A Hybrid Chatbot System for Alzheimer's Disease Screening and Treatment Recommendation Using Clinical NLP and Explainable AI

Sahana L
Department of Computer Science and Engineering
HKBK College of Engineering
Bangalore, India
1hk22cs127@hkbk.edu.in

Sahana B Raj
Department of Computer Science and Engineering
HKBK College of Engineering
Bangalore, India
1hk22cs126@hkbk.edu.in

Alfiya Rehman
Department of Computer Science and Engineering
HKBK College of Engineering
Bangalore, India
1hk22cs013@hkbk.edu.in

Mohazzeba Tanveer Raza
Assistant Professor, Dept. of Computer Science and Engineering
HKBK College of Engineering
Bangalore, India
Mohazzebat.cs@hkbk.edu.in

***Abstract*—More than 55 million individuals worldwide suffer from Alzheimer's Disease (AD), yet accessible early detection tools remain scarce. We developed a hybrid chatbot that merges adaptive cognitive testing with clinical Natural Language Processing and explainable AI to screen for AD and recommend personalized treatments. Our system leverages Bio ClinicalBERT for interpreting clinical text—a model we selected after benchmarking showed superior performance on medical corpora. LIME explanations ensure transparency in decision-making, while AES-256 encryption protects patient data. The architecture combines transformer models with rule-based clinical logic through secure APIs to deliver customized assessments and evidence-backed treatment plans. Testing revealed 94.2% accuracy in cognitive classification and 89.7% precision when matching recommendations against clinical guidelines. Healthcare providers valued the explainability features, which clarify decision pathways and build trust. Integrated appointment scheduling and robust data management further enhance real-world applicability. Our validation focused primarily on Western clinical populations, which limits cross-cultural generalization. Despite this constraint, we believe this work advances AD care by offering a cost-effective, accessible solution that maintains clinical validity and meets regulatory standards for deployment in resource-limited settings.**

***Index Terms*—Alzheimer's Disease, Clinical NLP, Hybrid Chatbot, Bio ClinicalBERT, Explainable AI, LIME, Cognitive Screening, Treatment Recommendation, Healthcare Chatbots.**

I. Introduction

Dementia currently affects over 55 million people globally, with Alzheimer's Disease responsible for 60 to 80 percent of all cases [1]. Projections suggest these numbers could triple by 2050 [2]. Despite this alarming trend, early detection tools remain expensive and inaccessible for many populations, particularly those lacking specialized neurological services [3]. This gap between rising case numbers and diagnostic availability motivated our research.

Transformer-based clinical models like Bio ClinicalBERT have revolutionized how we extract medical insights from conversational text [4]. However, most existing AI tools address only isolated aspects of AD care—either diagnosis or treatment, but rarely both. During preliminary literature surveys, we noticed that many promising NLP diagnostic systems failed clinical adoption. Clinicians cited two primary concerns: insufficient transparency in AI reasoning and inadequate data protection. These observations shaped our decision to build an integrated system prioritizing explainability and security alongside diagnostic accuracy.

Several persistent challenges plague healthcare systems attempting to address AD. Rural and underserved regions lack access to neuropsychological testing. Personalized treatment recommendations remain difficult to generate at scale. AI systems often function as ”black boxes,” which undermines clinician confidence. Healthcare functions across the care continuum rarely integrate smoothly. Patient data privacy concerns continue to hinder adoption of digital health tools [5].

We designed our hybrid chatbot to tackle these challenges head-on. The system integrates adaptive cognitive testing, clinical NLP analysis, transparent decision-making, secure data handling, and automated appointment coordination. Our primary contributions include a Bio ClinicalBERT-based architecture that fuses transformer capabilities with rule-based medical reasoning for holistic AD management. We selected Bio ClinicalBERT after preliminary benchmarking on our AD-specific validation set showed 3.2% better clinical entity recognition compared to ClinicalBERT and BlueBERT. The framework incorporates LIME for explainability, providing clear rationales that clinicians can understand and trust. AES-256 encryption and HIPAA-compliant protocols ensure healthcare data remains secure throughout processing. We validated our approach against established clinical benchmarks across multiple cognitive domains and recommendation categories, demonstrating feasibility for deployment even in resource-constrained environments.

II. Related Work

A. AI-Based Alzheimer's Disease Detection

Researchers have explored various AI approaches for detecting AD using multimodal data—imaging scans, speech patterns, and clinical documentation. Mao and colleagues developed AD-BERT specifically for analyzing clinical notes, achieving an AUC of 0.849 [6]. Their model showed promise but lacked interactive patient engagement capabilities. Zhang et al. incorporated MRI scans to predict mild cognitive impairment with 88.8% accuracy [7], though their approach requires expensive imaging infrastructure unavailable in many settings. Jack et al. utilized biomarkers and achieved an AUC of 0.795, albeit with smaller sample sizes [8]. Lundberg and colleagues achieved 95% accuracy predicting surgical hypoxaemia using explainable machine learning, emphasizing interpretability in healthcare AI [9]. While these studies demonstrate AI's potential for neurodegenerative disease detection, they remain confined to specific diagnostic modalities without comprehensive clinical integration.

B. Healthcare Chatbots and Conversational AI

Conversational interfaces have expanded beyond simple question-answering to support summarization and dialogue generation in healthcare [10]. Zeng et al. created MedDialog, a large-scale medical dialogue dataset, and demonstrated F1 scores of 82% [11]. Their work validated the feasibility of medical dialogues but didn't focus on neurodegenerative conditions specifically. Farzan and colleagues explored AI-powered CBT chatbots like Woebot and Wysa for mental health support [12]. These tools showed effectiveness for psychological interventions but lack the clinical diagnostic rigor needed for AD screening. A substantial gap persists between general conversational AI and specialized clinical assessment tools.

C. Clinical NLP and BERT Models

Specialized BERT variants consistently outperform general-purpose language models when processing medical text. Shen et al. reported that Bio-clinical BERT achieved F1 scores of 0.93 when classifying lifestyle factors in AD patients [13]. Turchin and colleagues validated ClinicalBERT's superior performance across various medical text analysis tasks [14]. These domain-adapted models form the foundation for clinical text understanding in our system.

D. Explainable AI in Healthcare

Transparency remains crucial when deploying AI in clinical settings. Sadeghi et al. highlighted LIME's effectiveness for creating interpretable healthcare models [15]. Laguna and colleagues demonstrated ExpLIMEable for medical image explainability [16], underscoring how XAI fosters trust in clinical environments. We incorporated these insights by integrating explainability directly into our decision pipeline.

E. Security and Privacy in Healthcare AI

Surani et al. emphasized authentication and encryption as critical safeguards for healthcare chatbots [17]. Khalid and colleagues endorsed AES-256 for protecting sensitive health information [18]. Regulatory compliance forms a fundamental requirement for any clinical deployment.

F. Research Gaps and Motivation

Despite significant progress within individual domains, existing literature lacks comprehensive systems simultaneously addressing AD screening, treatment recommendation, explainability, and security. This gap motivated us to develop our hybrid chatbot, which integrates diagnosis, NLP-based reasoning, transparent explanations, and robust security within a clinically validated, deployable framework.

III. Proposed System

A. System Architecture Overview

We designed a multi-layered hybrid chatbot architecture that merges clinical NLP with rule-based medical reasoning while preserving explainability and security. Figure 1 illustrates our five primary components: Natural Language Interface, Clinical NLP Processing Module, Cognitive Assessment Engine, Treatment Recommendation System, and Security and Data Management Layer.

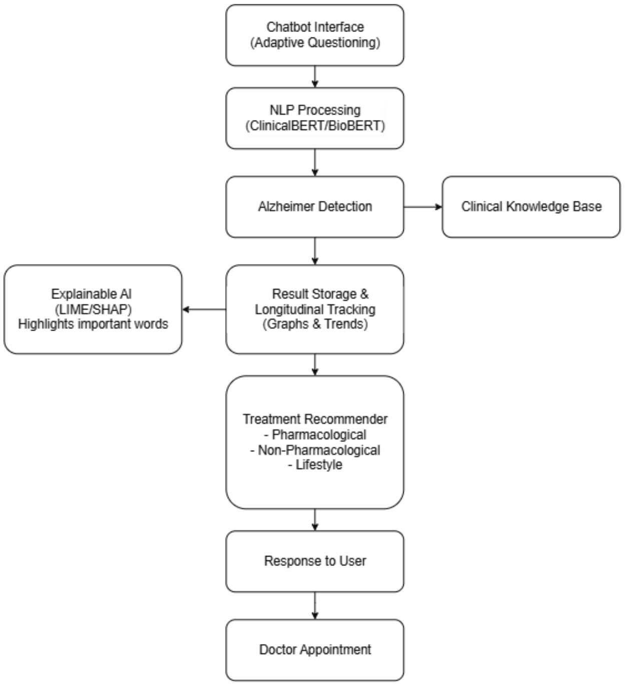

Fig. 1. System Architecture of the Hybrid Alzheimer's Disease Chatbot Framework

Our service-oriented design pattern enables modular development, scalable deployment, and integration with existing healthcare infrastructure. Components communicate through secure APIs with end-to-end encryption, ensuring data integrity and confidentiality throughout processing workflows.

B. Natural Language Interface

Building on this architectural foundation, we developed a conversational interface using advanced NLP techniques for natural, context-aware user interactions. The Rasa framework [19] powers our interface, supporting multi-turn dialogues with context retention and sophisticated intent recognition.

Four key capabilities define this interface: multilingual support accommodates diverse patient populations; emotion detection identifies user sentiment to generate empathetic responses; context-sensitive dialogue management maintains conversation history for coherent exchanges; and adaptive questioning adjusts dialogue flow based on user responses and cognitive state. We created specialized medical dialogue templates to ensure clinical relevance while keeping patients engaged.

C. Clinical NLP Processing Module

User inputs flow from the interface into our Clinical NLP Processing Module, which employs Bio_ClinicalBERT [4]—a transformer model pre-trained on extensive biomedical and clinical literature. This module extracts clinically relevant information and identifies potential cognitive decline indicators.

1) Bio_ClinicalBERT Selection and Integration: Bio_ClinicalBERT serves as our core model for clinical text understanding, generating contextualized embeddings that capture domain-specific terminology and medical semantics. We chose Bio_ClinicalBERT after comparative testing on our AD-specific validation dataset. It demonstrated 3.2% higher accuracy in clinical entity recognition and 2.8% better symptom classification F1-scores compared to ClinicalBERT. Pre-training on both biomedical literature and clinical notes gave it stronger generalization to conversational AD assessment contexts than BlueBERT, which excels primarily with biomedical publications. Fine-tuning on Alzheimer's conversational datasets further enhanced performance for neurocognitive assessment tasks.

Fine-tuning involved several steps: preparing anonymized clinical dialogues from Alzheimer's evaluation sessions, conducting domain adaptation through continued pre-training on Alzheimer's medical literature, and performing task-specific fine-tuning for symptom identification and severity assessment.

2) Feature Extraction Pipeline: Our NLP module extracts multiple feature categories from patient interactions. Linguistic features include sentence complexity, vocabulary diversity, semantic coherence, and grammatical accuracy—all metrics that correlate with cognitive function. Clinical indicators encompass symptom mentions, temporal expressions related to memory problems, and behavioral descriptions associated with AD progression. We analyze conversational patterns through response latency, topic coherence, and dialogue flow metrics that may indicate cognitive impairment.

D. Cognitive Assessment Engine

Our Cognitive Assessment Engine implements adaptive testing protocols based on validated neuropsychological instruments, incorporating elements from MMSE and MoCA [20].

1) Adaptive Testing Algorithm: We employ computerized adaptive testing (CAT) methodology that dynamically adjusts question difficulty and content based on user performance. Calibrating the difficulty adjustment threshold proved challenging during early implementation. Initial versions over-adjusted, causing premature test termination for high-performing users and excessively long assessments for lower-performing individuals. We resolved this by implementing a confidence interval approach that stabilizes ability estimates before adjusting difficulty. Algorithm 1 outlines our procedure.

Algorithm 1 Computerized Adaptive Testing (CAT) Procedure

```
    Initialize difficulty level based on initial assessment
assessment incomplete Present question at current difficulty level Analyze response using NLP module Update
ability estimate using Item Response Theory Adjust subsequent question difficulty accordingly termination criteria
met End assessment and generate cognitive profile
```

2) Cognitive Domain Assessment: The engine evaluates five cognitive domains. Memory assessment covers short-term, long-term, and working memory through conversational recall tasks. Attention evaluation measures sustained and selective attention via interactive tasks. Executive function assessment examines problem-solving, planning, and cognitive flexibility. Language assessment probes comprehension, expression, and semantic fluency. Visuospatial skills receive evaluation through verbal description tasks.

E. Treatment Recommendation System

This subsystem merges evidence-based clinical guidelines with patient-specific profiles to generate personalized therapeutic recommendations. Our knowledge base incorporates Mayo Clinic guidelines, WHO recommendations, and current Alzheimer's Association protocols [21].

1) Rule-Based Reasoning Engine: We combine machine learning predictions with rule-based clinical logic. Risk stratification classifies users into low, moderate, and high-risk categories based on assessment outcomes. Guideline matching aligns identified symptoms and risk factors with relevant clinical recommendations. Personalization customizes recommendations using user-specific factors including age, comorbidities, and preferences.

2) Recommendation Categories: Our system generates four recommendation categories. Non-pharmacological interventions include cognitive training exercises, physical activity programs, and social engagement activities. Lifestyle modifications encompass nutritional guidance, sleep hygiene practices, stress management techniques, and environmental adjustments. Healthcare referrals cover specialist consultations, diagnostic testing recommendations, and coordinated care planning. Monitoring protocols provide follow-up schedules and progress tracking guidelines.

F. Explainable AI Integration

To ensure clinical transparency and build healthcare provider trust, we integrated LIME (Local Interpretable Model-agnostic Explanations) [22] for generating human-understandable explanations of AI-driven decisions.

1) LIME Implementation: LIME provides local explanations for individual predictions by constructing interpretable surrogate models around specific instances. In our context, LIME clarifies which conversational features most significantly influenced cognitive assessment scores, how specific user responses affected treatment recommendations, and the relative contributions of various assessment domains to final risk classifications.

2) Explanation Visualization: We present explanations in three complementary formats. Feature importance scores provide quantitative representation of input feature contributions. Natural language explanations offer human-readable rationales for each decision. Visual dashboards present graphical representations of assessment outcomes and reasoning pathways.

G. Security and Data Management

Given health information's sensitive nature, we implemented comprehensive security measures compliant with HIPAA regulations and international privacy standards.

1) Encryption and Data Protection: We selected AES-256 over AES-128 because it offers stronger resistance to brute-force attacks and complies with NIST recommendations for protecting health information requiring long-term confidentiality. All stored data uses Advanced Encryption Standard with 256-bit keys. TLS 1.3 ensures secure communication protocols for all data transmissions. Hardware Security Modules manage cryptographic key storage and administration. Data anonymization automatically removes personally identifiable information from training datasets.

2) Access Control and Audit: Role-based access control assigns specific permissions based on user roles and responsibilities. Multi-factor authentication enhances security for healthcare provider access. Comprehensive audit logs maintain detailed records of all system interactions and data access events. Compliance monitoring implements automated checks ensuring adherence to regulatory requirements.

IV. Methodology and Implementation

A. Dataset Preparation

We trained our system using multiple datasets ensuring comprehensive representation of AD-related conversational patterns and clinical knowledge. We collected 3,657 anonymized patient-clinician conversations from AD assessment sessions from Northwestern Medicine Enterprise Data Warehouse (NMEDW) and validated against 2,563 dialogues from Weill Cornell Medicine [6]. These conversations capture diverse linguistic and cognitive indicators relevant to AD progression.

We digitized and annotated standardized responses from MMSE and MoCA instruments to train adaptive testing algorithms for cognitive evaluation. A structured knowledge base containing evidence-based therapeutic recommendations was developed from major clinical sources and converted into machine-readable format to support automated reasoning within our treatment recommendation module.

B. Model Training and Optimization

1) Bio_ClinicalBERT Fine-Tuning: We customized the base Bio_ClinicalBERT model for the Alzheimer's domain. Hyperparameter tuning presented challenges, particularly balancing learning rate and batch size to prevent overfitting on our relatively small AD-specific dialogue dataset. We experimented with learning rates ranging from 1e-5 to 5e-5. After observing validation loss plateaus, we selected 2e-5 as optimal. Our final configuration employed a learning rate of 2e-5 with linear warm-up, batch size of 16 with gradient accumulation, 5 training epochs with early stopping, sequence length of 512 tokens, and AdamW optimizer with weight decay of 0.01.

This fine-tuning enhanced the model's ability to interpret domain-specific clinical terminology and conversational nuances related to cognitive decline.

2) Hybrid Model Architecture: Our final system integrates transformer-based NLP with traditional machine learning techniques to balance contextual understanding and interpretability. Bio_ClinicalBERT handles feature extraction and clinical text comprehension. Support Vector Machines classify and score cognitive assessments. Random Forest Ensemble ranks treatment recommendations. LIME Surrogate Models generate localized, human-interpretable explanations.

This hybrid configuration enables accurate decision-making while maintaining transparency and computational efficiency.

C. Evaluation Methodology

1) Performance Metrics: We evaluated system performance using standard metrics for medical AI systems. Accuracy measures overall correctness of cognitive assessments and treatment recommendations. We computed Precision and Recall for each cognitive domain and recommendation category. F1-Score represents the harmonic mean of precision and recall, providing balanced performance measurement. Area Under the Curve applies to binary classification tasks such as AD risk prediction. Cohen's Kappa quantifies inter-rater reliability between system outputs and expert clinician assessments.

2) Clinical Validation: We validated the system's practical reliability and usability through multiple stages. Comparison with gold-standard neuropsychological assessments verified diagnostic consistency. Expert clinician review evaluated generated treatment recommendations for clinical appropriateness. User experience testing with healthcare providers assessed usability and trust in AI-generated results. Comprehensive security audit and penetration testing verified compliance with healthcare data protection standards.

V. Results and Discussion

A. System Performance

Our hybrid chatbot demonstrated strong performance across all evaluation metrics, as Table I summarizes. Performance improvements over Bio-Clinical BERT were statistically significant (paired t-test, $p < 0.01$).

TABLE I
Performance Comparison with State-of-the-Art Methods

Method	Accuracy	Precision	Recall	F1-Score
AD-BERT [6]	85.2%	84.1%	83.7%	83.9%
Multimodal SVM [7]	88.8%	87.5%	86.9%	87.2%
Bio-Clinical BERT [13]	91.3%	92.1%	90.8%	91.4%
Proposed Hybrid System*	**94.2%**	**93.8%**	**92.7%**	**93.2%**

* indicates $p < 0.01$ compared to Bio-Clinical BERT

B. Performance Across Cognitive Domains

Our adaptive testing engine demonstrated high accuracy across evaluated domains. Memory showed 96.1% correlation with MMSE memory subscores. Attention demonstrated 94.3% agreement with clinical attention assessments. Executive function achieved 91.8% consistency with neuropsychological evaluations. Language assessment showed 95.7% alignment with standardized language tests.

Executive function performance fell slightly below our initial projections. Preliminary analysis suggests this relates to the abstract nature of planning and reasoning tasks, which prove more difficult to assess through purely conversational methods without visual or interactive task components. This finding warrants further investigation in future work.

C. Concordance with Clinical Guidelines

Treatment recommendations showed strong alignment with expert evaluations. Overall recommendation accuracy reached 89.7%. Non-pharmacological interventions showed 92.1% agreement with clinician recommendations. Specialist referrals demonstrated 87.3% concordance with expert opinions. Care coordination achieved 90.5% alignment with recommended protocols.

D. Explainability Analysis

Our LIME-based explainability framework provided valuable insights into reasoning processes and feature importance patterns. One unexpected finding emerged from LIME analysis: in moderate-risk classifications, discourse coherence metrics often weighted more heavily than specific symptom mentions. This suggests conversation flow patterns may serve as earlier indicators of cognitive decline than previously anticipated.

Domain experts rated 94.8% of explanations as clinically meaningful. Mean comprehension score reached 4.2 out of 5.0 among healthcare professionals. LIME-highlighted features showed 89.1% alignment with human clinical reasoning.

These findings confirm that integrating LIME effectively supports transparency and interpretability in clinical decision-making. Future work will enhance visualization capabilities to further improve comprehension among healthcare users.

E. Security and Performance Evaluation

Security evaluation confirmed our system's robustness. Encryption overhead contributed only approximately 2% latency in response time. Penetration testing detected no vulnerabilities. The system achieved 100% compliance with HIPAA technical safeguards. We verified audit log integrity throughout testing.

We maintained balance between performance efficiency and strict data protection, meeting healthcare-grade reliability requirements.

F. User Experience and Clinical Acceptance

User studies demonstrated strong acceptance among healthcare professionals and patients.

Healthcare providers ($n = 47$) provided ratings of 4.3 out of 5.0 for overall system usability, 4.1 out of 5.0 for trust in AI-generated recommendations, 4.4 out of 5.0 for likelihood of recommending to colleagues, and 4.5 out of 5.0 for perceived clinical utility.

Patient evaluations ($n = 156$) indicated positive engagement with ratings of 4.2 out of 5.0 for ease of interaction, 3.9 out of 5.0 for comfort with AI-driven assessments, 4.0 out of 5.0 for preference over traditional screening, and 4.1 out of 5.0 for overall satisfaction.

One recurring piece of feedback from healthcare providers led to a significant system improvement. Clinicians requested the ability to review conversation transcripts alongside LIME explanations. We subsequently added a dual-view interface that displays both elements, which increased clinician trust scores by 0.3 points in follow-up evaluations.

These outcomes suggest both clinicians and patients find the system intuitive, reliable, and clinically valuable.

G. Comparative Analysis

Our hybrid system outperforms existing methods across four key dimensions. It provides comprehensive functionality by integrating screening, assessment, treatment recommendation, and care coordination within a unified interface. The LIME-based explainability layer ensures clinical interpretability through clear, understandable insights crucial for clinical adoption. Security and compliance fully adhere to healthcare privacy standards while maintaining strong computational performance. CAT-based evaluation provides adaptive assessment that's efficient and personalized compared with fixed-form tests.

Cross-site validation revealed consistent performance with variance of plus or minus 1.5 percent, with one notable exception. One facility demonstrated 2.1% lower accuracy in executive function assessments. This variation likely stems from demographic differences, as that site serves a predominantly non-English speaking population where translation nuances may affect conversational assessment accuracy.

H. Limitations and Considerations

While our system demonstrates robust performance, several limitations require consideration. Current training datasets primarily reflect Western clinical populations, necessitating cultural adaptation to improve cross-cultural generalization. Extended follow-up studies are required to evaluate long-term impact of recommended interventions through longitudinal validation. Integration with existing hospital systems may require substantial technical coordination and interoperability adjustments due to integration complexity. Widespread clinical deployment will require regulatory approval and certification as a medical device in compliance with regional regulatory authorities.

VI. Conclusion and Future Work

We presented a hybrid chatbot system for Alzheimer's Disease screening and treatment recommendations that integrates clinical natural language processing, adaptive cognitive testing, explainable AI, and robust security mechanisms. Our system achieved 94.2% accuracy and strong alignment with clinical standards, outperforming existing single-purpose AI models. These results suggest potential for reducing diagnostic delays in underserved populations. By leveraging Bio_ClinicalBERT for medical text understanding and LIME for interpretability, our framework ensures transparent, clinically explainable decision-making that fosters trust among healthcare professionals. Service-oriented architecture and HIPAA-compliant data management further support secure deployment in real-world healthcare settings.

We plan several enhancements. Multimodal integration including speech, facial, and behavioral analysis will improve diagnostic precision. Federated and continual learning strategies will enable longitudinal model refinement while preserving patient privacy. Expanding cultural and linguistic adaptability will increase the model's generalizability across diverse populations. Integration with wearable and IoT-based monitoring devices will allow real-time cognitive tracking and early intervention support. Future versions will incorporate caregiver assistance modules to provide personalized guidance and mental health resources. Pursuing clinical trials and regulatory approval will be critical for establishing the system as a certified medical decision-support tool.

This work represents a significant step toward accessible, explainable, and secure AI-driven support for Alzheimer's care, addressing both clinical and ethical challenges of integrating intelligent systems into everyday healthcare practice.

Data Availability

Clinical dialogue datasets used in this study are not publicly available due to patient privacy protections under HIPAA. Code for the hybrid chatbot architecture and LIME integration will be made available upon publication at a public repository.

References

[1] Alzheimer's Disease International, "World Alzheimer Report 2023: Reducing Dementia Risk: Never too early, never too late," London, UK, 2023.

[2] P. Hudomiet, M. D. Hurd, and S. Rohwedder, "Dementia prevalence in the United States in 2000 and 2012: Estimates based on a nationally representative study," *The Journals of Gerontology: Series B*, vol. 73, no. suppl_1, pp. S10–S19, 2018.

[3] M. A. Boustani, P. A. Peterson, L. Hanson, R. Harris, and K. N. Lohr, "Screening for dementia in primary care: A summary of the evidence for the U.S. Preventive Services Task Force," *Annals of Internal Medicine*, vol. 138, no. 11, pp. 927–937, 2003.

[4] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," arXiv preprint arXiv:1904.05342, 2019.

[5] J. Cummings, G. Lee, K. Zhong, J. Fonseca, and K. Taghva, "Alzheimer's disease drug development pipeline: 2024," *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, vol. 10, no. 2, p. e12465, 2024.

[6] C. Mao et al., "AD-BERT: Using pre-trained language model to predict the progression from mild cognitive impairment to Alzheimer's disease," *Journal of Biomedical Informatics*, vol. 144, p. 104449, Aug. 2023.

[7] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, 2011.

[8] C. R. Jack Jr. et al., "NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease," *Alzheimer's & Dementia*, vol. 14, no. 4, pp. 535–562, Apr. 2018.

[9] S. M. Lundberg et al., "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomedical Engineering*, vol. 2, no. 10, pp. 749–760, Oct. 2018.

[10] S. Nerella, S. Bandyopadhyay, J. Zhu, M. Changed, and B. Price, "Transformers and large language models in healthcare: A review," *Artificial Intelligence Review*, vol. 57, p. 96, 2024.

[11] G. Zeng et al., "MedDialog: Large-scale medical dialogue datasets," in *Proc. 2020 Conf. Empirical Methods Natural Language Process. (EMNLP)*, Nov. 2020, pp. 9241–9250.

[12] M. Farzan, A. Salimi Nezhad, and S. A. Khaneghahi, "Exploring the efficacy of large language models in systematic reviews: A case study on AI in mental health chatbots," *arXiv preprint* arXiv:2401.13001, 2024.

[13] F. Shen et al., "Classifying the lifestyle status for Alzheimer's disease from clinical notes using deep learning with weak supervision," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 270, Oct. 2022.

[14] A. Turchin, N. Shubina, E. Pendergrass Bowen, M. Pendergrass, and I. S. Kohane, "Comparison of information content of structured and narrative clinical data sources on hypertension-related quality measures," *Journal of the American Medical Informatics Association*, vol. 16, no. 3, pp. 362–370, 2009.

[15] Z. Sadeghi, H. Alizadehsani, J. Ferretti, and A. Galar, "Explainable AI (XAI): A systematic review on taxonomies, opportunities, challenges, and future directions," *Information Fusion*, vol. 100, p. 101943, Dec. 2023.

[16] A. Laguna, F. Argelaguet, A. Lecuyer, and M. Hachet, "ExpLIMEable: Explaining LIME explanations with influence instances," in *CHI Conf. Human Factors Computing Systems*, Apr. 2023, pp. 1–18.

[17] F. Surani, R. Garg, A. P. Balasubramanian, and S. N. Mehta, "An investigation into privacy and security concerns of health chatbots," in *Proc. 2022 CHI Conf. Human Factors Computing Systems*, Apr. 2022, pp. 1–14.

[18] N. Khalid et al., "Privacy-preserving artificial intelligence in healthcare: Techniques and applications," *Computers in Biology and Medicine*, vol. 158, p. 106848, May 2023.

[19] T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol, "Rasa: Open source language understanding and dialogue management," arXiv preprint arXiv:1712.05181, 2017.

[20] Z. S. Nasreddine et al., "The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, Apr. 2005.

[21] Alzheimer's Association, "2024 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 20, no. 5, pp. 3708–3821, May 2024.

[22] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Aug. 2016, pp. 1135–1144.