


REVIEW ARTICLE

Natural language processing in Alzheimer's disease research: Systematic review of methods, data, and efficacy

Arezo Shakeri  | Mina Farmanbar

Department of Electrical Engineering and Computer Science, Faculty of Science and Technology, University of Stavanger, Stavanger, Norway

Correspondence

Arezo Shakeri, Department of Electrical Engineering and Computer Science, Faculty of Science and Technology, University of Stavanger, Mailbox 8600, 4036 Stavanger, Norway.
Email: arezo.shakeri@uis.no

Abstract

INTRODUCTION: Alzheimer's disease (AD) prevalence is increasing, with no current cure. Natural language processing (NLP) offers the potential for non-invasive diagnostics, social burden assessment, and research advancements in AD.

METHOD: A systematic review using Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines explored NLP applications in AD, focusing on dataset types, sources, research foci, methods, and effectiveness. Searches were conducted across six databases (ACM, Embase, IEEE, PubMed, Scopus, and Web of Science) from January 2020 to July 2024.

RESULTS: Of 1740 records, 79 studies were selected. Frequently used datasets included speech and electronic health records (EHR), along with social media and scientific publications. Machine learning and neural networks were primarily applied to speech, EHR, and social media data, while rule-based methods were used to analyze literature datasets.

DISCUSSION: NLP has proven effective in various aspects of AD research, including diagnosis, monitoring, social burden assessment, biomarker analysis, and research. However, there are opportunities for improvement in dataset diversity, model interpretability, multilingual capabilities, and addressing ethical concerns.

KEYWORDS

Alzheimer's disease, dementia, natural language processing, systematic review, PRISMA guidelines

Highlights

- This review systematically analyzed 79 studies from six major databases, focusing on the advancements and applications of natural language processing (NLP) in Alzheimer's disease (AD) research.
- The study highlights the need for models focusing on remote monitoring of AD patients using speech analysis, offering a cost-effective alternative to traditional methods such as brain imaging and aiding clinicians in both prediagnosis and post-diagnosis periods.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2025 The Author(s). *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* published by Wiley Periodicals, LLC on behalf of Alzheimer's Association.

- The use of pretrained multilingual models is recommended to improve AD detection across different languages by leveraging diverse speech features and utilizing publicly available datasets.

1 | INTRODUCTION

Dementia is a clinical syndrome marked by a gradual, persistent, and progressive decline in cognitive function, which significantly impairs an individual's capacity for independent living.¹ Dementia ranks as the seventh leading cause of death and is a significant contributor to dependency and disability worldwide.²

Alzheimer's disease (AD) is the predominant form of dementia, impacting over 27 million individuals and accounting for 60% to 70% of all dementia cases.³ The predominant impact of AD on individuals is the creation of atrophy in diverse brain areas, causing the deterioration of brain cells and, in turn, cognitive impairment.⁴ AD hinders cognitive functions, resulting in symptoms such as memory decline, decision-making difficulties, impaired communication, reduced focus, planning issues, and changes in perception.⁵ AD's far-reaching prevalence has extensive societal and economic impacts, affecting more than just patients and caregivers.⁶

The rise of artificial intelligence (AI) introduces novel capabilities to assist AD patients in various capacities. One area showing great potential in AD diagnosis is natural language processing (NLP)-powered models, which are attracting considerable attention. The manual diagnosis of AD and other forms of dementia has become increasingly complex, and current diagnostic protocols rely heavily on specific clinical assessments and neuropsychological tests, which are both time-consuming and expensive. As a result, early detection remains a significant obstacle.⁴ However, recent studies, including the ADReSS Challenge at the Interspeech 2020 conference and the ADReSSo Challenge at the Interspeech 2021 conference, have proven the potential of the NLP-driven model for AD diagnosis and monitoring using patients' speech. These challenges offered research groups a platform to test their methods on two benchmark datasets of spontaneous speech from AD patients and healthy controls (HC). Researchers primarily competed using NLP and signal processing techniques to differentiate AD from non-AD in classification tasks and to evaluate cognitive impairment through a Mini-Mental State Examination (MMSE) score prediction task.

Several studies have reviewed NLP applications in AD-related fields. Shi et al.⁷ concentrated on deep learning (DL) approaches for dementia diagnosis based on speech and language data, specifically differentiating HC from dementia cases. Their review exclusively targeted studies that employed DL for dementia diagnosis. De la Fuente Garcia et al.⁶ also explored AI applications in monitoring AD based on patients' speech and language. They reviewed acoustic and linguistic features that AI-based approaches extract to monitor and diagnose AD stages using speech. A study by Petti et al.⁸ investigated automatic AD detection using speech data, concentrating on identifying

different categories of acoustic and linguistic features present in various patient cohorts, such as AD and mild cognitive impairment (MCI). They asserted that language and speech could be effectively utilized for the automatic detection of dementia. The review study carried out by Patra et al.⁹ presented a systematic review of state-of-the-art NLP approaches and tools designed to identify and extract social determinants of health such as smoking status, substance use, homelessness, and alcohol use from unstructured clinical text in electronic health records (EHR). They concluded that extracted data could aid in the development of screening tools, risk prediction models, and clinical decision support systems.

While the previous reviews provided valuable insights into dementia, our study differs in range and scope. We not only focus on language and speech data but also include studies utilizing other sources of textual data, such as EHR, comments posted by AD caregivers or their families on social media platforms, and medical publications extracted from scientific databases. However, the focus of our study is based on AD, which is the most predominant form of dementia. This systematic literature review (SLR) distinguishes itself from existing ones by proposing a thorough approach to recognizing the areas in which NLP can serve AD patients to provide an in-depth view of research in the field. This review addressed the following six research questions:

1. **RQ1** Which types of data are utilized for AD analysis using NLP?
2. **RQ2** What are the most popular datasets in each category?
3. **RQ3** What are the research goals?
4. **RQ4** What are emerging trends in the field?
5. **RQ5** What NLP approaches are employed in AD analysis?
6. **RQ6** How effective are NLP approaches in AD analysis?

The remainder of this paper is structured as follows. Section 2 describes the methodology and search and selection processes used in this SLR. Section 3 presents the results and provides the findings from the data synthesis. Section 4 discusses the findings and potential directions for future work, and Section 5 concludes the review.

2 | METHODOLOGY

We chose to implement a systematic approach as it is well suited for broadly evaluating diverse studies across multiple disciplines, particularly those that hold clinical significance.⁷ The review followed the PRISMA checklist as its protocol.¹⁰ The process began by identifying the need for the SLR, followed by a formulation of research questions. We then conducted an extensive search and selection of primary studies, assessed the quality of these studies, and extracted relevant data.

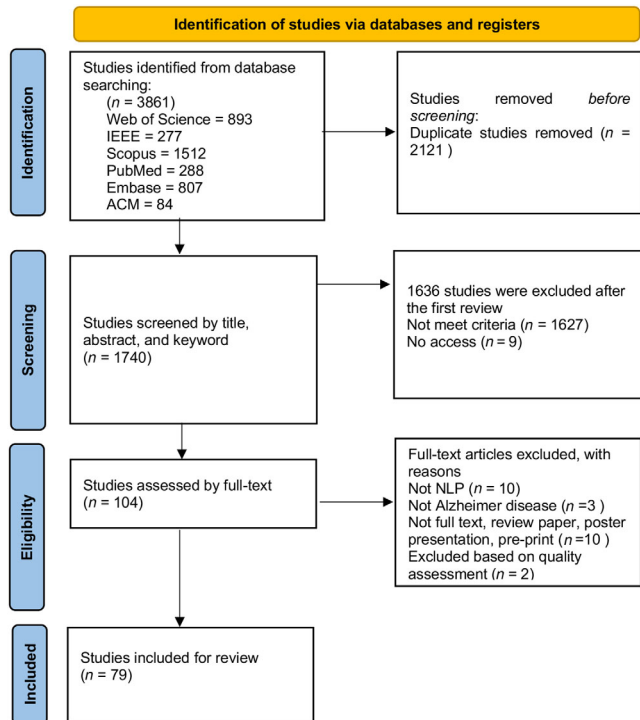


FIGURE 1 PRISMA workflow of reviewed studies. PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses flowchart. ¹⁰

The article assessment procedure is depicted in Figure 1. Finally, we interpreted the results and reported the findings of the SLR.

2.1 | Search process

We searched six scholarly databases – ACM, PubMed, Scopus, Web Of Science, Embase, and IEEE – to identify all relevant articles related to AD and NLP. We conducted searches within a defined timeframe to identify studies that met our criteria for inclusion. The first database query was executed from January 2020 to November 2023, and then the search was updated by July 2024. As a result, this SLR concentrates on the latest developments in AD analysis using NLP, with a specific focus on the developments from the past 4 years. This focus reflects the rapid advancements in NLP, aiming to provide a concise and current review of recent innovations, keeping researchers and clinicians updated on the latest developments in this fast-evolving field. To manage the search results effectively, we restricted the choice of keywords to two sets corresponding to AD and NLP. Those linked to AD encompassed “AD,” “Alzheimer’s,” “Alzheimer,” “Alzheimer’s disease,” “dementia,” and “Alzheimer’s Disease and Related Dementia,” and keywords related to NLP included “natural language processing,” “text mining,” “data mining,” “datamining,” “information storage and retrieval,” “information retrieval,” “NLP,” “medical language processing,” “information extraction”. When executing queries in a database, the keywords within each group were combined using the OR opera-

RESEARCH IN CONTEXT

- 1. Systematic review:** We performed a systematic review across ACM, Embase, IEEE, PubMed, Scopus, and Web of Science databases to assess recent advancements and potential uses of NLP in Alzheimer’s disease (AD) analysis. Adhering to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, this review is unique in its specific focus and scope.
- 2. Interpretation:** Analysis of 79 studies revealed four key trends in NLP applications for Alzheimer’s research: (1) detection and monitoring of AD using speech datasets, (2) identification of AD risk factors based on EHR, (3) summarizing the current state of knowledge in AD-related publications, and (4) exploring the social burden experienced by AD patients, caregivers, and their families based on social media comments.
- 3. Future directions:** The study suggests future research should address developing larger speech datasets for AD analysis, enhancing monitoring and remote AD diagnostic models, incorporating geo- graphically diverse datasets, and integrating privacy-preserving AI tools.

tor, while the two keyword sets related to AD and NLP were connected using the AND operator to refine the search results.

2.2 | Eligibility criteria

We excluded review papers and established our criteria to encompass a range of studies that fit the scope of this SLR. Our selection encompassed studies utilizing NLP approaches either independently or in conjunction with other methodologies. Papers selected for inclusion cover the period from January 2020 to July 2024. Table 1 outlines detailed inclusion and exclusion criteria.

2.3 | Selection process

A sole reviewer carried out the initial screening of studies for the review. Initially, automated searches were conducted across the six previously mentioned databases using the two sets of keywords described in Section 2.1 to identify relevant articles. The preliminary search yielded a collection of 3861 articles. Following the removal of duplicate entries, 1740 distinct articles remained. Subsequently, a rigorous screening process was initiated, wherein papers were evaluated based on predefined keywords, titles, and abstracts, identifying 104 pertinent papers. At this stage, a more inclusive approach was adopted when titles and abstracts did not explicitly specify the cognitive impairments addressed, the nature of the datasets processed,

TABLE 1 Exclusion and inclusion criteria for reviewed research articles.

Exclusion criteria

- Studies that did not employ datasets of patients with AD or MCI
- Conference proceedings and preprints
- Review articles
- Studies dependent on neuroimaging techniques like magnetic resonance imaging (MRI) but without relevance to language, speech, EHR, or other types of text-based datasets
- Conference abstracts, poster presentation
- Failure to meet the quality checklist criteria outlined in Table 2

Inclusion criteria

- Studies considered in the review utilized text-based data either in isolation or in conjunction with other data types
- Studies must incorporate NLP techniques independently or in connection with other methods such as signal processing or ML
- Full articles were accessible in the English language
- Peer-reviewed journal articles and conference papers

TABLE 2 Quality assessment questions for different parts of a quantitative study proposed by Kitchenham and Charters.¹¹

Question	Section
– Are the objectives clearly defined?	Design
– Was the study planned with their research questions as the focus?	Design
– What method was used to obtain the sample (eg, interview, web-based)?	Design
– Is the technology clearly described if the study includes an evaluation of a technology?	Design
– Are the variables in the study adequately measured to ensure they are likely valid and reliable?	Design
– Are the measures employed in the study clearly defined?	Design
– Are the methods for collecting data sufficiently detailed?	Conduct
– Do the researchers specify the data types (eg, continuous, categorical)?	Analysis
– Was there a sufficient description of the basic data?	Analysis
– Is the objective of the analysis clearly outlined?	Analysis
– Are all the study questions addressed?	Conclusion
– Are negative findings included in the study?	Conclusion
– How do the results enhance the current body of literature?	Conclusion

or the application of NLP methodologies. After this initial screening phase, a secondary round of screening was undertaken, involving an inclusive review of full-text articles, culminating in the retention of 79 papers. In the final stage, papers were excluded if they lacked full-text access, focused on dementia types other than MCI or AD, or did not fulfill the quality checklist criteria for bias risk assessment as outlined in Table 2. Studies focusing on MCI were included, given that MCI has the potential to progress into AD. Furthermore, a detailed examination of the papers not meeting the inclusion criteria presented in

Table 1 was undertaken to ensure a thorough selection process. Zotero and EndNote were essential tools to manage study records throughout this process.

2.4 | Bias risk assessment

Adherence to the PRISMA guidelines¹⁰ minimized selection bias and ensured a comprehensive review. For quality assessment, we used the checklist for quantitative studies from Kitchenham and Charters' guidelines for SLRs.¹¹ We applied the 12 questions listed in Table 2 to evaluate the design, implementation, analysis, and conclusions of the papers. This led to the exclusion of papers that failed to meet these criteria, reducing internal biases such as measurement and reporting biases and thereby enhancing the objectivity and reliability of our conclusions. Consequently, the identified biases were minor and did not have a major impact on the impartiality or quality of this review.

2.5 | Data extraction and synthesis

Data were directly gathered from the chosen papers and organized into MS Excel for further analysis. The extracted information comprised the publication year, methodology details, dataset details, performance metrics, objectives, key findings, future prospects, and limitations.

3 | RESULTS

Having discussed the identification of reviewed papers, we now address the six underlying research questions.

3.1 | RQ1: Which types of data are utilized for AD analysis using NLP?

NLP is a highly adaptable field, allowing for various applications in the analysis of AD. This flexibility means that NLP techniques can be employed in multiple, diverse ways to tackle the complexities of AD research. In the section being introduced, the objective is to highlight and discuss the primary directions in which NLP has been applied to AD analysis. This will be done by examining the types of data utilized in the studies that have been reviewed. We identified four major trends in how NLP is used for AD analysis. In addition to these four specific trends, we identified an additional category that encompasses approaches that do not fit neatly into the major classes and have been used infrequently. The details and citations of all reviewed papers are provided in Tables 3–5, each corresponding to a specific data type (Category). The detailed analysis of the 79 reviewed papers revealed that a total of 59 distinct datasets were utilized across these studies.

Speech dataset: The primary type of data utilized comprises speech datasets, accounting for 40.68% (24 out of 59) of the total dataset.

TABLE 3 Overview of research goals, datasets, NLP techniques, and evaluation metrics in studies utilizing speech datasets.

Ref	Goal	Dataset	NLP technique	Best result
28	AD detection based on Nepali speech dataset	Translated version of Pitt corpus: 255 transcripts from 168 AD patients and 244 transcripts from 98 HC	BoW, TF-IDF, Word2Vec, fastText, DT, KNN, SVM, NB, RF, AdaBoost, XGB, CNN, BiLSTM	Domain-specific Word2Vec and CNN: Acc = 0.968, F1-score = 0.968, Precision = 0.969, Recall = 0.968
51	Predicting cognitive status from neuropsychological test voice recordings	Custom dataset: 35 HC, 55 MCI, 110 participants with dementia	Acoustic, linguistic, and paralinguistic feature extraction, LR	Based on all types of features and LR classifier: AUROC = [0.80, 0.90]
52	MMSE score prediction and AD detection	ADReSS dataset: 54 AD and 54 HC in the training dataset and 48 unlabeled samples in the test dataset	TF-IDF, DistilBERT, DistilRoBERTa, BERT, RoBERTa, GBDT, SVM, CRF, LASSO regression, and LR	CRF, SVM, and DistilBERT: Acc = 0.81, DistilBERT and LASSO linear model: RMSE = 4.58
53	AD detection using multimodal feature fusion	ADReSS dataset: 78 AD and 78 non-AD	TF-IDF, linguistic and acoustic feature extraction, SVM, LR, linear regression	Audio and readability features and LR: Acc = 0.7708, audio and readability features and SVM: RMSE = 4.4388
54	MMSE score prediction and AD detection	ADReSS dataset: 78 AD and 78 non-AD	Linguistic and acoustic feature extraction, and fine-tuning BERT, SVM, NN, RF, NB	Fine-tuned BERT model: Acc = 0.833, 35 linguistic and acoustic features using ridge regression: RMSE = 4.56
12	Analysis of influence of different embedding models on AD detection	Pitt corpus: 104 HC and 208 dementia patients	Generic and domain-specific word embedding computation of fastText, Word2Vec, and GloVe algorithms, CNN, LSTM	Domain-specific fastText embeddings and a CNN+BiLSTM model: Acc = 0.91, Precision = 0.91, Recall = 0.91, F1-score = 0.91
55	AD detection	ADReSS challenge dataset: 78 AD and 78 non-AD	Wikipedia2Vec word embedding, linguistic feature extraction, BERT, DistilBERT, Bio-Clinical BERT, LR, DT, SVC, LDA, QDA, GNB, XG-Boost, AdaBoost, extra tree classifier	Leave-one-subject-out result based on Gaussian Naive Bayes, linguistic features: Acc = 0.90, F1-score = 0.90
56	Comparing text and audio modalities for AD detection	ADReSS dataset: 78 AD and 78 non-AD	X-vectors, BERT model, probabilistic linear discriminant analysis classifier, SVR, FFNN	BERT embedding, acoustic features, and probabilistic linear discriminant analysis: Acc = 0.75, BERT embedding, acoustic, and silence features, and SVR: RMSE = 5.32
57	MMSE score prediction	ADReSS dataset: 78 AD and 78 non-AD	Linguistic and acoustic feature extraction, linear regression	Linear regression and correlation-based feature selection: RMSE = 3.9
58	Comparing text and audio modalities for AD detection	ADReSSo dataset: 78 AD recordings and 79 HC recordings in the training set and 71 recordings in the test set	X-vectors, prosody vector extraction, BERT, LR, and ensemble method	Fusion of acoustic, linguistic, and BERT model: Acc = 0.845, fine-tuned BERT model: RMSE = 3.85
59	Data augmentation technique evaluation for AD detection	ADReSS dataset: 78 AD and 78 non-AD	Text augmentations: noise, lexical substitution, paraphrase, text generation, audio augmentations: Standard transformations, vocal tract length perturbation and generative models, BERT, audio spectrogram Transformer, TF-IDF, RF, and SVM	Paraphrasing using back-translation from Russian language and SVM: Acc = 0.85
60	Exploring the role of PoS features in AD detection	DementiaBank dataset: 208 AD and 104 HC	PoS feature extraction, Transformer-based classifier	PoS features and a transformer-based deep learning model: Acc = 0.922, F1-score = 0.955, Precision = 0.935, Recall = 0.971
61	Investigation of usefulness of the stopwords in AD detection	Pitt corpus: 194 dementia participants and 98 HC	CountVectorizer, TF-IDF, NB, SVM, DT, KNN, LR, ADB, XGB, RF, LGBM, CatBoost	SVM without removing stopwords and TF-IDF features, 10-fold CV: F1-score = 0.840

(Continues)

TABLE 3 (Continued)

Ref	Goal	Dataset	NLP technique	Best result
62	Exploring the influence of interviewer on AD detection	ADReSSo dataset: 87 AD and 79 HC recordings for the train set and 71 recordings for the test set	X-vectors, prosody, and emotional embeddings, linguistic feature extraction, SVM, linear regression	Fusion of prosody, valence, dominance, BERT embeddings, and RBF-SVM: Acc = 0.80, F1-score = 0.80. Perplexity scores, BERT embedding, and linear regression: RMSE = 4.56
63	Comparison of manually corrected transcription against automatic transcription for neurodegenerative disease classification	Custom dataset: data from 72 memory clinic patients and 77 HC	Feature extraction, text parsing using CoreNLP, Google speech-to-text software, LR, GNBs, RF	Linguistic and psycholinguistic features, and LR AUROC = 0.755
64	AD detection	DementiaBank dataset: 147 AD patients and 98 HC participants	Linguistic and pragmatic feature extraction, RF	Lexicosemantic and pragmatic features and RF: Acc = 0.855
65	Deriving digital voice biomarkers for AD patients	Custom dataset: 92 HC and 114 impaired participants	Meta-semantic and acoustic feature extraction, LR, NN, ensemble models	AUC = 0.80
26	AD detection based on Japanese language	Custom dataset: 42 AD patients and 52 HC	Automatic speech transcription using spaCy and GiNZA libraries, linguistic feature extraction, eXtreme gradient boosting	PoS and dependency parsing features, eXtreme gradient boosting: F1-score = 0.84, Recall = 0.84, Precision = 0.85
66	Developing speech-based diagnostic test for dementia	Custom dataset: 26 AD patients and 42 HC	Charniak Parser, linguistic and acoustic feature extraction, Kaldi software, HMM, DNN, GA-SVM, RF	MMSE score, pause and silence features, an ANN algorithm, and manual transcription: Acc = 0.958, sensitivity=0.962, specificity = 0.957
67	AD detection using a large language model	Pitt corpus: 194 dementia participants and 98 HC	Extraction of embedding based on BERTLarge model, LR	BERT embedding, LR: Acc = 0.880, F1-score = 0.872, Precision = 0.905, Recall = 0.843
68	Investigating the impact of different types of speech task for AD detection	Custom dataset: 25 HC, 13 mild AD, and 12 MCI	Linguistic and paralinguistic feature extraction, LR, SVM	Linguistic features, SVM: Acc = 0.90, AUC = 0.94, Sensitivity = 0.8, Specificity = 0.96 (AD vs HC), linguistic features, SVM Acc = 0.78, AUC = 0.82, Sensitivity = 0.67, Speifity = 0.90 (MCI vs HC)
69	Investigation of correlation of NLP and automated speech analysis extracted features with clinically identified language impairment	Subset of DementiaBank dataset: audio recordings of 30 participants with AD, MCI, and controls, linguistic and acoustic feature extraction	Exploratory factor analysis	–
24	Quantifying the role and contribution of disfluency and interactional features in AD detection	Custom dataset: 15 AD and 15 non-AD participants	Disfluency and interactional feature extraction, LR, SVM, MLP	Disfluency and interactional features and SVM with LOOCV: Acc = 0.90, Precision = 0.90, Recall = 0.90, F1-score = 0.90, AUC= 0.89
30	AD detection based on Chinese language	IFlytek dataset: 68 AD, 144 MCI, and 111 HC	Feature extraction using TF-IDF, word segmentation, keyword extraction, KNN	TF-IDF features, cosine similarity, and KNN using Acc = 0.98, Precision = 0.98, Recall = 0.98, F1-score = 0.98
70	AD detection	ADReSS dataset: 78 AD and 78 non-AD	Word-frequency network analysis, graph classification methods, spectral feature embedding method	Graph measures with RSVM with RBF kernel classifier: Acc = 0.667. MMSE score prediction using graph measures and RF: RMSE = 5.675
21	Examining the practicality of incorporating WLS dataset with ADReSS dataset for AD detection	WLS dataset: 839 HC and 98 cognitively impaired; ADReSS dataset: 78 AD and 78 non-AD	BERT model, NN	Result on ADReSS test dataset using BERT model classifier: Acc = 0.819, AUC = 0.912

(Continues)

TABLE 3 (Continued)

Ref	Goal	Dataset	NLP technique	Best result
27	Indication of potential of remote screening tool for early-stage cognitive decline diagnosis based on Greek language	Custom dataset: 11 MCI patients and 12 HC	Keystroke (CNN) and linguistic feature extraction, KNN, LR, ensemble methods	Keystroke dynamics and linguistic features and KNN classifier: Acc = 0.77, AUC = 0.78, Specificity = 0.64, Sensitivity = 0.92
71	Developing ADRD screening tool	Pitt corpus: 122 ADRD and 115 HC	iZotope RX8 for noise reduction, phonetic motor planning, semantic, syntactic, psycholinguistic features extraction, DistilBERT model, LR, RF, Extra Trees, XGB, AdaBoost, and SVM	Fusion of linguistic, phonetic, psycholinguistic features and embedding from DistilBERT and SVM: Acc = 0.901, F1-score = 0.895, Recall = 0.857, Precision = 0.937, and AUC-ROC = 0.938
72	AD detection	ADReSS dataset: 78 AD and 78 non-AD	Linguistic, acoustic, psycholinguistic and demographic feature extraction, CNN-LSTM	Linguistic, psycholinguistic, demographic features, and a DL model: Acc = 0.7292
73	Determining advantages of using domain knowledge versus pretrained transfer models for automatic AD detection	ADReSS dataset: 78 AD and 78 non-AD	Linguistic and acoustic feature extraction and BERT model, SVM, NN, RF, NB	Fine-tuned BERT model: Acc = 0.8332, Precision = 0.8389, Recall = 0.8333, F1-score = 0.8327, linguistic and acoustic features, Ridge regression: RMSE = 4.56
74	Characterize progressive speech changes in prodromal-to-mild AD cohort through a longitudinal study	Custom dataset: 130 dementia participants	Acoustic and linguistic feature extraction, spaCy, Stanford parser, Praat/Parselmouth27–30, statistical analysis, linear mixed-effects models, Pearson correlation computation	–
75	Determining transfer-learning approach advantages for AD detection	ADReSS dataset: 78 AD participants and 78 non-AD participants	DistilBERT, BERT, and ERNIE models, CNN, RF, AdaBoost, SVM, LR	DistilBERT embedding, and LR: Acc = 0.88, Precision = 0.88, Recall = 0.88, F1-score = 0.87
76	Developing NLP tool for identifying different dementia stages	Custom dataset: 410 HC and 387 MCI participants	Google Speech tool for transcription, ALBERT-xlargemodel for diarization, NER, ALBERT model, BERT model, MLP, and LR	Subtest sampling, random sampling, demographic information, and LR: AUC = 0.926, Acc = 87.1 (HC vs Dementia). AUC = 0.88, Acc = 0.831 (HC, MCI, and dementia). AUC = 0.744, Acc = 0.695 (HC vs MCI)
20	Proposing a novel feature purification network for transformer model for AD detection	Pitt corpus, ADReSS, and iFLY datasets	Transformer feature extractor, BERT, ERNIE, CNN, LR, RCNN, DPCNN	Transformer and feature purification network: Acc = 0.935, Precision = 0.94, Recall = 0.89, F1-score = 0.91 (Pitt), Acc = 0.786 (ADReSS), Acc = 0.83 (iFLY)
14	Development of remote automated story recall task for longitudinal assessment in older adults with and without MCI or mild AD	Custom dataset: 78 HC and 73 mild AD or MCI	Linguistic feature extraction, Google's speech-to-text automatic speech recognition system, LR, and linear regression	–
13	Exploring the potential of semantic fluency tasks in AD detection	Custom dataset: 42 HC, 24 amnesic-MCI, and 18 early AD	Word2vec, semantic fluency feature extraction, formulating the semantic space, t-SNE dimensionality reduction technique, multinomial LR	–
25	Analysis of speech data collected via mobile application for automatic AD detection	Custom dataset: 43 HC, 46 MCI patients, and 25 AD patients	Linguistic (information units, numbers of each PoS tag, and Brunet's index, word tokenization, PoS tagging, and word lemmatization using Japanese morphological analyzer Janome), prosodic and acoustic feature extraction, KNN, RF, LR, a light gradient boosting machine, SVM	Speech features and SVM with RBF kernel and nested leave-one-out procedure: Acc = 0.786 (AD, MCI, and HC), Acc = 0.912, F1-score = 93.5 for (AD vs HC) and Acc = 87.6, F1-score = 87.1 (MCI vs HC)

(Continues)

TABLE 3 (Continued)

Ref	Goal	Dataset	NLP technique	Best result
77	AD detection using multimodal data fusion	ADReSS dataset: 78 AD and 78 non-AD	BERT, Vision Transformer, Co-Attention, Multimodal Shifting Gate, a variant of the self-attention mechanism, NN	BERT, ViT, Gated Self-Attention Acc = 0.90, Precision = 0.90, Recall = 0.89, F-score = 0.89, RMSE = 3.61
78	Developing a personalized assistance system for mentally affected patients such as AD	Custom dataset: auditory sensors data from AD patients	Speech-to-text synthesizer, tokenization, stop words removal, lemmatization, PoS tagger, Naive Bayes classifier, an ensemble approach for abnormality tracking of AD patients, CNN-based emotion detection, IoT-based assistance mechanism	RNN (DeTrAs model) with Acc = 88.63, Precision = 0.83, Recall = 0.886 F-score = 0.84
29	Investigation of generalizable language features in AD detection	ADReSS dataset: 78 AD and 78 non-AD, French dataset: 22 ADRD and 25 non-dementia participants	Linguistic and Paralinguistic feature extraction, LR, SVM, MLP	–
79	AD detection	ADReSSo dataset: 87 AD and 79 non-AD	Generation of text embedding from GPT3 model and acoustic feature extraction, Wav2Vec2, RF, LR, SVC, SVR, RR	GPT-3 embedding (Babbage) with SVR: Acc = 0.8028, Precision = 0.723, Recall = 0.971, F1-score = 0.829, MMSE prediction: RMSE = 5.46
80	To create a multimodal ML model for early detection of cognitive decline based on remote, telephone-based interviews	Custom dataset: 29 HC, 30 AD, and 32 amnesic MCI	Traditional linguistic features, LSA, word2vec, GloVe, USE, BERT embeddings, DT	Acoustic features and DT: Acc = 0.65, Macro-avg-F1-score = 0.66, Macro-avg-Precision = 0.67, Macro-avg-Recall = 0.65
81	To enhance AD detection by employing NLP on spontaneous speech and leveraging audio enhancement techniques and innovative transcription methods	Pitt corpus: 234 probable AD, 21 possible AD, 42 MCI, and 242 HC	Synthetic Minority Over-sampling Technique, locally based Wav2Vec and Whisper transcription tools, GPT-based embeddings, SVM	AD versus HC using Wav2Vec and GPT-based embedding and K-NN: Acc = 0.99, F1-score = 0.99, Precision = 0.99, Recall = 0.99. AD versus MCI versus HC using manual transcription: Acc = 0.97, F1-score = 0.97, Precision = 0.97, Recall = 0.99.

Note: "Custom dataset" refers to datasets that are institutionally sourced and not publicly available.

Abbreviations: Acc, accuracy; AUROC, area under the receiver operating characteristic curve; DT, decision tree; NB, Naive Bayes; XGB, XGBoost; BiLSTM, bidirectional long short-term memory; GBDT, gradient boosted decision tree; CRF, conditional random field; RBF, radial basis function; RMSE, root mean square error; NPV, negative predictive value.

These speech datasets consist of recordings from patients with cognitive impairments such as AD and MCI, as well as HC. The speech samples are collected through various tasks and settings, including in-clinic interviews and remote self-assessment tools on smart devices. These tasks include the Cookie Theft description,¹² semantic fluency task, story recall, and sentence construction. The Cookie Theft description task is a component of the Boston Diagnostic Aphasia Examination,¹² specifically within the conversational and expository speech domain. This task involves depicting a domestic scene featuring a mother and her two children in a kitchen setting. Participants are instructed by the examiner to describe in detail everything occurring in the image. The semantic fluency task involves asking participants to generate as many words as they can within a specific semantic category (such as animals, fruits, and tools) within a set time period. This task can be understood as individuals navigating through a mental network of related words, where they move from one concept to the next based on the way their brains organize and store words and concepts.¹³ Story

recall is a cognitive assessment method designed to evaluate verbal episodic memory and is frequently utilized to monitor declines associated with AD.¹⁴ The story recall task is administered differently across various studies. For instance, in the approach applied by Skirrow et al.,¹⁴ participants are presented with a series of stories consecutively. After each story, they are required to immediately recount the story in as much detail as possible. Sentence construction is another task used in collecting speech samples for AD analysis, as individuals with probable AD often have difficulty with tasks that require generating specific nouns or verbs. Their responses frequently show signs of word-finding errors, dysfluent speech, and numerous lexical substitutions.¹⁵

EHR dataset: The second most frequently utilized data type is the EHR, accounting for 35.59% (21 out of 59) of the distinct datasets. EHR serve as a centralized repository of patient-centered records, encompassing records from healthcare providers across the full continuum of a patient's care.¹⁶ The content of an EHR dataset can vary significantly depending on the dataset provider, encompassing a

TABLE 4 Overview of research goals, datasets, NLP techniques, and evaluation metrics in studies utilizing EHR-based datasets.

Ref	Goal	Dataset	NLP technique	Results
34	Investigation of motor signs and predicting adverse outcomes in routine patient care based on NLP pipeline	11,106 AD, vascular dementia, and mixed-type dementia	Bespoke NLP algorithm using GATE software, Adverse Drug Event annotation Pipeline	–
82	Evaluation of documentation of cognitive tests and biomarkers in EHR	A 48,912 subset of AD and ADRD cohort from University Florida Health system	A rule-based NLP pipeline based on a total of 53 rules to extract all AD and ADRD cognitive test scores and biomarkers from clinical notes	F1-score = [0.83,0.96]
83	Developing HomeADScreen and risk-identifier tool based on clinical notes	Clinical notes of 15,973 HHC patients without ADRD diagnosis and 8901 ADRD cases	Word2Vec and topic modeling	F1-score = 0.63, AUC = 0.76, sensitivity = 0.75
31	Investigation of patterns of unplanned hospital admissions of people with dementia from diagnosis until death or study end	A combination of clinical datasets of SLAM BRC clinical record interactive search, hospital episode statistics, and a death registry	GATE	–
84	Analysis of EHR data to assess the prevalence and influencing factors of cognitive measures before dementia diagnosis or AD	Deidentified EHR of 111,125 of dementia and 30,203 of AD patients	Speech recognition, speech diarization, and a transformer-based sentence encoder, the Optum NLP system	–
85	Developing a rule-based NLP algorithm for identifying social determinants of health	1000 medical notes from 231 ADRD-diagnosed patients	Custom rule-based NLP algorithm	Acc = [0.98,1.0], specificity >0.99, F1-score >0.93
32	Investigation of link between weight loss and mortality and hospitalization in various dementia subtypes	Extensive dementia care health records of 11,607 AD, VD, or DLB cases	GATE	–
86	Investigation of reasons for cost factors associated with nurse practitioner dementia care	EHR of 856 individuals enrolled in an Alzheimer's and dementia care program and EHR of 3139 dementia cases not in program	Custom NLP algorithm	–
35	Assessing link between lithium use and dementia and its subtype incidence	EHR of 29,618 patients	CRATE, GATE	–
87	Investigation of association between covert cerebrovascular disease and dementia	EHR of 241,050 individuals enrolled in Kaiser Permanente Southern California health system	MedTagger	–
88	Analysis of clinical notes to estimate risk of developing ADRD	Deidentified clinical notes between 2009 and 2017	fastText	AUC = [0.85,0.94], PPV = [45.07%,68.32%]
89	Comparing NLP models for lifestyle status detection	Clinical notes of AD patients	Rule-based NLP, PubMedBERT, Unified Medical Language System BERT, Bio BERT, and Bio-clinical BERT	Precision = 0.93, Recall = 0.93, F1-score = 0.93
90	Developing NLP system for MCI identification from clinical text	143,153 clinical notes between 2004 and 2015	Locally developed NLP model called pyTAKES	AUC = 0.67, Sensitivity = 1.7%, Specificity = 99.7%, PPV = 70%, NPV = 70.5%
37	Comparing neuropsychiatric symptoms in AD patients by clinician versus proxy-based instruments	Combination of two memory clinic cohorts of Amsterdam UMC (count = 3001, 2004–2020) and Erasmus MC (count = 646, 1993–2020) with patients having MCI, AD dementia, or mixed AD/VaD dementia	NLP algorithm for mapping clinical text fragments onto ontology concepts	AUC=[0.51,0.93]
36	Investigating atypical antipsychotics' impact on mortality in dementia patients	Patient records in SHFT database (2013–2017)	Med-7	–

(Continues)

TABLE 4 (Continued)

Ref	Goal	Dataset	NLP technique	Results
91	Developing NLP algorithm for identifying neuropsychiatric symptoms in ADRD cases	Cohort of 2.6 million free-text notes for 89,459 patients admitted to non-profit HHC agency	NimbleMiner Software	F1-score = 0.88, Precision = 0.87, Recall = 0.91
33	Exploring factors associated with suicidal ideation at time of dementia diagnosis	EHR of 18,252 dementia patients	CRIS NLP SERVICE	–
17	Investigating antipsychotic prescription risks across neuropsychiatric syndromes	Cohort of 10,106 dementia cases	GATE	–
92	Assessing predictive potential of electronic dental records for AD and Parkinson's disease	Electronic dental records of Temple University School of Dentistry, including information on 27,138 patients from January 2017 to December 2021	Custom NLP pipeline for feature extraction using named entity recognition program	Acc= 97%
18	Validating a NLP system for identifying Alzheimer's signs in clinicians' notes and predicting new ADRD diagnoses within 4 years	OASIS assessment data linked with Medicare (a federal health insurance program) claims data for a 6-month look-back and 4-year follow-up	Custom NLP pipeline for ADRD-related concept extraction using Word2Vec and NimbleMiner	–
93	ML models for detecting activities of daily living and instrumental activities of daily living impairment in dementia clinical notes	Data from Research Patient Data Repository linked to Medicare fee-for-service records of more than 700,000 individuals spanning 2007 to 2017	TF-IDF, Bio+Clinical BERT	AUC >0.97
94	Determining availability and type of caregivers for patients with dementia using medical notes	2016–2019 telephone-encounter medical notes from Michigan Medicine	Custom rule-based NLP algorithm	F1-score = 0.94, Acc = 0.95, Sensitivity = 0.97, Specificity = 0.93

Abbreviation: PPV, positive predictive value.

broad range of information. This may include antipsychotic prescriptions, diagnoses, MMSE scores, mortality data, hospitalization records, clinicians' free-text notes, patient demographics, living arrangements, informal support systems, comorbidities, symptom severity, risk factors, prognosis, therapies, medication and equipment management, pain assessments, wound care, and neurocognitive and behavioral status.^{17,18}

Literature dataset: This data type constitutes 10.17% (six out of 59) of the total datasets. A literature dataset is generated from publications within scientific databases, such as PubMed, and is employed for a variety of purposes, including detecting potential drug-to-drug interactions, summarizing relevant information, and extracting insights from extensive text corpora.

Social dataset: Social datasets make up 8.47% (five out of 59) of the total datasets and are sourced from social media platforms. These datasets typically include user reviews of AD-related mobile applications and posts from AD stakeholders on social media. This type of dataset is often utilized to support disease management, gain insights into patients' needs, investigate the experiences of family caregivers, address behavioral and psychiatric challenges, and evaluate healthcare access and barriers.

Other dataset: As previously noted, this dataset category does not fit into the other four primary dataset sources and exhibits a diverse range of characteristics. Therefore, it is classified as "Other," comprising 5.08% (three out of 59) of the total datasets. This category includes

various types of data, such as free-text survey responses related to the development of new AD treatments and experimentally verified glycoprotein datasets. Glycoproteins are a prevalent class of proteins linked to a range of diseases, including AD.¹⁹

To summarize the answer to RQ1, we identified four primary types of data utilized in AD analysis with NLP: speech datasets, EHR datasets, literature datasets, and social datasets. Additionally, we recognized an additional type of dataset called "Other" that does not align with these primary categories, covering unique or non-standard datasets employed in AD research.

3.2 | RQ2: What are the most popular datasets in each category?

In this section, we aim to present the most frequently utilized dataset within each of the five dataset types.

Concerning studies focused on speech datasets, a frequently utilized source is the open-access, password-protected DementiaBank database. DementiaBank¹² is a collaborative database encompassing multimedia interactions in English, German, Mandarin, Spanish, and Taiwanese, specifically designed to examine communication in individuals with dementia. In the extensive exploration of 43 papers focused on speech datasets, 27 of them (62%) incorporated datasets from DementiaBank, with distribution as follows: Pitt corpus (10 papers),

TABLE 5 Overview of research goals, datasets, NLP techniques, and evaluation metrics in studies utilizing social media-based, literature, and “Other” types of datasets.

Ref	Goal	Dataset	NLP technique	Results
Studies using social media-based datasets				
44	Automated identification of X users with family members affected by dementia	10,733 tweets of 8846 users	BERT, DistilBERT, RoBERTa, BioBERT, Bio+ClinicalBERT, BERTweet-Large models	BERTweet-Large: F1-score = 0.962, Precision = 0.946, Recall = 0.979
47	User review analysis of 10 selected mobile apps for individuals with Alzheimer's	A collection of 1675 user reviews	TF-IDF, SVM, LR, and RF	SVM and TF-IDF: F1-score = 99.46, Acc = 99.43, Precision = 99.3, Recall = 99.70
48	Uncovering barriers in Alzheimer's patient journey through social media insights	225,977 AD-related social media posts from 112 public sources	RLS Reveal software platform	–
45	Analyzing emotional valence and caregiving perspectives based on Korean tweets on dementia/Alzheimer's	12413 Korean tweets on dementia/Alzheimer's	Afinn, Syuzhet, and Bing	–
46	Analyzing shifts in topics and sentiment in dementia/Alzheimer's caregiver tweets before and during COVID-19 pandemic	58,094 dementia-Alzheimer's-caregiver tweets	Topic modeling and sentiment analysis	–
Studies using literature datasets				
38	Leveraging knowledge graphs for insights into probable drug-to-drug interactions for AD and lung cancer	Biomedical publications from PubMed and disease ontology, Gene Ontology, and MeSH hierarchy	Project iASIS Open Data Graph generation pipeline, SemRep	–
39	Identify AD risk factors, potential therapeutics targeting these risk factor pathways	PubMed19 and Medical Subject Headings (MeSH) thesaurus, DrugBank, Therapeutic Target Database repositories	MeSH term extraction, manual filtering, custom relevance and confidence scoring based on search tool for retrieval of interacting genes/proteins	–
40	Enhance efficiency of literature-based discovery for end users	Corpus of 33+ million articles from PubMed	SemNet 2.0 framework	–
41	Developing automated NLP pipeline for extraction of information of randomized controlled trial (RCT) abstracts	EBM-NLPmod dataset, 150 coronavirus disease 2019 (COVID-19) RCT abstracts, 150 AD RCT abstracts	Named entity recognition (NER) model, prompt-based learning using hierarchical sequential labeling network, BERT, PubMedBERT, BioBERT, BioM-ELECTRA	Proposed prompt-based sentence classification model using PubMedBERT: F1-score = 0.962, Recall = 0.962, Precision = 0.963
42	Evaluation of semi-automatic method of generating knowledge graphs from biomedical texts in the scientific literature	A set of scientific paper abstracts on AD	PoS taggers, constituency and dependency parsers, and NER, pattern recognition using Stanford's TokensRegex, abbreviation detection utility for biomedical terms using ScispaCy, constituency parsing, dependency parsing, co-reference resolution, Stanford's neural co-reference model, sentence simplification	–
43	Creating a user-friendly web portal to access gene-specific AD information based on PubMed database	Publication information for 9983 genes from AD and other neurodegenerative disease-related studies from PubMed	Punkt sentence tokenization models, NLTK package, PoS tagging	–

(Continues)

TABLE 5 (Continued)

Ref	Goal	Dataset	NLP technique	Results
Studies using "Other" datasets				
49	Developing patient-centric app for pre-dementia AD clinical trials	Set of 80,000+ free text answers to self-reported clinical survey	GloVe	–
50	Integrating machine efficiency and human intelligence for clinical trial eligibility criteria conversion	Demographic and feature-specific questionnaires	Automated eligibility criteria prescreening, Valx System, EliIE System, Criteria2Query 1.0, Att-BiLSTM Model, BERT, BioBERT, and ClinicalBERT, DILBERT, PubMedBERT	Negation scope detection using Criteria2Query 2.0: Acc = 0.924, F1-score = 0.922, Precision = 0.963, Recall = 0.884
19	Enhancing prediction of glycosylation and glycation sites	Two public glycosylation and glycation site prediction datasets: Ngly and Kgly datasets	BERT, ProtBert-BFD, ProtBert, ProtAlbert, ProtXLNet (transformer model designed for protein sequences), ESM-1b, TAPE A transformer model designed for protein sequences	Proposed PTG-PLM model using ProtBert-BFD embedding: Acc = 0.64, Recall = 0.67, Precision = 0.64, F1-score = 0.65, AUC = 0.64, MCC = 0.28

ADReSS (15 papers), ADReSSo (three papers), and the Wisconsin Longitudinal Study (WLS) (one paper). Notably, one study²⁰ analyzed both Pitt and ADReSS datasets, while Guo et al.²¹ incorporated both ADReSS and WLS datasets.

The Pitt corpus was compiled over a 5-year period from 1983 to 1988 every year as part of the Alzheimer Research Program at the University of Pittsburgh.¹² The Pitt corpus comprises speech recordings from 104 HC, 208 dementia participants, and 85 individuals with an unknown diagnosis. Eligibility requirements included being over the age of 44, having completed at least 7 years of education, not having a history of neurological disorders or current use of neuroleptic medications, an initial MMSE score of 10 or higher, and the ability to provide informed consent.²² This dataset covers various tasks, including the Cookie Theft picture description, word fluency, story recall, and sentence construction tasks.

The ADReSS dataset, introduced as part of the Alzheimer's Dementia Recognition through Spontaneous Speech Challenge at the INTER-SPEECH 2020 conference, provides a standardized dataset for evaluating various methods of automated AD detection based on spontaneous speech. As a subset of the Pitt corpus, the ADReSS dataset is meticulously balanced with respect to age and gender, comprising recordings from 78 participants with AD and 78 HC. The dataset includes both audio recordings and transcriptions of participants' speech, specifically focusing on the Cookie Theft description task from the Boston Diagnostic Aphasia Exam.

The ADReSSo dataset, introduced as part of the ADReSSo Challenge 2021, serves as a benchmark for analyzing spontaneous speech in AD research. This dataset is acoustically pre-processed and balanced with respect to age and gender, making it ideal for comparing various approaches to AD recognition in spontaneous speech. The dataset consists of two parts. The first set includes 105 recordings of AD patients engaged in a semantic fluency task, while the second set comprises 237 recordings of both AD patients and HC participants describing the Cookie Theft picture.²³

Guo et al.²¹ combined the ADReSS Challenge 2020 dataset with the public dataset collected through the Wisconsin Longitudinal study. The WLS dataset is a longitudinal study of 10,317 graduates from Wisconsin high schools in 1957, containing cognitive test results and recordings of cognitive tests such as semantic verbal fluency. The Carolina Conversation Collection dataset is another open-access resource examined by Nasreen et al.²⁴ This dataset comprises over 200 interviews with elderly individuals experiencing 12 chronic diseases, as well as over 400 interviews with those facing cognitive impairment, collected in 2011. Access to this dataset requires an application process.

Several other studies tapped into non-English speech datasets, including Japanese datasets,^{25,26} a German dataset,¹³ a Greek dataset,²⁷ a Nepali speech dataset created using the DementiaBank dataset,²⁸ and a French dataset.²⁹ Two studies by Liu et al.^{20,30} utilized an open-access Chinese dataset from the Predictive Challenge of AD organized by iFlytek in 2019. This dataset encompasses 111 individuals in the HC group and 68 individuals in the AD group. Based on our analysis, most datasets used in the speech-based reviewed articles (36 out of 43) are in English. Some studies utilized multiple datasets in their analysis. For example, one study by Liu et al.²⁰ used both English and Chinese datasets, while Lindsay et al.²⁹ utilized English and French datasets.

Among the 79 studies reviewed, 22 employed EHR datasets. These EHR datasets were gathered from various locations: seven from the UK, 14 from the US, and one from the Netherlands. Notably, the South London and Maudsley NHS Foundation Trust (SLaM) clinical records, a major European provider of dementia and mental health services, were prominently featured, appearing in five of the 21 studies.^{17,31–34} Since 2006, SLaM has served 1.36 million residents across four South London boroughs (Lambeth, Lewisham, Southwark, and Croydon) with a comprehensive electronic health system. The Clinical Record Interactive Search (CRIS) platform provides research access to over 400,000 anonymized health records from SLaM, supported by a robust governance framework.¹⁷ Additionally, two studies, by Chen et al.³⁵ and

Phiri et al.³⁶, utilized EHR datasets from the secondary care mental health services of the Cambridgeshire and Peterborough NHS Foundation Trust (CPFT) and the Southern Health NHS Foundation Trust (SHFT) from the UK. In the US, EHR datasets are sourced from different institutional databases such as the University of Florida health system, the Outcome and Assessment Information Set (OASIS), Michigan Medicine EHR database, Kaiser Permanente Southern California health system, a the clinical data repository of the University of Minnesota. The OASIS comprises standardized data elements developed for the purpose of assessing and comparing patient outcomes within home healthcare environments. EHR data in the Netherlands have been collected from two sources: the Alzheimer Center Amsterdam at the Amsterdam University Medical Centers, spanning the period from March 1993 to December 2020, and the Alzheimer Center Erasmus MC at the Erasmus MC University Medical Center, covering data from January 2004 to April 2019.³⁷

Turning to literature datasets assessed in the six studies,^{38–43} four drew on data from PubMed, utilizing its extensive repository of publications and abstracts. These studies employed NLP pipelines to extract and analyze biomedical information from PubMed. PubMed, established in 1996 and maintained by the National Center for Biotechnology Information (NCBI) at the US National Library of Medicine (NLM), is a publicly accessible database offering over 36 million citations and abstracts of biomedical literature.

Among the five studies that analyzed social datasets, three primarily used X (formerly Twitter) as their data source,^{44–46} where researchers developed and annotated datasets by extracting tweets related to dementia or AD. Another study by Alroobaea et al.⁴⁷ analyzed comments from 10 applications designed for AD patients. Additionally, Tahami et al.⁴⁸ examined posts from AD stakeholders collected from 112 publicly accessible social media platforms indexed by major search engines such as Google and Bing.

The minor dataset category, labeled “Other,” comprises three out of the 79 reviewed papers. This category includes studies utilizing questionnaires^{49,50} and public protein language datasets for glycosylation and glycation site prediction.¹⁹ A comprehensive description of all datasets employed in the reviewed papers is detailed in the dataset column in Tables 3–5.

To address the second research question succinctly: ADReSS is the most frequently analyzed dataset among studies focusing on speech datasets. In the context of EHR datasets, SLaM, a prominent European EHR provider, is the dataset most commonly utilized. For studies concentrating on social data, X is the most frequently examined source. Regarding literature datasets, PubMed is the primary resource employed. In the “Other” category, the dataset sources exhibit considerable variability.

3.3 | RQ3: What are the research goals?

In this section, we briefly discuss the objectives of the papers reviewed. For detailed research goals and corresponding citations for all reviewed papers, please refer to the goal column in Tables 3–5.

The primary goal in the reviewed papers, particularly those utilizing speech datasets (40 out of 79 studies), was AD detection through classification tasks. Studies in Table 3 mainly extracted features from speech data to distinguish between AD and HC and to predict cognitive scores, such as the MMSE score. Another prevalent research aim involved identifying the optimal data modality for automatic AD detection – whether text, audio, or a combination of both – and the most effective methods for integrating these modalities. The findings of three reviewed papers^{53,56,58} indicate that combining linguistic and acoustic features enhances performance compared to unimodal approaches in AD detection. However, these studies report mixed conclusions regarding which modality – linguistic or acoustic – is more effective. For instance, Pappagari⁵⁸ found that linguistic models typically outperform acoustic ones in detecting AD and estimating MMSE scores. However, under certain conditions, acoustic models can outperform linguistic models, particularly when errors are present in automatic speech recognition transcriptions.⁵⁸

In studies analyzing EHR datasets (referenced in Table 4), the primary focus has been on employing rule-based NLP pipelines to extract risk factors associated with AD. These factors include motor issues, cognitive test results, biomarkers of AD and AD and related dementias (ADRD), uncooperative behavior, delusions or hallucinations, and neuropsychiatric symptoms. Furthermore, studies in this category explored the association between these factors and adverse outcomes in AD or dementia patients, with adverse outcomes being variably defined across studies ranging from death and hospitalization to weight loss.

The primary focus of studies assessing literature datasets (referenced in Table 5) has been the development and optimization of NLP pipelines aimed at summarizing information and generating insights from large text-based corpora derived from AD-related biomedical publications. These efforts are directed toward various objectives, such as identifying potential drug-to-drug interactions, uncovering preventive strategies for AD, and extracting AD-related gene-specific information. The underlying motivation in this area is the critical need for AD researchers to review previous work and remain up to date with the rapidly expanding body of AD literature, a task that is both essential and increasingly challenging.⁴³

Research on social datasets (referenced in Table 5) primarily focuses on understanding the needs of patients, providing disease management support outside of hospitals, and exploring the experiences and mental health needs of family caregivers of individuals with ADRD.

The studies categorized under the “Other” dataset (referenced in Table 5) had a wide range of focuses, including the prediction of specific proteins associated with AD using protein language models, the evaluation of patient awareness regarding their brain health in the predementia period, and various aspects of patient management.

In summary, the primary goals in the context of AD analysis span several key areas: automatic AD detection using speech data, identification of AD risk factors through EHR, summarization of AD-related literature, and examination of the social burden on AD patients and caregivers as expressed on social media platforms.

3.4 | RQ4: What are emerging trends in the field?

An emerging trend in AD detection using speech datasets is the shift toward more complex multiclass classification tasks. Only two of the 43 studies^{25,76} listed in Table 3 addressed the more complex task of multiclass classification, differentiating between AD, MCI, and HC. However, the remaining studies concentrated on binary classifications, such as AD versus HC or AD+MCI versus HC. The multiple-classification approach holds greater practical significance, as MCI represents a mild stage of cognitive impairment that may or may not progress to AD, making it challenging for models to differentiate between AD, MCI, and HC. Importantly, if a model can accurately detect MCI, it offers a critical opportunity for clinicians to take preventive measures to delay the onset of AD.

Five studies^{14,25,71,74,76} focused on developing automatic diagnostic tools using longitudinal speech datasets. Analyzing longitudinal speech data is another developing direction, as it may uncover temporal patterns in AD speech pathology, offering insights that single-time-point analyses cannot provide. The type of features used for AD detection is another crucial aspect. Some studies^{21,67,75} relied solely on embeddings extracted from pretrained large language models, which operate as black boxes. However, a rising tendency involves comparing and combining various feature types, such as linguistic, acoustic, and psycholinguistic features, alongside embeddings from pretrained models, as demonstrated in the studies by refs. [54, 55].

Data augmentation techniques are also attracting attention in the AD detection domain, particularly for speech datasets, which are often small and imbalanced, potentially compromising model reliability. Two studies^{59,81} explored data augmentation methods, with Runde et al.⁸¹ employing standard numerical techniques, over-sampling, after converting text to numerical representations, while Woszczyk et al.⁵⁹ have utilized audio-based or text-based augmentation methods, such as paraphrasing and lexical substitution. Furthermore, a recent trend involves directly inputting speech audio files into models, bypassing manual transcription, as investigated by Runde et al.⁸¹. They noted that for an automatic AD detection model to be practical, it must be capable of processing real-time speech and making continuous inferences. This necessitates research into the usability of automatic audio-to-text transcription tools and audio enhancement techniques.

An evolving direction in EHR-based research involves leveraging these datasets to identify social determinants of health, such as housing, transportation, food, medication access, and financial difficulties, in order to address the social needs of AD patients. However, a limitation in the current body of research is the predominant reliance on datasets collected from specific locations, which may limit the generalizability of findings. Expanding studies to include datasets from diverse geographical regions could provide a more comprehensive understanding of AD-related risk factors and outcomes.

According to Shao et al.,⁹⁵ a noteworthy advancement is the development of a transformer-based model architecture capable of jointly learning from both longitudinal and non-longitudinal inputs, which has demonstrated high performance in predicting AD outcomes. This study underscores the significant potential of EHR data not only for identifying AD risk factors but also for improving diagnostic accuracy.

Mahmoudi et al.⁹⁴ showed that EHR studies also have the potential to inform better care planning within health systems by identifying factors such as caregiver availability for dementia patients. Given the lack of effective therapies for AD, prevention through lifestyle changes and interventions has become increasingly important. Analyzing EHR data to understand the impact of lifestyle on AD, as examined by Shen et al.,⁸⁹ represents a valuable approach for advancing our understanding of how lifestyle modifications can influence the progression and prevention of AD.

A development in the AD literature summarization field is the work by Rossanez et al.,⁴² which facilitates a deeper understanding of AD and potential treatments by linking entities and relationships represented in knowledge graphs derived from biomedical publications to concepts from existing biomedical ontologies available on the web. However, this approach has certain limitations, such as the lack of direct accuracy assessment and the absence of qualitative analyses involving AD experts, which could further validate the findings.

In the domain of social datasets, a growing focus is on demonstrating how social media, such as X, can be harnessed to gain insights into the experiences and needs of dementia family caregivers. Klein et al.⁴⁴ identified a notable gap in leveraging social media for interventions. In response, they developed an annotated dataset and benchmark classification models to automatically identify X users who are family caregivers of AD patients, showcasing the platform's potential to facilitate large-scale interventions. However, this study assumes that all identified users are actual caregivers, which may not always be the case, highlighting a limitation in the definition used for family caregivers on X. Building on this, Sunmoo⁴⁶ provided an in-depth analysis of sentiments and topics expressed in AD and dementia-related tweets during the COVID-19 pandemic. This study provided critical insights into the mental health needs of family caregivers, suggesting the potential for social media analysis to extend beyond the pandemic period. Such research lays the groundwork for ongoing studies that could provide a more comprehensive understanding of the challenges and experiences faced by AD patients and their caregivers, illustrating the valuable role of social media data in addressing the broader impacts of AD.

Exploration of studies utilizing datasets in the "Other" category revealed a distinct trend, as exemplified by the work of Alkuhlani et al.,¹⁹ which integrates transformer-based language models into the proteomics field. Specifically, these researchers developed protein language models for the representation of protein sequences. This innovative approach adapts NLP principles to analyze protein sequences, paving new paths for biomarker discovery in AD research. This study demonstrates the potential of repurposing computational linguistic models to significantly advance biomedical research.

3.5 | RQ5: What NLP approaches are employed in AD analysis?

In the reviewed articles, researchers utilized diverse NLP techniques tailored to the datasets and research questions at hand. In what follows, we provide a concise overview of these methodologies, offering

a clear understanding of the approaches employed. Detailed information on the methodologies applied in each study is provided in the NLP technique column in Tables 3–5.

3.5.1 | Traditional linguistic features

Various hand-coded linguistic features have been employed to analyze speech in patients with AD. Chandler et al.⁸⁰ evaluated poverty of speech by utilizing the type-token ratio ($\frac{\text{count}(\text{word types})}{\text{count}(\text{word tokens})}$), part-of-speech (PoS) tag counts and Brunet's index ($\frac{\text{count}(\text{word tokens})}{\text{count}(\text{word types})^{0.165}}$) and poverty of content by measuring through the content density feature ($\frac{\text{count}(\text{verbs} + \text{nouns} + \text{adjectives} + \text{adverbs})}{\text{count}(\text{word tokens})}$). Furthermore, they evaluated verbigeration and language fluency by analyzing phrase- and word-level repetitions, including counts of filler words like “um” and “ah.” Syntactic complexity was examined using sentence parse trees and speech graphs. Additionally, semantic features were assessed by measuring coherence through cosine distances between adjacent text windows using embeddings.

Paralinguistic features, which refer to aspects of communication not related to the actual spoken words but rather to non-speech sounds and socially significant gradations that influence the meaning of utterances, were also examined for AD detection based on patients' speech.⁹⁶ These features are not confined to a rigidly defined set of units and include elements such as pitch contours, prominence, tempo, and rhythm.⁹⁶ For example, Tavabi et al.⁵¹ extracted paralinguistic features, including speech time variables such as total speaking time, the mean and standard deviation of word and pause lengths (in seconds), fraction of time spent pausing, total section time, and the fraction of that section in which the participant was speaking.

Psycholinguistic features have also been explored for diagnosing AD, as shifts in emotional expression among ADRD patients can occur alongside cognitive decline.⁷¹ These changes may affect the psycholinguistic elements of their speech, potentially impairing their ability to communicate effectively and convey their needs. Psycholinguistic indicators in speech can be evaluated through both non-verbal vocalizations (such as non-word utterances) and semantic language.

The reviewed articles also investigated cross-modal linguistic and task-specific features. Cross-modal linguistic features involve examining the interactions between language and acoustic elements, such as correlating acoustic features with linguistic features and analyzing time-aligned language and acoustic features.⁸⁰ Task-specific features, for example, include counting the unique animals mentioned in the animal fluency task.

3.5.2 | Sparse text representation

While sparse text representation is fundamentally a type of linguistic feature, we categorize it under sparse text representation due to the nature of the features generated, which result in sparse matrices (often containing zero elements) owing to the underlying algorithms. In the

reviewed studies, two primary sparse text representation techniques were frequently used for feature extraction, particularly with speech datasets, to convert text into numerical formats. The first one is the bag of words (BoW) technique, which represents a document as a vector of word frequencies, disregarding grammar and word order. This approach is commonly implemented using the CountVectorizer class in scikit-learn.⁹⁷ The second one is the term frequency-inverse document frequency (TF-IDF) approach, which extends the BoW model by incorporating a weighting mechanism that accentuates more informative words. This method evaluates the significance of a word by considering its commonality or rarity across the entire corpus, assigning greater weight to rarer words.⁹⁷

3.5.3 | Static dense text representation

Dense text representation is a type of linguistic feature usually generated through neural networks (NNs), allowing for meaningful interpretation of the resulting features. Dense representation involves text encoding where most elements in the resulting vectors are non-zero. These vectors typically have lower dimensionality compared to the full vocabulary size and are capable of capturing rich semantic information about words, phrases, or entire documents.⁹⁷ The reviewed articles frequently employed dense representation methods such as fastText, Word2Vec, and GloVe.

GloVe learns a single fixed embedding for each word in the vocabulary. GloVe generates dense word embeddings by factorizing a word co-occurrence matrix, thereby capturing global statistical relationships between words and encoding semantic meaning into vectors.⁹⁸ In the study by Jain et al.¹² both domain-specific and general-purpose versions of GloVe word embeddings were generated. The domain-specific embeddings were trained on a specialized corpus relevant to the area of interest, while the generic embeddings were trained on large, general-purpose corpora. These embeddings were then used for AD versus HC classification on the Pitt corpus. The results indicated that domain-specific embeddings outperformed generic word embeddings.

Word2Vec is an embedding technique that generates dense word embeddings by training on large text corpora. It utilizes two main algorithms: Skip-Gram, which predicts context words from a target word,⁹⁸ and Continuous BoW, which predicts a target word from its surrounding context. This model learns word vectors such that words appearing in similar contexts have similar embeddings, effectively capturing semantic relationships within the vector space. In the study by Saranpää et al.,¹³ Word2Vec was applied to extract feature-rich semantic vectors from participants' speech. These vectors were then visualized by reducing the 300-dimensional space to a two-dimensional map using a data-driven dimensionality reduction technique. This visualization helped identify and label semantic subcategories, providing insights into how patients with prodromal and early AD navigate the semantic space during fluency tasks compared to control participants.

fastText extends Word2Vec by addressing challenges related to unknown words and sparsity in morphologically rich languages. fastText represents each word as a combination of the word itself and its

constituent n-grams. A skip-gram model is then used to learn embeddings for each n-gram, with the word's final representation being the sum of these embeddings.⁹⁷ The study by Jain et al.¹² demonstrated that domain-specific fastText embeddings yielded the best results for dementia versus HC classification using the Pitt corpus.

3.5.4 | Contextualized dense text representation

Contextualized embedding models, which are a type of dense text representation, were frequently used to generate text features. In these models, the vector representation of a single word varies depending on its context. There was a particular focus on different versions of Bidirectional Encoder Representations from Transformers (BERT), such as DistilBERT, DistilRoBERTa, RoBERTa, Bio-clinical BERT, ALBERT, PubMed BERT, BioBERT, ProtBERT-BFD, ProtBERT, and ProtALBERT.¹⁹ Each of these models builds upon BERT and is tailored for specific tasks, domains, or improvements in efficiency through different training data, numbers of layers, attention heads, or hidden sizes. Twenty-two studies in the review examined BERT-derived embeddings either individually or in conjunction with other acoustic or linguistic features for AD detection through classification tasks (for a detailed description, see Tables 3–5). For instance, Shen et al.⁸⁹ used BERT models to classify lifestyle status in an EHR dataset of AD patients, while Klein⁴⁴ applied BERT to create a pipeline for identifying dementia-related families and caregivers on X. BERT is a transformer-based model that pretrains deep bidirectional representations by considering both left and right contexts across all layers, using large volumes of unlabeled text.⁹⁹

ESM-1B, TAPE, and ProtXLNet are pretrained transformer-based protein language models used to encode peptide sequences in ref. [19]. ESM-1B, developed by Facebook AI Research, is based on the BERT language model and trained on the UniRef50 dataset. TAPE, another BERT-based model, was trained on the Pfam database, which includes approximately 31 million protein sequences. ProtXLNet is built on the XLNet language model and trained on the UniRef100 database, containing around 216 million protein sequences.¹⁹ The embeddings obtained from these models were input into a conventional NN model to predict glycosylation and glycation sites, as detailed in ref. [19]. These predictions are vital because modifications at glycosylation and glycation sites are linked to various diseases, including AD.

GPT-3, OpenAI's Generative Pre-trained Transformer model, generated embeddings employed in two reviewed articles^{79,81} for AD detection. The embeddings of AD and HC participants were fed into machine learning (ML) classifiers for AD versus HC classification. Agbavor⁷⁹ demonstrated that the text embeddings generated by the GPT-3 model significantly outperformed conventional feature-based approaches.

3.5.5 | Acoustic features

In the reviewed articles, a wide range of acoustic features were utilized for AD detection.

Yeung et al.⁶⁹ extracted several acoustic features, including properties of the sound wave, speech rate, and the number of pauses, using automatic speech analysis tools. The study by Tavabi et al.⁵¹ computed a comprehensive set of acoustic features, including Mel-Frequency Cepstral Coefficients (MFCCs), fundamental frequency (F0), voicing probability, local jitter, difference of differences jitter, local shimmer, harmonics-to-noise ratio (HNR), power spectrum (audspec), relative spectral transform (RASTA), zero crossing rate, and root mean square (RMS) energy. These features were extracted using the Massachusetts Institute of Technology's featurization algorithm and the openSMILE toolkit.⁵¹

Pappagari et al.⁵⁶ employed x-vectors to characterize the speech signals of AD patients for AD versus HC classification. They noted that x-vector embeddings capture various speech characteristics including emotion, speaking rate, gender, and various articulatory, phonatory, and prosodic characteristics, which can be leveraged to identify neurological disorders such as Parkinson's disease.⁵⁶ X-vectors⁵⁶ are acoustic features primarily used for speaker recognition, generated by a deep NN (DNN) trained to distinguish between speakers. They convert variable-length speech utterances into fixed-dimensional embeddings, making them particularly useful for tasks like speaker verification and, in this context, detecting neurodegenerative diseases.

3.5.6 | NLP digital tools

A variety of NLP digital tools were employed in the reviewed articles to organize and analyze large volumes of unstructured text, with each tool typically tailored for specific text annotation purposes.

Emotional text annotation: Tools such as Affine, Syuzhet, and Bing were utilized for sentiment analysis in ref. [45] to assign emotional valence scores to tweets regarding AD. Affine and Syuzhet are dictionary-based algorithms, while Bing operates on a ML algorithm.

Medical text annotation: In ref. [48], RLS Reveal, a taxonomy-based and semantic text-mining NLP algorithm, was applied to extract medical, clinical, and functional terms from unstructured clinical data. Med-7, a transferable clinical NLP model for EHR, was employed in ref. [36] to identify medications within patient notes. NimbleMiner, an open-source nursing-sensitive NLP system based on word embeddings, is utilized to annotate neuropsychiatric symptoms in EHR datasets, as demonstrated in refs. [18,91]. The General Architecture for Text Engineering (GATE) toolkit was the most frequently utilized NLP toolkit for extracting relevant information from free-text clinical documentation, as seen in studies such as in refs. [31,32,34,35]. This toolkit can address a variety of text processing challenges. MedTagger, an open-source NLP pipeline, was applied in ref. [87] to identify individuals with covert brain infarction and white matter disease reported in neuroimaging reports. Additionally, the Adverse Drug Event annotation Pipeline (ADEPt), a semantically enriched pipeline for extracting adverse drug events from free-text EHR, was employed in ref. [34] to detect specific motor signs from clinical free text.

Custom NLP pipelines: The CRIS NLP SERVICE was leveraged in ref. [33] to extract symptoms, treatments, and patient outcomes from

EHR, enhancing mental health research by utilizing NLP techniques to mine unstructured clinical records. In a separate study, Chen et al.⁸² developed a custom rule-based NLP pipeline using the BRAT annotation tool to extract the most frequent cognitive measurements from clinical notes. A study by Zolnoori et al.⁸³ applied two NLP methods to develop pipelines for identifying potential risk factors: a semi-supervised Word2Vec method to extract specific predefined home health ADRD risk factors and unsupervised topic modeling to identify emerging topics from clinical notes. Finally, Wu et al.⁸⁵ developed a custom rule-based NLP algorithm to identify seven domains of social determinants of health, where the rule-based algorithm outperformed DL and regularized logistic regression (LR) methods.

3.5.7 | Machine learning models

A wide range of ML models are applied for regression and classification tasks in the reviewed papers. Examples of classification tasks include detecting AD or distinguishing between AD and HC based on speech or lifestyle data derived from EHR. Studies commonly employed ML classifiers, including support vector machine (SVM), random forest (RF), K-nearest neighbor (KNN), LR, convolutional neural network (CNN), and recurrent neural network (RNN). Detailed descriptions of ML models can be found in the NLP technique column of Tables 3–5.

In terms of regression tasks, such as predicting cognitive scores based on speech datasets or analyzing AD risk factors from EHR datasets, linear regression⁵⁷ and Cox regression³⁴ were the most frequently utilized methods. Aryal et al.⁵⁷ utilized 10 distinct linear regression models, leveraging a combination of handcrafted and learned acoustic-linguistic features to assess and score mental status. Al-Harrasi³⁴ employed the Cox regression model to explore the associations between survival, hospitalization, and a range of explanatory variables. The Cox regression model, often referred to as the proportional hazards model, is a statistical tool used to investigate the relationships between a time-to-event outcome and a set of covariates.

3.6 | RQ6: How effective are NLP approaches in AD analysis?

In the field of AD detection using speech data, studies typically evaluate performance metrics such as F1-score, recall, accuracy, precision, and area under the curve (AUC). These metrics are summarized in the results column of Table 3, where the most frequently reported metric is accuracy, noted in 32 out of 43 studies. The accuracy scores ranged from 0.65 to 0.99, with 27 studies reporting scores above 0.80.

The most effective results often employ transformer-based methods for contextualized embedding production, either alone or in conjunction with classical machine-learning approaches. Notably, Runde et al.⁸¹ reported an accuracy and an F1-score of 0.99 for differentiating AD from HC using a combination of GPT embeddings and a KNN classifier. However, non-public details of GPT models restrict their research utility. Transformer models, particularly various versions of BERT, are

extensively studied^{21,54,67,71,73,75,77} due to their capability to capture rich semantic information and effectively encode context-dependent meanings. However, a major limitation of transformer models is their “black box” nature, which poses challenges for interpretation and its high demand for computational power and extensive training data.

Among ML classifiers, SVM showed the best results in eight studies, followed by LR in five, with RF-, KNN-, and DL-based classifiers such as CNN, RNN, and BLSTM also being employed. In terms of static word embeddings, domain-specific adaptations of Word2Vec and fast-Text demonstrated superior performance, with the accuracy reaching 0.96 in ref. [28] and 0.91 in ref. [12], respectively. However, these static embedding models are limited compared to contextual models such as BERT and GPT, as they cannot adjust representations dynamically based on the text's context, potentially limiting their effectiveness in variable scenarios. TF-IDF, a method for sparse text representation, proved effective, as evidenced by Linu et al.³⁰ who achieved an accuracy and F1-score of 0.98 using TF-IDF features combined with a KNN classifier. This approach is advantageous in terms of computational resources and runtime, and it facilitates interpretation since it relies solely on the frequency of words in the text, making it particularly suitable for clinical applications. In contrast, TF-IDF disregards syntax and word order, leading to the potential loss of context and meaning. Its focus on rare words might overemphasize terms that are unique yet contextually irrelevant, potentially skewing the analytical outcomes.

Traditional linguistic features are frequently used alongside classical ML classifiers in AD detection. These features offer a clear, quantitative analysis of speech characteristics such as lexical diversity and syntax and are advantageous for their explainability. Conversely, they might fall short of capturing nuanced changes in semantics and pragmatics, in contrast to dense text representation models. Traditional linguistic features demonstrate moderate to low effectiveness in accurately detecting AD, as evidenced by the results shown in Table 3 and highlighted by ref. [80]. This shortcoming could be attributed to the methods used for feature selection, the relevance of the selected features, or the limited size of datasets.

Tavabi et al.⁵¹ demonstrated that paralinguistic features were the primary contributors to distinguishing between HC and dementia cases. A disadvantage of analyzing paralinguistic features is their susceptibility to external influences, such as stress and excitement, which might not be directly related to AD. Furthermore, analyzing psycholinguistic cues can enhance the accuracy of screening methods, according to Zolnoori et al.⁷¹ However, psycholinguistic analysis can be complex and may depend heavily on the context of the conversation. In accordance with results reported in several studies,^{53,56,62,80,81} the analysis of acoustic features such as x-vectors and MFCC features provides comprehensive information on speech dynamics and quality and has proven beneficial. In contrast, acoustic features demand high-quality audio recordings and can be susceptible to background noise and recording inconsistencies, which may complicate data collection and analysis in less controlled environments.

In studies utilizing EHR, 11 employed standard evaluation metrics such as F1-score, accuracy, and AUC, as referenced in the results column of Table 4. Six of them reported F1-scores ranging from 0.63 to

0.96, highlighting their models' capabilities in extracting relevant information from unstructured health data. The remaining studies, which did not explicitly present metrics, relied on NLP digital tools for data extraction, followed by various statistical analyses such as significance testing and Cox regression, indicating a descriptive rather than predictive approach common in EHR data analysis. A key advantage of Cox regression is that it does not require specific distributional assumptions and can effectively incorporate both baseline and time-varying clinical factors, making it ideal for analyzing complex data.¹⁰⁰

Regarding analyses of social datasets, Alroobaea et al.⁴⁷ achieved an accuracy and F1-score of 0.994 using TF-IDF features and a SVM classifier for classifying user reviews into positive and negative categories. Furthermore, a study by Klein et al.⁴⁴ demonstrated high performance with BERTweet-Large, attaining an F1-score of 0.962 for identifying tweets related to family members with dementia. The remaining three studies in this category did not explicitly report results. For literature datasets, the only reported study is the one by Hu et al.⁴¹ which utilized a prompt-based approach with PubMedBERT to extract relevant information from RCT abstracts, achieving an F1-score of 0.962.

In the "Other" category, Fang et al. achieved an accuracy of 0.924 and an F1-score of 0.922 in negation scope detection using their proposed model, Criteria2Query 2.0. This model integrates medical entity recognition and concept mapping from free-text eligibility criteria, employing a transfer learning approach with PubMedBERT. Additionally, Alkuhlani et al.¹⁹ utilized CNNs and embeddings generated from six protein language models, including ProtBert-BFD, ProtBert, ProtAlberty, ProtXlnet, ESM-1b, and TAPE for protein site prediction. The model leveraging ProtBert-BFD embeddings demonstrated the highest performance, achieving an accuracy of 0.64 and F1-score of 0.65. The reviewed papers collectively demonstrate that the transfer-learning approaches using BERT models are effective not only in AD detection using speech but also in analyzing EHR and in social-driven and literature datasets related to AD.

4 | DISCUSSION AND FUTURE WORK PROPOSALS

To contribute to the progress of the field, enhance comparability across studies, and improve research reproducibility, we make the following recommendations for future studies:

Improving remote monitoring: Only three studies^{14,27,80} have focused on developing remote, automated AD detection tools using speech. According to ref. [27], substantial work remains to be done to make these models clinically viable, particularly regarding privacy and security of speech and medical data. We believe advancing models for AD patient monitoring through speech analysis could enhance both pre- and post-diagnosis stages. Remote monitoring also offers a cost-effective alternative to brain imaging, accessible via smartphones, but demands careful attention to ethical and privacy standards.

Addressing dataset limitations: Studies by refs. [53,63] highlight the limitations of small datasets and the outdated collection period of the ADReSS dataset from the mid-1980s, which may not align with modern

diagnostic standards, potentially reducing the effectiveness of contemporary ML models. A critical factor for developing reliable models is access to large, balanced datasets. Enhancing dataset utility through augmentation, merging multiple sources, and standardizing collection protocols, particularly for speech data, could foster robust model development and facilitate more accurate performance comparisons across studies.

Utilizing pretrained multilingual models: Our review of NLP techniques, as listed in the NLP technique column in Table 3, reveals that no reviewed papers employed pretrained multilingual models for AD detection using speech. We recommend utilizing models such as M-BERT, which is trained across over 100 languages, to overcome the scarcity of extensive training datasets in multiple languages. Such models could facilitate the identification of speech features that are generalizable across languages and cross-language AD detection.

Expanding classification focus: Only two studies^{25,76} conducted a multiple AD versus MCI versus HC classification task described in Section 3.4. Future research should explore multiclass classification, given its complexity and relevance over binary tasks and its potential benefits for diagnosis and preventive measures in early-stage AD.

Improving model interpretability and error analysis: Chandler et al.⁸⁰ attempted to address the interpretability issue of automatic AD detection by developing dashboards to interpret the results of multiclass cognitive decline prediction models. Additionally, Martinc et al.⁵³ emphasized the importance of detailed error analysis to comprehend the model's behavior. Future research should conduct a thorough analysis of a speech-based assessment tool to facilitate informed decision-making and clinical integration of automated AD detection tools. Focusing on these aspects may illuminate previously hidden facets of the disease, thereby offering valuable insights and enhancing models' overall reliability and effectiveness.

Exploring alternative architectures: Despite considerable progress in large language models, it is important to note that various versions of the BERT model, an encoder-only architecture, dominate among studies employing transformer architectures. Specifically, as shown in Table 3, 16 out of 43 studies primarily used different versions of the BERT model for AD detection through speech analysis. Future research could explore alternative approaches by evaluating the performance of advanced models such as the encoder-decoder or the decoder-only model, for example, Gemini, which can handle different modalities, including audio, that allows us to use audio as well as text for AD detection and is available through the Application Programming Interface.

Addressing language diversity in speech datasets: The primary language for speech sample datasets predominantly relies on English datasets, as discussed in Section 3.2. This highlights a significant gap in non-English datasets for AD diagnostic and monitoring tools. The development of models that work across various languages is essential. Notably, only two studies among the reviewed articles^{20,29} utilized datasets from more than one language, indicating a constraint on the practical applications of such a model.

Standardizing metrics: According to the result column in Tables 3–5, studies have taken varying approaches to reporting their work. We

advocate for using a standardized set of evaluation metrics in classification tasks, rather than relying on a single metric such as AUC or accuracy. Reporting a comprehensive set of metrics, including accuracy, F1-score, precision, recall, and AUC, enables more meaningful comparisons across studies and ensures balanced performance evaluation.

Training rule-based algorithms on larger datasets: As stated by Wu et al.⁸⁵ a significant challenge is training rule-based algorithms on larger datasets of social worker notes, which contain detailed insights on crucial issues such as housing and medication insecurities. Expanding datasets is crucial to enhancing the algorithm's ability to accurately identify social determinants associated with AD risk factors.

Integrating privacy-preserving generative AI tools: Another unresolved issue is the potential for improving rule-based algorithms by integrating privacy-preserving generative AI tools, as discussed in ref. [85]. Such tools could significantly enhance the handling and analysis of sensitive data, ensuring confidentiality while improving the overall effectiveness of the algorithms in identifying and analyzing critical social determinants within EHR data.

Enhancing model generalizability: Given the challenges outlined in ref. [82] related to accurately identifying AD or ADRD and monitoring their severity through EHR, there is an evident need for models that can both detect dementia and track disease progression using cognitive test results and biomarkers derived from clinical narratives. EHR datasets, which are more readily accessible and available in larger quantities compared to speech datasets, provide a crucial resource for such developments. Additionally, the potential of EHR data in biomarker analysis for dementia and AD is underscored by findings from Chen et al.³⁵ which suggest that lithium treatment may reduce dementia risk. However, the study's focus on a specific mental health patient cohort curtails its broader applicability. Expanding research to include larger and more varied populations would help further elucidate the relationship between lithium levels and dementia outcomes.

Utilizing computational NLP pipelines: Four out of six studies^{38–40,43} analyzing the literature datasets focused on a single source of publications, PubMed. While this source is valuable, there is considerable scope for broadening the range of literature sources analyzed to provide a more comprehensive view of the field. According to Vitali et al.³⁹ there is significant potential in utilizing a computational NLP pipeline to analyze existing literature systematically and rank specific therapeutics for AD prevention. This approach could help address open research questions, such as how demographic data can be effectively integrated into computational models to improve the stratification and efficacy of therapeutic interventions in disease prevention.

Enhancing social analysis of AD using NLP: The study by Alrooba et al.⁴⁷ underscores the need for more nuanced research in the social analysis of AD using NLP. Beyond binary classification of feedback as positive or negative, future research should explore specific AD applications, such as gaming or caregiving assistants, to pinpoint their strengths and weaknesses. This could include extracting significant patterns or keywords from user reviews to refine these applications. Recognizing that feedback is not always clearly positive or negative –

often being neutral or inquisitive – can provide a more accurate reflection of user sentiment. Further, Klein et al.⁴⁴ demonstrated a method for identifying dementia stakeholders on X, which presents an opportunity to extend this approach to other platforms, such as Reddit, to boost the robustness and applicability of the models.

Advancing research on dementia subtypes and caregiver insights: Sunmoo et al.⁴⁶ investigated the sentiment and topics discussed by ADRD caregivers on X before and during the COVID-19 pandemic to offer insights into patients' mental well-being during the pandemic. This study could be furthered into a post-pandemic era to obtain updated insights into patients' well-being. One more illustrative approach could be to focus on dominant dementia subtypes such as AD and vascular AD to get a better understanding of challenges specific to each subtype.

5 | CONCLUSION

In this study, we followed the PRISMA protocol to conduct a comprehensive review of publications utilizing NLP methodologies for AD analysis, covering the period from January 2020 to July 2024. We systematically synthesized and examined the findings from 79 selected studies, identifying four primary trends in AD analysis: (1) detection and monitoring of AD using speech datasets, (2) identification of AD risk factors based on EHR, (3) summarizing the current state of knowledge in AD-related publications, and (4) exploring the social burden experienced by AD patients, caregivers, and their families.

The studies leveraging speech and EHR datasets frequently employed NN, ML classifiers, and various linguistic features, while rule-based NLP approaches were primarily used for extracting relevant information from EHR data. Analyses of social datasets predominantly utilized sentiment analysis and topic modeling, whereas studies focusing on literature datasets developed custom NLP pipelines for generating knowledge graphs and conducting network analysis.

Overall, NLP techniques have demonstrated effectiveness in detecting AD, as indicated by reported performance metrics. However, significant gaps remain, as outlined in Section 4, including dataset limitations, model interpretability, and privacy concerns. To address these issues, we propose several future research directions: developing larger speech datasets for AD, enhancing monitoring and remote AD detection models, incorporating geographically diverse datasets, and integrating privacy-preserving AI tools.

ACKNOWLEDGMENTS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

CONSENT STATEMENT

Consent was not necessary for this study.

ORCID

 Arezo Shakeri  <https://orcid.org/0000-0003-3823-2619>

REFERENCES

- Duong S, Patel T, Chang F. Dementia: what pharmacists need to know. *Can Pharmacists J*. 2017; 150: 171516351769074. doi: [10.1177/1715163517690745](https://doi.org/10.1177/1715163517690745)
- Wimo A, Seeher K, Cataldi R, et al. The worldwide costs of dementia in 2019. *Alzheimers Dement*. 2023; 19(7): 2865-2873.
- Silva MVF, Loures CdMG, Alves LCV, de Souza LC, Borges KBG, Carvalho MdG. Alzheimer's disease: risk factors and potentially protective measures. *J Biomed Sci*. 2019; 26:1-11.
- Adhikari S, Thapa S, Singh P, Huo H, Bharathy G, Prasad M. A comparative study of machine learning and NLP techniques for uses of stop words by patients in diagnosis of Alzheimer's disease. In: *2021 International Joint Conference on Neural Networks (IJCNN)*. 2021:1-8. doi: [10.1109/IJCNN52387.2021.9534449](https://doi.org/10.1109/IJCNN52387.2021.9534449)
- Aditya Shastri K, Sanjay H. A. Artificial intelligence techniques for the effective diagnosis of Alzheimer's disease: a review. *Multimed Tools Appl*. 2023;83:40057-40092.
- de la Fuente Garcia S, Ritchie CW, Luz S. Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. *J Alzheimers Dis*. 2020; 78(4): 1547-1574.
- Shi M, Cheung G, Shahamiri SR. Speech and language processing with deep learning for dementia diagnosis: a systematic review. *Psychiatry Res*. 2023; 329: 115538.
- Petti U, Baker S, Korhonen A. A systematic literature review of automatic Alzheimer's disease detection from speech and language. *J Am Med Inform Assoc*. 2020; 27(11): 1784-1797.
- Patra BG, Sharma MM, Vekaria V, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Inform Assoc*. 2021; 28(12): 2716-2727.
- Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev*. 2015; 4:1-9.
- Kitchenham B, Charters SM. Guidelines for performing systematic literature reviews in software engineering. Ver. 2.3 *EBSE Technical Report*. EBSE; 2007.
- Jain M, Doshi R, Sehra V, Sethia D. Exploring the effects of different embedding algorithms and neural architectures on early detection of Alzheimer's disease. In: *ISIC*. 2021:376-383.
- Saranpää AM, Kivisaari SL, Salmelin R, Krumm S. Moving in semantic space in prodromal and very early Alzheimer's disease: an item-level characterization of the semantic fluency task. *Front Psychol*. 2022; 13: 777656.
- Skirrow C, Meszaros M, Meepegama U, et al. Validation of a remote and fully automated story recall task to assess for early cognitive impairment in older adults: longitudinal case-control observational study. *JMIR Aging*. 2022; 5(3): e37090.
- Altmann LJ. Constrained sentence production in probable Alzheimer's disease. *Appl Psycholinguist*. 2004; 25(2): 145-173.
- Hovenga E, Hovenga H, Leslie H. Emerging digital health ecosystems. *Roadmap to Successful Digital Health Ecosystems*. Elsevier; 2022:555-567.
- Mueller C, John C, Perera G, Aarsland D, Ballard C, Stewart R. Antipsychotic use in dementia: the relationship between neuropsychiatric symptom profiles and adverse outcomes. *Eur J Epidemiol*. 2021; 36: 89-101.
- Ryvicker M, Barrón Y, Song J, et al. Using natural language processing to identify home health care patients at risk for diagnosis of Alzheimer's disease and related dementias. *J Appl Gerontol*. 2024:07334648241242321.
- Alkuhlani A, Gad W, Roushdy M, Voskoglou MG, Salem A-BM. PTG-PLM: predicting post-translational glycosylation and glycation sites using protein language models and deep learning. *Axioms*. 2022; 11(9): 469.
- Liu N, Yuan Z, Tang Q. Improving Alzheimer's disease detection for speech based on feature purification network. *Front Public Health*. 2022; 9: 835960.
- Guo Y, Li C, Roan C, Pakhomov S, Cohen T. Crossing the "cookie theft" corpus chasm: applying what BERT learns from outside data to the ADReSS challenge dementia detection task. *Front Comput Sci*. 2021; 3: 642517.
- Haider F, De La Fuente Garcia S, Albert P, Luz S. Affective speech for Alzheimer's dementia recognition. In: *LREC: Resources and Processing of Linguistic, Para-linguistic and Extra-linguistic Data From People With Various Forms of Cognitive/Psychiatric/Developmental Impairments (RaPID)*. 2020;67-73.
- Zhu Y, Obyat A, Liang X, Batsis JA, Roth RM. WavBERT: exploiting semantic and non-semantic speech using Wav2Vec and BERT for dementia detection. *Interspeech*. 2021; 2021: 3790-3794.
- Nasreen S, Rohanian M, Hough J, Purver M. Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features. *Front Comput Sci*. 2021; 3: 640669.
- Yamada Y, Shinkawa K, Nemoto M, Nemoto K, Arai T. A mobile application using automatic speech analysis for classifying Alzheimer's disease and mild cognitive impairment. *Comput Speech Lang*. 2023; 81: 101514.
- Momota Y, Liang K-C, Horigome T, et al. Language patterns in Japanese patients with Alzheimer's disease: a machine learning approach. *Psychiatry Clin Neurosci*. 2023; 77(5): 273-281.
- Ntracha A, Iakovakis D, Hadjimitriou S, Charisis VS, Tsolaki M, Hadjileontiadis LJ. Detection of mild cognitive impairment through natural language and touchscreen typing processing. *Front Digit Health*. 2020; 2: 567158.
- Adhikari S, Thapa S, Naseem U, et al. Exploiting linguistic information from Nepali transcripts for early detection of Alzheimer's disease using natural language processing and machine learning techniques. *Int J Hum Comput Stud*. 2022; 160: 102761.
- Lindsay H, Tröger J, König A. Language impairment in Alzheimer's disease-robust and explainable evidence for AD-related deterioration of spontaneous speech through multilingual machine learning. *Front Aging Neurosci*. 2021; 13: 642033.
- Liu N, Yuan Z. Spontaneous language analysis in Alzheimer's disease: evaluation of natural language processing technique for analyzing lexical performance. *J Shanghai Jiao Tong Univ (Sci)*. 2021; 27: 160-167.
- Yorganci E, Stewart R, Sampson EL, Sleeman KE. Patterns of unplanned hospital admissions among people with dementia: from diagnosis to the end of life. *Age Ageing*. 2022; 51(5): afac098.
- Soysal P, Tan SG, Rogowska M, et al. Weight loss in Alzheimer's disease, vascular dementia and dementia with Lewy bodies: impact on mortality and hospitalization by dementia subtype. *Int J Geriatr Psychiatry*. 2021;37(2):1-8. doi:[10.1002/gps.5659](https://doi.org/10.1002/gps.5659)
- Naismith H, Howard R, Stewart R, Pitman A, Mueller C. Suicidal ideation in dementia: associations with neuropsychiatric symptoms and subtype diagnosis. *Int Psychogeriatr*. 2022; 34(4): 399-406.
- Al-Harrasi AM, Iqbal E, Tsamakis K, et al. Motor signs in Alzheimer's disease and vascular dementia: detection through natural language processing, co-morbid features and relationship to adverse outcomes. *Exp Gerontol*. 2021; 146: 111223.
- Chen S, Underwood BR, Jones PB, Lewis JR, Cardinal RN. Association between lithium use and the incidence of dementia and its subtypes: a retrospective cohort study. *PLoS Med*. 2022; 19(3): e1003941.
- Phiri P, Engelthaler T, Carr H, Delanerolle G, Holmes C, Rathod S. Associated mortality risk of atypical antipsychotic medication in individuals with dementia. *World J Psychiatry*. 2022; 12(2): 298.

37. Eikelboom WS, Singleton EH, van den Berg E, et al. The reporting of neuropsychiatric symptoms in electronic health records of individuals with Alzheimer's disease: a natural language processing study. *Alzheimers Res Ther.* 2023;15(1):1-12.
38. Bougiatiotis K, Aisopos F, Nentidis A, Krithara A, Paliouras G. Drug-drug interaction prediction on a biomedical literature knowledge graph. In: *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18.* Springer; 2020:122-132.
39. Vitali F, Branigan GL, Brinton RD. Preventing Alzheimer's disease within reach by 2025: targeted-risk-AD-prevention (trap) strategy. *Alzheimers Dement.* 2021; 7(1): e12190.
40. Kirkpatrick A, Onyeze C, Kartchner D, et al. Optimizations for computing relatedness in biomedical heterogeneous information networks: SemNet 2.0. *Big Data and Cogn Comput.* 2022; 6(1): 27.
41. Hu Y, Keloth VK, Raja K, Chen Y, Xu H. Towards precise pico extraction from abstracts of randomized controlled trials using a section-specific learning approach. *Bioinformatics.* 2023; 39(9): btad542.
42. Rossanez A, Dos Reis JC, Torres RdS, de Ribaupierre H. Kgen: a knowledge graph generator from biomedical scientific literature. *BMC Med Inf Decis Making.* 2020; 20(4):1-24.
43. Liu J, Wu H, Robertson DH, Zhang J. Text mining and portal development for gene-specific publications on Alzheimer's disease and other neurodegenerative diseases. *BMC Med Inf Decis Making.* 2024; 24(suppl 3): 98.
44. Klein AZ, Magge A, O'Connor K, Gonzalez-Hernandez G. Automatically identifying twitter users for interventions to support dementia family caregivers: annotated data set and benchmark classification models. *JMIR Aging.* 2022; 5(3): e39547.
45. Sunmoo Y, Broadwell P, Frederick FS, Jang SJ, Haeyoung L. Comparing emotional valence scores of twitter posts from manual coding and machine learning algorithms to gain insights to refine interventions for family caregivers of persons with dementia. *Stud Health Technol Inform.* 2022; 295: 253.
46. Sunmoo Y, Broadwell P, Alcantara C, et al. Analyzing topics and sentiments from twitter to gain insights to refine interventions for family caregivers of persons with Alzheimer's disease and related dementias (ADRD) during COVID-19 pandemic. *Stud Health Technol Inform.* 2022; 289: 170.
47. Alroobaea R, Haoues M, Rubaiee S, Ahmed A, Almansour F. Machine-learning based analysis of mobile apps for people with Alzheimer's disease. *SSRG Int J Eng Trends Technol.* 2021; 69(2): 126-133.
48. Tahami Monfared AA, Stern Y, Doogan S, Irizarry M, Zhang Q. Understanding barriers along the patient journey in Alzheimer's disease using social media data. *Neurol Ther.* 2023; 12(3): 899-918.
49. Saunders S, Muniz-Terrera G, Sheehan S, Ritchie C, Luz S. A UK-wide study employing natural language processing to determine what matters to people about brain health to improve drug development: the electronic person-specific outcome measure (ePSOM) programme. *J Prev Alzheimers Dis.* 2021; 8: 448-456.
50. Fang Y, Ilday B, Sun Y, et al. Combining human and machine intelligence for clinical trial eligibility querying. *J Am Med Inform Assoc.* 2022; 29(7): 1161-1171.
51. Tavabi N, Stüch D, Signorini A, et al. Cognitive digital biomarkers from automated transcription of spoken language. *J Prev Alzheimers Dis.* 2022; 9(4): 791-800.
52. Searle T, Ibrahim Z, Dobson R. Comparing natural language processing techniques for Alzheimer's dementia prediction in spontaneous speech. *Proc Interspeech.* 2020;2192-2196. doi:10.21437/Interspeech.2020-2729
53. Martinc M, Pollak S. Tackling the ADReSS challenge: a multimodal approach to the automated recognition of Alzheimer's dementia. In: *Interspeech.* ISCA; 2020:2157-2161.
54. Balagopalan A, Eyre B, Rudzicz F, Novikova J. To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection. *arXiv preprint arXiv:2008.01551* 2020
55. Taghibeyglou B, Rudzicz F. Who needs context? Classical techniques for Alzheimer's disease detection. In: *Proceedings of the 5th Clinical Natural Language Processing Workshop.* 2023:102-107.
56. Pappagari R, Cho J, Moro-Velazquez L, Dehak N. Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity. *Proc Interspeech.* 2020:2177-2181. doi: 10.21437/Interspeech.2020-2587
57. Aryal SK, Prioleau H, Burge L. Acoustic-linguistic features for modeling neurological task score in Alzheimer's. In: *Pacific Symposium on Biocomputing 2023: Kohala Coast, Hawaii, USA, 3–7 January 2023.* World Scientific; 2022:335-346.
58. Pappagari R, Cho J, Joshi S, et al. Automatic detection and assessment of Alzheimer's disease using speech and language technologies in low-resource scenarios. *Interspeech.* 2021;2021:3825-3829.
59. Woszczyk D, Hedlikova A, Akman A, Demetriou S, Schuller B. Data augmentation for dementia detection in spoken language. 2022.
60. Wen B, Wang N, Subbalakshmi K, et al. Revealing the roles of part-of-speech taggers in Alzheimer's disease detection: scientific discovery using one-intervention causal explanation. *JMIR Form Res.* 2023; 7(1): e36590.
61. Adhikari S, Thapa S, Singh P, Huo H, Bharathy G, Prasad M. A comparative study of machine learning and NLP techniques for uses of stop words by patients in diagnosis of Alzheimer's disease. In: *2021 International Joint Conference on Neural Networks (IJCNN).* IEEE; 2021:1-8.
62. Pérez-Toro PA, Bayerl SP, Arias-Vergara T, et al. Influence of the interviewer on the automatic assessment of Alzheimer's disease in the context of the ADReSSo challenge. In: *Interspeech.* ISCA; 2021:3785-3789.
63. Soroski T, da Cunha Vasco T, Newton-Mason S, et al. Evaluating web-based automatic transcription for Alzheimer speech data: transcript comparison and machine learning analysis. *JMIR Aging.* 2022; 5(3): e33460.
64. Pompili A, Abad A, de Matos DM, Martins IP. Pragmatic aspects of discourse production for the automatic identification of Alzheimer's disease. *IEEE J Sel Top Signal Process.* 2020; 14(2): 261-271.
65. Hajjar I, Okafor M, Choi JD, et al. Development of digital voice biomarkers and associations with cognition, cerebrospinal biomarkers, and neural representation in early Alzheimer's disease. *Alzheimers Dement.* 2023; 15(1): e12393.
66. Sadeghian R, Schaffer JD, Zahorian SA. Towards an automatic speech-based diagnostic test for Alzheimer's disease. *Front Comput Sci.* 2021; 3: 624594.
67. Roshanzamir A, Aghajan H, Soleymani Baghshah M. Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech. *BMC Med Inf Decis Making.* 2021; 21:1-14.
68. Clarke N, Barrick TR, Garrard P. A comparison of connected speech tasks for detecting early Alzheimer's disease and mild cognitive impairment using natural language processing and machine learning. *Front Comput Sci.* 2021; 3: 634360.
69. Yeung A, Iaboni A, Rochon E, et al. Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and Alzheimer's dementia. *Alzheimers Res Ther.* 2021; 13(1): 109.
70. Millington T, Luz S. Analysis and classification of word co-occurrence networks from Alzheimer's patients and controls. *Front Comput Sci.* 2021; 3: 649508.
71. Zolnoori M, Zolnour A, Topaz M. Adscreen: a speech processing-based screening system for automatic identification of patients with Alzheimer's disease and related dementia. *Artif Intell Med.* 2023; 143: 102624.

72. Mahajan P, Baths V. Acoustic and language based deep learning approaches for Alzheimer's dementia detection from spontaneous speech. *Front Aging Neurosci.* 2021; 13: 623607.
73. Balagopalan A, Eyre B, Robin J, Rudzicz F, Novikova J. Comparing pre-trained and feature-based models for prediction of Alzheimer's disease based on speech. *Front Aging Neurosci.* 2021; 13: 635945.
74. Robin J, Xu M, Balagopalan A, et al. Automated detection of progressive speech changes in early Alzheimer's disease. *Alzheimers Dement.* 2023; 15(2): e12445.
75. Liu N, Luo K, Yuan Z, Chen Y. A transfer learning method for detecting Alzheimer's disease based on speech and natural language processing. *Front Public Health.* 2022; 10: 772592.
76. Amini S, Hao B, Zhang L, et al. Automated detection of mild cognitive impairment and dementia from voice recordings: a natural language processing approach. *Alzheimers Dement.* 2023; 19(3): 946-955.
77. Ilias L, Askounis D. Multimodal deep learning models for detecting dementia from speech and transcripts. *Front Aging Neurosci.* 2022; 14: 830943.
78. Sharma S, Dudeja RK, Aujla GS, Bali RS, Kumar N. DeTrAs: deep learning-based healthcare framework for IoT-based assistance of Alzheimer's patients. *Neural Comput & Applic.* 2020;1-13. doi:10.1007/s00521-020-05327-2
79. Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. *PLOS Digit Health.* 2022; 1(12): e0000168.
80. Chandler C, Diaz-Asper C, Turner RS, Reynolds B, Elvevåg B. An explainable machine learning model of cognitive decline derived from speech. *Alzheimers Dement.* 2023; 15(4): e12516.
81. Runde BS, Alapati A, Bazan NG. The optimization of a natural language processing approach for the automatic detection of Alzheimer's disease using GPT embeddings. *Brain Sci.* 2024; 14(3): 211.
82. Chen Z, Zhang H, Yang X, et al., Assess the documentation of cognitive tests and biomarkers in electronic health records via natural language processing for Alzheimer's disease and related dementias. *Int J Med Inf.* 2023;170:104973.
83. Zolnoori M, Barrón Y, Song J, et al. Homeadscreen: developing Alzheimer's disease and related dementia risk identification model in home healthcare. *Int J Med Inf.* 2023; 177: 105146.
84. Maserejian N, Krzywy H, Eaton S, Galvin JE. Cognitive measures lacking in EHR prior to dementia or Alzheimer's disease diagnosis. *Alzheimers Dement.* 2021; 17(7): 1231-1243.
85. Wu W, Holkeboer KJ, Kolawole TO, Carbone L, Mahmoudi E. Natural language processing to identify social determinants of health in Alzheimer's disease and related dementia from electronic health records. *Health Serv Res.* 2023; 58(6): 1292-1302.
86. Jennings LA, Hollands S, Keeler E, Wenger NS, Reuben DB. The effects of dementia care co-management on acute care, hospice, and long-term care utilization. *J Am Geriatr Soc.* 2020; 68(11): 2500-2507.
87. Kent DM, Leung LY, Zhou Y, et al. Association of incidentally discovered covert cerebrovascular disease identified using natural language processing and future dementia. *J Am Heart Assoc.* 2023; 12(1): e027672.
88. Hane CA, Nori VS, Crown WH, Sanghavi DM, Bleicher P. Predicting onset of dementia using clinical notes and machine learning: case-control study. *JMIR Med Inform.* 2020; 8(6): e17819.
89. Shen Z, Schutte D, Yi Y, et al. Classifying the lifestyle status for Alzheimer's disease from clinical notes using deep learning with weak supervision. *BMC Med Inf Decis Making.* 2022; 22(1):1-11.
90. Penfold RB, Carrell DS, Cronkite DJ, et al. Development of a machine learning model to predict mild cognitive impairment using natural language processing in the absence of screening. *BMC Med Inf Decis Making.* 2022; 22(1):1-13.
91. Topaz M, Adams V, Wilson P, Woo K, Ryvicker M. Free-text documentation of dementia symptoms in home healthcare: a natural language processing study. *Gerontol Geriatr Med.* 2020; 6: 2333721420959861.
92. Patel J, Wu H. Utilizing electronic dental records to predict neurodegenerative diseases in a dental setting: a pilot study. *Stud Health Technol Inform.* 2024; 310: 1322-1326.
93. Laurentiev J, Kim DH, Mahesri M, et al. Identifying functional status impairment in people living with dementia through natural language processing of clinical documents: cross-sectional study. *J Med Internet Res.* 2024; 26: e47739.
94. Mahmoudi E, Wu W, Najarian C, et al. Identifying caregiver availability using medical notes with rule-based natural language processing: retrospective cohort study. *JMIR Aging.* 2022; 5(3): e40241.
95. Shao Y, Cheng Y, Nelson SJ, et al. Hybrid value-aware transformer architecture for joint learning from longitudinal and non-longitudinal clinical data. *J Pers Med.* 2023;13(7):1070.
96. Crystal D, Quirk R. *Systems of Prosodic and Paralinguistic Features in English.* Vol 39. Walter de Gruyter GmbH & Co KG; 2021.
97. Liu Y, Zhang M. *Neural network methods for natural language processing.* 2018.
98. Martin JH. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Pearson/Prentice Hall; 2009.
99. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics; 2019: 4171-4186.
100. Dawson DV, Blanchette DR, Pihlstrom BL. Application of biostatistics in dental public health. *Burt and Eklund's Dentistry, Dental Practice, and the Community.* Elsevier; 2021:131-153.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Shakeri A, Farmanbar M. Natural language processing in Alzheimer's disease research: Systematic review of methods, data, and efficacy. *Alzheimer's Dement.* 2024;17:e70082.
<https://doi.org/10.1002/dad2.70082>