,4]. The early diagnosis of AD is, therefore, crucial as it enables a greater efficiency of pharmaceutical treatments that may mitigate the AD side effects [5–7]. Further, early stages of cognitive decline could be stabilized or curtailed in some cases [8–11]. Accordingly, there has been a call by concerned associations for replacing the initial traditional pen-and-paper screening tests with effective automated diagnostic procedures [2,12].

* Corresponding author.
E-mail address: y2.li@qut.edu.au (Y. Li).

https://doi.org/10.1016/j.jbi.2021.103803
Received 9 December 2020; Received in revised form 6 April 2021; Accepted 3 May 2021
Available online 19 May 2021

In this context, Machine learning (ML) and natural language processing (NLP) could provide insights into linguistic mechanisms of AD patients, which have been underlined to be disrupted by early AD [13], leading to patterns and deficits that can be used for AD diagnosis. Subsequently, various language and speech analysis studies have recently investigated such patterns and deficits and asserted their interrelation with and contribution to early diagnosis of AD [14–23]. Among these patterns and deficits, lexicosyntactic (i.e., lexical and syntactic) processing has been recommended for further investigations owing to its ascertained deterioration in people with cognitive decline in general and in people with AD, specifically [24,25]. On that basis, this study aims to develop robust diagnostic models for predicting AD, employing modern ML algorithms trained with patterns and deficits extracted from spoken and written-based language samples. We particularly aim to develop multiple heterogeneous ensemble models for predicting AD based on ensemble methods that combine the advantages of several single algorithms towards more robust predictive models.

Ensemble methods have emerged to overcome challenges associated with single classifiers such as limited performance and increased prediction errors by exploiting several individual classifiers and combining their predictions [26]. There are generally two types of ensemble methods: homogeneous and heterogeneous ensembles [27]. Homogeneous ensemble consists of a single-type base learning algorithm, achieved by partitioning the training dataset and parallelly using the same base classifier on these partitions, with the most common examples being bootstrap aggregation and boosting. While the earlier is constructed by the usage of different samples for building distinct trees for the same learning algorithm then generating an aggregate prediction [28], boosting refers to the usage of different weights to sequentially train learning algorithms that are generally considered weak [29]. A heterogeneous ensemble, on the other side, consists of members with different base learning algorithm and is achieved by employing such diverse types of base classifiers, where different strategies can be used to combine these features such as voting and stacked generalization. Voting ensemble methods mainly average the predictions of sub-models to construct a stronger model [30], and contrarily stacked generalization uses the predictions of a pool of base classifiers to train another classifier based on these predictions [31]. In this study, a heterogeneous ensemble method is built and evaluated using several diverse classifiers as the base learning algorithms. We select a heterogeneous ensemble method over the homogeneous method due to its diversity reflected from the different nature of base classifiers, which increases the generalizability and reduces the uncertainty of the predictive model. Besides, its performance was seen to outperform that of homogeneous ensemble methods [32]. Despite the recent growing interest in using ML algorithms and NLP for predicting AD [4], the usage of ensemble methods in this context remains unexplored to date. Therefore, a novelty of this study lies in developing heterogeneous ensemble models that embed different classifiers via a *meta*-classifier for the prediction of AD.

Moreover, it is worth noting that the diagnosis of AD from written language has received less attention compared to that from spoken language even though writing could reveal the early signs of AD through evident lower grammatical complexity as well as lower idea density [15,33,34]. Especially, the use of written language samples in the automatic identification of AD by training ML algorithms has rarely been attempted [35]. Consequently, there has been a recent attempt to predict the disease from routine blogs written by AD patients, where Masrani et al. [35] crawled and used several blogs belonging to AD patients as well as healthy controls (HC) to train ML models. This initial step in using written language for ML-based identification of AD was followed by a further recent study, where Kong et al. [36] used the same blogs to train neural networks models. However, both studies have not explored ensemble methods to improve the results. They also have not addressed the effect of linked spoken and written language in the automatic identification of AD. As such, we investigate the potential of developing a hybrid heterogeneous ensemble model that can predict AD from spoken and written language by recrawling these blogs and use it alongside a spoken-based language datasets; the Cookie Theft Picture Corpus (CTPC) from DementiaBank[1] dataset, which is the largest publicly available spoken-based dataset for such diagnostic task [37] and the major focus of most concerned researchers. Our recent work presents a comprehensive review of the work conducted using DementiaBank dataset [4].

For efficient learning of the base as well as heterogeneous ensemble classifiers, two distinct feature spaces are extracted from these datasets. We firstly investigate a few new lexicosyntactic features and examine them along with other lexicosyntactics, where we propose a correlation-based feature selection technique to optimize the entire lexicosyntactic feature space. Afterwards, we generate character-based vocabulary spaces to represent the stylistic properties of the language samples, motivated by their effectiveness in related tasks [38,39]. A fusion of the optimized feature spaces forms the optimal feature set for both the base as well as heterogeneous ensemble classifiers, leading to benchmark results on both datasets achieved with our stacked fusion models.

The rest of the paper is structured as follow. We firstly discuss the proposed stacked fusion model is presented in section 2 alongside the feature spaces and optimization techniques. Afterwards, the experimental results and discussion are presented in section 3. Section 4 presents a comparative analysis of our stacked fusion models against the state-of-the-art models on both datasets. Finally, we address the potential drawbacks of our models before concluding this work.

2. Methods and materials

2.1. Datasets

2.1.1. DementiaBank

DementiaBank dataset is being the main openly available dataset for evaluating spoken language of patients with AD at present, collected via a longitudinal study from 1983 to 1988 with 45 to 90 years old English-speaking participants [37]. The Cookie Theft Picture description task, from the Boston Diagnostic Aphasia Examination “BDAE”, was used as a part of the study to elicit language samples from the participants by requesting each participant to describe the event happing in the picture and audio-recording the provided description. Afterward, a manual transcription was applied to these recordings by the dataset custodian using CHAT transcription protocol [40].This protocol is a part of the TalkBank[2] project and was created via computational tools established to accelerate the automatic transcription of audio recordings especially for research purposes. All participants had received extensive neuropsychological screening prior to performing the task. The corpus of the Cookie Theft Picture description Corpus (CTPC) comprises 548 samples in total; out of which, 243 samples belong to 98 HC, 236 samples retrieved from 189 CE patients, 43 samples deduced by 19 patients with Mild cognitive Impairment, 21 samples belong to patients with possible AD, and 5 samples belong to Vascular dementia patients.

Since our study is concerned with the diagnosis of AD, we retained the language samples of AD alongside that of HC and discarded all the remaining samples. The number of samples belonging to HC group was down-sampled to equalize that of the AD group based on demographic attributes of participants such as age and level of education, following Orimaye et al. [41].The dataset preprocessing was initiated by extracting the word-level sentences then removing the annotations involved in the CHAT transcriptions protocol. Besides, we discarded the participants’ demographic as our approach is exclusive to linguistic patterns.

2.1.2. Alzheimer’s disease Blog corpus (ADBC)

In contrary to DementiaBank dataset, the Alzheimer’s Disease Blog

[1] https://dementia.talkbank.org/
[2] https://talkbank.org/

Table 1
Blog corpus information.

URL (http://*.blogspot.ca)	Number of Posts		Diagnosis
	Masrani et al.	*Ours* ADBC	
living-with-alzhiemers	344	656	AD
creatingmemories	618	706	AD
parkblog-silverfox	692	-	Lewy body
journeywithdementia	201	274	HC
Earlyonset	452	609	HC
helpparentsagewell	498	509	HC
Total posts	2805	AD 1362	-
		HC 1392	
		2754	

Table 2
Number of samples from CTPC and ADBC.

Dataset	Number of samples		D_{train}		D_{test}	
	AD	HC	AD	HC	AD	HC
CTPC	236	236	189	189	47	47
ADBC	1362	1392	1090	1114	272	278

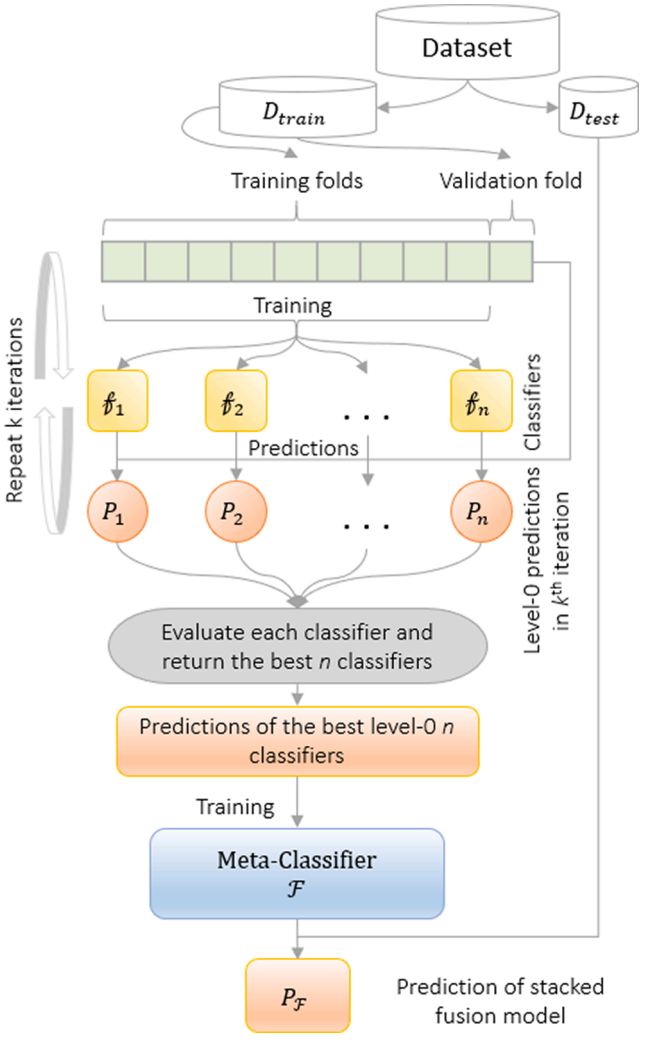


Fig. 1. The proposed stacked fusion model framework.

Corpus (ADBC) was originated from written language of people diagnosed with dementia as well as healthy caregivers (HC) to demented people. Masrani et al. [35] initiated this corpus by scraping a total of 2805 posts from six public blogs; two written by AD patients, one belongs to a patient diagnosed with Lewy body, and three written by HC. Aligning with the focus of our study, our ADBC croups is inherited only from blogs belonging to AD and HC groups, where we recrawled these blogs as to enlarge the corpus volume, as some of these bloggers update their blogs regularly. Table 1 illustrates this corpus blogs alongside that of Masrani et al.

2.1.3. Datasets notation

The dataset D utilized in our study represents a set of Npairs of language samples and corresponding target class as $D = (x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ in which, $x_i \in X$ and $y_i \in Y$. Let $X = \{x_1, x_2, ..., x_n\}$ represents an input space of the language samples produced by $Y = \{y_1, y_2, ..., y_n\}$ which, the later, denotes the HC and AD target classes. Wherein, $x_i = (x_{i_1}, x_{i_2}, ..., x_{i_m})$for $\forall i \in (1, 2, \cdots, n)$ is the input vector defining the m-dimensional feature values of the i^{th} language sample in X, and $y_i \in \{0, 1\} \forall i \in (1, 2, \cdots, n)$ is the corresponding target class for the n^{th} language sample. In such supervised classification task, a classifier is a mapping function $\not{f}$ trained to label the input vector to a target class as $\not{f} : X \rightarrow Y$ so that given unseen language samples, it is able to predict the corresponding target class as illustrated in Eq. (1).

$$\not{f}(x) = \hat{y} = \begin{cases} 1, if x \in AD \\ 0, if x \in HC \end{cases} \quad (1)$$

Each of our datasets D was shuffled and then randomly split into D_{train} as the training set and D_{test} as the test set, applying the common 80/20 split to train and test the proposed stacked fusion model in this work, detailed in Table 2, with examples illustrated in table (15) in the appendix.

2.1.4. Proposed methodology

In order to optimize the prediction of AD, we propose a stacked fusion model optimized via two different filter-based feature reduction techniques. Our experimental design of the proposed model adopts 10-fold cross validation (CV) approach to evaluate the performance of the model in classifying people with AD from HC using spoken and written languages. The framework of the stacked fusion model is illustrated in Fig. 1, which comprises four phases detailed in below sections and illustrated as follow:

- Datasets preprocessing and feature extraction: In this phase, raw data from original datasets is preprocessed and transformed into a suitable format for analysis and machine learning algorithms. This phase also involves extracting lexicosyntactic features and character-level vocabulary spaces.
- Dimensionality reduction: This phase presents the optimization techniques of our feature spaces.
- Classifiers training: The development of the proposed stacked fusion model involves two-level learning process; level-0 for training the single individual classifiers as base learners using K-fold cross validation approach on the training set D_{train} and level-1for training the stacked fusion models on top of base learners, using the output of level-0 as the training set.
- Alzheimer's disease prediction: During this phase, the proposed stacked fusion models are tested using a held-out test set D_{test}. This model is based on stacked generalization, which is an ensemble technique that aims to reduce the generalization error by using the predictions of a pool of base classifiers to train another classifier based on these predictions [31].

2.1.5. Feature engineering

Our feature space in this study incorporates two widely used patterns of connected speech: lexicosyntactic features and n-gram vocabulary spaces. Lexicosyntactic features are established as promising patterns in

the AD diagnostic task [4,42] and have been recommended for further investigations given its association with early cognitive decline [24,25]. On the other side, *n*-gram vocabulary spaces have attracted several NLP researchers concerned with dementia diagnosis [41–44]. This two-level feature space is detailed as follows:

Lexicosyntactic: Our lexicosyntactic feature space involves several features, amongst them are a few features investigated by other researchers, consisting of type-token ratio (TTR) [45], character, word, and sentence counts [42], content density and idea density [17], functional words, noun to verb and verb to noun indexes, active and passive proposition densities as well as open-class and closed-class ratios [4]. In Addition, this space was extended with other lexicosyntactic features that have rarely been applied to AD diagnostic task including sentence count, average lengths of word and sentences, type-token count (TTC). We also investigated new measures, namely lengths of open and closed classes, averaged open and closed classes, and averaged idea to content density, motivated by previous findings asserting the increased usage of propositions and verbs compared to nouns as AD progresses [15,17,46,47]. The lexicosyntactic feature space is described as follows:

- Character count: Measures the absolute character count.
- Word count: Measures the absolute number of words, taking repeated words into account.
- Sentence count: Measures the absolute sentences count.
- Average word length: Measures the total word length (i.e., character count) to the word count.
- Average sentence length: Measures the total sentence length (i.e., word count) to the sentence count.
- Stopwords count: Also named functional words, which represent a collection of words that appear in most documents and typically considered noisy in NLP tasks. Nevertheless, they could aid the AD diagnostic task given the deteriorated language production in AD patients [4].
- Stopwords ratio: Measures the total count of functional words (e.g., the, is, and are) to the word count.
- Type-token count: Type-token count (TTC) is the absolute number of unique POS tagged tokens (e.g., noun count, verb count, and adjective count).
- Type-token ratio: Type-token ratio (TTR) is the absolute number of unique POS tagged tokens to the total word count (e.g., noun ratio, pronoun ratio, and adverb ratio).
- Open class and closed class counts: At a higher level of word classes, POS tags are grouped into two main classes; open class and closed class. Open class denotes an infinite total of new words that can be added or created including nouns, verbs, adjectives, adverbs, and interjections. Closed class, on the other hand, is a small fixed set that includes pronouns, conjunctions, prepositions, auxiliary verbs, adpositions , determiners, particles, and modals [48,49]. We measure the total counts of open class and closed class, respectively.
- Open class and closed class ratios: In contrary to open class and closed class counts, open class ratio measures the total of open class words to the absolute word count as open class ratio and, likewise, closed class ratio is the total of closed class words to the absolute word count as closed class ratio.
- Idea density: Idea density is a measure of language complexity that calculates the total expressed propositions to the total word count, where verbs and all their arguments (i.e., adjectives, adverbs, conjunctions, and prepositions) constitute one proposition [17].
- Content density: Content density is another language complexity measure that calculates the open class ratio to the closed class ratio [17].
- Noun to verb index: Measures the noun ratio to the verb ratio, by calculating the total number of nouns to the total word count then dividing the result by the total number of verbs to the word count.
- Verb to noun index: Measures the verb ratio to the noun ratio, by calculating the total number of verbs to the total word count then dividing the result by the total number of nouns to the word count.
- Active proposition density: Measures the total ratios of verbs, adverbs, and adjective to the noun ratio.
- Passive proposition density: Measures the noun ratio to the total ratios of verbs, adverbs, adjective.
- Word class count: Measures the absolute count of open and closed classes, by calculating the total number of words that belong to either open or closed classes.
- Word class ratio: Measures the total ratios of open class and closed class to the word count.
- Mean content density: Measures the ratio of propositions, where verbs and all their arguments (i.e., adjectives, adverbs, conjunctions, and prepositions) constitute one proposition [17], to the content density (i.e., the open class ratio to the closed class ratio).

The POS annotation of the utilized datasets was performed using NLTK[3] POS tagger, which was trained with maximum entropy on the corpus of Penn Treebank [50]. NLTK POS tagger uncovered 30 tags from the CTPC and 36 tags from the ADBC.

2.1.6. *N-gram Character spaces:*

N-gram spaces refer to an adjacent series of n tokens typically extracted from a written or spoken language sample, which could be characters, words, or phonemes. Character n-grams are effective and recognized better than content words, especially on data inherited from blogs and connected language corpora [51]. In a series of characters tokens $T = (t_1, t_2, \cdots, t_{N+(n-1)})$, an n-gram is any n-length sequential tokens, where the i^{th} n-gram of T is $(t_i, t_{i+1}, \cdots, t_{i+n-1})$. For instance, a series of *1*-grams "i.e., unigrams" extracted from the sentence "AD_is_a_progressive_disease" would be: A, D, _, i, s, _, a, _, p, r, o, g,…; a *2*-grams "i.e., bigrams" series from the same sentence would be: AD, _i, s_, a_, pr, og, re, ss,…; *3*-grams "i.e., trigrams" are: AD_, is_, a_p, rog, res, siv,…; and so forth. Owing to their optimal performance in a similar task [51,52], our n-gram character spaces consists of low-level bigrams and trigrams extracted after the removal of stopwords.

3. Feature scaling

Feature scaling transforms the feature space into a unified scale with a mean $\widehat{x} = 0$ and $\sigma = 1$. It is considered an important step especially when fitting learners that rely on gradient descent (e.g., MLP) which lead to a faster converge. As such, we rescaled the lexicosyntactic feature space extracted from the utilized datasets using RobustScaler due to its robusness to outliers, where it eliminates the bottom and top quartiles. Eq. (2) represents this normalization technique, where x' is the scaled feature value, Q_1 is the 1st quarile which equals to 25%, and Q_3 is the 3rd quartile that equals to 25%.

$$x' = \frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)} \tag{2}$$

This scaler has been recommnded for normalizing such feature space where detecting outliers entials defining an acceptable range of each feature by a medical espcailist and thus sometimes not applicable [53].

3.1. Feature selection

Reducing the feature space without sacrificing the performance is generally preferred. Specifically, with textual feature space that is typically prone to the high dimensionality issue given the numerical representation of texts which not only leads to an increased

[3] http://www.nltk.org/

computational complexity, but also has a direct negative effect on the performance of ML algorithms [54]. Alleviating this issue is commonly achieved via reduction of feature space using either feature extraction or feature selection. While the earlier reduces the space by generating a smaller new feature space out of the original one, feature selection aims at selecting an optimal subset of the feature space that optimize the learning algorithm. Due to the difficulty of interpreting the new feature space resulted from feature extraction given the loss of original space's meaning [55], feature selection would be more appropriate to such diagnostic tasks. Feature selection methods can be grouped into filter, wrapper, and embedded categories, where filter methods are more general and have attracted many NLP researchers in consequence of their relatively-low computational complexity and nature of being independent of any learning algorithms [56]. Accordingly, we selected filter methods for revealing the most informative space. Filter methods have been effective in optimizing the learning robustness of learning algorithms with small feature spaces in related studies [42]. The feature selection process is explained as follows:

3.2. Correlation based feature selection (CFS)

Correlation-based feature selection (CFS) is a well-known filtering technique for selecting the optimal subset of features by calculating the feature-class correlation and feature-feature inter-correlation [57]. While a feature is generally considered relevant if it is correlated with the target class, it is also considered redundant if it highly correlates with one or more feature in the feature space [58]. Eliminating such irrelevant and redundant features is crucial to enhancing the performance of learning algorithms. Subsequently, we present a correlation-based feature selection method that takes these two main principals into account towards selecting the optimal subset of our lexicosyntactic feature space. In the proposed method, we fuse two well-established correlation evaluation algorithms, namely the Pearson's correlation [59] and mutual information [60], to reveal the importance of the lexicosyntactic features with an overall aim to select a reduced feature space that contains only highly predictive features of the target class and, on the other hand, within a certain predetermined correlation threshold.

The Pearson's correlation is a widely used feature-feature and feature-class relationship measure that indicates the direction and strength of the linear relationship between two variables [61]. Its absolute value ranges between −1 to1, where a value less than zero represents a negative relationship and, contrarily, a value greater than zero indicates a positive relationship [62]. The mutual information, on the other side, is an efficient information theory-based measurement for feature-class inter-dependency [63]. It is considered the most comprehensive measure of the total dependence between two different variables given its virtue of measuring both linear and nonlinear dependencies, which can handle complex relationships [64]. Unlike Pearson's correlation, the value of mutual information is more open-ended, ranging from zero for total independence to infinity for complete correlation.

Table 3
A contingency table of feature t and class c.

	Class c	Different c	Total c
Term t	a	b	a + b
Different term t	c	d	c + d
Total t	a + c	b + d	Grand Total N

In our proposed method, we firstly assign pairwise inter-correlation coefficients to the feature space using Pearson's correlation (r) as given in Eq. (3) where,$r(x,y\ )$ represents the correlation of the linear relationship between two features × and y, x_i and y_i are the i^{th} observations in the variables x and y whereas $\overline{x}$ and $\overline{y}$ are the means of variables x and y, respectively, and N is the total number of observations. S_x and S_y are the standard deviations for variables x and y, respectively.

$$r(x,y\ ) = \sum \frac{(x_i - \overline{x})(y_i - \overline{y})}{(N-1)(s_x)(s_y)} \tag{3}$$

Afterwards, mutual information (I) is used for measuring the inter-dependency of a feature and the target class as demonstrated in Eq. (4) in which, $P(c\ ,f\ )$ signifies the joint probability function of C and F, $P(f\ )$ and $P(c\ )$ are the probability density functions of C and F, respectively.

$$I(F,C) = \sum_{f\in F}\sum_{c\in C} P(c,f) log \frac{P(f,c)}{P(f)P(c)} \tag{4}$$

The fusion of these two measures forms an intelligent block that captures the most predictive features while maintain a given feature-feature inter-correlation threshold. Considering the diagnostic task in this work, we follow the guideline to appropriate usage of correlation in medical research presented by a previous work [59], which suggests that a correlation coefficient below (-)0.30 is considered neglected relation and, contrarily, over (-)0.70 is a high correlation. It also concluded that a

Algorithm 1. Correlation Based Feature Selection

Input: $F \leftarrow$ lexicosyntactic features, $F = \{f_1, \cdots f_n\}$ in D
$C \leftarrow$ the set of Classes, $C = c_1, c_2$
$L \leftarrow$ Corrleation level (0.5)
Output: $OFS \leftarrow$ Optimal Feature Set *in* F
initialize a pairwise correlation coefficient matrix S of all features in F and let $S_{ij} = 0$; $OFS = \varnothing$
for $i = 1$ *to* n
  for $j = 1$ *to* n
    $S_{ij} = r\left(f_i, f_j\right) \big/ /$Eq.(3)
    if $S_{ij} \geq L$
      if $I(\{f_i\}, C) \geq I(\{f_j\}, C) //$Eq.(4)
        $OFS = OFS \cup \{f_i\}$
      else
        $OFS = OFS \cup \{f_j\}$
      endif
    endif
  endfor
endfor
Return OFS

Table 4
Results of the base classifiers on the CTPC.

Feature space	Model	AUC	Acc	F1
35 LS	RF	89.88	80	79.12
	LDA	89.53	78.95	78.26
	MLP	90.64	77.89	77.42
	LR	90.78	82.11	81.72
	SVC	91.67	83.16	82.98
	GNB	83.16	78.95	79.59
	XGB	90.37	76.84	76.6
14 LS Sig.	RF	83.58	77.89	76.92
	LDA	90.82	83.16	82.98
	MLP	92.03	80	79.57
	LR	91	84.21	83.87
	SVC	91.04	83.16	82.98
	GNB	86.54	81.05	80.85
	XGB	89.8	76.84	76.09
Top 200 Bi.-Tri.	RF	94.3	84.21	83.52
	LDA	89.8	80	79.12
	MLP	95.19	88.42	87.64
	LR	93.05	83.21	83.58
	SVC	95.5	86.32	85.06
	GNB	86.5	82.11	81.32
	XGB	91.67	83.16	82.22
Top combined 200 LS-Bi.-Tri.	RF	95.86	85.26	84.78
	LDA	90.37	82.11	81.32
	MLP	97.26	90.53	89.89
	LR	93.81	84.21	83.87
	SVC	94.79	84.21	83.52
	GNB	86.5	82.79	83.17
	XGB	95.14	88.42	87.64
Top combined 200 LS Sig.-Bi.-Tri.	RF	95.72	87.37	87.23
	LDA	90.46	82.05	81.32
	MLP	98.31	94.74	94.25
	LR	95.05	85.26	85.11
	SVC	95.68	85.26	85.42
	GNB	86.32	83.16	84
	XGB	95.59	88.42	87.64

LS., Lexicosyntactic feature space; LS. Sig., Significant lexicosyntactic feature space; Bi.-Tri., Bigrams and trigrams space; LS-Bi.-Tri., Fusion of lexicosyntactic, bigrams and trigrams spaces; LS Sig.-Bi.-Tri, Fusion of significant lexicosyntactic, bigrams and trigrams spaces

Table 5
Results of the base classifiers on the ADBC.

Feature space	Model	AUC	Acc	F1
35 LS	RF	92.39	84.78	84.96
	LDA	89.88	84.75	84.84
	MLP	93.39	85.02	85.17
	LR	89.89	84.57	84.68
	SVC	88.99	83.82	83.62
	GNB	82.65	76.77	75
	XGB	92.78	85.75	85.92
14 LS sig.	RF	92.58	86.03	86.18
	LDA	89.41	83.85	84.02
	MLP	93.54	86.93	87.14
	LR	89.42	84.39	84.59
	SVC	89.71	85.3	85.25
	GNB	85.24	78.04	78.73
	XGB	93.39	86.03	86.32
Top 200 Bi.-Tri.	RF	95.11	90.83	91.2
	LDA	94.05	93.99	94.01
	MLP	96.33	93.33	93.75
	LR	93.69	89.93	89.83
	SVC	95.11	91.67	92.06
	GNB	92.59	91.5	91.03
	XGB	95.47	91.83	91.96
Top combined 200 LS-Bi.-Tri.	RF	95.5	94.56	94.58
	LDA	94.78	94.56	94.55
	MLP	96.42	95.01	95.01
	LR	95.69	90.93	91.04
	SVC	95.9	93.47	93.57
	GNB	93.87	91.37	91.45
	XGB	95.49	92.1	92.9
Top combined 200 LS sig.-Bi.-Tri.	RF	96.56	94.74	94.76
	LDA	94.99	94.74	94.77
	MLP	96.89	95.19	95.17
	LR	95.23	91.47	91.65
	SVC	96.11	93.56	93.76
	GNB	94.9	91.56	91.58
	XGB	95.51	92.47	93.1

LS., Lexicosyntactic feature space; LS. Sig., Significant lexicosyntactic feature space; Bi.-Tri., Bigrams and trigrams space; LS-Bi.-Tri., Fusion of lexicosyntactic, bigrams and trigrams spaces; LS Sig.-Bi.-Tri, Fusion of significant lexicosyntactic, bigrams and trigrams spaces

value ranging between (-)0.30 to (-)0.50 indicates a low relation and (-)0.50 to (-)0.70 represents a moderate relation. As such, we set the threshold value to (-)0.50. Accordingly, in the proposed method, we firstly create a correlation coefficients matrix of feature-feature using Pearson's correlation (r). In case a coefficient of a pair of features overlaps with the given threshold value, it is then passed to mutual information (I), which returns the feature-class dependencies of each pairwise features in the matrix. Finally, the most predictive feature out of this pair is retained. This procedure is iterated until the feature space is neglectable-to-moderately correlated. This method is demonstrated in algorithm 1.***N-gram vocabulary spaces:*** The chi-square (X^2) statistic is an information-theoretic based feature selection method commonly used for revealing the most relevant feature space [65–67]. It presumes different distributions of top terms (i.e., character spaces in this study) t_p in negative and positive samples of class c_i. The X^2 assesses how dependent a term t and a class c are, where an increased score indicates a strong relevancy. The calculation of the score between a term t and a class c is exemplified in the illustrated contingency table (i.e., Table 3), wherein each cell denotes an observed value and the observed values are then used to compute the estimated value "E" as $E = (totalt^{*}totalc)/grandtotalN$. Eq. (5) represents the X^2 statistic.

$$X^2(t_p, c_i) = \frac{N(ad-bc)^2}{(a+b)(c+d)(b+d)(a+c)} \quad (5)$$

Whereby: the total samples in a dataset is represented by N, the total samples in class c_i that involves the term t_p is represented by a; the total samples of other classes that contain the term t_p is represented by b; the total samples in class c_i that do not include the term t_p is represented by c; and the total samples belong to other classes that do not contain the term t_p is denoted by d. A score then is allocated for each feature per each class, where all given values are then framed into a final X^2 value.

3.3. The proposed fusion model

In supervised classification, a classifier is considered a mapping function $\mathcal{f}$ trained to label the input vector to a target class as $\mathcal{f} : X \rightarrow Y$ so that given unseen language samples, it is able to predict the corresponding target class. An ensemble classifier $\mathcal{F}$ is built using multiple classifiers as $\mathcal{F}(x) = \mathcal{f}'(\mathcal{f}_1(x), \mathcal{f}_2(x), \ldots, \mathcal{f}_T(x))$, which typically outperforms single classifiers.

3.4. The development of the proposed stacked fusion model is described as follows:

3.4.1. Single classifier models

Our set of single classifiers includes Random Forest (RF), Extreme Gradient Boosting (XGB), Linear discriminant analysis (LDA), Support Vector Machine (SVC), Gaussian Naïve Bayes (GNB), Logistic Regression (LR), and Multi-layer Perceptron (MLP), which are predetermined in this study given their well performance in related studies [68,69]. In this context, we refer to this set as level-0 learners $\mathcal{f}'$. Each of these classifiers was optimized by cross-validated grid-search over a parameter grid and then assigned the best performing parameters.

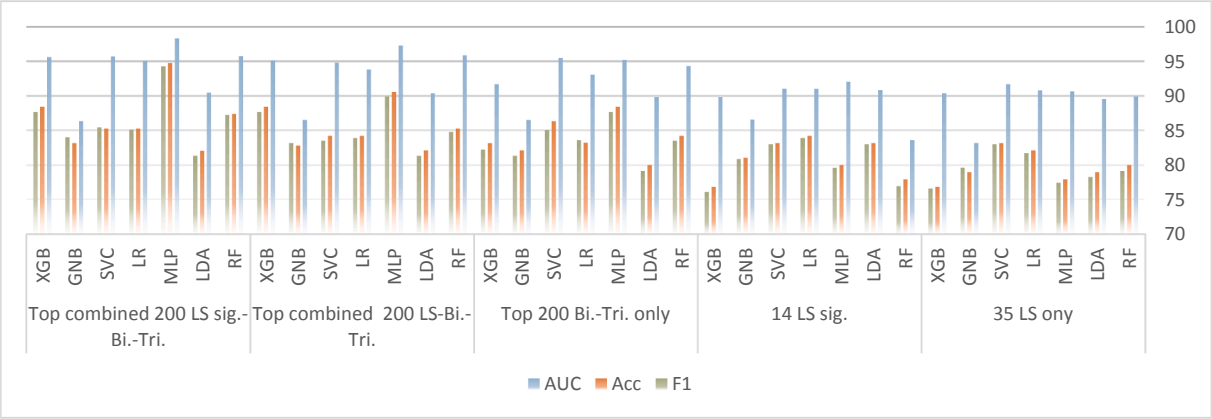


Fig. 2. Results of the base classifiers on the CTPC.

3.4.2. Stacked fusion model

Stacked generalization is a fusion learning technique driven by the principle of ensemble methods which entails combining the estimates of multiple base predictors to enhance the performance and generalizability over a single predictor [31,70]. In this paper, a stacked fusion model that assembles several diverse base classifiers is presented. The learning process from the training set D_{train} of this fusion model consists of two levels; a *meta*-classifier at a higher level (i.e., level-1) is trained on out-of-fold estimates of the base classifiers (i.e., level-0), producing a final estimate of the overall stacked model, as depicted in Fig. 1. To avoid overfitting that may result from training both levels on the same training set D_{train}, K-fold cross validation (CV) is incorporated in stacking, which is the most widely used evaluation technique of classification performance. It partitions D_{train} into predefined K disjoint subsets and runs the learning algorithm over K iterations, each of which entails training the learning algorithm on K-1 subsets and then using the trained model to predict on the remaining subset then return the model score. We assign $K = 10$ for eliminating the difference between true and predicted values as to reduce the bias [71]. The average score of all iterations is then returned as the final prediction score. This process is applied to each of the base classifiers $\mathscr{f}$, and then they are evaluated in parallel. Considering the associated computational complexity, we purify the set of base classifiers by selecting only the best performing classifiers. The output of the purified set of base classifiers $\mathscr{f}^t$ then forms the input feature space for the *meta*-classifier $\mathscr{F}$. Algorithm 2 illustrates this training process.

The develomnet of our stacked fusion model was carried out using Scikit-Learn[4] library, Python version 3.7.3.

Algorithm 2. Stacked Generalization with K-fold Cross Validation

Input:	Training dataset $D_{train} = \{x_i, y_i\}_{i=1}^{m} (x_i \in \mathbb{S}^n, y_i \in Y$, *where* $\mathbb{S}^n$ *is the feature space,* Y *is the class label set, and* m *is the number of trainng exambles*
Output:	*Stacking fusion model* F
1:	Step1 : Adopt cross validation approach in preparing a training set for second−level classifier (level−1)
2:	Randomly split D_{train} *into* K equal−size subsets : $D_{train} = \{D_{train_1}, D_{train_2}, \ldots, D_{train_K}\}$
3:	for $k \leftarrow 1$ *to* K do
4:	Step 1.1 : Learn first−level classifier
5:	for $t \leftarrow 1$ *to* T do, *where* T *is the total number of training sets*
6:	Learn a classifier $\mathscr{f}_{kt}$ from $D_{train} \backslash D_{train_K}$, *where* $\mathscr{f}$ *is a base classifier*
7:	endfor
8:	Step 1.2 : Evaluate each of 1^{st}−level classifiers and select only the best n predictive classifiers, where $n = 4$
9:	Step 1.3 : Construct a training set for 2^{nd}−level classifier using the selected n classifiers
10:	for $x_i \in D_{train_K}$ do
11:	Get a record $\{x_i^t, y_i\}$, where $x_i^t = \{\mathscr{f}_{k1}(x_i), \mathscr{f}_{k2}(x_i), \ldots, \mathscr{f}_{kT}(x_i)\}$
12:	endfor
13:	endfor
14:	Step2 : Learn second−level classifier
15:	Learn a new classifier $\mathscr{f}^t$ from the collection of $\{x_i^t, y_i\}$
16:	return $\mathscr{F}(x) = \mathscr{f}^t(\mathscr{f}_1(x), \mathscr{f}_2(x), \ldots, \mathscr{f}_T(x))$

This stacked fusion model can be deemed as a collective estimator of the errors resulting from the base classifiers against a specific training set that purifies these errors. On the other hand, it can be considered as a meta MLP that harnesses the base models (i.e., level-0) as neural units of a hidden layer and the meta model (i.e., level-1) as units of the output layer with an overall aim of increasing the prediction performance and generalizability of the model.

3.4.3. Evaluation

The performance evaluation of models presented in this work was carried out using a commonly used measurement in biomedical research, Area Under the Receiver Operating Characteristic curve (AUC). The AUC efficiently summarizes the performance of a diagnostic

[4] https://scikit-learn.org/stable/

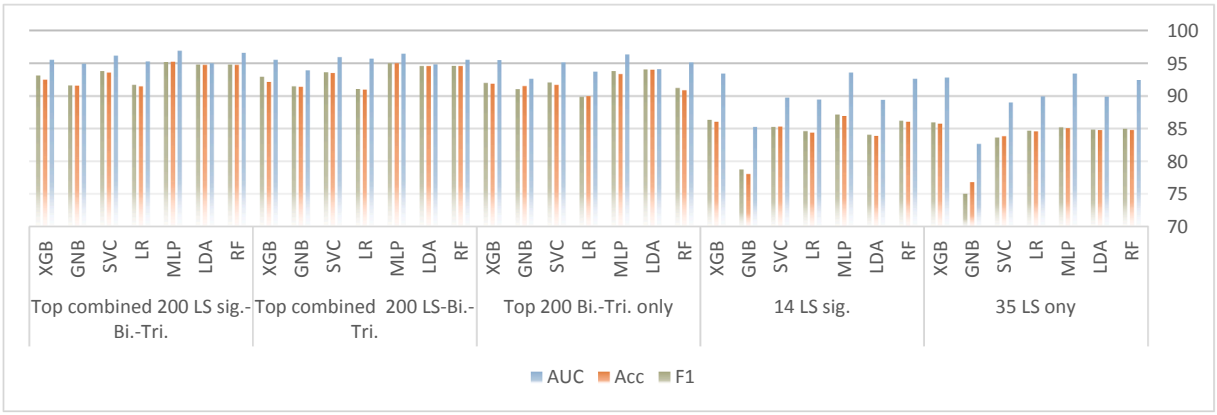


Fig. 3. Results of the base classifiers on the ADBC.

Table 6
Comparison of stacked fusion models with entire and reduced base classifiers on the CTPC.

Feature space	Model	AUC	Acc	F1
35 LS	SFM – 7*Cs*	91.4	77.89	77.42
	SFM – 4*Cs*	91.7	78.95	78.72
14 LS sig.	SFM – 7*Cs*	89.3	80	80
	SFM – 4*Cs*	92.8	82.11	82.11
Top 200 Bi.-Tri.	SFM – 7*Cs*	91.9	86.01	86
	SFM – 4*Cs*	95.3	87.64	88.42
Top combined 200 LS-Bi.-Tri.	SFM – 7*Cs*	97.11	91.95	92.63
	SFM – 4*Cs*	97.9	93.68	93.18
Top combined 200 LS sig.-Bi.-Tri.	SFM – 7*Cs*	97.5	91.58	90.91
	SFM – 4*Cs*	98.1	95.35	95.79

LS., Lexicosyntactic feature space; LS. Sig., Significant lexicosyntactic feature space; Bi.-Tri., Bigrams and trigrams space; LS-Bi.-Tri., Fusion of lexicosyntactic, bigrams and trigrams spaces; LS Sig.-Bi.-Tri, Fusion of significant lexicosyntactic, bigrams and trigrams spaces; SFM, Staked fusion model; Cs, Base classifiers

model and has been used across various diagnostic studies [72]. Eq. (6) illustrates this metric, where i runs over all m samples labeled 1 (i.e., AD), and j runs over all n samples labeled 0 (i.e., HC); p_i and p_j symbolize the probability score given by the classifier to sample i and j, sequentially. One (i.e., 1) is the output if the condition $p_i > p_j$ is satisfied.

$$AUC = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} 1_{p_i > p_j} \tag{6}$$

Besides, we use two other measurements widely utilized for assessing classification models, namely Accuracy and F1 score. Eq. (7–8) represent these metrics wherein, the total of correctly identified samples as positive signifies the True Positives (TP), the total of incorrectly identified samples as positive corresponds to the False Positives (FP), and the False Negative (FN) is the total of samples that have incorrectly been classified as negative. The weighted results are reported for F1 score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$F1 = 2^{*}\frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{8}$$

4. Results and discussion

The evaluation of our proposed fusion system entailed extensive experiments using both the ADBC and CTPC datasets to determine the extent to which it is efficient, initiated with lexicosyntactic features followed by character-based *n*-grams and a fusion of both feature spaces.

Table 7
Comparison of stacked fusion models with entire and reduced base classifiers on the ADBC.

Feature space	Model	AUC	Acc	F1
35 LS	SFM – 7*Cs*	92.99	84.84	84.84
	SFM – 4*Cs*	93.4	86.01	86.7
14 LS sig.	SFM –7*Cs*	92.5	85.3	86.11
	SFM – 4*Cs*	93.8	86.75	87.03
Top 200 Bi.-Tri.	SFM –7*Cs*	96.8	93.11	93.56
	SFM – 4*Cs*	97.5	94.17	94.49
Top combined 200 LS-Bi.-Tri.	SFM –7*Cs*	97.01	96.49	96.47
	SFM – 4*Cs*	98.11	97.01	97.03
Top combined 200 LS sig.-Bi.-Tri.	SFM –7*Cs*	98.03	96.26	96
	SFM – 4*Cs*	99.47	97.37	97.67

LS., Lexicosyntactic feature space; LS. Sig., Significant lexicosyntactic feature space; Bi.-Tri., Bigrams and trigrams space; LS-Bi.-Tri., Fusion of lexicosyntactic, bigrams and trigrams spaces; LS Sig.-Bi.-Tri, Fusion of significant lexicosyntactic, bigrams and trigrams spaces; SFM, Staked fusion model; Cs, Base classifiers

Table 8
Results of the ensemble methods on the CTPC.

Feature space	Model	AUC	Acc	F1
35 LS	Voting	90.69	78.11	78.03
	SFM	91.7	78.95	78.72
14 LS sig.	Voting	92.07	81.16	81.98
	SFM	92.8	82.11	82.11
Top 200 Bi.-Tri.	Voting	94.63	87.37	87.96
	SFM	95.3	87.64	88.42
Top combined 200 LS-Bi.-Tri.	Voting	95.99	91.11	91.32
	SFM	97.9	93.68	93.18
Top combined 200 LS Sig.-Bi.-Tri.	Voting	97.37	91.58	91.3
	SFM	98.1	95.35	95.79

LS., Lexicosyntactic feature space; LS. Sig., Significant lexicosyntactic feature space; Bi.-Tri., Bigrams and trigrams space; LS-Bi.-Tri., Fusion of lexicosyntactic, bigrams and trigrams spaces; LS Sig.-Bi.-Tri, Fusion of significant lexicosyntactic, bigrams and trigrams spaces

We firstly experimented with the entire lexicosyntactic space, including 35 features, which then was optimized by reducing the redundant features via a proposed CFS technique that eliminates inter-correlated features based on their relation to the target class. Afterwards, we experimented with different character *n*-grams to disclose the most discriminative spaces, where a combination of bigrams and trigrams was the most discriminative of the AD and HC on both ADBC and CTPC datasets. Given the scope of this paper, we reported only results of models based on this combination of bigrams and trigrams. Besides, we present a new approach of identifying AD through linked data, where we

Table 9
Results of the ensemble methods on the ADBC.

Feature space	Model	AUC	Acc	F1
35 LS	Voting	93.01	85.99	86.11
	SFM	93.4	86.01	86.7
14 LS Sig.	Voting	93.19	86.11	86.48
	SFM	93.8	86.75	87.03
Top 200 Bi.-Tri.	Voting	96.11	93.33	93.65
	SFM	97.5	94.17	94.49
Top combined 200 LS-Bi.-Tri.	Voting	97.24	96.01	96.03
	SFM	98.11	97.01	97.03
Top combined 200 LS sig.-Bi.-Tri.	Voting	98	96.55	96.55
	SFM	99.47	97.37	97.67

LS., Lexicosyntactic feature space; LS. Sig., Significant lexicosyntactic feature space; Bi.-Tri., Bigrams and trigrams space; LS-Bi.-Tri., Fusion of lexicosyntactic, bigrams and trigrams spaces; LS Sig.-Bi.-Tri, Fusion of significant lexicosyntactic, bigrams and trigrams spaces; SFM, Stacked fusion model

use the optimal feature space extracted from a fusion of both ADBC and CTPC datasets for training our base classifiers and the stacked fusion model. These results are presented in the following subsections.

4.1. Results of base classifiers

This section presents the evaluation of our seven base classifiers RF, XGB, LDA, SVC, GNB, LR, and MLP, employing different feature spaces extracted from both CTPC and ADBC datasets, with each space classified with base and stacked fusion models. Tables 4, 5 shows the performance of these models on a held-out dataset, presenting the AUC, accuracy, and F1 score of each model.

For CTPC lexicosyntactic feature space, we observed that SVC generated the best values of 91.67%, 83.16%, and 82.98% for AUC accuracy, and F1 score, respectively. For the optimized lexicosyntactic feature space, LR presented the best accuracy and F1 score of 84.21% and 83.87%, respectively. except for AUC. However, MLP slightly surpassed others with an AUC of 92.03%. This is mirrored in differentiating AD and HC groups from written language (i.e., ADBC), where MLP resulted in the best AUC on the entire as well as optimized lexicosyntactic spaces with values of 93.39% and 93.54%, respectively. While XGB showed the highest performance on the ADBC entire lexicosyntactic feature space with an accuracy of 85.75% and F1 score of 85.92%, MLP took the lead with the optimized lexicosyntactic space, achieving 86.93% accuracy and 87.14% F1 score.

Our *n*-gram space was seen to be effective in classifying AD and HC groups using both spoken and written language. For the CTPC testing set, SVC presented the best AUC of 95.5%, followed by MLP with a value of 95.19%, which, the later, also resulted in the best accuracy and F1 score with values of 88.42% and 87.64%, respectively. Likewise, we observed that the ADBC testing set was better classified with MLP with 96.33% AUC, but LDA outperformed others in terms of accuracy and F1 score of 93.99% and 94.01% in a row. On the other hand, while the fusion of lexicosyntactic and *n*-gram spaces enhanced all the base classifiers, we observed that MLP responded best to this fusion among these classifiers, scoring improved values on CTPC with all metrics with AUC of 97.26%, accuracy of 90.53%, and F1 score of 89.89%. This is interestingly mirrored on the ADBC testing set, where MLP scored the best response to the fusion of lexicosyntactic and *n*-gram spaces by achieving the highest AUC of 96.89%, accuracy of 95.19%, and F1 score of 95.01%. Likewise, fusing the optimized lexicosyntactic space with the *n*-gram space led to the best results on both datasets achieved with MLP, which could classify the CTPC testing set with 98.31% AUC, 94.74% accuracy, and 94.25 F1 score. Despite the slight difference between the base classifiers for differentiating the ADBC testing set, MLP generated

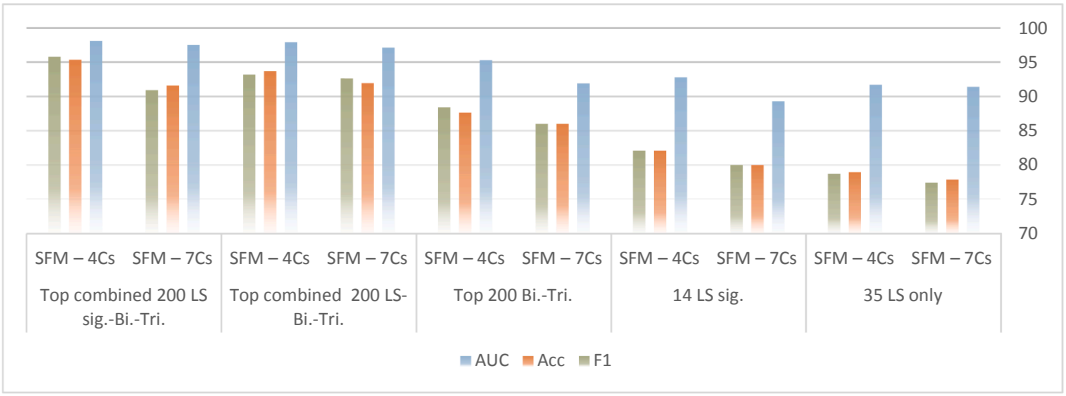


Fig. 4. Comparison of stacked fusion models with entire and reduced base classifiers on the CTPC.

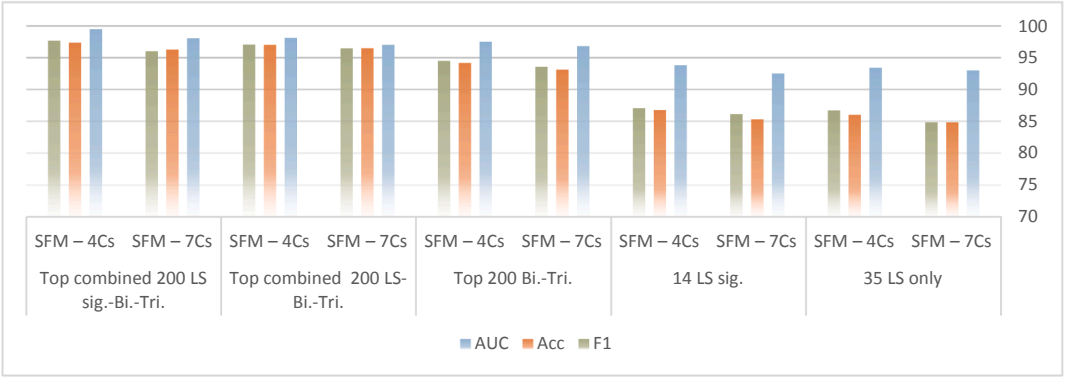


Fig. 5. Comparison of stacked fusion models with entire and reduced base classifiers on the ADBC.

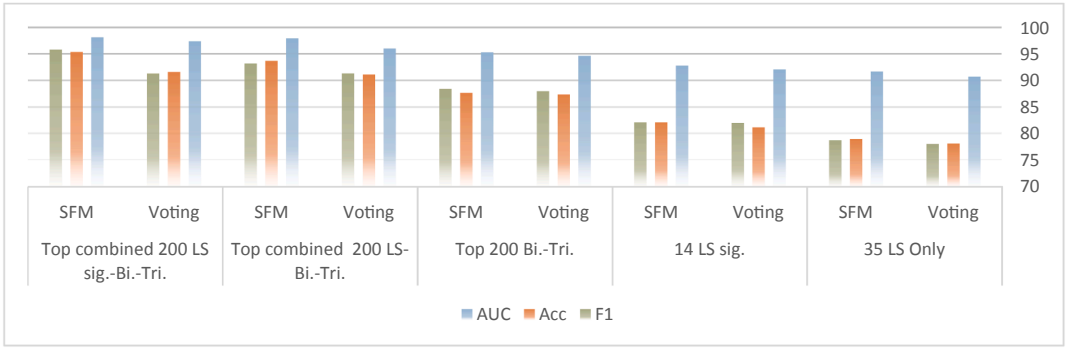


Fig. 6. Results of the ensemble methods on the CTPC.

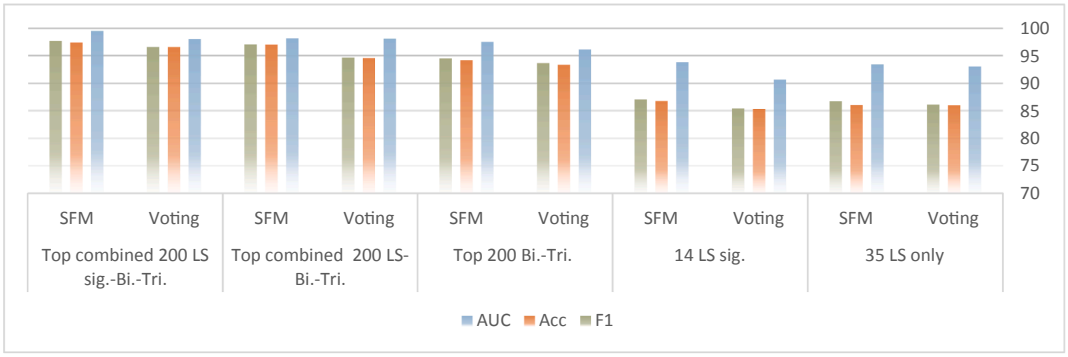


Fig. 7. Results of the ensemble methods on the ADBC.

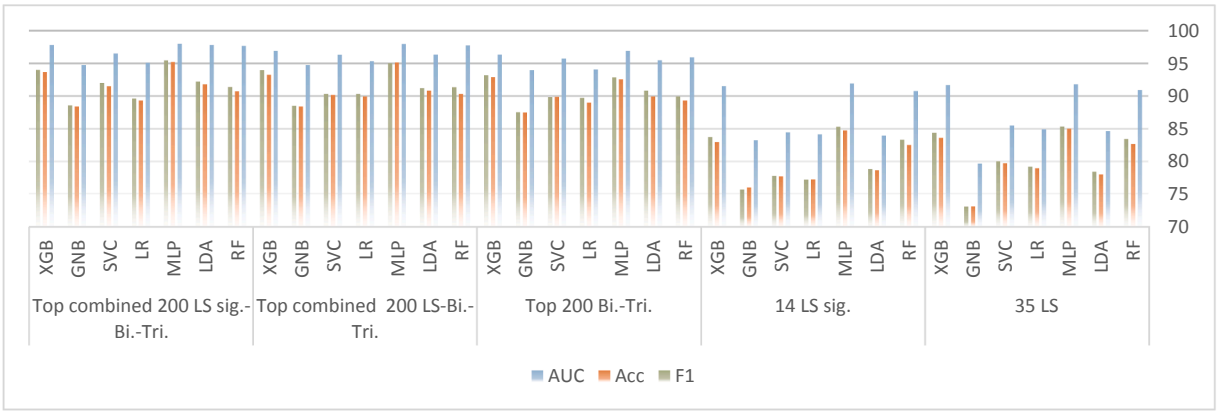


Fig. 8. Results of the base classifiers on linked data.

the best results with an AUC of 96.89%, accuracy of 95.19%, and F1 score of 95.17%.

From the short review above, key observations emerge: firstly, the optimized lexicosyntactic space fused with the *n*-gram space presents the optimal space for our base classifiers, leading to an improved classification performance on both spoken and written language. We also observed that the performance of all the classifiers improved as we optimized and fused the feature spaces with an exemption to that being the CTPC optimized lexicosyntactic feature space, where the performance dropped for some classifiers. Figs. 2, 3 show a direct comparison of the feature spaces and the corresponding performance of our base classifiers.

4.2. Results of stacked fusion models

It is an intuitive anticipation to see an improved performance resulted from ensemble classifiers. Accordingly, we expect our stacked fusion models to increase the performance over the single base classifiers. Several experiments were conducted to evaluate and compare our proposed models, where we firstly compare our stacked fusion models that were built based on entire set of base classifiers alongside that of the purified set of base classifiers. Afterwards, we compare these stacked fusion models against soft voting, which is another popular and effective ensemble method. Each feature space was utilized separately prior to feature fusion. The results are summarized in this section and illustrated in Tables 6-9.

Our experiments showed gradual improvements of stacked fusion models based on purified set of base classifiers (i.e., RF, MLP, SVC, and

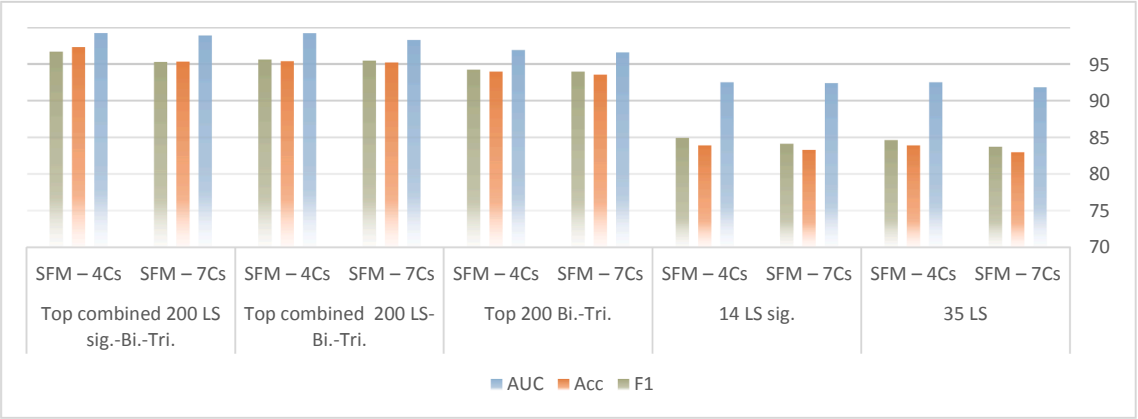


Fig. 9. Comparison of stacked fusion models with entire and reduced base classifiers on linked data.

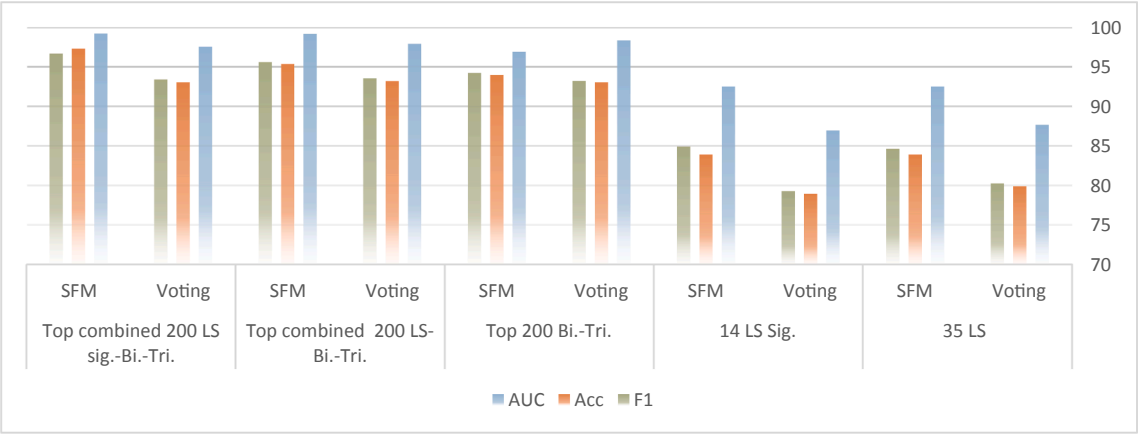


Fig. 10. Results of the ensemble methods on the linked data.

XGB) against the entire base classifiers as we optimized and fused feature spaces. As illustrated in Tables 5, 6, reducing the set of classifiers enhanced the learning process of the *meta*-classifier resulting in an enhanced performance over all the feature spaces. For instance, while there was an improvement of AUC by 0.3 – 0.4% on the lexicosyntactic space of both datasets, the values of accuracy and F1 score improved by a larger margin of 1 – 1.5%. This was the case with the optimized lexicosyntactic space on both datasets, where we observed improvement by 1.5 – 3.5% on AUC and 1.5 – 2% on accuracy and F1 score. The behavior of the stacked fusion models using character *n*-grams was relatively similar, with the fusion of the optimized lexicosyntactic and *n*-grams spaces led to the best results achieved with the stacked fusion models based purified set of base classifiers. Figs. 4, 5 depict the behaviours of these stacked fusion models.

In the comparison of our best performing stacked fusion models against the voting ensemble models, we observed that the stacked fusion models surpassed the voting models on all spoken and written-based feature spaces. For the CTPC original lexicosyntactic space, the stacked fusion model scored 91.7% AUC, 78.95% accuracy, and 78.72% F1 score against 90.69% AUC, 78.11% accuracy, and 78.03 F1 score achieved with the voting model. This is the case with ADBC, showing the relative effectiveness of stacked model over the voting model. As shown in Tables 7, 8, using the optimized lexicosyntactic space improved the values of these metrics by 0.45–1% on the CTPC and 0.4–0.75% on the ADBC. As expected, slightly superior results were achieved with the *n*-gram space, where the stacked fusion model could classify the spoken language samples with an AUC of 95.3%, accuracy of 87.64%, and F1 score of 88.42%. Similarly, it achieved 97.5% AUC, 94.17% accuracy, and 94.49% F1 score when differentiating the written language samples. The fusion of lexicosyntactic and *n*-gram spaces improved the performance of the stacked fusion model, increasing its AUC to 97.9% on the CTPC and 98.11% on the ADBC. Interestingly, the accuracy and F1 score of the model were significantly improved to 93.68% and 93.18%, respectively, on the CTPC testing set. This is mirrored on the ADBC, where it achieved an accuracy of 97.01% and F1 score of 97.03%. At last, our stacked fusion model scored its highest performance with the fusion of the optimized lexicosyntactic and *n*-gram spaces with 98.1% AUC, 95.35% accuracy, and 95.79% F1 score on the CTPC and 99.47% AUC, 97.37% accuracy, and 97.67% F1 score on the ADBC. This comparison is depicted in Figs. 6, 7.

Together, the present results suggest that the fusing the optimized lexicosyntactic and *n*-grans spaces forms the optimal feature space for our stacked fusion model beside being effective for the voting classifier as well as the base classifiers, which may be explainable given the optimization of the lexicosyntactic space.

4.3. Results of base and stacked models on linked data

Another suggestion can be drawn from these findings is the fact that the performance of the stacked fusion model could be positively impacted by the number of samples given the improved results on the ADBC testing set over that of the CTPC. Consequently, we examined the effect of linking these data sources on our stacked fusion model, were we fused our feature spaces into the base and stacked fusion classifiers with an overall aim of developing a robust hybrid model for foreseeing AD from spoken and written language given the promising results achieved

Table 10
Results of the base classifiers on linked data.

Feature space	Model	AUC	Acc	F1
35 LS	RF	90.91	82.66	83.43
	LDA	84.64	78.02	78.42
	MLP	91.79	84.98	85.33
	LR	84.91	78.95	79.2
	SVC	85.49	79.72	80
	GNB	79.67	73.11	73.08
	XGB	91.66	83.59	84.37
14 LS sig.	RF	90.76	82.51	83.31
	LDA	83.92	78.64	78.83
	MLP	91.91	84.76	85.29
	LR	84.14	77.24	77.21
	SVC	84.45	77.71	77.78
	GNB	83.22	76.01	75.67
	XGB	91.53	82.97	83.73
Top 200 Bi.-Tri.	RF	95.9	89.32	89.93
	LDA	95.47	89.93	90.83
	MLP	96.89	92.57	92.86
	LR	94.05	88.99	89.71
	SVC	95.72	89.88	89.83
	GNB	93.95	87.46	87.52
	XGB	96.33	92.88	93.18
Top combined 200 LS-Bi.-Tri.	RF	97.73	90.31	91.35
	LDA	96.33	90.83	91.2
	MLP	97.94	95.11	95.01
	LR	95.31	89.94	90.31
	SVC	96.3	90.18	90.31
	GNB	94.74	88.39	88.48
	XGB	96.88	93.22	93.95
Top combined 200 LS sig.-Bi.-Tri.	RF	97.64	90.71	91.38
	LDA	97.81	91.8	92.19
	MLP	98.01	95.2	95.45
	LR	95.1	89.32	89.62
	SVC	96.49	91.49	91.99
	GNB	94.75	88.39	88.58
	XGB	97.79	93.65	94

LS., Lexicosyntactic feature space; LS. Sig., Significant lexicosyntactic feature space; Bi.-Tri., Bigrams and trigrams space; LS-Bi.-Tri., Fusion of lexicosyntactic, bigrams and trigrams spaces; LS Sig.-Bi.-Tri, Fusion of significant lexicosyntactic, bigrams and trigrams spaces

Table 11
Comparison of stacked fusion models with entire and reduced base classifiers on linked data.

Feature space	Model	AUC	Acc	F1
35 LS	SFM – 7*Cs*	91.8	82.97	83.73
	SFM – 4*Cs*	92.5	83.9	84.62
14 LS sig.	SFM – 7*Cs*	92.4	83.28	84.12
	SFM – 4*Cs*	92.5	83.9	84.92
Top 200 Bi.-Tri.	SFM – 7*Cs*	96.56	93.54	93.95
	SFM – 4*Cs*	96.89	93.96	94.22
Top combined 200 LS-Bi.-Tri.	SFM – 7*Cs*	98.26	95.2	95.45
	SFM – 4*Cs*	99.17	95.36	95.59
Top combined 200 LS sig.-Bi.-Tri.	SFM – 7*Cs*	98.88	95.31	95.27
	SFM – 4*Cs*	99.2	97.29	96.66

LS., Lexicosyntactic feature space; LS. Sig., Significant lexicosyntactic feature space; Bi.-Tri., Bigrams and trigrams space; LS-Bi.-Tri., Fusion of lexicosyntactic, bigrams and trigrams spaces; LS Sig.-Bi.-Tri, Fusion of significant lexicosyntactic, bigrams and trigrams spaces; SFM, Staked fusion model; Cs, Base classifiers

using each dataset. We initiated this approach with the single base classifiers followed by the stacked fusion model to better understand the overall behaviour of our models with these features and data fusions. As illustrated in Figs. 8, 9 fusing the written and spoken language resulted in an encouraging performance across both the base classifiers and the stacked fusion model. For instance, as we anticipated, MLP took the lead almost across all feature spaces, scoring the highest AUC of 98.01% using the top combined *n*-grams and significant lexicosyntactic features, with corresponding accuracy of 95.2% and F1 score of 95.45%. Despite being slight difference, XGB outperformed MLP in terms of accuracy and F1 score with the fusion of *n*-grams space with 92.88% and 93.18%, consecutively, which was not the case with the disjointed datasets. We also noticed that the stacked fusion algorithm (i.e., algorithm 2) returned different purified set of base classifiers with this fusion of datasets, where the selected set of classifiers included LDA, MLP, SVC, and XGB. Similar to that of separate datasets, this purified set led to the highest performance with 99.2% AUC, 97.29% accuracy, and 96.66% F1 score, which is better than that of the entire set of classifiers. These scores present improvements by approximately 2% on AUC and 1% on accuracy and F1 score compared to using the spoken language alone. Table 11 presents the results of this data fusion. Besides, the stacked fusion model behaved similarly on the linked data by outperforming the soft voting across all feature spaces by around 1.3 – 6%, which emphasizes the effectiveness of stacked generalization compared to other ensemble methods even on linked data, as shown in Fig. 10. Tables 10-12 present the results of these data and feature fusions.

Table 12
Results of the ensemble methods on liked data.

Feature space	Model	AUC	Acc	F1
35 LS	Voting	87.65	79.88	80.24
	SFM	92.5	83.9	84.62
14 LS Sig.	Voting	86.93	78.95	79.27
	SFM	92.5	83.9	84.92
Top 200 Bi.-Tri.	Voting	98.33	93.03	93.21
	SFM	96.89	93.96	94.22
Top combined 200 LS-Bi.-Tri.	Voting	97.89	93.19	93.55
	SFM	99.17	95.36	95.59
Top combined 200 LS sig.-Bi.-Tri.	Voting	97.54	93.03	93.39
	SFM	99.2	97.29	96.66

LS., Lexicosyntactic feature space; LS. Sig., Significant lexicosyntactic feature space; Bi.-Tri., Bigrams and trigrams space; LS-Bi.-Tri., Fusion of lexicosyntactic, bigrams and trigrams spaces; LS Sig.-Bi.-Tri, Fusion of significant lexicosyntactic, bigrams and trigrams spaces; SFM, Stacked fusion model

Table 13
Comparison of baselines against the stacked fusion model on the ADBC.

	Model	AUC	Accuracy	F1 score
Baselines	Masrani et al.	84.4	–	–
	Kong et al.	–	93.4	94.4
Ours	Stacked Fusion Model	99.47	97.37	97.67

This study extends previous work on AD diagnosis based on natural language processing and machine learning by exploring lexicosyntactic features and combining them with character *n*-grams then using these features with carefully selected single classifiers with an overall aim to develop a robust stacked fusion model. Additionally, it examines the fusion of spoken and written language samples and its effect on the base as well as ensemble learning algorithms. The initial interpretation drawn from our experimental results would be the effectiveness of the proposed approach in diagnosing AD, given the gradual improvements achieved with both base and ensemble models as we optimized and fused the feature spaces. A possible explanation could be our preprocessing of these feature spaces; for example, lexicosyntactic space was normalized using outlier-proof normalization technique, which could have highly contributed to the performance. Moreover, the proposed CFS method was designed to eliminate redundant features to an extent most suitable to this diagnostic task thus may have enhanced the learning process of the ML algorithms. Interestingly, the feature-feature and feature-class inter-correlation have rarely been investigated together from medical point of view for such task. Besides, it is also worth mentioning that the preprocessing involved the removal of functional as well as non-informative words prior to generating the *n*-gram space, resulting in a low-level character-based vocabulary space. The selection of our top *n*-gram space using the X^2 is also a potential

Table 14
Comparison of baselines against the stacked fusion model on the CTPC.

	Model	AUC	Acc.	F1
Baselines	**Orimaye et al.**SVM	93	–	–
	Di Palo and NatalieCNN-LSTM	92.07	84.95	91.07
	CNN-LSTM-ATT	95.03	84.66	91.58
	CNN-LSTM-ATT-W	94.98	88.20	93.05
	Chen et al.ATT-CNN + ATT-BiGRU	–	97.42	–
Ours	Stacked Fusion Model	98.1	95.35	95.79

reason given its effectiveness in related work [4]. On the other side, the hyperparameter optimization performed using CV grid search are seen to be effective in a similar task thus a probable reason behind the achieved performance [27]. At last, using the fusion of spoken and written language samples with our learning algorithms led to an optimal hybrid stacked fusion model.

4.4. Comparison with related work

Previous studies have introduced promising ML models and asserted the importance of linguistic analysis for the early diagnosis of AD. Given the challenges to comparison of such approach against these studies addressed in our recent work [4], we selected a few related studies in contrast to our results and highlighted how they differ from our approach. In terms of written language, we evaluate our work alongside the work of Masrani et al. [35] and the work of Kong et al. [36], where both used the same blogs for building their predictive models. Masrani et al. [35] have initiated the ADBC by crawling the six blogs highlighted in Table (1), where they extracted over 100 lexicosyntactic features and fused them into several ML algorithms. An AUC of 84.8% was reported as the best score achieved with MLP. Kong et al. [36], on the other hand, used the same corpus for training hierarchical attention networks as an attempt to optimize the prediction of AD via written language samples without task-specific feature engineering. Interestingly, they reported their best average accuracy and F1 of 93.4% and 94.4%, respectively.

Our approach can be distinguished from a few perspectives; firstly, while both baselines have used the original blogs, we have recrawled them which resulted in a bigger dataset. Besides, they used a blog belonging to a Lewy body patient along that of AD which was ignored in our case given the AD diagnostic task. Unlike both, we conducted extensive experiments with and without optimization of the feature space. Another significant difference is the robustness of our stacked model compared to single models in both baselines given the reduced errors of ensemble models in general. Moreover, we examine the effect of linked data on our best modelling feature space. Finally, we evaluate our models using three different metrics in parallel, where both baselines were outperformed on each of these metrics. For example, the stacked fusion model remarkably surpassed that of Masrani et al. (i.e., MLP) by 15% with an AUC of 99.47% against 84.4%. Likewise, it exceeded the hierarchical attention networks model of Kong et al. by approximately 4% with an accuracy of 97.37% against 93.4% and F1 score of 97.67% against 94.4%. Table 13 presents this comparison.

For spoken language, on the other hand, we have chosen the work of Orimaye et al. [42], Di Palo and Parde [73], and Chen et al. [74] since they present the best performing models on the CTPC. The details of these baselines were discussed alongside most of related studies on DementiaBank dataset in our recent work [4]. Firstly, Orimaye et al. [42] introduced a few models trained with similar feature spaces, where they investigated lexicosyntactic features and fused them with *n*-grams then selected the top 1000 for training an SVC, scoring an AUC of 93%. Di Palo and Parde [73] integrated Convolutional Neural Networks (CNNs) and Long Short-Term Memory-Recurrent Neural Networks (LSTM-RNNs) with added attention mechanism and class weight, with an overall aim to optimize the performance of CTPC classification. They reported 95.03%, 88%, and 93.05% for AUC, accuracy, and F1 score, respectively. Likewise, Chen et al. [74] combined CNNs with a bidirectional GRU (BiGRU) and added attention mechanism, introducing a hybrid neural model that achieved an accuracy of 97.42%.

These baselines differ from our work from a few main points. firstly, while our feature space may be similar to that of Orimaye et al., our lexicosyntactic space is inherited from the production of POS tags and the vocabulary space is character-based that has been extracted from low-level language samples. Besides, they used the originally annotated language samples which may limit the generalizability of their models. Another difference is the involvement of demographic information in their feature space while our feature space is mainly focusing on language patterns. In light of these differences, our stacked fusion model outperformed their model by 5%. Regarding the work of Di Palo and Parde, we can distinguish our by stating a few perspectives; initially, while their feature space included POS tags, these POS tags were annotated by the CTPC custodian, meaning that the generalizability of their model to unseen data may be affected. Furthermore, unlike them, we dealt with the skewness of the classes by reducing the majority to equalize the minority, taking into consideration the participants' age and level of education. Another potential drawback of their approach would be the sentence-based classification given the connected speech-based diagnostic task [74]. Finally, our stacked fusion model surpassed their best models by 2–7% on AUC, accuracy, and F1 score. The third baseline of Chen et al. [74], which presents the state-of-the-art model on CTPC, can be distinguished from our approach from three main points; at first, they evaluated their work with the accuracy metric only despite using an imbalanced dataset, which has been asserted to be misleading in such a case [75]. Furthermore, our approach examines two different feature spaces and a fusion of them, which could shed light on the behaviour on our models towards the results. Another point worthwhile to mention is the time complexity of RNNs given its nature of being unable to be parallelly processed [76]. Table 14 presents these baselines against our models.

Overall, a main difference that makes our approach stands against these baselines is the investigation of linked spoken and written data towards a hybrid diagnostic model, which, to the best of our knowledge, has not been explored previously. Moreover, stacked generalization is known to be robust and stable given its ability of selecting the best models out of the base algorithms and avoiding the corresponding errors [77].

5. Conclusion

Even though many ML approaches have been proposed to enhance the automation diagnosis of AD, it still requires robust models towards fully automated process that can replace the initial manual clinical-based diagnosis. This study investigates the development of multiple heterogeneous stacked fusion models for predicting AD from written and spoken-based language samples. Specifically, it explores the representation of lexicosyntactics in spoken and written language samples of AD and HC groups and examine their efficacy in identifying AD patients from healthy adults. Besides, it introduces character-based language models and explores the fusion of lexicosyntactics into these language models. We also introduce a CFS based on the fusion of two well-known algorithms (i.e., Pearson's correlation and mutual information) to

optimize the lexicosyntactic feature space. Using a fusion of the optimized lexicosyntactic space and character-based vocabulary spaces, we introduce a hybrid stacked fusion model that could classify the linked written and spoken language samples of AD and HC groups with an AUC of 99.1%, accuracy of 96.51, and F1 score of 96.74%. Likewise, our stacked fusion model puts benchmarks on the ADBC with 99.47%, 97.37%, and 97.67% of AUC, accuracy, and F1 score, respectively. For CTPC, the stacked fusion model achieved benchmarks on AUC with 98.1% and F1 score with 95.79%. Our models emphasize the effectiveness of ensemble methods for AD diagnosis, suggesting the replacement of traditional screening tests with such ML models.

Acknowledgment

This work uses a subset of the cookie theft picture corpus (CTPC) from DementiaBank dataset, which is being presently supported by NICHD, SBE NIDCD, RIDIR, and NSF and maintained at Carnegie Mellon University.

Appendix

(See Table 15)

Table 15
Examples of AD and HC language samples from CTPC and ADBC.

CTPC	HC	uh badly damaged sink. mothers drying dishes. uh is it action if she is standing in a puddle of water ? no, water is running over f is overflowing the sink. the window is open. it is blowing the curtain. uh billys about to fall off the stool while handing cookies to his sister. that is another thing. and uh . waters running on the floor as well as out of the sink . waters running out of the faucet. does that count too ? okay. i do not see anything else happening.
	AD	first of all the little girls saying "/. " shh. and and he is climbing up to get a cookie. and he is going to fall. and the stool is on tipping. the water is running over in the sink. uh . the towel seems to go in one side and out the other side of the dish. oh that is part of the curtain i guess. that is what it is. that is part of the curtain . it looked like it c gone through here and come out here . ah. two cups and a dish. i do not see anything unusual.
ADBC	HC	Today I took mom out to Target to buy a few household essentials and to shop for dad's birthday, which is this upcoming weekend. I wasn't sure how this would go. For those faithful readers of mine, you can understand why I would be a little apprehensive about gift shopping with mom, after our Christmas shopping experience. As usual, mom had a list ready to go when I picked her up. She generally makes her list by looking at the labels of the product she wants; sometimes there is no empty bottle to look at and she has to write down from memory what it is she wants. This usually requires much interpretation.
	AD	It is beocming harder each day to remember yesterday, let alone try today, even this monring. Here we are mid October per the calendar and I am somewere in September. I feel like I am altogether but I am totallly confused as to what to do next. This is not fun. But it is the hand I was dealt so I play the cards and see what they do and try to understand them. The new meds seem to help the pains in the chest and the sweating whrn walking or doing some work. So I gues complaining is just whimpy on my part. I wish I had the words to discribe what it is like to be in this World that I and others like me live in. It reallly sucks. First you know then you forget and then you do not remember forgetting what it was that you knew. If tha makes no sense to you ok, but it does to me.

Although we attempted to overcome some challenges in the given diagnostic task such as the sparse nature of related datasets by combining two different datasets, a weakness of this work would be the small volume of the spoken language dataset, aligning with other computational diagnostic studies that suffer from this out of control limitation [78,79]. Further, this study is mainly concerned with English language, thus; models may not be suitable for other languages. Additionally, we reduce our character *n*-grams space to the top 200, which warrants further investigations of different thresholds that may increase the overall performance. Despite being slightly imbalanced, our inclusion of ADBC without handling its skewness could be another limitation if we ignore the drawbacks of subsampling technique; yet the effect may be neglectable given the large number of samples compared to CTPC. Our future work would investigate a potential solution to balance these datasets while maintain the associated limitations. It will also address how variant spoken and written language samples of AD patients by addressing lexicosyntactics alongside other linguistic patterns. Another future work would be to estimate the extent to which each of these lexicosyntactic features relate to AD. It would also be interesting to see the behaviour of these stacked fusion model over other related datasets such as dem@care [80], which is a planned investigation of our future work.

CRediT authorship contribution statement

Ahmed H. Alkenani: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Yuefeng Li:** Supervision, Methodology. **Yue Xu:** Review. **Qing Zhang:** Review.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N. Dey, A.S. Ashour, S. Borra, Classification in BioApps: automation of decision making, vol. 26:, Springer, 2017.
[2] M.A. Myszczynska, P.N. Ojamies, A.M. Lacoste, D. Neil, A. Saffari, R. Mead, et al., Applications of machine learning to diagnosis and treatment of neurodegenerative diseases, Nature Reviews Neurology 16 (2020) 440–456.
[3] J. Xu, Y. Zhang, C. Qiu, F. Cheng, Global and regional economic costs of dementia: a systematic review, The Lancet 390 (2017) S47.
[4] A. Alkenani, Y. Li, Y. Xu, Q. Zhang, Predicting Prodromal Dementia Using Linguistic Patterns and Deficits, IEEE Access (2020) 1.
[5] N. Herrmann, K. L. Lanctôt, and D. B. Hogan, "Pharmacological recommendations for the symptomatic treatment of dementia: the Canadian Consensus Conference on the Diagnosis and Treatment of Dementia 2012," Alzheimer's research & therapy, vol. 5, p. S5, 2013.
[6] NHS, "What are the treatments for dementia?," 2018.
[7] D. Kempler, "Language changes in dementia of the Alzheimer type," Dementia and communication, pp. 98-114, 1995.
[8] A.J. Mitchell, M. Shiri-Feshki, Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies, Acta Psychiatrica Scandinavica 119 (2009) 252–265.
[9] M. Canevelli, G. Grande, E. Lacorte, E. Quarchioni, M. Cesari, C. Mariani, et al., Spontaneous reversion of mild cognitive impairment to normal cognition: a systematic review of literature and meta-analysis, Journal of the American Medical Directors Association 17 (2016) 943–948.
[10] M. Malek-Ahmadi, Reversion from mild cognitive impairment to normal cognition, Alzheimer Disease & Associated Disorders 30 (2016) 324–330.
[11] M. Ganguli, Y. Jia, T.F. Hughes, B.E. Snitz, C.C.H. Chang, S.B. Berman, et al., Mild Cognitive Impairment that Does Not Progress to Dementia: A Population-Based Study, Journal of the American Geriatrics Society 67 (2019) 232–238.
[12] M.S. Albert, S.T. DeKosky, D. Dickson, B. Dubois, H.H. Feldman, N.C. Fox, et al., The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease, Alzheimer's & dementia 7 (2011) 270–279.
[13] V. Taler, N.A. Phillips, Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review, Journal of clinical and experimental neuropsychology 30 (2008) 501–556.
[14] E. Giles, K. Patterson, J.R. Hodges, Performance on the Boston Cookie Theft picture description task in patients with early dementia of the Alzheimer's type: missing information, Aphasiology 10 (1996) 395–408.
[15] X. Le, I. Lancashire, G. Hirst, R. Jokel, Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists, Literary and Linguistic Computing 26 (2011) 435–461.
[16] S. Ahmed, C.A. de Jager, A.-M. Haigh, P. Garrard, Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease, Neuropsychology 27 (2013) 79.

[17] B. Roark, M. Mitchell, J. Hosom, K. Hollingshead, J. Kaye, Spoken Language Derived Measures for Detecting Mild Cognitive Impairment, Audio, Speech, and Language Processing, IEEE Transactions on 19 (2011) 2081–2090.

[18] M. Lehr, I. Shafran, E. Prud'hommeaux, and B. Roark, "Discriminative joint modeling of lexical variation and acoustic confusion for automated narrative retelling assessment," in Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 211-220.

[19] A. Satt, A. Sorin, O. Toledo-Ronen, O. Barkan, I. Kompatsiaris, A. Kokonozi, et al., Evaluation of speech-based protocol for detection of early-stage dementia, ed:, International Speech and Communication Association, 2013, pp. 1692–1696.

[20] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, et al., Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease, Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 1 (2015) 112–124.

[21] M. Asgari, J. Kaye, H. Dodge, Predicting mild cognitive impairment from spontaneous spoken utterances, Alzheimer's & Dementia: Translational Research & Clinical Interventions 3 (2017) 219–228.

[22] L. Toth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatloczki, Z. Banreti, et al., A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech, Current Alzheimer Research 15 (2018) 130–138.

[23] M. Lehr, E. Prud'hommeaux, I. Shafran, B. Roark, Fully automated neuropsychological assessment for detecting mild cognitive impairment. in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[24] M.J. Ball, M.R. Perkins, N. Müller, S. Howard, The handbook of clinical linguistics, Wiley Online Library, 2008.

[25] V. Rentoumi, G. Paliouras, E. Danasi, D. Arfani, K. Fragkopoulou, S. Varlokosta, et al., Automatic detection of linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: A computational linguistics analysis, in: in *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2017, pp. 000033–000038.

[26] B. Wang, Z. Mao, A dynamic ensemble outlier detection model based on an adaptive k-nearest neighbor rule, Information Fusion (2020).

[27] N. El-Rashidy, S. El-Sappagh, T. Abuhmed, S. Abdelrazek, H.M. El-Bakry, Intensive Care Unit Mortality Prediction: An Improved Patient-Specific Stacking Ensemble Model, IEEE Access 8 (2020) 133541–133564.

[28] H.-J. Yoon, H.B. Klasky, J.P. Gounley, M. Alawad, S. Gao, E.B. Durbin, et al., Accelerated training of bootstrap aggregation-based deep information extraction systems from cancer pathology reports, Journal of Biomedical Informatics 110 (2020), 103564.

[29] M.H.D.M. Ribeiro, L. dos Santos Coelho, Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series, Applied Soft Computing 86 (2020), 105837.

[30] U. K. Kumar, M. S. Nikhil, and K. Sumangali, "Prediction of breast cancer using voting classifier technique," in 2017 IEEE international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM), 2017, pp. 108-114.

[31] D.H. Wolpert, Stacked generalization, Neural networks 5 (1992) 241–259.

[32] M.J. Siers, M.Z. Islam, Software defect prediction using a cost sensitive decision forest and voting, and a potential solution to the class imbalance problem, Information Systems 51 (2015) 62–71.

[33] K.P. Riley, D.A. Snowdon, M.F. Desrosiers, W.R. Markesbery, Early life linguistic ability, late life cognitive function, and neuropathology: findings from the Nun Study, Neurobiology of aging 26 (2005) 341–347.

[34] S. Kemper, L.H. Greiner, J.G. Marquis, K. Prenovost, T.L. Mitzner, Language decline across the life span: Findings from the nun study, Psychology and aging 16 (2001) 227.

[35] V. Masrani, G. Murray, T. Field, G. Carenini, Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia, BioNLP 2017 (2017) 232–237.

[36] W. Kong, H. Jang, G. Carenini, T. Field, A Neural Model for Predicting Dementia from Language, Machine Learning for Healthcare Conference (2019) 270–286.

[37] J.T. Becker, F. Boiler, O.L. Lopez, J. Saxton, K.L. McGonigle, The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis, Archives of Neurology 51 (1994) 585–594.

[38] E. Ouyang, Y. Li, L. Jin, Z. Li, X. Zhang, Exploring n-gram character presentation in bidirectional RNN-CRF for chinese clinical named entity recognition, CEUR Workshop Proc (2017) 37–42.

[39] I. Solti, C. R. Cooke, F. Xia, and M. M. Wurfel, "Automated classification of radiology reports for acute lung injury: comparison of keyword and machine learning based natural language processing approaches," in 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop, 2009, pp. 314-319.

[40] B. MacWhinney, The CHILDES project: The database vol. 2: Psychology Press, 2000.

[41] S. O. Orimaye, J. S.-M. Wong, and J. S. G. Fernandez, "Deep-Deep Neural Network Language Models for Predicting Mild Cognitive Impairment," in BAI@ IJCAI, 2016, pp. 14-20.

[42] S.O. Orimaye, J.S. Wong, K.J. Golden, C.P. Wong, I.N. Soyiri, Predicting probable Alzheimer's disease using linguistic deficits and biomarkers, BMC bioinformatics 18 (2017) 34.

[43] S. Wankerl, E. Nöth, S. Evert, An N-gram based approach to the automatic diagnosis of Alzheimer's disease from spoken language. in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2017.

[44] S. O. Orimaye, J. S.-M. Wong, and K. J. Golden, "Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances," in Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2014, pp. 78-87.

[45] G. Kavé, A. Dassa, Severity of Alzheimer's disease and language features in picture descriptions, Aphasiology 32 (2018) 27–40.

[46] R.S. Bucks, S. Singh, J.M. Cuerden, G.K. Wilcock, Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance, Aphasiology 14 (2000) 71–91.

[47] S.V. Pakhomov, G.E. Smith, D. Chacon, Y. Feliciano, N. Graff-Radford, R. Caselli, et al., Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration, Cognitive and Behavioral Neurology 23 (2010) 165.

[48] C. Rao, V.N. Gudivada, Computational Analysis and Understanding of Natural Languages: Principles, Elsevier, Methods and Applications, 2018.

[49] J. Eisenstein, Introduction to natural language processing, Mit Press, 2019.

[50] N. Hardeniya, NLTK essentials, Packt Publishing Ltd (2015).

[51] J. Kruczek, P. Kruczek, and M. Kuta, "Are n-gram Categories Helpful in Text Classification?," in International Conference on Computational Science, 2020, pp. 524-537.

[52] Y. HaCohen-Kerner, Z. Ido, and R. Ya'akobov, "Stance classification of tweets using skip char ngrams," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2017, pp. 266-278.

[53] S. Gupta and R. Sedamkar, "Machine Learning for Healthcare: Introduction," in Machine Learning with Health Care Perspective, ed: Springer, 2020, pp. 1-25.

[54] M. Labani, P. Moradi, F. Ahmadizar, M. Jalili, A novel multivariate filter method for feature selection in text classification problems, Engineering Applications of Artificial Intelligence 70 (2018) 25–37.

[55] E. Hancer, B. Xue, M. Zhang, Differential evolution for filter feature selection based on information theory and feature ranking, Knowledge-Based Systems 140 (2018) 103–119.

[56] G. Forman, An extensive empirical study of feature selection metrics for text classification, Journal of machine learning research 3 (2003) 1289–1305.

[57] I. Jain, V.K. Jain, R. Jain, Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification, Applied Soft Computing 62 (2018) 203–215.

[58] D. Chutia, D.K. Bhattacharyya, J. Sarma, P.N.L. Raju, An effective ensemble classification framework using random forests and a correlation based feature selection technique, Transactions in GIS 21 (2017) 1165–1178.

[59] M.M. Mukaka, A guide to appropriate use of correlation coefficient in medical research, Malawi medical journal 24 (2012) 69–71.

[60] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on pattern analysis and machine intelligence 27 (2005) 1226–1238.

[61] J. Xu, B. Tang, H. He, H. Man, Semisupervised feature selection based on relevance and redundancy criteria, IEEE transactions on neural networks and learning systems 28 (2016) 1974–1984.

[62] Y. Mu, X. Liu, L. Wang, A Pearson's correlation coefficient based decision tree and its parallel implementation, Information Sciences 435 (2018) 40–58.

[63] X. Su, L. Li, F. Shi, H. Qian, Research on the fusion of dependent evidence based on mutual information, IEEE Access 6 (2018) 71839–71845.

[64] R. Smith, A mutual information approach to calculating nonlinearity, Stat 4 (2015) 291–303.

[65] A.K. Uysal, S. Gunal, A novel probabilistic feature selection method for text classification, Knowledge-Based Systems 36 (2012) 226–235.

[66] M.L. McHugh, The chi-square test of independence, Biochemia medica: Biochemia medica 23 (2013) 143–149.

[67] Y. Zhai, W. Song, X. Liu, L. Liu, X. Zhao, A Chi-Square Statistics Based Feature Selection Method in Text Classification, in: in *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, 2018, pp. 160–163.

[68] S.O. Orimaye, K.Y. Tai, J.S.-M. Wong, C.P. Wong, Learning linguistic biomarkers for predicting mild cognitive impairment using compound skip-grams, Workshop on Machine Learning in Healthcare, 2015.

[69] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, et al., "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2014, pp. 27-37.

[70] K. M. Ting and I. H. Witten, "Stacked Generalization: when does it work?," 1997.

[71] M. Kuhn and K. Johnson, "Over-fitting and model tuning," in Applied predictive modeling, ed: Springer, 2013, pp. 61-92.

[72] J.N. Mandrekar, Receiver operating characteristic curve in diagnostic test assessment, Journal of Thoracic Oncology 5 (2010) 1315–1316.

[73] F. Di Palo, N. Parde, in: Enriching Neural Models with Targeted Features for Dementia Detection, Student Research Workshop, 2019, pp. 302–308.

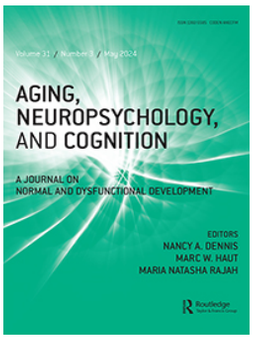
[74] J. Chen, J. Zhu, J. Ye, An Attention-Based Hybrid Network for Automatic Detection of Alzheimer's Disease from Narrative Speech, Proc. Interspeech 2019 (2019) 4085–4089.

[75] J. Akosa, "Predictive accuracy: a misleading performance measure for highly imbalanced data," in Proceedings of the SAS Global Forum, 2017, pp. 2-5.

[76] Z. Yu, G. Liu, Sliced recurrent neural networks, in: in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 2953–2964.

[77] S. Bhatt, E. Cameron, S.R. Flaxman, D.J. Weiss, D.L. Smith, P.W. Gething, Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization, Journal of The Royal Society Interface 14 (2017) 20170520.

[78] H.R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, et al., Improving computer-aided detection using convolutional neural networks and random view aggregation, IEEE transactions on medical imaging 35 (2015) 1170–1181.
[79] J. Shi, S. Zhou, X. Liu, Q. Zhang, M. Lu, T. Wang, Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset, Neurocomputing 194 (2016) 87–94.
[80] A. Karakostas, A. Briassouli, K. Avgerinakis, I. Kompatsiaris, and M. Tsolaki, "The dem@ care experiments and datasets: a technical report," arXiv preprint arXiv: 1701.01142, 2016.



Routledge
Taylor & Francis Group

Aging, Neuropsychology, and Cognition

A Journal on Normal and Dysfunctional Development

ISSN: 1382-5585 (Print) 1744-4128 (Online) Journal homepage: www.tandfonline.com/journals/nanc20

Distinguishable features of spontaneous speech in Alzheimer's clinical syndrome and healthy controls

Erin Burke, John Gunstad, Olesia Pavlenko & Phillip Hamrick

To cite this article: Erin Burke, John Gunstad, Olesia Pavlenko & Phillip Hamrick (2024) Distinguishable features of spontaneous speech in Alzheimer's clinical syndrome and healthy controls, Aging, Neuropsychology, and Cognition, 31:3, 575-586, DOI: 10.1080/13825585.2023.2221020

To link to this article: https://doi.org/10.1080/13825585.2023.2221020

AGING, NEUROPSYCHOLOGY, AND COGNITION
2024, VOL. 31, NO. 3, 575–586
https://doi.org/10.1080/13825585.2023.2221020

Routledge
Taylor & Francis Group

Distinguishable features of spontaneous speech in Alzheimer's clinical syndrome and healthy controls

Erin Burke[a], John Gunstad[a], Olesia Pavlenko[b] and Phillip Hamrick[a]

[a]Department of Psychological Sciences, Kent State University, Kent, Ohio, United States of America;
[b]Department of English, Kent State University, Kent, Ohio, United States of America

ABSTRACT

There is growing evidence that subtle changes in spontaneous speech may reflect early pathological changes in cognitive function. Recent work has found that lexical-semantic features of spontaneous speech predict cognitive dysfunction in individuals with mild cognitive impairment (MCI). The current study assessed whether Ostrand and Gunstad's (OG) lexical-semantic features extend to predicting cognitive status in a sample of individuals with Alzheimer's clinical syndrome (ACS) and healthy controls. Four additional (New) speech indices shown to be important in language processing research were also explored in this sample to extend prior work. Speech transcripts of the Cookie Theft Task from 81 individuals with ACS (M_{age} = 72.7 years, SD = 8.80, 70.4% female) and 61 healthy controls (HC) (M_{age} = 63.9 years, SD = 8.52, 62.3% female) from Dementia Bank were analyzed. Random forest and logistic machine learning techniques examined whether subject-level lexical-semantic features could be used to accurately discriminate those with ACS from HC. Results showed that logistic models with the New lexical-semantic features obtained good classification accuracy (78.4%), but the OG features had wider success across machine learning model types. In terms of sensitivity and specificity, the random forest model trained on the OG features was the most balanced. Findings from the current study suggest that features of spontaneous speech used to predict MCI may also distinguish between individuals with ACS and healthy controls. Future work should evaluate these lexical-semantic features in pre-clinical persons to further explore their potential to assist with early detection through speech analysis.

ARTICLE HISTORY
Received 20 January 2023
Accepted 29 May 2023

KEYWORDS
Spontaneous speech; Alzheimer's disease (AD); Alzheimer's clinical syndrome; machine learning

Introduction

Machine learning techniques have become increasingly utilized to examine spontaneous speech in persons with Mild Cognitive Impairment (MCI) and Alzheimer's disease (AD) (Clarke, Foltz, & Garrard, 2020; Fraser et al., 2016; Hernández-Domínguez et al., 2018; Lindsay et al., 2021; Rentoumi et al., 2014), as disruptions across multiple linguistic levels (i.e., phonetic, phonological, lexical-semantics, morphosyntactic, and pragmatic) are well-established in these conditions (Bayles et al., 1992; Boschi et al., 2017; Fleming & Harris, 2008; Ostrand & Gunstad, 2021; Pistono et al., 2016, 2019; Roark et al., 2011; Szatloczki

CONTACT Phillip Hamrick phamric1@kent.edu

et al., 2015; Taler & Phillips, 2008). Specifically, word finding difficulty (Nelson & O'Connor, 2008; Yeung et al., 2021) and decreased lexical diversity (Ostrand & Gunstad, 2021) are characteristic features of MCI, and additional declines in semantic, syntactic, and lexical functions are found as these individuals progress to AD (Ahmed et al., 2013). AD is further characterized by decreased semantic content and syntactical complexity, empty speech (i.e., a lack of descriptive specificity and the use of "thing," or "stuff") (Forbes McKay et al., 2013), reduced speech rate (Hoffmann et al., 2010), greater use of pronouns and indefinite terms (Ahmed et al., 2013; Lai, 2014) and repetition (Sajjadi et al., 2012).

Recent work raises the possibility that these linguistic indices derived from spontaneous speech might be more sensitive to pathological cognitive decline than traditional language tests of confrontation naming and semantic fluency (Bird et al., 2000; Forbes McKay et al., 2013; Henry et al., 2004; Kavé & Levy, 2003; Loewenstein et al., 2018; Mueller et al., 2016; Ostrand & Gunstad, 2021; Sajjadi et al., 2012; Slegers et al., 2018; Szatloczki et al., 2015; Taler & Phillips, 2008; Vonk et al., 2022). However, despite being identified as an important next step for the field (e.g., Boschi et al., 2017; Bucks et al., 2000; Filiou et al., 2020; Mueller et al., 2016), very few studies have examined whether lexical-semantic features found to be sensitive to MCI can be cross validated to AD samples. This lack of replication is a limitation to understanding how speech indices change over time in the context of cognitive decline.

The current study had two primary goals. First, lexical-semantic features found to distinguish persons with MCI from healthy controls in past work (Ostrand & Gunstad, 2021) were examined in a new sample of persons with Alzheimer's clinical syndrome to determine the stability and generalizability of these features. In addition, several other lexical-semantic features not previously investigated in neurodegenerative populations were utilized to help provide a more comprehensive assessment of spontaneous speech in this population. Although conceptually similar to some of the OG features, these novel features have been shown to be particularly powerful in other studies of language in AD (e.g., Hernández-Domínguez et al., 2018) and in language processing more broadly (e.g., Hamrick & Pandža, 2020). It was hypothesized that the lexical-semantic features shown to be sensitive to MCI would also be sensitive to Alzheimer's clinical syndrome and that the inclusion of additional speech indices of lexical-semantic functioning would increase predictive validity. If confirmed, these findings would provide further evidence for the idea that spontaneous speech assessment can be useful in differentiating between normative speech changes and neurodegenerative conditions.

Methods

Corpus

We used publicly available data from the Pitt Corpus (Becker et al., 1994) of the DementiaBank database. This corpus contains manual transcriptions of participants who completed a standard version of the Cookie Theft picture description task from the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983). This archive contains data from healthy controls, subjects with a variety of dementias, including "probable" and "possible" AD, MCI, vascular dementia, and Parkinson's disease. Participant history of cognitive and functional decline obtained via extensive interviews, medical workup,

and neuropsychological performance completed at the time of assessment were used to inform diagnostic groups (see Becker et al., 1994 for more details regarding diagnostic procedures). To avoid confounds from other diagnoses on language production, we only examined data from participants who were either labeled by the original researchers as "probable AD" with no other diagnoses ($n = 81$) and healthy controls ($n = 61$), and biodata for these participants can be found in Table 1. Since there is a lack of reported biomarkers associated with the "probable AD" classification in the Pitt Corpus, we use the term "Alzheimer's clinical syndrome" (ACS) here instead of "AD" or "probable AD." Moreover, because the aim of this research is to ultimately be able to predict AD as early as possible, we opted to focus on participants' first visit transcripts rather than looking at all their transcripts across several visits.

Within this sample, healthy controls (63.90 ± 8.53 years) were younger than persons in the ACS group (72.70 ± 8.80); $t(140) = 5.98$, $p < .001$), but did not differ in gender [62.3% women vs 70.4% women $t(140) = 1.009$, $p = .32$]. Given this pattern, the potential contribution of age to study findings was directly examined as part of analysis.

Lexical-semantic features

We were able to generate 13 of the 16 lexical-semantic features computed by Ostrand and Gunstad (2021). Three of their lexical-semantic features (Empty Words, Speech Rate, Filler Rate) were not able to be computed on the Pitt Corpus data due to missing information. The 13 variables included were Total Number of Words, Filler Words, Word Frequency, Type-Token Ratio, Honore's Statistic, Brunet's Index, Definite Articles, Content Words, Indefinite Articles, Pronouns, Nouns, Verbs, and Determiners. Their definitions and the details of their computation can be found in Table 2 and will be hereafter be referred to as the OG features.

In addition to attempting to replicate the OG features in persons with ACS, we also included four variables (New features) which have been shown to be useful in both AD research as well as psycholinguistic research. These four lexical-semantic features were: Hapax Legomena, HD Type-Token Ratio (computed following Hernández-Domínguez et al., 2018), Root Type-Token Ratio, and Semantic Distinctiveness. Hapax Legomena refers to the number of words (lemmas) used only once within a participant's speech transcript, with a larger value reflecting greater lexical diversity. HD Type-Token Ratio was computed following Hernández-Domínguez et al. (2018) and was the ratio of each participant's Hapax Legomena score to their total number of unique lemmas. Root Type-Token Ratio was computed as the number of unique word types divided by the square root of the total number of tokens. Finally, semantic distinctiveness was introduced as an alternative word frequency metric. In short, traditional word frequency metrics are built around the idea that a word's strength in memory is a function of the number of times it is repeated; however, increasing evidence suggests that the number of contexts in which

Table 1. Biographical data M(SD).

	Age at first visit	MMSE Score	Sex (female/male)	Education
ACS	72.70 (8.80)	19.42 (4.47)	57/24	11.80 (2.91)
Healthy Controls	63.90 (8.52)	29.08 (1.05)	38/23	13.89 (2.30)

Abbreviations: ACS, Alzheimer's clinical syndrome; MMSE, Mini Mental State Examination.

Table 2.

OG Speech Indices	Operational Definition
Total Words	total number of words spoken by the subject
Fillers	number filler words (e.g., um, uh, hmm) spoken by the subject, scaled by the total word count
Definites	total number of definite articles (the), scaled by the total word count
Indefinites	total number of indefinite articles (a, an), scaled by the total word count
Pronouns	number of pronouns (calculated by the Penn Treebank POS tags), scaled by the total word count
Nouns	number of nouns (calculated by the Penn Treebank POS tags), scaled by the total word count
Verbs	number of verbs (calculated by the Penn Treebank POS tags), scaled by the total word count
Determiners	number of determiners (as calculated by the Penn Treebank POS tags), scaled by the total word count
Content Words	number of content words (defined by the words NOT in NLTK's list of stop words), scaled by word count
Frequency	mean of the log of the frequency of all the words spoken by the subject
Token Ratio	number of different word types accounting for total number of words (a measure of vocabulary size and lexical richness)
Honore's Statistic	a measure of lexical richness/diversity (number of words produced exactly once). It is calculated as: (100 * log(tokens))/(1 − V1/types), where V1 is the number of words spoken exactly once
Brunet's index	a length-insensitive measure of lexical diversity/richness. It is calculated as: tokens ^ types ^ (−.165)
New Speech Indices	Operational Definition
Hapax Legomena	another way of measuring lexical richness of vocabulary (simple count of the number of words that are produced exactly once)
HD Type-Token Ratio	ratio of each participant's Hapax Legomena score to the total number of unique base words
Root Type Token Ratio	number of unique words divided by the square root of the total number of words (like TTR, but addresses text length variation)
Semantic Distinctiveness	a measure of semantic diversity of a word (number of different semantic contexts in which a word appears). SD = e−λ*cos(context,wordi)

a word occurs is a better predictor (Adelman et al., 2006) of a word's strength in memory, and this value is made even more predictive if it is weighted by the semantic distinctiveness of those contexts (Jones et al., 2012), and this holds in monolinguals (Jones et al., 2012), bilinguals (Hamrick & Pandža, 2020), and in aging (Johns et al., 2016). These lexical-semantic features will hereafter be referred to as the "New" features.

Classification modeling with machine learning

Machine learning is a subfield of artificial intelligence that employs automatic model construction to solve, among other things, classification problems. In this study, we employed two popular model types: logistic regression and random forest classifiers.[1] The code and data for these models is available at https://dementia.talkbank.org/. We used two types of classifiers because each has complementary strengths and weaknesses. Moreover, if we found similar results across the classifier types, it could be taken as evidence of more robustness in the roles of our lexical-semantic features in ACS classification. The model training and testing procedure was as follows. The entire dataset consisted of subject-level values for the 17 lexical-semantic features along with each participant's diagnosis data (ACS vs HC). The entire dataset from all 142 participants was randomly split into two sets: a training/evaluation dataset (75% of all data) and a testing dataset (the remaining 25% of the data). The training/evaluation dataset was then randomly divided into 25 bootstrap resampled datasets. Within each of these resample training/evaluation datasets, we bootstrap resampled another 75/25% split, with 75% of each resample being used for model training purposes and the other 25% being used to

evaluate the overall model fit. This procedure was repeated across each of the 25 training/evaluation datasets. In the final step of modeling, we then tested the best overall model from these fits against the testing dataset, which consisted exclusively of data that had been held aside at the outset and that none of the models had ever been exposed to prior. Note, that this modeling procedure establishes thresholds for classification of participants as ACS and HC based on the models that best fit the training data, and no further model tuning or boosting was used to alter these thresholds, so as to avoid potential issues with overfit. In all cases, resampling was stratified to preserve similar proportions of ACS and HC participants in all datasets. These final model fits to the test data set were evaluated based on the accuracy of classification (ACS vs HC) as well as ROC-AUC. All analyses were conducted using the tidymodels package (Kuhn & Wickham, 2020) within the R statistical programming language (R Core Team, 2021), and the code for reproducing this modeling procedure is publicly available (https://osf.io/qc3fw/?view_only=8406d0372b68472486f915ce25e455cf).

Results

The models' abilities to accurately classify test (i.e., generalization) participants as either ACS or HC were evaluated on the basis of their accuracy, area under the curve of receiver operating characteristics (AUC, see Figures 1 and 2), their sensitivity (i.e., the % of time the model classifies ACS correctly), and their specificity (i.e., the % of time HC was classified correctly). The results are shown in Table 3, which shows better performance for random

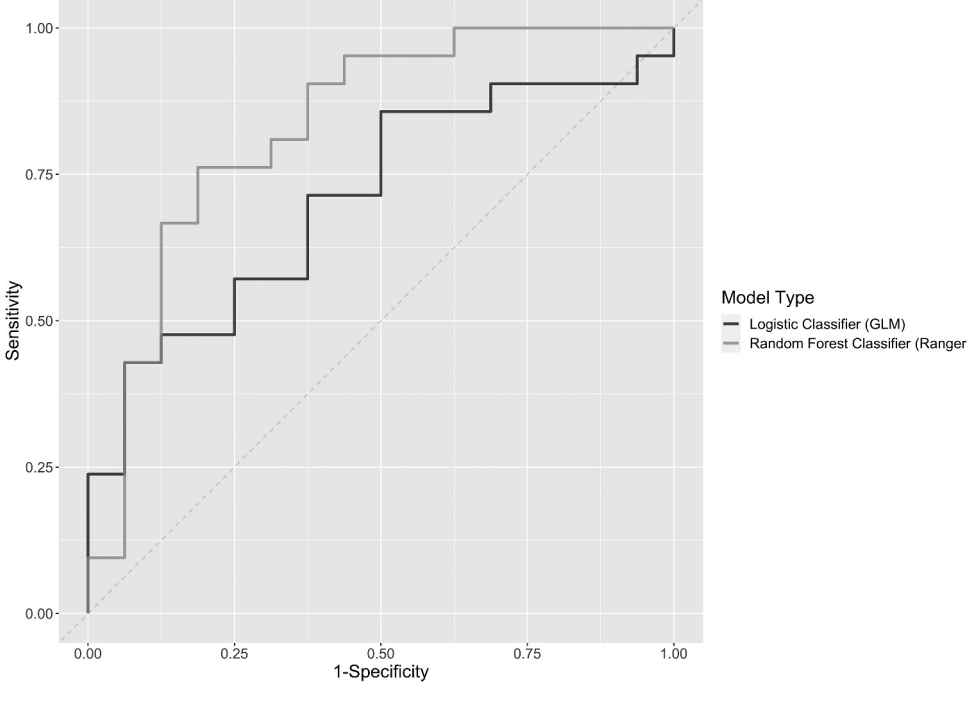


Figure 1. ROC-AUC for the logistic and random forest classifier models for the OG features.

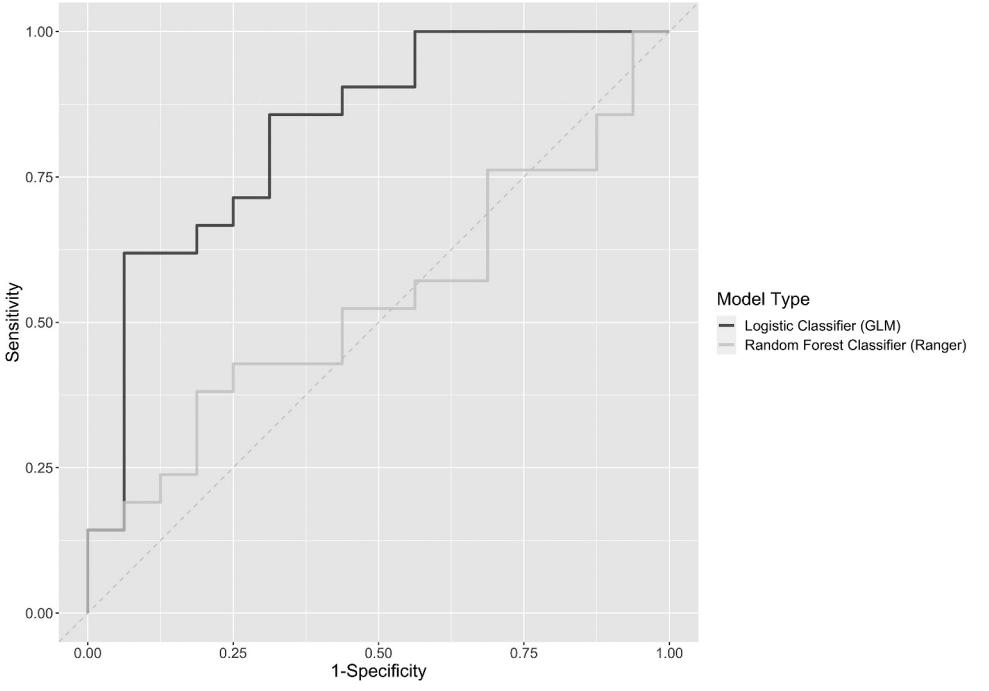


Figure 2. ROC-AUC for the logistic and random forest classifier models for the New features.

Table 3. Performance of classifier models at distinguishing ACS patients and HCs.

Model Type	Features	Accuracy	AUC	Sensitivity	Specificity
Logistic	OG	0.649	0.708	0.571	.750
Random Forests	OG	0.757	0.804	0.761	.750
Logistic	New	0.784	0.830	0.857	.687
Random Forests	New	0.486	0.533	0.523	.437

Abbreviations: ACS, Alzheimer's clinical syndrome; AUC = area under the curve of receiver operating characteristics; HCs, healthy elderly controls; OG = 13 semantic features used in Ostrand and Gunstad (2021) re-computed over the current dataset.

forest over logistic models for the OG lexical-semantic features, as well as better performance for logistic models over random forest models in the New lexical-semantic features.

Accuracy and AUC values were highest for the logistic model based on the New lexical-semantic features, suggesting that those lexical-semantic features may have particularly promising value for future research (see Figure 1 for a visual of these findings). At the same time, the model with the most balanced sensitivity and specificity was the random forest model trained on the OG lexical-semantic features.

Because age differed between the ACS and HC groups, we also ran the entire modeling procedure described above while including age as a covariate in the models (see Table 4). This approach resulted in a similar overall pattern of findings (e.g., random forest models were superior to logistic models for the OG features, but the converse was true for the

Table 4. Performance of classifier models at distinguishing ACS patients and HCs, with age included as a covariate in the model.

Model Type	Features	Accuracy	AUC	Sensitivity	Specificity
Logistic	OG	0.703	0.810	0.667	.750
Random Forests	OG	0.865	0.836	0.857	.875
Logistic	New	0.703	0.839	0.761	.625
Random Forests	New	0.595	0.685	0.571	.625

Abbreviations: ACS, Alzheimer's clinical syndrome; AUC = area under the curve of receiver operating characteristics; HCs, healthy elderly controls; OG = 13 semantic features used in Ostrand and Gunstad (2021) re-computed over the current dataset.

New features), with generally better fit to the data – which is unsurprising given that age is correlated with ACS diagnosis.

Discussion

Findings from the current study both replicate and extend past work. Features of spontaneous speech shown to distinguish persons with MCI from healthy older adults in past work (Ostrand & Gunstad, 2021) successfully identified persons with ACS in the current sample. The inclusion of additional indices from more recent work in language processing research increased predictive validity over previously used lexical-semantic features. Several aspects of these findings warrant brief discussion.

Finding that OG speech indices successfully predicted group status in the current study (i.e., ACS vs healthy control) is a modest but valuable contribution to the literature. As described above, though a growing number of studies have identified lexical-semantic features that can distinguish persons with and without neurological conditions (e.g., Boschi et al., 2017; Fraser et al., 2016; Ostrand & Gunstad, 2021; Roark et al., 2011), very few studies have examined whether the same set of speech features found to be sensitive to impairment in one sample generalize to another (Bucks et al., 2000). This makes interpretation of inconsistent findings difficult, as differences across studies may then be due to either conceptual issues (e.g., incomplete or incorrect understanding of language in AD) or any number of methodological choices (e.g., specific lexical-semantic features being examined, sample composition, statistical techniques, etc.). Our results showing lexical-semantic features that predict MCI also predict ACS support the notion that these indices have the capacity to detect lexical-semantic features characteristic of cognitive decline (Ahmed et al., 2013; Forbes McKay et al., 2013; Ostrand & Gunstad, 2021) and that MCI and ACS share similar speech characteristics (Boschi et al., 2017; Taler & Phillips, 2008). Future work should continue to cross-validate lexical-semantic indices across samples to further support their use in identifying changes in speech associated with neurodegenerative conditions.

In addition to generalizing findings from past work, the current analyses also revealed that four other markers of lexical and semantic performance were sensitive to ACS and distinguished these persons from healthy controls. These indices (i.e., Hapax Legomena, HD Type-Token Ratio, Root Type-Token Ratio, and Semantic Distinctiveness) are similar to the OG indices in that they also measure lexical-semantic diversity in spoken language. The fact that these lexical-semantic features were also disrupted in the context of neurodegeneration represents an important confirmation that both MCI and ACS affect

lexical-semantic processing broadly, not just for some lexical-semantic features. Such findings are consistent with past work showing numerous aspects of speech and language function may rely on the declarative memory system (Hamrick et al., 2018) that are disrupted in persons with MCI/AD and encourage continued examination of other elements including acoustic features such as voice quality and pauses (Gosztolya et al., 2019; Pistono et al., 2016; Themistocleous et al., 2020).

The fact that the lexical-semantic features had different levels of classification performance is not surprising, as logistic and random forest models often differ in their success rates, largely as a function of the datasets used, the nature of the predictor variables and their multicollinearity, and a host of other factors (Couronné et al., 2018). Future research should examine the differential efficacy of random forest and logistic models (as well as other common machine learning techniques, such as support vector machines) in using lexical-semantic features to predict AD classification, especially since modern machine learning techniques involve a range of tuning parameters. At their best, principle tuning of these parameters may result in optimal classification performance, but at their worst, parameter tuning could be fishing for statistical significance. More research on this issue is needed.

Using existing, archived data limited the characterization of diagnostic groups. Cognition of participants was not comprehensively assessed and biomarker confirmation of diagnostic groups was not available from that original project, but would serve as key components to further understanding in future work (Graff-Radford et al., 2021). For example, individuals with AD recruit different brain regions on tasks of semantic and working memory compared to healthy controls (Hirni et al., 2013; Teipel et al., 2015), raising the possibility that neuroimaging correlates of spontaneous speech may be detectable at various disease stages. Using neuroimaging to investigate the association between atrophy or amyloid deposits to spontaneous speech performance may shed key insight into the biological representations of changes in spontaneous speech. Some evidence for these relationships already exists. Past work reveals atrophy in cortical regions critical for language processes (i.e., naming, semantic fluency and word retrieval) such as the hippocampus and Broadmann's areas 37 and 40, as well as reduced activation in prefrontal and parietal regions (Harasty et al., 1999; Hirni et al., 2013; McGeown et al., 2009; Venneri et al., 2008). Reduced activation and gray matter volume existing in these language areas may reflect impaired spontaneous speech that occurs in relation to deficits in semantic memory seen in AD (Zannino et al., 2015).

Future work should also prospectively investigate the sensitivity and specificity of these lexical-semantic features to AD in larger, age- and demographically-balanced samples. Such work would provide key insight into speech changes associated with normal and pathological cognitive aging, as well as provide a better understanding of possible influence of age, gender, and race/ethnicity. For example, the wide age range of the current sample (i.e., 47 to 88) limits the opportunity to generate insight into speech within narrow age ranges (e.g., 40–50 years of age vs 50–60 years of age). Further, the lack of racial/ethnic diversity in the current sample raises concerns for understanding known cultural influences on language (Gutchess et al., 2010). Similarly, further work into possible sex differences is also needed, as women exhibit greater atrophy in brain regions important for language (Harasty et al., 1999) and the potential contribution of these changes

to spontaneous speech is poorly understood. Additionally, though picture description tasks such as Cookie Theft are successful in revealing linguistic impairments (Cummings, 2019), it is possible that other tasks may generate speech samples even more sensitive to AD. Picture description tasks are highly structured and rely on a lower cognitive load relative to expository speech tasks (i.e., responding to open-ended questions in an interview or telling a story). Examination of the ability of the speech indices utilized in the current study to predict ACS status from other speech samples appears warranted.

In brief summary, lexical-semantic features sensitive to MCI in past work were also sensitive to ACS in the current sample and novel markers of lexical and semantic distinctiveness showed incremental validity in the identification of persons with ACS. The current findings encourage further examination of the possible utility of automated lexical-semantic analyses to aid in early detection of MCI and AD.

Note

1. These two methods were used as complementary. Some studies have suggested superior performance for random forest over logistic models (Maroco et al., 2011), while others have shown better performance in logistic models (Kirasich et al., 2018). Because random forest models are difficult to interpret in terms of the structure of their resulting models and involve fine-tuning of parameters (which we left standard in our analyses, including the default number of trees), we opted to also use logistic models, given that they are more interpretable and may provide, in some cases, superior fit.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work was supported by the National Institute on Aging [AG03705, AG05133]; National Institutes of Health [R01AG065432]. Data for the current study was extracted from DementiaBank, which is funded by NIH grants NIA AG03705 and AG05133. Dr. Gunstad funded in part by NIH (NIA R01AG065432).

ORCID

Phillip Hamrick http://orcid.org/0000-0003-4910-5455

References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Ahmed, S., Haigh, A. M., de Jager, C. A., & Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain: A Journal of Neurology*, *136*(Pt 12), 3727–3737. https://doi.org/10.1093/brain/awt269

Bayles, K. A., Tomoeda, C. K., & Trosset, M. W. (1992). Relation of linguistic communication abilities of Alzheimer's patients to stage of disease. *Brain and Language*, *42*(4), 454–472. https://doi.org/10.1016/0093-934x(92)90079-t

Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, *51*(6), 585–594. https://doi.org/10.1001/archneur.1994.00540180063015

Bird, H., Ralph, M. A. L., Patterson, K., & Hodges, J. R. (2000). The rise and fall of frequency and imageability: Noun and verb production in semantic dementia. *Brain and Language*, *73*(1), 17–49. https://doi.org/10.1006/brln.2000.2293

Boschi, V., Catricalà, E., Consonni, M., Chesi, C., Moro, A., & Cappa, S. F. (2017). Connected speech in neurodegenerative language disorders: A review. *Frontiers in Psychology*, *8*, 269. https://doi.org/10.3389/fpsyg.2017.00269

Bucks, R. S., Singh, S., Cuerden, J. M., & Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, *14*(1), 71–91. https://doi.org/10.1080/026870300401603

Clarke, N., Foltz, P., & Garrard, P. (2020). How to do things with (thousands of) words: Computational approaches to discourse analysis in Alzheimer's disease. *Cortex*, *129*, 446–463. https://doi.org/10.1016/j.cortex.2020.05.001

Couronné, R., Probst, P., & Boulesteix, A. L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, *19*(1), 270. https://doi.org/10.1186/s12859-018-2264-5

Cummings, L. (2019). Describing the cookie theft picture: Sources of breakdown in alzheimer's dementia. *Pragmatics and Society*, *10*(2), 153–176. https://doi.org/10.1075/ps.17011.cum

Filiou, R. P., Bier, N., Slegers, A., Houzé, B., Belchior, P., & Brambati, S. M. (2020). Connected speech assessment in the early detection of Alzheimer's disease and mild cognitive impairment: A scoping review. *Aphasiology*, *34*(6), 723–755. https://doi.org/10.1080/02687038.2019.1608502

Fleming, V. B., & Harris, J. L. (2008). Complex discourse production in mild cognitive impairment: Detecting subtle changes. *Aphasiology*, *22*(7–8), 729–740. https://doi.org/10.1080/02687030701803762

Forbes McKay, K., Shanks, M. F., & Venneri, A. (2013). Profiling spontaneous speech decline in Alzheimer's disease: A longitudinal study. *Acta Neuropsychiatrica*, *25*(6), 320–327. https://doi.org/10.1017/neu.2013.16

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease: JAD*, *49*(2), 407–422. https://doi.org/10.3233/JAD-150520

Goodglass, H., & Kaplan, E. (1983). *Boston diagnostic aphasia examination booklet*. Lea & Febiger.

Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J., & Hoffmann, I. (2019). Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. *Computer Speech & Language*, *53*, 181–197. https://doi.org/10.1016/j.csl.2018.07.007

Graff-Radford, J., Yong, K. X. X., Apostolova, L. G., Bouwman, F. H., Carrillo, M., Dickerson, B. C., Rabinovici, G. D., Schott, J. M., Jones, D. T., & Murray, M. E. (2021). New insights into atypical Alzheimer's disease in the era of biomarkers. *Lancet Neurology*, *20*(3), 222–234. https://doi.org/10.1016/S1474-4422(20)30440-3

Gutchess, A. H., Hedden, T., Ketay, S., Aron, A., & Gabrieli, J. D. (2010). Neural differences in the processing of semantic relationships across cultures. *Social Cognitive and Affective Neuroscience*, *5*(2–3), 254–263. https://doi.org/10.1093/scan/nsp059

Hamrick, P., Lum, J. A. G., & Ullman, M. T. (2018). Child first language and adult second language are both tied to general-purpose learning systems. *Proceedings of the National Academy of Sciences*, *115*(7), 1487–1492. https://doi.org/10.1073/pnas.1713975115

Hamrick, P., & Pandža, N. B. (2020). Contributions of semantic and contextual diversity to the word frequency effect in L2 lexical access. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, *74*(1), 25–34. https://doi.org/10.1037/cep0000189

Harasty, J. A., Halliday, G. M., Kril, J. J., & Code, C. (1999). Specific temporoparietal gyral atrophy reflects the pattern of language dissolution in Alzheimer's disease. *Brain: A Journal of Neurology*, *122*(*Pt 4*), 675–686. https://doi.org/10.1093/brain/122.4.675

Henry, J. D., Crawford, J. R., & Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: A meta-analysis. *Neuropsychologia*, *42*(9), 1212–1222. https://doi.org/10.1016/j.neuropsychologia.2004.02.001

Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G., & Roche-Bergua, A. (2018). Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's & Dementia (Amsterdam, Netherlands)*, *10*(1), 260–268. https://doi.org/10.1016/j.dadm.2018.02.004

Hirni, D. I., Kivisaari, S. L., Monsch, A. U., & Taylor, K. I. (2013). Distinct neuroanatomical bases of episodic and semantic memory performance in Alzheimer's disease. *Neuropsychologia*, *51*(5), 930–937. https://doi.org/10.1016/j.neuropsychologia.2013.01.013

Hoffmann, I., Nemeth, D., Dye, C. D., Pákáski, M., Irinyi, T., & Kálmán, J. (2010). Temporal parameters of spontaneous speech in Alzheimer's disease. *International Journal of Speech-Language Pathology*, *12*(1), 29–34. https://doi.org/10.3109/17549500903137256

Johns, B. T., Sheppard, C. L., Jones, M. N., & Taler, V. (2016). The role of semantic diversity in word recognition across Aging and Bilingualism. *Frontiers in Psychology*, *7*, 703. https://doi.org/10.3389/fpsyg.2016.00703

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology = Revue Canadienne de Psychologie Experimentale*, *66*(2), 115–124. https://doi.org/10.1037/a0026727

Kavé, G., & Levy, Y. (2003). Morphology in picture descriptions provided by persons with Alzheimer's disease. *Journal of Speech, Language, & Hearing Research*, *46*(2), 341–352. https://doi.org/10.1044/1092-4388(2003/027)

Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: Binary classification for heterogeneous datasets. *SMU Data Science Review*, *1*(3), article 9 Available at: https://scholar.smu.edu/datasciencereview/vol1/iss3/9

Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles*. https://www.tidymodels.org/

Lai, Y. H. (2014). Discourse features of Chinese-speaking seniors with and without Alzheimer's disease. *Language and Linguistics*, *15*(3), 411–434. https://doi.org/10.1177/1606822X14520665

Lindsay, H., Tröger, J., & König, A. (2021). Language Impairment in Alzheimer's disease-robust and explainable evidence for AD-Related deterioration of spontaneous speech through multilingual machine learning. *Frontiers in Aging Neuroscience*, *13*, 642033. https://doi.org/10.3389/fnagi.2021.642033

Loewenstein, D. A., Curiel, R. E., Duara, R., & Buschke, H. (2018). Novel cognitive paradigms for the detection of memory impairment in preclinical Alzheimer's disease. *Assessment*, *25*(3), 348–359. https://doi.org/10.1177/1073191117691608

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & Mendonça, A. (2011). Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes*, *4*(1), 299. https://doi.org/10.1186/1756-0500-4-299

McGeown, W. J., Shanks, M. F., Forbes McKay, K. E., & Venneri, A. (2009). Patterns of brain activity during a semantic task differentiate normal aging from early Alzheimer's disease. *Psychiatry Research*, *173*(3), 218–227. https://doi.org/10.1016/j.pscychresns.2008.10.005

Mueller, K. D., Koscik, R. L., Turkstra, L. S., Riedeman, S. K., LaRue, A., Clark, L. R., Hermann, B., Sager, M. A., & Johnson, S. C. (2016). Connected language in late middle-aged adults at risk for Alzheimer's disease. *Journal of Alzheimer's Disease*, *54*(4), 1539–1550. https://doi.org/10.3233/JAD-160252

Nelson, A. P., & O'Connor, M. G. (2008). Mild cognitive impairment: A neuropsychological perspective. *CNS Spectrums*, *13*(1), 56–64. https://doi.org/10.1017/s1092852900016163

Ostrand, R., & Gunstad, J. (2021). Using automatic assessment of speech production to predict current and future cognitive function in older adults. *Journal of Geriatric Psychiatry and Neurology*, *34*(5), 357–369. https://doi.org/10.1177/0891988720933358

Pistono, A., Jucla, M., Barbeau, E. J., Saint-Aubert, L., Lemesle, B., Calvet, B., Köpke, B., Puel, M., & Pariente, J. (2016). Pauses during autobiographical discourse reflect episodic memory processes in early Alzheimer's disease. *Journal of Alzheimer's Disease*, *50*(3), 687–698. https://doi.org/10.3233/JAD-150408

Pistono, A., Pariente, J., Bézy, C., Lemesle, B., Le Men, J., & Jucla, M. (2019). What happens when nothing happens? An investigation of pauses as a compensatory mechanism in early Alzheimer's disease. *Neuropsychologia*, *124*, 133–143. https://doi.org/10.1016/j.neuropsychologia.2018.12.018

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rentoumi, V., Raoufian, L., Ahmed, S., de Jager, C. A., & Garrard, P. (2014). Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. *Journal of Alzheimer's Disease: JAD*, *42*(*Suppl 3*), S3–S17. https://doi.org/10.3233/JAD-140555

Roark, B., Mitchell, M., Hosom, J. P., Hollingshead, K., & Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 2081–2090. https://doi.org/10.1109/TASL.2011.2112351

Sajjadi, S. A., Patterson, K., Tomek, M., & Nestor, P. J. (2012). Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. *Aphasiology*, *26*(6), 847–866. https://doi.org/10.1080/02687038.2012.654933

Slegers, A., Filiou, R.-P., Montembeault, M., & Brambati, S. M. (2018). Connected speech features from picture description in Alzheimer's Disease: A systematic review. *Journal of Alzheimer's Disease*, *65*(2), 519–542. https://doi.org/10.3233/JAD-170881

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., & Pakaski, M. (2015). Speaking in Alzheimer's Disease, is that an early sign? Importance of changes in language abilities in Alzheimer's Disease. *Frontiers in Aging Neuroscience*, *7*, 195. https://doi.org/10.3389/fnagi.2015.00195

Taler, V., & Phillips, N. A. (2008). Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. *Journal of Clinical and Experimental Neuropsychology*, *30*(5), 501–556. https://doi.org/10.1080/13803390701550128

Teipel, S., Ehlers, I., Erbe, A., Holzmann, C., Lau, E., Hauenstein, K., & Berger, C. (2015). Structural connectivity changes underlying altered working memory networks in mild cognitive impairment: A three-way image fusion analysis. *Journal of Neuroimaging: Official Journal of the American Society of Neuroimaging*, *25*(4), 634–642. https://doi.org/10.1111/jon.12178

Themistocleous, C., Eckerström, M., & Kokkinakis, D. (2020). Voice quality and speech fluency distinguish individuals with mild cognitive impairment from healthy controls. *PLos One*, *15*(7), e0236009. https://doi.org/10.1371/journal.pone.0236009

Venneri, A., McGeown, W. J., Hietanen, H. M., Guerrini, C., Ellis, A. W., & Shanks, M. F. (2008). The anatomical bases of semantic retrieval deficits in early Alzheimer's disease. *Neuropsychologia*, *46*(2), 497–510. https://doi.org/10.1016/j.neuropsychologia.2007.08.026

Vonk, J. M., Geerlings, M. I., Avila-Rieger, J., Qian, C. L., Schupf, N., Mayeux, R., & Manly, J. (2022). Semantic item-level metrics relate to future memory decline beyond existing cognitive tests in older adults without dementia.

Yeung, A., Iaboni, A., Rochon, E., Lavoie, M., Santiago, C., Yancheva, M., Novikova, J., Xu, M., Robin, J., Kaufman, L. D., & Mostafa, F. (2021). Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and Alzheimer's dementia. *Alzheimer's Research & Therapy*, *13*(1), 109. https://doi.org/10.1186/s13195-021-00848-x

Zannino, G. D., Caltagirone, C., & Carlesimo, G. A. (2015). The contribution of neurodegenerative diseases to the modelling of semantic memory: A new proposal and a review of the literature. *Neuropsychologia*, *75*, 274–290. https://doi.org/10.1016/j.neuropsychologia.2015.06.023

HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed

***Abstract*—Self-supervised approaches for speech representation learning are challenged by three unique problems: (1) there are multiple sound units in each input utterance, (2) there is no lexicon of input sound units during the pre-training phase, and (3) sound units have variable lengths with no explicit segmentation. To deal with these three problems, we propose the Hidden-Unit BERT (HuBERT) approach for self-supervised speech representation learning, which utilizes an offline clustering step to provide aligned target labels for a BERT-like prediction loss. A key ingredient of our approach is applying the prediction loss over the masked regions only, which forces the model to learn a combined acoustic and language model over the continuous inputs. HuBERT relies primarily on the consistency of the unsupervised clustering step rather than the intrinsic quality of the assigned cluster labels. Starting with a simple k-means teacher of 100 clusters, and using two iterations of clustering, the HuBERT model either matches or improves upon the state-of-the-art wav2vec 2.0 performance on the Librispeech (960 h) and Libri-light (60,000 h) benchmarks with 10 min, 1 h, 10 h, 100 h, and 960 h fine-tuning subsets. Using a 1B parameter model, HuBERT shows up to 19% and 13% relative WER reduction on the more challenging dev-other and test-other evaluation subsets.[1][2]**

***Index Terms*—Self-supervised learning, BERT.**

I. Introduction

THE north star for many research programs has been learning speech and audio representations through listening and interaction, similar to how babies learn their first language. High fidelity speech representation includes disentangled aspects of the spoken content along with non-lexical information of how it is delivered, e.g., speaker identity, emotion, hesitation, interruptions. Furthermore, reaching a complete situational understanding requires modeling structured noise interleaving and overlapping with the speech signal, e.g., laughter, coughing, lip-smacking, background vehicle engine, birds chirping, or food sizzling sounds.

The need for such high-fidelity representations drove research in self-supervised learning for speech and audio where the targets driving the learning process of a designed pretext task are drawn from the input signal itself. Examples of pretext tasks for self-supervised speech representation learning include distinguishing near-by features from temporally distant ones [2]–[4], next-step prediction of audio features [5], masked prediction of audio features given unmasked context [6], [7]. Besides, self-supervised learning methods do not rely on any linguistic resources during training, allowing them to learn universal representations since labels, annotations, and text-only material ignores rich information in the input signal.

Learning speech representations without reliance on large volumes of labeled data is crucial for industrial applications and products with ever-increasing coverage of new languages and domains. The time needed to collect large labeled datasets covering each of these scenarios is the real bottleneck in the current fast-moving AI industry, with time-to-market playing a critical role for product success. Building more inclusive applications covering spoken-only dialects and languages is another significant benefit of reducing dependence on linguistic resources. Given their non-standard orthographic rules, many of these languages and dialects have very little or no resources at all.

Pseudo-labeling (PL), also known as self-training, belongs to the family of semi-supervised learning techniques, and has been the dominant approach for utilizing unlabeled speech and audio with successful applications dating back to the mid-1990s [8]–[11]. PL starts with some supervised data to train a "teacher" model in one specific downstream task. Pseudo-labels are then generated for the unlabeled data using the teacher model. Next, a student model is trained using the combined supervised and teacher-labeled data either using the standard cross-entropy [10] loss or using a contrastive loss [12] to account for noise in teacher-generated labels. The pseudo-labeling process may be repeated multiple times to improve teacher label quality [13] iteratively.

Without discounting the immense success of pseudo-labeling techniques, self-supervised representations offer two unique advantages: (1) Pseudo-label methods force student models to merely mimic a teacher model, which is limited by its supervised data size and the provided annotation quality. On the other hand,

Manuscript received June 7, 2021; revised September 27, 2021; accepted October 10, 2021. Date of publication October 26, 2021; date of current version November 26, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rohit Prabhavalkar. *(Corresponding author: Wei-Ning Hsu.)*

Wei-Ning Hsu is with the Facebook Inc., New York, NY 10003 USA (e-mail: wnhsu@fb.com).

Benjamin Bolte, Kushal Lakhotia, and Abdelrahman Mohamed are with Facebook AI Research, Menlo Park, CA 94025 USA (e-mail: ben@bolte.cc; kushall@fb.com; abdo@fb.com).

Yao-Hung Hubert Tsai and Ruslan Salakhutdinov are with the Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213-3815 USA (e-mail: yaohungt@cs.cmu.edu; rsalakhu@cs.cmu.edu).

Digital Object Identifier 10.1109/TASLP.2021.3122291

[1]The code, pre-trained and fine-tuned models are available at https://github.com/pytorch/fairseq/tree/master/examples/hubert.

[2]This manuscript is an extended version of [1].

self-supervised pretext tasks force the model to represent the entire input signal by compressing many more bits of information into the learned latent representation. (2) In pseudo-labeling, the supervised data of the teacher model forces the whole learning to be geared towards a single downstream task. On the contrary, self-supervised features show better generalization to a multitude of downstream applications.

There have been impressive successes for self-supervised learning in Computer Vision (CV) [14]–[16] and Natural Language Processing (NLP) [17]–[19] applications. Learning representations of discrete input sequences, such as in Natural Language Processing (NLP) applications, uses either masked prediction [20], [21] or auto-regressive generation [19], [22] of input sequences with partial obfuscation. For continuous inputs, such as in Computer Vision (CV) applications, representations are often learned through instance classification, in which each image and its augmentations are treated as a single output class to be pulled together [15], [16] or contrasted against other negative samples [23].

Speech signals differ from text and images in that they are *continuous-valued sequences*. Self-supervised learning for the speech recognition domain faces unique challenges from those in CV and NLP. Firstly, the presence of multiple sounds in each input utterance breaks the instance classification assumption used in many CV pre-training approaches. Secondly, during pre-training, there is no prior lexicon of discrete sound units available, as in NLP applications in which words or word pieces are used, hindering the use of predictive losses. Lastly, the boundaries between sound units are not known, which complicates masked prediction pre-training.

In this paper, we introduce **H**idden **u**nit **BERT** (HuBERT) that benefits from an offline clustering step to generate noisy labels for a BERT-like per-training. Concretely, a BERT model consumes masked continuous speech features to predict predetermined cluster assignments. The predictive loss is only applied over the masked regions, forcing the model to learn good high-level representations of unmasked inputs to infer the targets of masked ones correctly. Intuitively, the HuBERT model is forced to learn both acoustic and language models from continuous inputs. First, the model needs to model unmasked inputs into meaningful continuous latent representations, which maps to the classical acoustic modeling problem. Second, to reduce the prediction error, the model needs to capture the long-range temporal relations between learned representations. One crucial insight motivating this work is the importance of consistency of the targets, not just their correctness, which enables the model to focus on modeling the sequential structure of input data. Our approach draws inspiration from the DeepCluster method for self-supervised visual learning [24]; however, HuBERT benefits from the masked prediction loss over speech sequences to represent their sequential structure.

When the HuBERT model is pre-trained on either the standard Librispeech 960h [25] or the Libri-Light 60 k hours [26], it either matches or improves upon the state-of-the-art wav2vec 2.0 [7] performance on all fine-tuning subsets of 10mins, 1 h, 10 h, 100 h, and 960 h. We present systematic results on three model sizes pre-trained with HuBERT: BASE (90 M parameters), LARGE (300 M), and X-LARGE (1B). The X-LARGE model shows up to 19% and 13% relative WER improvement from LARGE models on dev-other and test-other evaluation subsets when pre-trained on the Libri-Light 60 k hours.

TABLE I
MODEL ARCHITECTURE SUMMARY FOR BASE, LARGE, AND X-LARGE HUBERT MODELS

		BASE	LARGE	X-LARGE
CNN Encoder	strides	5, 2, 2, 2, 2, 2, 2		
	kernel width	10, 3, 3, 3, 3, 2, 2		
	channel	512		
Transformer	layer	12	24	48
	embedding dim.	768	1024	1280
	inner FFN dim.	3072	4096	5120
	layerdrop prob	0.05	0	0
	attention heads	12	16	16
Projection	dim.	256	768	1024
Num. of Params		95M	317M	964M

II. METHOD

A. Learning the Hidden Units for HuBERT

An acoustic model trained on text and speech pairs provides pseudo-phonetic labels for each frame via forced alignment in semi-supervised learning. On the contrary, the self-supervised representation learning setup has access to speech-only data. Nevertheless, simple discrete latent variable models such as k-means and Gaussian mixture models (GMMs) infer hidden units that exhibit non-trivial correlation with the underlying acoustic units [27] (see also Table V). More advanced systems can achieve better acoustic unit discovery performance using better graphical models [28], [29] or parameterizes the distributions with more powerful neural network models [30]–[34].

Inspired by this, we propose to use acoustic unit discovery models to provide frame-level targets. Let X denote a speech utterance $X = [x_1, \ldots, x_T]$ of T frames. Discovered hidden units are denoted with $h(X) = Z = [z_1, \ldots, z_T]$, where $z_t \in \{1, \ldots, C\}$ is a C-class categorical variable and h is a clustering model, e.g. k-means.

B. Representation Learning via Masked Prediction

Let $M \subset \{1, \ldots, T\}$ denote the set of indices to be masked for a length-T sequence X, and $\tilde{X} = r(X, M)$ denote a corrupted version of X where x_t is replaced with a mask embedding $\tilde{x}$ if $t \in M$. A masked prediction model f takes as input $\tilde{X}$ and predicts a distribution over the target indices at each timestep $p_f(\cdot \mid \tilde{X}, t)$. There are two decisions to be made for masked prediction: *how to mask* and *where to apply the prediction loss*.

Regarding the first decision, we adopt the same strategies used in SpanBERT [35] and wav2vec 2.0 [7] for mask generation, where $p\%$ of the timesteps are randomly selected as start indices, and spans of l steps are masked. To address the second decision, we denote the cross-entropy loss computed over masked and unmasked timesteps as L_m and L_u, respectively. L_m is defined

TABLE II
RESULTS AND COMPARISON WITH THE LITERATURE ON LOW RESOURCE SETUPS (10-MIN, 1-HOUR, 10-HOUR, AND 100-HOUR OF LABELED DATA)

Model	Unlabeled Data	LM	dev-clean	dev-other	test-clean	test-other
		10-min labeled				
DiscreteBERT [52]	LS-960	4-gram	15.7	24.1	16.3	25.2
wav2vec 2.0 BASE [7]	LS-960	4-gram	8.9	15.7	9.1	15.6
wav2vec 2.0 LARGE [7]	LL-60k	4-gram	6.3	9.8	6.6	10.3
wav2vec 2.0 LARGE [7]	LL-60k	Transformer	4.6	7.9	4.8	8.2
HUBERT BASE	LS-960	4-gram	9.1	15.0	9.7	15.3
HUBERT LARGE	LL-60k	4-gram	6.1	9.4	6.6	10.1
HUBERT LARGE	LL-60k	Transformer	**4.3**	7.0	4.7	7.6
HUBERT X-LARGE	LL-60k	Transformer	4.4	**6.1**	**4.6**	**6.8**
		1-hour labeled				
DeCoAR 2.0 [51]	LS-960	4-gram	-	-	13.8	29.1
DiscreteBERT [52]	LS-960	4-gram	8.5	16.4	9.0	17.6
wav2vec 2.0 BASE [7]	LS-960	4-gram	5.0	10.8	5.5	11.3
wav2vec 2.0 LARGE [7]	LL-60k	Transformer	2.9	5.4	2.9	5.8
HUBERT BASE	LS-960	4-gram	5.6	10.9	6.1	11.3
HUBERT LARGE	LL-60k	Transformer	**2.6**	4.9	2.9	5.4
HUBERT X-LARGE	LL-60k	Transformer	**2.6**	**4.2**	**2.8**	**4.8**
		10-hour labeled				
SlimIPL [55]	LS-960	4-gram + Transformer	5.3	7.9	5.5	9.0
DeCoAR 2.0 [51]	LS-960	4-gram	-	-	5.4	13.3
DiscreteBERT [52]	LS-960	4-gram	5.3	13.2	5.9	14.1
wav2vec 2.0 BASE [7]	LS-960	4-gram	3.8	9.1	4.3	9.5
wav2vec 2.0 LARGE [7]	LL-60k	Transformer	2.4	4.8	2.6	4.9
HUBERT BASE	LS-960	4-gram	3.9	9.0	4.3	9.4
HUBERT LARGE	LL-60k	Transformer	2.2	4.3	2.4	4.6
HUBERT X-LARGE	LL-60k	Transformer	**2.1**	**3.6**	**2.3**	**4.0**
		100-hour labeled				
IPL [13]	LL-60k	4-gram + Transformer	3.19	6.14	3.72	7.11
SlimIPL [55]	LS-860	4-gram + Transformer	2.2	4.6	2.7	5.2
Noisy Student [62]	LS-860	LSTM	3.9	8.8	4.2	8.6
DeCoAR 2.0 [51]	LS-960	4-gram	-	-	5.0	12.1
DiscreteBERT [52]	LS-960	4-gram	4.0	10.9	4.5	12.1
wav2vec 2.0 BASE [7]	LS-960	4-gram	2.7	7.9	3.4	8.0
wav2vec 2.0 LARGE [7]	LL-60k	Transformer	1.9	4.0	2.0	4.0
HUBERT BASE	LS-960	4-gram	2.7	7.8	3.4	8.1
HUBERT LARGE	LL-60k	Transformer	1.8	3.7	2.1	3.9
HUBERT X-LARGE	LL-60k	Transformer	**1.7**	**3.0**	**1.9**	**3.5**

TABLE III
COMPARISON WITH THE LITERATURE ON HIGH RESOURCE SETUPS USING ALL 960 HOURS OF LABELED LIBRISPEECH DATA

Model	Unlabeled Data	LM	dev-clean	dev-other	test-clean	test-other
		Superivsed				
Conformer L [63]	-	LSTM	-	-	1.9	3.9
		Self-Training				
IPL [13]	LL-60k	4-gram + Transformer	1.85	3.26	2.10	4.01
Noisy Student [62]	LV-60k	LSTM	1.6	3.4	1.7	3.4
		Pre-Training				
wav2vec 2.0 LARGE [7]	LL-60k	Transformer	1.6	3.0	1.8	3.3
pre-trained Conformer XXL [41]	LL-60k	LSTM	1.5	3.0	1.5	3.1
		Pre-Training + Self-Training				
wav2vec 2.0 + self-training [64]	LL-60k	Transformer	1.1	2.7	1.5	3.1
pre-trained Conformer XXL + Noisy Student [41]	LL-60k	LSTM	1.3	2.6	1.4	2.6
		This work (Pre-Training)				
HUBERT LARGE	LL-60k	Transformer	1.5	3.0	1.9	3.3
HUBERT X-LARGE	LL-60k	Transformer	1.5	2.5	1.8	2.9

TABLE IV
STABILITY OF K-MEANS AS AN UNSUPERVISED UNIT DISCOVERY ALGORITHM WITH RESPECT TO DIFFERENT FEATURES, NUMBERS OF CLUSTERS, AND TRAINING DATA SIZES. PNMI STANDS FOR PHONE-NORMALIZED MUTUAL INFORMATION

feature	C	PNMI (mean ± std) with K-means Training Size = 1h	10h	100h
MFCC	100	0.251 ± 0.001	0.253 ± 0.001	0.253 ± 0.001
	500	0.283 ± 0.001	0.285 ± 0.000	0.287 ± 0.001
BASE-it1-L6	100	0.563 ± 0.012	0.561 ± 0.012	0.575 ± 0.008
	500	0.680 ± 0.005	0.684 ± 0.003	0.686 ± 0.004

TABLE V
THE EFFECT OF THE TRAINING OBJECTIVE AND CLUSTERING QUALITY. C REFERS TO THE NUMBER OF UNITS, AND α IS THE WEIGHT FOR MASKED FRAMES

teacher	C	PNMI	dev-other WER (%) $\alpha = 1.0$	$\alpha = 0.5$	$\alpha = 0.0$
Chenone (supervised top-line)	8976	0.809	10.38	9.16	9.79
GMM on MFCC	100	0.303	16.95	-	-
K-means on MFCC	50	0.227	18.68	31.07	94.60
	100	0.243	17.86	29.57	96.37
	500	0.276	18.40	33.42	97.66
K-means on BASE-it1-layer6	500	0.637	11.91	13.47	23.29
K-means on BASE-it2-layer9	500	0.704	10.75	11.59	13.79

as:

$$L_m(f; X, M, Z) = \sum_{t \in M} \log p_f(z_t \mid \tilde{X}, t), \tag{1}$$

and L_u is of the same form except that it sums over $t \notin M$. The final loss is computed as a weighted sum of the two terms: $L = \alpha L_m + (1 - \alpha) L_u$. In the extreme case when $\alpha = 0$, the loss is computed over the unmasked timesteps, which is similar to acoustic modeling in hybrid speech recognition systems [36]–[39]. In our setup, this limits the learning process to mimicking the clustering model.

In the other extreme with $\alpha = 1$, the loss is only computed over the masked timesteps where the model has to predict the targets corresponding to the unseen frames from context, analogous to language modeling. It forces the model to learn both the acoustic representation of unmasked segments and the long-range temporal structure of the speech data. We hypothesize that the setup with $\alpha = 1$ is more resilient to the quality of cluster targets, which is demonstrated in our experiments (see Table V).

C. Learning With Cluster Ensembles

A simple idea to improve target quality is to utilize multiple clustering models. While an individual clustering model may perform terribly, cluster ensembles can provide complementary information to facilitate representation learning. For example, an ensemble of k-means models with different codebook sizes can create targets of different granularity, from manner classes (vowel/consonant) to sub-phone states (senones). To extend the proposed framework, let $Z^{(k)}$ be the target sequences generated by the k-th clustering model. We can now re-write L_m as:

$$L_m(f; X, \{Z^{(k)}\}_k, M) = \sum_{t \in M} \sum_k \log p_f^{(k)}(z_t^{(k)} \mid \tilde{X}, t), \tag{2}$$

and similarly for the unmasked loss L_u. This is analogous to multi-task learning, but with tasks created by unsupervised clustering.

Additionally, ensembling is intriguing because it can be used alongside product quantization (PQ) [40], where a feature space is partitioned into multiple subspaces, and each subspace is quantized separately. PQ allows effective Euclidean distance-based quantization such as k-means for high-dimensional features and heterogeneous features whose scale differs significantly between subspaces. In this case, the theoretical size of the target space is the product of all codebooks' sizes.

D. Iterative Refinement of Cluster Assignments

In addition to using cluster ensembles, another direction for improved representation is *refining* the cluster assignments throughout the learning process. Since we expect a pre-trained model to provide better representations than the raw acoustic feature such as MFCCs, we can create a new generation of clusters by training a discrete latent model over the learned latent representations. The learning process then proceeds with the newly discovered units.

E. Implementation

Our pre-trained models follows the wav2vec 2.0 architecture [7], with a convolutional waveform encoder, a BERT encoder [20], a projection layer and a code embedding layer. We consider HuBERT in three different configurations: BASE, LARGE, and X-LARGE. The fisrt two follow the architectures of wav2vec 2.0 BASE and LARGE closely. The X-LARGE architecture expands the model size to about 1 billion parameters, similar to the size of the Conformer XXL model in [41]. The waveform encoder is identical for all the three configurations, which is composed of seven 512-channel layers with strides [5,2,2,2,2,2,2] and kernel widths [10,3,3,3,3,2,2]. The BERT encoder consists of many identical transformer blocks, whose parameters along with the parameter of the subsequent projection layer are specified in Table I.

The convolutional waveform encoder generates a feature sequence at a 20 ms framerate for audio sampled at 16 kHz (CNN encoder down-sampling factor is 320x). The audio encoded features are then randomly masked as described in Section II-B. The BERT encoder takes as input the masked sequence and outputs a feature sequence $[o_1, \ldots, o_T]$. The distribution over codewords is parameterized with

$$p_f^{(k)}(c \mid \tilde{X}, t) = \frac{\exp(\text{sim}(A^{(k)} o_t, e_c)/\tau)}{\sum_{c'=1}^{C} \exp(\text{sim}(A^{(k)} o_t, e_{c'})/\tau)}, \tag{3}$$

where A is the projection matrix, e_c is the embedding for codeword c, $\text{sim}(\cdot, \cdot)$ computes the cosine similarity between two vectors, and τ scales the logit, which is set to 0.1. When cluster ensembles are used, one projection matrix $A^{(k)}$ is applied for each clustering model k.

After HuBERT pre-training, We use the connectionist temporal classification (CTC) [42] loss for ASR fine-tuning of the whole model weights except the convolutional audio encoder, which remains frozen. The projection layer(s) is removed and replaced with a randomly initialized softmax layer. The CTC target vocabulary includes 26 English characters, a space token, an apostrophe, and a special CTC blank symbol.

III. Related Work

We discuss recent studies on self-supervised speech representation learning by grouping them by training objective. The earliest line of work learns representations by postulating a generative model for speech with latent variables, which are assumed to capture the relevant phonetic information. Training of these models amounts to likelihood maximization. Different latent structures have been applied to encode the prior assumption, such as continuous [30], discrete [32], [43], or sequential [29], [31], [33], [34], [44].

Prediction-based self-supervised learning has gathered increasing interests recently, where a model is tasked to predict the content of the unseen regions [5], [45]–[51] or to contrast the target unseen frame with randomly sampled ones [2]–[4], [7]. Some models combine both the predictive and the contrastive losses [6], [52]. These objectives can usually be interpreted as mutual information maximization [53]. Other objectives do not belong to these categories, for example, [54].

This work is most related to DiscreteBERT [52]: both HuBERT and DiscreteBERT predict discrete targets of masked regions. However, there are several crucial differences. First, instead of taking quantized units as input, HuBERT takes raw waveforms as input to pass as much information as possible to the transformer layers, which was shown to be important in [7]. Furthermore, in the experiment section, we show that our model, with simple k-means targets, can achieve better performance than DiscreteBERT that uses vq-wav2vec [6] learned units. Second, we also present many techniques to improve teacher quality instead of using a single fixed teacher as done in DiscreteBERT.

HuBERT is also related to wav2vec 2.0 [7]. However, the latter employs a contrastive loss that requires careful design of where to sample negative frames from, an auxiliary diversity loss to encourage the discrete unit usage, and demands a proper Gumbel-softmax temperature annealing schedule. In addition, it only explores quantizing the waveform encoder output, which may not be the best feature for quantization due to the limited capacity of the convolutional encoder, as suggested by our ablation studies in Figure 2. Concretely, our proposed method adopts a more direct predictive loss by separating the acoustic unit discovery step from the masked prediction representation learning phase and achieves the state-of-the-art results that match or outperform wav2vec 2.0 on different fine-tuning scales.

Finally, the idea of iterative refinement target labels is similar to iterative pseudo labeling for semi-supervised ASR [13], [55], which leverages an improving student model to generate better pseudo-labels for the next iteration of training. The HuBERT approach can be seen as extending this method to the self-supervised setup with a masked prediction loss.

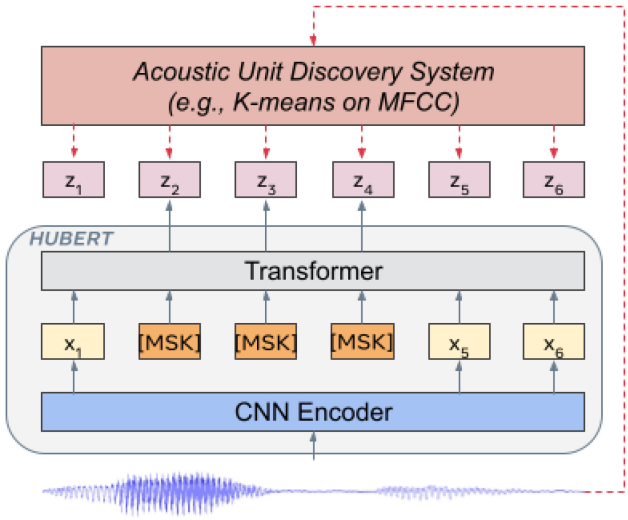


Fig. 1. The HuBERT approach predicts hidden cluster assignments of the masked frames (y_2, y_3, y_4 in the figure) generated by one or more iterations of k-means clustering.

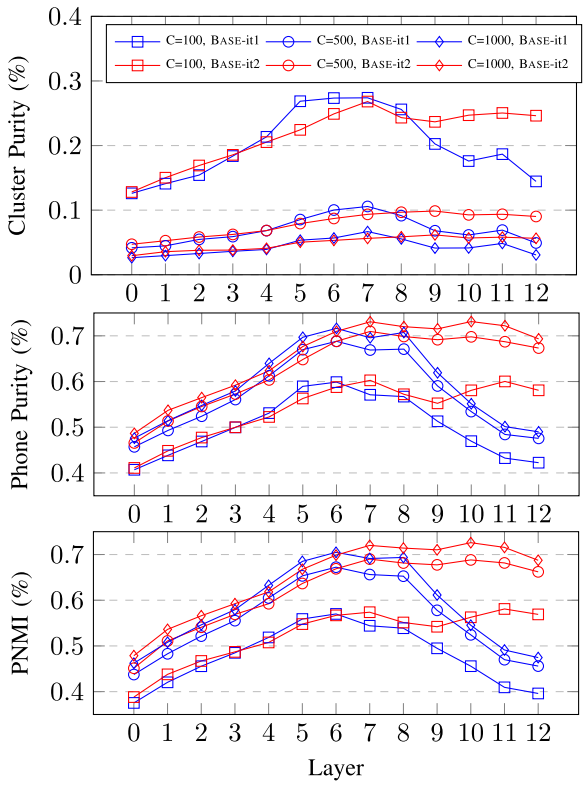


Fig. 2. Quality of the cluster assignments obtained by running k-means clustering on features extracted from each transformer layer of the first and the second iteration BASE HuBERT models.

IV. Experimental Details

A. Data

For unsupervised pre-training, we use the full 960 hours of LibriSpeech audio [25] or 60,000 hours of Libri-light [26] audio, both of which are derived from the LibriVox project that contains

English recordings of copyright-free audiobooks by volunteers from the Internet. For supervised fine-tuning, five different partitions are considered: Libri-light 10-minute, 1-hour, 10-hour splits and LibriSpeech 100-hour (`train-clean-100`) and 960-hour (`train-clean-100`, `train-clean-360`, `train-other-500` combined) splits. The three Libri-light splits are subsets of the the LibriSpeech training split, and each of them contain half of the audio from `train-clean-*` and the other from `train-other-500`.

B. Unsupervised Unit Discovery

To demonstrate the effectiveness of the proposed method on utilizing low-quality cluster assignments, we consider the k-means algorithm [56] for acoustic unit discovery by default. It is one of the most naive unit discovery models that can be treated as modeling an isotropic Gaussian with the same scalar variance for each acoustic unit. To generate labels for the first iteration HuBERT training over the 960 h LibriSpeech training set, we run k-means clustering with 100 clusters on 39-dimensional MFCC features, which are 13 coefficients with the first and the second-order derivatives.

To generate better targets for the subsequent iterations, we run k-means clustering with 500 clusters on the latent features extracted from the HuBERT model pre-trained in the previous iteration (not fine-tuned) at some intermediate transformer layer. Since the feature dimension at the transformer output is much higher than the MFCC features (768-D for HuBERT BASE), we cannot afford to load the entire 960 h training split to the memory. So instead, we randomly sample 10% of the data for fitting the k-means model.

The `MiniBatchKMeans` algorithm implemented in the `scikit-learn` [57] package is used for clustering, which fits a mini-batch of samples at a time.[3] We set the mini-batch size to be 10,000 frames. k-means++ [58] with 20 random starts is used for better initialization.

C. Pre-Training

We train the BASE model for two iterations on the 960 hours of LibriSpeech audio on 32 GPUs, with a batch size of at most 87.5 seconds of audio per GPU. The first iteration is trained for 250 k steps, while the second iteration is trained for 400 k steps using labels generated by clustering the 6-th transformer layer output of the first iteration model. The third iteration brings very marginal improvement when fine-tuned on the 10-hour split (from 9.0% to 8.9% on `dev-other`) and hence we stop at the second iteration. Training for 100 k steps takes about 9.5 hours.

Next we train HuBERT LARGE and X-LARGE for one iteration on 60,000 hours of Libri-light audio on 128 and 256 GPUs, respectively, for 400 k steps. The batch sizes are reduced to 56.25 and 22.5 seconds of audio per GPU due to memory constraints. Instead of restarting the iterative process from clustering MFCC features, we extract features from the 9-th transformer layer of the second iteration BASE HuBERT for clustering and use those labels for training these two models. Hence, these two models can also be seen as the third iteration models.

For all HuBERT configurations, mask span is set to $l = 10$, and $p = 8\%$ of the waveform encoder output frames are randomly selected as mask start if not otherwise mentioned. Adam [59] optimizer is used with $\beta = (0.9, 0.98)$, and the learning rate ramps up linearly from 0 to the peak learning rate for the first 8% of the training steps, and then decays linearly back to zero. The peak learning rates are 5e-4/1.5e-3/3e-3 for BASE/LARGE/X-LARGE models.

D. Supervised Fine-Tuning and Decoding

We fine-tune each model on 8 GPUs on the labeled splits described in Section IV-A. The batch sizes per GPU are at most 200/80/40 seconds of audio for BASE/LARGE/X-LARGE models. During fine-tuning, the convolutional waveform audio encoder parameters are fixed. Like wav2vec 2.0, we introduce a *freeze-step* hyperparameter to control how many fine-tuning steps the transformer parameters are fixed, and only the new softmax matrix is trained. We sweep over peak learning rate ([1e-5, 1e-4]), learning rate schedule (percentage of steps for linear ramp-up and decay), number of fine-tuning steps, freeze step, and waveform encoder output masking probability for each model size and fine-tuning split combination using the word error rate (WER) on the `dev-other` subset as a criterion for model selection.

We use the wav2letter++ [60] beam search decoder wrapped in Fairseq [61] for language model-fused decoding, which optimizes:

$$\log p_{CTC}(Y \mid X) + w_1 \log P_{LM}(Y) + w_2|Y|, \qquad (4)$$

where Y is the predicted text, $|Y|$ is the length of the text, and w_1 and w_2 denote the language model weight and utterance length weight. The decoding hyperparameters are searched with Ax, a Bayesian optimization toolkit,[4]. In this work, we consider both n-gram and transformer language models trained on the official Librispeech language modeling data.

E. Metrics of Target Quality

For analysis, we derive frame-level forced-aligned phonetic transcripts using a hybrid ASR system to measure the correlation between the k-means cluster assignments and the actual phonetic units. Given aligned frame-level phonetic labels $[y_1, \ldots, y_T]$ and k-means labels $[z_1, \ldots, z_T]$, the joint distribution between the two variables $p_{yz}(i, j)$ can be estimated by counting the occurrences:

$$p_{yz}(i, j) = \frac{\sum_{t=1}^{T} [y_t = i \wedge z_t = j]}{T}, \qquad (5)$$

where i denotes the i-th phoneme class and j denotes the j-th k-means label class. The marginal probabilities are computed as $p_z(j) = \sum_i p_{yz}(i, j)$ and $p_y(i) = \sum_j p_{yz}(i, j)$.

[3]It still requires loading the entire dataset to the memory first.

[4][Online]. Available: https://github.com/facebook/Ax

For each phone class i, we further compute the most likely target label as:

$$z^*(i) = \arg\max_j p_{yz}(i, j). \tag{6}$$

Likewise, for each k-means class j, we compute the most likely phone label as:

$$y^*(j) = \arg\max_i p_{yz}(i, j). \tag{7}$$

Three metrics are considered:

1) *Phone Purity (Phn Pur.):*

$$\mathbb{E}_{p_z(j)}[p_{y|z}(y^*(j) \mid j)], \tag{8}$$

where $p_{y|z}(i \mid j) = p_{yz}(i, j)/p_z(j)$ denotes the conditional probability of phone given a k-means label. This metric measures the average phone purity within one class, which can be interpreted as the frame-level phone accuracy if we transcribe each k-means class with its most likely phone label. When comparing different sets of target labels with the same number of units, higher purity indicates better quality. However, this metric is less meaningful when comparing two sets with different numbers of units: in the extreme case where each frame is assigned a unique target label, the phone purity would be 100%.

2) *Cluster Purity (Cls Pur.):*

$$\mathbb{E}_{p_y(i)}[p_{z|y}(z^*(i) \mid i)], \tag{9}$$

where $p_{z|y}(j \mid i) = p_{yz}(i, j)/p_y(i)$ denotes the conditional probability of a k-means label given phone label. Cluster purity is the counterpart of phone purity, whose value would typically decrease when the number of units increases. When comparing target labels with the same number of units, higher cluster purity also indicates a better quality, as frames of the same phone are more likely labeled as the same k-means label class.

3) *Phone-Normalized Mutual Information (PNMI):*

$$\frac{I(y; z)}{H(y)} = \frac{\sum_i \sum_j p_{yz}(i, j) \log \frac{p_{yz}(i, j)}{p_y(i)p_z(j)}}{\sum_i p_y(i) \log p_y(i)} \tag{10}$$

$$= \frac{H(y) - H(y \mid z)}{H(y)} \tag{11}$$

$$= 1 - \frac{H(y \mid z)}{H(y)}. \tag{12}$$

PNMI is an information-theoretic metric that measures the percentage of uncertainty about the phone label y eliminated after observing the k-means label z. Higher PNMI also indicates better k-means clustering quality.

V. Results

A. Main Results: Low- and High-Resource Setups

Table II presents results for the low-resource setup, where pre-trained models are fine-tuned on 10 minutes, 1 h, 10 hours, or 100 hours of labeled data. We include comparison with semi-supervised (iterative pseudo labeling (IPL) [13], slimIPL [55], noisy student [62]) and self-supervised approaches (DeCoAR 2.0 [51], DiscreteBERT [52], wav2vec 2.0 [7]) in the literature. Increasing the amount of unlabeled data and increasing the model size improve performance, demonstrating the scalability of the proposed HuBERT self-supervised pre-training method. In the ultra-low resource setup with just 10 minutes of labeled data, the HuBERT LARGE model can achieve a WER of 4.7% on the test-clean set and 7.6% on the test-other set, which is 0.1% and 0.6% WER lower, respectively than the state-of-the-art wav2vec 2.0 LARGE model. By further scaling up the model size to 1B parameters, the HuBERT X-LARGE model can further reduce the WER to 4.6% and 6.8% on test-clean and test-other. The superiority of HuBERT persists across most setups with different amounts of labeled data, with the exceptions being fine-tuning on 100 hours of labeled data, where HuBERT LARGE is 0.1% higher than wav2vec 2.0 LARGE on test-clean and HuBERT BASE is 0.1% higher than wav2vec 2.0 BASE on test-other, and the test-clean performance of HuBERT BASE models when fine-tuned on 10-min and 1-hour splits. In addition, HuBERT also outperforms DiscreteBERT by a large margin in all setups, while both are trained with a virtually identical objective - masked prediction of discovered units. The considerable performance gap suggests two things. First, using waveform as the input to the model is crucial for avoiding loss of information during quantization. Second, while vq-wav2vec [6], the units that DiscreteBERT uses for training, may discover better units than k-means clustering of MFCC features, the proposed iterative refinement benefits from the improving HuBERT model and learn better units eventually. We will verify these statements in the ablation study sections.

We report results of fine-tuning HuBERT models on the full 960 hours of Librispeech data and compare with the literature in Table III. Prior studies using additional unpaired speech are classified into:

1) self-training: first train an ASR on labeled data to annotate unlabeled speech, and then combine both golden and ASR-annotated text-speech pairs for supervised training.
2) pre-training: first use unlabeled speech for pre-training a model, and then fine-tune the model on labeled data with a supervised training objective.
3) pre-training + self-training: first pre-train and fine-tune a model, and then use it to annotate unlabeled speech for self-training combined with supervised data.

HuBERT outperforms the state-of-the-art supervised and self-training methods and is on par with the two best pre-training results in the literature; both are based on wav2vec 2.0 contrastive learning. In contrast, it lags behind methods combining pre-training with self-training. However, as observed in [64] and [41], we expect that HuBERT can achieve comparable or better performance after combining with self-training, since the pre-trained HuBERT model is on par or better than the pre-trained model those two methods use for pseudo labeling.

B. Analysis: K-Means Stability

To better understand why masked prediction of discovered units is effective, we conduct a series of analyses and ablation studies. We start with probing the stability of the k-means clustering algorithm concerning different numbers of clusters and

different sizes of its training data. Two features are considered: 39-dimensional MFCC features and 768-dimensional output from the 6-th transformer layer of the first iteration HuBERT-BASE model. These two features are used to produce cluster assignments for the first and the second iteration HUBERT training, respectively.

For k-means clustering, we consider $K = \{100, 500\}$ clusters fitted on {1, 10, 100} hours of speech sampled from the LibriSpeech training split. Each combination of the hyperparameters and the features are trained for 10 trials, and the mean and standard deviation of the supervised PNMI metric on the development set (combining dev-clean and dev-other from LibriSpeech) is reported in Table IV. The results show that the k-means clustering is reasonably stable given the small standard deviations across different hyperparameters and features. Furthermore, increasing the amount of data used for fitting k-means models improves PNMI in general, but the gain is only as much as 0.012, suggesting the feasibility of using k-means for unit discovery even with limited CPU memory relative to the feature matrix size. Lastly, the PNMI score is much higher when clustering on HuBERT features than clustering on MFCC features, and the gap is even larger with 500 clusters, indicating that iterative refinement significantly improves the clustering quality.

C. Analysis: Clustering Quality Across Layers and Iterations

We next study how each layer of the HuBERT model from each iteration performs when used for clustering to generate training targets. The two BASE HuBERT models from the first two iterations as described in Section IV-C are considered, which are referred to as BASE-it1 and BASE-it2, respectively. There are 26 features representing 12 transformer layers plus the input to the first transformer layer (denoted as "Layer 0") from the two HuBERT models. For each feature, we fit three k-means models ($K = \{100, 500, 1000\}$ clusters) on a 100 h subset randomly sampled from the LibriSpeech training data. The teacher quality measured in cluster purity, phone purity, and phone normalized mutual information (PNMI) is shown in Figure 2. As a baseline, MFCC achieves (cluster purity, phone purity, PNMI) = (0.099, 0.335, 0.255) for $K = 100$ and (0.031, 0.356, 0.287) for $K = 500$.

Both BASE-it1 and BASE-it2 features result in significantly better clustering quality on all three metrics than MFCC with the same number of clusters. On the other hand, the best BASE-it2 feature is better than the best BASE-it1 on phone purity and PNMI, but slightly worse on cluster purity. Finally, we observe different trends across layers from BASE-it1 and BASE-it2: while BASE-it2 model features generally improve over layers, BASE-it1 has the best features in the middle layers around the 6th layer. Interestingly, the quality of the last few layers degrades dramatically for BASE-it1, potentially because it is trained on target assignments of worse quality, and therefore the last few layers learn to mimic their bad label behavior.

D. Ablation: The Importance of Predicting Masked Frames

We present a series of ablation studies in the following sections to learn how pre-training objective, cluster quality, and hyperparameters affect the performance. The models for ablation studies are pre-trained for 100 k steps and fine-tuned on the 10-hour libri-light split using fixed hyperaprameters. MFCC-based k-means units with C=100 are used if not otherwise mentioned. We report WERs on the dev-other set decoded with the n-gram language model using fixed decoding hyperparameters.

To understand the importance of our proposal to predict the masked frames only, we compare three conditions: 1) predicting masked frames, 2) predicting all frames, and 3) predicting unmasked frames, which can be simulated by setting α to 1.0, 0.5, and 0.0, respectively. We are comparing three k-means models learned from clustering MFCC teachers with 50, 100, 500 clusters, one learned from clustering HuBERT-BASE-it1 6th transformer layer features, and supervised labels obtained from the forced-alignment of character-based HMM models (chenone) [65].

Results shown in Table V indicate that when learning from bad cluster assignments, computing loss only from the masked regions achieves the best performance, while the inclusion of unmasked loss results in significantly higher WERs. However, as the clustering quality improves, the model would suffer less when computing losses on the unmasked frames (BASE-it1-layer6) or even achieve better performance as the case of chenone.

Replacing the k-means teacher with GMM clustering provides less than 1% of WER reduction as shown in Table V. It is an encouraging sign that a better clustering method improves the learned HuBERT representation; however, the scale of the observed gain is much less than the 6% obtained from another iteration of HuBERT training from learned hidden representation.

E. Ablation: The Effect of Cluster Ensembles

To understand the effect of combining multiple k-means models for generating targets, we consider two setups. The first one has k-means models of different numbers of clusters presented in Table V, denoted with "K-means on MFCC, K=*." The second one has k-means models trained on spliced MFCC features with a window of three; hence, each input feature is represented as a 117-dimensional vector. In this second case, we apply product quantization on the spliced features, where dimensions are split into the coefficients of the zeroth, first, and second-order derivatives. This is equivalent to quantizing the spliced zeroth, first, and second order MFCC coefficients separately. We denote these codebooks with "K-means on S-MFCC-*, K=100." By comparing the results from Table V and Table VI, we observe that combining teachers obtained from clustering the same feature with different number of clusters achieves better performance (17.81%/17.56%) than using each teacher individually (18.68%/17.86%/18.40%). Similarly, combining teachers obtained from clustering different features (16.73%) also leads to improvement over using individual teachers (19.26%/17.64%/18.46%). Note that while predicting cluster ensemble leads to improvement, the gain is not as significant as that achieved from iterative clustering (from 18.40% to 11.91%

TABLE VI
CLUSTER ENSEMBLES WITH K-MEANS ON MFCC AND SPLICED MFCC FEATURES. S-MFCC-N REFERS TO THE N-TH ORDER DERIVATIVES OF THE SPLICED MFCC FEATURES

teacher	WER
K-means on MFCC, K={50,100}	17.81
K-means on MFCC, K={50,100,500}	17.56
K-means on S-MFCC-0, K=100	19.26
K-means on S-MFCC-1, K=100	17.64
K-means on S-MFCC-2, K=100	18.46
K-means on S-MFCC-{0,1,2}, K=100	16.73

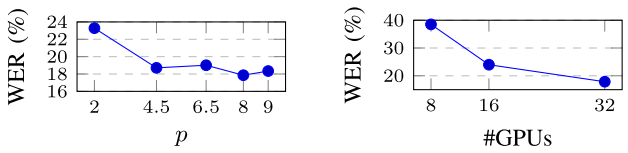


Fig. 3. Varying masking probability p (left) and effective batch size through the number of GPUs (right).

TABLE VII
VARYING THE NUMBER OF HUBERT PRE-TRAINING STEPS. p IS SET TO 6.5%

teacher	C	dev-other WER (%) steps=100k	250k	400k	800k
K-means	50	18.68	13.65	12.40	11.82
	100	17.86	12.97	12.32	11.68
[52]	13.5k	26.6			

as shown in Table V. Hence, for simplicity we present results with a single K-means teacher in Table II and III.

F. Ablation: Impact of Hyperparameters

Figure 3 and Table VII studies how hyperparameters affect HuBERT pre-training. It is shown that

1) the portion of frames selected as mask start is optimal at p =8%;
2) increasing the batch size can significantly improve the performance;
3) training for longer consistently helps for both k-means models with C={50, 100}, and the best model achieves a WER of 11.68%.

These findings are also consistent with those from BERT-like models [21]. In addition, we include a comparable result from DiscreteBERT [52] in Table VII which applies k-means to quantize the same MFCC features into 13.5 k units, used as both the output and the *input* to the BERT model. Besides using continuous speech input rather than discrete units, we hypothesize that HuBERT achieves significantly better performance because its fewer k-means clusters of 100 or 500 help capture broad phonetic concepts without delving into inter/intra-speaker variation.

VI. CONCLUSION

This paper presents HuBERT, a speech representation learning approach that relies on predicting K-means cluster assignments of masked segments of continuous input. On both the Librispeech 960 hours and the 60,000 hours Libri-light pre-training setups, HuBERT matches or outperforms the state-of-the-art systems over all fine-tuning subsets of 10mins, 1 h, 10 h, 100 h, and 960 h. Furthermore, the learned representation quality improves dramatically with iteratively refining K-means cluster assignments using learned latent representations for a previous iteration. Finally, HuBERT scales well to a 1B transformer model showing a relative reduction in WER of up to 13% on the test-other subset. For future work, we plan to improve the HuBERT training procedure to consist of a single phase. Furthermore, given the high quality of its representations, we will consider using HuBERT pre-trained representations for multiple downstream recognition and generation tasks beyond ASR.

REFERENCES

[1] W.-N. Hsu, Y.-H. H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "HuBERT: How much can a bad teacher benefit ASR pre-training?," in *Proc. Neural Inf. Process. Syst. Workshop Self-Supervised Learn. Speech Audio Process. Workshop*, 2021, pp. 6533–6537. [Online]. Available: https://arxiv.org/pdf/2106.07447.pdf
[2] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
[3] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," 2019, *arXiv:1904.05862*.
[4] E. Kharitonov *et al.*, "Data augmenting contrastive learning of speech representations in the time domain," 2020, *arXiv:2007.00991*.
[5] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," 2019, *arXiv:1904.03240*.
[6] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," 2019, *arXiv:1910.05453*.
[7] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020, *arXiv:2006.11477*.
[8] G. Zavaliagkos and T. Colthurst, "Utilizing untranscribed training data to improve performance," in *Proc. DARPA Broadcast. News Trans. Understanding Workshop*, Landsdowne, VA, 1998, pp. 301–305.
[9] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised training on large amounts of broadcast news data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Toulouse, May 2006, pp. 1056–1059.
[10] J. Kahn, A. Lee, and A. Hannun, "Self-training for end-to-end speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7084–7088.
[11] W.-N. Hsu, A. Lee, G. Synnaeve, and A. Hannun, "Semi-supervised speech recognition via local prior matching," 2020, *arXiv:2002.10336*.
[12] A. Xiao, C. Fuegen, and A. Mohamed, "Contrastive semi-supervised learning for ASR," 2021, *arXiv:2103.05149*.
[13] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, "Iterative pseudo-labeling for speech recognition," 2020, *arXiv:2005.09267*.
[14] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," 2020, *arXiv:2006.09882*.
[15] X. Chen and K. He, "Exploring simple siamese representation learning," 2020, *arXiv preprint arXiv:2011.10566*.
[16] J. Grill *et al.*, "Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning," 2020, *arXiv:2006.07733*.
[17] T. B. Brown *et al.*, "Language models are few-shot learners," 2020, *arXiv:2005.14165*.
[18] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[19] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[21] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," 2020, *arXiv:2003.10555*.

[22] M. E. Peters *et al.*, "Deep contextualized word representations," *Proc. 28th Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, pp. 2227–2237, Jun. 2018.

[23] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.

[24] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2018, pp. 132–149.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5206–5210.

[26] J. Kahn *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7669–7673.

[27] C.-Y. Lee and J. Glass, "A nonparametric bayesian approach to acoustic model discovery," in *Proc. Assoc. Comput. Linguistics*, 2012, pp. 40–49.

[28] L. Ondel, L. Burget, and J. Černockỳ, "Variational inference for acoustic unit discovery," *Procedia Comput. Sci.*, vol. 81, pp. 80–86, 2016.

[29] J. Ebbers, J. Heymann, L. Drude, T. Glarner, R. Haeb-Umbach, and B. Raj, "Hidden Markov model variational autoencoder for acoustic unit discovery," in *Proc. InterSpeech*, 2017, pp. 488–492, doi: 10.21437/Interspeech.2017-1160.

[30] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," in *Proc. InterSpeech*, 2017, pp. 1273–1277, doi: 10.21437/Interspeech.2017-349.

[31] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," *Adv. Neural Inf. Process. Syst.*, vol. 33, 2017.

[32] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2041–2053, Dec. 2019.

[33] S. Khurana, S. R. Joty, A. Ali, and J. Glass, "A factorial deep Markov model for unsupervised disentangled representation learning from speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6540–6544.

[34] S. Khurana *et al.*, "A convolutional deep markov model for unsupervised speech representation learning," 2020, *arXiv:2006.02547*.

[35] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: Improving pre-training by representing and predicting spans," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 64–77, 2020.

[36] S. Young, "Large vocabulary continuous speech recognition: A review," *IEEE Signal Process. Mag.*, vol. 13, no. 5, pp. 45–57, Sep. 1996.

[37] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech signal Process.*, 2012, pp. 4277–4280.

[38] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Univ. Cambridge, Cambridge, U.K., 2005.

[39] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, vol. 247, Berlin, Germany: Springer, 2012.

[40] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998, doi: 10.1109/18.720541.

[41] Y. Zhang *et al.*, "Pushing the limits of semi-supervised learning for automatic speech recognition," 2020, *arXiv:2010.10504*.

[42] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[43] A. van den Oord *et al.*, "Neural discrete representation learning," *Adv. Neural Inf. Process. Syst.*, vol. 33, 2017.

[44] T. Glarner, P. Hanebrink, J. Ebbers, and R. Haeb-Umbach, "Full Bayesian hidden Markov model variational autoencoder for acoustic unit discovery," in *Proc. InterSpeech*, 2018, pp. 2688–2692, doi: 10.21437/Interspeech.2018-2148.

[45] Y.-A. Chung and J. Glass, "Generative pre-training for speech with autoregressive predictive coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 3497–3501.

[46] Y. A. Chung, J. Glass, "Improved speech representations with multi-target autoregressive predictive coding," 2020, *arXiv:2004.05274*.

[47] S. Ling, Y. Liu, J. Salazar, and K. Kirchhoff, "Deep contextualized acoustic representations for semi-supervised speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6429–6433.

[48] W. Wang, Q. Tang, and K. Livescu, "Unsupervised pre-training of bidirectional speech encoders via masked reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6889–6893.

[49] A. T. Liu, S.-W. Yang, P.-H. Chi, P.-C. Hsu, and H.-Y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6419–6423.

[50] P.-H. Chi, P.-H. Chung, T.-H. Wu, C.-C. Hsieh, S.-W. Li, and H.-Y. Lee, "Audio albert: A lite bert for self-supervised learning of audio representation," 2020, *arXiv:2005.08575*.

[51] S. Ling and Y. Liu, "Decoar 2.0: Deep contextualized acoustic representations with vector quantization," 2020, *arXiv:2012.06659*.

[52] A. Baevski, M. Auli, and A. Mohamed, "Effectiveness of self-supervised pre-training for speech recognition," 2019, *arXiv:1911.03912*.

[53] Y.-H. H. Tsai, Y. Wu, R. Salakhutdinov, and L.-P. Morency, "Self-supervised learning from a multi-view perspective," 2020, *arXiv:2006.05576*.

[54] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," in *Proc. InterSpeech*, 2019, pp. 161–165, doi: 10.21437/Interspeech.2019-2605.

[55] T. Likhomanenko, Q. Xu, J. Kahn, G. Synnaeve, and R. Collobert, "slimipl: Language-model-free iterative pseudo-labeling," 2020, *arXiv:2010.11524*.

[56] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982, doi: 10.1109/TIT.1982.1056489.

[57] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–30, 2011.

[58] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982, doi: 10.1109/TIT.1982.1056489.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[60] V. Pratap *et al.*, "wav2letter : The fastest open-source speech recognition system," 2018, *arXiv:1812.07625*.

[61] M. Ott *et al.*, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2019, pp. 48–53.

[62] D. S. Park *et al.*, "Improved noisy student training for automatic speech recognition," 2020, *arXiv:2005.09629*.

[63] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," 2020, *arXiv:2005.08100*.

[64] Q. Xu *et al.*, "Self-training and pre-training are complementary for speech recognition," 2020, *arXiv:2010.11430*.

[65] D. Le, X. Zhang, W. Zheng, C. Fügen, G. Zweig, and M. L. Seltzer, "From senones to chenones: Tied context-dependent graphemes for hybrid speech recognition," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 457–464.